UNIVERSIDAD DE DEUSTO

Desarrollo de un modelo de toma de decisiones para el posicionamiento activo frente a la aparición de nuevas tecnologías

Decision-making predictive model development for the organization active positioning faced with the appearance of new technologies

Tesis doctoral presentada por LUIS MIGUEL ARIAS
dentro del Programa de Doctorado en INGENIERÍA PARA LA SOCIEDAD DE LA INFORMACIÓN Y DESARROLLO SOSTENIBLE

Dirigida por el Dr. Javier Nieves Acedo, el Dr. Garikoitz Artola Beobide y la Dra. Igone Porto Gómez

Bilbao, junio de 2024

UNIVERSIDAD DE DEUSTO

Desarrollo de un modelo de toma de decisiones para el posicionamiento activo frente a la aparición de nuevas tecnologías

Decision-making predictive model development for the organization active positioning faced with the appearance of new technologies

Tesis doctoral presentada por LUIS MIGUEL ARIAS
dentro del Programa de Doctorado en INGENIERÍA PARA
LA SOCIEDAD DE LA INFORMACIÓN Y DESARROLLO
SOSTENIBLE

Dirigida por el Dr. Javier Nieves Acedo, el Dr. Garikoitz Artola Beobide y
la Dra. Igone Porto Gómez

El doctorando

Los directores

Bilbao, junio de 2024

A mis padres, sin ellos no habría llegado hasta aquí.

Resumen

Vivimos actualmente en un mundo en el que el fenómeno de la globalización ha cambiado el orden económico mundial. Nos enfrentamos además a diferentes e importantes retos como son el cambio climático, el paradigma del cambio de modelo energético y la inclusión social entre otros.

En este contexto, surge como figura clave la innovación. Se trata de un proceso dinámico y polifacético impulsor del progreso, la competitividad y el crecimiento. Es un elemento fundamental para abordar los desafíos de nuestra sociedad en continua evolución, impulsando cambios y mejorando, en definitiva, la calidad de vida de la sociedad. La innovación es un proceso complejo que implica la creación, desarrollo e implementación de ideas, métodos, productos o servicios nuevos o significativamente mejorados. Su propósito se fundamenta en generar valor, resolver problemas, satisfacer necesidades o aprovechar las diferentes oportunidades que aparecen en diversos ámbitos.

Asimismo, podemos observar que el ámbito de aplicación de la innovación no se limita únicamente al desarrollo de productos o servicios, sino que también abarca procesos, modelos de negocio, estrategias organizativas y cambios en la mentalidad o cultura para fomentar la creatividad y la capacidad de adaptación. La manifestación de este proceso puede desarrollarse de forma incremental, impulsando mejoras graduales, o disruptivas, generando cambios radicales capaces de transformar completamente las industrias.

Considerando este escenario, surge la necesidad de introducir el término predicción como la posibilidad de realizar afirmaciones sobre algo concreto con suficiente antelación. En este caso, la predicción de la innovación permite conocer las tendencias que se darán en el futuro, su impacto en la sociedad o en nuestra organización, analizar si nuestro sistema de I+D+i está bien orientado posibilitando la reorientación de la investigación y facilitar la toma de decisiones en la alta dirección de las empresas o instituciones.

Este trabajo de investigación plantea un nuevo enfoque aplicado a la generación de modelos predictivos de innovación, concretamente la creación de nuevas patentes. Se basa en una combinación tanto de artículos científicos como de patentes, susceptibles de análisis y asociados a una determinada tecnología. La solución propuesta, ha sido desarrollada para la tecnología de estampación en caliente y, posteriormente, validada con la tecnología de hierro fundido. Este procedimiento combina métodos estadísticos con técnicas de aprendizaje automático, dando como resultado tanto una red Bayesiana, como un modelo de árbol de decisión que nos ayudan a identificar patrones de comportamiento en cuanto a la producción de propiedad intelectual protegida. En definitiva, al análisis llevado a cabo es capaz de detectar, con alta probabilidad, la proclividad de un investigador a la generación de nuevas patentes.

Abstract

We are now living in a world where the globalisation phenomenon has changed the world economic order. We are also facing other important challenges such as climate change, the shift paradigm in the energy model, social inclusion among others.

In this context, innovation emerges as a key figure. It is a dynamic and multifaceted process that drives progress, competitiveness and growth. It is a fundamental element to facing the challenges of our constantly evolving society, promoting change and, in short enhancing the quality of life of society. Innovation is a complex process that involves the creation, development and implementation of ideas, methods, new or significantly improved products or services. Its purpose is based on generating value, solving problems, satisfying needs or taking advantage of the different opportunities that arise in various fields.

We can also see that the scope of innovation is not only limited to the development of products or services, but also involves processes, business models, organisational strategies and changes in mindset or culture to foster creativity and adaptability. The expression of this process can be incremental, driving gradual improvements, or disruptive, generating radical changes capable of completely transforming industries.

In view of this scenario, it is necessary to introduce the concept of forecasting as the possibility of making statements about something specific sufficiently in advance. In this case, forecasting innovation

allows us to know the trends that will emerge in the future, their impact on society or on our organisation, to analyse whether our R&D system is well oriented, to redirect research if it is necessary and to facilitate decision-making in the senior management of companies or institutions.

This research work proposes a new approach applied to the generation of predictive models of innovation, specifically the creation of new patents. It is based on a combination of both scientific articles and patents, which can be analysed and associated with a given technology. The proposed solution has been developed for hot stamping technology and subsequently validated with cast iron technology. This procedure combines statistical methods with machine learning techniques, resulting in both a Bayesian network and a decision tree model that help us to identify patterns of behaviour in the production of protected intellectual property. In summary, the analysis performed is able to detect, with a high degree of probability, a researcher's propensity to generate new patents.

The proposed solution, developed for hot stamping technology and, subsequently, validated for cast iron technology, combines statistical methods with machine learning techniques, resulting in both a Bayesian network and a decision tree model, which help us to identify patterns of behaviour in the production of protected intellectual property. In summary, the analysis performed is able to detect, with a high degree of probability, a researcher's propensity to generate new patents.

Agradecimientos

Llegados a este punto no queda sino dar las gracias a todas aquellas personas que de una u otra forma han colaborado en la realización de esta tesis.

En primer lugar, no puedo sino comenzar agradeciendo a mi familia por todo el apoyo prestado a lo largo de todo este proceso, en especial a Juana, mi madre, sin cuya ayuda incondicional esto hubiese sido mucho más complicado, a Alex, mi sobrino, que espero que algún día siga estos mismos pasos, a Inés, mi sobrina, que no deja de preguntarme "Tío, cuando vas a terminar" así como a mi hermana, Gema, y a mi cuñado, Javi.

En segundo lugar, no puedo olvidarme de mis amigos que han aguantado mis quejas y mis protestas durante estos años. Mención especial para el Dr. José Luis Arroyo que ha estado conmigo día a día, siempre apoyando y empujando en los momentos más difíciles. La lista puede ser más o menos larga, pero vamos a ello. Gracias a Aída Mª, a Bibi, a Cris, a Esther, a Fernando, a Gerardo, a Gorka, a Inés, a Lauri (toda una vida juntos) a Maite, a Nora, a Patxi, a Pepe, a Rebeca y, por último, pero no por ello menos importante, a Verónica (tú y yo sabemos porque). Gracias a todos por estar ahí.

En tercer lugar, es necesario dar las gracias a mis directores de tesis. Al Dr. Garikoitz Artola Beobide con el que comencé esta aventura con un café y unos *post-it* que, a día de hoy, aún conservo. Gracias por tu apoyo incondicional y constante sin que importase ni el día ni la hora. A la

Dra. Igone Porto Gómez, por su apoyo, consejos, esfuerzo, tiempo, estímulo y, sobre todo, por esa paciencia infinita que ha demostrado conmigo. Al Dr. Javier Nieves Acedo, que mucho antes de convertirse en uno de mis directores comenzó a apoyarme y ayudarme de forma totalmente altruista. Gracias Javi por introducirme en el mundo de la inteligencia artificial y por todo lo que me has enseñado.

En cuarto lugar, quiero agradecer a mis compañeros que se han interesado no solo por mi sino también por este trabajo. Gracias a David G., a Enara, a Iñaki, a Javi L., a María, a Txemari y al Sr. Gorordo que se empeñó en abandonase la programación con VB. Seguramente me dejaré a alguien en el tintero, pero gracias de verdad a todos lo que de una u otra forma me habéis apoyado.

En quinto lugar, he de agradecer a Azterlan, Centro Tecnológico de Investigación en Metalurgia, por todas las facilidades que me han proporcionado para hacer esta investigación posible.

Por último, agradecer a Jesús Mancha, de la empresa Balder, por su apoyo y generosidad al ayudarme con los datos necesarios sobre las patentes para este proyecto. Así mismo, al equipo de PatBase® (https://www.patbase.com) por su ayuda desinteresada para proporcionarme los datos esenciales para esta investigación.

Índice general

INTRODUCCIÓN 1

1.1 Innovación. 3

1.2 Propiedad intelectual. 12
1.2.1 Patentes. 13

1.3 La importancia de la predicción de la innovación. 20
1.3.1 Retos a superar. 22

1.4 Importancia en el campo de la investigación 24

1.5 Solución propuesta. 25
1.5.1 Hipótesis de trabajo. 26
1.5.2 Objetivos. 26
1.5.3 Diseño global del modelo. 27

1.6 Desarrollo de la tesis. 29
1.6.1 Metodología. 29
1.6.2 Métricas utilizadas. 30

1.7 Estructura del documento. 32

MODELOS PREDICTIVOS DE INNOVACIÓN 35

2.1 Fuentes de información. 36
2.1.1 Orígenes de datos inherentes a artículos científicos indexados. 37
2.1.2 Orígenes de datos referentes a patentes. 39

2.2 Técnicas utilizadas en análisis predictivo. 42

2.2.1 Técnicas basadas en metadatos de patentes. 42
2.2.2 Técnicas basadas minería de datos y minería de textos. 45
2.2.2.1 Procesamiento del lenguaje natural. 47
2.2.2.2 Técnicas de minería de textos basadas en reglas. 52
2.2.2.3 Técnicas de minería de textos basadas en análisis semántico. . 56
2.2.2.4 Enfoques de visualización. 70

2.3 Métodos estadísticos. 72

2.4 Inteligencia artificial. 76
2.4.1 Aprendizaje automático. 81
2.4.1.1 Aprendizaje automático supervisado. 84
2.4.1.2 Aprendizaje automático no supervisado. 91
2.4.1.3 Aprendizaje automático semisupervisado o híbrido. 93
2.4.1.4 Aprendizaje por refuerzo. 94

2.5 Sumario. 95

CASO DE ESTUDIO: ESTAMPACIÓN EN CALIENTE 97

3.1 Una breve introducción a la tecnología de estampación en caliente. 98

3.2 Los últimos 10 años de estampación en caliente. 102
3.2.1 Planificando la estrategia. 105
3.2.2 Mostrando los resultados. 106
3.2.3 Estableciendo conclusiones. 116

3.3 Evolución de la tecnología de estampación en caliente. 117
3.3.1 Diseñando mapas científicos. 120
3.3.2 Analizando mapas científicos. 122
3.3.2.1 Analizando el contenido de los artículos publicados. 123
3.3.2.1.1 Primer periodo (1950-2009) 128
3.3.2.1.2 Segundo periodo (2010-2015) 129

3.3.2.1.3 Tercer periodo (2016-2019) 132
3.3.2.2 Analizando el mapa de evolución conceptual. 133
3.3.3 Valorando los resultados. 136

3.4 Una nueva metodología para predecir la innovación a través de las patentes. 138
3.4.1 Adquisición y preprocesamiento de datos. 138
3.4.2 Método de cálculo. 142
3.4.2.1 Discretización de datos. 148
3.4.3 Resultados. 149
3.4.3.1 Red Bayesiana. 149
3.4.3.2 Árbol de decisión. 152
3.4.3.3 Evolución en el tiempo. 158
3.4.4 Discusión y conclusiones. 159

3.5 Sumario. 162

CASO DE ESTUDIO: FUNDICIÓN DE HIERRO 165

4.1 Una breve introducción a la tecnología de fundición de hierro. 167

4.2 Aplicación del nuevo método de predicción de innovación a otras tecnologías. 170
4.2.1 Adquisición y preprocesamiento de datos. 171
4.2.2 Método de cálculo. 173
4.2.2.1 Discretización de datos. 178
4.2.3 Resultados. 178
4.2.3.1 Red Bayesiana. 179
4.2.3.2 Árbol de decisión. 181
4.2.3.3 Evolución en el tiempo. 183
4.2.4 Discusión y conclusiones. 184

4.3 Sumario..... 186

CONCLUSIONES 187

5.1 Resumen y resultados de la investigación. 188

5.2 Validación del modelo. 199

5.3 Limitaciones del modelo. 201

5.4 Aplicaciones de la investigación. 204

5.5 Trabajo futuro. 206

5.6 Reflexión final. 208

ANEXO 1 209

ANEXO 2 225

BIBLIOGRAFÍA 241

Índice de figuras

Figura 1: Estructura de la estrategia Europa 2020 dónde se muestran las 5 áreas prioritarias de actuación cuyo objetivo ha sido asegurar la competitividad europea a nivel mundial en el año 2020.. 8
Figura 2: Estructura del programa Horizon 2020 basado en tres grandes bloques, la excelencia en la ciencia, el liderazgo industrial y los retos de la sociedad. 9
Figura 3: Objetivos de desarrollo sostenible marcados en la Agenda 2030 de la Asamblea General de la Organización de las Naciones Unidas (ONU) en 2015 [ONU15]. 11
Figura 4: Estructura del programa Horizon Europe, herramienta de la cual se dota la CE para fortalecer su base científica y tecnológica, desarrollar soluciones cuyo objetivo es lograr una vida más saludable, impulsar la transformación digital y luchar contra el cambio climático. 12
Figura 5: Esquema de la concesión de una patente entendida como un monopolio temporal otorgado por un estado al inventor o empresa tras su solicitud. 14
Figura 6: Esquema que indica los tiempos y estados en la solicitud de una patente española. Dicha patente puede obtener una extensión territorial país por país. 15
Figura 7: Esquema que indica los tiempos y estados en la extensión de una patente española a una patente europea o EP (de la acepción inglesa European Patent). 17
Figura 8: Esquema que indica los tiempos y estados en la extensión de una patente española a una patente internacional o PCT (de la acepción inglesa Patent Cooperation Treatment). 18

Figura 9: Diseño global del modelo predictivo de innovación planificado, que incluye todos los pasos necesarios desde la captación inicial de datos, hasta la generación de conocimiento y monetización del ciclo. 27
Figura 10: Proceso genérico de un modelo predictivo de innovación dividido en tres fases, (i) selección de fuentes de información y adquisición de datos, (ii) procesado de estos y (iii) análisis de resultados y elaboración de informes orientados a la toma de decisiones. 36
Figura 11: Clasificación de los diferentes tipos de inteligencia artificial existentes en la actualidad. 79
Figura 12: Clasificación esquemática de los diferentes tipos de aprendizaje automático que muestra su encaje en el entorno de la inteligencia artificial y los métodos y modelos utilizados en cada caso. 81
Figura 13: Evolución del uso del acero en componentes para carrocería [CZEAFC21]. 99
Figura 14: Componentes pertenecientes a la carrocería del automóvil fabricador con la tecnología de estampación en caliente [LSDFLW21]. 100
Figura 15: Primer uso de la estampación en caliente en un automóvil de serie. SAAB 9000. Viga de impacto lateral [SHFGTD21]. 101
Figura 16: Número de publicaciones científicas en revistas indexadas por año para el periodo 2009-2019. 107
Figura 17: Número de artículos científicos publicados en las diez revistas indexadas más influyentes para el periodo 2009-2019. 108

Figura 18: Evolución del número de artículos científicos publicados por año en las cinco revistas indexadas más influyentes para el periodo 2009-2019. 108
Figura 19: Número de artículos científicos publicados por los diez autores más relevantes en el periodo 2009-2019. 109
Figura 20: Número de artículos científicos publicados por las diez instituciones/empresas más relevantes en el periodo 2009-2019. .. 110
Figura 21: Número de citaciones por año que han recibido los artículos científicos sobre estampación en caliente durante el periodo 2009-2019.....111
Figura 22: Valor del índice h para el conjunto de artículos científicos sobre estampación en caliente durante el periodo 2009-2019.....111
Figura 23: Número de citaciones por autores de artículos científicos sobre estampación en caliente durante el periodo 2009-2019..... 112
Figura 24: Número de citas por revista que ha publicado artículos científicos sobre estampación en caliente durante el periodo 2009-2019..... 113
Figura 25: Número de publicaciones de artículos científicos por país sobre estampación en caliente durante el periodo 2009-2019..... 113
Figura 26: Red formada por las palabras clave sobre estampación en caliente utilizadas por los autores durante el periodo 2009-2019. .. 114
Figura 27: Red formada por los autores de artículos científicos en revistas indexadas sobre estampación en caliente durante el periodo 2009-2019..... 115
Figura 28: Esquema representativo de un diagrama estratégico y de un gráfico de evolución temática en el que se muestran diferentes tipos de conexiones entre temas..... 119

Figura 29: Metodología empleada en el diseño y creación de mapas científicos. 120
Figura 30: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 1950-2009. 128
Figura 31: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 2010-2015. 130
Figura 32: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 2016-2019. 132
Figura 33: Mapa de evolución conceptual y áreas temáticas para que comprende tres periodos definidos, 1950-2009, 2010-2015 y 2015-2019. 135
Figura 34: Metodología utilizada para procesar el conjunto de datos de artículos científicos y patentes con el objeto de obtener un nuevo conjunto de datos mediante el cálculo de (i) la inversión acumulada, (ii) la inclinación hacia la generación de propiedad intelectual y (iii) la generación real de esta. 143
Figura 35: Modelo de red bayesiana con TAN obtenido para la variable objetivo PASA1. La figura muestra las relaciones entre variables que permiten la generación real de propiedad intelectual, que es la generación de nuevas patentes considerando un solo inventor. 150
Figura 36: Monitores utilizados para comprobar no sólo el comportamiento de la variable objetivo, en este caso, PASA1, sino también grupos de variables no dependientes (ANA y PENA). 151
Figura 37: Modelo de red bayesiana para la variable objetivo PASA2. La figura revela la combinación de variables que generan PI real considerando pares de inventores. 151

Figura 38: Modelo de árbol de decisión para la variable objetivo PASA1. La figura muestra las condiciones en las que se genera la propiedad intelectual real, lo que nos permite esbozar un conjunto de reglas de comportamiento. 153
Figura 39: Modelo de árbol de decisión para la variable objetivo PASA2. La figura muestra las situaciones en las que se genera el PI real que permite establecer un conjunto de reglas de comportamiento 157
Figura 40: Evolución de la precisión para los modelos de red Bayesiana y árbol de decisión durante los 20 años de vigencia de las patentes. 158
Figura 41: Diagrama de fases Fe-C estable y metaestable según Caesar [ICPD19]..... 168
Figura 42: Modelo de red bayesiana obtenido para la variable objetivo PASA1 para la tecnología de hierro fundido. La figura muestra las relaciones entre variables que permiten la generación real de propiedad intelectual para un inventor. 179
Figura 43: Monitores utilizados para comprobar no sólo el comportamiento de la variable objetivo, en este caso, PASA1, sino también grupos de variables no dependientes (ANA y PENA) en el caso de la tecnología de fundición de hierro. 180
Figura 44: Modelo de red bayesiana para la variable objetivo PASA2 para la tecnología de hierro fundido. La figura revela la combinación de variables que generan propiedad intelectual real considerando pares de inventores. 180
Figura 45: Modelo de árbol de decisión para la variable objetivo PASA1. La figura muestra las condiciones en las que se genera la propiedad intelectual real. 181

Figura 46: Evolución de la precisión para los modelos de red Bayesiana y árbol de decisión para la tecnología de fundición de hierro durante los 20 años de vigencia de las patentes. 184

Índice de tablas

Tabla 1: Estados miembros de la Organización Europea de Patentes 17
Tabla 2: Etapas o fases necesarios indicando los plazos aconsejables o necesarios en el proceso de solicitud de una patente. 19
Tabla 3: Resumen de métodos estadísticos aplicados en conjunción con diferentes técnicas de análisis utilizadas en la predicción de la innovación. 73
Tabla 4: Propiedades estadísticas del conjunto de datos resultante del tratamiento de los juegos de datos primarios de artículos científicos y patentes. 144
Tabla 5: Valores resultantes de la discretización y etiquetado para los tres grupos de variables del proceso, ANA (inversión acumulada), PENA (inclinación a la generación de propiedad intelectual) y PASA (generación real de propiedad intelectual). 148
Tabla 6: Escenarios con producción real de propiedad intelectual considerando como variable objetivo la variable PASA1, es decir, la generación de patentes. 155
Tabla 7: Escenarios con producción real de propiedad intelectual considerando como variable objetivo la variable PASA1, es decir, la generación de patentes. 157
Tabla 8: Propiedades estadísticas del conjunto de datos resultante del tratamiento de los juegos de datos primarios de artículos científicos y patentes para el caso de fundición de hierro. 174
Tabla 9: Valores resultantes de la discretización y etiquetado para los tres grupos de variables del proceso, ANA (inversión acumulada),

PENA (inclinación a la generación de propiedad intelectual) y PASA (generación real de propiedad intelectual). 178
Tabla 10: Escenarios con producción real de propiedad intelectual para la tecnología de fundición de hierro considerando como variable objetivo la variable PASA1, es decir, la generación de patentes. 182
Tabla 11: Nº de intervalos de discretización en los que se han dividido las variables ANA, PENA y PASA para los 35 juegos de datos que se han analizado. 210
Tabla 12: Porcentaje de precisión (o acierto), correctitud, obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 213
Tabla 13: Error absoluto medio (Mean Absolute Error, MAE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 215
Tabla 14: Raíz del error cuadrado medio medio (Root Mean Square Error, RMSE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 217
Tabla 15: Precisión obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 219
Tabla 16: Exhaustividad (del vocablo inglés Recall) obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 221
Tabla 17: Valor-F obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente..... 223
Tabla 18: Nº de intervalos de discretización en los que se han dividido las variables ANA, PENA y PASA para los 35 juegos de datos que se han analizado. 226

Tabla 19: Porcentaje de precisión (o acierto), correctitud, obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 229
Tabla 20: Error absoluto medio (Mean Absolute Error, MAE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 231
Tabla 21: Raíz del error cuadrado medio medio (Root Mean Square Error, RMSE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente..... 233
Tabla 22: Precisión obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 235
Tabla 23: Exhaustividad (del vocablo inglés Recall) obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 237
Tabla 24: Valor-F obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente. 239

1

Introducción

La capacidad de predecir o prever el futuro ha sido una búsqueda constante a lo largo de la historia de la humanidad. Desde las antiguas civilizaciones hasta la era moderna, el deseo de entender y anticipar lo que está por venir ha sido una preocupación inherente a nuestra naturaleza. Esta inquietud ha dado lugar a diversas prácticas, desde métodos místicos y supersticiones hasta enfoques científicos y tecnológicos, en un intento por desentrañar los misterios del mañana.

En los albores de la humanidad, la predicción del futuro estaba vinculada a creencias religiosas, a la interpretación de fenómenos naturales y a la observación de patrones en la naturaleza.

Sin embargo, con el desarrollo del pensamiento racional y el surgimiento de la ciencia, la predicción comenzó a basarse en el análisis sistemático de datos y en la construcción de modelos explicativos. La estadística, las teorías probabilísticas y la extrapolación de tendencias se convirtieron en herramientas fundamentales para prever fenómenos naturales, económicos, sociales y científicos.

El advenimiento de la Revolución Industrial marcó un punto de inflexión en la capacidad de predecir el futuro. El avance tecnológico y el acceso a una cantidad sin precedentes de datos llevaron a la creación de modelos matemáticos más sofisticados.

En el actual panorama empresarial y tecnológico, la capacidad de prever tendencias y cambios futuros se ha vuelto fundamental para el desarrollo sostenible de las organizaciones. En este contexto, surge también la innovación como un elemento crucial para mantener la competitividad en un mercado globalizado y dinámico, siendo también el pilar sobre el cual se construyen nuevos productos, servicios y procesos que transforman la forma en que interactuamos con el mundo [RICCA20].

La importancia de la innovación radica en su capacidad no solo para adaptarse a un entorno en constante evolución, sino también para liderar y moldear ese mismo entorno. Sin embargo, la naturaleza intrínsecamente incierta del futuro plantea un desafío: ¿cómo pueden las empresas y los investigadores anticiparse a las demandas y cambios que aún no han surgido?

En este sentido, la predicción se revela como una herramienta invaluable. La capacidad de prever tendencias, identificar oportunidades emergentes y entender los posibles escenarios futuros se ha convertido en una habilidad esencial para los líderes empresariales, investigadores y estrategas.

Con el desarrollo de la informática y, más recientemente, con el auge de la inteligencia artificial y el aprendizaje automático se ha revolucionado la forma en que abordamos la predicción. Los modelos predictivos, en particular, han desempeñado un papel fundamental en esta búsqueda por anticipar el curso de la innovación. La evolución de estos modelos ha sido constante, impulsada por la creciente disponibilidad de datos, la capacidad computacional y la sofisticación de los algoritmos.

Hoy en día, con la vasta cantidad de información generada constantemente, los modelos predictivos han alcanzado niveles de complejidad y precisión impensables décadas atrás. Desde la predicción del clima y la evolución económica hasta la anticipación de tendencias en el mercado y el comportamiento humano, los modelos actuales son capaces de procesar grandes volúmenes de datos y encontrar patrones ocultos, brindando así pronósticos más certeros [INNSM20].

Sin embargo, a pesar de todos estos avances, la predicción del futuro sigue siendo un desafío complejo y lleno de incertidumbre. El futuro está influido por una

multitud de variables interconectadas y dinámicas, y factores imprevistos pueden alterar drásticamente cualquier predicción. La evolución en el tiempo de los métodos de predicción refleja el constante esfuerzo humano por comprender y controlar lo desconocido, pero también subraya la necesidad de reconocer los límites de nuestra capacidad para prever el futuro.

Esta tesis doctoral se adentra en el campo de la predicción de la innovación. A lo largo de estas páginas, se analizará la importancia de la innovación en el contexto actual, abordando también la predicción de ésta mediante el amplio abanico de técnicas empleadas para dicho fin.

El resto del capítulo está organizado de la siguiente forma. La sección 1.1 describe la importancia que tiene actualmente la innovación en nuestra sociedad, no solo a nivel europeo (avalado por la Unión Europea), sino a nivel mundial (impulsado por la Organización de las Naciones Unidas). La sección 1.2 describe el concepto de propiedad intelectual analizando en profundidad el mundo de las patentes. La sección 1.3 analiza la importancia de la predicción de la innovación identificando aspectos como los retos a superar, la hipótesis de trabajo, los objetivos y la solución propuesta. La sección 1.4 establece la hipótesis general de esta tesis doctoral y enumera los objetivos a cumplir en esta investigación. La sección 1.5 enumera la hipótesis de trabajo, objetivos y diseño global del modelo. La sección 1.6 describe la metodología de investigación utilizada y las métricas empleadas. Por último, la sección 1.7 detalla cómo se ha estructurado esta tesis doctoral.

1.1 Innovación.

En un intervalo de tiempo inusualmente corto, el rápido desarrollo del fenómeno de la globalización ha cambiado el orden económico mundial previamente establecido, haciendo surgir en este nuevo entorno tanto nuevas oportunidades como nuevos retos. En este actual escenario, Europa no puede competir a menos que sea más ingeniosa, reaccione mejor a las necesidades y preferencias de los consumidores e innove más [COTEC06].

En la primera década del siglo XXI los ciudadanos europeos empezaron a estar preocupados por temas que consideraron importantes como, por ejemplo:

- el cambio climático,
- la reducción del uso de los recursos no renovables,
- el cambio demográfico y,
- las nuevas necesidades de seguridad.

Dichas reflexiones posibilitaron la demanda de una acción colectiva cuyo objeto era salvaguardar el estilo de vida europeo combinando, por un lado, la prosperidad económica, y por otro, la solidaridad. Estas legítimas preocupaciones se convirtieron en una oportunidad para mejorar la competitividad económica global de Europa.

La Unión Europea ha tenido, y aún tiene, un extraordinario potencial en el ámbito de la innovación. No obstante, tanto las condiciones marco, como la permanente infravaloración de la innovación como valor importante en la sociedad, que se dieron en los primeros años del presente siglo, contribuyeron a la no explotación de esta.

Sin embargo, en el año 2008, la crisis económica mundial dañó enormemente las economías de los estados miembros. Como consecuencia, las pequeñas y medianas empresas (PYMEs) tuvieron importantes dificultades de financiación y se produjeron altas tasas de desempleo en diferentes países. Teniendo en cuenta todos estos hechos, la innovación fue considerada como la clave para luchar contra aquella fuerte recesión económica, ayudando a las empresas a crecer y creando puestos de trabajo para contrarrestar los despidos. Para promover la innovación en la Unión Europea de la manera más eficaz posible, el apoyo a la misma se basó en una política claramente justificada y se demostró la capacidad de marcar una diferencia real.

Por estos motivos, la Unión Europea, a lo largo de su historia, ha ido tomando diferentes medidas en un importante intento de modernizar su economía. La Estrategia de Lisboa para el Crecimiento y el Empleo [TLS22], establecida para el periodo de 2000 a 2010, estableció un conjunto completo de normas, procedimientos y reformas destinadas a crear un marco económico en Europa cercano a la innovación. Los acuerdos dentro del nuevo marco financiero (el 7º Programa Marco de Investigación y Desarrollo y el Programa Marco de

Competitividad e Innovación) son ejemplos importantes del nuevo camino que se emprendió en la dirección de la innovación proporcionando así mismo el necesario apoyo económico.

Dentro del marco de la Estrategia de Lisboa para el Crecimiento y el Empleo [EDL24], la mayoría de los Estados miembros realizaron en aquellos años un importante esfuerzo con el objetivo de mejorar sus mecanismos de apoyo a la innovación, invirtiendo en investigación y aplicando nuevos o mejores instrumentos de apoyo a las PYMEs innovadoras. Los gráficos de tendencia de las políticas de innovación de aquellos años [NTPI2011] identificaban más de 1000 medidas horizontales y específicas de apoyo a la innovación en toda Europa, que apoyan la transferencia de tecnología, la incubación y el acceso a la financiación entre otras.

La crisis económica mundial que se desató en 2008 debido al colapso de la burbuja inmobiliaria en los Estados Unidos en el año 2006, provocó, aproximadamente en octubre de 2007, la llamada crisis de las hipotecas *subprime*. En consecuencia, se ejerció una mayor presión sobre los presupuestos públicos, por lo que el apoyo a la innovación tuvo que demostrar un impacto económico positivo para justificar, tanto una mayor financiación, como una total transparencia.

En los siguientes años, Europa se enfrentó a un periodo de transformación que intentó recuperar los años perdidos de progreso económico y social y reforzar las debilidades estructurales existentes en la economía europea. Se necesitó una estrategia para convertir a la Unión Europea en una economía inteligente, sostenible e integradora para poder ofrecer altos niveles de empleo, productividad y cohesión social. Esta fue la estrategia Europa 2020 [E202022].

Los principales objetivos considerados por la UE fueron los siguientes.

- El 75% de la población de entre 20 y 64 años debería estar empleada.
- El 3% del PIB de la UE debe invertirse en I+D.
- Los objetivos climáticos/energéticos “20/20/20”, es decir, reducir al menos en un 20% las emisiones de gases de efecto invernadero, aumentando el porcentaje de las fuentes de energía renovables en

nuestro consumo final de energía hasta un 20% y en un 20% la eficacia energética, deberían ser cumplidos (incluyendo un aumento al 30% de la reducción de emisiones siempre y cuando se den las condiciones adecuadas).

- El porcentaje de abandono escolar prematuro debería ser inferior al 10% y al menos un 40% de la generación más joven debería poseer un título universitario.
- Al menos 20 millones de personas deberían quedar fuera del umbral de la pobreza.

Una vez definido el ámbito de actuación, Europa 2020 estableció tres ejes estratégicos que se reforzaban entre ellos.

- **Crecimiento inteligente**: desarrollar una economía basada en el conocimiento y la innovación.
- **Crecimiento sostenible**: promover una economía más eficiente en el uso de los recursos, más ecológica y más competitiva.
- **Crecimiento inclusivo**: fomento de una economía con alto nivel de empleo que aporte tanto cohesión social como territorial.

Estos tres focos estratégicos fueron divididos en siete iniciativas emblemáticas para impulsar los avances que habían de producirse en todos y cada uno de los temas considerados prioritarios.

- Unión por la innovación para mejorar las condiciones marco, así como el acceso a la financiación de la investigación y la innovación, a fin de garantizar que las ideas innovadoras pudieran convertirse en productos y servicios que generasen crecimiento y empleo [UPLI22].
- Juventud en movimiento cuyo objeto era mejorar el rendimiento de los sistemas educativos y facilitar la incorporación de los jóvenes al mercado laboral [JEM22].
- Una agenda digital para Europa que facilitase y acelerase el despliegue del Internet de alta velocidad y aprovechase las ventajas de un mercado único digital tanto para los hogares como para las empresas [UADE22].

- Una Europa que utilice eficazmente los recursos que ayudase a disociar el crecimiento económico del uso de los recursos, apoyase el cambio hacia una economía con bajas emisiones de carbono, aumentase el uso de fuentes de energía renovables, modernizase el sector del transporte y promoviese la eficiencia energética [EUER22].
- Una política industrial para la era de la globalización que mejorase el entorno empresarial, especialmente para las PYMEs, y apoyase el desarrollo de una base industrial fuerte y sostenible capaz de competir a nivel mundial [PIIG22].
- Una agenda de nuevas habilidades y empleos que modernizase los mercados de trabajo y capacitase a las personas mediante el desarrollo de sus habilidades a lo largo del ciclo de vida, con el fin de aumentar la participación laboral y adecuar mejor la oferta y la demanda de trabajo, incluso mediante la movilidad laboral [ANCE22].
- Una plataforma europea contra la pobreza que garantizase la cohesión social y territorial de manera que los beneficios del crecimiento y el empleo se repartiesen ampliamente y que las personas en situación de pobreza y exclusión social pudiesen vivir con dignidad y participar activamente en la sociedad [PECP22].

En la siguiente figura, **Figura 1**, se presenta la estructura de la estrategia Europa 2020 por la cual se definieron cinco áreas prioritarias de actuación cuyo objetivo no era otro sino asegurar la competitividad europea a nivel mundial en 2020: empleo, innovación, educación, inclusión social y clima/energía [GBH202022].

Como puede observarse también en la **Figura 1**, bajo el epígrafe de Unión por la Innovación se han definido una serie de iniciativas de ámbito europeo denominadas Asociaciones de Innovación Europeas (EIP, acrónimo extraído del nombre inglés *European Innovation Partnerships*) que constituyeron un mecanismo de participación y gobernanza cuyo objetivo era movilizar a los actores participantes en todo el ciclo de la innovación, es decir, administraciones públicas, industria, comunidad científica y sociedad civil.

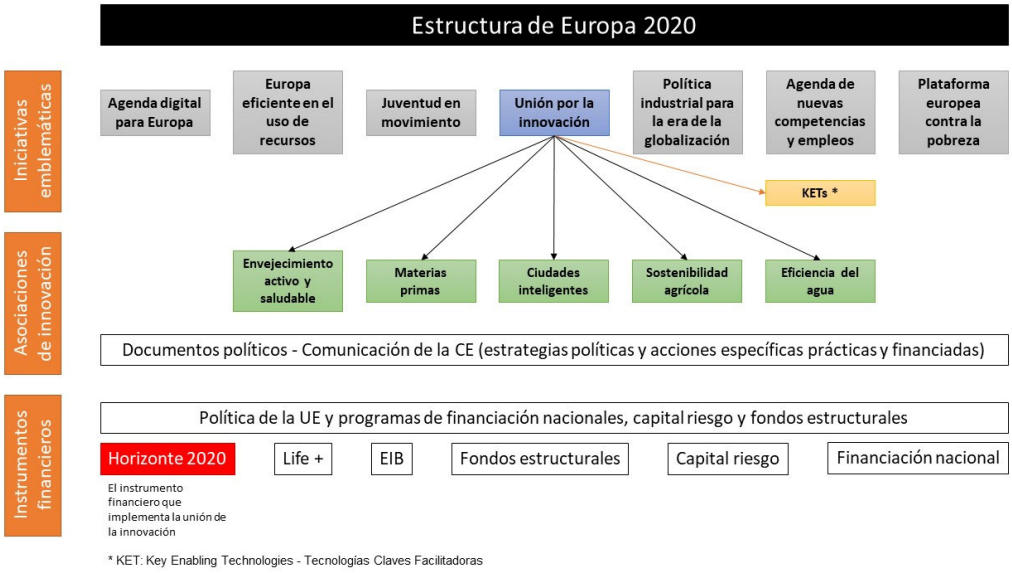

Figura 1: Estructura de la estrategia Europa 2020 dónde se muestran las 5 áreas prioritarias de actuación cuyo objetivo ha sido asegurar la competitividad europea a nivel mundial en el año 2020.

Con el planteamiento realizado, se articuló Horizon 2020 como instrumento financiero, durante el periodo 2014-2020, para implementar el apartado correspondiente Unión por la innovación.

La **Figura 2** muestra la estructura que se estableció para Horizon 2020, que se basó en tres grandes bloques, la excelencia en la ciencia, el liderazgo industrial y los retos de la sociedad. Así, el presupuesto definitivo del programa Horizon 2020 ascendió a 70.000 millones de euros repartidos de la siguiente forma [AH20AIF22].

- Excelencia en la ciencia 31,73%
- Liderazgo industrial..... 22,09%
- Retos de la sociedad..... 38,53%
- Otros.....7,65%

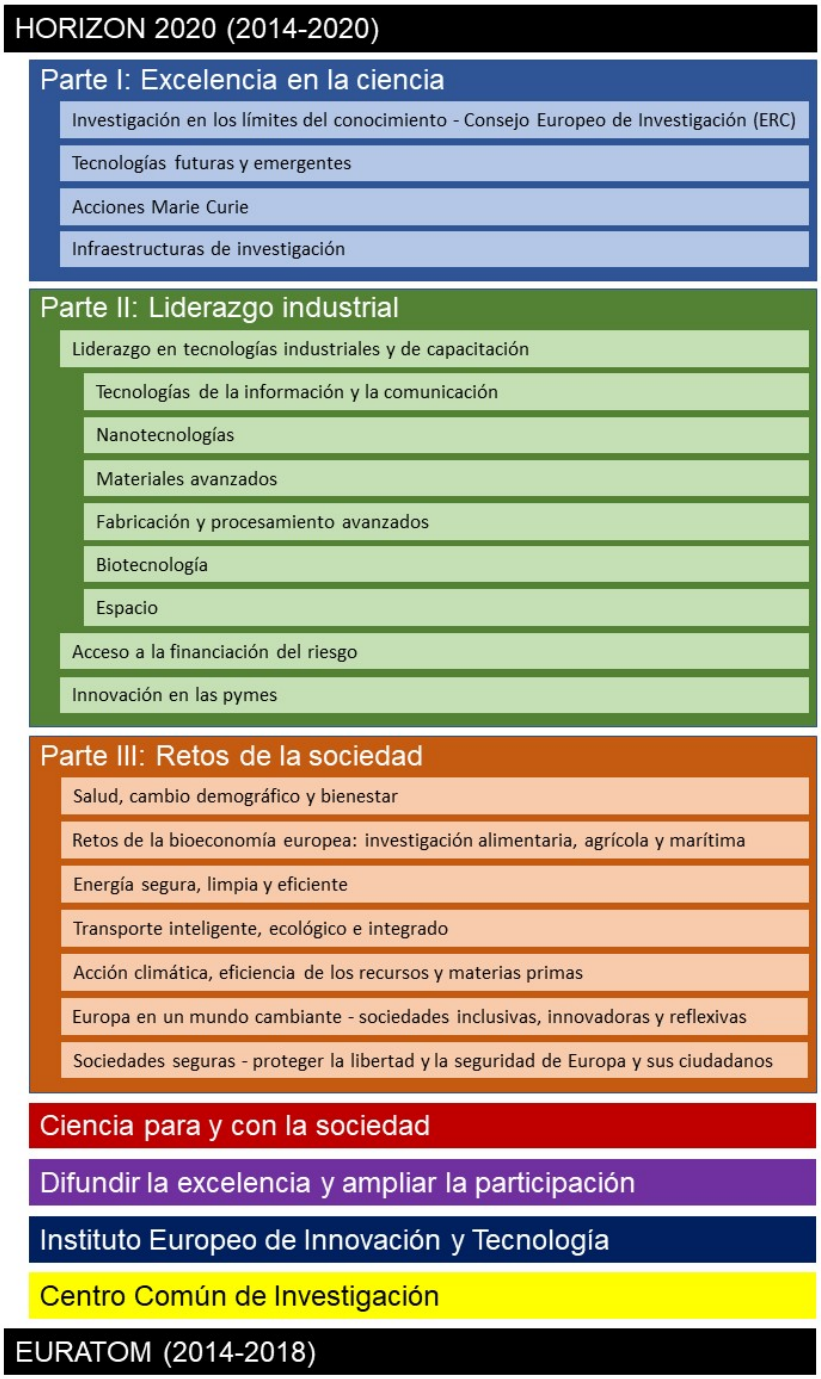

Figura 2: Estructura del programa Horizon 2020 basado en tres grandes bloques, la excelencia en la ciencia, el liderazgo industrial y los retos de la sociedad.

De forma simultánea a estos Programas Marco para acciones de investigación, desarrollo tecnológico y demostración - Horizonte 2020, (del vocablo inglés *The*

Framework Programmes for research, technological development and demonstration activities – Horizon 2020), se desarrolló otro programa: el Programa Marco de la Comunidad Europea de la Energía Atómica (Euratom) para acciones de investigación y formación en materia nuclear (del inglés *The Framework Programme of the European Atomic Energy Community (Euratom) for nuclear research and training activities*). Ambos estuvieron regidos no solo por distintos instrumentos jurídicos, sino por diferentes tratados, habiendo sido desarrollados en distintos períodos presupuestarios.

El Programa Marco Euraton, concebido como programa complementario al Horizon 2020, estaba dotado con una financiación de 1.603 millones de euros repartidos según el siguiente esquema [CCGPE22].

- Acciones indirectas DG-RTD
  - Energía de fusión 45,42%
  - Fisión nuclear, seguridad y protección contra las radiaciones 19,68%
- Acciones directas JRC
  - Seguridad nuclear, protección y control de seguridad.... 34,90%

En septiembre de 2015, la Asamblea General de la Organización de las Naciones Unidas (ONU) adoptó la Agenda 2030 para el desarrollo sostenible. Dicha Agenda 2030 plantea 17 objetivos, véase **Figura 3**, con 169 metas de carácter integrado e indivisible que abarcan las esferas económica, social y ambiental.

En línea con esta estrategia de la ONU, que da libertad a los países miembros para adaptarla a sus propias circunstancias, se enmarca el programa de investigación e innovación de la Unión Europea (UE) (2021-2027) conocido como *Horizon Europe* (HE) [HE24].

Esta es la principal herramienta de la cual se dota la Comunidad Europea (CE) para fortalecer su base científica y tecnológica, desarrollar soluciones cuyo objetivo es lograr una vida más saludable, impulsar la transformación digital y luchar contra el cambio climático.

Figura 3: Objetivos de desarrollo sostenible marcados en la Agenda 2030 de la Asamblea General de la Organización de las Naciones Unidas (ONU) en 2015 [ONU15].

Una vez más, la investigación y la innovación proporcionan nuevos conocimientos y soluciones para superar nuestros retos sociales, ecológicos y económicos. HE, véase **Figura 4**, ayuda a los investigadores y a los innovadores de alto nivel a desarrollar y a desplegar sus ideas, reúne a los mejores talentos y los equipa con infraestructuras de investigación de primer orden y, además, apoya la innovación y ayuda a crear nuevos servicios y mercados [ECDGRI21].

El presupuesto de este programa HE se estima en unos 95.500 millones de euros para el periodo 2021-2027. Se incluyen en esta cantidad 5.400 millones de euros para impulsar la recuperación europea, haciendo que esta sea más resistente en el futuro. Se ha dotado a esta iniciativa con un refuerzo adicional de 4.600 millones de euros, lo cual hace que se sitúe en el entorno de un total 100.000 millones de euros.

Figura 4: Estructura del programa Horizon Europe, herramienta de la cual se dota la CE para fortalecer su base científica y tecnológica, desarrollar soluciones cuyo objetivo es lograr una vida más saludable, impulsar la transformación digital y luchar contra el cambio climático.

Se ha abierto la participación de terceros países ajenos a la UE dentro del programa HE. Es necesario que estos países compartan valores comunes con la UE y demuestren, así mismo, una buena capacidad en ciencia, tecnología e innovación. Japón es un país candidato basándose en el excelente nivel de cooperación en materia de ciencia y tecnología ya demostrado en programas anteriores, con bastante éxito, como el *Horizon 2020*. Se están llevando a cabo también conversaciones exploratorias con Canadá y Nueva Zelanda, así como con la República de Corea [EUJHE22].

1.2 Propiedad intelectual.

La propiedad intelectual (a partir de ahora PI) emerge como uno de los pilares fundamentales en la era moderna, donde la creatividad, la innovación y el conocimiento desempeñan roles centrales en el progreso y desarrollo de las sociedades. En un mundo impulsado por avances tecnológicos vertiginosos y un intercambio globalizado de ideas, el reconocimiento y la protección de la propiedad intelectual han cobrado una importancia sin precedentes.

Es por ello que la PI abarca un vasto espectro de creaciones humanas [PBPI16] [PIIC22], desde invenciones tecnológicas y obras artísticas, hasta marcas comerciales y diseños industriales. Se convierte en un valioso instrumento que resguarda los derechos y el reconocimiento de los creadores, permitiéndoles beneficiarse justamente de sus creaciones y fomentando así un entorno propicio para la innovación continua.

De esta forma, el concepto de PI se fundamenta en la noción de que las ideas y la creatividad merecen protección y reconocimiento legal, equiparándolas en muchos aspectos a la propiedad física. Esta protección no solo estimula la creatividad y el pensamiento innovador, sino que también promueve la difusión del conocimiento al garantizar, tanto beneficios, como reconocimiento a aquellos que contribuyen con nuevos descubrimientos, invenciones y expresiones artísticas.

Debe señalarse que la evolución de la PI a lo largo del tiempo refleja la complejidad de los desafíos que enfrentan las sociedades contemporáneas. Desde las primeras leyes de patentes y derechos de autor hasta los tratados internacionales y las regulaciones actuales [ADPIC22] [BLD14] [USPN00] [QEPI20] [UIP16], se ha buscado constantemente equilibrar la protección de los creadores con el acceso público al conocimiento y la información.

Cabe considerar, por otra parte, que en un mundo cada vez más interconectado, donde la información y las ideas fluyen a través de fronteras sin restricciones, la propiedad intelectual se convierte en un elemento crucial para asegurar la equidad, la justicia y el avance continuo de la sociedad.

Por su parte, la legislación protege la PI, por ejemplo, mediante las patentes, el derecho de autor y las marcas. Se permite así la obtención de un reconocimiento o unos beneficios por las invenciones o creaciones realizadas. Al equilibrar el interés de los innovadores y el interés público, el sistema de PI procura fomentar un entorno propicio para que prosperen la creatividad y la innovación [QELPI22].

1.2.1 Patentes.

La historia de las patentes se remonta a civilizaciones antiguas que reconocían y protegían las innovaciones. Sin embargo, fue en el Renacimiento y la Revolución

Industrial cuando las legislaciones específicas para la protección de invenciones comenzaron a desarrollarse. A lo largo del tiempo, el concepto de las patentes ha evolucionado, abarcando una amplia gama de campos que van desde la ingeniería y la biotecnología hasta el software y los productos farmacéuticos.

Relacionado con la PI, las patentes se erigen como un elemento básico de protección de las invenciones, promoviendo la innovación y fomentando el progreso tecnológico. En un mundo donde la competencia y el desarrollo tecnológico son motores clave del crecimiento económico, las patentes juegan un papel crucial al reconocer y salvaguardar los logros creativos e inventivos de individuos y empresas.

Una patente, tal y como se muestra en la **Figura 5**, es un título que reconoce el derecho a explotar en exclusiva la invención patentada, impidiendo a otros su fabricación, venta o utilización sin consentimiento del titular. Como contrapartida, la patente se pone a disposición del público para general conocimiento. El monopolio sobre la patente es concedido por un máximo de 20 años [PIGBP03].

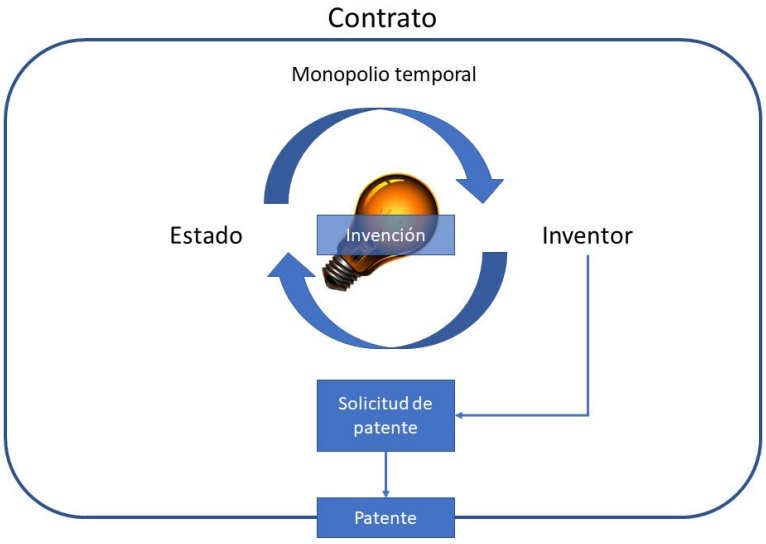

Figura 5: Esquema de la concesión de una patente entendida como un monopolio temporal otorgado por un estado al inventor o empresa tras su solicitud.

La obtención de una patente puede hacerse por la vía española, europea o internacional. En el caso de la vía nacional se realiza a través de la Oficina Española de Patentes y Marcas [OEPM22], pudiendo ir a los centros regionales de propiedad industrial o a las oficinas de correos para poder obtener la documentación. En la **Figura 6** se esquematiza la solicitud de una patente española marcando los tiempos y estados de la misma. Dicha patente es susceptible de ser ampliada a otros países.

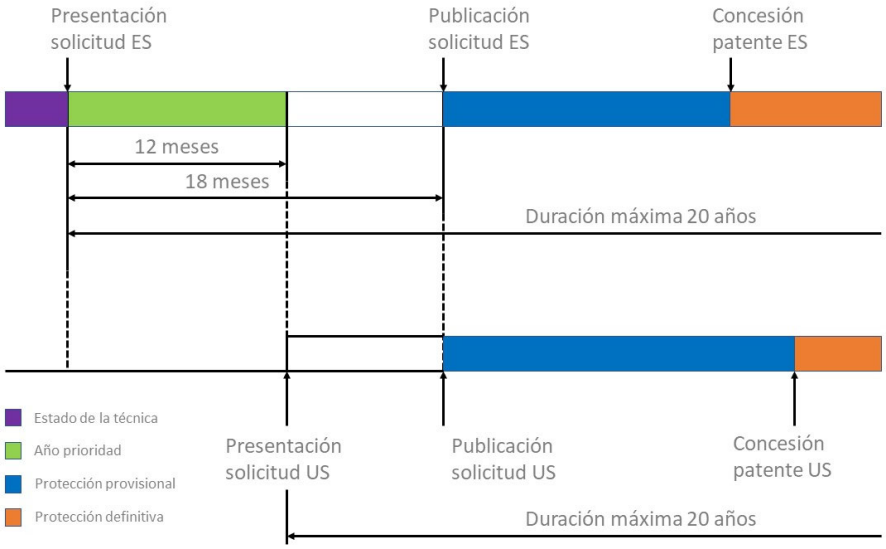

Figura 6: Esquema que indica los tiempos y estados en la solicitud de una patente española (ES). Dicha patente puede obtener una extensión territorial país por país como Estados Unidos (US).

La solicitud de una patente es un procedimiento jurídico regulado por plazos estrictos y generalmente inamovibles. Para optimizar la probabilidad de obtener una buena patente, se deberá:

- Estudiar el procedimiento de solicitud en detalle.
- Intentar no solicitarla con prisas, sino de una manera estratégica, en el momento y por las razones que mejor convengan a sus planes de explotación.
- Hacer uso de un abogado de patentes para evitar el riesgo de cometer errores.

Con respecto a la solicitud de una patente ante una oficina de PI nacional es, en líneas generales, un procedimiento que ha de presentarse en el idioma local [PTSP12] e incluye los siguientes puntos:

1. Descripción de la invención.
   1.1. Sector de la técnica.
   1.2. Estado anterior de la técnica.
   1.3. Exposición de la invención.
   1.4. Breve descripción de las figuras.
   1.5. Exposición detallada de la invención.
2. Reivindicaciones: determinan el alcance de la protección.
   2.1. Normalmente hay una reivindicación principal y varias.
   2.2. reivindicaciones dependientes.
3. Figuras.

Hay que tener en cuenta que:

- Cuando se presenta una patente en un país, el solicitante dispone de un año de prioridad para solicitar protección en otros países.
- Las solicitudes se publican a los 18 meses, tal y como se muestra en la Figura 6, de la fecha de prioridad, confiriendo una protección provisional.
- El procedimiento de concesión es independiente para cada solicitud.
- La protección definitiva se obtiene con la publicación de la patente concedida.

La solicitud de una Patente Europea se realiza al amparo del Convenio sobre la Patente Europea (CPE) [HAEP22]. Constituye un único procedimiento de concesión para todos los países miembros del CPE [MEPO22]. Una vez concedida, la patente europea se valida en los países del CPE elegidos por el titular.

La Oficina Europea de Patentes (OEP) se encarga de la tramitación de las solicitudes de patentes europeas. Dichas solicitudes son concedidas con arreglo a un derecho único, esto es, unos requisitos de patentabilidad uniformes. La **Figura 7** muestra los tiempos necesarios y los estados por los que pasa una patente española al ser extendida a una patente europea.

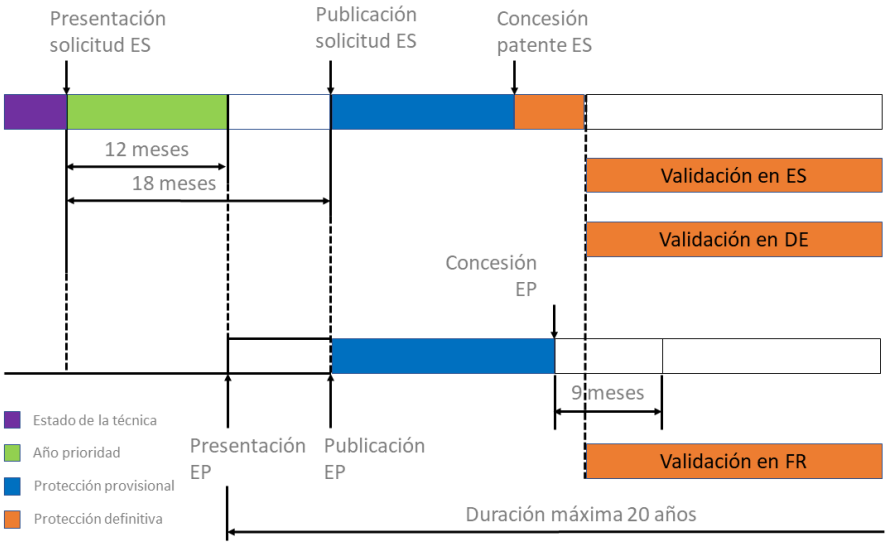

Figura 7: Esquema que indica los tiempos y estados en la extensión de una patente española a una patente europea o EP (de la acepción inglesa *European Patent*).

La siguiente tabla, **Tabla 1**, muestra la lista de los 39 estados miembros de la Organización Europea de Patentes. Los 7 primeros países (léase la tabla de izquierda a derecha y de arriba abajo) se incorporaron en 1977, siendo el último, Montenegro, que se incorporó en 2022.

Tabla 1: Estados miembros de la Organización Europea de Patentes

Bélgica	Alemania	Francia	Luxemburgo
Holanda	Suiza	Reino Unido	Suecia
Italia	Austria	Liechtenstein	Grecia
España	Dinamarca	Mónaco	Portugal
Irlanda	Finlandia	Chipre	Turquía
Bulgaria	República Checa	Estonia	Eslovaquia
Eslovenia	Hungría	Rumanía	Polonia
Islandia	Lituania	Latvia	Malta
Croacia	Noruega	Macedonia del norte	San Marino
Albania	Serbia	Montenegro	

Presentar solicitudes internacionales al amparo del Tratado de Cooperación en materia de patentes (PCT, por sus siglas en inglés) supone un único procedimiento para las fases 1 a 4, pero transcurridos 30 meses desde la presentación de la solicitud, se pasa a las fases 5 y 6 en cada una de las oficinas nacionales o regionales en las que desee obtener la protección [PCTSIP22]. Una única solicitud equivale a una solicitud en cada uno de los países miembros del PCT [LPCTCS22]. No representa un procedimiento de concesión, la solicitud internacional se acaba convirtiendo en solicitudes nacionales. La **Figura 8** especifica los tiempos y las fases requeridas para llevar a cabo la extensión de una patente española a PCT.

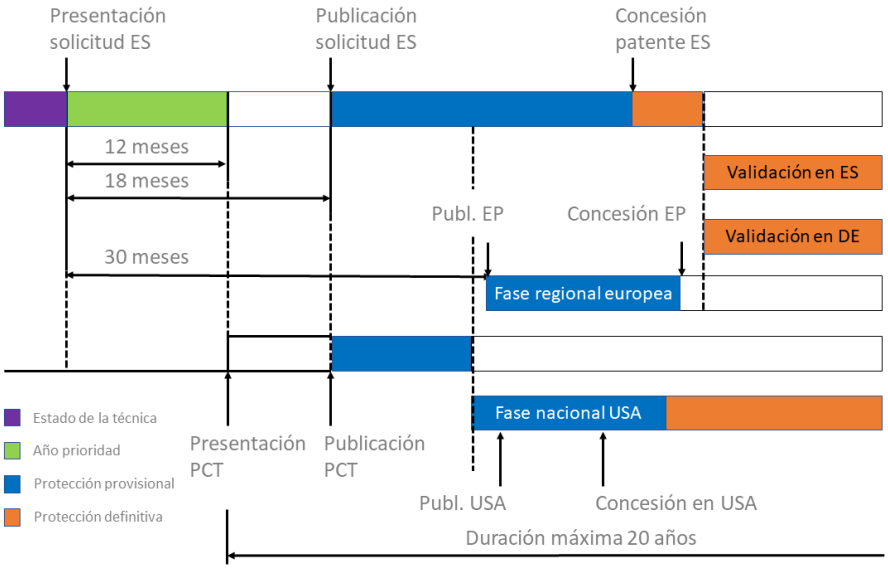

Figura 8: Esquema que indica los tiempos y estados en la extensión de una patente española a una patente internacional o PCT (de la acepción inglesa Patent Cooperation Treatment).

Si bien, como se ha establecido, los procesos de solicitud de patentes son similares, estos difieren en función del ámbito territorial en el que van a ser efectivas, es decir, si vamos a solicitar una patente española, una patente europea o una PCT. En la siguiente tabla, **Tabla 2**, y a modo de resumen, se muestran las diferentes etapas y los plazos de tiempo necesarios desde que se toma la decisión de solicitar una patente hasta que esta es concedida [EPI22].

Tabla 2: Etapas o fases necesarios indicando los plazos aconsejables o necesarios en el proceso de solicitud de una patente.

Etapa	Plazo
Evaluación inicial: Póngase en contacto con su agente de patentes tan pronto como haya desarrollado una idea interesante. Decida si se solicita una patente. Si el tiempo lo permite, encargue una búsqueda provisional para aumentar su conocimiento de la tecnología de referencia.	N/A
Primera presentación: Presentación de la primera solicitud a la oficina de patentes, la rapidez es esencial, estableciendo una fecha de prioridad.	Lo más pronto posible, la rapidez es esencial.
Presentación en el extranjero: Una vez redefinida y reforzada la invención y las reivindicaciones de la solicitud, presentar la solicitud final a la oficina de patentes. Presentar solicitudes en países adicionales utilizando la ruta nacional o internacional. La OEP cubre muchos países europeos con una sola presentación.	12 meses después de la primera presentación.
Publicación: La oficina de patentes publica de forma completa la solicitud en la "Publicación A"	18 meses después de la primera presentación.
Tramitación: Argumentar los méritos de su solicitud ante el examinador de patentes. Si la oficina de patentes le concede su solicitud, emitirá la "Publicación B". Si su solicitud es rechazada, usted tiene la posibilidad de recurrir.	Puede durar varios años después de la publicación.
Oposición: Cualquier tercero puede cuestionar su patente utilizando el proceso de oposición de la OEP. Cualquiera de las partes puede recurrir si la decisión es contraria a sus intereses.	Dentro de los 9 meses siguientes a la concesión.

Etapa	Plazo
Confirmación de la concesión: Según la decisión sobre la apelación referente a la oposición, la concesión queda confirmada o revocada. No se pueden hacer otros recursos a través de la OEP.	Puede durar varios años después del final del periodo de oposición (9 meses).

1.3 La importancia de la predicción de la innovación.

En la sección 1.1 se ha descrito el concepto de innovación y la importancia que esta tiene en el entorno socioeconómico actual.

En cuanto al término innovación, el manual de Oslo [OM18] es un documento clave, en el cual se hace referencia tanto a la definición como a la medida de esta. Este término, recogido en dicho documento, puede significar tanto una actividad como su resultado. Se ofrecen aquí ambas definiciones. La definición general de innovación es la siguiente:

> *Una innovación es un producto o proceso nuevo o mejorado (o una combinación de ellos) que difiere significativamente de los productos o procesos anteriores de la unidad y que ha sido puesto a disposición de los usuarios potenciales (producto) o puesto en uso por la unidad (proceso).*

Esta definición utiliza el término genérico unidad para describir al agente responsable de las innovaciones. Concretamente, se refiere a cualquier unidad institucional de cualquier sector, incluidos los hogares y sus miembros individuales.

En el marco de los modelos de gestión, de necesaria aplicación por parte de las organizaciones, con el objeto de dirigir los entornos de I+D+i, surge el concepto de predicción de la innovación. Podemos encontrar artículos como *The Challenges While Measuring Enterprise Innovative Activities - the Case from a Developing Country* [CMEIA18] que nos presenta la evaluación de las actividades de innovación por medio de un cuestionario, a cuyos resultados se aplica el test Q de

Cochran y el test de McNemar, test estadísticos que muestran las actividades de innovación en las que se centran las empresas, así como aquellas otras que se consideran pertinentes, pero no se utilizan. Así mismo, el artículo *Knowledge trends in the subfields of manufacturing engineering at the platform of ISO/IEC standardization* [MTSME13], plantea la predicción de la innovación a partir de encuestas.

La predicción suele interpretarse como la capacidad de ver el futuro lejano. Se define, así mismo, como la posibilidad de realizar afirmaciones sobre algo concreto con suficiente antelación, especialmente si se realiza sobre la base de conocimientos específicos [BFPIDI12].

Si aplicamos estos conceptos a las empresas podemos afirmar que [IBC08],

> *una ejecución exitosa, a su vez, requiere de la predicción; toda industria y gobierno se enfrentan a un mundo rápidamente cambiante en el que es difícil distinguir las tendencias importantes a largo plazo del ruido.*

El documento *New Approaches to Predicting Cluster* [NAPC18], que también plantea la predicción de la innovación, en concreto analiza la creación de nuevos clústeres a partir de masas críticas por medio de un estudio realizado a partir de la base de datos de I+D del gobierno alemán con más de 110.000 proyectos de I+D. Podría realizarse también un planteamiento similar con la base de datos CORDIS de proyectos europeos.

Artículos como *A Predictive Model of Technology Transfer Using Patent Analysis* [PMTT15] tratan de predecir la innovación a partir de un análisis de patentes. En este caso lo que se pretende es determinar aquellas patentes que son realmente significativas, conocidas en inglés como *hot patents*, debido a la gran cantidad de patentes que se generan anualmente.

La importancia de la predicción de la innovación radica en su capacidad para ofrecer perspectivas fundamentales que orienten la toma de decisiones estratégicas en empresas, instituciones académicas y entornos de investigación. La predicción efectiva de la innovación no solo proporciona una visión anticipada de los posibles

desarrollos futuros, sino que también permite una planificación más precisa, identificación de oportunidades y mitigación de riesgos en el ámbito de la innovación.

En un contexto empresarial altamente competitivo, como el actual, donde la innovación se considera un motor clave del crecimiento y la sostenibilidad, poder anticipar tendencias, cambios en la demanda del mercado, avances tecnológicos o incluso las necesidades emergentes de los consumidores, brinda a las organizaciones una ventaja estratégica significativa. La capacidad para prever innovaciones potenciales no solo facilita la creación de productos y servicios disruptivos, sino que también permite a las empresas adaptarse proactivamente a un entorno cambiante.

En el ámbito académico y de investigación, la predicción de la innovación ofrece una valiosa oportunidad para identificar áreas de estudio prometedoras y direccionar recursos hacia investigaciones que puedan tener un impacto significativo en el futuro. Permite a los investigadores y académicos enfocar sus esfuerzos en áreas que pueden generar avances revolucionarios o soluciones innovadoras a problemas complejos.

Además, la predicción de la innovación puede tener implicaciones en políticas gubernamentales, ya que proporciona información esencial para la formulación de estrategias de apoyo a la investigación y el desarrollo, así como para la planificación de políticas de fomento a la innovación en diversos sectores económicos. En resumen, la predicción de la innovación no solo tiene un valor instrumental en la planificación estratégica y el desarrollo de productos, sino que también representa un área crucial para la investigación académica y contribuye al avance del conocimiento en el campo de la innovación y la gestión empresarial.

1.3.1 Retos a superar.

Dada la gran cantidad de información accesible hoy en día, el proceso de selección y tratamiento de datos para llevar a cabo esta investigación llega a resultar complejo. Surgen en este punto una serie de retos que hemos de superar. Nos centraremos en los siguientes aspectos:

- **Selección y adquisición de datos**. Este trabajo de investigación se centra en dos orígenes de datos diferentes, datos relativos a publicaciones científicas, concretamente artículos, y datos relativos a patentes. El problema principal, que es necesario abordar, es la selección de una fuente de datos para cada uno de estos dos ámbitos asumiendo que existe una relación entre ambos, que a priori, no podrá establecerse fácilmente debido a la heterogeneidad de los datos procedentes de las distintas fuentes de información.

  Es esencial llevar a cabo un estudio pormenorizado de la calidad y fiabilidad de la información obtenida ya que puede no solo complicar en exceso su posterior tratamiento sino también introducir excesivo ruido en los cálculos.

- **Correlación de datos**. Se parte de la base de que se está obligado a trabajar con datos procedentes de diferentes bases de datos, es decir, sistemas heterogéneos en los cuales la misma información se encuentra de diferente forma. Es necesario analizar y buscar fuentes de información que a pesar de presentar los datos de forma distinta permitan establecer una correlación entre los mismos evitando, en la medida de lo posible, un exceso de trabajo realizado de forma manual.
- **Conocimiento**. Esta investigación busca la predicción de innovación aplicada a diferentes tecnologías o líneas tecnológicas referidas al ámbito metalmecánico. Se han utilizado concretamente las tecnologías de estampación en caliente y hierro fundido para el desarrollo de este trabajo, siendo necesario un conocimiento experto de ambas desde el inicio del proceso con la selección de las palabras claves que determinarán los resultados de la búsqueda, la interpretación de resultados y la validación de estos.
- **Predicción**. El último, pero no por ello menos importante reto hace referencia a la predicción de la innovación, siendo el objetivo final articular una solución que nos permita predecir la aparición de una nueva patente ligada a una tecnología específica. El desarrollo de este procedimiento nos conduce a la generación de modelos predictivos de innovación, utilizando

tanto métodos estadísticos como técnicas de aprendizaje automático, tratando de conseguir la precisión más alta posible que nos permita establecer la validez del modelo

1.4 Importancia en el campo de la investigación.

Los modelos predictivos de innovación desempeñan un papel fundamental en el ámbito de la investigación al proporcionar herramientas y enfoques que permiten anticipar y comprender mejor las tendencias, los patrones y las posibles direcciones futuras de la innovación. Estos modelos ofrecen una estructura analítica que ayuda a los investigadores a explorar, comprender y predecir el cambio, lo que resulta crucial en varios aspectos:

- **Identificación de oportunidades.**
  - **Exploración de nuevas ideas**: Los modelos predictivos pueden ayudar identificando áreas de oportunidad para investigaciones futuras, permitiendo la exploración de nuevas ideas que puedan tener un impacto significativo en el desarrollo de la ciencia, la tecnología o la sociedad en general.
  - **Detección de tendencias emergentes**: Estos modelos pueden analizar datos históricos y actuales para identificar patrones emergentes en la evolución de la innovación, lo que permite a los investigadores enfocarse en áreas que podrían ser cruciales en el futuro próximo.
- **Optimización de recursos.**
  - **Uso eficiente de recursos**: Los modelos predictivos ayudan a los investigadores a dirigir sus recursos, ya sean financieros o de tiempo, hacia áreas que tienen un mayor potencial de impacto y contribución, evitando inversiones en direcciones que puedan ser menos prometedoras.
- **Planificación estratégica.**
  - **Apoyo en la toma de decisiones**: Los modelos predictivos ofrecen información valiosa que respalda la toma de decisiones estratégicas. Esto puede incluir desde decisiones sobre el enfoque

de investigación hasta la planificación de políticas y estrategias organizativas.

- **Avance del conocimiento.**
  - **Contribución al conocimiento**: El desarrollo y perfeccionamiento de modelos predictivos específicos para el campo de la innovación pueden contribuir al avance del conocimiento en sí mismo. La investigación sobre la creación y mejora de estos modelos puede ser una valiosa contribución al campo de la ciencia de datos, la estadística o la inteligencia artificial.
- **Comprender la dinámica de la innovación.**
  - **Análisis de factores clave**: Los modelos predictivos permiten identificar y analizar los factores que influyen en el proceso de la innovación, lo que proporciona una mejor comprensión de cómo se generan y desarrollan las ideas innovadoras en diferentes contextos, en este caso aplicado a las diferentes líneas de investigación.

Con los puntos anteriormente expuestos, esta investigación pretende conseguir un modelo predictivo de innovación mejorando los existentes, de forma que:

1. Presente un alto porcentaje de precisión en cuanto los resultados obtenidos, permitiendo predecir la aparición e inminente comercialización de una innovación a medio plazo.
2. Ayude a las organizaciones a determinar si su propia estrategia de investigación está bien orientada facilitando la toma de decisiones a alto nivel para protegerla, reorientarla o incluso abandonarla si se considerase necesario, es decir, facilitando una planificación estratégica.
3. Facilite la detección de ideas y/o tendencias emergentes que contribuyan a una distribución eficaz de los recursos, no solo económicos sino también recursos humanos.

1.5 Solución propuesta.

Considerando los desafíos expuestos en la sección 1.3.1, se trabajará en la resolución de, al menos, algunas de las limitaciones existentes. Así mismo, se

definirá la hipótesis inicial de este trabajo de investigación y se especificarán los objetivos principales de la solución, así como su arquitectura.

1.5.1 Hipótesis de trabajo.

Una vez analizados los diferentes tratamientos de la información relativos, tanto a publicaciones científicas, como patentes se pretende ir un paso más allá. Se planteará la formulación de un modelo de predicción de la innovación que combine ambos tipos de datos. Se emplearán, para ello, no solo métodos estadísticos, sino también el aprendizaje automático. La hipótesis fundamental en la que se basa esta investigación puede formularse de la siguiente manera.

> Es posible generar un modelo predictivo de innovación, que varía en función del ámbito científico-tecnológico analizado, basado en la correlación existente entre el histórico de producción científica elaborado en base a artículos publicados en revistas indexadas, y la subsiguiente generación de propiedad intelectual que debe ser protegida.

Se formula dicha hipótesis bajo la convicción de la existencia de la ya mencionada correlación entre ambos conjuntos de datos. En consecuencia, se podrá desarrollar una metodología de trabajo que contribuya, mediante el análisis de los resultados de estas técnicas predictivas, a analizar el estado de nuestra investigación, permitiendo su reorientación si fuese necesario y facilitando la toma de decisiones a la alta dirección.

1.5.2 Objetivos.

Teniendo en cuenta la hipótesis de trabajo definida pueden identificarse los siguientes objetivos:

- Validar la existencia de la correlación efectiva entre autores de artículos científicos en revistas indexadas e inventores que protegen su propiedad intelectual en forma de patentes.
- Confirmar que existe la masa crítica suficiente susceptible de ser analizada en función del ámbito científico-tecnológico examinado.

- Demostrar que esta correlación puede ser modelada en base a métodos estadísticos y técnicas de aprendizaje automático.
- Verificar que el modelo es lo suficiente preciso tanto a la hora de predecir la aparición de innovaciones como para justificar su protección intelectual en forma de patente.
- Identificar patrones de comportamiento de la actividad científica susceptibles de producir nueva propiedad intelectual, que pueden ser utilizados para tomar decisiones en el ámbito de la inteligencia competitiva.

1.5.3 Diseño global del modelo.

Una vez definida la hipótesis de investigación, así como los diferentes objetivos a alcanzar en el desarrollo de este trabajo se procede a explicar el ciclo de trabajo, esquematizado en la **Figura 9**, que conforma el modelo predictivo de innovación.

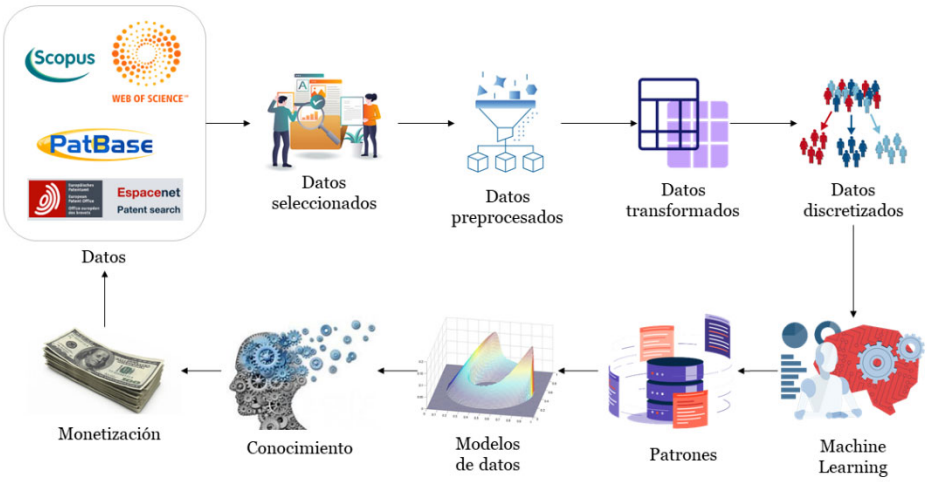

Figura 9: Diseño global del modelo predictivo de innovación planificado, que incluye todos los pasos necesarios desde la captación inicial de datos, hasta la generación de conocimiento y monetización del ciclo.

Se pasa a continuación a explicar los diferentes apartados que componen este diseño.

- Partimos de que existe una gran variedad de bases de datos que contienen información tanto de artículos científicos publicados en revistas indexadas como de patentes.
- Utilizaremos, al menos, dos bases de datos diferentes, una para literatura científica y otra para patentes. Es necesario analizar en este punto cual va a ser nuestro mecanismo de selección de datos, teniendo en cuenta que vamos a trabajar con bases de datos no homogéneas en cuanto a la presentación de la información. Una vez llevado a cabo este proceso obtendremos el conjunto de datos heterogéneos seleccionados.
- El tercer bloque nos lleva a realizar un preprocesamiento, normalización de datos, consistente, en una revisión de estos para verificar su coherencia. Así mismo, se realiza la transformación necesaria para garantizar la homogeneidad entre sistemas de bases de datos diferentes en lo referente a la forma de presentar el dato.
- A continuación, se aplica una transformación de estos datos por medio de un procedimiento de cálculo estadístico que contempla la inversión o esfuerzo acumulado en la publicación de artículos científicos en revistas indexadas, la proclividad a la generación de propiedad intelectual y la generación real de esta. Dicha transformación se realizará mediante un programa de cálculo estadístico desarrollado para esta investigación.
- Se aplicará a continuación un proceso de discretización de los datos de las variables obtenidos en el paso anterior realizando los cálculos con un conjunto de clasificadores que nos indicarán la discretización óptima a utilizar en el siguiente proceso.
- Aplicaremos técnicas de aprendizaje automático al conjunto de datos discretizado anteriormente obteniendo como resultado unos modelos de red Bayesiana y de árbol de decisión con una determinada precisión que habremos de analizar.
- Analizando los modelos obtenidos en el apartado anterior, podemos inferir una serie de reglas o patrones de comportamiento asociados a la tecnología analizada.

- El estudio del modelado de estos datos nos permitirá establecer la validez o no de los modelos.
- Tanto los modelos obtenidos como los patrones que se infieren de los mismos facilitan la adquisición de nuevo conocimiento.
- El conocimiento generado, es susceptible de ser monetizado. Dicha monetización permitirá invertir en la adquisición de nuevos datos o procesos garantizándose la sostenibilidad del ciclo.

1.6 Desarrollo de la tesis.

En esta sección se detallará como se ha planificado este trabajo de investigación presentando en primer lugar, la metodología de investigación empleada. Así mismo, también se identificarán las métricas utilizadas en la validación de la solución propuesta.

1.6.1 Metodología.

Con el objeto de demostrar la hipótesis de este trabajo de tesis doctoral, sección 1.5.1, que es confirmar y modelar la relación existente entre la producción científica basada en artículos indexados y la propiedad intelectual establecida en base a patentes, se ha planteado la siguiente metodología.

1. **Identificación del problema y de los retos a superar**: en este primer paso se trata de tomar conciencia de que nos vamos a mover tanto en el ámbito de la innovación como en el de su predicción, teniendo en cuenta como puede afectar a la toma de decisiones a nivel de la alta dirección en la empresa.
2. **Adquisición de conocimiento**: en esta segunda etapa comenzaremos con una obtención de conocimiento a un nivel general que nos permita ir centrando las etapas iniciales de la investigación. Superadas estas primeras etapas será necesario ir profundizando hacia un conocimiento más específico propio de los diferentes procesos que se irán desarrollando.
3. **División en retos a superar**: definiremos en este punto la metodología a emplear en la resolución de los diferentes retos planteados en el punto 1.3.1. Esto nos permitirá ir abordando esta investigación de una forma más

sencilla disminuyendo la complejidad de las tareas y facilitando la consecución del objetivo final.

4. **Análisis e interpretación de resultados**: una vez resuelto cada uno de los diferentes retos planteados se irán analizando e interpretando los resultados parciales obtenidos. Con la finalización de todos ellos se realizará una interpretación final de estos.
5. **Difusión de los resultados**: esta es la etapa final de este proceso por la cual se ponen a disposición de la comunidad científica los resultados de esta investigación.

1.6.2 Métricas utilizadas.

Una parte importante del desarrollo de esta investigación está sustentada por técnicas de aprendizaje automático. Las métricas utilizadas, que se describen a continuación, son las habituales en este campo, que se encuentran referenciadas en [PRML6].

- **Correctitud**: utilizaremos dicha métrica para saber el grado de corrección y precisión alcanzado en el modelo desarrollado. Este dato, dado en forma de porcentaje, nos indicará la relación de instancias correctamente clasificadas frente a las incorrectas.
- **Tasas de error**: utilizaremos las siguientes tasas de error que nos ayudarán a identificar el comportamiento anómalo del modelo.
    - Error absoluto medio: se define como la tasa de error entre el conjunto de valores predichos X y el conjunto de valores reales Y (ambos con el tamaño del conjunto de datos de pruebas m). Se le conoce como MAE (de la acepción inglesa *Mean Absolute Error*) y su ecuación aparece representada debajo.

$$MAE(X,Y) = \sum_{i=1}^{m} \frac{|X_i - Y_i|}{m}$$

Ecuación 1: Fórmula para el cálculo del error absoluto medio.

    - Raíz del error cuadrado medio: se utiliza habitualmente para evaluar las diferencias entre los valores pronosticados por el modelo

y los valores observados a partir de los que se modela. Se conoce como RMSE (del vocablo inglés *Root Mean Square Error*) y se muestra en la siguiente ecuación.

$$RMSE(X,Y) = \frac{1}{m} \cdot \sqrt{\sum_{i=1}^{m} (X_i - Y_i)^2}$$

Ecuación 2: Fórmula para el cálculo de la raíz del error cuadrado medio.

- **Precisión:** Este concepto fue denominado por Kent [MLS55], como factor de pertinencia. Otros autores que se han referido a él, como ratio de aceptación. Para Salton [IMIR83], la precisión es la proporción de material recuperado realmente relevante, del total de los documentos recuperados. A esta definición Frakes [IRDSAA92] añade que el resultado de esta operación está entre 0 y 1. Así, la recuperación perfecta es aquella en la que únicamente se recuperan los documentos relevantes y por lo tanto tiene un valor de 1.

$$Precisión = \frac{|\{documentos\ relevantes\}| \cap |\{documentos\ recuperados\}|}{|\{documentos\ recuperados\}|}$$

Ecuación 3: Fórmula para el cálculo de la precisión.

- **Exhaustividad:** Este término es conocido por su acepción inglesa *Recall***.** Es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no. Esta medida es inversamente proporcional a la precisión. Fue formulada, al igual que la de precisión por Kent [MLS55], con el nombre de factor de exhaustividad. Años más tarde, Swet [IRSS63] la llamó probabilidad condicional de un item, y Goffman y Newil [GNM64] la denominaron sensibilidad (del vocablo inglés *sensitivity*).

$$Exhaustividad = \frac{|\{documentos\ relevantes\}| \cap |\{documentos\ recuperados\}|}{|\{documentos\ relevantes\}|}$$

Ecuación 4: Fórmula para el cálculo de la exhaustividad.

Si el resultado de esta fórmula arroja como valor 1, se tendrá la exhaustividad máxima posible, y esto viene a indicar que se ha encontrado todo documento relevante que residía en la base de datos, por lo tanto, no se tendrá ni ruido, ni silencio informativo: siendo la recuperación de documentos entendida como perfecta. Por el contrario, en el caso de que el valor de la exhaustividad sea igual a cero, se tiene que los documentos obtenidos no poseen relevancia alguna.

- **Valor-F:** El Valor-F (denominada también *F-score* o medida-F) en estadística es la medida de precisión que tiene un análisis. Se emplea en la determinación de un valor único ponderado de la precisión y la exhaustividad [UCWQ06].

  El valor F se considera como una media armónica que combina los valores de la precisión y de la exhaustividad. La fórmula general para un número real β es:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precisión \cdot Exhaustividad}{(\beta^2 \cdot Precisión) + Exhaustividad}$$

Ecuación 5: Fórmula para el cálculo del valor-F.

  Si β es igual a uno, se está dando la misma ponderación (o importancia) a Precisión que, a la Exhaustividad, si β es mayor que uno de damos más importancia a Exhaustividad, mientras que si es menor que uno se le da más importancia a la Precisión.

1.7 Estructura del documento.

El presente trabajo de tesis doctoral está dividido en una serie de capítulos que se detallan a continuación.

- **Capítulo 1: Introducción**. En este capítulo se introduce el concepto de innovación y la importancia que ha tenido y tiene hoy en día como motor económico y social. Se realiza un pequeño recorrido temporal de la apuesta realizada, no solo por la Unión Europea sino también por la Asamblea General de la Organización de las Naciones Unidas, en pro de la innovación.

Se analiza también el significado de propiedad intelectual profundizando en la modalidad de patentes como elemento de esta. Así mismo, se considera el término predicción aplicado a la innovación y la relevancia que tiene como elemento de toma de decisiones para las empresas. Se plantean, a continuación, los retos a superar en esta investigación, se establece la hipótesis principal de esta, así como los objetivos a cumplir, planteándose el diseño global del modelo a elaborar. Finalmente, se establece la metodología y métricas de este trabajo de investigación concluyendo el capítulo con el resumen y estructura del documento.

- **Capítulo 2: Modelos predictivos de innovación**. En este segundo capítulo se lleva a cabo una revisión de la literatura relativa a los modelos predictivos de innovación. Se comienza haciendo un análisis de las distintas fuentes de información que podemos encontrar para incorporar a nuestro proceso. Se continúa exponiendo las diferentes técnicas de análisis que se emplean tanto en el tratamiento de datos procedentes de artículos científicos indexados como de datos procedentes de patentes, bien de forma individual o de forma conjunta, en aras de predecir la innovación. Se extiende dicha revisión a los métodos estadísticos utilizados en este tipo de procesos detallando cada uno de ellos, la tecnología a la que se aplica y conjunto de datos utilizado. Se finalizará este capítulo con un estudio sobre aprendizaje automático que incluye los conceptos de inteligencia artificial y tipos de esta, así como los distintos tipos de aprendizaje automático, supervisado, no supervisado y semi supervisado.
- **Capítulo 3: Caso de estudio: Estampación en caliente**. En este capítulo se realiza una introducción sobre la tecnología de estampación en caliente (utilizándose también el vocablo inglés *hot stamping* para referirse a esta). Posteriormente se realizará un abordaje bibliométrico del problema con diferentes objetivos, analizar la evolución de dicha tecnología en el tiempo por un lado y, por otro, tratar de clarificar las variables utilizadas en la construcción del modelo predictivo en función de la información ofrecida por las diversas fuentes de información previamente analizadas. Tras este análisis bibliométrico se presentan las bases sobre las que se asienta el nuevo modelo predictivo de innovación desarrollado en este proyecto de

investigación que se compone tanto de un componente estadístico como de un proceso de aprendizaje automático. Se explica con detalle todo el proceso realizado que va desde la adquisición de los datos iniciales, su preprocesado, transformación, discretización, hasta la obtención final del modelo.

- **Capítulo 4: Caso de estudio: Fundición de hierro**. En este capítulo se realizará primero una introducción sobre la tecnología de fundición de hierro (se utiliza también el término inglés *cast iron* para referirse a esta). Es una tecnología mucho más asentada que la tecnología de estampación en caliente lo cual, a priori, indica que aparecerán diferencias de comportamiento a la hora de plantear los modelos. Se analiza en este punto si la metodología que concluye con la obtención de un modelo en el caso de estampación en caliente es también aplicable al caso de la tecnología de fundición de hierro. Se describe cómo se repite el proceso para llegar a la obtención de un modelo específico para esta tecnología, que presenta diferencias con el caso anterior, pero que en cualquier caso valida la sistemática utilizada.
- **Capítulo 5: Conclusiones**. En este último capítulo se realiza la evaluación de los resultados obtenidos comparándolos con la hipótesis y los objetivos de este trabajo de investigación. Se analizan, así mismo, las limitaciones del modelo y se plantean diversas soluciones a los diferentes problemas encontrados abriendo de esta forma futuras líneas de investigación. Se discute también las aplicaciones de este trabajo en lo referente al mundo empresarial, aplicaciones prácticas que ayudarán a la toma de decisiones en cuanto a las diferentes líneas de investigación desarrolladas por cada entidad, indicando la conveniencia o no de continuarla y de qué forma.

2

Modelos predictivos de innovación

Bajo el concepto de modelos predictivos de innovación se agrupa un conjunto de herramientas analíticas y metodologías que mediante la utilización de un conjunto de datos, generalmente heterogéneo, correspondiente a hechos que han tenido lugar en el pasado, más o menos reciente, son capaces de predecir, con cierta precisión, los posibles escenarios futuros en el área de estudio analizada. Se emplean diferentes técnicas estadísticas, algoritmos de aprendizaje automático e inteligencia artificial con el objeto de detectar patrones ocultos a simple vista, identificar tendencias y evidenciar oportunidades que nos indiquen el futuro curso de la innovación en el marco de la tecnología examinada.

Los cálculos realizados por este tipo de modelos se caracterizan por utilizar grandes conjuntos de datos que son procesados mediante complejos algoritmos. Los resultados de este análisis realizado facilitan, a menudo, la toma de decisiones estratégicas tanto a empresas, instituciones académicas como a gobiernos. Esta cualidad, que presentan estos modelos, de facilitar información relativa a posibles escenarios futuros permite tanto a dirigentes empresariales como a investigadores y gobernantes una adaptación proactiva a un entorno en constante evolución impulsando el avance y la mejora en sus ámbitos de actuación.

Este capítulo se estructura de la siguiente forma. La sección 2.1 analiza las diferentes fuentes de información de las que se nutre esta investigación. La sección 2.2 por las diferentes técnicas de análisis que se utilizan en el diseño y construcción de los modelos predictivos de innovación. A continuación, la sección 2.3 hace un recorrido por los diferentes métodos estadísticos utilizados, generalmente de forma complementaria, en este tipo de estudios. Se analiza posteriormente, en la sección 2.4, los conceptos de inteligencia artificial y de aprendizaje automático. Por último, en la sección 2.5 se hace un pequeño resumen de los conceptos revisados en este capítulo.

2.1 Fuentes de información.

Analizando la diversidad de modelos predictivos de innovación existentes en la literatura, vemos que todos ellos responden a un esquema común estructurado en fases [HATMPN19], [TMPET19] y [ICPM19].

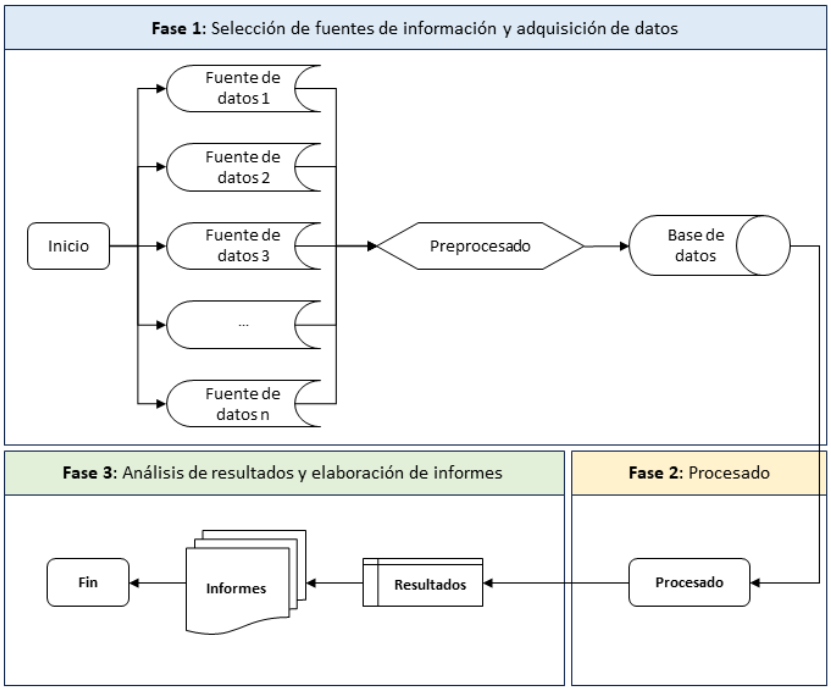

Figura 10: Proceso genérico de un modelo predictivo de innovación dividido en tres fases, (i) selección de fuentes de información y adquisición de datos, (ii) procesado de estos y (iii) análisis de resultados y elaboración de informes orientados a la toma de decisiones.

El proceso comienza con la selección de una o varias fuentes de información y el proceso de adquisición de datos desde estas. Se continua con un preprocesado de estos a fin de poder ser posteriormente tratados mediante las técnicas seleccionadas. El resultado obtenido mediante el procesado realizado es susceptible de ser analizado y está orientado principalmente a la elaboración de informes utilizados en la toma de decisiones.

Si nos centramos en la fase 1 del esquema presentado en la **Figura 10**, vemos que cobra especial relevancia la selección de los orígenes de datos. En función de la opción elegida los resultados de la investigación podrían variar.

Existe una gran diversidad de orígenes de datos que podemos considerar en el desarrollo de esta investigación. En [FETDL17] se propone la utilización de registros bibliográficos basados en la literatura científica existente para llevar a cabo este tipo de análisis. Otros estudios como [NDPMA19] y [TMGPD19] utilizan solamente datos relativos a patentes en el punto de partida de sus procesos, e incluso en [IMDTET19] se lleva a cabo el proceso de análisis de datos mediante el uso de mensajes en Twitter (actualmente X) [X24] para prever las tendencias de las tecnologías emergentes. Como se explicará en el capítulo 3, el desarrollo de esta tesis doctoral está basado en artículos publicados en revistas indexadas, y propiedad intelectual que debe ser protegida en forma de patentes.

2.1.1 Orígenes de datos inherentes a artículos científicos indexados.

Como fuentes de datos principales para la utilización de literatura científica encontramos Web of Science (WoS), Scopus, Google Académico (del vocablo inglés *Google Scholar*) y Dimensions.

La base de datos de Web of Science (WoS) [WoS21] contiene más de 174 millones de registros, más de 34.586 revistas más libros, actas, patentes, y conjuntos de datos que cubren un periodo de tiempo que va desde 1800 hasta nuestros días para literatura científica publicada en revistas y desde 1963 hasta el presente para patentes. WoS también incluye 1,9 billones de referencias de citas desde 1900 hasta el día de hoy. El análisis de las citas incluye el seguimiento de las mismas, su recuento y el cálculo del índice h del autor.

Scopus [SCO21], otra importante base de datos relativa a publicaciones científicas, lanzada en 2004, tiene más de 76,8 millones de registros clave, 51,3 millones de registros con referencias posteriores a 1955, 25,3 millones de registros anteriores a 1996, y 43,7 millones de registros de patentes. Con más de 1.7 billones de referencias de citas, solo los documentos que se remontan a 1970 las contienen. El análisis de citas incluye recuentos de citas, citas por autor y por revista, así como el índice h.

Google Académico [GA21] es un motor de búsqueda propiedad de Google en funcionamiento desde noviembre de 2004 y especializado en la búsqueda de bibliografía académica en general. Proporciona información relativa a muchas disciplinas, así como diversas fuentes, artículos, tesis, libros, resúmenes y opiniones de tribunales, de editoriales académicas, sociedades profesionales, repositorios en línea, universidades y otros sitios web.

Dimensions [DIM21] es una base de datos parcialmente gratuita propiedad de Digital Science que opera desde enero de 2018 [DSWOS18]. Dimensions incluía en 2019 más de 106 millones de publicaciones en revistas académicas, es decir, alrededor de un 30% más que las bases de datos comparables, libros y capítulos de libros, preprints y actas de conferencias. Todas las publicaciones están contextualizadas con conjuntos de datos vinculados como datos referentes a subvenciones, ensayos clínicos, patentes y documentos políticos [DBDBBSD20]. También pueden analizarse aspectos como las categorías asociadas, los financiadores, las instituciones y los perfiles de los investigadores.

En el estudio realizado para las bases de datos Google Académico, Scopus y Web of Science [GSSWoS16] se proporciona una comparación sistemática y exhaustiva de la cobertura de estas. La comparación longitudinal realizada muestra un crecimiento trimestral consistente y razonablemente estable tanto para las publicaciones como para las citas en las tres bases de datos. Esto sugiere que las tres bases de datos proporcionan suficiente estabilidad de cobertura para ser utilizadas para comparaciones interdisciplinarias. comparación interdisciplinaria de cuatro métricas clave de investigación (publicaciones, citas, índice h e índice h, anual, un índice h individual anualizado) en cinco disciplinas principales

(Humanidades, Ciencias Sociales, Ingeniería, Ciencias y Ciencias de la Vida) mostró que tanto la fuente de datos como las métricas específicas utilizadas cambian las conclusiones que se pueden extraer de las comparaciones interdisciplinarias.

Además, teniendo en cuenta solamente WoS y Scopus, se puede concluir que no existen grandes diferencias entre ambas (Bakkalbasi, Bauer, Glover, & Wang, 2006) y ambas son complementarias (Escalona, Lagar, & Pulgarín, 2010), varios autores (Bakkalbasi et al., 2006; Neuhaus & Daniel, 2008) recomiendan realizar un análisis preliminar con ambas bases de datos para identificar posibles diferencias en los resultados dependiendo de la disciplina.

2.1.2 Orígenes de datos referentes a patentes.

Por otro lado, vamos a considerar también los datos relativos a patentes como la otra fuente de información necesaria para el desarrollo de este estudio. Las patentes son, por naturaleza, documentos muy técnicos que pueden ser difíciles de leer y que condensan sólo la información más importante. Constituyen otra fuente de datos susceptible de ser analizada en el proceso de la predicción de la innovación [DI21]. Derwent Innovation [PMLNA18] es una plataforma de investigación y análisis de patentes que ofrece acceso a patentes y literatura científica de confianza a nivel mundial. Clarivate Analytics mejora la capacidad de búsqueda y descubrimiento de los datos de patentes añadiendo metadatos al registro de patentes:

- **Título descriptivo**: títulos concisos que describen la invención y su novedad.
- **Resumen**: descripción de 250-500 palabras en inglés sobre la reivindicación y la novedad de la invención.
- **Familia de patentes**: las solicitudes de la misma invención en países de todo el mundo están vinculadas en un solo registro.
- **Códigos de clase Derwent**: permite al usuario recuperar rápidamente una categoría de invenciones.
- **Códigos manuales Derwent**: indica los aspectos técnicos novedosos de la invención.

Otro recurso en el cual podemos encontrar información acerca de patentes, dada la gran cantidad de estudios analizados que lo mencionan, es el Servicio de Información sobre Derechos de Propiedad Intelectual de Corea (traducción de la voz inglesa *Korea Intellectual Property Rights Information Service* (KIPRIS)) [KIPRIS21]. KIPRIS [IMDTET19] es el servicio gratuito de búsqueda de información sobre propiedad intelectual más completo de Corea disponible en internet. Cubre toda la información de propiedad intelectual coreana de la Oficina de la Propiedad Intelectual de Corea (del vocablo inglés *Korean Intellectual Property Office* (KIPO)) y está gestionado por el Instituto Coreano de Información sobre Patentes (del vocablo inglés *Korea lnstitute of Patent Information* (KIPI)) en nombre de la KIPO [KIPRISS21]. Entre los servicios que ofrece encontramos:

- Servicio de búsqueda de información de propiedad intelectual. Patentes, modelos de utilidad, diseños, marcas, resúmenes de patentes coreanas (*Korean Patent Abstracts*, KPA).
- Servicio de búsqueda de información administrativa de la propiedad intelectual. Examen, registro y estado del proceso.

Sofean et Al. [WPMAF18] utilizan los servicios de la plataforma comercial [STN21] que proporcionan un acceso integrado a una colección actualizada y completa de contenido científico y técnico, tanto de patentes como de no patentes. A partir de dicha plataforma utilizan para su repositorio patentes de la Organización Mundial de la Propiedad Intelectual (del vocablo inglés *World Intelectual Property Organization* (WIPO)), patentes de la Oficina Europea de Patentes (de la acepción inglesa *European Patent Office* (EPO)) y patentes de la Oficina de Patentes y Marcas de los Estados Unidos (del término inglés *United States Patent and Trademark Office* (UPSTO)).

La Organización Mundial de la Propiedad Intelectual [WIPO99] ha asumido el compromiso de proporcionar servicios, conocimientos y datos de propiedad intelectual de alta calidad que aporten valor a quienes utilizan dicho sistema en todo el mundo a través de su portal WIPO IP portal [WIPOIP21]. Proporciona acceso a más de 96 millones de documentos de patente disponibles en PATENTSCOPE [PS21], más de 46 millones de marcas, denominaciones de origen y emblemas disponibles en la Base Mundial de Datos sobre Marcas (del término

inglés *Global Brand Database*) [GBD21] y a más de 13 millones de dibujos y modelos disponibles en la Base Mundial de Datos sobre Dibujos y Modelos (de la expresión inglesa *Global Design Database*) [GDD21].

El 5 de octubre de 1973, tras más de 20 años de negociaciones y debates, 16 países firman en Múnich el Convenio sobre la Patente Europea. Este tratado multilateral crea la Organización Europea de Patentes y la Oficina Europea de Patentes, (de la voz inglesa *European Patent Office* (EPO)) [EPO21] y establece un sistema jurídico autónomo para la revisión y concesión de patentes europeas. La colección de la Oficina Europea de Patentes contiene más de 100 millones de documentos de patentes de todo el mundo y está a disposición del público a través del servicio gratuito de internet Espacenet [ENET21].

La Oficina de Patentes y Marcas de los Estados Unidos, (de la expresión inglesa *United States Patent and Trademark Office* (USPTO)) [UPSTO21] es la agencia federal encargada de conceder patentes y registrar marcas en los Estados Unidos. Ofrece un servicio tanto para la búsqueda de patentes como de marcas. Se puede utilizar el servicio UPSTO Base de datos de patentes completas – Texto e imágenes (del vocablo inglés *UPSTO Patent full – Text and image database*) [UPSTOPF21] y el servicio Sistema de Búsqueda Electrónica de Marcas (del inglés *Trademark electronic search system* (TESS)) [TESS21]. En los estudios realizados por Noh, Jo, y Lee [KSTMPA15], No, An y Park [SACATM15] y Comins [DMGFPPS15] se utiliza dicha base de datos de patentes.

Existen en el mercado soluciones comerciales como es el caso de PatBase® [PATB23] entre otras. Esta base de datos es capaz de proporcionar acceso a documentos de patentes de más de 100 autoridades emisoras de todo el mundo y contiene más de 47 millones de familias de patentes. Si las patentes contienen prioridades comunes con otras patentes, PatBase® las agrupa en familias. Estas familias ampliadas son utilizadas por la Oficina Europea de Patentes (OEP) y tienen la ventaja de producir resultados deduplicados y preagrupados [TTHRPA20].

2.2 Técnicas utilizadas en análisis predictivo.

La predicción de la innovación basada en la prospectiva tecnológica, en la previsión tecnológica y en el análisis de oportunidades tecnológicas, ha sido ampliamente estudiada desde que la RAND Corporation desarrolló el método Delphi en la década de 1950, para analizar nuevos sistemas militares [FOTA08]. Los investigadores han clasificado estos métodos en áreas específicas, como el análisis de tendencias [TPTTFP12], la previsión tecnológica [TFEE01] [RRATF03] [PDTF11], el análisis del futuro de la tecnología [TFATIFM04], la detección de nuevas oportunidades para las tecnologías [TOI14] y la inteligencia tecnológica [TIMM05].

En otro orden de ideas, podemos establecer que la predicción del futuro, y por lo tanto de la innovación como parte de este, es una tarea compleja, siendo el futuro incierto por naturaleza. No obstante, existen diversas técnicas y herramientas que pueden ayudarnos en tareas como la anticipación de las tendencias del mercado, la identificación de nuevas oportunidades y el aumento de las probabilidades de éxito en el desarrollo tanto de productos como de procedimientos.

En lo que respecta a las técnicas de análisis predictivo, estas desempeñan un papel fundamental en estos ámbitos. Mediante la investigación realizada sobre conjuntos de datos, tanto de carácter histórico como más cercanos al presente, cabe la posibilidad encontrar patrones y/o tendencias que nos permitan establecer un posible comportamiento futuro para el área de estudio analizada. Puede conseguirse con ello una minimización del riesgo en la toma de decisiones, así como una anticipación al mercado. Se presentan a continuación una serie de técnicas, con distintos enfoques, relativas al análisis predictivo.

2.2.1 Técnicas basadas en metadatos de patentes.

Las patentes constituyen la mayor fuente de información tecnológica disponible. Actualmente existen, aproximadamente, unos 50 millones de documentos de patentes en el mundo, que contienen el 90% de los logros científicos y tecnológicos. El uso eficaz de la información sobre patentes puede acortar el 60% del tiempo de I+D+i y ahorrar el 40% de la financiación de I+D+i según [TMAPD09]. Si se analizan detenidamente, pueden mostrar detalles tecnológicos, revelar tendencias

empresariales, inspirar soluciones industriales novedosas o ayudar a elaborar políticas de inversión [PTTFT83] [IUPI03]. Es por ello por lo que el análisis de patentes ha sido adoptado para este tipo de estudios debido a la relación natural existente entre el texto de las patentes y la innovación, constituyéndose como el recurso más importante aplicado en el ámbito de los análisis tecnológicos orientados a la predicción.

En este contexto, podemos hablar también de innovación recombinativa que surge de nuevas combinaciones a partir de tecnologías o características tecnológicas ya existentes [IIS97]. La recombinación de elementos tecnológicos está reconocida como una actividad crucial de innovación [TPRI16] debido a que gran parte de las innovaciones tecnológicas proceden de la combinación de elementos tecnológicos ya existentes [WFRCCF13] [MPITE15].

Relacionado con este aspecto debemos preguntarnos ¿Cómo podemos detectar las tecnologías recombinantes? ¿Cómo podemos agruparlas para facilitar su análisis? En gran medida solamente la experiencia y el juicio de las personas expertas es lo que facilita esta labor [RRFKS98]. Existen, no obstante, métodos cuantitativos para apoyar dichas labores. En cuanto al tipo de datos utilizados, los metadatos de las patentes han demostrado ser especialmente útiles como fuente de inteligencia técnica [HSMMPA16]. De esta forma, el análisis de redes de patentes se ha establecido como una técnica para generar y visualizar los campos tecnológicos establecidos como objetivo a un nivel macro [PKNA14]. Sin embargo, a pesar de los avances conseguidos en las técnicas de minería de textos, que permiten extraer información sobre las tecnologías a un nivel de detalle mucho más preciso [SAONAP11], no se ha logrado que el análisis de redes de patentes se convierta en una técnica consolidada aplicada en el campo de la innovación recombinativa.

Tal y como ya se ha comentado en el punto 2.1, el rápido despliegue de tecnologías basadas en Internet, sobre todo en lo referente al acceso a la información y las facilidades de búsqueda de esta, han facilitado a las empresas el acceso a los documentos de las patentes. Así mismo, oficinas como la WIPO [WIPO99], la USPTO [UPSTO21] y la EPO [EPO21], entre otras, prestan gratuitamente este servicio.

Han ido surgiendo, a lo largo de los años, herramientas y servicios de software susceptibles de ser aplicadas al ámbito de las patentes [MAPA02] [PACTIIT05] [SCPD08] cuyo objeto principal es analizar las patentes utilizando métodos de clasificación, agrupación y estadística para encontrar relaciones entre patentes con contenido/estructura similar.

La información obtenida de las patentes puede analizarse tanto de forma cuantitativa como cualitativa [LPANSE03]. Las medidas cuantitativas se basan en el procesamiento estadístico e indican el nivel de actividad de patentes [DPISME15]. Las medidas cualitativas se calculan en función de la información sobre citas y se utilizan para evaluar la calidad de una patente [DPCC00].

Debido a esto, se han definido una serie de indicadores con el objeto de determinar el valor de las patentes. Se parte de la base de que son una fuente de información tecnológica y competitiva y además el acceso a las mismas es relativamente fácil [IFPA91] [IPVMP04] [PCIPV07].

- **Edad de la patente**: La edad en años calculada desde la fecha de solicitud de esta.
- **Citación realizada (citas retrospectivas)**: Número de patentes citadas por la patente objetivo.
- **Índice de citas (citas hacia delante)**: El número de citas recibidas por la patente objetivo. Es una medida del impacto de la patente objetivo.
- **Originalidad**: La originalidad de una patente objetivo indica la diversidad de patentes citadas. Es decir, si durante el proceso de examen, una patente hace referencia o cita muchas otras patentes, puede ser un indicio de que la invención es original y podría representar una contribución significativa al estado de la técnica.
- **Generalidad**: La generalidad de una patente se refiere tanto a la relevancia como a la aplicabilidad de una invención. La generalidad de una patente indica la diversidad de otras patentes que hacen referencia a la patente en cuestión, es decir, la cantidad y variedad de patentes que

citan la patente objetivo como una referencia importante en su campo tecnológico.

- **Duración del ciclo tecnológico (del inglés *Technology Cycle Time* (TCT))**: El TCT de una patente objetivo es la edad media de las patentes citadas por la patente objetivo. Es una medida del progreso tecnológico.

2.2.2 Técnicas basadas minería de datos y minería de textos.

Un conjunto de técnicas que ha logrado una gran relevancia en el procesado de la información son la minería de datos (de la expresión inglesa *data mining*) y la minería de textos (del inglés *text mining*). La minería de datos se utiliza tanto en el ámbito de la literatura científica como en el de las patentes. Se trabaja con un conjunto de datos estructurados, homogéneos y organizados obtenidos habitualmente de una base de datos. Por el contrario, en la minería de textos los datos utilizados, es decir, los datos procedentes de textos son heterogéneos, no se presentan de forma estructurada, presentan formatos diversos y contenidos distintos.

El principal objetivo de la minería de textos no es otro sino extraer información significativa de un conjunto de datos no estructurados. Se utiliza en una gran variedad de campos de investigación así como en campos relativos a la ciencia de la información [MTICRM09] [DMIS04] [MARCA07] [DMTCRM09] [DMADK06]. Su uso extendido se debe, sobre todo, a su capacidad para trabajar con grandes volúmenes de texto.

La aplicación de la minería de textos puede realizarse empleando palabras clave (se utiliza también el vocablo ingles *keywords*) o bien puede estar basado en el análisis de palabras de forma general [IETT11]. Song et al. [DNTOBP17] han aplicado esta tecnología de minería de textos junto con el análisis de palabras clave en un intento de descubrir nuevas oportunidades tecnológicas. Walter et al. [TBOBB17] han utilizado el análisis de minería de textos en estudios relacionados con el azufre, concretamente con la belleza de la mariposa de azufre, buscando novedades en las patentes mediante exploraciones del entorno cercano, es decir, relaciones entre palabras. Se han desarrollado herramientas de minería de textos

expresamente para la recuperación de información a partir de conjuntos de patentes [TMTIRP17]. Kayser et al. [EKBFTM17] han llevado a cabo un estudio basado en la minería de textos que ha permitido ampliar la base de los conocimientos prospectivos. Se ha utilizado también esta tecnología en conjunción con la técnica de similitud del coseno en la búsqueda de oportunidades relativas a productos [CDCPPO17]. En un intento de mejorar la gestión de la tecnología sostenible [PKESTM18] se ha llevado a cabo un estudio mediante la utilización de palabras clave aplicadas a la minería de textos. Así mismo, Roh et al. [DMSLTIP17] han desarrollado una metodología de estructuración y estratificación de la información tecnológica aplicando la minería de textos a conjuntos de documentos de patentes.

Como ya hemos establecido la minería de textos es una técnica multidisciplinar capaz de analizar grandes volúmenes de texto, es decir, información heterogénea o datos no estructurados. Se combinan en este proceso complejos algoritmos informáticos con métodos lingüísticos y estadísticos con el objetivo final de descubrir patrones ocultos no accesibles a simple vista, nuevas tendencias o cualquier tipo de información y/o conocimiento que se considere útil, tal y como ha quedado referenciado en párrafos anteriores. Se emplea en campos tan diversos como, por ejemplo, el análisis de opiniones y retroalimentación del cliente, el análisis de redes sociales, la detección del fraude y/o anomalías, la gestión del conocimiento y el análisis de la satisfacción y el sentimiento de los empleados.

Todos estos ámbitos, susceptibles de ser analizados mediante la aplicación de la tecnología de minería de textos, hacen que podamos establecer la siguiente clasificación que comprende las diferentes técnicas que la componen.

1. Procesamiento del lenguaje natural (del inglés *Natural Language Processing* (NLP)).
    a. Técnica basada en palabras clave.
    b. Técnica Sujeto-Acción-Objeto (del término inglés *Subject-Action-Object* (SAO)).
    c. Técnicas de análisis sintáctico.
2. Técnicas basadas en reglas.

   a. Minado por reglas de asociación.
   b. Sistema de inferencia difusa (de la expresión inglesa *Fuzzy Inference System* (FIS)).
3. Técnicas basadas en análisis semántico.
   a. Método de modelo documental basado en vectores.
   b. Métodos de similitud basados en corpus.
   c. Métodos híbridos.
   d. Método jerárquico de vectores de palabras clave.
   e. Similitud semántica de textos.
4. Enfoques de visualización.

2.2.2.1 Procesamiento del lenguaje natural.

El procesamiento del lenguaje natural, también conocido como la comprensión del lenguaje natural, es un campo de las ciencias de la computación que explora cómo este puede ser procesado de forma mecánica mediante la utilización de sistemas informáticos. Se utiliza en dicho proceso tanto el aprendizaje automático como la lingüística computacional intentando que la interacción entre las personas y las máquinas sea fácil pero eficiente. El ordenador aprende la sintaxis y el significado del lenguaje humano, lo procesa y le da la salida al usuario independientemente del idioma utilizado [NLP18]. Está basado tanto en un conjunto de teorías como de tecnologías cuyo fin último es realizar un análisis de textos. El NLP investiga el diseño y desarrollo de modelos matemáticos para las estructuras del lenguaje natural, así como su implementación [NLPPP10].

No existe una única definición consensuada sobre NLP que satisfaga a todo el mundo, pero si existen algunos aspectos que formarían parte de la definición proporcionada por cualquier persona con estos conocimientos. Concretamente, la definición ofrecida por [NLP01] es:

> *El procesamiento del lenguaje natural está formado por un conjunto teórico de técnicas computacionales, desarrolladas para analizar y representar textos de forma natural en uno o más niveles de análisis lingüístico, con el propósito de lograr un*

> *procesamiento del lenguaje similar al humano, utilizado en un amplio rango de tareas o aplicaciones.*

Como ya se ha explicado, los documentos utilizados en los análisis realizados mediante modelos predictivos de innovación, tanto los procedentes de literatura científica como de patentes, presentan una naturaleza no estructurada utilizándose lenguaje natural para su redacción. Si bien las personas pueden entender fácilmente este lenguaje, diferenciando entre la ortografía del texto, los errores ortográficos y el significado contextual de las palabras, un procesamiento manual de este tipo de documentos además de ser muy lento requiere mucho tiempo y esfuerzo. Es por esto, por lo que es necesario la creación de algoritmos complejos que hagan a los ordenadores *inteligentes* para distinguir entre la sintaxis y la semántica del texto. En este contexto, el NLP se ha convertido en una tecnología esencial en la construcción de modelos y procesos que utilizan fragmentos de información como datos de entrada, ya sea en forma de voz, texto o ambos, y los transforman de acuerdo con el algoritmo utilizado.

El NLP presenta diferentes metodologías o técnicas de aplicación. La primera de ellas está basada en el uso de palabras clave o *keywords*. Como ya se ha constatado en el punto 2.1, se dispone de un amplio abanico de recursos que nos van a proporcionar información tanto relativa a las publicaciones científicas como de patentes. La utilización de estos servicios, habitualmente disponibles en Internet, se realiza mediante el uso de palabras clave que dictaminan la parametrización de la búsqueda y con el objeto de recuperar la información susceptible de ser analizada [DCIMPT10]. Esta técnica se basa en un modelo de espacio vectorial [VSMAI75] que representa la documentación buscada en términos de frecuencia de la aparición de las palabras clave. Se conoce como TF-IDF (del inglés *Term Frequency – Inverse Document Frequency*), frecuencia de término – frecuencia inversa de documento (o sea, la frecuencia de ocurrencia del término en la colección de documentos), es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Se utilizan, sobre todo, palabras clave de tipo tecnológico con el objetivo de obtener mejores resultados a partir de la consulta realizada. Si bien es un método sencillo de utilizar, eficiente y con una alta disponibilidad, no proporciona por sí solo información relativa a las relaciones

tecnológicas existentes, es decir, no describe cómo se utilizan los diferentes componentes pertenecientes una tecnología dada. Además, presenta una serie de limitaciones sustanciales como:

1. La definición tanto de las palabras como de las frases clave depende en gran medida de la experiencia y de los conocimientos sobre la materia que pueden aportarse al proceso [SAONAP11] [IETT11] [DSNTO12].
2. Las frecuencias y coocurrencias de los patrones pueden no reflejar características sustanciales de las invenciones [TCOPDP08].
3. Los mapas de patentes ser sustancialmente diferentes dependiendo de la persona o grupo de personas que hayan seleccionado tanto las palabras como las frases clave [DSNTO12].

Otra técnica utilizada dentro del NLP es la conocida como Sujeto-Acción-Objeto o SAO. El concepto de estructura SAO está basado en la teoría de resolución creativa de problemas (de la expresión rusa *Tieoriya Riesheniya Izobrietatielskij Zadach* (TRIZ)). Genrich Altshuller utilizó el concepto de función y desarrolló TRIZ tras realizar un extensivo análisis de más de 200.000 patentes cuyo objetivo era generalizar y abstraer tecnologías [CESIP84]. Una estructura SAO nos permite la representación de los *conceptos clave* en lugar de las *palabras clave* [NLPPTD04] pudiendo de esta forma exponer las relaciones de tipo medio-fin [PBIP05]. Los investigadores Cascini, G., Fantechi, A., y Spinicci, E. [NLPPTD04] desarrollaron PAT Analyzer con el objeto de extraer los *conceptos clave* en lugar de las *palabras clave*. Cascini, G. llevó a cabo una investigación en la cual se buscaban contradicciones en la teoría TRIZ [CAAP07] y para medir la similitud de las patentes por comparación con los árboles funcionales creados a partir de los inventos [MPSCIFT08].

Un enfoque diferente consiste en utilizar las estructuras SAO para analizar similitudes entre patentes y para mejorar la precisión de los métodos de evaluación de estas [SMDM09] [MTPS10]. En este sentido, Moehrle et al. [PBIP05] plantearon una metodología que utiliza la información de los perfiles de los inventores, contenida en los documentos de las patentes, con el objeto de dar soporte a las decisiones que debía tomar los departamentos de recursos humanos.

Así mismo, Bergmann et al. [ERPISPA08] utilizaron este concepto de similitud entre patentes [CPL01] para analizar las posibles infracciones de estas [IPISAO12], mientras que Yoon y Kim [IETT11] plantearon la utilización de una red de patentes basada en el análisis semántico mediante la utilización de estructuras SAO. La utilización de esta tecnología se justifica en base a que las técnicas basadas solamente en el uso palabras clave, muy extendida en el ámbito del análisis de patentes [DTIITO08] [VPAET08] [USPN00] [TMPN04], no son capaces de identificar las infracciones cometidas por estas. En otras palabras, no pueden determinar aquellas situaciones en las que una patente se utiliza o explota por terceros violando los derechos que se otorgan al titular de la misma, es decir, los derechos de propiedad intelectual. Esta imposibilidad se justifica porque los vectores de palabras clave se basan únicamente en la frecuencia de estas, sin tener en cuenta ni las claves tecnológicas específicas, ni las relaciones estructurales de los componentes [IPFNAP11].

Las estructuras SAO son también susceptibles de ser utilizadas en las técnicas de análisis de redes. Choi et al. [SAONAP11] trataron de identificar tendencias tecnológicas utilizando dicho planteamiento para medir la densidad y la cohesión de las subredes SAO. Del mismo modo Yoon y Kim [IETT11], midieron la similitud de las estructuras SAO mediante la construcción de redes de patentes basadas en esta técnica. Calcularon la densidad y la cohesión entre subredes para identificar el grado de novedad de las patentes, es decir la innovación, a nivel de subdominios no de patentes individuales.

En este sentido Choi et al. [ASAOTMA12] han generado los llamados árboles tecnológicos (del vocablo inglés *TechTree*), que no son sino diagramas con estructura de árbol utilizados en la representación de las relaciones existentes entre las diferentes tecnologías [MTPS10] [SMDM09] [DSNTO12] [OFMTR11]. Los árboles tecnológicos se han convertido en herramientas esenciales para la toma de decisiones en el contexto de la planificación tecnológica [MOTSMS06] siendo ampliamente utilizadas para dicho fin [DRDSTAA86] [TTDVND09] [DTT92] [TNPESA10] [ESAARRD08] [MATRTM08]. Sin embargo, la creación de estos árboles tecnológicos presenta ciertas limitaciones.

1. Se depende de personas expertas tanto en el proceso de creación como en la actualización y mantenimiento de estos debido, en gran medida, a un entorno cambiante, aumentando considerablemente los costes tanto en términos de tiempo como de recursos humanos.
2. Representan una información tecnológica limitada y solo es de utilidad para aquellas áreas para las cuales ha sido desarrollado.
3. No pueden representar información subsidiaria de la tecnología analizada que es, a todos los efectos, tan importante como la tecnología misma.

Con todo lo anteriormente expuesto sobre la metodología SAO, puede concluirse que, si bien es muy eficaz en el proceso de extracción de información relativa a la tecnología analizada partiendo de un gran volumen de información, pueden presentarse problemas relativos a las ambigüedades tanto semánticas como gramaticales [LRPA14] sobre todo cuando se analizan patentes dónde las relaciones semánticas no aparecen siempre correctamente representadas.

Otro grupo de técnicas englobadas dentro del procesamiento de lenguaje natural son aquellas que utilizan métodos de análisis sintáctico. El objeto de estas técnicas PNL no es otro sino analizar el texto de los documentos seleccionados, realizando una extracción selectiva de este, facilitando la búsqueda de patrones ocultos que puedan resultar de interés. Se analizan, es este proceso, las diferentes frases del texto y se construyen los denominados árboles de análisis. Esta herramienta proporciona mejores estructuras si vamos a realizar nuestra investigación, por ejemplo, mediante las técnicas de minería de texto combinado con aprendizaje automático [DAESKI13] [ITMT13] [OUOG13].

Esta técnica, al igual que sucedía con la del NLP, tiene un elevado costo tanto en personal como en tiempo. La elaboración de los árboles de análisis sintáctico [AUP03] o estructuras similares [GTDP06] pueden llevarse a cabo tanto mediante técnicas de aprendizaje automático como de forma manual elaborados por expertos. Incluso cuando se utiliza el aprendizaje automático, este se realiza por comparación con patrones que previamente han debido ser construidos de forma manual por expertos, que constituye una ardua y larga tarea.

2.2.2.2 Técnicas de minería de textos basadas en reglas.

El segundo gran conjunto de técnicas analíticas empleadas en la minería de textos corresponde a aquellas basadas en reglas. Se puede distinguir, básicamente, entre dos tipos de reglas, las de inferencia y las de asociación, que no son más que un conjunto de criterios establecidos de la siguiente forma,

SI (condición) ENTONCES (sentencias)

que proviene de la acepción inglesa *IF (condition) THEN (statements)*. El objeto de este tipo de reglas no es otro sino la agrupación de documentos en función de su interdependencia.

Está técnica tiene multitud de aplicaciones que van desde un sencillo análisis de la cesta de la compra hasta un complejo análisis de patentes utilizando los códigos de clasificación internacional de estas en un intento de realizar predicciones tecnológicas [IPCPD11] [NPAARM13]. Estas reglas establecen distintas correlaciones entre los ítems analizados [ESLDMIP01].

Se parte de un gran conjunto de datos que serán analizados por medio de esta metodología buscando asociaciones o relaciones significativas entre los diferentes elementos que lo componen [MARLD01]. Una vez aplicadas las reglas de asociación se mostrarán las diferentes regularidades o patrones existentes. Con el objeto de determinar si el minado con las reglas aplicadas constituye un patrón regular se emplean dos tipos de métricas, la de soporte o apoyo y la de confianza [DMPMLT00]. La medida de apoyo determina la probabilidad de que una transacción contenga tanto la parte condicional como la parte resultante de una regla, mientras que la medida de confianza es la probabilidad condicional de que una transacción que contenga la parte condicional de una regla también contenga la parte resultante.

Existen algunas limitaciones cuando aplicamos este método, en especial cuando nuestro objeto de análisis son las patentes. Se corre el riesgo de obtener una representación incompleta o vaga acerca del conocimiento extraído de los documentos. Así mismo, es posible que se establezcan relaciones falsas entre los diferentes objetos de análisis.

Las reglas de asociación pueden utilizarse también para detectar cambios, es decir, determinar los cambios existentes entre dos conjuntos de datos comparando las reglas extraídas con este método [MCCBSM01] [MCCBRM05] [DSFRC01]. Existen varios tipos de patrones que nos permiten identificar los cambios:

- **Patrones emergentes**: El concepto de patrones emergentes capta cambios significativos entre los conjuntos de datos. Un patrón emergente es un patrón de reglas cuya influencia aumenta significativamente de un conjunto de datos a otro.
- **Cambios consecuentes inesperados**: Estos cambios se encuentran en reglas de asociación recién descubiertas cuyos resultados difieren de los de los patrones de reglas anteriores.
- **Cambios de condición inesperados**: Estos cambios se encuentran en reglas de asociación recién descubiertas cuyas partes condicionales difieren de las de los patrones de reglas anteriores.
- **Reglas añadidas**: Son reglas nuevas que sólo existen en el conjunto de datos actual.
- **Reglas perdidas**: Son reglas que sólo existen en el conjunto de datos anterior.

Esta estrategia de utilización de reglas de asociación aplicada a la detección de cambios en los conjuntos de datos se emplea para identificar cambios de tendencias en las patentes.

Los sistemas de inferencia difusa o *Fuzzy Inference System*, FIS, funcionan también mediante la aplicación de reglas SI-ENTONCES. Cuando el resultado de aplicación de estas reglas indica que existe similitud en el conjunto de datos analizados, por ejemplo, similitud entre patentes, se produce un agrupamiento de estos para someterlos a un análisis posterior. De esta forma vamos realizando un ajuste progresivo de las reglas utilizadas logrando una mejor identificación de la concurrencia. En el caso del análisis de datos de patentes, este método permite obtener buenos resultados incluso trabajando solo con información parcial de estas. Sin embargo, existen también limitaciones cuando aplicamos esta sistemática. En el caso del estudio de las patentes, se requieren, a menudo, los

conocimientos de uno o varios expertos para poder establecer el conjunto de reglas necesario y adecuado que sea capaz de desentrañar las relaciones ocultas existentes.

Así mismo, cuando utilizamos sistemas de análisis de patentes asistidos por ordenador obtenemos los atributos críticos necesarios que sirven como parámetros de entrada en los sistemas FIS, que son:

- la cantidad de patentes [BTCPCMA06],
- el índice de especialización RPA (del vocablo inglés *Revealed Patent Advantage*) [ETSC95] [WDSG13],
- la actividad de la patente [PAKT02]
- el índice de citas [BRDCPA04] y,
- el índice relativo de citas [TPSAM07].

A partir del algoritmo de aprendizaje de Kohonen [SOAM12] así como de la heurística proporcionada por la técnica del primer y más cercano vecino (del vocablo inglés *first-nearest-neighbour*) [NNFLCDS91] se construyen los diferentes términos de las sentencias SI-ENTONCES tomando como referencia los valores de los parámetros relativos a empresas del mismo ámbito tecnológico. Considerando también los valores de los atributos de entrada, pertenecientes a las patentes analizadas, los sistemas de inferencia difusa son capaces de deducir y sugerir las estrategias tecnológicas más adecuadas.

Estás técnicas presentan una serie de pros y contras en lo que se refiere a su aplicación. En cuanto a las ventajas podemos enumerar las siguientes.

1. **Interpretación explícita**: Las reglas utilizadas en estas técnicas son explícitas y transparentes, lo que facilita la comprensión de cómo se realizan las extracciones de información. Esto permite a los usuarios comprender y validar fácilmente el proceso de análisis de texto.
2. **Flexibilidad**: Las reglas pueden ser ajustadas y adaptadas según las necesidades específicas del problema o del dominio de aplicación. Esto proporciona flexibilidad para personalizar el proceso de minería de textos y mejorar la precisión y relevancia de los resultados.

3. **Eficacia en dominios específicos**. En ciertos dominios o industrias con características particulares del lenguaje, las reglas pueden ser altamente efectivas para identificar patrones y extraer información relevante. Por ejemplo, en el ámbito legal o médico, donde se utilizan terminologías y estructuras de lenguaje específicas, las reglas pueden ser especialmente útiles.
4. **Control sobre el proceso de extracción**. Las técnicas basadas en reglas brindan un alto grado de control sobre el proceso de extracción de información. Los usuarios pueden definir reglas específicas para capturar ciertos tipos de información, lo que permite un ajuste fino del proceso de minería de textos según los requisitos del proyecto.
5. **Desempeño predecible**. Debido a su naturaleza determinística, las técnicas basadas en reglas tienden a tener un desempeño predecible y consistente en una amplia variedad de situaciones. Esto facilita la evaluación y comparación de diferentes enfoques de minería de textos.
6. **Costo computacional**. En comparación con enfoques más avanzados como el aprendizaje automático, las técnicas basadas en reglas suelen tener un menor costo computacional, lo que las hace adecuadas para aplicaciones donde se requiere eficiencia en términos de recursos computacionales.

En cuanto a las desventajas de las técnicas de minería de textos basadas en reglas podemos señalar las siguientes.

1. **Dependencia de la calidad de las reglas**. La efectividad de las técnicas basadas en reglas depende en gran medida de la calidad y exhaustividad de las reglas definidas por los usuarios. Si las reglas no capturan todos los casos relevantes o si son demasiado restrictivas, pueden perder información importante o generar resultados incompletos o inexactos.
2. **Dificultad para capturar complejidades lingüísticas**. El lenguaje natural es intrínsecamente complejo y a menudo ambiguo, lo que puede dificultar la creación de reglas exhaustivas para capturar todas las variaciones lingüísticas y contextuales. Esto puede limitar la capacidad de las técnicas basadas en reglas para manejar de manera efectiva la diversidad lingüística y semántica en diferentes tipos de texto.

3. **Mantenimiento y actualización constantes**. Mantener y actualizar las reglas a medida que cambian los datos o los requisitos del proyecto puede ser una tarea laboriosa y costosa. A medida que el corpus de texto evoluciona o se expande, las reglas pueden volverse obsoletas o necesitar ajustes para mantener su relevancia y precisión.
4. **Escalabilidad limitada**. A medida que aumenta la complejidad y el volumen de los datos de texto, puede volverse más difícil y costoso definir y mantener reglas efectivas para la minería de textos. Las técnicas basadas en reglas pueden tener dificultades para escalar para manejar grandes cantidades de datos de manera eficiente.
5. **Dificultad para generalizar**. Las reglas diseñadas para un conjunto específico de datos pueden no ser fácilmente generalizables a otros conjuntos de datos o dominios. Esto puede requerir la creación de nuevas reglas o ajustes significativos para adaptar las técnicas basadas en reglas a diferentes contextos o tipos de texto.

2.2.2.3 Técnicas de minería de textos basadas en análisis semántico.

Este grupo de metodologías se caracteriza por la categorización de los documentos en general, y de patentes en particular, que contienen el mismo significado en términos de similitud de texto, para lo cual se requiere un amplio conocimiento del ámbito o del dominio estudiado.

El primer método de este grupo es el vector basado en documentos modelo (del término inglés *Vector-based Document Model Method*) [TIRS00]. Partimos de una consulta parametrizada que nos dará como resultado un conjunto de documentos relativos a las patentes. Esta parametrización nos permite llevar a cabo la localización de documentos similares en cuanto al texto que contienen y palabras clave que los describen. No obstante, y debido al gran tamaño del vector de documentos, su representación no es demasiado eficaz. Salton y Lesk [CEITP68] ya presentaron métricas similares a las de este procedimiento utilizando el sistema SMART para efectuar una recuperación de documentos de forma totalmente automática. El sistema no se basaba en palabras clave asignadas de forma manual o en términos previamente indizados para la identificación de documentos y/o peticiones de búsqueda. Tampoco utiliza la frecuencia de

aparición de ciertas palabras o frases incluidas en el documento, sino que utilizaba una gran variedad de servicios avanzados como eran los diccionarios de sinónimos, ordenamientos jerárquicos de identificadores de temas, métodos de generación de frases, etc., a fin de obtener los localizadores de contenido necesarios y poder llevar a cabo el proceso de recuperación.

Las ventajas de este método se resumen a continuación.

1. **Representación densa de documentos**. En lugar de representar documentos como vectores dispersos de términos, como en los modelos de bolsa de palabras (del inglés *Bag of Words*), el modelo de documento basado en vectores utiliza representaciones densas que capturan mejor la información semántica y contextual de los documentos. Esto permite una representación más rica y precisa de la información contenida en los documentos.
2. **Captura de relaciones semánticas**. Al utilizar representaciones densas, el modelo de documento basado en vectores puede capturar relaciones semánticas entre palabras y documentos. Esto permite la detección de similitudes semánticas entre documentos, incluso si no comparten exactamente las mismas palabras clave.
3. **Dimensionalidad reducida**. Las representaciones densas de documentos suelen tener una dimensionalidad mucho menor que las representaciones dispersas de bolsa de palabras, lo que reduce la complejidad computacional y el consumo de recursos. Esto facilita el procesamiento y la manipulación de grandes colecciones de documentos.
4. **Mejora de la generalización**. Las representaciones densas de documentos tienden a generalizar mejor a nuevos datos, lo que significa que el modelo puede aplicarse con éxito a documentos que no estaban presentes en el conjunto de entrenamiento. Esto aumenta la capacidad del modelo para manejar documentos nuevos y variados.
5. **Incorporación de contexto**. Al utilizar representaciones densas, el modelo de documento basado en vectores puede capturar el contexto en el que se utilizan las palabras en los documentos. Esto mejora la precisión de

la representación de los documentos y facilita la extracción de información relevante.

6. **Compatibilidad con técnicas de aprendizaje automático**. Las representaciones densas de documentos son compatibles con una variedad de técnicas de aprendizaje automático, como clasificación, agrupamiento (del inglés *clustering*) y recuperación de información. Esto permite la integración del modelo de documento basado en vectores en sistemas más complejos y sofisticados para tareas específicas de procesamiento de texto.

Las desventajas de este modelo se resumen a continuación.

1. **Requiere grandes cantidades de datos de entrenamiento**. Para obtener representaciones densas de alta calidad, el modelo de documento basado en vectores generalmente requiere grandes cantidades de datos de entrenamiento. Esto puede ser un desafío en entornos donde los datos etiquetados son escasos o costosos de obtener.
2. **Dependencia de la calidad de los datos de entrenamiento**. La calidad de las representaciones de documentos aprendidas por el modelo de documento basado en vectores está directamente relacionada con la calidad y la representatividad de los datos de entrenamiento. Si los datos de entrenamiento contienen sesgos o errores, esto puede afectar negativamente el rendimiento del modelo.
3. **Sensibilidad a la dimensionalidad y al espacio de características**. La elección de la dimensionalidad y el espacio de características en el que se representan los documentos puede afectar significativamente el rendimiento del modelo. En algunos casos, encontrar la configuración óptima de dimensionalidad y espacio de características puede ser un desafío.
4. **Dificultad para capturar relaciones semánticas complejas**. Aunque el modelo de documento basado en vectores puede capturar relaciones semánticas entre palabras y documentos, puede tener dificultades para capturar relaciones semánticas complejas o abstractas. Esto puede limitar su capacidad para comprender completamente el significado de los documentos en contextos complejos.

5. **Costo computacional**. Dependiendo de la dimensionalidad de las representaciones de vectores y el tamaño del conjunto de datos, el entrenamiento y la utilización del modelo de documento basado en vectores pueden requerir recursos computacionales significativos, especialmente en aplicaciones con grandes cantidades de datos de texto.
6. **Interpretación de características latentes**. Las características latentes aprendidas por el modelo de documento basado en vectores pueden ser difíciles de interpretar y visualizar, lo que puede dificultar la comprensión del funcionamiento interno del modelo y la explicación de sus decisiones.

La siguiente técnica dentro del citado grupo corresponde a los métodos basados en corpus, entendiendo por corpus un conjunto de (fragmentos de) textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico [HCL14]. El Análisis Semántico Latente, (de la voz inglesa *Latent Semantic Analysis* (LSA)) [ILSA98], y el modelo de Análogos Hiperespaciales del Lenguaje (del vocablo inglés *Hyperspace Analogues to Language model* (HAL)) [ECSWSD98], son ejemplos conocidos de los procesos de similitud basada en corpus.

El análisis semántico latente analiza un gran corpus de texto, basado en lenguaje natural, buscando similitudes fundamentadas en relaciones estadísticas tanto entre las palabras como entre los diferentes párrafos del texto. Este análisis se fundamenta en la concepción que cuanto más parecidos son los documentos más palabras tienen en común, es decir, toma como base la semántica de las palabras. Es especialmente útil cuando se analizan textos largos. LSA utiliza la descomposición del valor singular, (del vocablo inglés *Singular Value Decomposition* (SVD)) [LHKSVD98], que puede considerarse un método para entrenamiento no supervisado de una red que asocia, de forma recíproca, dos clases de eventos mediante conexiones lineales a través de una única capa oculta. Se utiliza tanto para aprender como para representar relaciones, en textos escritos con lenguaje natural, que contienen un gran número de palabras.

Las ventajas de la utilización del LSA se resumen a continuación.

1. **Reducción de dimensionalidad**. LSA reduce la dimensionalidad del espacio vectorial de términos y documentos al capturar las relaciones semánticas latentes entre términos y documentos. Esto simplifica la representación de los documentos y facilita el procesamiento de grandes volúmenes de texto.
2. **Captura de relaciones semánticas**. LSA captura las relaciones semánticas entre términos y documentos al analizar la co-ocurrencia de palabras en contextos similares. Esto permite identificar términos sinónimos, relacionados y conceptos similares, incluso si no comparten las mismas palabras clave.
3. **Mejora de la precisión en la recuperación de información**. Al utilizar un enfoque basado en la semántica en lugar de en la coincidencia exacta de palabras clave, LSA puede mejorar la precisión en la recuperación de información al recuperar documentos relevantes que contienen términos relacionados o conceptos similares a la consulta.
4. **Reducción del ruido**. LSA puede ayudar a reducir el ruido en los datos de texto al identificar y filtrar palabras irrelevantes o poco informativas que no contribuyen al significado general de los documentos. Esto mejora la calidad de los resultados al centrarse en la información más relevante.
5. **Adaptabilidad a diferentes dominios**. LSA es una técnica general que puede aplicarse a una variedad de dominios y tipos de texto sin necesidad de conocimiento específico del dominio. Esto lo hace adecuado para aplicaciones en una amplia gama de campos, desde la recuperación de información hasta el análisis de sentimientos.
6. **Escalabilidad**. LSA es escalable y puede aplicarse eficientemente a grandes colecciones de documentos. Esto lo hace adecuado para su uso en sistemas que manejan grandes volúmenes de texto, como motores de búsqueda web o sistemas de análisis de datos.

En cuanto a las desventajas de la técnica LSA, se detallan a continuación.

1. **Sensibilidad al preprocesamiento de texto**. LSA es sensible al preprocesamiento de texto, incluyendo la tokenización, eliminación de stop words, y el proceso de stemming o lematización. La calidad del preprocesamiento puede afectar significativamente los resultados de LSA, y encontrar la configuración óptima puede requerir experimentación y ajuste.
2. **Pérdida de información semántica específica**. Debido a su naturaleza de reducción de dimensionalidad, LSA puede perder cierta información semántica específica en los documentos. Esto puede conducir a una pérdida de detalles y sutilezas semánticas en el proceso de análisis.
3. **Incapacidad para manejar polisemia y homonimia**. LSA puede tener dificultades para manejar términos polisémicos (con múltiples significados) y homónimos (términos diferentes con la misma forma, pero diferentes significados). Esto puede resultar en agrupaciones inadecuadas de términos similares o conceptos relacionados.
4. **Necesidad de grandes cantidades de datos de entrenamiento**. Para obtener representaciones semánticas precisas, LSA generalmente requiere grandes conjuntos de datos de entrenamiento. La disponibilidad y la calidad de estos datos pueden ser un desafío en ciertos dominios o industrias.
5. **Falta de interpretabilidad**. Aunque LSA puede capturar relaciones semánticas entre términos y documentos, las dimensiones latentes obtenidas pueden ser difíciles de interpretar directamente por humanos. Esto puede dificultar la comprensión de cómo se representan los documentos y los términos en el espacio semántico.
6. **No tiene en cuenta el contexto**. LSA no tiene en cuenta el contexto en el que se utilizan los términos en los documentos, lo que puede afectar su capacidad para capturar el significado completo de los documentos en contextos específicos.

El modelo de análogos hiperespaciales del lenguaje, funciona construyendo una matriz en la que los puntos representan palabras, siendo la distancia entre estos la que determina la relación existente entre los diferentes términos. Si comparamos

ambos métodos HAL ofrece peores resultados que LSA en el cálculo de similitudes en el texto.

Las ventajas del modelo HAL se exponen a continuación.

1. **Captura de relaciones semánticas**. HAL puede capturar relaciones semánticas entre palabras basadas en su co-ocurrencia en contextos similares. Esto permite identificar términos relacionados, sinónimos y conceptos similares, lo que mejora la comprensión del significado semántico del lenguaje.
2. **Representación densa de palabras**. HAL utiliza representaciones densas de palabras en un espacio vectorial de alta dimensionalidad, lo que permite una representación más rica y precisa de la información semántica de las palabras. Esto facilita el procesamiento y la manipulación de las palabras en aplicaciones de procesamiento de texto y minería de información.
3. **Incorporación de contexto**. HAL tiene en cuenta el contexto en el que se utilizan las palabras al calcular sus representaciones en el espacio vectorial. Esto mejora la precisión de las representaciones al capturar el significado y la connotación de las palabras en diferentes contextos lingüísticos.
4. **Flexibilidad en el modelado de relaciones**. HAL es flexible y puede ser adaptado para modelar diferentes tipos de relaciones semánticas, como relaciones de sinonimia, antonimia, hiperonimia e hiponimia. Esto permite una representación más completa y detallada de la estructura semántica del lenguaje.
5. **Eficiencia computacional**. Aunque HAL utiliza representaciones densas de palabras en un espacio vectorial de alta dimensionalidad, su implementación computacional puede ser eficiente y escalable, lo que lo hace adecuado para su uso en sistemas que manejan grandes volúmenes de datos de texto.
6. **Interpretabilidad**. Aunque las representaciones en el espacio vectorial de HAL pueden ser de alta dimensionalidad, pueden ser interpretadas y visualizadas para comprender mejor las relaciones semánticas entre

palabras y conceptos. Esto facilita la interpretación y la comprensión de los resultados obtenidos con HAL en aplicaciones de análisis de texto.

En lo referente a las desventajas del método HAL se resumen debajo.

1. **Dependencia del contexto**. Aunque HAL tiene en cuenta el contexto en el que se utilizan las palabras para calcular sus representaciones, sigue siendo sensible al contexto específico de los datos de entrenamiento. Esto puede limitar su capacidad para generalizar a diferentes contextos o dominios lingüísticos.
2. **Dimensionalidad alta**. Las representaciones densas de palabras en un espacio vectorial de alta dimensionalidad pueden resultar en modelos computacionalmente intensivos y con una mayor demanda de recursos. Esto puede afectar la eficiencia y la escalabilidad del modelo, especialmente en aplicaciones que manejan grandes volúmenes de texto.
3. **Interpretación y visualización complejas**. Las representaciones de palabras en un espacio vectorial de alta dimensionalidad pueden ser difíciles de interpretar y visualizar. Esto puede dificultar la comprensión de las relaciones semánticas entre palabras y conceptos, lo que limita la interpretación y la explicación de los resultados obtenidos con HAL.
4. **Necesidad de grandes cantidades de datos de entrenamiento**. Para obtener representaciones semánticas precisas, HAL generalmente requiere grandes conjuntos de datos de entrenamiento. La disponibilidad y la calidad de estos datos pueden ser un desafío en ciertos dominios o industrias.
5. **Sensibilidad a la distribución de palabras**. HAL puede ser sensible a la distribución de palabras en los datos de entrenamiento, lo que puede afectar la calidad de las representaciones aprendidas. En algunos casos, ciertas palabras pueden tener una influencia desproporcionada en las representaciones semánticas, lo que puede distorsionar los resultados.
6. **Dificultad para capturar relaciones semánticas complejas**. Aunque HAL puede capturar relaciones semánticas entre palabras basadas en su co-ocurrencia en contextos similares, puede tener dificultades para capturar relaciones semánticas complejas o abstractas. Esto puede limitar

su capacidad para comprender completamente el significado de los documentos en contextos complejos.

Dentro de este grupo de técnicas basadas en las similitudes semánticas encontramos los denominados métodos híbridos que combinan las medidas basadas en el corpus con las medidas de similitud semántica entre palabras para establecer así la afinidad de los textos. Se utiliza tanto la información semántica como la sintáctica de la frase estableciendo comparaciones con la base de datos del corpus buscando la equivalencia de las frases. Mihalcea, R., Corley, C. y Strapparava, C. [CBKBM06] propusieron un método combinado con el objeto de medir la similitud semántica de los textos, tomando como base seis diferentes métricas cuyo objeto es la explotación de la información extraída a partir de la semejanza de las palabras.

1. La similitud definida por Leacock, C. y Chodorow, M. [CLCWSS98]

$$Sim_{lch} = -\log \frac{longitud}{2 \times D}$$

Ecuación 6: Similitud de Leacock, C. y Chodorow, M.

donde longitud es la longitud del camino más corto entre dos conceptos mediante el recuento de nodos, y D es la profundidad máxima de la taxonomía.

2. La similitud de dos conceptos de Lesk se define como una función del solapamiento entre las definiciones correspondientes proporcionadas por un diccionario. Se basa en un algoritmo propuesto Lesk, M. [ASDMRD86] como solución para la desambiguación del sentido de las palabras.
3. La métrica de similitud de Wu, Z. y Palmer, M. [VSALA94] mide la profundidad de dos conceptos dados en la taxonomía WordNet [UMSRWD03], y la profundidad de la subsumidor menos común (del inglés *Least Common Subsumer* (LCS)), y combina estas cifras en una puntuación de similitud:

$$Sim_{wup} = \frac{2 \times profundidad(LCS)}{profundidad(concept_1) + profundidad(concept_2)}$$

Ecuación 7: Similitud de Wu, Z. y Palmer, M.

4. La medida introducida por Resnik, P. [UICESS95] devuelve el contenido de la información (del inglés *Information Content* (IC)) del LCS de dos conceptos:

$$Sim_{res} = IC(LCS)$$

Ecuación 8: Similitud de Resnik, P.

donde IC se define como:

$$IC(c) = -\log P(c)$$

siendo P(c) la probabilidad de encontrar una instancia con el concepto c en un corpus grande.

5. La siguiente medida es la métrica in Lin, D. [AITDS98], que se basa en la medida de similitud de Resnik, P. añadiendo un factor de normalización consistente en el contenido de información de los dos conceptos de entrada:

$$Sim_{lin} = \frac{2 \times IC(LCS)}{IC(concepto_1) + IC(concepto_2)}$$

Ecuación 9: Similitud de Lin, D.

6. La última métrica es la similitud considerada por Jiang, J.J y Conrath, D.W. [SSCSLT97]:

$$Sim_{jnc} = \frac{1}{IC(concepto_1) + IC(concepto_2) - 2 \times IC(LCS)}$$

Ecuación 10: Similitud de Jiang, J.J y Conrath, D.W.

El hecho de que sean necesarias seis métricas para calcular dicha semejanza hace que este método resulte ineficiente.

Los métodos híbridos, que combinan medidas basadas en el corpus con medidas de similitud semántica entre palabras, presentan las siguientes ventajas.

1. **Mayor precisión**. Al combinar diferentes enfoques, los métodos híbridos pueden aprovechar lo mejor de ambos mundos, mejorando la precisión en la medición de la similitud entre textos. Esto permite obtener resultados más precisos y confiables en aplicaciones como la recuperación de información, la agrupación de documentos o la detección de plagio.
2. **Robustez ante la variabilidad lingüística**. Los métodos híbridos pueden ser más robustos ante la variabilidad lingüística y las ambigüedades del lenguaje natural. Al combinar medidas basadas en el corpus con medidas de similitud semántica, pueden capturar mejor las relaciones entre palabras y conceptos, incluso en contextos lingüísticos complejos o ambiguos.
3. **Generalización a diferentes dominios**. Los métodos híbridos pueden generalizar mejor a diferentes dominios y tipos de texto. Al combinar medidas basadas en el corpus con medidas de similitud semántica, pueden adaptarse a una amplia gama de aplicaciones y escenarios, desde la búsqueda de información en la web hasta el análisis de texto en dominios especializados.
4. **Reducción del ruido**. La combinación de diferentes enfoques puede ayudar a reducir el ruido en los datos de texto y mejorar la calidad de las medidas de similitud. Al filtrar la información redundante o poco relevante, los métodos híbridos pueden proporcionar resultados más limpios y significativos.
5. **Flexibilidad y personalización.** Los métodos híbridos pueden ser altamente flexibles y personalizables según las necesidades específicas del problema o del dominio de aplicación. Los investigadores y los practicantes pueden ajustar y combinar diferentes componentes del método para adaptarse a diferentes requisitos y escenarios.
6. **Eficiencia computacional**. Aunque los métodos híbridos pueden ser más complejos que los enfoques individuales, pueden ser implementados de manera eficiente y escalable, especialmente con el avance de la tecnología y el procesamiento distribuido. Esto permite su aplicación en sistemas que manejan grandes volúmenes de datos de texto.

En lo referente a las desventajas, se enumeran a continuación.

1. **Mayor complejidad**. La combinación de diferentes enfoques puede aumentar la complejidad del método, lo que puede dificultar su comprensión, implementación y ajuste. Esto puede requerir un mayor esfuerzo en términos de desarrollo y mantenimiento del sistema.
2. **Dificultad para determinar los pesos óptimos**. La combinación de medidas basadas en el corpus y medidas de similitud semántica implica la determinación de los pesos óptimos para cada componente del método. En algunos casos, encontrar los pesos adecuados puede ser difícil y puede requerir experimentación y ajuste.
3. **Dependencia de la calidad de los datos**. Los métodos híbridos son sensibles a la calidad de los datos de entrada, incluidos los datos del corpus y las medidas de similitud semántica. La presencia de ruido, errores o sesgos en los datos puede afectar negativamente la precisión y confiabilidad de los resultados.
4. **Limitaciones en la generalización**. Aunque los métodos híbridos pueden ser efectivos en ciertos dominios o conjuntos de datos, pueden tener dificultades para generalizar a otros dominios o conjuntos de datos que difieren significativamente en términos de contenido, estilo de escritura o estructura lingüística.
5. **Costo computacional**. La combinación de diferentes enfoques puede aumentar el costo computacional del método, especialmente si implica el uso de algoritmos o técnicas computacionalmente intensivas. Esto puede afectar la eficiencia y la escalabilidad del método, especialmente en aplicaciones que manejan grandes volúmenes de datos de texto.
6. **Interpretación y explicación complicadas**. Los resultados obtenidos mediante métodos híbridos pueden ser difíciles de interpretar y explicar debido a la combinación de diferentes componentes y la complejidad del método. Esto puede dificultar la comprensión de cómo se calculan las medidas de similitud y cómo se relacionan con los textos de entrada.

Un enfoque diferente dentro de este conjunto de técnicas es el llamado vector jerárquico de palabras clave (de la acepción inglesa *Hierarchical Keyword Vector*

(HKV)). Lee, C., Song. B. y Park, Y. [HAPIR13] desarrollaron este método de análisis semántico para su uso con el apartado de las explicaciones de los inventos de los documentos completos de las patentes. El objetivo era encontrar dependencias en el texto no estructurado de esta sección, aunque se encontraron con el inconveniente de que funciona solamente con algunas tecnologías específicas.

En cuanto a las ventajas que ofrece este método encontramos las siguientes.

1. **Estructura jerárquica**. El HKV agrupa los datos en una estructura jerárquica, lo que permite visualizar y comprender las relaciones entre los diferentes grupos de manera más intuitiva. Esto facilita la exploración y la interpretación de los resultados del agrupamiento.
2. **Flexibilidad en la cantidad de clústeres**. El HKV no requiere que el usuario especifique previamente el número de clústeres. En su lugar, el usuario puede ajustar la estructura jerárquica del dendrograma mediante la selección de diferentes niveles de corte, lo que proporciona flexibilidad en la interpretación de los resultados.
3. **Captura de diferentes niveles de similitud**. El HKV puede capturar diferentes niveles de similitud entre los datos, desde similitudes más altas en grupos más grandes hasta similitudes más bajas en grupos más pequeños. Esto permite identificar estructuras complejas y detalladas en los datos.
4. **Robustez ante ruido y outliers**. El HKV puede ser robusto ante ruido y outliers en los datos, ya que utiliza una estrategia aglomerativa que fusiona gradualmente los puntos de datos en grupos más grandes. Esto permite que los outliers sean absorbidos por grupos más grandes en lugar de formar su propio clúster.
5. **No requiere especificar el número de clústeres de antemano**. A diferencia de otros métodos de agrupamiento que requieren que el usuario especifique el número de clústeres de antemano, el HAC no tiene esta limitación. Esto hace que el HAC sea especialmente útil en situaciones donde el número de clústeres no es conocido de antemano o puede variar según el contexto.

6. **Escalabilidad**. El HKV es relativamente escalable y puede manejar grandes conjuntos de datos con eficiencia. Esto lo hace adecuado para aplicaciones que requieren el análisis de grandes volúmenes de datos, como la minería de texto en colecciones extensas de documentos.

En cuanto a las desventajas que este método presenta podemos señalar las siguientes.

1. **Sensibilidad a la métrica de distancia**. El rendimiento del HKV puede verse afectado por la elección de la métrica de distancia utilizada para calcular la similitud entre los puntos de datos. Diferentes métricas pueden producir resultados significativamente diferentes, lo que puede influir en la estructura y la interpretación de los clústeres.
2. **Costo computacional**. El HKV tiene un alto costo computacional, especialmente para conjuntos de datos grandes. Esto se debe a que el algoritmo debe calcular la distancia entre todos los pares de puntos de datos y fusionar los clústeres en cada paso, lo que puede ser computacionalmente costoso, especialmente para conjuntos de datos grandes.
3. **Sensibilidad a la inicialización**. El HKV es sensible a la inicialización, lo que significa que diferentes configuraciones iniciales pueden llevar a diferentes resultados finales. Esto puede hacer que el algoritmo sea menos robusto y reproducible en comparación con otros métodos de agrupamiento.
4. **Dificultad para manejar conjuntos de datos grandes.** Debido a su alto costo computacional, el HKV puede tener dificultades para manejar conjuntos de datos grandes o de alta dimensionalidad. Esto puede hacer que el algoritmo sea menos práctico en aplicaciones que involucran grandes volúmenes de datos, como la minería de texto en colecciones extensas de documentos.
5. **Falta de escalabilidad vertical.** A medida que el número de puntos de datos aumenta, el HKV puede tener dificultades para mantener su eficiencia computacional. Esto puede limitar su aplicabilidad en situaciones donde se requiere un análisis en tiempo real o una respuesta rápida a cambios en los datos.

6. **Sensibilidad a la forma de los clústeres.** El HKV es sensible a la forma de los clústeres, lo que significa que puede tener dificultades para detectar clústeres de formas no convencionales o de densidades variables. Esto puede resultar en la formación de clústeres subóptimos o la división incorrecta de clústeres en estructuras complejas.

El método conocido como similitud semántica del texto [STSCBWS08], es utilizado para encontrar similitudes entre palabras y textos combinando la similitud de cadenas, la similitud semántica (palabras con el mismo significado) y la similitud de orden de palabras comunes con normalización. El hecho de utilizar la similitud entre cadenas de palabras hace que requiera una menor cantidad de búsquedas en la base de datos del corpus, lo cual hace que el tiempo de respuesta sea bajo aumentando así la eficacia del método. Así mismo, se reduce la necesidad de la dependencia del dominio en el que se encuentra la palabra clave jerarquizada.

2.2.2.4 Enfoques de visualización.

El último gran bloque son las técnicas basadas en redes neuronales, descrito por Sarkar, Nasipuri y Ghose [NAKNN10], cuyo enfoque está basado en la extracción de frases clave. Este enfoque ha demostrado ser efectivo para procesar y comprender la información textual, plasmando la complejidad semántica y sintáctica del lenguaje natural.

Las redes neuronales, como las redes neuronales recurrentes (del inglés *Recurrent Neural Network* (RNN)) [RNRDS21], las redes neuronales convolucionales (del término inglés *Convolutional Neural Network* (CNN)) [CNNAAP21] o incluso modelos más avanzados, son utilizadas para aprender representaciones de palabras y secuencias de texto. Estas redes son capaces de capturar la estructura y las relaciones complejas entre las palabras y oraciones en el texto, permitiendo así una comprensión más profunda y contextualizada del contenido.

Los enfoques de visualización basados en redes neuronales presentan varias ventajas que se enumeran a continuación.

1. **Capacidad para capturar estructuras no lineales**. Las redes neuronales tienen la capacidad de capturar relaciones no lineales entre

datos, lo que les permite modelar estructuras complejas y no lineales en conjuntos de datos de alta dimensión. Esto les permite representar y visualizar datos de manera más precisa y efectiva.

2. **Representaciones de alta dimensión**. Las redes neuronales pueden procesar y aprender representaciones de alta dimensión de los datos de entrada. Esto les permite capturar características y relaciones intrínsecas en los datos que pueden ser difíciles de visualizar con enfoques lineales o de baja dimensión.
3. **Aprendizaje no supervisado**. Los enfoques de visualización basados en redes neuronales pueden aprender de manera no supervisada, lo que significa que pueden extraer automáticamente patrones y estructuras interesantes de los datos sin necesidad de etiquetas de clase. Esto facilita la exploración y la comprensión de datos no etiquetados.
4. **Adaptabilidad a diferentes tipos de datos**. Las redes neuronales pueden adaptarse a una amplia variedad de tipos de datos, incluidos datos numéricos, de texto, de imágenes y de secuencias. Esto les permite aplicarse a una variedad de problemas de visualización en diferentes dominios y aplicaciones.
5. **Escalabilidad**. Con los avances en hardware y técnicas de entrenamiento, las redes neuronales pueden ser escaladas para manejar grandes conjuntos de datos con eficiencia. Esto las hace adecuadas para aplicaciones que involucran grandes volúmenes de datos, como la visualización de datos a gran escala.
6. **Flexibilidad en la arquitectura**. Las redes neuronales ofrecen una amplia variedad de arquitecturas y modelos que pueden adaptarse a diferentes tipos de datos y problemas de visualización. Esto proporciona flexibilidad para elegir la arquitectura más adecuada para el problema específico que se está abordando.

Las desventajas que presentan este tipo de sistemas se especifican a continuación.

1. **Complejidad computacional**. Los enfoques basados en redes neuronales pueden ser computacionalmente intensivos, especialmente para conjuntos de datos grandes o modelos muy profundos. Esto puede

requerir hardware especializado o recursos computacionales significativos para entrenar y ejecutar los modelos de visualización.

2. **Dificultad en la interpretación**. A menudo, los modelos de redes neuronales son cajas negras, lo que significa que pueden ser difíciles de interpretar y entender cómo funcionan internamente. Esto puede hacer que sea difícil explicar los resultados de la visualización a los usuarios o comprender la relación entre las características de entrada y la representación visual resultante.
3. **Necesidad de grandes conjuntos de datos de entrenamiento**. Los modelos de redes neuronales generalmente requieren grandes conjuntos de datos de entrenamiento para aprender patrones significativos en los datos. Esto puede ser un desafío en algunos casos, especialmente cuando los datos son escasos o costosos de obtener.
4. **Sensibilidad a la inicialización y los hiperparámetros**. Los modelos de redes neuronales son sensibles a la inicialización de los pesos y los hiperparámetros del modelo. En algunos casos, encontrar la configuración óptima de inicialización e hiperparámetros puede requerir experimentación y ajuste.
5. **Posibilidad de sobreajuste**. Los modelos de redes neuronales pueden ser propensos al sobreajuste, especialmente cuando se utilizan en conjuntos de datos pequeños o ruidosos. Esto puede llevar a modelos que no generalizan bien a nuevos datos y producen resultados poco confiables en entornos del mundo real.
6. **Requisitos de almacenamiento y memoria**. Los modelos de redes neuronales pueden requerir grandes cantidades de almacenamiento y memoria para almacenar los pesos del modelo y los datos intermedios durante el entrenamiento y la inferencia. Esto puede ser un desafío en sistemas con recursos limitados.

2.3 Métodos estadísticos.

En el ámbito de la predicción de la innovación, la combinación de técnicas de análisis descritas en el punto 2.2 con métodos estadísticos ha surgido como un enfoque complementario y eficaz para anticipar y comprender el desarrollo futuro

en diversos sectores. La necesidad de predecir la dirección de la innovación ha llevado a la integración de métodos estadísticos avanzados en conjunto con enfoques de análisis de datos para capturar con mayor precisión los patrones y tendencias que moldean el horizonte de la innovación.

Los métodos estadísticos ofrecen herramientas analíticas fundamentales que permiten modelar y cuantificar la probabilidad de ocurrencia de innovaciones potenciales, al tiempo que identifican relaciones entre variables, evalúan la incertidumbre y ofrecen perspectivas valiosas para la toma de decisiones estratégicas. Desde modelos de regresión hasta análisis de series temporales pasando por métodos probabilísticos más avanzados, la combinación de enfoques estadísticos complementa y fortalece las técnicas de análisis al ofrecer una profundidad adicional en la interpretación de los datos y en la generación de pronósticos más precisos.

La **Tabla 3** representa un ejemplo de esta combinación de métodos estadísticos aplicada a la predicción de tecnologías emergentes o bien a nichos de oportunidad que podamos encontrar en tecnología ya asentadas.

Tabla 3: Resumen de métodos estadísticos aplicados en conjunción con diferentes técnicas de análisis utilizadas en la predicción de la innovación.

Autor	Método	Tecnología	BBDD
Altuntas et al. [APDWAR15]	• Regla de asociación ponderada	• Aplicación de BBDD • Teoría de BBDD	UPSTO
Caviggioli [TFIAPD16]	• Bibliometría • Modelo de duración	• Todos los campos de la tecnología	EPO
Choi and Jun [VTFBPC14]	• Bibliometría • Modelo de duración	• Sistemas de robots humanoides	USPTO
Daim et al. [FETBPA06]	• Bibliometría • Método Delphi • Curva de crecimiento	• Seguridad alimentaria • Células de combustible • Almacenamiento óptico	USPTO
Joung and Kim [METKBA17]	• Algoritmo de agrupamiento jerárquico • Análisis de tendencia de palabras clave • TF-IDF	• Biosensor electroquímico de glucosa	USPTO
Jun et al. [PMTF12]	• Minado de reglas de asociación • Agrupación K-means • Análisis de series de tiempo	• Biotecnología	USPTO

Autor	Método	Tecnología	BBDD
Jun et al. [TFMMPC12]	• Agrupación K-medoides • Mapa de matriz • Agrupación de vectores de apoyo	• Gestión de la tecnología	USPTO
Kim and Bae [NAFTPA17]	• Bibliometría • Agrupación K-means	• Cuidado del bienestar	USPTO
Kim et al. [TFTBPA15]	• Distribución latente de Dirichlet • Análisis de componentes principales	• Energías renovables	USPTO
Kyebambe et al. [FETSLPA17]	• Aprendizaje supervisado • Ciclo de vida tecnológico	• Todos los campos de la tecnología	USPTO
Lee et al. [ADTO09]	• Mapa de patentes • Análisis de componentes principales	• Asistente personal digital (PDA)	USPTO
Lee et al. [EIETMLPI18]	• Red neuronal multicapa	• Farmacéutico	USPTO
Song et al. [DNTOP17]	• Similitud del coseno • Curva S	• Sistemas de frenado	JPO
Trappey et al. [UPDFTF11]	• Agrupación por el contenido de la patente • Ciclo de vida tecnológico	• RFID	CNIPA
Yoon and Magee [ETOPITM18]	• Mapeo topológico generativo • Predicción de enlaces	• Impresión 3D • Fusión nuclear • Purificación del agua	USPTO

La utilización de los métodos estadísticos, por sí mismos, como técnica de predicción de la innovación presenta las siguientes ventajas.

1. **Objetividad y rigor**. Los métodos estadísticos se basan en datos concretos y análisis rigurosos, lo que permite obtener predicciones objetivas y confiables.
2. **Identificación de patrones**. Los métodos estadísticos permiten identificar patrones y tendencias en los datos, lo que puede ser útil para predecir futuras innovaciones.
3. **Cuantificación del riesgo**. Los métodos estadísticos permiten cuantificar el riesgo asociado a las predicciones de innovación, lo que ayuda a tomar decisiones informadas.
4. **Flexibilidad**. Existen diversos métodos estadísticos disponibles, lo que permite elegir el más adecuado para cada caso específico.

5. **Integración con otras técnicas**. Los métodos estadísticos pueden integrarse con otras técnicas de predicción de la innovación, como el análisis de expertos o el aprendizaje automático.
6. **Mejora continua**. Los métodos estadísticos pueden refinarse y actualizarse a medida que se dispone de nuevos datos, lo que permite mejorar la precisión de las predicciones.
7. **Comunicación efectiva**. Los resultados de los análisis estadísticos pueden comunicarse de manera efectiva a través de tablas, gráficos y otros recursos visuales.
8. **Evidencia para la toma de decisiones**. Los métodos estadísticos proporcionan evidencia sólida para respaldar las decisiones relacionadas con la innovación.
9. **Asignación eficiente de recursos**. Las predicciones de innovación basadas en métodos estadísticos pueden ayudar a asignar los recursos de manera eficiente a las áreas con mayor potencial de innovación.
10. **Fomento de una cultura de innovación**. El uso de métodos estadísticos para la predicción de la innovación puede fomentar una cultura de innovación dentro de una organización.

En cuanto a las limitaciones podemos encontrar las siguientes.

1. **Dependencia de datos históricos**. Los métodos estadísticos a menudo se basan en datos históricos para predecir eventos futuros. Esto puede ser problemático en el caso de la innovación, donde los cambios pueden ser rápidos e impredecibles, y los datos históricos pueden no reflejar completamente las tendencias emergentes.
2. **Limitaciones en la captura de factores cualitativos**. Los métodos estadísticos suelen basarse en datos cuantitativos y pueden tener dificultades para capturar factores cualitativos importantes que influyen en la innovación, como la creatividad, la visión empresarial y el entorno cultural.
3. **Sesgo en los datos históricos**. Los datos históricos pueden estar sesgados hacia ciertos tipos de innovación o industrias, lo que puede llevar

a predicciones inexactas si las condiciones cambian o surgen nuevos factores influyentes.

4. **Falta de flexibilidad**. Los modelos estadísticos suelen ser menos flexibles que otros enfoques de predicción, lo que puede limitar su capacidad para adaptarse a cambios rápidos o eventos inesperados en el entorno de innovación.
5. **Riesgo de subestimación de la disrupción**. Los métodos estadísticos pueden tener dificultades para predecir la aparición de innovaciones disruptivas que cambian radicalmente un mercado o industria existente, ya que estas innovaciones suelen surgir de manera impredecible y pueden no tener precedentes históricos.
6. **Interpretación limitada de resultados**. Los modelos estadísticos pueden producir resultados difíciles de interpretar, especialmente cuando se utilizan en conjunción con conjuntos de datos complejos o de gran escala. Esto puede dificultar la identificación de los factores clave que impulsan la innovación.

2.4 Inteligencia artificial.

El campo de la inteligencia artificial (IA) ha evolucionado mucho desde sus inicios hasta convertirse en un campo imprescindible en nuestros días. La definición de inteligencia artificial y el contenido de esta ha ido cambiando con el tiempo.

Concretamente, Kaplan y Haenlein definen la IA como *la capacidad de un sistema para interpretar correctamente datos externos, aprender de esos datos y utilizar esos aprendizajes para lograr objetivos y tareas específicas mediante una adaptación flexible.* [SSIMH19]. Poole y Mackworth indican que es *el campo que estudia la síntesis y el análisis de agentes computacionales que actúan de forma inteligente,* [AIFCA10].

Teniendo en cuenta el concepto de agente introducido en la definición anterior, un agente es algo (o alguien) que actúa [WIAI21]. Un agente es inteligente cuando:

1. sus acciones son adecuadas a sus circunstancias y objetivos
2. es flexible ante entornos y objetivos cambiantes

3. aprende de la experiencia, y
4. toma las decisiones adecuadas dadas sus limitaciones perceptivas y computacionales.

Otros autores, Russell y Norvig, definen la inteligencia artificial como *el estudio de agentes [inteligentes] que reciben preceptos del entorno y actúan. Cada agente de este tipo se implementa mediante una función que asigna percepciones a acciones, representándose las funciones de formas diferentes, como pueden ser los sistemas de producción, los agentes reactivos, los planificadores lógicos, las redes neuronales y los sistemas teóricos de decisión,* [AIMA10].

De esta manera, ya en 1950, Alan Turing [AT14] sugirió que *sería posible establecer si una máquina es inteligente o no basándose en su capacidad para mostrar un comportamiento inteligente que no se distinguiera del comportamiento de un ser humano inteligente*. Turing describió un agente conversacional que sería entrevistado por un humano. Si el humano era incapaz de determinar si la máquina era o no una persona, se consideraría que la máquina había superado la prueba. El Test de Turing marcó un importante intento de evitar términos vagos mal definidos como *pensar* y, en su lugar, definió la inteligencia artificial con respecto a una tarea o actividad comprobable.

Posteriormente, John Searle separó la inteligencia artificial en dos áreas distintas. Por un lado, una inteligencia artificial débil que se limita a una sola tarea única que está perfectamente definida. La mayoría de los sistemas actuales pertenecen a esta categoría. Son sistemas se pretenden resolver un único problema, tarea o cuestión y, por lo general, no son aplicables a otros problemas ni, aunque exista relación entre ellos. Por otro lado, Searle define la inteligencia artificial fuerte de la siguiente forma: *El ordenador adecuadamente programado con las entradas y salidas correctas tendría por tanto una mente exactamente en el mismo sentido en que los seres humanos tienen mentes,* [MBP80].

Existen muchos tipos de sistemas de inteligencia artificial. Entre los más importantes podemos destacar:

- **Procesamiento del lenguaje natural**. El procesamiento del lenguaje natural ayuda a las máquinas a comunicarse con las personas en su propio idioma, realizando además otras tareas relacionadas con el lenguaje. El procesamiento del lenguaje natural hace posible, por ejemplo, que los ordenadores lean textos, escuchen el habla, la interpreten, midan pensamientos y emociones y determinen qué partes son importantes.
- **Visión**: En los últimos años, el coste de adquisición e identificación de grandes conjuntos de datos se ha reducido gracias a los avances en el Internet Industrial de las Cosas (de la acepción inglesa *Industrial Internet od Things* (IIoT)), haciendo que el aprendizaje automático sea más accesible para las aplicaciones de inspección y reconocimiento, que es la forma en que se utiliza la inteligencia artificial en los sistemas de visión.
- **Vehículos autónomos**. Los coches autónomos obtienen datos de su entorno, que son introducidos y procesados por medio de un agente inteligente, que toma decisiones y permite a un vehículo autónomo realizar las actividades específicas necesarias para la conducción.
- **Aprendizaje automático**. El aprendizaje automático se corresponde con una categoría de algoritmos, generalmente estadísticos, que permite a las aplicaciones informáticas llevar a cabo predicciones de forma más o menos precisa y específica sin necesidad de programarlas explícitamente.

La siguiente figura, **Figura *11***, muestra una clasificación de los tipos de inteligencia artificial existentes. Entre las grandes áreas encontramos el procesamiento del lenguaje natural que se subdivide en (i) generación de texto, (ii) respuesta a las preguntas, (iii) traducción automática, (iv) clasificación y (v) extracción de contenido. Así mismo, el aprendizaje automático (que será desarrollado posteriormente en la sección 2.4.1) representa un área de gran interés que se distribuye en (i) aprendizaje supervisado, (ii) aprendizaje no supervisado, (iii) aprendizaje semisupervisado o híbrido y (iv) aprendizaje profundo.

Por otro lado, otras áreas como planificación, sistemas expertos, robótica, voz y visión cobran cada día más importancia debido a sus campos de aplicación.

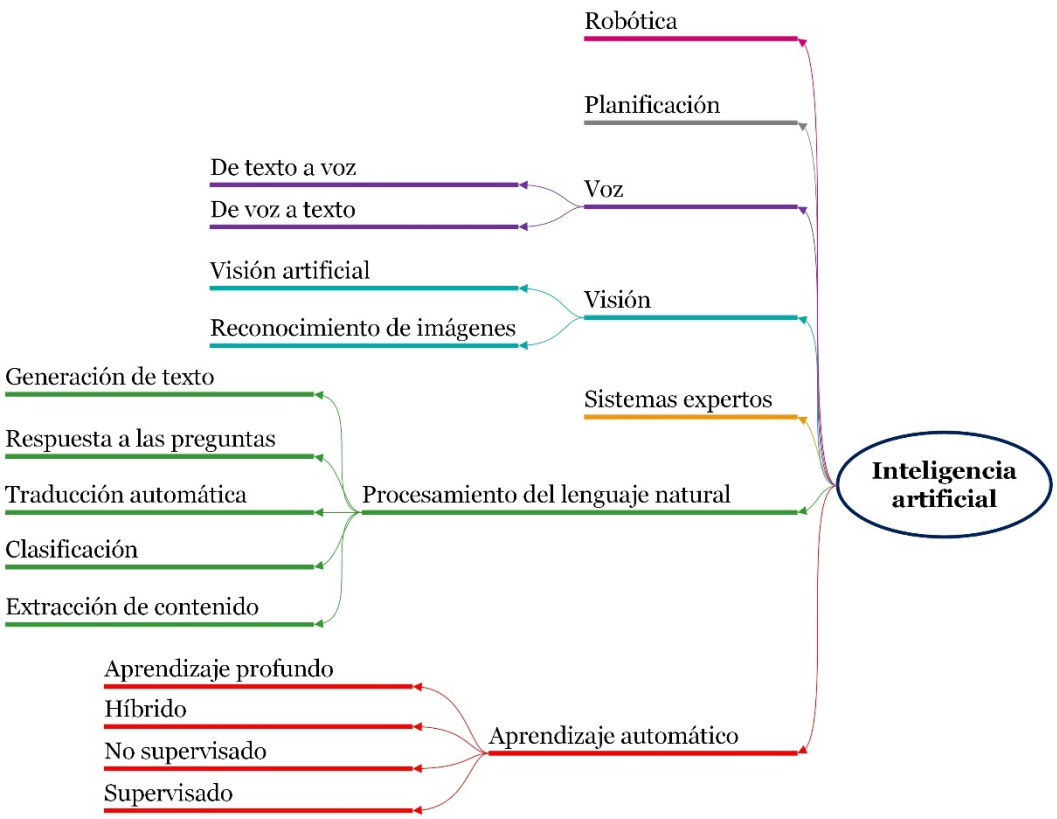

Figura 11: Clasificación de los diferentes tipos de inteligencia artificial existentes en la actualidad.

Otro aspecto muy importante de la inteligencia artificial es el relativo a sus implicaciones éticas. El grupo AI4People del Atomium European Institute, establece una serie de principios aplicados a entornos digitales [EFGAIS21]. A modo de resumen, podemos establecer los siguientes puntos.

1. **Marcos éticos**: Desarrollar directrices y marcos éticos para guiar el desarrollo y la aplicación de la inteligencia artificial, haciendo hincapié en la equidad, la transparencia, la responsabilidad y la inclusión.
2. **Concienciación y educación del público**: Concienciar al público sobre el potencial y los retos de la inteligencia artificial, fomentar la comprensión y promover debates informados sobre su impacto social.
3. **Recomendaciones políticas**: Proporcionar recomendaciones a los responsables políticos y a los organismos reguladores para desarrollar leyes

y reglamentos adecuados que rijan el uso de la inteligencia artificial, garantizando su alineación con los valores y objetivos sociales.

4. **Inclusión y diversidad**: Fomentar diversas perspectivas e inclusividad en el desarrollo de la inteligencia artificial para minimizar los prejuicios y mejorar la comprensión y representación de todos los individuos por parte de los sistemas de inteligencia artificial.
5. **Privacidad y protección de datos**: Abogar por medidas sólidas de privacidad de datos y enfatizar la importancia de asegurar y proteger los datos de los individuos en las aplicaciones de inteligencia artificial.
6. **Colaboración e intercambio de conocimientos**: Facilitar la colaboración entre las partes interesadas, compartiendo las mejores prácticas, los resultados de la investigación y los conocimientos para fomentar un esfuerzo colectivo hacia la inteligencia artificial responsable.

En el artículo [SSEGAIS23] de *Google DeepMind* se presenta un marco sociotécnico de tres capas para la evaluación de la seguridad de los sistemas de IA generativa, que son también aplicables a otros tipos de IA. Dichas capas son:

1. **Capa 1**: Capacidad.
2. **Capa 2**: Interacción humana.
3. **Capa 3**: Impacto sistémico.

Las capas no son secuenciales ni dependientes entre sí, pero su conexión va añadiendo progresivamente más contexto. Las áreas de riesgo de alto nivel son detectables y pueden evaluarse en cada capa. La integración de los resultados de cada capa proporciona una nueva capa de seguridad en un sistema de IA generativa.

La siguiente figura, **Figura 12**, muestra un esquema general de lo que se conoce como inteligencia artificial, en la cual una parte importante corresponde al aprendizaje automático, que se subdivide en cuatro categorías, aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo y aprendizaje profundo. Cada una de estas categorías se clasifica en métodos que, a su vez, se reparte en diferentes modelos.

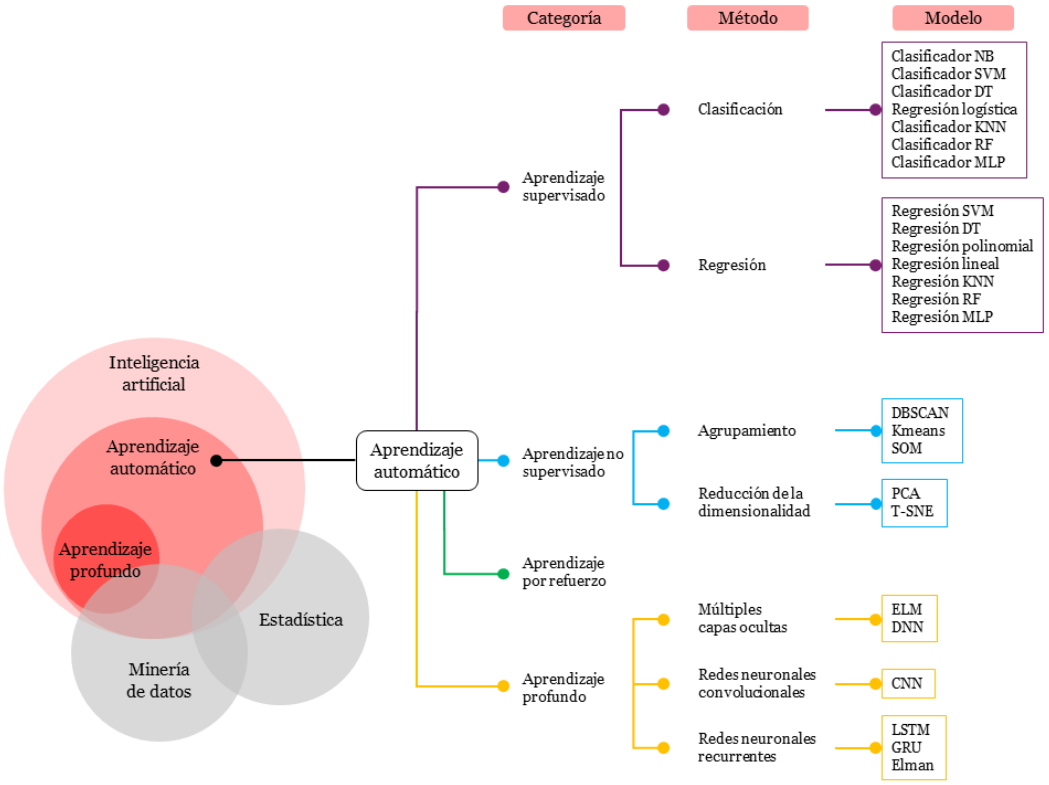

Figura 12: Clasificación esquemática de los diferentes tipos de aprendizaje automático que muestra su encaje en el entorno de la inteligencia artificial y los métodos y modelos utilizados en cada caso.

2.4.1 Aprendizaje automático.

El aprendizaje automático (del término inglés *Machine Learning* (ML)) constituye un subcampo de la inteligencia artificial que se centra en la creación de algoritmos que utilizan la experiencia con respecto a una clase de tareas y la retroalimentan con medidas con el objetivo de mejorar el rendimiento de esta [AIID96].

Como podemos ver en la **Figura 12**, se definen varias categorías de aprendizaje automático como son el supervisado, el no supervisado, el aprendizaje por refuerzo y el aprendizaje profundo (del inglés *Deep Learning* (DL)). Existe además un tipo de aprendizaje semisupervisado o híbrido que combina técnicas de ambos métodos, supervisado y no supervisado.

De igual forma, el aprendizaje automático supervisado se divide en dos métodos que son la clasificación y la regresión dando lugar a un conjunto de modelos

diferentes. Dentro del conjunto de modelos de clasificación encontramos, por ejemplo, la regresión logística [DABTIF22] que no es sino una adaptación de la regresión lineal. Dentro de este conjunto de modelos encontramos los árboles de decisión (del vocablo inglés *Decision Tree* (DT)) [DTEPRAI23] y los bosques aleatorios (del vocablo inglés *Random Forest* (RF)) [DDSMLA22]. Se construyen en base a reglas de decisión que facilita su representación en forma de árbol, lo cual hace más sencilla su interpretación. Son modelos precisos y que pueden representar relaciones complejas.

Además, en lo referente a métodos de regresión quizás el más conocido sea la regresión lineal [MLIRC19] (o regresión de mínimos cuadrados). Se trata de modelos bastante simples que no ofrecen buenos resultados para conjuntos de datos con presentan comportamientos más complejos.

De la misma forma, el aprendizaje automático no supervisado se subdivide en dos métodos, agrupamiento y reducción de la dimensionalidad. Su finalidad es encontrar patrones intrínsecos en los datos y conseguir una estructura que resulte de utilidad [SRSUML20]. Uno de los algoritmos más comunes de agrupamiento es el KMeans que agrupa los datos en un número predeterminado de clústeres por similitud, minimizando las distancias cuadradas al punto centroide del clúster asignado [UKMCA20]. Asimismo, la técnica PCA (del inglés *Principal Component Analysis*) es capaz de transformas los datos que contienen una alta dimensionalidad en otro conjunto más reducido de variables manteniendo la mayor cantidad de información posible [PCAIIS18].

En relación con las redes neuronales, están relacionadas con todo es espectro del aprendizaje automático, desde el aprendizaje supervisado, el aprendizaje no supervisado, el aprendizaje por refuerzo e incluyendo el aprendizaje profundo. Una red neuronal artificial (del vocablo inglés *Artificial Neural Network* (ANN)) es un algoritmo de aprendizaje automático inspirado en las redes neuronales biológicas [ISLWAR13] [ESLDMIP09] [APP50]. Su campo de aplicación es muy amplio incluyendo la clasificación de imágenes, textos, procesamiento del lenguaje natural y reconocimiento de voz entre otros.

Existen una gran variedad de productos y/o servicios que utilizan algoritmos de aprendizaje automático, como son los que se muestran en la siguiente lista.

- Vehículos no tripulados que se conducen solos, [AMLVRR23].
- Brazo robótico que juega al ajedrez, [GACPRS11].
- Reconocimiento facial de Facebook para identificar contactos, [FRT21].
- Microsoft Cortana, asistente personal inteligente de Microsoft para diferentes dispositivos, [NGVPA18].
- Motores de búsqueda que ofrecen información de acuerdo con las preferencias de los usuarios, [AIMLA22].
- Traducción automática usada por el traductor de Google, que reconoce palabras en más de 100 idiomas humanos, [AGBMT20].
- Google Trends, que son las tendencias de búsquedas en Google, [TARTGT20].
- Google N Gram Viewer, que indexa libros que tiene Google escaneados y sus términos gramaticales, [BDHGN16].
- Siri, que convierte conversaciones habladas a texto (del vocablo inglés *Speech To Text* (STT)), [CASIT22].

Así mismo, existen una gran cantidad de procesos que hacen uso del aprendizaje automático, como puede verse a continuación.

- Detectar fraudes en transacciones bancarias, [MLAOFD19].
- Detectar intrusiones en una red de comunicaciones de datos, [MLMNID18].
- Predecir fallos en equipos tecnológicos, [PEDP21].
- Prever qué proyectos serán más rentables el próximo año y con un menor riesgo, [RIAP21].
- Seleccionar clientes potenciales basándose en comportamientos en las redes sociales e interacciones en la web, [RBMSASM14].
- Predecir el tráfico urbano y dar rutas alternativas, [CMPTD22].
- Conocer anticipadamente qué partido político ganará las próximas elecciones analizando los comentarios de los usuarios en las redes sociales, [PBUSE20].

- Saber cuál es el mejor momento para publicar *twits*, actualizaciones de Facebook o enviar boletines (del vocablo inglés *newsletters*), [RSME18].
- Prevenir la deserción de clientes en una empresa de telefonía, [SLRCDP22].
- Predecir las ventas de los años siguientes analizando comportamiento actual de los clientes, [MLMSTSF19].
- Conocer las preferencias de los clientes a través de sus operaciones en la red, [ARWARD21].
- Hacer prediagnósticos médicos basados en síntomas del paciente, [MDUML21].
- Cambiar el comportamiento de una aplicación móvil (App) para adaptarse a las costumbres y necesidades de cada usuario, [MGCGUB15].

2.4.1.1 Aprendizaje automático supervisado.

Utilizando esta metodología, el proceso comienza entrenando a la máquina mediante datos de entrada o características asociadas a un resultado o etiqueta ya conocido, que ha sido determinado previamente por expertos humanos [MLIM19] [CAFIMLR18] [MLIM15] [MLFMI17]. Estos algoritmos pretenden obtener reglas generales que relacionan entradas y salidas [CAFIMLR18] [MLIM15].

Los procesos de aprendizaje automático constan de dos fases. En la primera fase, o fase de entrenamiento, un subconjunto del total de datos se separa para poder entrenar al algoritmo de forma sea capaz de encontrar los patrones que posteriormente servirán para hacer predicciones. La segunda fase, o fase de prueba, es aquella en la que preguntaremos al algoritmo, pudiendo evaluar si las respuestas son o no correctas, es decir, si el algoritmo está aprendiendo o no, lo que nos indicará si es necesario cambiar el método de entrenamiento utilizado.

Supongamos que un concesionario de coches quiere predecir el precio de un coche basándose en características específicas de este. El concesionario reuniría un amplio conjunto de datos relativos a los coches [ISLWAR13] [ESLDMIP09] [IML20]. Cada instancia representa una observación singular del coche y sus características asociadas. Las características no son más que las propiedades del coche que pueden resultar útiles para predecir los precios (por ejemplo, si el coche es gasolina, diésel o eléctrico, si es nuevo o de segunda mano o la marca de este).

El objetivo es la característica que se quiere predecir, en este caso el precio del coche. Los conjuntos de datos se dividen en datos de entrenamiento, validación y prueba. El aprendizaje supervisado utiliza patrones en el conjunto de datos de entrenamiento para asignar características al objetivo, de forma que un algoritmo pueda realizar predicciones del precio del coche en conjuntos de datos futuros. Este enfoque es supervisado porque el modelo deduce un algoritmo a partir de pares característica-objetivo y es informado, por el objetivo, de si la predicción se ha logrado correctamente [ISLWAR13] [NBSVU24]. Las características, x, se asignan al objetivo, Y, mediante el aprendizaje de la función de asignación, f, de modo que los precios futuros de un automóvil pueden aproximarse utilizando el algoritmo $Y = f(x)$. Los pasos del aprendizaje automático supervisado son:

1. adquirir un conjunto de datos y dividirlo en tres conjuntos, entrenamiento, validación y prueba;
2. utilizar los conjuntos de datos de entrenamiento y validación para indicar al modelo la relación existente entre las características y el objetivo; y
3. evaluar el modelo a través del conjunto de datos de prueba para determinar lo bien que se predicen los precios de los automóviles para instancias no vistas. En cada iteración, el rendimiento del algoritmo en los datos de entrenamiento se compara con el rendimiento en el conjunto de datos de validación. De este modo, el algoritmo se ajusta mediante el conjunto de validación.

Las técnicas de aprendizaje supervisado más comunes son la regresión y la clasificación. La regresión implica predecir datos numéricos, como en el ejemplo el precio del coche. La clasificación, implica predecir a qué categoría pertenece un ejemplo, es decir, una de las instancias. Si lo que queremos es deducir un rango de precios por el cual se venderá un automóvil, la variable numérica objetivo se transformará en una variable categórica dividiendo los precios de los coches en clases separadas. Estas clases serían ordinales, lo que significa que hay un orden natural asociado a las categorías. Si lo que queremos es determinar si los coches son blancos, negros o rojos, las clases serían nominales; son independientes entre sí y no tienen un orden natural.

El propósito de una regresión lineal no es otro sino encontrar relaciones y dependencias entre variables. Representa una relación de modelado entre una variable dependiente escalar continua y (etiqueta u objetivo en la terminología del aprendizaje automático) una o más (vector D-dimensional) variables explicativas (variables independientes, variables de entrada, características, datos observados, observaciones, atributos, dimensiones y punto de datos) denotadas X utilizando una función lineal. El objetivo del análisis de regresión es predecir una variable objetivo continua, mientras que el área denominada clasificación, consiste en predecir una etiqueta a partir de un conjunto finito. El modelo para una regresión múltiple implica la combinación lineal de variables de entrada y tiene la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Ecuación 11: Fórmula utilizada en el cálculo de una regresión múltiple.

La regresión lineal [TRNMLR07] pertenece a la categoría de algoritmos de aprendizaje supervisado. Esto implica, como ya se ha mencionado, que entrenamos el modelo en un conjunto de datos etiquetados (datos de entrenamiento) y luego utilizamos el modelo para predecir etiquetas en datos no etiquetados (datos de prueba).

Por otro lado, los clasificadores bayesianos [MLTPP15] son redes bayesianas utilizadas para tareas de clasificación supervisada construidas mediante aprendizaje automático. Uno de los más conocidos en el clasificador bayesiano ingenuo (del vocablo inglés *Naïve Bayes* (NB)) [PCSA73] [AAOBC92] que se basa en dos supuestos:

- Todos los atributos de una clase son independientes entre sí.
- Todos los atributos de la clase influyen sobre esta.

Este clasificador, a pesar de su gran simplicidad, presenta un comportamiento realmente bueno en cuanto a la precisión de las predicciones realizadas. Obtiene mejores resultados que otros clasificadores más complejos, especialmente cuando no existe una fuerte correlación entre los atributos. [AAOBC92] [PSIRCT94].

La probabilidad de que el j–ésimo ejemplo pertenezca a la clase i-ésima de la variable a clasificar C, puede calcularse, sin más que aplicar el teorema de Bayes [ETSPSC91], de la siguiente manera:

$$P\left(C = c_i \middle| X_1 = x_{i_j}, \cdots, X_p = x_p\right) \propto P(C = c_i) \times P\left(X_1 = x_{i_j}, \cdots, X_p = x_p \middle| C = c_i\right)$$

Ecuación 12: Teorema de Bayes.

Cuando las variables predictoras son condicionalmente independientes de la variable C, se obtiene:

$$P\left(C = c_i \middle| X_1 = x_{i_j}, \cdots, X_p = x_p\right) \propto P(C = c_i) \times \prod_{r=1}^{P} P\left(X_r = x_{r_j} \middle| C = c_i\right)$$

Ecuación 13: Teorema de Bayes con variables predictoras independientes de la variable C.

Si bien el clasificador *Naïve Bayes* es muy simple y teniendo en cuenta la restricción que este modelo presenta respecto a los atributos, existe una gran cantidad de literatura [ISBC94] que muestra el alto grado de eficacia que se obtiene en diferentes ámbitos, especialmente en medicina.

Existen además del *Naïve Bayes* otro conjunto de clasificadores bayesianos como son.

- Estructura de árbol aumentado (del vocablo inglés *Tree Augmented Naïve Bayes* (TAN)) que elimina la restricción de la independencia de los atributos en aquellos casos en los cuales la correlación es fuerte [BNC97]. Se obtiene mejores resultados que con *Naïve Bayes* y sigue siendo un método robusto que no penaliza especialmente la simplicidad de cálculo del método anterior.
- Clasificador bayesiano k-dependiente (del vocablo inglés *K-dependence Bayesian network classifier* (KDB)). Este clasificador KDB [KDBNC20] puede lograr un equilibrio entre la precisión de la clasificación y la complejidad estructural por medio del ajuste del parámetro k, que indica un mayor o menor grado de dependencia, por lo que ha recibido gran

atención aplicándose con éxito en el campo del aprendizaje automático y la minería de datos.

- Naïve Bayes aumentado a red bayesiana (del vocablo inglés *Bayesian Network augmented naïve Bayesian classifier* (BAN)) [CBNC13]. Este clasificador elimina la restricción impuesta por la cual los atributos eran independientes entre sí para una clase dada.
- Semi Naïve Bayes. Se intenta mejorar la condición de independencia entre las variables predictoras que posee el clasificador Naïve Bayes, sin complicar la estructura existente entre estas. Los modelos propuestos por Kononenko [SNBC91] y Pazzani [SFDBC95] mezclan distintos atributos correlacionados en nuevos atributos compuestos. Como sucede con el clasificador Naïve Bayes, los atributos compuestos se consideran independientes entre ellos para una clase dada, pero no se consideran independientes aquellos atributos que forman atributo compuesto.
- Naïve Bayes selectivo (del vocablo inglés *Selective Naïve Bayes*). Se pretende cambiar el hecho de que todos los atributos tienen influencia en una clase dada, utilizándose solo aquellos que son relevantes [ISBC94] [IFSSP94].
- Clasificadores bayesianos sin restricciones estructurales. Se trata en este punto de eliminar las restricciones acerca de la estructura, utilizando cualquier red bayesiana como clasificador. Este tipo de clasificadores son conocidos como redes bayesianas sin restricciones (del vocablo inglés *Unrestricted Bayesian Networks* (UBN)) [BNC97] o redes bayesianas generales (del vocablo inglés *General Bayesian Network* (GBN)) [CBNC13].
- Multiredes bayesianas. Las multiredes bayesianas no son más que una extensión de las redes bayesianas que nos permiten representar independencias asimétricas. Heckerman en [PSN90] distingue entre dos tipos de independencias asimétricas, las de subconjunto (relación entre la variable a clasificar y los atributos) y las de hipótesis específica (relación entre atributos únicamente).

- Árbol Naïve Bayes (del vocablo inglés *Naïve Bayes Tree* (NBTree))., propuesto por Kohavi [SUANBC96] y se puede considerar un tipo especial de multired bayesiana recursiva donde las hojas son clasificadores Naïve Bayes.

Además de los modelos descritos, tenemos que un perceptrón es también un algoritmo de clasificación que toma una serie de características y sus objetivos como entrada e intenta encontrar una línea, un plano o un hiperplano que separe las clases en un espacio bidimensional, tridimensional o hiperdimensional, respectivamente [ESLDMIP09] [LRBPE86] [PNPTBM62]. Estas características se transforman utilizando la función sigmoidea [SWLNNF18]. Se aplica, por ejemplo, en el reconocimiento de imágenes y videos, que constituye un proceso complejo que puede representarse de manera correcta con este tipo de redes. Son algoritmos complejos y lentos de entrenar necesitándose una gran capacidad computacional.

Podemos incluir también dentro de los modelos de aprendizaje automático aquellos que tienen un comportamiento dual, es decir, pueden tener el comportamiento de un modelo de regresión o bien puede actuar como clasificador en función del problema que se esté abordando y de la configuración del algoritmo. Dentro de este grupo tenemos el algoritmo de K vecinos más próximos (del vocablo inglés *K-Nearest Neighbour* (KNN)) [DANDSSP52]. Se trata de un algoritmo utilizado en casos simples que basa su método de clasificación en la clase a la que corresponden la mayoría de los casos más cercanos a la instancia desconocida, la que va a ser clasificada.

En igual forma, las Redes Neuronales Artificiales (del término inglés *Artificial Neural Networks* (ANN)) son modelos matemáticos utilizados en el aprendizaje automático supervisado que tratan de imitar el comportamiento y funcionalidad de las redes neuronales biológicas, es decir, el cerebro humano [NNFPR95]. Cada red neuronal artificial contiene nodos que se comunican con otros nodos mediante conexiones. Las conexiones entre nodos de una red neuronal artificial se equilibran en función de su capacidad para proporcionar un resultado deseado.

Cuando se abordan tareas de reconocimiento de imágenes, los datos de entrada de la red neuronal artificial corresponden a cada uno de los píxeles de la imagen. En

este contexto no hay conexiones entre los nodos de una capa, es decir, pierde el contexto espacial de las características de la imagen [DL15] [ICDCNN12] [RDDNN06]. En otras palabras, es probable que los píxeles cercanos entre sí en una imagen estén más correlacionados que los píxeles situados en lados opuestos de la imagen, pero una ANN no tiene esto en cuenta.

Por su parte, una red neuronal convolucional (del inglés *Convolutional Neural Network* (CNN)) es un caso especial de red neuronal artificial que supera este problema preservando la relación espacial entre los píxeles de una imagen [DL15] [ICDCNN12] [RDDNN06].

De la misma forma, las máquinas de soporte vectorial (de la expresión inglesa *Support Vector Machines* (SVM)) fueron desarrolladas por Vapnik et al. [AOASLT99] así como sus colaboradores en el marco de la teoría de aprendizaje estadístico. Son entrenadas por algoritmos de optimización convexa (existe una única solución) y construidas a partir de una estructura que depende de un subconjunto de vectores soporte, que ayudan a la interpretación del modelo.

Con respecto a la tarea de clasificación existen dos fases: la fase de aprendizaje automático y la fase de reconocimiento. En la primera se selecciona el conjunto de datos de entrenamiento, se extraen los atributos y características del espacio de entrada y se entrena el clasificador. El resultado del entrenamiento es un conjunto de parámetros que definen al clasificador y a la función discriminante que representa la frontera entre clases o regiones. En la fase de reconocimiento, el modelo del clasificador entrenado asigna a los nuevos datos de entrada una de las clases según la similitud de sus características [AMSVEP17].

En igual forma, un árbol de decisión (del inglés *Decision Tree* (DT)) es una técnica de aprendizaje supervisado, utilizada principalmente para tareas de clasificación, pero también puede utilizarse para regresión [ISLWAR13] [ESLDMIP09]. Un árbol de decisión comienza con un nodo raíz, el primer punto de decisión para dividir el conjunto de datos, y contiene una única característica que divide mejor los datos en sus respectivas clases [ISLWAR13] [ESLDMIP09]. Cada división tiene una arista que conecta con un nuevo nodo de decisión que contiene otra característica para dividir aún más los datos en grupos homogéneos o con un nodo

terminal que predice la clase. Este proceso de separación de datos en dos particiones binarias se conoce como partición recursiva [ISLWAR13] [ESLDMIP09].

Finalmente, un bosque aleatorio (del término inglés *Random Forest* (RF)) es una extensión del método anterior, conocido como método de conjunto, que produce múltiples árboles de decisión [ISLWAR13] [ESLDMIP09]. En lugar de utilizar cada característica para crear cada árbol de decisión en un bosque aleatorio, se utiliza una submuestra de características para crear cada árbol de decisión. A continuación, los árboles predicen un resultado de clase, y el voto mayoritario entre los árboles se utiliza como predicción de clase final del modelo [IMLNNDL20].

2.4.1.2 Aprendizaje automático no supervisado.

En los procesos en el aprendizaje automático no supervisado, a diferencia del caso anterior, no se proporciona ningún tipo de información al sistema de forma inicial. Se proporcionan grandes cantidades de datos no etiquetados siendo el sistema, mediante la técnica adecuada, el responsable de encontrar tendencias o patrones ocultos separando automáticamente la información en los grupos correspondientes [BDAMLMI19] [MLIM19] [CAFIMLR18] [MLIM15].

La agrupación jerárquica (de la expresión inglesa *Hierarchical Clustering* (HC)) [HC16] es un método de análisis utilizado en minería de datos que crea una representación jerárquica de los clústeres de un conjunto de datos proporcionado. El método comienza tratando cada punto de datos como un clúster independiente y, a continuación, combina iterativamente los clústeres más cercanos hasta que se llega a un punto en el que el proceso se detiene. El resultado de la agrupación jerárquica es una estructura en forma de árbol, denominada dendrograma, que muestra las relaciones jerárquicas establecidas entre los clústeres.

La agrupación jerárquica tiene ciertas ventajas sobre otros métodos de agrupación.

- La capacidad de utilizar clústeres no convexos y clústeres de diferentes tamaños y densidades.
- La capacidad de utilizar datos ausentes en el clúster y datos que producen ruido.

- Posibilita la visualización de la estructura jerárquica de los datos, que puede ser útil para comprender las diferentes relaciones entre los clústeres.

Otro algoritmo de aprendizaje automático no supervisado es el agrupamiento de K-medios (del vocablo ingles *K-Means Clustering*) que realiza la división de objetos en clústers que comparten similitudes y son disímiles a los objetos que pertenecen a otro clúster [UKMCA20] [AEKMCA02]. El término "K" es un número que indica al sistema cuántos clústeres debe crear.

Las ventajas de este algoritmo son:

- ***Sencillo y fácil de aplicar***. El algoritmo K-means es fácil de entender e implementar, lo que lo convierte en una opción muy utilizada en tareas de agrupamiento.
- ***Rápido y eficaz***. K-means es un algoritmo eficiente desde el punto de vista de consumo de recursos informáticos pudiendo gestionar grandes conjuntos de datos con alta dimensionalidad.
- ***Escalabilidad***. El método K-means puede utilizarse con conjuntos grandes de datos y puede escalarse fácilmente para emplear conjuntos de datos aún mayores.
- ***Flexibilidad***. La técnica K-means puede adaptarse fácilmente a diferentes usos o aplicaciones y puede utilizarse con diferentes métricas y métodos de inicialización.

Así mismo existen una serie de desventajas como son:

- ***Sensibilidad a los centroides iniciales***. El algoritmo K-means es sensible a la selección inicial de centroides y puede converger a una solución insatisfactoria.
- ***Requiere especificar el número de clústeres***. Es necesario especificar el número de clústeres K antes de ejecutar el algoritmo, lo que puede resultar complicado en algunas aplicaciones.
- ***Sensible a los valores atípicos***. El método K-means es sensible a los valores atípicos, que pueden tener un impacto significativo en los clústeres resultantes.

El análisis lineal discriminante (del inglés *Linear Discriminant Análisis* (LDA)) es uno de los métodos de aprendizaje supervisado de subespacios más utilizados. Sin embargo, el LDA es infructuoso ante una situación de ausencia de etiquetas. En el artículo [ULDA19] se plantea el análisis lineal discriminante no supervisado (Un-LDA) que se formula como una optimización de objetivos unificada sin fisuras que garantiza la convergencia durante el proceso de resolución alternativa iterativa. La extensión de LDA a Un-LDA permite no sólo completar el aprendizaje no supervisado del subespacio a través de la matriz de proyección del subespacio presentada explícitamente, sino también terminar simultáneamente la agrupación e incluso la agrupación de datos fuera de la muestra a través de la matriz de transformación presentada explícitamente.

2.4.1.3 Aprendizaje automático semisupervisado o híbrido.

En el aprendizaje automático, el enfoque de hibridación ha sido un área de investigación muy utilizada para mejorar la clasificación/predicción respecto a los enfoques de aprendizaje individual [HMACSM05] [FPSOM06] [HANNGA07] [CSHNDT02] [GCBCNFS02]. En general, se basa en la combinación de dos técnicas de aprendizaje de aprendizaje automático. Por ejemplo, un modelo de clasificación híbrido puede estar compuesto por un método no supervisado, clusterización o agrupamiento, para preprocesar los datos de entrenamiento y un método supervisado, clasificación, para estudiar y entender el resultado de la agrupación o viceversa [DVHIS98].

Se han utilizado modelos híbridos de aprendizaje automático para la predicción de enfermedades cardiacas [HDPHML21]. En el citado artículo se diseñó una nueva técnica implementada mediante diferentes algoritmos de aprendizaje automático, (i) bosque aleatorio, (ii) árboles de decisión y (iii) modelo híbrido desarrollado a partir de bosque aleatorios y árboles de decisión. Los resultados experimentales mostraron una precisión del 88,7%.

Otro estudio [ESDHML19] también relacionado con el campo de la medicina trató de establecer un modelo híbrido para la detección de ataques epilépticos mediante un algoritmo genético (del vocablo inglés *Genetic Algorithm* (GA)) y optimización de enjambre o nube de partículas (del vocablo inglés *Particle Swarm Optimization*

(PSO)) para determinar los parámetros óptimos de las máquinas de soporte vectorial que clasificaban los datos de los electroencefalogramas (del vocablo inglés *electroencephalogram* (EEG)). La máquina de soporte vectorial híbrida propuesta llegó a alcanzar una precisión de clasificación de hasta el 99,38% para los conjuntos de datos de EEG, lo cual la convierte en una herramienta eficaz.

2.4.1.4 Aprendizaje por refuerzo.

Otro método de aprendizaje automático es el aprendizaje por refuerzo dónde el sistema recibe tanto datos etiquetados como no etiquetados. Dicho sistema, al interactuar con el entorno recibe respuestas por parte de este que pueden ser positivas o negativas permitiendo el perfeccionamiento de este y desarrollando caracterizaciones y clasificaciones que mejoran según avanza el proceso [BDAMLMI19] [CAFIMLR18] [MLFMI17], es decir, el ordenador aprende sin tener instrucciones expresamente dadas para ello.

En el problema del aprendizaje por refuerzo, un agente, robot, máquina u ordenador explora todo el conjunto de estrategias posibles recibiendo información sobre el resultado obtenido a partir de las decisiones que ha tomado. A partir de los datos obtenidos debe deducirse una buena política de actuación [RLIRS13].

El aprendizaje por refuerzo puede entenderse cuando se contrasta el problema con otras áreas de estudio del aprendizaje automático. En el aprendizaje supervisado [RLPCP05], se presenta directamente a un agente una secuencia de ejemplos independientes de predicciones correctas que debe realizar en diferentes circunstancias. En el aprendizaje por imitación, se proporcionan a un agente demostraciones de acciones de una buena estrategia a seguir en situaciones dadas [SRLD09] [IILRHR99].

Para ayudar a comprender el problema del aprendizaje por refuerzo y su relación con técnicas ampliamente utilizadas en robótica hay que tener en cuenta dos ejes de variabilidad del problema: la complejidad de la interacción secuencial y la complejidad de la estructura de recompensa. Esta jerarquía de problemas, y las relaciones entre ellos, es compleja, y varía en múltiples atributos siendo difícil de condensar en algo parecido a una simple ordenación lineal de los problemas.

El modelo básico de aprendizaje por refuerzo consiste en:

1. Un conjunto de estados de entorno S.
2. Un conjunto de acciones A.
3. Reglas de la transición entre los estados.
4. Reglas que determinan la recompensa inmediata escalar de una transición.
5. Reglas que describen lo que observa el agente.

Existen diferentes planteamientos para desarrollar un aprendizaje por refuerzo como pueden ser el método de Montecarlo [MDM83], métodos de diferencias temporales [LPMTD88], etc.

2.5 Sumario.

A lo largo de este capítulo se han analizado, en primer lugar, las diferentes fuentes de información relativas tanto a literatura científica como a patentes. En el primer caso encontramos diferentes bases de datos como pueden ser WoS, Scopus, Google Scholar o Dimensions siendo más relevantes las dos primeras ya que son utilizadas por la mayor parte de los autores a la hora de realizar sus investigaciones. Este trabajo utilizará ambas en diferentes fases del proceso. En cuanto a información relativa a patentes se han citado otras bases de datos como WIPO, Espacenet, USPTO, PatBase® o KIPRIS. Para el desarrollo de esta investigación se utilizará PatBase® debido a que aglutina todo tipo de patentes, establece una buena clasificación de estas y facilita en gran medida el trabajo a desarrollar sin perder la calidad del dato.

A continuación, se han revisado las diferentes técnicas utilizadas en análisis predictivo de innovación que utilizan datos relativos a artículos científicos publicados en revistas indexadas y a patentes. El primer grupo de técnicas utiliza los metadatos de las patentes para realizar las predicciones mientras que el segundo grupo utiliza técnicas de minería de datos y minerías de textos. El procesamiento del lenguaje natural, las técnicas de minería de textos basadas en reglas, las técnicas de minería de textos basadas en análisis semántico y los enfoques de visualización forman parte de este amplio grupo. Para el desarrollo de esta investigación no se considera, al menos en sus inicios, la utilización de estas

técnicas que, en general, necesitan del texto completo del documento a analizar, ya sea artículo científico o patente.

Se mencionan después un conjunto de métodos estadísticos que se utilizan también, de forma complementaria, en el ámbito de la predicción innovativa como pueden ser la regla de asociación ponderada, bibliometría, método Delphi, curvas de crecimiento, análisis de tendencias de palabras clave o análisis de series temporales indicando su ámbito de aplicación y base de datos utilizada en los estudios realizados.

Finalmente se hace una introducción a la inteligencia artificial para centrarnos en el aprendizaje automático, que se divide en, supervisado, no supervisado, híbrido y aprendizaje por refuerzo. Se examinan, las diferentes técnicas y algoritmos correspondientes a cada una de estas categorías.

Para concluir, esta investigación se desarrollará en varias fases que incluyen una aproximación bibliométrica y una creación de modelos de predicción de la innovación a partir de métodos estadísticos junto con varios algoritmos de aprendizaje automático supervisado.

3

Caso de estudio: Estampación en caliente

La tecnología de estampación en caliente (del término inglés *Hot Stamping* (HS)) ha mostrado un importante rendimiento científico en la últimos años. La actividad investigadora en este campo se ha extendido a diversas disciplinas, como la ciencia de los materiales, la mecánica, la ingeniería de procesos, la instrumentación, la física o la ingeniería de diseño de herramientas.

La industria automovilística actual requiere innovaciones continuas. En este sentido, cada día aparecen numerosas publicaciones sobre nuevos materiales o desarrollos que pueden aportar nuevas reducciones de peso o mejoras de la seguridad. En este sentido, la estampación en caliente, ha ido ganando popularidad en los últimos años por sus ventajas en este sector.

Teniendo esto en cuenta, se ha elegido la estampación en caliente como la tecnología sobre la cual desarrollar un nuevo modelo predictivo de innovación. El objetivo es ser capaces de anticipar las posibles nuevas patentes que surgirán de los trabajos de investigación relacionados con este campo.

Este capítulo está organizado de la siguiente manera. En primer lugar, en la sección 3.1, se hará una breve introducción a la tecnología de estampación en caliente en

la que se pondrá de manifiesto la importancia de esta en el sector del automóvil. A continuación, en la sección 3.2, se desarrollará una aproximación de tipo bibliométrico básico utilizando para ello la información obtenida de artículos científicos, a fin de establecer las tendencias de esta tecnología. La sección 3.3 presenta un nuevo abordaje bibliométrico basado en el rendimiento y los mapas evolutivos a fin de detectar y visualizar los temas o áreas conceptuales, así como su desarrollo a lo largo del tiempo. La sección 3.4 examina la creación de patentes con un nuevo enfoque basado en una combinación de artículos científicos y patentes susceptibles de ser analizadas, mediante la combinación tanto de métodos estadísticos como de aprendizaje automático, dando como resultado una red bayesiana y un modelo de árbol de decisión que nos permitirán identificar patrones de comportamiento de producción científica y protección de la propiedad intelectual capaces de detectar, con alta probabilidad, la probabilidad de que el investigador se dedique a patentar. Por último, la sección 3.5 contiene un pequeño resumen de todo lo analizado en este capítulo.

3.1 Una breve introducción a la tecnología de estampación en caliente.

El sector automovilístico, sumido en constante desarrollo debe dar respuesta a las distintas exigencias del mercado como son la reducción de peso de los componentes, la minimización del consumo de combustible y la emisión de CO_2 (debido en gran parte a los cambios de las políticas medioambientales), así como los diferentes avances en materia de seguridad, principalmente en lo que se refiere a impactos o choques de forma que la seguridad de los pasajeros no resulte comprometida [MTSMFA13] [IAHSS04].

En consecuencia, existen varios acuerdos y regulaciones internacionales con el fin de mantener y mejorar la seguridad [TPPQHF11]. En este sentido, están presentes las exigencias de la normativa europea Programa Europeo de Evaluación de Automóviles Nuevos (del inglés European New Car Assessment Programme (NCAP)) [TENCAP16] apoyado por varios gobiernos europeos y organizaciones de automovilismo y consumidores de varios países de la Unión Europea.

Para satisfacer todas estas demandas del mercado, el sector de la automoción ha centrado sus esfuerzos en la fabricación de componentes para la carrocería (del inglés *body-in-white* (BIW)) a partir de aceros de ultra alta resistencia (del término inglés *Ultra High Strength Steel* (UHSS)) como pone de manifiesto la **Figura 13**. El uso de este tipo de materiales ha permitido fabricar componentes con mayor resistencia y menor espesor [HSMF11].

La estampación en caliente de chapa se utilizó por primera vez en 1973 para la fabricación de herramientas agrícolas por la compañía sueca Plannja Hard Tech [HHBSNS08]. Sin embargo, esta tecnología es de reciente implantación en la industria automovilística. La razón por la que se desarrolló esta tecnología fue la necesidad de fabricar componentes de aceros de elevada resistencia y menor peso, requerimientos similares a los que en la actualidad demandan los estándares de seguridad y resistencia al choque desde el sector de automoción, junto con la necesidad de reducir la huella de carbono de los vehículos.

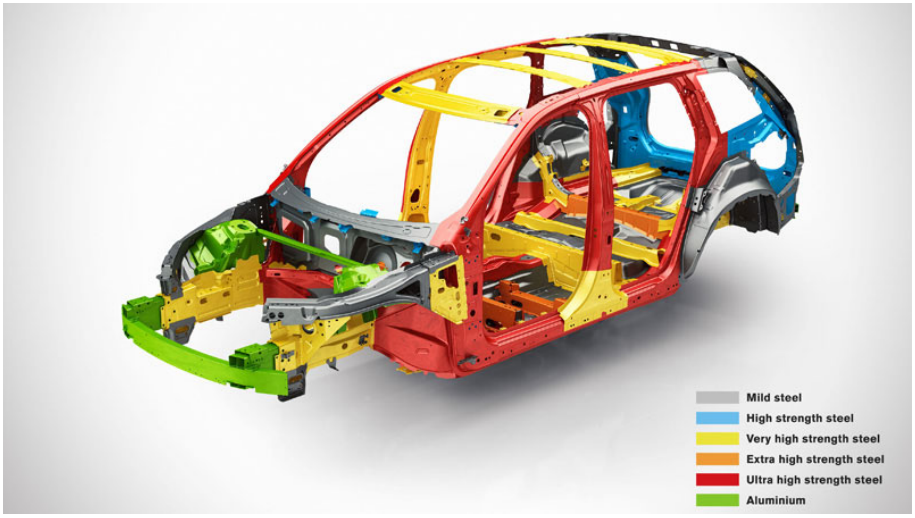

Figura 13: Evolución del uso del acero en componentes para carrocería [CZEAFC21].

Las piezas estampadas en caliente están sustituyendo paulatinamente a las piezas estampadas en frío, fabricadas con aceros de alta resistencia, para su uso en zonas críticas del automóvil, especialmente en aquellos componentes que forman parte de la carrocería. Esto implica que, hoy en día, no hay ningún diseño de carrocería,

para producciones en serie, que no tenga piezas estampadas en caliente, incluso en los modelos de gama alta. Por lo general, se observa un uso de piezas de estampación en caliente en las zonas de la carrocería que son muy sensibles a los impactos.

Por ejemplo, las vigas de impacto laterales, los pilares A y B, las barras de impacto, los parachoques delanteros y traseros, el marco del techo, los refuerzos transversales y longitudinales, así como partes del revestimiento del habitáculo delantero y trasero se fabrican con componentes estampados en caliente como se muestra en la **Figura 14**. El objetivo es mantener la integridad estructural mediante una mayor resistencia a la intrusión, una reducción del pandeo y una mayor estabilidad.

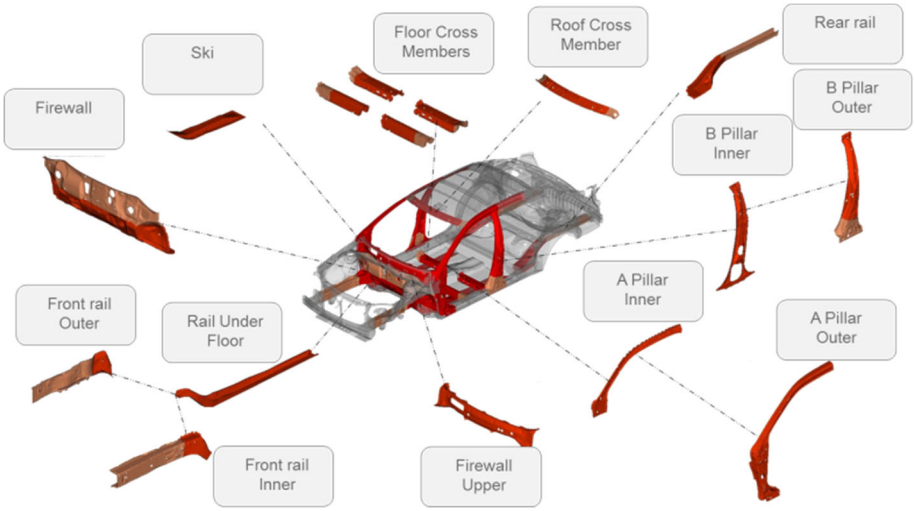

Figura 14: Componentes pertenecientes a la carrocería del automóvil fabricador con la tecnología de estampación en caliente [LSDFLW21].

En el año 1984, Saab Automobile, junto con la empresa siderúrgica sueca SSAB, con sede en Lulea, fue el primer fabricante de automóviles en producir un componente de acero al boro endurecido por medio de esta tecnología para uno de sus vehículos [HSAAA22]. El primer vehículo en el que se introdujeron estos materiales fue el SAAB 9000, concretamente en los refuerzos de las 4 puertas (del inglés *Side impact Beam*). En la **Figura 15** pueden verse las vigas de impacto lateral.

Figura 15: Primer uso de la estampación en caliente en un automóvil de serie. SAAB 9000. Viga de impacto lateral [SHFGTD21].

La mayor dificultad ha sido introducir estos componentes de acero al boro endurecido en el competitivo mundo de la producción automovilística, extremadamente exigente debido a las altas velocidades de producción y a la fiabilidad requerida [PHICT17].

Para hacerse una idea de la importancia de esta tecnología, hay que tener en cuenta que el número de piezas producidas aumentó de 3 millones de piezas al año en 1987, a 8 millones en 1997 y aproximadamente hasta los 107 millones de piezas por año en 2007 [AROHS10]. Para el 2015 se pronosticó una demanda anual de 450 millones componentes [CIHFT11]. Para el año 2018, los expertos pronosticaron un fuerte crecimiento en la demanda de componentes fabricados mediante la tecnología de estampación en caliente en el sector de automoción, con una estimación de 574 millones de piezas [FFPH21]. Dicho incremento está directamente relacionado con las ventajas en cuanto al aumento de la resistencia y disminución del espesor de la pieza fabricada y, en consecuencia, a la reducción de peso de elementos o subconjuntos estructurales que representa.

La presencia de boro en estos aceros garantiza una alta templabilidad [AROHS10]. Esta característica permitió desarrollar un proceso industrial que (i) partiendo de una chapa caliente en estado austenítico, se (ii) deforma y (iii) se templa para obtener una pieza de gran resistencia, 1500 MPa, en lo que se refiere a impactos y

choques, en definitiva, seguridad. La alta resistencia así obtenida es única, de hecho, no se puede obtener un material estampado en frío para las geometrías típicas de automoción. Actualmente se está considerando el uso de aceros de hasta 2000 MPa [HSUHSSP17] para reducir aún más el peso de los vehículos, mejorando incluso aún más la resistencia a impactos.

Gracias a la mayor resistencia del material, los diseñadores de automóviles pueden proyectar piezas con menor grosor y, por tanto, menor peso para la misma resistencia. En el Golf 5 [SHSU13], Volkswagen introdujo por primera vez piezas 22MnB5 en el subconjunto del pilar B, reduciendo el peso en 1,5 kg respecto al modelo anterior. El coche tiene tres pilares para sostener el techo y, por tanto, son clave para la seguridad en caso de vuelco, siendo el pilar B el intermedio entre las puertas. Em consecuencia, Volkswagen introdujo la tecnología a gran escala en el modelo Passat B6. La reducción de peso lograda en Volkswagen en el Passat B6 catapultó el desarrollo de la estampación en caliente al resto de fabricantes de automóviles [HFOHS19].

El objetivo para 2020 era que todos los modelos tuvieran al menos un 50 % de la carrocería, *body-in-white*, compuesta por piezas estampadas en caliente. Esta situación se refleja en las previsiones de crecimiento de esta tecnología, entre ellas las del Dr. Ralf Hund, de la empresa alemana de matrices de estampación en caliente Braum Cartec [ATSHSCC12], y las de J. Schimit, T. Lung [NDAHSA18], T. Taylor [CRAHSSS18] y J. Fekete, R. Hall [DABMP17]. Todos ellos indicaban que el número de piezas obtenidas por estampación en caliente aumentaría de unos 100 millones en 2020 a más de 600 millones de piezas en 2022. Las previsiones de Volvo indicaban, así mismo, que hasta el 50 % de la carrocería puede consistir en piezas obtenidas por estampación en caliente de aceros al boro [MPEC20].

3.2 Los últimos 10 años de estampación en caliente.

Cunando se comenzó este trabajo de investigación se tomó la decisión de realizar los análisis pertinentes sobre la tecnología de estampación en caliente. Uno de los primeros abordajes fue realizar un estudio bibliométrico que complementase las revisiones que ya existían sobre esta, es decir, las revisiones de Karbasian, H. [AROHS10] y Merklein, M. [HSBSSTP16].

La bibliometría es una ciencia que se centra esencialmente en el cálculo y en el análisis de los valores de lo que es cuantificable en la producción y en el consumo de la información científica [AESLC72] [DEBCI96]. En consecuencia, un análisis bibliométrico utiliza datos bibliográficos y estadísticos para evaluar la producción científica, su visibilidad, así como el impacto de la investigación. Debido al carácter cuantitativo de este análisis es posible evaluar aspectos como la productividad de los investigadores, detectar las tendencias que marcan la investigación y evaluar, por ejemplo, la relevancia de determinadas revistas e incluso congresos.

Un proceso de análisis bibliométrico consta de una serie de etapas que comienzan con la definición del ámbito de estudio a analizar procediendo a una recopilación de datos relacionados con este. A continuación, se seleccionará el conjunto de indicadores bibliométricos necesarios para llevar a cabo el análisis que incluirá la producción científica, las citas o las redes de colaboración. Posteriormente seleccionaremos un método de visualización de datos que nos ayudará a realizar un análisis de tendencias, evaluando además revistas y congresos. Finalmente se procederá a hacer una interpretación de los resultados obtenidos.

De hecho, existe una gran variedad de indicadores que pueden utilizarse en un análisis bibliométrico. Podemos encontrar, entre otros, los siguientes.

1. **Indicadores de producción científica**.
    a. *Número de publicaciones*. Indica el número de publicaciones, revistas, libros, etc. publicados por uno o varios autores, una o varias instituciones, un país, un periodo determinado o la combinación de todas estas variables.
    b. *Índice h* (Índice h de Hirsch) [BIB12]. Significa el número de artículos que han sido citados al menos h veces.
    c. *Índice g* (Índice g de Egghe) [TPgI06]. Cuantifica la productividad bibliométrica basada en el historial de publicaciones. Se calcula ordenando en orden descendente los artículos producidos por un investigador según el número de citas recibidas y elevando al cuadrado el orden del artículo.

2. **Indicadores de citación**.
    a. *Número total de citas*. Proporciona la cantidad de citas recibidas por un autor, institución, artículo o libro en particular, etc.
    b. *Promedio de citas por artículo*. Representa el valor promedio de citas que recibe un artículo publicado en un periodo concreto.
    c. *Índice de citas*. Indica la relación entre las citas recibidas y las esperadas aplicado a un conjunto de publicaciones.
3. **Indicadores de colaboración**.
    a. *Índice de colaboración*. Expresa la colaboración entre autores, instituciones o países. Se expresa como un porcentaje que indica el grado de colaboración.
    b. *Factor de impacto de colaboración*. Evalúa tanto la calidad como la visibilidad de la colaboración realizada.
4. **Indicadores de revistas y congresos**.
    a. *Factor de impacto*. Mide la frecuencia en que los artículos publicados en una revista son citados para un periodo de tiempo establecido.
    b. *Tasa de aceptación*. Indica el porcentaje de artículos aceptados por una revista o por un congreso respecto al total de artículos presentados.
5. **Indicadores de redes de citación**.
    a. *Densidad de citación*. Calcula la cantidad de conexiones entre los documentos de una red de citas.
    b. *Coeficiente de agrupamiento*. Indica la proporción de conexiones entre los nodos vecinos de un nodo determinado perteneciente a una red de citas.
6. **Indicadores temporales**.
    a. *Índice de crecimiento*. Calcula el cambio, en términos de porcentaje, en el número de publicaciones o citas a lo largo del tiempo.
    b. *Edad media de las citas*. Proporciona el promedio de la antigüedad de las citas recibidas para un conjunto de documentos concreto.

En lo que respecta a la literatura científica existente ha sido abordada desde distintos ángulos: (i) geográfico, (ii) colaborativo, (iii) divulgativo y (iv) basado en palabras clave. La primera aproximación implica trazar un mapa que describa la participación de cada región del mundo en el avance de la tecnología de estampación en caliente en términos de volumen de producción científica. El segundo enfoque permite identificar las redes más productivas que se han establecido entre las instituciones y los agentes más influyentes en este campo. El tercero clasifica las revistas y eventos más influyentes en función de los índices de citación, lo que indica dónde publicar para obtener un mayor impacto. Y el último de ellos pretende inferir tendencias de investigación a partir de la evaluación de las palabras clave empleadas en la literatura científica publicada.

3.2.1 Planificando la estrategia.

Como parte de un proyecto de investigación de tesis doctoral, se ha discutido ya en el punto 2.1 Fuentes de información el uso de bases de datos tanto de literatura científica como de patentes. Para este análisis bibliométrico se eligió Scopus como base de datos de referencia debido a, por un lado, la fiabilidad de los datos mostrados y por otro la ayuda que presta para su análisis e interpretación.

Utilizando la información extraída de Scopus y considerando los indicadores de tipo bibliométrico se han podido analizar conceptos como el número de artículos por año, por autor y por fuente. Así mismo se estudiará la relevancia de la estampación en caliente a lo largo del tiempo, cuáles son los autores más relevantes, que revistas son más influyentes para el sector, cuáles son los países más activos, etc. Se ha elegido el periodo comprendido entre 2009 y 2019 que incluye las dos revisiones más importantes publicadas sobre esta tecnología [AROHS10] [HSBSSTP16].

Para conseguirlo, la estrategia de investigación comenzó con la selección de las palabras clave necesarias para realizar la búsqueda en la base de datos de Scopus. Se seleccionaron los siguientes términos en inglés, (i) *Die quenching*, (ii) *Hot stamping*, (iii) *Press hardening* y (iv) *Press quenching*. El resultado de la consulta realizada mostró un total de 2.316 documentos. Después de seleccionar solamente aquellos documentos publicados en inglés, la cifra se redujo a 1.837 documentos.

A continuación, se filtraron los datos para seleccionar solo aquellos documentos publicados en una revista, es decir, solo artículos y se obtuvieron 965 documentos y, excluyendo el año 2020, se obtuvo un resultado de 854 documentos. Finalmente, se seleccionaron las áreas de interés eligiendo ingeniería, ciencia de los materiales, física y astronomía, informática, química, ciencias ambientales, matemáticas, ingeniería química, energía, empresa, gestión y contabilidad, ciencias de la decisión y multidisciplinar de entre todas las posibles descartándose aquellos no relacionados directamente con la tecnología. Se obtuvo así un resultado final de 851 documentos susceptibles de ser analizados.

Scopus proporciona una ecuación que resume la búsqueda realizada, dicha ecuación se muestra a continuación y fue utilizada en julio de 2020.

```
( TITLE-ABS-KEY ( "Die quenching" ) OR TITLE-ABS-KEY ( "Hot stamping" ) OR
TITLE-ABS-KEY ( "Press hardening" ) OR TITLE-ABS-KEY ( "Press quenching" ) )
AND ( LIMIT-TO ( LANGUAGE , "English" ) ) AND ( LIMIT-TO ( SRCTYPE , "j"
) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) ) AND ( LIMIT-TO ( SUBJAREA ,
"MATE" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA ,
"PHYS" ) OR LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA ,
"CHEM" ) OR LIMIT-TO ( SUBJAREA , "CENG" ) OR LIMIT-TO ( SUBJAREA ,
"BUSI" ) OR LIMIT-TO ( SUBJAREA , "DECI" ) OR LIMIT-TO ( SUBJAREA ,
"ENVI" ) OR LIMIT-TO ( SUBJAREA , "ENER" ) OR LIMIT-TO ( SUBJAREA ,
"MATH" ) OR LIMIT-TO ( SUBJAREA , "MULT" ))
```

3.2.2 Mostrando los resultados.

Los indicadores bibliométricos incluidos en este apartado muestran información sobre los artículos científicos, los autores de estos y su afiliación, las revistas en las cuales han publicado sus trabajos, las citas recibidas, etc.

Con la información obtenida, se ha analizado la producción científica relativa al área de la estampación en caliente, estableciendo aspectos importantes como son los autores y las revistas más relevantes del sector. Se han tenido también en cuenta las citas en aras de mostrar los investigadores más significativos.

Para concluir, se ha realizado un análisis descriptivo utilizando VosViewer [VVVSL21] para mostrar, mediante redes bibliométricas, las relaciones entre

autores y entre los términos más relevantes de la tecnología de estampación en caliente.

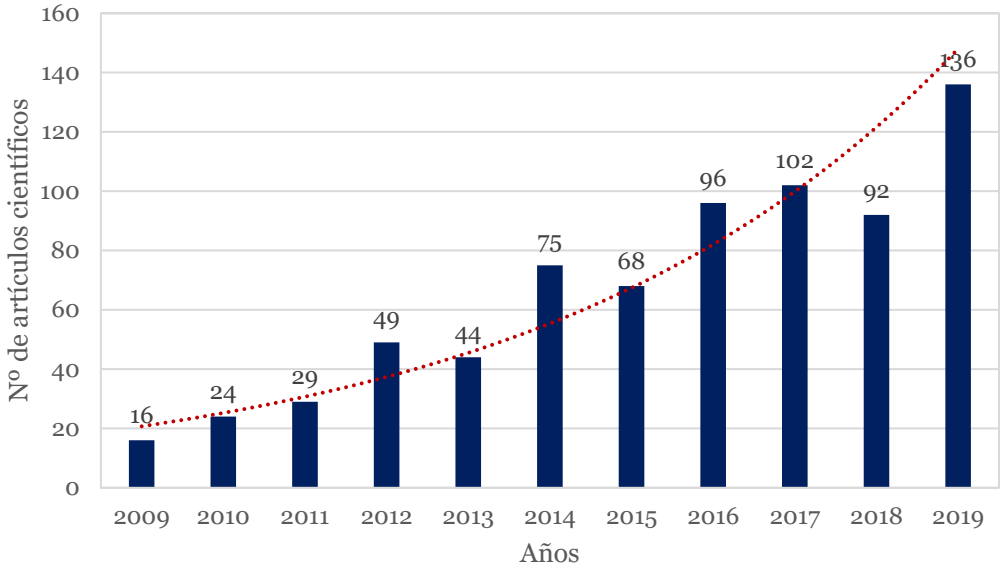

Figura 16: Número de publicaciones científicas en revistas indexadas por año para el periodo 2009-2019.

Como puede apreciarse en la **Figura 16**, la tecnología de estampación en caliente ha suscitado un gran interés durante la década comprendida entre 2009 y 2019. Este hecho se pone de manifiesto por el crecimiento exponencial de esta, mostrado por la línea de tendencia, pasando de 16 publicaciones en 2009 a 136 en 2019.

Completando esta información, la **Figura 17**, muestra las 10 revistas más productivas, es decir, las que más artículos han publicado, referentes a la tecnología de estampación en caliente. Analizando, de forma conjunta, estos resultados junto con los presentados en la **Figura *24***, número de citas por revista, se puede concluir que *Journal of Materials Processing Technology* es, sin duda alguna, la revista de referencia para la tecnología de estampación en caliente con 60 artículos en la última década que han sido citados un total de 3095 veces. También es destacable, en este contexto, la contribución de *CIRP Annals - Manufacturing technology* con sólo 17 publicaciones, pero 1359 citas. Este hecho, la convierte también en una revista pertinente en este campo.

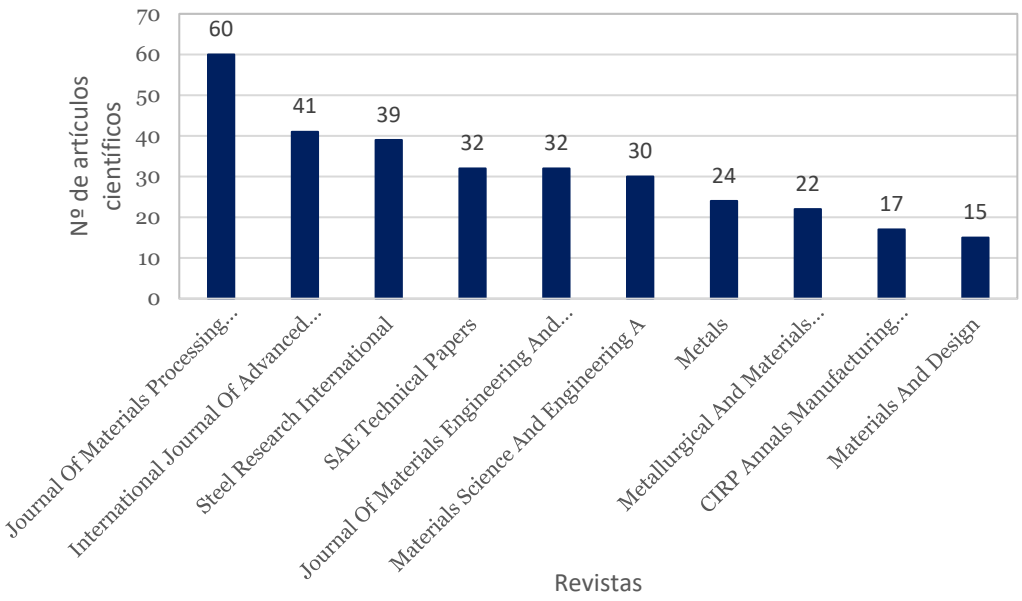

Figura 17: Número de artículos científicos publicados en las diez revistas indexadas más influyentes para el periodo 2009-2019.

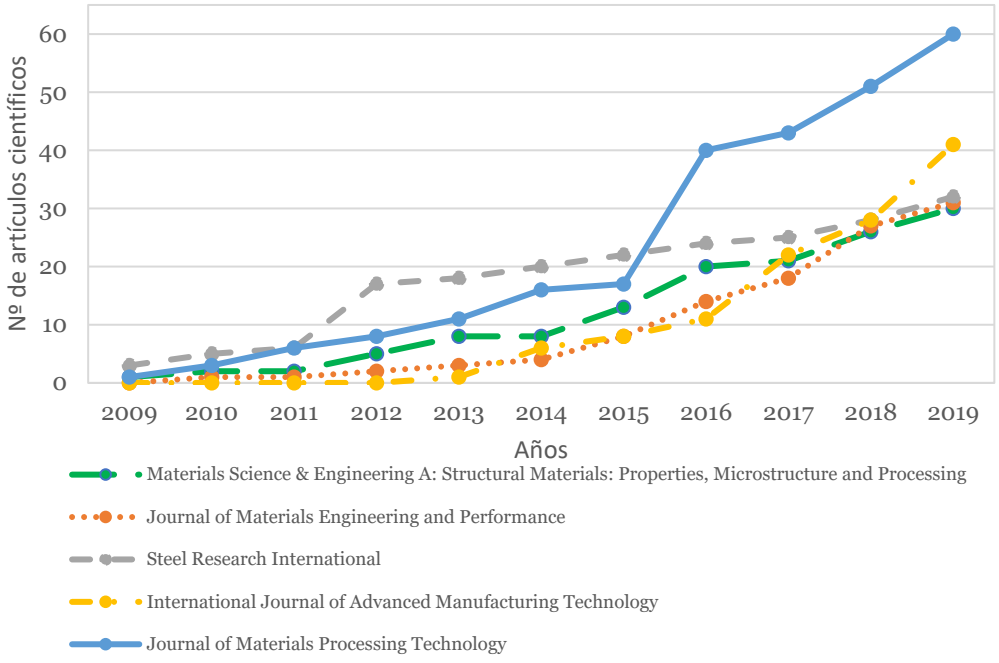

Figura 18: Evolución del número de artículos científicos publicados por año en las cinco revistas indexadas más influyentes para el periodo 2009-2019.

También la **Figura 18**, muestra el número de publicaciones anuales para las cinco revistas más relevantes del sector en el periodo comprendido entre 2009 y 2019. Destaca, entre ellas, la evolución de *International journal of advanced manufacturing technology*, una de las últimas en incorporarse en este sector.

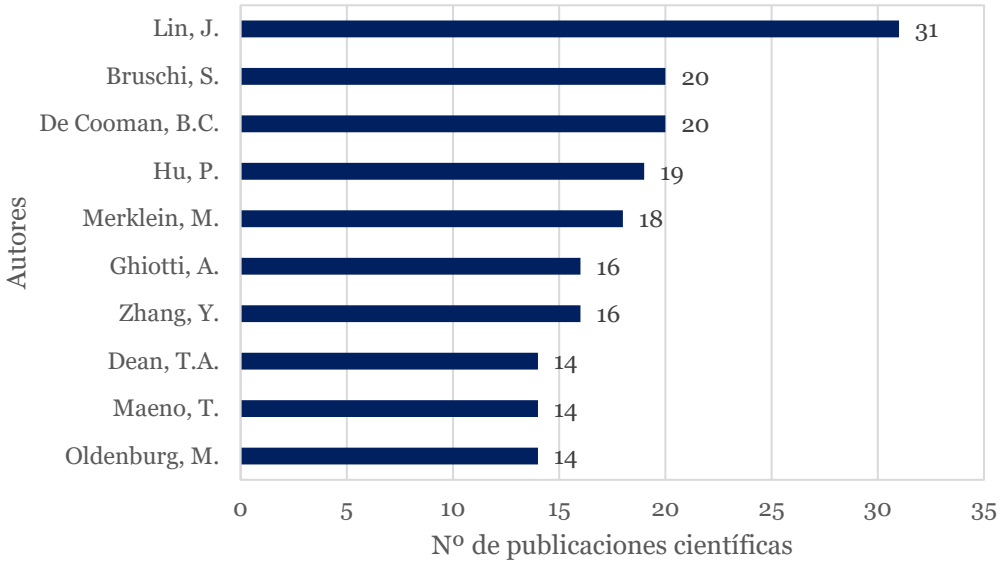

Figura 19: Número de artículos científicos publicados por los diez autores más relevantes en el periodo 2009-2019.

Centrándonos ahora en los autores, la **Figura 19** muestra los 10 autores con más publicaciones científicas, durante el periodo 2009 a 2019, relacionadas con el área de la estampación en caliente. El investigador más productivo es, sin duda, J. Lin con 31 publicaciones. Los investigadores más relevantes, como Merklein, M., Bruschi, S. y Ghiotti, A. son también los autores de las revisiones más relevantes en este campo, lo que confirma que, en la tecnología de estampación en caliente, los autores con mayor producción literaria son también los referentes científicos.

La **Figura 20** muestra las 10 instituciones con mayor número de trabajos publicados sobre la estampación en caliente. Las instituciones chinas se encuentran, por supuesto, entre las organizaciones más productivas.

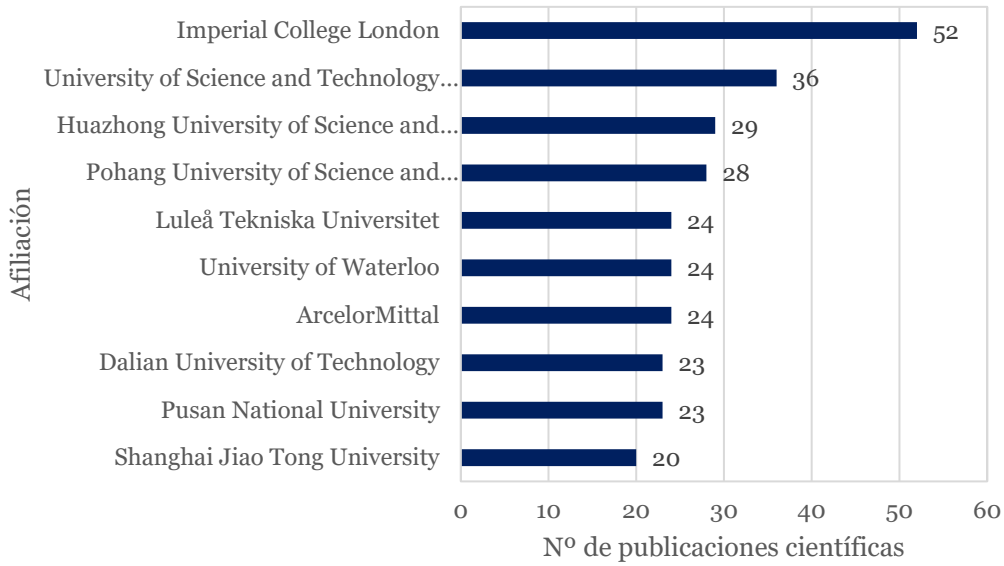

Figura 20: Número de artículos científicos publicados por las diez instituciones/empresas más relevantes en el periodo 2009-2019.

Las instituciones más relevantes en la tecnología de la estampación en caliente son en su mayoría universidades. Llama la atención que la Universidad de Lulea, la ciudad donde nació el proceso de estampación en caliente, haya mantenido sus vínculos con esta tecnología durante más de 40 años. Destaca también la presencia en esta lista de Arcelor Mittal, un agente industrial, entre las instituciones o entidades que más publican. Esto subraya el carácter aplicado que posee la investigación sobre la estampación en caliente.

Si ahora nos centramos en las citaciones, la **Figura 21** muestra la evolución del número de citaciones que han recibido los artículos sobre estampación en caliente por año. Este gráfico ratifica los resultados mostrados en la **Figura 15** en cuanto al interés mostrado por esta tecnología. No obstante, es importante tener en cuenta que este valor absoluto del número de citas por año, no refleja la relevancia real tal y como lo hacen otros indicadores como el índice h.

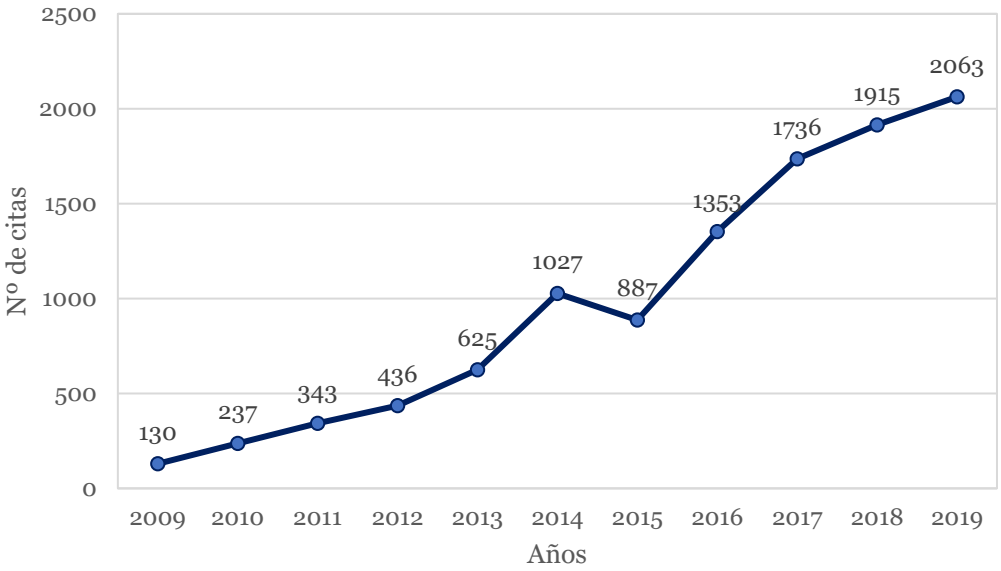

Figura 21: Número de citaciones por año que han recibido los artículos científicos sobre estampación en caliente durante el periodo 2009-2019.

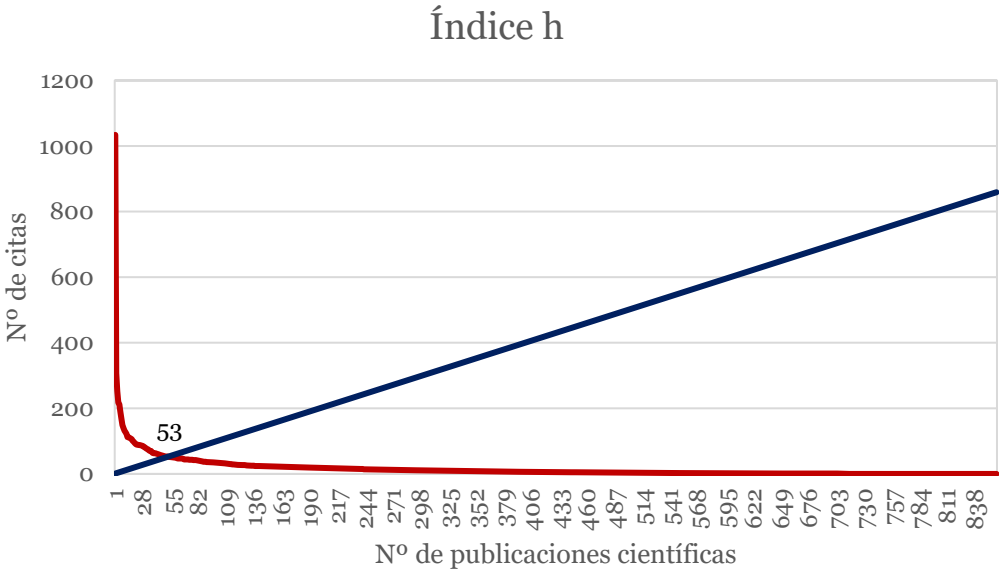

Figura 22: Valor del índice h para el conjunto de artículos científicos sobre estampación en caliente durante el periodo 2009-2019.

La **Figura 22** muestra el valor del índice h establecido en 53. Esto significa que, entre 2009 y 2019, 53 publicaciones han sido citadas al menos 53 veces. Si se

compara con una tecnología de referencia, como puede ser la estampación en frío, cuyo índice h es de 30, para el mismo periodo, es evidente que la estampación en caliente denota un gran interés por parte de la comunidad científica.

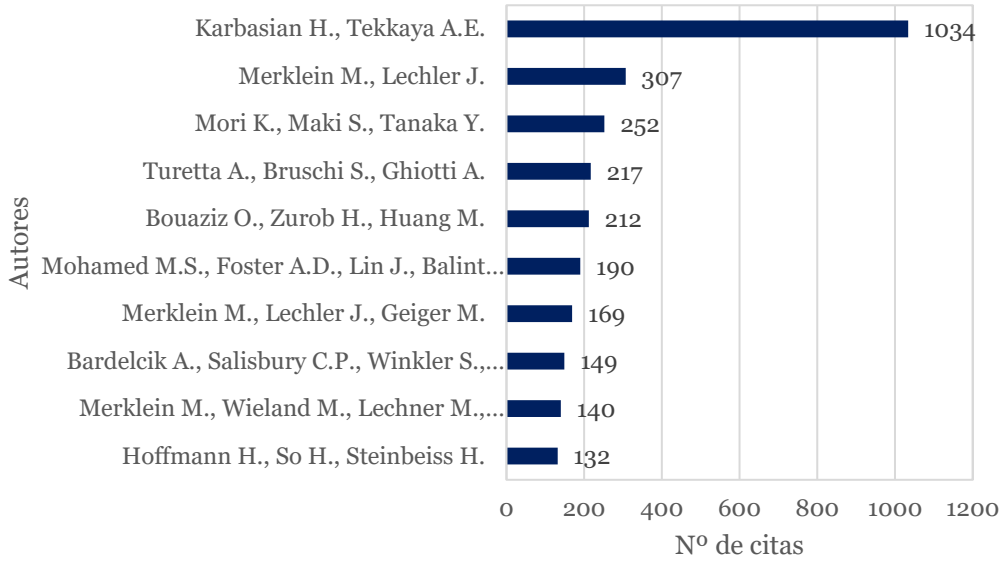

Figura 23: Número de citaciones por autores de artículos científicos sobre estampación en caliente durante el periodo 2009-2019.

Ahora, focalizándonos en el impacto, la **Figura 23** muestra los 10 autores o grupos de estos más influyentes o que han tenido un mayor impacto sobre esta tecnología. H. Karbasian y A.E. Tekkaya [AROHS10] son los autores más citados con su revisión en el año 2010, seguidos de M. Merklein y J. Lechler [HSBSSTP16] con una segunda revisión publicada en 2016. Otros autores relevantes en este campo, como A. Ghiotti y S. Bruschi, también figuran entre los autores de mayor influencia debido a su número de citaciones.

Por otra parte, la **Figura 24** muestra el número de citaciones obtenido por revista, siendo *Journal of Materials Processing Technology* y *CIRP Annals - Manufacturing Technology* las más influyentes en el aspecto referente a la difusión científica de esta tecnología. Esta observación concuerda con la lista de revistas más prolíficas, lo que significa que la productividad y la calidad, medidas como citas, están relacionadas en la literatura sobre estampación en caliente.

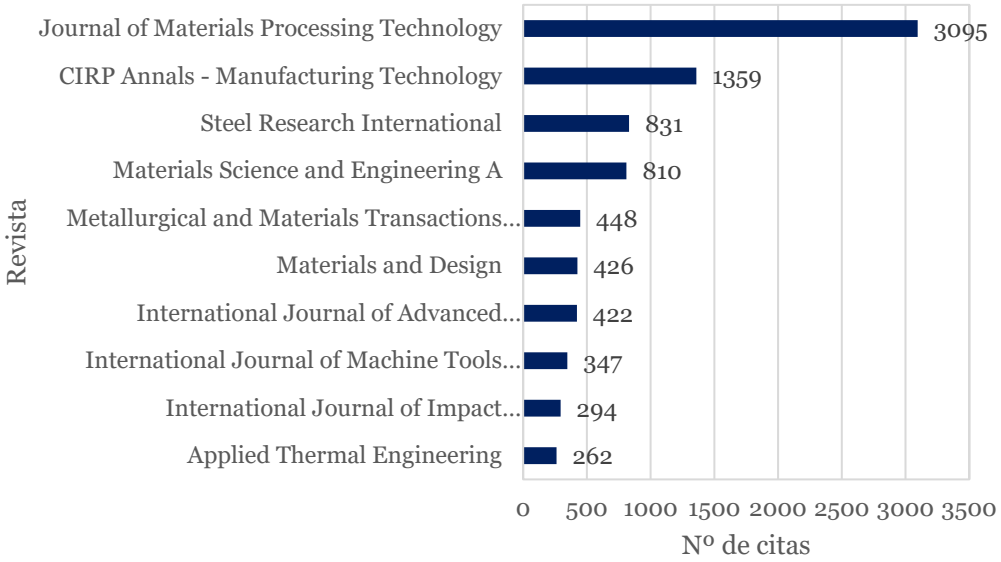

Figura 24: Número de citas por revista que ha publicado artículos científicos sobre estampación en caliente durante el periodo 2009-2019.

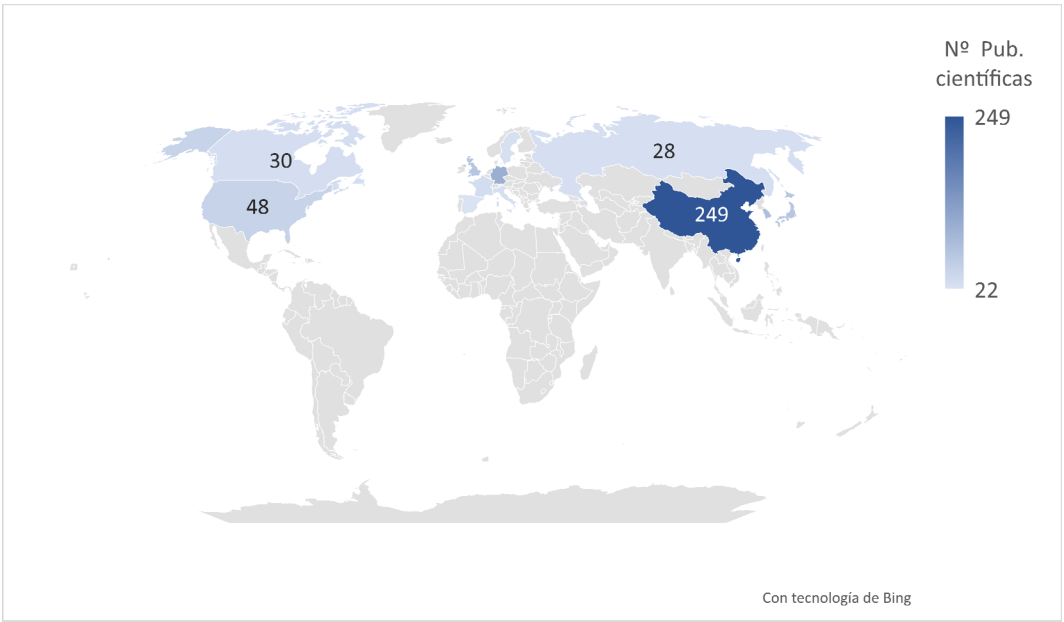

Figura 25: Número de publicaciones de artículos científicos por país sobre estampación en caliente durante el periodo 2009-2019.

Con respecto al número de publicaciones por país la **Figura 25** muestra que la mayoría de las publicaciones proceden de China, seguida de lejos por Alemania con

105 y Corea del Sur con 78. Estados Unidos, junto con Canadá, desarrollan también una importante actividad investigadora. Estos datos concuerdan con los de la **Figura 19**, dónde J. Lin es el autor con más publicaciones y, la **Figura 20**, en la cual las instituciones chinas se encuentran entre las más productivas.

La **Figura 26** muestra una visualización de red de las palabras clave utilizadas por los autores para establecer cuáles son las más relevantes. Hay un total de 207 artículos agrupados en 6 clústeres diferentes. Las palabras clave más relevantes se encuentran en el clúster verde con términos como estampación en caliente, estampación, máquinas de forja, etc. También es relevante el clúster rojo con términos como microestructura, propiedades mecánicas, acero de alta resistencia o temple.

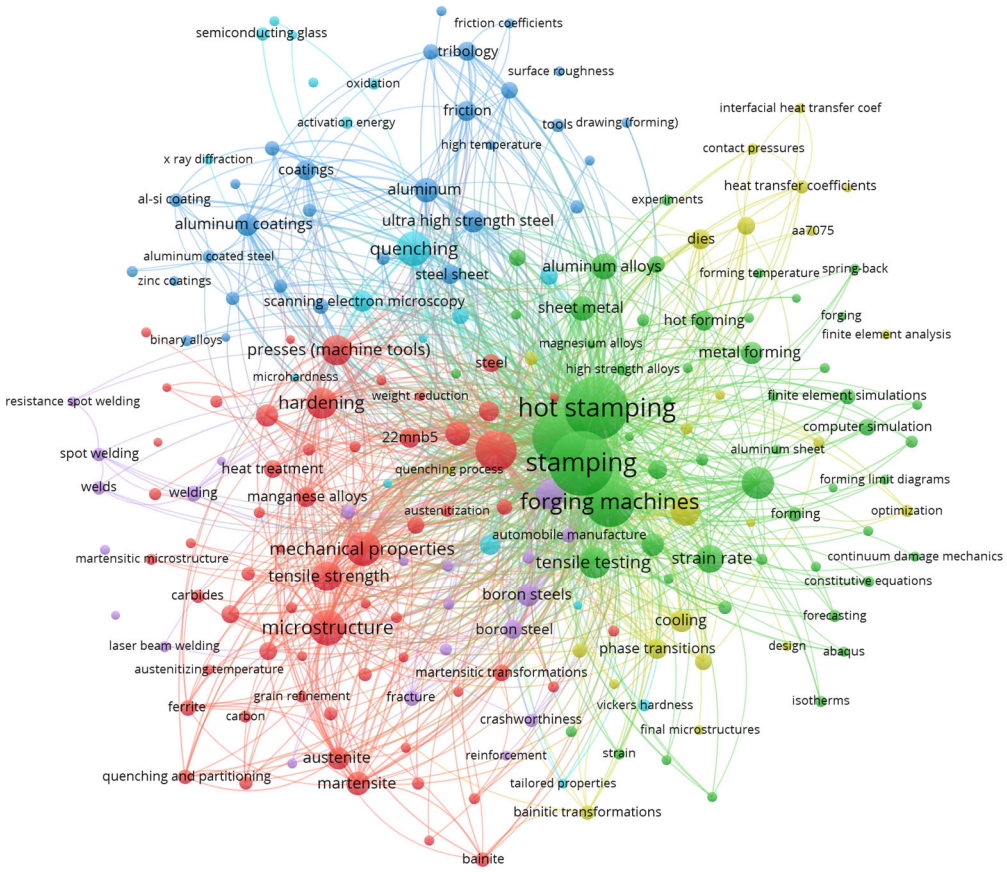

Figura 26: Red formada por las palabras clave sobre estampación en caliente utilizadas por los autores durante el periodo 2009-2019.

Si en lugar de las palabras, ahora nos centramos en los autores, la **Figura 27** muestra una red formada por los autores de artículos científicos relacionados con la tecnología de estampación en caliente que tienen al menos 2 publicaciones. Hay un total de 316 autores agrupados en 23 clústeres que definen los autores que colaboran más estrechamente. J. Lin, el investigador con más publicaciones en esta década, aparece en el clúster rosa. S. Brushi, el segundo investigador relevante aparece en el clúster amarillo.

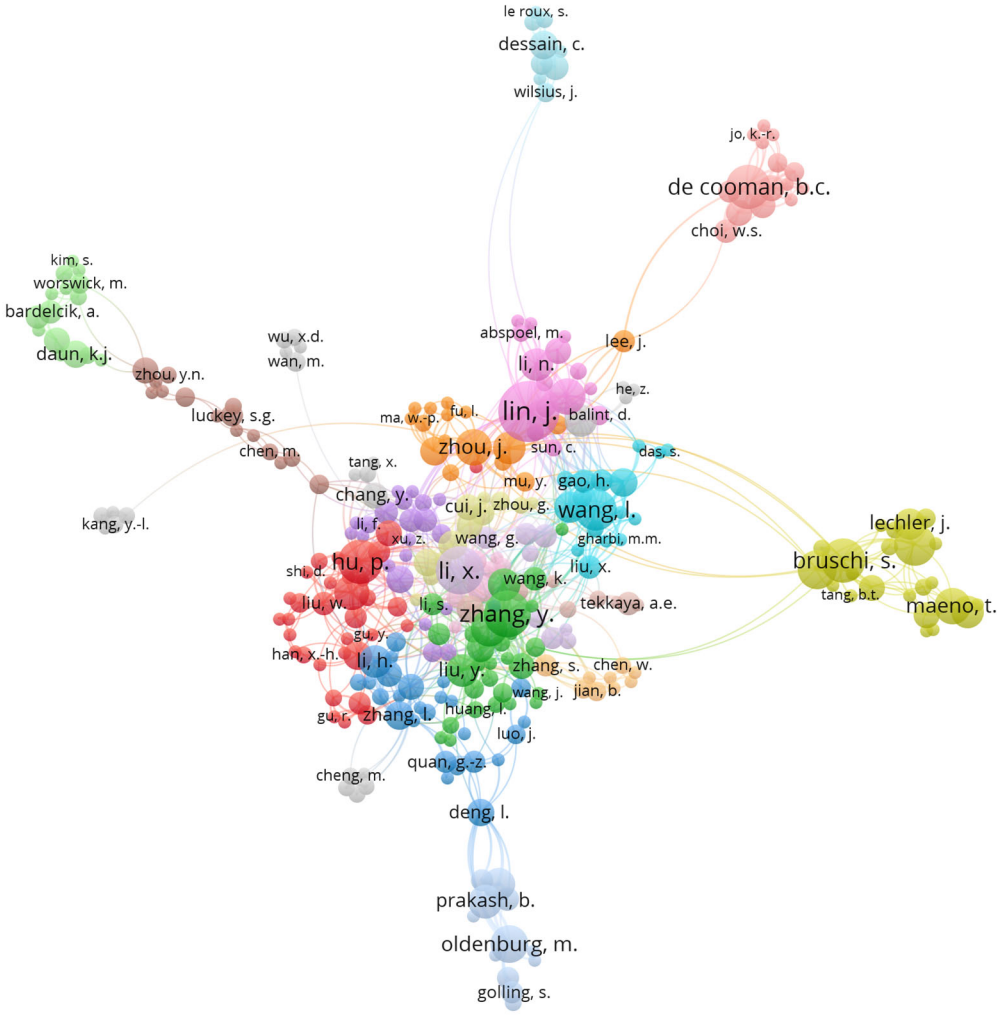

Figura 27: Red formada por los autores de artículos científicos en revistas indexadas sobre estampación en caliente durante el periodo 2009-2019.

3.2.3 Estableciendo conclusiones.

Teniendo en cuenta los análisis bibliográficos anteriores puede concluirse que:

- La estampación en caliente es una tecnología con una fuerte y constante evolución, con una producción científica creciente y un nivel de impacto medido en términos de índice h por encima de sus competidores tecnológicos, como puede ser la estampación en frío.
- Existe un conjunto establecido de fuentes científicas que concentra la mayor parte tanto de la producción como de las citas. Las revistas más prolíficas tienen también el mayor número de citas, pero en cuanto al número de citas por número de documentos totales publicados, Anales CIRP - Tecnología de fabricación destaca como fuente de alta relevancia.
- Los autores más relevantes, tanto por su producción como por su volumen de publicaciones, son también referentes en las principales revistas del sector.
- Cabe destacar que la investigación industrial en estampación en caliente aparece entre las diez primeras instituciones en términos de producción científica, lo que indica que la industria es un motor importante para la generación de nuevos conocimientos en este campo.
- Geográficamente, la producción científica se concentra en China, con otros dos focos en Europa (liderada por Alemania) y América del norte (liderada por EEUU).
- El gráfico de red de palabras clave muestra que existen seis grupos temáticos diferentes relativos a la investigación de la estampación en caliente, aquellos relacionados con los parámetros del proceso, las propiedades de los materiales, las tecnologías de ensamblaje, el propio proceso de temple, las alternativas a la estampación en caliente del acero y la refrigeración de las matrices.
- En cuanto a las redes de investigadores, se identifican varios grupos de estables, con una gran concentración en torno a los investigadores chinos y otros grupos a su alrededor en países como Suecia, Italia, Alemania, Reino Unido y Canadá.

Por tanto, se puede afirmar que la estampación en caliente es una tecnología prometedora y en constante desarrollo, con una actividad muy intensa impulsada por el gran interés industrial que suscita.

Con la aproximación realizada, aunque se puede obtener información valiosa como, por ejemplo, conocer los grupos de investigadores que trabajan juntos, la importancia relativa de estos, cuál es su área de investigación e incluso tratar de predecir futuras tendencias, no es suficiente para abordar el planteamiento desarrollado en esta tesis doctoral.

3.3 Evolución de la tecnología de estampación en caliente.

Como se ha establecido en el punto 3.2, el análisis bibliométrico es una de las metodologías utilizadas para el estudio y la evaluación de la actividad científica. Según Norton [ICIIS00], la bibliometría no es más que una medida tanto de los textos como de la información [DNTO09], así como un estudio cuantitativo de la literatura [BCIRCWA01]. Garfield [CIFS95] afirma que las técnicas de análisis bibliométrico son un conjunto de métodos matemáticos y estadísticos utilizados para analizar y medir, no sólo la cantidad sino también la calidad de las publicaciones científicas, cuyo resultado se utiliza para la toma de decisiones [BIQMSP10]. El análisis bibliométrico ayuda a explorar, organizar y analizar grandes cantidades de datos, permitiendo a los investigadores encontrar patrones no evidentes en la investigación y el desarrollo científico [CMPUBM05].

Para analizar la evolución de la tecnología de estampación en caliente, de nuevo, se ha utilizado un enfoque bibliométrico basado en la creación de mapas evolutivos [ADQVRF11a]. Este enfoque ayuda no sólo a analizar el campo de interés, sino también a detectar y visualizar los temas o áreas conceptuales, así como su desarrollo a lo largo del tiempo. Para llevar a cabo esta tarea, se utilizó el análisis de redes de palabras conjuntas (del término inglés *co-word*), una técnica para analizar las co-ocurrencias de palabras clave, así como para identificar las relaciones e interacciones entre los temas investigados y las tendencias de investigación emergentes, considerando sólo el título, el resumen (del término

inglés *abstract*) y las palabras clave proporcionadas por los autores [MTTOS94]. La técnica *co-word* es útil para [CTCOC18] dibujar una red de conocimiento teniendo en cuenta las palabras identificadas en el documento y su co-ocurrencia. Esta técnica, proporcionada por SciMAT [SciMAT12], no sólo sirve para dibujar un mapa sobre un área de conocimiento científico específico, sino que también ayuda a establecer relaciones entre los autores y los centros a los que pertenecen.

Existen diferentes herramientas con las cuales desarrollar un análisis de mapas científicos [MSIPPP15]. Se ha seleccionado SciMAT tanto por su disponibilidad como por la sencillez del procedimiento a realizar [SciMAT12]. Al aplicar este enfoque bibliométrico se pueden identificar diferentes etapas [ADQVRF11a]:

1. **Extracción global de palabras**. Búsqueda en la base de datos, descarga de resultados y extracción de todas las palabras de los *abstracts*, títulos y palabras clave.

2. **Tratamiento con SciMAT**. Cribado automático de palabras, agrupación preliminar y proceso de clasificación manual.

3. **Identificación de temas**. Aplicación de un algoritmo de agrupamiento, para cada periodo establecido, en el contexto de un análisis *co-word* [ASICW98].

4. **Visualización de temas y redes temáticas**. Los diferentes temas obtenidos se caracterizan por dos parámetros: la densidad y la centralidad que permiten clasificarlos en cuatro grupos diferentes [ASICW98]. La centralidad, c, mide el grado de interacción de un nodo con otros nodos; en este caso cada palabra constituye un nodo. La centralidad puede definirse como $c = 10 \times \sum e_{kh}$, donde k es una palabra perteneciente al tema y h es una palabra perteneciente a otros temas. La densidad, d, mide la fuerza interna de cada nodo, y puede definirse como $d = 100\ (\sum(e_{ii} / w))$, donde i y j son palabras clave pertenecientes al tema y w es el número de palabras clave del tema. Teniendo en cuenta el valor de estas dos variables, podemos encontrar cuatro tipos de temas según el cuadrante del diagrama estratégico en el que se encuentran, **Figura 28**, [ADQVRF11a].

- **Temas motores**. El cuadrante superior derecho muestra aquellos temas, caracterizados por una fuerte centralidad y alta densidad, que son relevantes para el desarrollo y la estructuración del campo y presentan relaciones externas con otros temas.
- **Temas muy desarrollados y aislados**. En el cuadrante superior izquierdo se encuentran los temas que no tienen vínculos externos con otros temas, pero sí internos, por lo que son marginales para el campo, muy especializados y periféricos.
- **Temas emergentes o en declive**. En el cuadrante inferior izquierdo se encuentran los temas con baja densidad y baja centralidad. Estos temas representan tanto los temas emergentes como los que están desapareciendo.
- **Temas básicos y transversales**. Los temas del cuadrante inferior derecho son significativos para el campo de investigación, pero no están bien desarrollados.

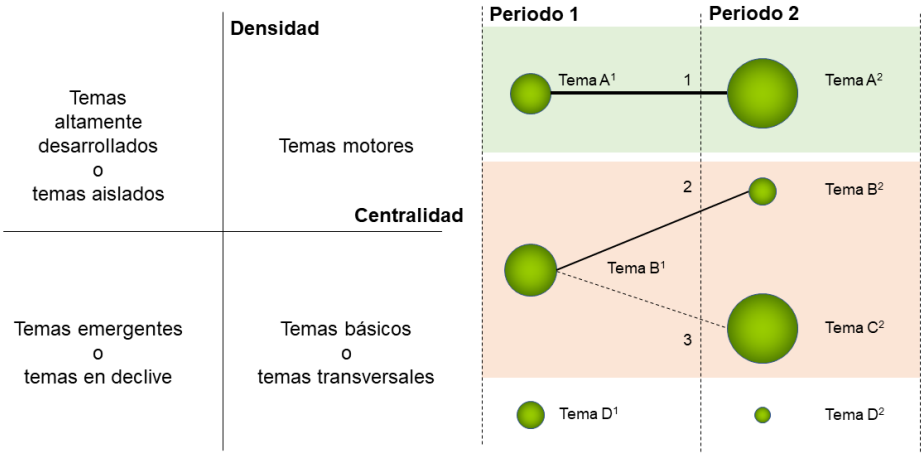

Figura 28: Esquema representativo de un diagrama estratégico y de un gráfico de evolución temática en el que se muestran diferentes tipos de conexiones entre temas.

5. **Identificación de áreas temáticas**. Cuando el conjunto de datos se separa en diferentes periodos, la progresión del campo de investigación puede estudiarse mediante un nexo conceptual (elementos en común,

palabras clave). La **Figura 28** incluye el diagrama de evolución temática que representa las áreas temáticas fundadas en el análisis y el nexo existente.

6. **Análisis del rendimiento**. Para estudiar la productividad de los temas y áreas temáticas descubiertos se pueden utilizar medidas cuantitativas como el número de documentos, autores, revistas, etc.

3.3.1 Diseñando mapas científicos.

Para llevar a cabo esta fase del estudio se seleccionó la base de datos WoS no sólo por la calidad de la información sino para mantener la misma línea de investigación que se ha utilizado en investigaciones anteriores [WIRRA20].

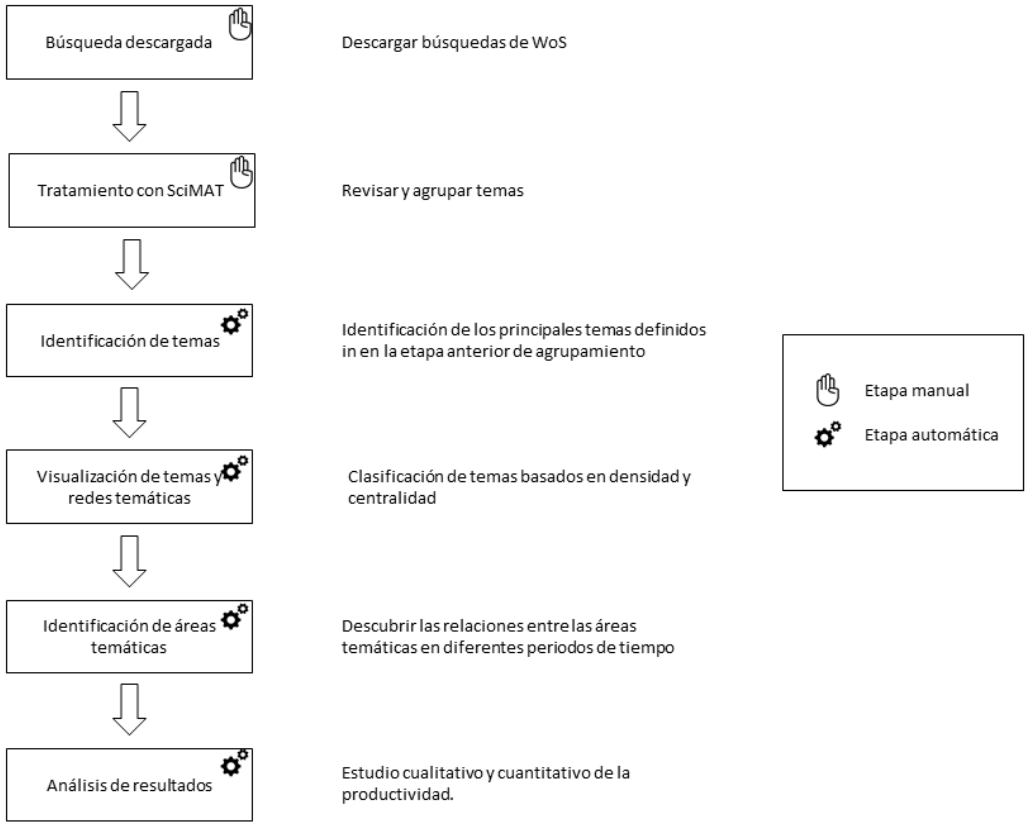

Figura 29: Metodología empleada en el diseño y creación de mapas científicos.

El proceso que se va a llevar a cabo en este análisis se resume en la **Figura 29**.

El conjunto de datos se ha obtenido utilizando la base de datos científica Web of Science (WoS), seleccionando concretamente *Web of Science Core collection*. Se seleccionaron cuatro temas, (i) *Die quenching*, (ii) *Hot stamping*, (iii) *Press hardening* y (iv) *Press quenching*, que son los empleados en la literatura para referirse a la misma tecnología, es decir, estampación en caliente tal y como como confirman diferentes estudios [CRAHSSS18] [AROHS10]. El resultado de esta consulta mostró 1.337 resultados. Se refinó la búsqueda seleccionando solamente aquellos trabajos escritos en inglés reduciéndose a 1.294 los resultados obtenidos. Se introdujo un criterio de fechas para seleccionar solamente los artículos publicados en revistas indexadas hasta el año 2019 incluido, quedando un total de 1.248 registros. Se seleccionaron, a continuación, áreas de investigación como ingeniería, ciencia de los materiales, metalurgia e ingeniería metalúrgica, física, mecánica, otros temas de ciencia y tecnología, termodinámica, sistemas de control de automatización, química, ciencia de los polímeros, transporte, matemáticas, ciencias de la imagen y tecnología fotográfica, ciencias ambientales y ecología e investigación operativa y ciencias de la gestión. Se excluyeron de esta lista aquellas áreas de investigación que utilizan los procesos de estampación en caliente en otros ámbitos. Esta reducción dio como resultado 1.233 artículos. El último filtro consistió en la selección de artículos dentro de la categoría tipos de documentos, con la única opción de artículos publicados en revistas JCR. Esto generó un resultado final de 689 publicaciones. La consulta final utilizada para recoger los datos se muestra a continuación.

> Hot stamping (Topic) or Hot stamping (Topic) or Press hardening (Topic) or Press quenching (Topic) and English (Languages) and 2019 or 2018 or 2017 or 2016 or 2015 or 2014 or 2013 or 2012 or 2011 or 2010 or 2009 or 2008 or 2007 or 2006 or 2005 or 2004 or 2003 or 2002 or 2001 or 2000 or 1999 or 1998 or 1997 or 1996 or 1995 or 1994 or 1993 or 1992 or 1991 or 1990 or 1989 or 1988 or 1987 or 1986 or 1985 or 1984 or 1983 or 1981 or 1980 or 1979 or 1978 or 1977 or 1976 or 1975 or 1974 or 1973 or 1972 or 1970 or 1969 or 1968 or 1967 or 1950 (Publication Years) and Articles (Document Types) and Engineering or Materials Science or Polymer Science or Metallurgy Metallurgical Engineering or Physics or Transportation or Mechanics or Mathematics or Science Technology Other Topics or Imaging Science Photographic Technology or Thermodynamics or Automation Control Systems or Environmental

Sciences Ecology or Chemistry or Operations Research Management Science (Research Areas)

Los resultados así obtenidos fueron revisados para ganar más integridad llevándose a cabo un análisis que redujo el conjunto de datos hasta 573 artículos. Esto significa, por ejemplo, que se han excluido artículos relacionados con la estampación litográfica que aparecían en la búsqueda original.

El siguiente paso consistió en el establecimiento de tres periodos de tiempo consecutivos con el objeto de cubrir el periodo global desde 1950 a 1919. Los criterios seleccionados incluyen dos revisiones importantes: *A review on hot stamping* [AROHS10], que marcó un punto de inflexión, y *Hot stamping of boron steel sheets with tailored properties: A review* [HSBSSTP16] que marcó otro hito. La estampación en caliente ha sido una tecnología que ha evolucionado rápidamente y aunque se han examinado otras formas de establecer periodos, como las patentes o la comercialización de vehículos, ésta se ha considerado la mejor. Los periodos seleccionados han sido 1950-2009 con 34 artículos, 2010-2015 con 211 y 2016-2019 con 329.

3.3.2 Analizando mapas científicos.

A continuación, se llevará a cabo un análisis de mapas científicos basado en la técnica de co-word. El objetivo de este análisis se realizó con el objetivo de descubrir la evolución de los temas clave en el área de la estampación en caliente, así como las relaciones entre estos.

Este apartado se divide en dos secciones diferentes. La primera de ellas se centra en el análisis del conjunto de artículos científicos publicados en revistas indexadas, así como todos los metadatos proporcionados por WoS (autores, instituciones y palabras clave entre otros) para obtener un corpus relacionado con la tecnología de estampación en caliente. A continuación, se ha realizado la agrupación de todos y cada uno de los artículos a fin de entender su relación con la tecnología de estampación en caliente comprobando las relaciones existentes con otras características o temas. Este análisis inicial proporciona una fotografía estática, mientras que el segundo enfoque, con un carácter más evolutivo, ayuda a

comprender y analizar la progresión de esta tecnología, basándose en las relaciones establecidas entre los diferentes temas.

3.3.2.1 Analizando el contenido de los artículos publicados.

Una vez importados los resultados de la búsqueda realizada en WoS a SciMAT, el primer paso ha sido establecer los períodos para realizar el análisis. Como ya se ha mencionado, se han creado tres periodos, basados en criterios relacionados con las revisiones descriptivas de la tecnología de estampación en caliente.

- El primer periodo comprende desde el año 1950 hasta el 2009 y se corresponde con el inicio de esta tecnología. Solo 34 artículos pertenecientes a este periodo han sido incluidos en WoS.
- El segundo periodo comienza en 2010 y finaliza en 2015. En 2010, Karbasian, H. [AROHS10], escribió una importante revisión sobre la tecnología de estampado en caliente. Este evento ha sido considerado como un hito para esta tecnología y representa el comienzo del segundo periodo. La consulta realizada en WoS para este periodo obtuvo 211 artículos, que corresponden al 36.8% del juego de datos completo.
- El periodo final corresponde a publicaciones realizadas entre 2016 y 2019. En 2016 una segunda revisión descriptiva fue publicada por Merklein, M. [HSBSSTP16], hecho que establece el final del segundo periodo y el comienzo del tercero. Se encontraron un total de 329 artículos en este periodo, es decir un 57.4% del total de la muestra.

Para realizar el análisis de tipo *co-word*, el paso inicial consiste en identificar el conjunto de palabras reconocidas como esenciales para la tecnología analizada, con el fin de agruparlas y determinar así los temas clave. A partir de una identificación inicial, realizada por SciMAT, en la que reconoció 1.966 grupos iniciales, basándonos en conocimientos expertos sobre la tecnología de estampación en caliente, se obtuvieron un total de 282 grupos.

El siguiente paso consistió en la parametrización del análisis, tal y como se indica a continuación. Para ello, se optó por realizar un análisis de co-ocurrencias, teniendo en cuenta una reducción de frecuencias en función de las publicaciones disponibles en ese periodo. Así, mientras que sólo se indicó una palabra como

referencia para el primer periodo, se obligó al sistema a proporcionar dos réplicas para el segundo y tercer periodos. Además, en base a la elección de la co-ocurrencia, se recurrió al índice de equivalencia ya que aparece en la literatura como el más adecuado para la normalización de frecuencias en este tipo de análisis [BA20YRS16]. El uso del índice h [ThIROM10] y de la suma de citas [BIOAL15] como medidas de calidad para el análisis se justifica en diversos artículos bibliométricos [ABIIS13]. Por último, para realizar el análisis longitudinal, se recurre a los índices de Jaccard e Inclusión. Considerando la popularidad del índice de Jaccard, se sigue la sugerencia de Leydesdorff [NVACCS08] en la que proponía la adición del número total de citas como medida de calidad de las publicaciones. Una ventaja del uso del índice de Jaccard está relacionada con el enfoque en la intersección entre las palabras de co-ocurrencia. A su vez, el índice de inclusión se utiliza para detectar el vínculo conceptual entre los temas de investigación de los diferentes períodos [BAITSR13]. Finalmente, la parametrización del subsiguiente análisis se muestra a continuación.

- **Períodos**: 1950-2009; 2010-2015; 2016-2019
- **Unidades de análisis**: Palabras: Palabras del autor; Palabras del origen; Palabras extraídas
- **Reducción de datos**: SciMAT permite filtrar los datos utilizando una frecuencia mínima como umbral. Por lo tanto, sólo se consideran las unidades de análisis con una frecuencia mayor o igual (en cada periodo) al umbral seleccionado. La reducción de frecuencia utilizada fue 1;2;2.
- **Tipo de matriz**: En la fase de construcción de la red, utilizando las palabras como unidad de análisis y la co-ocurrencia como relación, se construirá una red bibliométrica a partir de *co-words*. Se seleccionó la co-ocurrencia.
- **Reducción del ruido**: Cada arista del grafo tendrá un valor, que puede ser la co-ocurrencia o el acoplamiento entre los nodos correspondientes. SciMAT permite filtrar la red utilizando un valor de borde de umbral mínimo. 1; 2; 2 fueron los valores seleccionados.

- **Normalización**: SciMAT permite al usuario elegir las medidas de similitud comúnmente utilizadas en la literatura para normalizar las redes. El valor seleccionado fue el Índice de equivalencia [CWATDN91].
- **Algoritmo de agrupamiento o clustering**: Para obtener el mapa y sus clústeres o subredes asociadas se seleccionó el algoritmo de centros simples [ADQVRF11a] [ASICW98] utilizando los valores 8 y 2 para establecer el tamaño máximo y mínimo de la red.
- **Mapeado de documentos**: SciMAT permite seleccionar los documentos utilizados en cada clúster para desarrollar el análisis de rendimiento basado en medidas cuantitativas y cualitativas. Se seleccionó el mapeado principal y el secundario.
- **Medidas de calidad**: Una vez asociados los conjuntos de documentos a cada clúster, se puede añadir a cada conjunto una serie de medidas bibliométricas de rendimiento. Se ha elegido el índice h y la suma de citas.
- **Longitudinal:** Permite al usuario descubrir la evolución conceptual, social o intelectual del campo. También permite elegir diferentes medidas para calcular el peso del "nexo de evolución" entre los artículos de dos períodos consecutivos. Se han seleccionado el índice de Jaccard y el índice de inclusión.

Los resultados del análisis se muestran en la **Figura 30**, **Figura 31** y **Figura 32** proporcionando un gráfico diferente para cada periodo. Las figuras aportadas aclaran la relevancia de la estampación en caliente, por su inclusión como tema clave, en cuanto a número de publicaciones identificado por el tamaño de la bola verde. Apareció como tema básico de investigación en el primer periodo, y resultó ser un tema motor en el segundo y el tercero.

Para aclarar la inclusión de nuevos temas en la bibliografía, se ha proporcionado un esquema de colores, según las relaciones de los temas identificados en la bibliografía con 8 grupos diferentes: metalurgia del acero, propiedades de los materiales, tecnología de recubrimiento, modelización y simulación, diseño de

piezas, tecnología de fabricación, aleaciones de aluminio y tecnología de moldes. Se muestran, a continuación, los temas incluidos en cada familia tecnológica.

1. **Metalurgia del acero**.
    a. Aceros de fase compleja (del inglés *CP steels*).
    b. Material estampado.
    c. Microaleación.
    d. Aceros de doble fase (del inglés *DP steels*).
    e. Acero de alta resistencia (del inglés *HSS*).
    f. Aceros al silicio (del inglés *Si Steels*).
    g. Aceros aleados.
2. **Propiedades de los materiales**.
    a. Endurecimiento.
    b. Distorsión dimensional.
    c. Microestructuras.
    d. Flujo.
    e. Fluctuación.
    f. Propiedades físico-químicas.
    g. Conformabilidad.
    h. Austenita.
    i. Martensita.
    j. Ferrita-Perlita.
    k. Dureza.
3. **Tecnología de recubrimiento**.
    a. Eliminación del recubrimiento.
    b. Recubrimiento intermetálico.
    c. Recubrimiento AlSi.
    d. Difusión.
4. **Modelización y simulación**.
    a. Simulación asistida por ordenador.
    b. Modelos de materiales.
    c. Modelos constitutivos.
    d. Modelo.

   e. Fallo por fragilidad (rotura)
   f. Análisis de límites de conformado.
   g. Predicción.
   h. Fallo.
5. **Diseño de piezas**.
   a. Piezas a medida.
   b. Piezas/Productos.
   c. Diseño.
   d. Paredes finas.
6. **Tecnología de fabricación**.
   a. Estampación en caliente (del inglés *Hot stamping*).
   b. Soldadura láser.
   c. Optimización.
   d. Conformado.
   e. Estampación semicaliente (del inglés *Warm stamping*).
   f. Temperatura.
   g. Temple en prensa.
   h. Tratamientos térmicos en disolución.
   i. Enfriamiento.
7. **Aleaciones de aluminio**.
   a. Series 2XXX.
8. **Tecnología de moldes**.
   a. Aceros para herramientas de trabajo en caliente.
   b. Contacto.

Este nivel de agrupación superior basado en familias tecnológicas permite analizar la evolución de la tecnología de estampación en caliente. La **Figura *29*** muestra que el grupo de propiedades de los materiales, así como los grupos de metalurgia del acero, se establecen principalmente como temas emergentes y motores. Esto indica las fases iniciales del desarrollo de la estampación en caliente, incluida en el grupo de tecnología de fabricación, que aparece como un importante tema básico o transversal.

3.3.2.1.1 Primer periodo (1950-2009)

En este primer periodo, que comprende desde 1950 a 2009, encontramos doce temas de investigación, tal y como se muestra en la **Figura 30**.

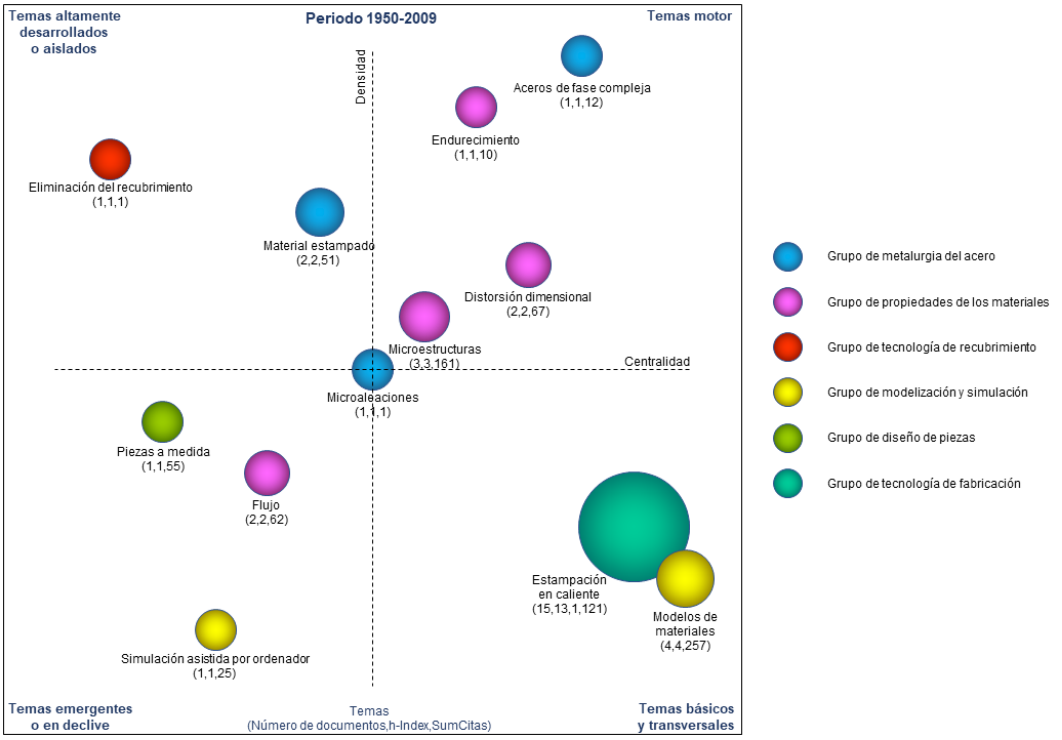

Figura 30: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 1950-2009.

A pesar de la menor producción científica, en el diagrama estratégico pueden identificarse varios grupos relacionados con las disciplinas científicas cuya evolución puede observarse en los dos períodos siguientes. Es destacable que la posición de los grupos en el diagrama coincide con la situación de la tecnología de estampación en caliente entre 1950 y 2009. A continuación, se enumeran los grupos, sus componentes y su relación con el estado de la técnica en el periodo analizado:

- El grupo de metalurgia del acero se sitúa en la mitad superior del diagrama, lo que refleja la exploración de alternativas a los aceros al boro en los primeros años.

- El grupo de propiedades materiales se distribuye en dos cuadrantes del diagrama, lo que demuestra que se trata de un campo de investigación muy activo, con materias y temas motores tanto emergentes como en declive.
- El grupo de tecnología de recubrimiento es un elemento aislado, lo que se entiende porque en este primer periodo los aceros recubiertos aún no estaban totalmente industrializados. Se observa que en los siguientes periodos el acero recubierto se convierte en un tema motor debido a la generalización del uso del material recubierto de AlSi patentado por ArcelorMittal.
- El grupo de modelización y simulación se encuentra en la zona de muy baja densidad, ya que las capacidades de simulación por ordenador no se utilizaban ampliamente. No se disponía de ningún software comercial que pudiera realizar una simulación termo-mecánica-microestructural.
- El grupo de diseño de piezas es un grupo emergente impulsado por la demanda del mercado. Su posición en este periodo está relacionada con el mercado de diseños de piezas monolíticas (no adaptadas), que aún no estaba firmemente establecido.
- El grupo de tecnología de fabricación. es un tema básico/transversal en la zona inferior derecha coincide con el hecho de que la tecnología aún no estaba fuertemente establecida como campo de investigación. Este grupo ascendió hasta convertirse en un tema motor en los siguientes periodos, como puede observarse en la Figuras 30 y 31.

3.3.2.1.2 Segundo periodo (2010-2015)

En el segundo periodo, que va desde 2010 a 2015, aparecen 22 temas diferentes, tal y como muestra la **Figura 31**.

Al centrarse en la evolución de las disciplinas científicas del primer al segundo periodo, el análisis bibliométrico coincide con las tendencias de implantación de la estampación en caliente en la industria del automóvil. Esto significa que la tecnología de estampación en caliente es un ejemplo de alineación entre el motor industrial y la producción científica asociada.

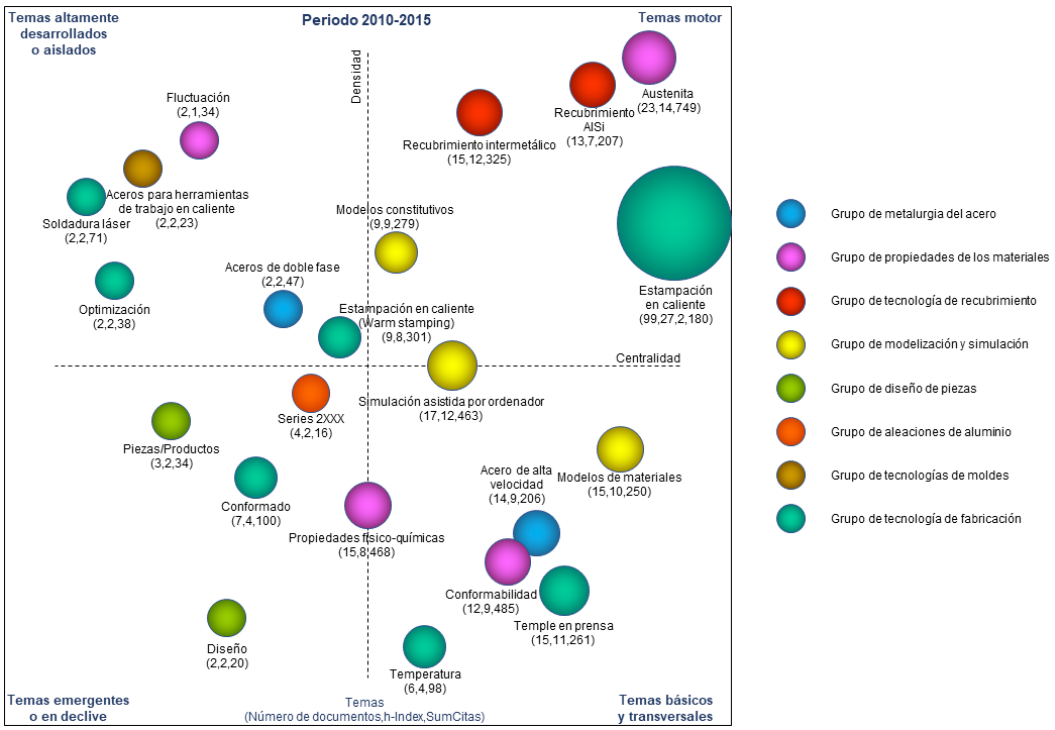

Figura 31: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 2010-2015.

Se explican, a continuación, de forma detallada todos los grupos.

- El grupo metalúrgico del acero aparece en la parte izquierda e inferior del diagrama. Esto se ajusta a la consolidación industrial del 22MnB5 como el principal grado de acero para la estampación en caliente. Así, se observa una reducción de la centralidad y la densidad.

- El grupo de propiedades de los materiales se sitúa en los cuatro cuadrantes del diagrama debido a su gran actividad.

- El grupo de tecnología de recubrimiento se ha desplazado desde la posición de temas aislados en este periodo debido a la amplia implantación de los aceros recubiertos de AlSi en la industria del automóvil. Este grupo es un tema motor de acuerdo con las demandas reales recibidas de la industria.

- El grupo de modelado y simulación muestra un cambio hacia una mayor centralidad y densidad. El periodo coincide con el desarrollo de paquetes

de software como Autoform® y Pamstamp®, para resolver escenarios de estampación en caliente con software de simulación multifísica.

- El grupo de diseño de piezas se mantiene en la zona emergente/declive. Este comportamiento se interpreta como una producción baja pero constante en las nuevas aplicaciones de la tecnología. La creciente cantidad de piezas estampadas en caliente en la carrocería en blanco (del término inglés *Body In White* (BIW)) apoya esta interpretación.

- Las aleaciones de aluminio son un nuevo grupo que coincide con la popularidad de los desarrollos de BIW de aluminio de Jaguar-Land Rover y los primeros estudios de la industria aeroespacial sobre la estampación en caliente de aluminio 2024. Su posición en el cuadrante inferior izquierdo implica que es un tema emergente que decae en el próximo periodo debido a la importancia relativa de los aceros en la tecnología.

- La tecnología de las matrices es un nuevo elemento debido a un creciente interés por la optimización de los tiempos de ciclo, la productividad de las matrices y la geometría de las piezas. En retrospectiva, el crecimiento de la producción científica sobre tecnología de troqueles está claramente vinculado al aumento de la productividad en las líneas de producción.

- El grupo de tecnologías de fabricación se extiende durante este periodo en los diferentes cuadrantes. Los términos “estampación en caliente” y “endurecimiento en prensa” aparecen en diferentes zonas. Este doble posicionamiento de palabras clave con el mismo significado está relacionado con la evolución terminológica del campo. Los términos “soldadura láser” y “optimización” deben interpretarse como temas muy desarrollados. Esta afirmación se ve corroborada por la explosión del número de líneas de producción de calidad para la automoción y de instalaciones de corte por láser para la estampación en caliente en todo el mundo durante este periodo. “Temperatura” y “Estampación en caliente” son sólo términos transitorios.

3.3.2.1.3 Tercer periodo (2016-2019)

El tercer y último periodo, desde 2016 a 2019, presenta otros 22 temas, tal y como muestra la **Figura 32**.

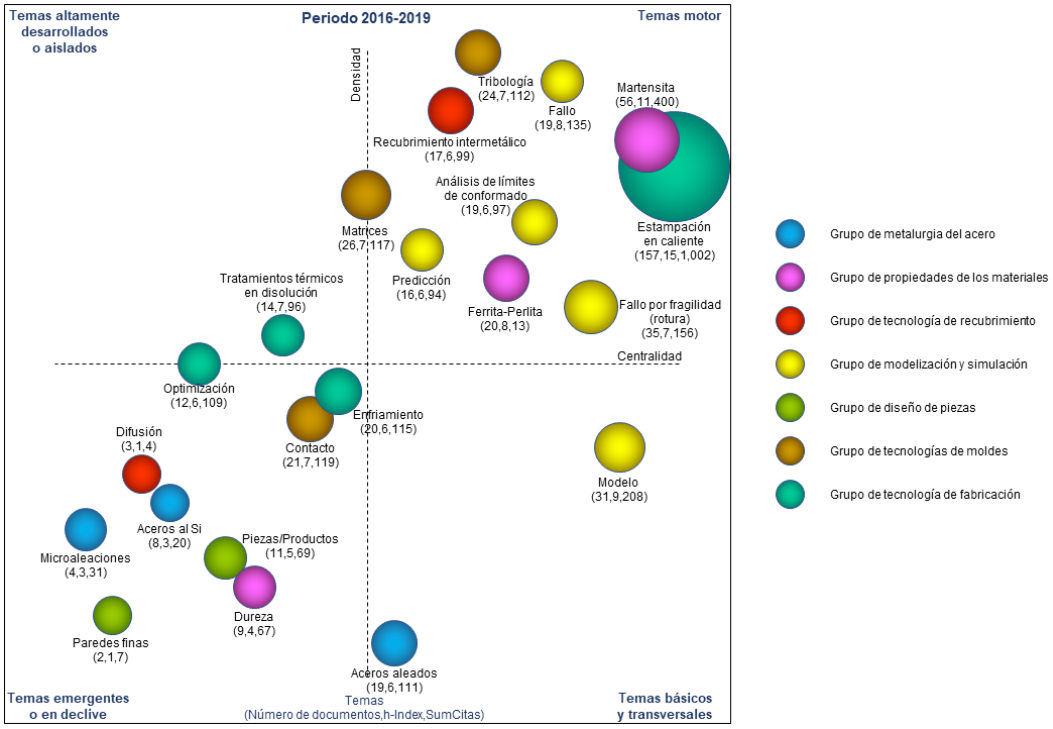

Figura 32: Diagrama estratégico para la tecnología de estampación en caliente en el periodo 2016-2019.

Se interpretan a continuación los grupos pertenecientes a este periodo.

- El grupo de metalurgia del acero se adentra en la posición de los temas en declive. Vuelven a aparecer temas como las microaleaciones, pero esto sólo refleja que se han hecho pocos esfuerzos de publicación en el desarrollo de los aceros para la estampación en caliente. Esto no implica, sin embargo, que la actividad haya sido escasa, ya que han aparecido en el mercado productos de ArcelorMittal (aceros para estampación en caliente de 2.000 MPa de resistencia) y de Kobe Steel (aceros de largo retardo en la formación de ferrita y perlita). Las pruebas apuntan a una tendencia que explica porque las innovaciones patentadas se mantuvieron en secreto hasta su completo desarrollo.

- El grupo de propiedades de los materiales gana relevancia en este periodo desplazando su peso al primer cuadrante y apuntando a que la ciencia de los materiales se posicione como tema motor.
- El grupo de tecnología de recubrimientos mantiene su papel de motor en este campo. Esto no solo se debe al uso extendido de los revestimientos de AlSi, sino a la investigación en la aplicación potencial de los revestimientos de Zn y ZnNi como posibles sustitutos que podrían superar al AlSi.
- El grupo de modelización y simulación mantiene su posición como tema motor, a pesar de que aparece una palabra clave "modelo" en el cuadrante de temas básicos. Esta observación coincide con el uso generalizado de software comercial para la estampación en caliente y el desarrollo de modelos de daños para la evaluación de fallos.
- El grupo de diseño de piezas se sitúa en el cuadrante inferior izquierdo, que se explica por la actividad permanente en el desarrollo de nuevos componentes y aplicaciones.
- La tecnología de matrices pasa a valores de centralidad y densidad más altos, lo que indica que podría ser un tema motor potencial para el futuro próximo, en consonancia con el creciente interés por la tecnología de fabricación aditiva y su potencial aplicación a la refrigeración en matrices de estampación en caliente.
- El grupo de tecnología de fabricación es un fuerte representante de los grupos motores "estampación en caliente" y está rodeado de grupos de densidad media. Esto subraya el reconocimiento de la "estampación en caliente" como una disciplina en sí misma con varias tecnologías satélite relativas a proceso.

3.3.2.2 Analizando el mapa de evolución conceptual.

Los mapas de evolución conceptual se emplean para mostrar las relaciones entre los distintos temas a lo largo del tiempo. El tamaño de los círculos es proporcional al número de artículos y su color representa un nivel de agrupación adicional. Este nivel superior de agrupación asigna el mismo color a los temas que pertenecen a

una disciplina científica común, basada en la experiencia de los autores en el campo de la estampación en caliente. También hay líneas sólidas y líneas discontinuas que conectan las esferas. Una línea continua representa un vínculo temático entre los distintos clústeres (nexo conceptual), mientras que una línea discontinua significa que sólo se comparten algunas palabras clave entre ellos (nexo no conceptual). El grosor de cada línea representa la intensidad de la relación, siendo proporcional al índice de inclusión [SciMAT12].

En concreto, la **Figura *32*** muestra un mapa de evolución conceptual de la tecnología de estampación en caliente basado en el conjunto de datos analizados. Si se observa el primer periodo (1950-2009), las familias tecnológicas predominantes son el grupo de propiedades de los materiales, el grupo de la metalurgia del acero y el grupo de la tecnología de fabricación. Existen fuertes conexiones entre el grupo de propiedades de los materiales y el grupo de tecnología de fabricación con los temas clave de estampación en caliente y estampación en caliente incluidos en el segundo periodo. Esto significa la conexión entre el inicio y la consolidación de la tecnología de estampación en caliente. También hay algunos vínculos menos intensos en el grupo de propiedades de los materiales, lo que implica la evolución y la búsqueda de nuevos materiales de producción.

Por otra parte, el segundo periodo (2010-2015) muestra, en ambas direcciones, fuertes relaciones en los temas clave incluidos en el grupo de tecnologías de fabricación, para periodos pasados y futuros. La estampación en caliente es un tema clave que evoluciona con el tiempo y adquiere gran relevancia. Además, los términos clave relacionados con el grupo de modelización y simulación están adquiriendo mayor importancia y estableciendo relaciones con conceptos similares a los de los últimos periodos. Estas relaciones están vinculadas al desarrollo de nuevos programas informáticos de simulación y a la capacidad de cálculo del nuevo hardware.

De esta manera, se muestra en la **Figura 33**, el gráfico correspondiente al mapa de evolución conceptual desarrollado para esta tecnología desde sus comienzos en 1950 hasta 2019, dividiéndolo en los tres periodos anteriormente descritos. Se muestra, igualmente, el conjunto de áreas temáticas que la definen.

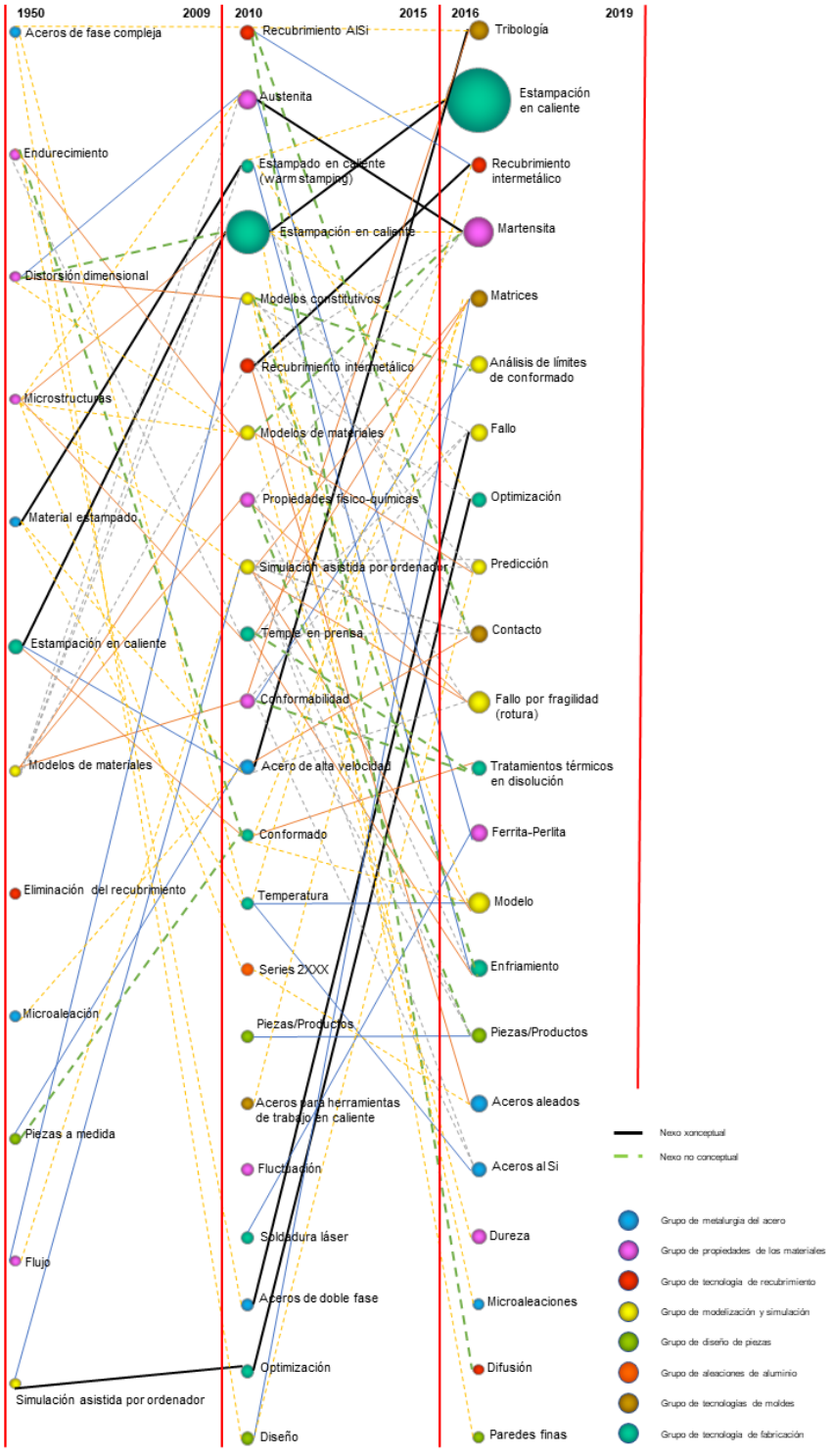

Figura 33: Mapa de evolución conceptual y áreas temáticas para que comprende tres periodos definidos, 1950-2009, 2010-2015 y 2015-2019.

Utilizando este tipo de representación gráfica para mostrar la evolución de los conceptos clave, se puede afirmar que:

- Aunque está presente en los tres periodos, la «estampación en caliente» adquiere sin duda relevancia como estado del arte, teniendo en cuenta el tamaño de las esferas.
- Existe una intensa relación en la «estampación en caliente» como tema clave a lo largo de los tres periodos, y observándose un mayor tamaño de esfera en las etapas más actuales. Esto marca la identificación de la «estampación en caliente» como una tecnología con buenas perspectivas de futuro.
- Las relaciones entre el «grupo de propiedades de los materiales» y el «grupo de tecnologías de fabricación» son un caso similar, que muestra la relevancia de los diferentes tipos de aceros y aleaciones en el proceso industrial.
- El «Grupo de modelización y simulación» también muestra la evolución en términos de capacidad de cálculo y se puede observar cómo este factor ha afectado al proceso de optimización.

3.3.3 Valorando los resultados.

El estudio bibliométrico desarrollado con SciMAT ofrece una visión de la evolución de la estampación en caliente desde mediados de la década de 1950 hasta la actualidad, no sólo en cuanto a los temas clave, sino también a sus relaciones y evolución a lo largo de los tres períodos definidos (1950-2009, 2010-2015 y 2016-2019). Además, muestra dos niveles diferentes de agrupación, uno de ellos proporcionado por el propio software y otro mediante la agrupación de conceptos relacionados.

El estudio actual ha identificado los temas motores que contribuirán al desarrollo futuro de la estampación en caliente, véase la **Figura 33**. Cabe destacar que la mayoría de ellos se encuentran en el grupo de «modelado y simulación» que, junto con los grupos de «propiedades de los materiales» y «tecnología de fabricación»,

establecen las líneas de futuro considerando el aumento de las capacidades informáticas.

Este análisis, no obstante, presenta varias limitaciones. En primer lugar, nos enfrentamos a una limitación temporal relativa al periodo definido en esta investigación. Podríamos afirmar que estos resultados podrían definirse como estáticos, debido a que la fotografía ha sido extraída del SciMAT bajo una configuración específica de la base de datos para un periodo de tiempo concreto. Sería recomendable realizar un análisis temporal en tiempo real [SADTC20] para obtener un mapeo dinámico que pudiera proporcionar un resumen gráfico evolutivo del estado del arte. Esta limitación inicial se convierte en una potencial línea de investigación futura.

Otra limitación se deriva del uso del SciMAT es que depende, en gran parte, de la calidad de los datos de entrada. Si bien es posible obtener gran cantidad de información de las bases de datos de literatura científica. SciMAT sólo utiliza el título, las palabras clave y el resumen para caracterizar la evolución de esta. Además, si los datos son inexactos, incompletos o mal etiquetados, los resultados pueden verse afectados.

Teniendo esto en cuenta, recurrir a técnicas de Procesamiento del Lenguaje Natural, como hacen Doloreux et al. [TIM19], permite la extracción automática de los temas reales mencionados a partir de todo el documento, en lugar de limitarse a utilizar unas pocas palabras clave definidas por los autores para posicionar el artículo en un campo específico de la literatura.

Por último, debemos mencionar el uso de artículos científicos para seguir la evolución de la tecnología de estampación en caliente, lo que se identifica como nuestra última limitación. El trabajo presentado aquí podría considerarse como el primer intento de contribuir a la predicción de la innovación en este campo específico. Sin embargo, deberíamos considerar otros posibles medios como conferencias, patentes, informes técnicos de empresas o proyectos subvencionados de I+D.

3.4 Una nueva metodología para predecir la innovación a través de las patentes.

Si bien los métodos bibliográficos analizados en los puntos 3.2 y 3.3 aportan valor en el contexto de la predicción de la innovación ya que podemos observar tendencias, grupos de trabajo, instituciones, analizar palabras clave, etc., cabe en este punto realizar un abordaje diferente. Para alcanzar el objetivo principal, se ha aplicado la siguiente metodología.

1. **Adquisición y preprocesamiento del conjunto de datos**. Se define la base de datos que se va a utilizar, tanto para los artículos como para las patentes utilizadas en esta investigación, las tareas de preprocesamiento de datos con las cuales se desarrollarán los cálculos necesarios, así como las diferentes variables utilizadas.
2. **Metodología de cálculo**. Se establecerán las bases de los cálculos estadísticos utilizados en esta investigación y se explicarán las diferentes variables, no sólo las proporcionadas por WoS y PatBase®, sino también otras variables resultantes de este procedimiento.
3. **Modelos de predicción**. A continuación, se describen los algoritmos de aprendizaje supervisado empleados, utilizando entre otros las redes bayesianas y los árboles de decisión.

3.4.1 Adquisición y preprocesamiento de datos.

Para comenzar esta investigación se ha considerado que la literatura científica, por un lado, es capaz de proporcionar un amplio abanico de información relativo a investigaciones previas, descubrimientos, avances y tendencias en diferentes áreas de conocimiento. Por otro lado, las patentes proporcionan información relativa al panorama competitivo, a las estrategias de innovación y a las tendencias emergentes de un campo específico. Cuando se combina esta información, literatura científica y patentes, dicho enfoque proporciona una perspectiva global del panorama científico y tecnológico. Este enfoque integrado constituye una base sólida que permite la toma de decisiones estratégicas en el ámbito de la I+D y la

identificación de oportunidades que favorecen tanto la colaboración como la comercialización.

En este trabajo se propone un enfoque basado tanto en métodos estadísticos como de aprendizaje automático [MLAI23]. Ambos son alimentados con datos extraídos de diferentes fuentes. Se utiliza, por un lado, literatura científica [HNLP18] y, por otro, información relacionada con patentes [TODDC19].

Se ha realizado el análisis de fuentes de datos teniendo en cuenta lo argumentado en la sección 2.1 Fuentes de información. y consideramos que no existen grandes diferencias entre las bases de datos WoS y Scopus [TOFCT06]. Además, dado que ambas son complementarias [WoSvsS10], varios autores [DSFPCA08] recomiendan realizar un análisis previo con ellas para identificar posibles diferencias en los resultados en función de la disciplina.

Finalmente, WoS ha sido la base de datos de datos para extraer artículos de investigación. Este hecho viene avalado por la calidad de los datos y la precisión que permite analizar la información científica desde mediados del siglo XX hasta la actualidad, tal y como se ha hecho en otra investigación anterior [EHSTBR22].

Además, se utilizó la base de datos de patentes en línea PatBase® [PATB23] para gestionar los datos de creación de patentes. También se consideró el uso de la base de datos *WoS Derwent Innovation Index* [PMLNA18] para garantizar cierto grado de homogeneidad entre ambas fuentes de información, pero esta base de datos dejó de utilizarse en 2010. De hecho, la base de datos PatBase®, como ya se ha indicado en el punto 2.1, es capaz de proporcionar acceso a documentos de patentes de más de 100 autoridades emisoras de todo el mundo y contiene más de 47 millones de familias de patentes agrupándolas en familias [TTHRPA20]. Esto hace que sea adecuada para los propósitos de esta investigación.

Partiendo de WoS, se ha creado una consulta utilizando las siguientes palabras clave (i) *Hot stamping*, (ii) *Die quenching*, (iii) *Press hardening*, (iv) *Press quenching*, (v) *Die hardening*, (vi) *Hot forming*, (vii) *Hot pressing* y (viii) *Hot press forming*. La consulta se realizó manualmente utilizando su sitio web y tecleando este código. Posteriormente, se refinó la consulta configurando los siguientes filtros en base al conocimiento experto sobre estampación en caliente.

- Tipo de documento: sólo artículos.
- Se incluyeron los años de publicación de 1950 a 2021.
- Sólo documentos en lengua inglesa.
- Áreas de investigación seleccionadas. Ingeniería, Ciencia de los materiales, Metalurgia, Ingeniería metalúrgica, Física, Mecánica, Ciencia y tecnología de otros temas, Química, Ciencia de los polímeros, Transporte, Matemáticas, Ciencia de la imagen, Tecnología fotográfica, Ciencias medioambientales, Ecología e Investigación de operaciones, Investigación Operativa y Ciencia de la Gestión.

Por último, la consulta arrojó un resultado de 12479 artículos. El resultado de esta consulta es un conjunto de diferentes campos como título, autores, año de publicación, institución del autor, etc., es decir un conjunto de metadatos susceptibles de ser utilizados en el modelo de análisis.

Antes de iniciar la tarea de análisis de datos, y después de una revisión de estos, se tomó la decisión de utilizar una serie de filtros adicionales que mejorarían sustancialmente la calidad de los mismos. En concreto, y por razones obvias, se tomó la decisión de seleccionar:

- Todos los artículos que tuvieran al menos un autor.
- Todos los artículos que tuvieran un año de publicación.

Esto se debe a que cuando se extraen los datos de WoS aparecen publicaciones sin autor, no hablamos de publicaciones dónde el autor tiene el valor “anónimo” sino de aquellos en los cuales el valor del campo autor aparece vacío. Lo mismo sucede con el año de publicación ya que se han extraído registros en los cuales la fecha de publicación del artículo no aparece registrada.

Se eliminaron, por lo tanto, todos los artículos que no cumplían estos dos requisitos. A continuación, se procesó el conjunto de artículos restantes, asignando a cada uno un valor de ID secuencial y, posteriormente, se dividieron los autores para identificar, finalmente, la posición de cada uno en cada artículo. Por último, se asignó un ID único a cada autor para identificarlo inequívocamente.

Se utilizó un método similar para recuperar los datos relacionados con las patentes de la base de datos PatBase® [CCFD14]. En esta ocasión, se utilizaron también las mismas palabras clave que en el caso de la búsqueda de artículos científicos. Se aplicó, a continuación, el siguiente proceso de filtrado:

- Se excluyeron las patentes publicadas después del 31 de diciembre de 2021.
- Se excluyeron los grupos principales A, D y H de la Clasificación Internacional de Patentes [WIPO99], es decir, necesidades humanas, textiles y papel y electricidad. Ello se debe a que, por ejemplo, el estampado en caliente en una camiseta no es el objeto de análisis de esta investigación.

Una vez más, se procesaron los datos extraídos, estableciéndose los siguientes requisitos:

- Todas las patentes debían tener al menos un inventor.
- Todas las patentes debían tener un año de publicación.
- La estructura del nombre de los inventores tenía que ser la misma que la de los autores, por lo que fue necesario transformar estos datos a "Apellido, Inicial del nombre", por ejemplo, "Doe, J.".

La justificación, en este caso, es similar a la de los datos extraídos de WoS. Cuando extrajimos el conjunto de datos de PatBase®, pudo también comprobarse que existían registros de patentes sin inventor y sin fecha de publicación. Esta afirmación no se refiere a que, por ejemplo, el inventor fuera una empresa, sino que ese campo para algunas patentes estaba vacío al igual que sucedía con la fecha de publicación de la patente.

Se excluyeron todas aquellas patentes que no cumplían estos requisitos. Se asignó un ID secuencial de patente a cada una de las patentes y se procesó la lista de inventores asignando un ID secuencial de inventor que indica el orden del inventor en la patente. También se asignó un ID único a cada inventor para identificarlo.

Los datos extraídos de las bases de datos WoS y PatBase® se procesaron para obteniéndose un nuevo conjunto de datos que contiene las siguientes variables divididas en tres diferentes grupos. El primer grupo, las variables ANA1 a ANA8, representan la inversión acumulada realizada por los autores de un artículo en la

producción de nuevos conocimientos. El segundo grupo, las variables PENA1 a PENA8, corresponden a la tendencia de los inventores a generar propiedad intelectual (PI). Por último, las variables PASA1 a PASA8 indican la generación real de PI. Los números, del 1 al 8, representan el número de autores o inventores implicados en el cálculo de cada variable. Estas variables son esenciales en el proceso de generación del modelo y nos permiten establecer las reglas de comportamiento que implican la generación de nuevas patentes. El proceso de cálculo se explica en la siguiente sección.

3.4.2 Método de cálculo.

Una vez extraído el conjunto de datos, este nuevo enfoque considera la publicación de un artículo como un evento estadístico. Para calcular la inversión acumulada en generación de nuevo conocimiento mediante la investigación, el proceso calcula para la variable $N_a = 1$ (donde N_a significa número de autores y 1 significa un solo autor) la suma desde $y = 0$ hasta z (donde y y z son el primer y el último registro en nuestro conjunto de datos de artículos), es decir, la suma de todas las veces que este autor aparece en todos los artículos anteriores al año de publicación del artículo, evento, que se está analizando. De forma similar, el resto de las variables, desde $N_a = 2$ hasta $N_a = n$ (en este caso $2, 3, \ldots, n$ significa el número de autores implicados en el cálculo de la variable), se calculan analizando no la aparición de un único autor sino las combinaciones no repetitivas [MDOC18] de los autores de 2 a n, siempre que el número de autores del artículo lo permita, y teniendo en cuenta el margen temporal indicado anteriormente.

Esta metodología considera la publicación de una patente como una propiedad asociada a un evento estadístico. Las variables $N_{p_i} = 1, N_{p_i} = 2, \cdots, N_{p_i} = p$ (donde N_{p_i} significa número de patentes de entrada, y los números $1, 2, 3, \cdots, p$ significan el número de inventores implicados en el cálculo de la variable), de $y = 1$ a z (donde y y z son el primer y el último registro de nuestro conjunto de datos de patentes), se calculan de forma similar al proceso descrito anteriormente. Se busca un autor o una combinación de ellos no repetida en el conjunto de datos de inventores considerando un periodo de tiempo que se calcula restando 2 al año de publicación del artículo y analizando las patentes sólo hasta el año resultante. La

razón por la cual restamos 2 años al año de publicación del artículo es el retraso de 18 meses entre la fecha de prioridad de las patentes y su fecha de publicación real. Las conclusiones indican la inclinación de los autores a generar propiedad intelectual.

Para calcular la generación real de propiedad intelectual, este procedimiento repite el mismo proceso anterior para obtener $N_{p_o} = 1, N_{p_o} = 2, \cdots, N_{p_o} = p$ (donde N_{p_o} significa número de patentes de salida, y el número $1, 2, 3, \cdots, p$ significa el número de inventores que intervienen en el cálculo de la variable), desde $y = 0$ hasta z (donde y y z son el primer y el último registro de nuestro conjunto de datos de patentes), cambiando el periodo de tiempo y analizando las patentes a partir del año de publicación del artículo al cual restaremos 2, es decir, consideraremos las patentes a partir de este año.

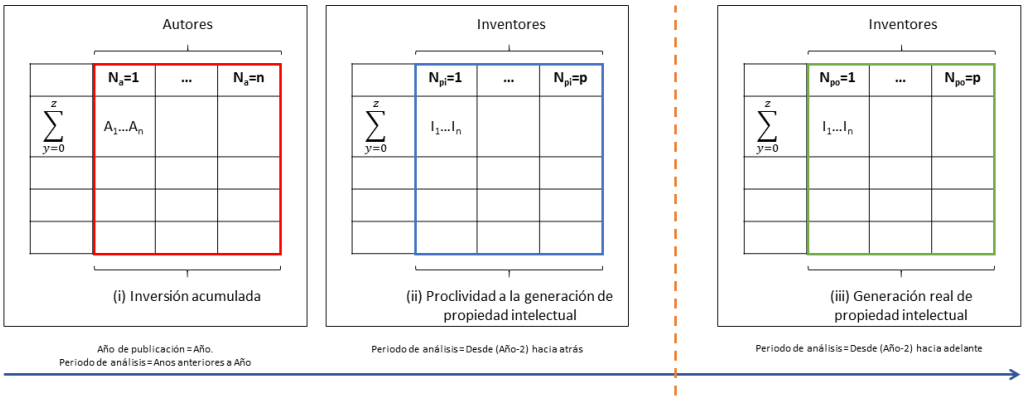

Figura 34: Metodología utilizada para procesar el conjunto de datos de artículos científicos y patentes con el objeto de obtener un nuevo conjunto de datos mediante el cálculo de (i) la inversión acumulada, (ii) la inclinación hacia la generación de propiedad intelectual y (iii) la generación real de esta.

Toda la metodología queda reflejada en la **Figura 34**. Concretamente, (i) ilustra cómo se realiza el proceso de cálculo de la inversión acumulada, variables ANA1 a ANA8, (ii) explica cómo se determina la proclividad hacia la generación de propiedad intelectual teniendo en cuenta los 18 meses de prioridad de la patente, variables PENA1 a PENA8 y, por último, pero no menos importante, (iii) evalúa la

generación real de propiedad intelectual teniendo en cuenta también este concepto.

Tabla 4: Propiedades estadísticas del conjunto de datos resultante del tratamiento de los juegos de datos primarios de artículos científicos y patentes.

Var.	Min	Max	Mediana	Nº ceros	Nº intervalos discretización	Discretización/Frecuencia segmentación				
ANA1	0	320	7	2508	5	1994	3988	5983	7977	9971
ANA2	0	347	0	6499	4	1495	2990	4485	5980	-
ANA3	0	660	0	8991	3	1163	2325	3488	-	-
ANA4	0	1320	0	10830	2	825	1649	-	-	-
ANA5	0	1848	0	11718	2	381	761	-	-	-
ANA6	0	1848	0	12129	2	175	350	-	-	-
ANA7	0	1320	0	12347	2	66	132	-	-	-
ANA8	0	660	0	12425	2	27	54	-	-	-
PENA1	0	272	0	8639	3	1280	2560	3840	-	-
PENA2	0	86	0	12153	3	109	217	326	-	-
PENA3	0	9	0	12455	2	12	24	-	-	-
PENA4	0	0	0	12479	**0**	-	-	-	-	-
PENA5	0	0	0	12479	**0**	-	-	-	-	-
PENA6	0	0	0	12479	**0**	-	-	-	-	-
PENA7	0	0	0	12479	**0**	-	-	-	-	-
PENA8	0	0	0	12479	**0**	-	-	-	-	-
PASA1	0	374	0	8084	2	2198	4395	-	-	-
PASA2	0	136	0	11977	2	-	-	-	-	-
PASA3	0	34	0	12451	2	-	-	-	-	-
PASA4	0	11	0	12472	2	-	-	-	-	-
PASA5	0	1	0	12476	2	-	-	-	-	-
PASA6	0	0	0	12479	**0**	-	-	-	-	-
PASA7	0	0	0	12479	**0**	-	-	-	-	-
PASA8	0	0	0	12479	**0**	-	-	-	-	-

La **Tabla 4** contiene información estadística relativa al conjunto de datos obtenido mediante el procesamiento de la información extraída de las bases de datos WoS y PatBase®.

Todos los valores son inicializados a 0. Significa que siempre hay un autor o inventor que no tiene publicaciones anteriores al año en que se produjo el hecho estadístico ni patentes anteriores o posteriores.

Puede observarse también que el valor máximo depende de cada variable. Los valores van de 0 (valor mínimo) a 1848 (valor máximo) en el caso de PASA1. Generalizando para todas las variables, esto significa que el máximo recoge el número de artículos anteriores al año en que se produjo el hecho estadístico o patentes anteriores o posteriores a ese año.

Así mismo, la mediana mide la tendencia central, que es la ubicación del centro de un grupo de números en una distribución estadística, es prácticamente 0 para todas las variables excepto para ANA1, cuyo valor es 7. Si se presta atención al número de ceros, se pone de manifiesto que es diferente para cada variable. El valor más bajo es 2508 y corresponde a ANA1 y justifica el valor 7 para la mediana de ANA1. En el extremo opuesto, hay variables para las que todos los valores son 0. Estas variables no son representativas y se rechazarán porque no aportan ninguna información útil.

Antes de procesar este conjunto de datos se realiza un procedimiento que consiste en establecer el número de intervalos de discretización y configurar los distintos intervalos para generar modelos. Para ello, se utilizó el conjunto de datos obtenido al procesar los datos de artículos científicos y patentes. Sé calculó (i) la inversión acumulada (variables ANA1 a ANA8), (ii) la proclividad a la generación de propiedad intelectual (variables PENA1 a PENA8) y (iii) la generación real de propiedad intelectual (variables PASA1 a PASA8). Se analizaron los valores obtenidos de las diferentes variables y se aplicaron diferentes segmentaciones a estas en función del número de ceros obtenidos para cada una. Se descartaron en algunas de las segmentaciones variables como el intervalo PENA4 a PENA8 ya que no arrojan valores como se pone de manifiesto en la **Tabla 4**. Se probó también la influencia de otras variables, que si bien no eran todo ceros si presentaban una gran cantidad de estos, preparando diferentes escenarios en los cuales en unos eran tenidas en cuenta y en otras no. Se prepararon un total de 35 discretizaciones diferentes de los datos que fueron posteriormente testeados en busca de la mejor discretización posible. Se detallarán posteriormente tanto el conjunto de test como los resultados obtenidos. Como ya es sabido, se necesita algo de ruido cuando intentamos crear modelos de aprendizaje automático. Esto evita el sobreajuste, lo que nos proporciona modelos más precisos [DWNP19].

Los valores de los distintos intervalos se han obtenido por generación de frecuencias, conocimiento experto, habiéndose etiquetado todos ellos. El etiquetado de estos intervalos se muestra a continuación en la **Tabla 5**. Se ha utilizado en esta un conjunto de términos en inglés, utilizados también a la hora de publicar artículos, que se explican a continuación.

- La variable ANA1 hace referencia a la producción científica de un investigador en forma de artículos publicados en revistas indexadas. Se han utilizado los términos *No activity* (0-1994), *Starters* (1995-3988), *Low activity* (3989-5893), *Medium activity* (5984-7977) e *Intense activity* (7978-9971), que no es sino una categorización, en cinco intervalos, de estos autores, desde los que no realizan publicaciones hasta aquellos que realizan una intensa actividad publicadora. Los intervalos se han establecido utilizando frecuencias iguales.
- La variable ANA2 hace referencia a la producción científica por parejas de investigadores que publican juntos. Hablamos en este punto de cooperación entre investigadores habiéndose establecido la clasificación de la siguiente forma, *No cooperation* (0-1495), *Low cooperation* (1496-2990), *Medium cooperation* (2991-4485) y *Strong cooperation* (4486-5980), que indica el grado de cooperación existente.
- La variable ANA3 indica publicaciones entre 3 autores. Hablamos entonces de redes de investigadores que publican juntos y establecemos la clasificación como *No networking* (0-1163), *Low-medium networking* (1164-2325) y *Strong networking* (2326-3488), que nos indica el grado de cooperación entre grupos de investigadores.
- Las variables ANA4, …, ANA8 indican publicaciones entre 4, …, 8 autores. Seguimos hablando de redes de investigadores que publican juntos y establecemos la clasificación como *No networking*, *Networking*, que nos indica el grado de cooperación entre grupos de investigadores. Este valor va a depender de los valores de la variable para establecer los intervalos.
    - ANA4 *No networking* (0-825) *Networking* (826-1649)
    - ANA5 *No networking* (0-381) *Networking* (382-761)
    - ANA6 *No networking* (0-175) *Networking* (176-350)

  - *ANA7* *No networking* (0-66) *Networking* (67-132)
  - ANA8 *No networking* (0-27) *Networking* (28-54)
- Las variables PENA1 y PENA2 indican la proclividad de los investigadores a patentar su trabajo. PENA1 hace referencia a las patentes individuales, con un único inventor, y PENA2 hace referencia a las patentes realizadas por parejas de investigadores. En ambos casos se ha establecido la clasificación de la siguiente forma, *No patenting culture*, *Low-medium patenting culture* y *Strong patenting culture*, que establece intervalos regulares que enmarcan al investigador por su tendencia a patentar.
  - PENA1
    - *No patenting culture* (0-1280)
    - *Low-medium patenting culture* (1281-2560)
    - *Strong patenting culture* (2561-3840)
  - PENA2
    - *No patenting culture* (0-109)
    - *Low-medium patenting culture* (110-217)
    - *Strong patenting culture* (218-326)
- La variable PENA3 indica que tres inventores patentan de forma conjunta. El número de intervalos establecidos para esta variable es dos, *No patenting culture* (0-12) y *Patenting culture* (13-24), que indican si se produce o no la patente, pero no en que grado.
- La variable PASA1, variable objetivo, se establece como una variable binaria, se produce o no se produce la generación de propiedad intelectual. Esto se dispone con dos intervalos cuyos valores son, *No IP production* (0-2198) y *IP production* (2199-4395), que indican si se ha creado nueva propiedad intelectual.

Tabla 5: Valores resultantes de la discretización y etiquetado para los tres grupos de variables del proceso, ANA (inversión acumulada), PENA (inclinación a la generación de propiedad intelectual) y PASA (generación real de propiedad intelectual).

Variables			
ANA1 (5)	**ANA2 (4)**	**ANA3 (3)**	**ANA4 (3)**
• *No activity* • *Starters* • *Low activity* • *Medium activity* • *Intense activity*	• *No cooperation* • *Low cooperation* • *Medium cooperation* • *Strong cooperation*	• *No networking* • *Low-medium networking* • *Strong networking*	• *No networking* • *Networking*
ANA5 (2)	**ANA6 (2)**	**ANA7 (2)**	**ANA8 (2)**
• *No networking* • *Networking*	• *No networking* • *Networking*	• *No networking* • *Networking*	• *No networking* • *Networking*
PENA1 (3)	**PENA2 (3)**	**PENA3 (3)**	**PASA1 (2)**
• *No patenting culture* • *Low-medium patenting culture* • *Strong patenting culture*	• *No patenting culture* • *Low-medium patenting culture* • *Strong patenting culture*	• *No patenting culture* • *Patenting culture*	• *No IP production* • *IP production*

La variable objetivo de este estudio es PASA1, que permite obtener un modelo de predicción de la innovación, en este caso a través de la predicción de patentes. También se han creado y analizado, entre otros, un modelo de red bayesiana y un modelo de árbol de decisión para prever el comportamiento de esta variable dependiente. En la sección de resultados se ofrecerá más información sobre el rendimiento de ambos modelos.

3.4.2.1 Discretización de datos.

En el Anexo 1 se explica con detalle cómo se ha llevado a cabo del proceso de discretización. Se han preparado un conjunto de 35 juegos de datos cuyas variables ANA, PENA y PASA han sido segmentadas en intervalos similares, es decir, por frecuencia. A estos conjuntos de datos se les ha aplicado un conjunto de 24 algoritmos cuyos resultados, correspondientes a las métricas descritas en el punto 1.6.2 Métricas utilizadas., se presentan también en dicho Anexo 1. Dichas métricas son (i) el porcentaje de precisión, (ii) el error absoluto medio, MAE, (iii) la raíz del error cuadrado medio, RMSE, (iv) la precisión, (v) la exhaustividad y (vi) el valor-F.

3.4.3 Resultados.

Si analizamos las diferentes tablas del Anexo 1, que contienen los resultados de las métricas utilizadas en esta investigación podemos observar que, por ejemplo, en la **Tabla 12** que muestra el porcentaje de precisión o correctitud, los mejores resultados se obtienen para las discretizaciones D7-3 t D7-5 con el algoritmo C4.5 con un valor de 82,50. Teniendo en cuenta solamente los valores de MAE y RMSE, **Tabla 13** y **Tabla 14**, la mejor discretización sería las D6-4 con los algoritmos ANN en el primer caso con un valor de 0,09 y con TAN, KNN, C4.5 y *Random Forest* con un valor 0,23 en el segundo. Si miramos las tablas de precisión, exhaustividad y valor-F, **Tabla 15**, **Tabla 16** y **Tabla 17**, lo que obtenemos es que, para la precisión, la discretización D1-1-5 con el valor 0,84 y con los algoritmos SVM (*Polynomial*), SVM (*Normalized Polynomial*) y SVM (RFB) es la que presenta mejores resultados. Analizando la exhaustividad, la discretización D3-1-5 obtiene el mejor valor 0.99 para el algoritmo SVM (RBF) y, asimismo, la discretización D6-4 obtiene el mismo valor con los algoritmos SVM (*Normalized Polynomial*) y SVM (RBF). Para El valor-F, se alcanza el valor de 0,89 en las discretizaciones D4-4 y D6-4. La primera de ellas lo obtiene para los algoritmos ANN (MLP), ANN (MLP CS), SVM (*Normalized Polynomial*), SVM (Pearson VII), KNN, C4.5 y *Random Forest*, mientras que para la segunda se obtiene dicho valor con los algoritmos ANN (MLP), ANN (MLP CS) y *Random Forest*.

Para la explicación del modelo, que se presenta a continuación, se ha tomado como referencia la discretización D5-1, es decir, la que presenta un porcentaje de 82,49, prácticamente el más elevado, para poder analizar la influencia de todas las variables. Se han construido sobre ella, por su fácil representación e interpretabilidad, los modelos de red Bayesiana y árbol de decisión que se exponen en los siguientes puntos.

3.4.3.1 Red Bayesiana.

La red bayesiana[1], **Figura 35**, para PASA1 como variable objetivo ha alcanzado una precisión del 84,97%. Los valores de MAE, que muestra un valor de 0,21, y RMSE,

[1] La red Bayesiana se ha calculado con un algoritmo anterior al utilizado en las tablas del Anexo 1 por lo que los valores son ligeramente diferentes.

con un valor de 0,33, indican que no tenemos un mal modelo pero que hay margen de mejora [RMSE22].

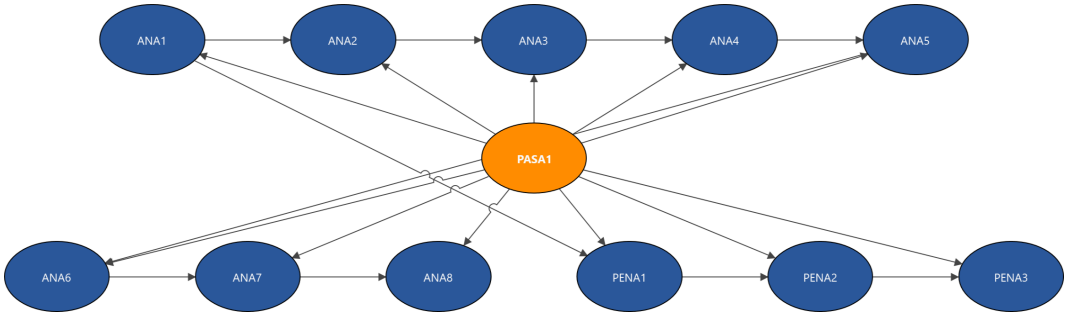

Figura 35: Modelo de red bayesiana con TAN obtenido para la variable objetivo PASA1. La figura muestra las relaciones entre variables que permiten la generación real de propiedad intelectual, que es la generación de nuevas patentes considerando un solo inventor.

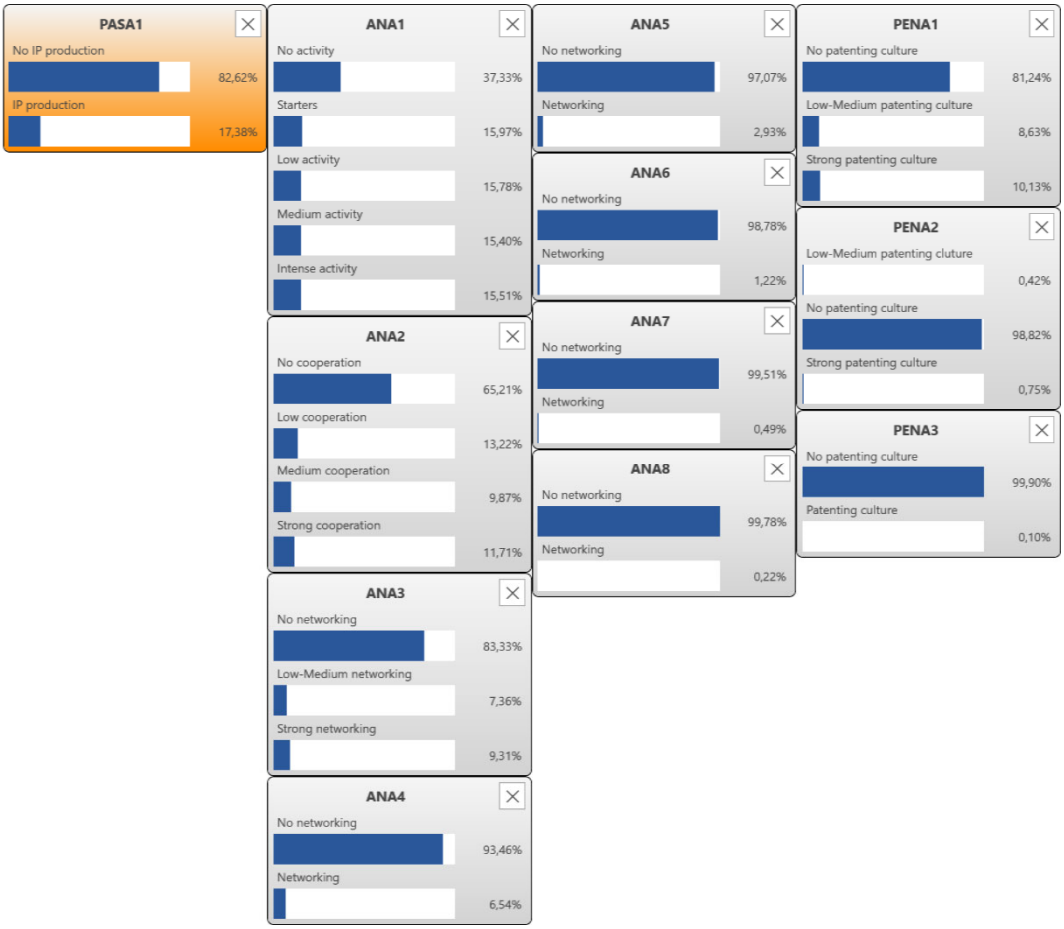

Figura 36: Monitores utilizados para comprobar no sólo el comportamiento de la variable objetivo, en este caso, PASA1, sino también grupos de variables no dependientes (ANA y PENA).

Utilizando esta red Bayesiana y la propagación de probabilidad condicional, podemos probar diferentes comportamientos para nuestra variable objetivo. De hecho, esta prueba puede realizarse de dos formas distintas. Podemos establecer diferentes valores para las variables independientes, ANA1 a ANA8 y PENA1 a PENA4. Como resultado, podemos ver y analizar los diferentes valores de la producción de PI para la variable dependiente PASA1. Por el contrario, podemos fijar el valor para la variable PASA1 y analizar el comportamiento del resto de variables independientes (ver **Figura 36**).

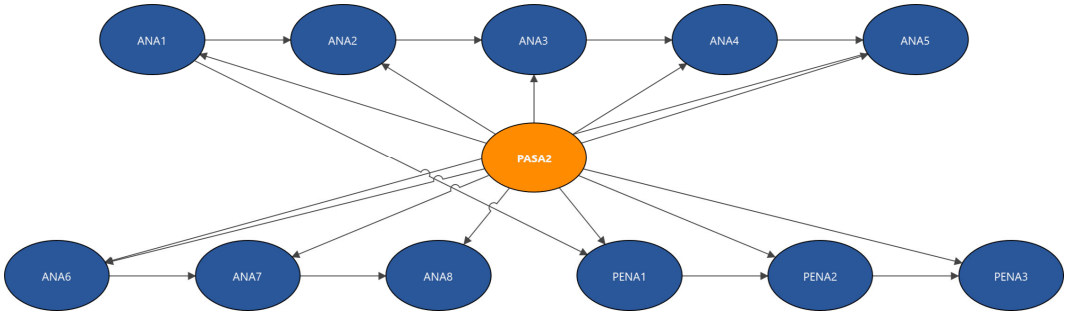

Figura 37: Modelo de red bayesiana para la variable objetivo PASA2. La figura revela la combinación de variables que generan PI real considerando pares de inventores.

La **Figura 37** muestra el modelo de red bayesiana para la variable objetivo PASA2 con una precisión del 98,55%. Los valores de MAE y RMSE son 0,03 y 0,12 respectivamente. A pesar de la elevada precisión, el modelo no es realmente representativo, aunque permite extraer algunas conclusiones. La razón principal es que el número de ceros es muy elevado para esta variable, 11977 de 12479 valores, como hemos explicado anteriormente. Esto indica un modelo muy preciso, pero poco útil.

Se ha probado también con el resto de las variables objetivo desde PASA3 a PASA8, pero los resultados obtenidos con estos análisis no son relevantes para esta investigación al presentar unos valores poco coherentes.

3.4.3.2 Árbol de decisión.

De la misma forma se planteó el desarrollo de un modelo de árbol de decisión[2]. Se estableció como variable objetivo la variable PASA1 obteniéndose los siguientes resultados.

- **Correctitud**: 85,26% frente al 84,97% del modelo anterior, red Bayesiana.
- **MAE**: 0,22 frente al 0,21, prácticamente idéntico, en el modelo de red Bayesiana.
- **RMSE**: 0,33 en ambos modelos.

Como puede verse en la **Figura 38**, que representa el modelo de árbol de decisión, todas las variables ANA, desde ANA1 hasta ANA8 (excepto ANA7) aparecen implicadas, de una u otra forma en el proceso, lo cual indica que la publicación de artículos científicos en revistas indexadas tiene gran relevancia. La ausencia de la variable ANA7, no es fácil de explicar, ya que la variable ANA8, con características muy similares, si está presente. Puede explicarse su ausencia por la gran cantidad de valores cero obtenidos el procesar los datos iniciales, es decir, es poco significativa al igual que la variable ANA8. No obstante, observando el comportamiento de esta última variable, vemos que contribuye ligeramente a la generación de patentes, pero solo cuando existe proclividad media-alta de los inventores a patentar, es decir, la clave de la rama donde se encuentra dicha variable es el hecho en sí de patentar un descubrimiento, no de publicar artículos junto con otros autores. Así mismo aparecen todas las variables PENA que hemos considerado, es decir, PENA1, PENA2 y PENA3 lo cual revela que la proclividad a patentar sus descubrimientos por parte de los inventores, a su vez autores de artículos científicos, también es relevante.

[2] El árbol de decisión se ha calculado con un algoritmo anterior al utilizado en las tablas del Anexo 1 por lo que los valores son ligeramente diferentes.

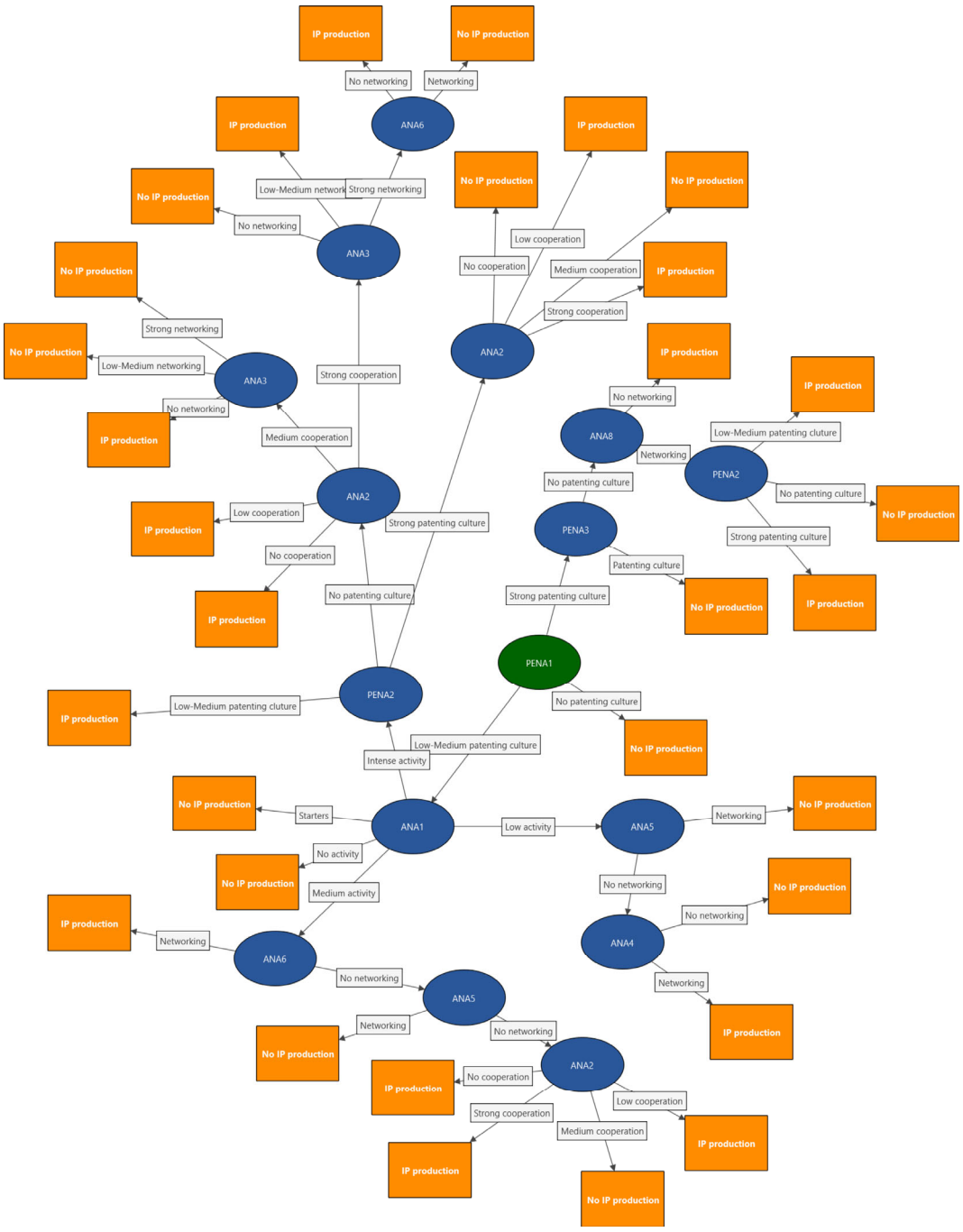

Figura 38: Modelo de árbol de decisión para la variable objetivo PASA1. La figura muestra las condiciones en las que se genera la propiedad intelectual real, lo que nos permite esbozar un conjunto de reglas de comportamiento.

Analizando el árbol de decisión, observamos una serie de diferentes escenarios que representaremos y analizaremos en la **Tabla 6**. Se ha estimado, también, el

porcentaje de producción de propiedad intelectual correspondiente a cada uno de los escenarios en los que se produce generación real de propiedad intelectual. El resto de las situaciones se descarta por ser irrelevante en términos de futura producción de propiedad intelectual.

Se presentan a continuación, en la **Tabla 6**, los diferentes escenarios, representados por la primera columna, ID, de la tabla. Se ha utilizado como origen el modelo de árbol de decisión junto con un cálculo estimativo de la producción real de propiedad intelectual, es decir, la generación de patentes.

Se han utilizado en la **Tabla 6**, por motivos de espacio, las siguientes abreviaturas.

- Variables PENA1, PENA2 y PENA 3.
  - No PC. *No patenting culture*.
  - L-M PC. *Low-Medium patenting culture*.
  - S PC. *Strong patenting culture*.
- Variable ANA1.
  - L A. *Low activity*.
  - M A. *Medium activity*.
  - I A. *Intense activity*.
- Variable ANA2.
  - No C. *No cooperation*.
  - L C. *Low cooperation*.
  - M C. *Medium cooperation*.
  - S C. *Strong cooperation*.
- Variables ANA3, ANA4, ANA5, ANA6 y ANA8.
  - No N. *No networking*.
  - N. *Networking*.

El escenario 13, con un 38,97% de producción de propiedad intelectual (PI), indica que un inventor que posee una cultura baja-media para patentar, con una baja actividad publicando de forma individual, pero cooperando con otros autores en algunas publicaciones puede dar lugar a la generación de PI, aunque no en gran medida.

Por otra parte, los escenarios 8, 11 y 14 se caracterizan por una cultura baja-media para patentar por parte del inventor y una actividad media a la hora de publicar artículos de forma individual. Hay un cierto *networking* en el escenario 8 lo que eleva el porcentaje de producción de IP al 46,07%, pero no lo hay en los escenarios 11 y 14. En estos dos últimos aparece una media y fuerte cooperación para las publicaciones por parejas lo cual disminuye la producción de IP a valores de 41,83% y 35,83% respectivamente. Aparentemente estos autores/inventores se centran más en difundir sus conocimientos que en proteger sus inventos.

Tabla 6: Escenarios con producción real de propiedad intelectual considerando como variable objetivo la variable PASA1, es decir, la generación de patentes.

ID	PENA1	PENA2	PENA3	ANA1	ANA2	ANA3	ANA4	ANA5	ANA6	ANA8	% IP Production
1	L-M PC	No PC		I A	No C						70,81
2	L-M PC	No PC		I A	L C						68,82
3	S PC		No PC							No N	60,12
4	L-M PC	S PC		I A	L C						59,55
5	L-M PC	No PC		I A	M C	No N					55,41
6	L-M PC	No PC		I A	S C	S N			No N		52,75
7	S PC	L-M PC	No PC							N	52,10
8	L-M PC			M A					N		46,07
9	L-M PC	No PC		I A	S C	L-M N					44,94
10	S PC	S PC	No PC							N	44,30
11	L-M PC			M A	M C			No N	No N		41,83
12	L-M PC	S PC		I A	S C						40,22
13	L-M PC			L A			N	No N			38,97
14	L-M PC			M A	S C			No N	No N		35,83
15	L-M PC	L-M PC		I A							34,97

En el caso del escenario 15 se caracteriza por una baja-media cultura para patentar tanto de forma individual como en parejas, pero una intensa actividad a la hora de publicar artículos. Esto reduce la generación de PI al valor más bajo de todos los escenarios alcanzando únicamente un 34,97%.

En lo concerniente a los escenarios 1, 2, 5, 9 y 6, estos son similares en cuanto a que existe una cultura baja-media a la hora de patentar y esto solo se hace de forma individual, no por parejas ni en grupos de 3. Poseen una alta actividad a la hora de realizar publicaciones científicas de forma individual, mientras que el escenario de publicación por parejas es diverso cubriendo todo el espectro. Los mayores

porcentajes de producción de PI se dan en los escenarios 1 y 2 dónde no existe la publicación por parejas o esta es baja con un 70,81 y un 68,82% respectivamente. Según la cooperación para publicar de forma conjunta aumenta disminuyen los porcentajes de producción de PI. Así, en el escenario 5 con una cooperación media y sin *networking* para el caso de 3 autores estamos en un 55,41%, en el escenario 9 con una cooperación fuerte para ANA2 y un bajo-medio *networking* para ANA3 bajamos al 44,94% y en el escenario 6, con una cooperación fuerte en ANA2 y un fuerte grado de *networking* en ANA3 estamos en un 52,75%.

Los escenarios 4 y 12 presentan similitudes estando caracterizados por una cultura baja-media de cara a la generación de patentes de forma individual, una fuerte cultura de patentar por parejas y una intensa actividad para la publicación de artículos científicos. Cuando la cooperación para publicar por pares es baja, como sucede en el escenario 4, obtenemos un 59,55%. Sin embargo, cuando este tipo de cooperación aumenta, como sucede en el escenario 12, el valor decrece hasta el 40,22% de generación de propiedad intelectual.

Por último, los escenarios 3, 7 y 10 también presentan similitudes. Se caracterizan por una fuerte cultura para patentar en solitario, mientras que la cultura para patentar por parejas va desde baja-media a fuerte. Se caracterizan también por no presentar actividad relativa a la publicación de artículos científicos ni de forma individual ni en colaboración. El escenario 3 no presenta ningún tipo de colaboración, e incluso presenta un no networking en cuanto a la variable ANA8. Se consigue así un 60,12% de generación de PI. Los escenarios 7 y 10 presentan *networking* en la variable ANA8 lo cual produce un 52,10% y un 44,30% de generación de propiedad intelectual respectivamente.

En el mismo orden de ideas, se ha generado un modelo de árbol de decisión para la variable objetivo PASA2, mostrado en la **Figura *39***, tiene una precisión del 98,52%. Los valores de MAE y RMSE son 0,03 y 0,12 respectivamente.

De la misma forma que sucede con el modelo de red Bayesiana, a pesar de conseguirse una elevada precisión, el modelo no es realmente representativo, aunque permite muestra dos escenarios de los cuales puede extraerse cierta información. Como en el caso anterior, la razón principal es que el número de ceros

es muy elevado para esta variable, 11977 de 12479 valores, lo que representa un 95,97% de los valores del conjunto de datos. Esto indica un modelo muy preciso, pero poco útil.

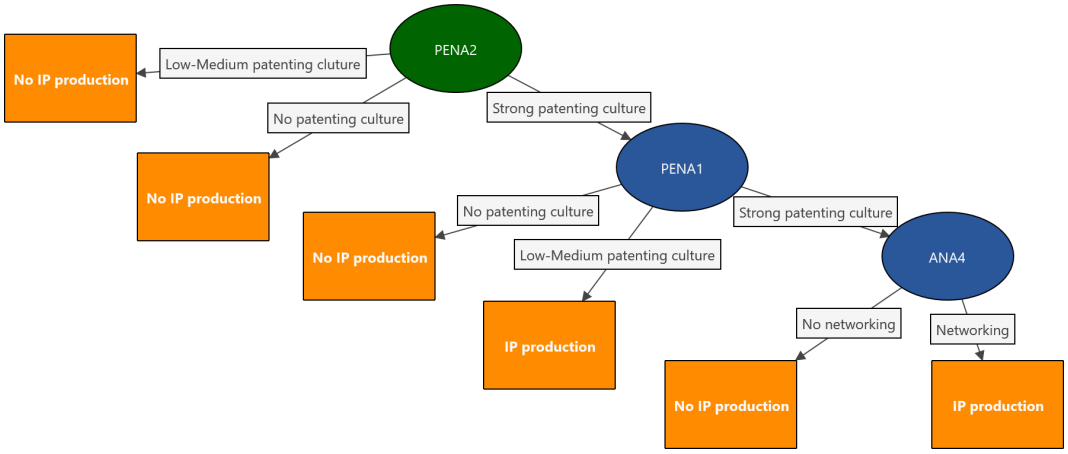

Figura 39: Modelo de árbol de decisión para la variable objetivo PASA2. La figura muestra las situaciones en las que se genera el PI real que permite establecer un conjunto de reglas de comportamiento

Las conclusiones extraídas se resumen, a continuación, en la **Tabla *12***.

Tabla 7: Escenarios con producción real de propiedad intelectual considerando como variable objetivo la variable PASA1, es decir, la generación de patentes.

ID	PENA1	PENA2	ANA4	% IP production
2	Strong patenting culture	Strong patenting culture	Networking	81,71
1	Low-Medium patenting culture	Strong patenting culture		76,30

Estos escenarios son muy similares e indican que cuando los inventores tienen cultura para patentar sus inventos de forma individual, pero también en colaboración con otros inventores para patentar por parejas el porcentaje de producción real de propiedad intelectual es considerablemente alto y se incrementa aún más cuando establecen relaciones de *networking* a la hora de publicar sus artículos científicos. Esto hace que lleguemos a porcentajes del 76,30% y 81,71% para los escenarios 2 y 1 respectivamente.

3.4.3.3 Evolución en el tiempo.

Otro de los aspectos relevantes para esta investigación es el periodo de tiempo para el cual las predicciones realizadas son válidas, ya que estamos hablando de tecnologías cuyo comportamiento, evidentemente, puede y va a evolucionar en el tiempo. Teniendo en cuenta este aspecto, se han realizado análisis repitiendo el cálculo estadístico inicial, que proporciona el conjunto de datos para la generación de modelos, y calculando la precisión tanto para el modelo de la red Bayesiana como para el de árbol de decisión durante los 20 años de vigencia de la patente. Los resultados obtenidos se presentan en la siguiente gráfica.

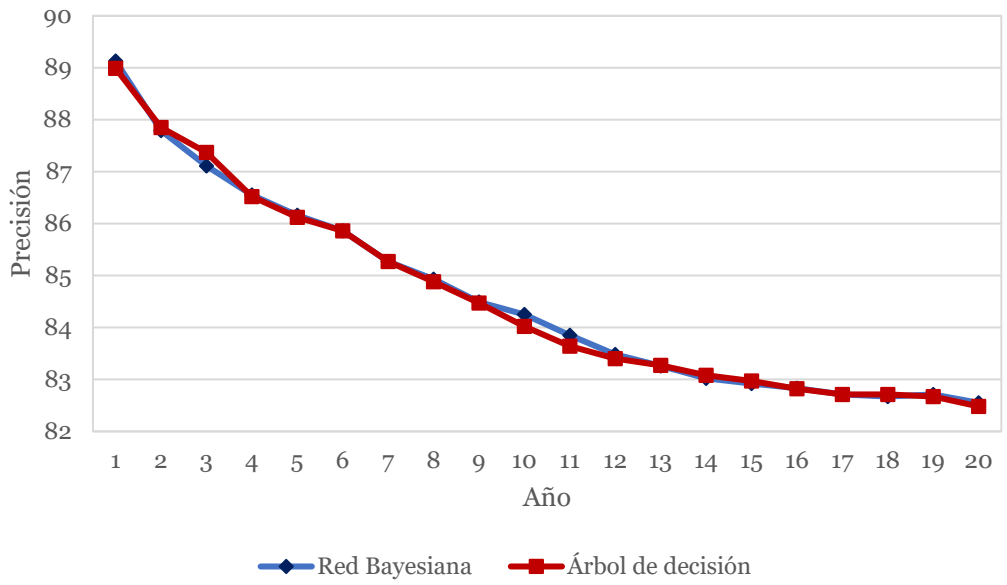

Figura 40: Evolución de la precisión para los modelos de red Bayesiana y árbol de decisión durante los 20 años de vigencia de las patentes.

Como puede verse al analizar la gráfica la precisión va decreciendo con el tiempo, pasando de casi un 90% en el primer año hasta un 82% aproximadamente al final de la vigencia de la patente. Teniendo en cuenta, por un lado, la rápida evolución de la tecnología analizada, estampación en caliente, y por otro lado los planes estratégicos que en las empresas suelen hacerse con un horizonte aproximado de 4 años, estaríamos hablando de precisiones por encima de un 86,5%, bajando a un 85% en torno a los 7 u 8 años, por lo cual podemos hablar de un buen modelo.

3.4.4 Discusión y conclusiones.

En esta investigación, se ha puesto el foco en la predicción de la innovación como creación de una patente. La razón principal para hacerlo es que cuando se han analizado los trabajos relacionados, no se ha encontrado una metodología capaz de predecir la aparición de una patente basada en la generación de literatura científica y el análisis de la tendencia de los inventores a patentar. El objetivo ha consistido en establecer un modelo, así como un conjunto de reglas que contribuyan a predecir la generación real de PI en base a la creación de patentes. Se ha elegido la tecnología de estampación en caliente, debido a que está claramente reconocida como una tecnología relevante [NGUSBS14] y al conocimiento de esta por parte de uno de los directores de esta tesis.

En este trabajo de investigación se han utilizado específicamente técnicas estadísticas y de inteligencia artificial con el objeto de predecir la innovación, concretamente la generación de nuevas patentes. Como fuente de datos se ha utilizado no sólo literatura científica, sino también información sobre patentes. Para llevar a cabo esta investigación se extrajeron varios artículos científicos de WoS, centrándose en el periodo desde el inicio de esta tecnología hasta 2021 incluido. Del mismo modo, las patentes se extrajeron de PatBase®. De todos los metadatos existentes tanto para los artículos científicos como para las patentes, sólo se han utilizado para esta investigación los nombres del autor del artículo o inventor de la patente, así como el año de publicación del artículo y de la patente.

En este punto hay que tener en cuenta que hay diferentes autores con el mismo apellido, pero diferentes nombres que empiezan por la misma letra, por ejemplo, John Doe y Jane Doe. Son personas diferentes, pero cuando extraemos los metadatos de nuestras bases de datos seleccionadas ambos son Doe, J. Este punto implica un cierto grado de error cuando se procesan los datos.

En relación con el nombre del inventor, es importante señalar que, en teoría, la estructura del nombre completo del inventor es "Nombre Primer apellido Segundo apellido". Esto es sólo teórico, ya que al analizar el conjunto de datos obtenido se puede observar una estructura diferente que contiene, por ejemplo, el nombre de la empresa propietaria de la patente, diferentes títulos como Dr., Ing. o Lic., el país

de la empresa o del inventor y cualquier otra palabra que no tenga cabida en dicho contexto. Además, teniendo en cuenta la necesidad de adaptar los nombres de los autores de los artículos y de los inventores de las patentes a la estructura "Apellido, Inicial del nombre", se requiere la pericia del investigador y también una gran cantidad de tiempo para optimizar el primer conjunto de datos. Este problema se repite con los inventores de las patentes.

Como desarrollo posterior en futuras investigaciones se ha considerado, en primer lugar, identificar todas las palabras, o conjuntos de estas, descritas anteriormente y que complican sobremanera la identificación del inventor, aplicando un sistema automático de limpieza o preprocesado para nuestro conjunto de datos, que evite la realización de este procedimiento de forma manual.

En lo que se refiere al proceso realizado, el primer paso consistió en el cálculo de la inversión acumulada, la proclividad a la generación de PI y la generación real de PI mediante un software desarrollado a tal efecto que, partiendo de los conjuntos de datos adquiridos de WoS y PatBase®, genera un nuevo conjunto de datos que contiene las variables ANA1 a ANA8, PENA1 a PENA8 y PASA1 a PASA8 antes mencionadas. Posteriormente, el conjunto de datos resultante se procesa de nuevo para adaptar estos datos de salida a nuestros modelos de aprendizaje automático. En primer lugar, creamos una red bayesiana y, a continuación, reprodujimos el método para elaborar un árbol de decisión. Ambos nos han permitido extraer ciertas reglas. La precisión o correctitud de estos modelos, tal y como se muestra en la **Tabla 12**, se mantiene entre el 64,91% y el 82,50% en función de la discretización y el algoritmo utilizado, obteniéndose los mejores resultados se obtienen para las discretizaciones D7-3 t D7-5 con el algoritmo C4.5 con un valor de 82,50%. Para los valores de MAE, **Tabla 13**, la mejor discretización sería las D6-4 con los algoritmos ANN con un valor de 0,09. En el caso de RMSE, **Tabla 14**, sería también la mejor discretización las D6-4 con un valor 0,23 para TAN, KNN, C4.5 y *Random Forest*. Si examinamos el valor-F, dependiente de los valores de precisión y exhaustividad, **Tabla 17**, se alcanza el valor de 0,89 en las discretizaciones D4-4 y D6-4. La primera de ellas lo obtiene para los algoritmos ANN (MLP), ANN (MLP CS), SVM (Normalized Polynomial), SVM (Pearson VII),

KNN, C4.5 y *Random Forest*, mientras que para la segunda se obtiene dicho valor con los algoritmos ANN (MLP), ANN (MLP CS) y *Random Forest*.

Se trata de modelos que, examinando la literatura científica, las patentes, los autores y los inventores, han demostrado que existe una interrelación entre ellos. Estos son los escenarios resultantes que producen la generación de PI utilizando PASA1 como variable objetivo.

Teniendo en cuenta que la tecnología analizada es la de estampación en caliente, los principales escenarios utilizando PASA1 como variable objetivo indican que una cultura de patentes es en cierta medida esencial para continuar generando PI, combinada con una alta producción de literatura científica. Algunos otros escenarios produjeron PI en menor grado. Utilizando PASA2 como variable objetivo podemos confirmar que una cultura de patentes es crucial para continuar generando PI y la cooperación en la redacción de artículos aumenta el proceso de generación de PI. En este punto, no podemos afirmar que este comportamiento sea similar cuando analicemos tecnologías diferentes.

En conclusión, se han logrado algunos modelos, pero aún hay posibilidades de mejora. A continuación, se describen algunos puntos que podrían conducir a mejores resultados.

1. El conjunto de datos con el que hemos trabajado tiene un número muy elevado de registros (eventos) que quizás podrían conducir a otros resultados si las consultas de partida fuesen más específicas.
2. Hay un gran número de ceros en algunas variables.
3. Hay más metadatos que pueden utilizarse en este tipo de análisis.

Obtener conjuntos de datos más específicos podría ser crucial para el primer punto. Además, dar menos datos a nuestros modelos aprendizaje automático hará que Cambie su funcionamiento y el tiempo de procesamiento manual de los datos necesario para llevar a cabo la investigación sería considerablemente menor.

Para abordar la última mejora, se podrían utilizar más metadatos, como la entidad a la que pertenece el autor y/o inventor, la combinación autor-institución e inventor-institución, entre otros. Incluir esta nueva información en nuestros

modelos permitirá descubrir nuevas relaciones y mejorar nuestro método de predicción de la innovación.

Además, debemos considerar otras posibles fuentes de datos como conferencias, patentes, informes técnicos o proyectos de I+D.

3.5 Sumario.

A lo largo de este capítulo se ha analizado con gran exhaustividad la tecnología de estampación en caliente. Se comenzó realizando una introducción de esta, en la cual se puso de relevancia la gran importancia que esta tecnología tiene actualmente, a pesar de ser una tecnología joven, impulsada principalmente por la industria del automóvil. Los factores que han influido en su desarrollo son, por ejemplo, la necesidad de reducción de peso de los componentes de fabricación de los vehículos, la minimización del consumo de combustible y la emisión de CO_2 y diferentes avances en materia de seguridad.

A continuación, se realizó un análisis bibliométrico que analizaba la evolución de los 10 últimos años de la estampación en caliente. Se analizaron aspectos relevantes como el número de publicaciones científicas en revistas indexadas por año, el número de artículos científicos publicados en las diez revistas indexadas más influyentes, el número de artículos científicos publicados por los diez autores más relevantes, el número de artículos científicos publicados por las diez instituciones/empresas más relevantes, el número de citas por año, el valor del índice h, el número de citas por autor, por institución o revista y el número de publicaciones de artículos científicos por país entre otros. Se llegó incluso a crear una red de palabras clave y otra formada por los autores de artículos científicos en revistas indexadas. Se concluyó que a pesar de obtenerse información valiosa como conocer los grupos de investigadores que trabajan juntos, la importancia relativa de estos, cuál es su área de investigación e incluso tratar de predecir futuras tendencias no era suficiente para abordar el problema de la predicción de la generación de propiedad intelectual.

Posteriormente, se planteó un nuevo análisis bibliométrico basado en la creación y análisis de mapas evolutivos. Que ayudan a analizar el campo de interés estudiado,

en este caso la tecnología de estampación en caliente, así como a detectar y visualizar los temas o áreas conceptuales y su desarrollo a lo largo del tiempo. Este estudio se ha desarrollado con la ayuda del programa SciMAT y nos ha ofrecido una visión de la evolución de esta tecnología desde mediados de la década de 1950 hasta la actualidad, no sólo en cuanto a los temas clave, sino también a sus relaciones y evolución a lo largo de los tres períodos definidos (1950-2009, 2010-2015 y 2016-2019), marcados por dos hitos importantes que han sido dos importantes revisiones llevadas a cabo por Karbasian, H. (2010) y Merklein, M. (2016). Este análisis presenta ciertos inconvenientes como la limitación temporal relativa al periodo definido en esta investigación que podría eliminarse llevando a cabo un análisis temporal en tiempo real y hay que considerar también que SciMAT sólo utiliza el título, las palabras clave y el resumen para caracterizar la evolución de la literatura. Si añadimos el alto coste en términos de tiempo que se necesita para realizar el análisis y el conocimiento experto requerido, se decidió abandonar esta vía a pesar de la valiosa información proporcionada.

Finalmente, se propone una nueva metodología para la predicción de la innovación basada en la utilización conjunta de métodos estadísticos y aprendizaje automático. Se propone utilizar datos relativos a publicaciones científicas, concretamente artículos publicados en revistas indexadas, y datos referentes a patentes para la generación de modelos. Estos datos se preprocesan para obtener un conjunto que pueda ser procesado mediante técnicas estadísticas, cuya salida será la entrada al proceso de aprendizaje automático. Se generan dos modelos que alcanzan precisiones en el rango del 82,50% que dan lugar a una serie de patrones de comportamiento que establecen cuando va a producirse la generación de propiedad intelectual. Se analiza, por último, la evolución en el tiempo de estas predicciones ante un posible futuro en constante cambio, garantizando su validez en el entorno de los 4 años, que generalmente es el periodo de validez de un plan estratégico.

Para concluir, el análisis elaborado con la tecnología de estampación en caliente demuestra la hipótesis de partida, no obstante, para ver la transversalidad de la solución propuesta se realizará un segundo análisis con la tecnología de función de hierro que se expone en el siguiente capítulo.

4

Caso de estudio: Fundición de hierro

El hierro fundido (del término inglés *Cast Iron*) es uno de los materiales más antiguos y versátiles en lo que a metalurgia se refiere. Hacia finales del año 5000 a.C., los humanos aprendieron a fundir minerales de cobre para obtener cobre casi puro. El cobre refinado se vertía en moldes abiertos para dar forma al objeto deseado. Una vez solidificado, se forjaba para obtener la forma definitiva. En Anatolia se han encontrado crisoles de arcilla y ejemplos de este tipo de fundiciones [ACOC04]. Hacia el 3500 a.C., los primeros metalúrgicos aprendieron a producir la primera aleación fundida, el bronce arsenical. El hierro fundido tiene al menos 2.500 años de antigüedad, aunque aún no se había inventado la palabra solidificación y siguió siendo un arte durante muchísimos años. Desempeñó un importante papel en la revolución industrial [SEM98] convirtiéndose en un elemento esencial para el desarrollo de la humanidad.

A medida que la especie humana ha evolucionado de forma gradual desde la edad de hierro a la era de los materiales de ingeniería, de todos los procesos de conformado de metales, el proceso de fundición sigue siendo la ruta más directa y corta desde el diseño del componente hasta el producto acabado. Esto convierte a la fundición en uno de los principales procesos de fabricación, mientras que las

aleaciones de fundición son algunos de los materiales más utilizados. Entre 2012 y 2013, mientras que la economía mundial se estancó en su mayor parte, la producción de fundición aumentó un 2,4% y un 3,4%, respectivamente [SAEOCS15].

Las principales razones de la longevidad del proceso de fundición son la amplia gama de propiedades mecánicas y físicas de las aleaciones de fundición, la versatilidad del proceso (peso desde gramos a cientos de toneladas, fundición de cualquier metal que pueda ser corrugado, formas complejas que no pueden producirse con otros métodos de fabricación) y el precio competitivo de los de los productos fabricados. Aunque las piezas de fundición son invisibles en muchas de sus aplicaciones, ya que pueden formar parte de equipos complejos, se utilizan en el 90% de los productos manufacturados.

En consecuencia, ha desempeñado, y aún lo hace en la actualidad, un papel relevante en diversas industrias como pueden ser la fabricación de herramientas y maquinaria, fabricación de armas e incluso un uso importante en arquitectura e ingeniería.

Este capítulo está organizado de la siguiente manera. En primer lugar, en la sección 4.1, se hará una breve introducción a la tecnología de fundición de hierro o hierro fundido destacando su aplicabilidad en los diferentes tipos de industria. A continuación, en la sección 4.2, se volverán a aplicar las técnicas ya explicadas en la sección 3.4 volviendo a examinar la creación de patentes a partir de artículos científicos y patentes ya existentes, mediante el empleo de técnicas estadísticas combinadas con aprendizaje automático, para generar modelos. Del conjunto de resultados obtenidos, utilizaremos la red bayesiana y el árbol de decisión para identificar patrones referentes a la protección de la propiedad intelectual capaces de detectar, con alta probabilidad, la aparición de una nueva patente. Por último, la sección 4.3 contiene un pequeño resumen de todo lo analizado en este capítulo.

4.1 Una breve introducción a la tecnología de fundición de hierro.

El hierro fundido se define como un grupo de aleaciones ferrosas que presentan un contenido de carbono superior al 2,06% e inferior al 6,67% en peso o masa [CBMCI90]. Esta es la principal diferencia que presenta respecto el acero, el otro miembro principal de las aleaciones con base de hierro. El acero presenta un contenido de carbono inferior al 2,06 % [EMEA08]. Debido a este hecho, el hierro fundido solidifica presentando dos fases diferenciadas, el grafito o carburos de hierro (Fe_3C) y la matriz metálica. En cambio, el acero, debido a su menor contenido en carbono, sólo está formado por la matriz metálica, ya que todo el contenido en carbono presente en la aleación se disuelve en esta [FMSCI08].

El hierro fundido puede ser blanco (del inglés *White Cast Iron* (WCI)) o grafítico, dependiendo de la forma de precipitación del carbono. Si el proceso de solidificación tiene lugar de acuerdo con el diagrama metaestable Fe-C, **Figura 41**, entonces todo el carbono presente en la masa fundida líquida precipitará en forma de carburos de hierro. Si, por el contrario, la solidificación tiene lugar siguiendo el diagrama Fe-C estable, entonces todo el carbono presente en el metal líquido precipitará en forma de grafito. En este caso, el material se denomina hierro fundido grafítico [ACADIM22]. El grafito puede adoptar, en este caso, diferentes formas y basándose en su morfología, se pueden distinguir varios tipos de hierro fundido:

- Hierro fundido grafítico laminar o gris (del inglés *Lamellar Graphite Cast Iron* (LGI)) [TGNLGCI09].
- Hierro fundido compactado (del inglés *Compacted Graphite Cast Iron* (CGI)) [MWCGCI09].
- Hierro fundido grafítico dúctil o esferoidal (del inglés *Spheroidal Graphite Cast Iron* (SGI)) [CGFSGCI20].
- Hierro fundido maleable (del inglés *Malleable Graphite Cast Iron* (MCI)) [ICH81] [CSGGCI18].

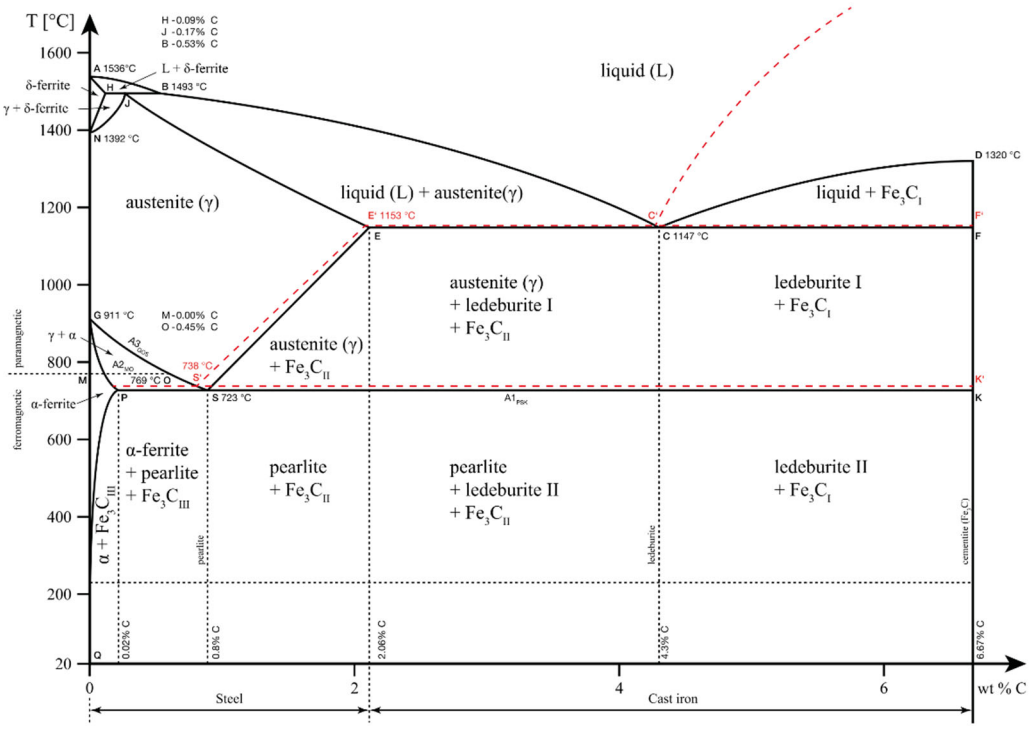

Figura 41: Diagrama de fases Fe-C estable y metaestable según Caesar [ICPD19].

Las propiedades mecánicas de los materiales de fundición presentan una gama muy amplia. La morfología del grafito es uno de los puntos clave que definirá estas propiedades. Así, SGI, LGI, CGI y MCI no pueden considerarse como un único grupo y deben tratarse por separado, ya que las diferentes formas de grafito que presentan influirán en las propiedades mecánicas con más fuerza que cualquier otra variable [ACIED16].

En cuanto a las fundiciones grafíticas, los requisitos SGI se definen en la norma EN 1563 [EN156318], los grados especiales de fundición dúctil denominados ADI (del término inglés *Austempered Ductile Iron*) se definen en la norma EN 1564 [EN156411], los LGI en la norma EN 1561 [EN156111] y los CGI en la norma ISO 16112 [ISO1611216].

Considerando las fundiciones grafíticas, una vez definida la morfología grafítica, las propiedades mecánicas vienen dadas por la microestructura. Dicha microestructura o matriz metálica formada tras la transformación sólido-sólido,

puede estar constituida por diferentes fases que se enumeran a continuación [FMSCI08] [EOMEA08].

- **Austenita**: Es una fase que presenta una ordenación cristalina cúbica centrada en las caras o FCC (del inglés *Face Centered Cubic*). Es la fase que presentan las fundiciones de hierro tras la solidificación y en aleaciones convencionales, no es estable a temperatura ambiente.
- **Ferrita**: Es una fase que presenta una ordenación cristalina cúbica centrada en el cuerpo o BCC (del inglés *Based Centered Cubic*). Es el resultado de una transformación eutectoide estable.
- **Perlita**. Es una estructura de doble fase que consiste en capas alternas de ferrita y carburo de hierro (Fe_3C). Es el resultado de una transformación eutectoide metaestable.
- **Ausferrita**. Es una microestructura de fase dual, formada por ferrita acicular o agujas de ferrita y austenita enriquecida en carbono o austenita reaccionada. La formación de esta fase tiene lugar cuando, partiendo de una microestructura austenítica, se lleva a cabo un temple isotermo, manteniendo la temperatura constante en un rango comprendido entre 250-450 ºC.
- **Martensita**. Se trata de una solución sólida sobresaturada de carbono en hierro y su estructura cristalográfica es una forma tetragonal centrada en el cuerpo (del inglés *Body Centred Tetragonal* (BCT)). Suele obtenerse por un enfriamiento rápido de la austenita, sin difusión atómica, sino por un proceso súbito de cizallamiento sin difusión y la consiguiente deformación de la red matriz en la de la martensita.
- **Bainita**. Es la fase que se obtiene cuando se prolonga el tiempo del periodo de transformación isotérmica para obtener ausferrita. Entonces, la austenita enriquecida en carbono no puede disolver más carbono y comienza a descomponerse en ferrita y pequeños carburos de hierro. La combinación de agujas de ferrita y carburos constituye la fase bainítica.

En cuanto a las aplicaciones del hierro fundido podemos encontrar las siguientes [CIT88].

- **Sector automoción.** Componentes para motores como pueden ser bloques y cabezas de cilindros, así como del sistema de frenado (calipers, horquillas y discos) y turbocompresores, colectores de escape o cajas diferenciales.
- **Sector ferrocarril:** Componentes del sistema de frenado, cajas de grasa y cambios de vía entre otros.
- **Vehículo industrial:** Sistemas de frenado, del motor, portamanguetas o componentes de la dirección.
- **Maquinaria.** Bases de máquinas, troqueles y engranajes.
- **Tuberías.** Sistemas de conducción de agua, gas y saneamiento.
- **Sector de agricultura y minería:** Componentes de los tractores como pueden ser las cajas de cambios o del sistema de frenado, así como horquillas para las cosechadoras y palas para las excavadoras.
- **Electrodomésticos**. Diferentes partes de los electrodomésticos como en estufas o fregaderos para las cocinas.
- **Sector energético**: Principalmente en lo referente a la energía eólica. Los componentes de hierro fundido tienen presencia tanto en los bujes como en los portasatélites.

En resumen, podemos concluir que la tecnología de hierro fundido ha tenido y sigue teniendo hoy en día una gran importancia en lo referente al desarrollo industrial al cual ha proporcionado gran cantidad de soluciones de forma continua y rentable para un amplio abanico de aplicaciones. Debido a su versatilidad, resistencia, así como la facilidad de producción a gran escala, la tecnología del hierro fundido constituye un pilar esencial en la fabricación moderna y en la ingeniería.

4.2 Aplicación del nuevo método de predicción de innovación a otras tecnologías.

La metodología que se va a aplicar a continuación para la tecnología de fundición de hierro es la ya explicada en el punto 3.4 para la tecnología de estampación en caliente. Procederemos de la siguiente forma.

1. **Adquisición y preprocesamiento del conjunto de datos**. Seleccionaremos para este análisis la base de datos de WoS para la recuperación de artículos científicos y PatBase® para la recuperación de información relativa a las patentes. Se analizarán los datos obtenidos para llevar a cabo un preprocesamiento de estos con el fin de homogeneizar el conjunto. Asimismo, se establecerán las variables a utilizar en el proceso.
2. **Metodología de cálculo**. La metodología que se aplicará consiste en el uso de un cálculo estadístico inicial que nos proporcionará las variables necesarias para la generación del modelo.
3. **Modelos de predicción**. Se utilizarán una serie de modelos de aprendizaje automático y se analizarán los resultados obtenidos. Esto determinará el mejor modelo obtenido, utilizándose redes bayesianas y árboles de decisión, por su simplicidad, para explicar el comportamiento de la tecnología.

4.2.1 Adquisición y preprocesamiento de datos.

Como ya se explicó con anterioridad en el punto 3.4.1, esta metodología se nutre, por un lado, de datos relativos a publicaciones científicas que extraeremos de la base de datos WoS, por la calidad de los datos, por la precisión de estos y por coherencia con las investigaciones desarrolladas con anterioridad. Por otro lado, se ha utilizado la base de datos de patentes PatBase® por la gran cantidad y calidad de la información en ella contenida, siguiendo así la línea de investigación iniciada en el caso de estampación en caliente.

Partiendo de WoS, se ha creado una consulta utilizando las siguientes palabras clave *Cast Iron*. La consulta se realizó manualmente utilizando su sitio web y tecleando este código. Posteriormente, se refinó la consulta configurando los siguientes filtros en base al conocimiento experto sobre esta tecnología.

- Tipo de documento: sólo artículos.
- Se incluyeron los años de publicación de 1900 a 2022.
- Sólo documentos en lengua inglesa.
- Áreas de investigación seleccionadas. Acústica, Agricultura, Anatomía y morfología, Arqueología, Arquitectura, Arte, Automatización y control de

sistemas, Bioquímica y biología molecular, Biofísica, Biotecnología y microbiología aplicada, Negocios y economía, Sistema cardiovascular y cardiología, Química, Ciencias de la computación, Construcción y tecnología de construcción de edificios, Cristalografía, Odontología, Cirugía oral y medicina, Educación e investigación educacional, Electroquímica, Energía y combustibles, Ingeniería, Ciencias medioambientales y ecología, Ciencia y tecnología alimentarias, Silvicultura, Geoquímica y geofísica, Geología, Instrumentos e instrumentación, Ciencias de la vida y biomedicina – Otros temas, Biología marina y de agua dulce, Ciencia de los materiales, Biología matemática y computacional, Métodos matemáticos en ciencias sociales, Matemáticas, Mecánica, Metalurgia e Ingeniería Metalúrgica, Meteorología y Ciencias Atmosféricas, Microscopía, Mineralogía, Minería y tratamiento de minerales, Ciencia y tecnología nuclear, Oceanografía, Investigación operativa y ciencias de la gestión, Óptica, Física, Ciencia de los polímeros, Salud pública, Medioambiental y laboral, Radiología, medicina nuclear e imagen médica, Robótica, Ciencia y tecnología - Otros temas, Ciencias sociales - Otros temas, Espectroscopia, Telecomunicaciones, Termodinámica, Transporte y Recursos hídricos.

En consecuencia, se obtuvieron 9847 artículos científicos publicados en revistas indexadas. Para este conjunto de artículos se adoptaron las siguientes decisiones para considerarlos susceptibles de análisis.

- Todos los artículos que tuvieran al menos un autor.
- Todos los artículos que tuvieran un año de publicación.

Del mismo modo, se procedió a extraer la información referente a las patentes en PatBase® utilizando las mismas palabras clave, es decir, *Cast Iron*. La consulta arrojó un total de 58002 patentes que fueron sometidas al siguiente proceso de filtrado:

- Se excluyeron las patentes publicadas después del 31 de diciembre de 2021.
- Se excluyeron los grupos principales A, D y H de la Clasificación Internacional de Patentes [WIPO99], es decir, necesidades humanas, textiles y papel y electricidad.

A continuación, se procesaron los datos extraídos, estableciéndose los siguientes requerimientos:

- Todas las patentes han detener al menos un inventor.
- Todas las patentes han de tener un año de publicación.
- La estructura del nombre de los inventores tenía que ser la misma que la de los autores, por lo que fue necesario realizar una adaptación de los nombres de los inventores en el juego de datos de las patentes.

En lo que respecta a ambos juegos de datos, artículos científicos y patentes, los requisitos establecidos para validar los registros se deben a que la información procedente de las bases de datos o bien está incompleta o presenta falta de coherencia en los datos presentados.

Posteriormente, estos datos son procesados para calcular la inversión acumulada realizada por los autores de un artículo en la producción de nuevos conocimientos, representada por el conjunto de variables ANA1 a ANA8, a la tendencia de los inventores a generar propiedad intelectual, variables PENA1 a PENA8 y la generación real de PI, variables PASA1 a PASA8. Los números, del 1 al 8, representan el número de autores o inventores involucrados en el cálculo de cada variable. Este conjunto de variables será utilizado en la generación del modelo.

4.2.2 Método de cálculo.

El método de cálculo aplicado, como en el caso del estampado en caliente detallado en la sección 3.4.2, parte de la consideración de que la publicación de un artículo científico constituye un evento estadístico. A partir de ese momento se puede calcular la inversión acumulada en generación de nuevo conocimiento para cada uno de los autores de forma individual o en colaboración. Se llega a analizar grupos de hasta 8 autores colaborando de forma conjunta. Esto nos permite la obtención del conjunto de variables desde ANA1 a ANA8.

Así mismo, se establece que la publicación de una patente es una propiedad asociada al evento estadístico. Se calcula, por lo tanto, la proclividad de un autor o conjunto de estos a patentar sus investigaciones convirtiéndose, de esta forma, en inventores. Teniendo en cuenta la fecha de prioridad de las patentes calcularemos

las variables PENA1 a PENA8 que indica la inclinación de los inventores a generar propiedad intelectual.

Por último, se calculará el conjunto de variables PASA1 a PASA8 que indicarán la generación real de propiedad intelectual, es decir, patentes publicadas por parte de los inventores.

Tabla 8: Propiedades estadísticas del conjunto de datos resultante del tratamiento de los juegos de datos primarios de artículos científicos y patentes para el caso de fundición de hierro.

Var.	Min	Max	Mediana	Nº ceros	Nº intervalos discretización	Discretización/Frecuencia segmentación				
ANA1	0	179	2	3677	5	1234	2468	3702	4936	6170
ANA2	0	146	0	6555	4	823	1646	2469	3292	-
ANA3	0	190	0	8131	3	572	1144	1716	-	-
ANA4	0	330	0	9127	2	360	720	-	-	-
ANA5	0	462	0	9582	2	133	265	-	-	-
ANA6	0	462	0	9748	2	50	99	-	-	-
ANA7	0	330	0	9815	2	16	32	-	-	-
ANA8	0	165	0	9832	2	85	165	-	-	-
PENA1	0	1350	0	7069	3	926	1852	2778	-	-
PENA2	0	60	0	9485	3	121	241	362	-	-
PENA3	0	12	0	9799	2	24	48	-	-	-
PENA4	0	3	0	9840	2	4	7	-	-	-
PENA5	0	0	0	9847	0	-	-	-	-	-
PENA6	0	0	0	9847	0	-	-	-	-	-
PENA7	0	0	0	9847	0	-	-	-	-	-
PENA8	0	0	0	9847	0	-	-	-	-	-
PASA1	0	1828	0	7210	2	1319	2637	-	-	-
PASA2	0	92	0	9290	2	279	557	-	-	-
PASA3	0	24	0	9752	0	-	-	-	-	-
PASA4	0	11	0	9824	0	-	-	-	-	-
PASA5	0	2	0	9842	0	-	-	-	-	-
PASA6	0	0	0	9847	0	-	-	-	-	-
PASA7	0	0	0	9847	0	-	-	-	-	-
PASA8	0	0	0	9847	0	-	-	-	-	-

El conjunto de variables utilizadas es siempre inicializado a 0, lo cual indica que hay autores o inventores que no tiene publicaciones previas al año de publicación del artículo ni patentes anteriores o posteriores.

Puede verse que el valor máximo depende de cada variable. Los valores van de 0 (valor mínimo) a 1828 (valor máximo) en el caso de PASA1. De la misma forma, la mediana, que mide la tendencia central, es prácticamente 0 para todas las variables excepto para ANA1, cuyo valor es 2. Asimismo, el número de ceros es diferente para cada variable. El valor más bajo es 3677 y corresponde a ANA1 y justifica el valor 2 para la mediana de ANA1. En el extremo opuesto, hay variables para las que todos los valores son 0 y no aportan ninguna información relevante.

Antes de procesar este conjunto de datos se realiza un procedimiento que consiste en establecer el número de intervalos de discretización y configurar los distintos intervalos a fin de generar modelos. Sé procedió al cálculo de (i) la inversión acumulada (variables ANA1 a ANA8), (ii) la proclividad a la generación de propiedad intelectual (variables PENA1 a PENA8) y (iii) la generación real de propiedad intelectual (variables PASA1 a PASA8) con los datos procedentes de WoS y PatBase®. Los resultados obtenidos, tras un proceso de análisis, fueron segmentados en distintos intervalos teniendo en cuenta el número de ceros obtenido para cada variable. No se han descartado las variables PENA3 y PENA4, a pesar del su elevado porcentaje de ceros, ya que produce resultados ligeramente mejores. Posteriormente, se prepararon un conjunto de 35 discretizaciones, a las cuales se ha aplicado un conjunto de modelos, que se detallarán posteriormente junto con los resultados obtenidos.

Los valores de los distintos intervalos se han obtenido por generación de frecuencias, conocimiento experto habiéndose etiquetado todos ellos tal y como se recoge en la **Tabla 9**. Se ha utilizado terminología en inglés, tal y como se ha realizado en artículos científicos.

- La variable ANA1 hace referencia a la producción científica de un investigador en forma de artículos publicados en revistas indexadas. Se han utilizado los términos *No activity* (0-1234), *Starters* (1234-2468), *Low activity* (2468-3702), *Medium activity* (3702-4936) e *Intense activity* (4936-6170), que no es sino una categorización, en cinco intervalos, de estos autores, desde los que no realizan publicaciones hasta aquellos que

realizan una intensa actividad publicadora. Los intervalos se han establecido utilizando frecuencias iguales.

- La variable ANA2 hace referencia a la producción científica por parejas de investigadores que publican juntos. Hablamos en este punto de cooperación entre investigadores habiéndose establecido la clasificación de la siguiente forma, *No cooperation* (0-823), *Low cooperation* (823-1646), *Medium cooperation* (1646-2469) y *Strong cooperation* (2469-3292), que indica el grado de cooperación existente.
- La variable ANA3 indica publicaciones entre 3 autores. Hablamos entonces de redes de investigadores que publican juntos y establecemos la clasificación como *No networking* (0-572), *Low-medium networking* (572-1144) y *Strong networking* (1144-1716), que nos indica el grado de cooperación entre grupos de investigadores.
- Las variables ANA4, …, ANA8 indican publicaciones entre 4, …, 8 autores. Seguimos hablando de redes de investigadores que publican juntos y establecemos la clasificación como *No networking*, *Networking*, que nos indica el grado de cooperación entre grupos de investigadores. Este valor va a depender de los valores de la variable para establecer los intervalos.
    - ANA4 *No networking* (0-360) *Networking* (360-720)
    - ANA5 *No networking* (0-133) *Networking* (133-265)
    - ANA6 *No networking* (0-50) *Networking* (50-99)
    - *ANA7* *No networking* (0-16) *Networking* (16-32)
    - ANA8 *No networking* (0-85) *Networking* (85-165)
- Las variables PENA1 y PENA2 indican la proclividad de los investigadores a patentar su trabajo. PENA1 hace referencia a las patentes individuales, con un único inventor, y PENA2 hace referencia a las patentes realizadas por parejas de investigadores. En ambos casos se ha establecido la clasificación de la siguiente forma, *No patenting culture*, *Low-medium patenting culture* y *Strong patenting culture*, que establece intervalos regulares que enmarcan al investigador por su tendencia a patentar.

  - PENA1
    - *No patenting culture* (0-926)
    - *Low-medium patenting culture* (926-1852)
    - *Strong patenting culture* (1852-2778)
  - PENA2
    - *No patenting culture* (0-121)
    - *Low-medium patenting culture* (121-241)
    - *Strong patenting culture* (241-362)
- La variable PENA3 indica que tres inventores patentan de forma conjunta. El número de intervalos establecidos para esta variable es dos,
  - *No patenting culture* (0-24)
  - *Patenting culture* (24-28)

  que indican si se produce o no la patente, pero no en que grado.

- La variable PENA4 indica que cuatro inventores patentan de forma conjunta. El número de intervalos establecidos para esta variable es también dos,
  - *No patenting culture* (0-4)
  - *Patenting culture* (4-7)

  que indican si se produce o no la patente.

- La variable PASA1, variable objetivo, se establece como una variable binaria, se produce o no se produce la generación de propiedad intelectual. Esto se dispone con dos intervalos cuyos valores son, *No IP production* (0-1319) y *IP production* (1319-1637), que indican si se ha creado nueva propiedad intelectual.

La variable objetivo de este estudio es PASA1, a partir de la cual se obtienen los modelos de predicción de patentes, entre otros, un modelo de red bayesiana y un modelo de árbol de decisión nos ayudarán a entender esta tecnología.

Tabla 9: Valores resultantes de la discretización y etiquetado para los tres grupos de variables del proceso, ANA (inversión acumulada), PENA (inclinación a la generación de propiedad intelectual) y PASA (generación real de propiedad intelectual).

Variables			
ANA1 (5)	**ANA2 (4)**	**ANA3 (3)**	**ANA4 (3)**
• *No activity* • *Starters* • *Low activity* • *Medium activity* • *Intense activity*	• *No cooperation* • *Low cooperation* • *Medium cooperation* • *Strong cooperation*	• *No networking* • *Low-medium networking* • *Strong networking*	• *No networking* • *Networking*
ANA5 (2)	**ANA6 (2)**	**ANA7 (2)**	**ANA8 (2)**
• *No networking* • *Networking*	• *No networking* • *Networking*	• *No networking* • *Networking*	• *No networking* • *Networking*
PENA1 (3)	**PENA2 (3)**	**PENA3 (3)**	**PASA4 (2)**
• *No patenting culture* • *Low-medium patenting culture* • *Strong patenting culture*	• *No patenting culture* • *Low-medium patenting culture* • *Strong patenting culture*	• *No patenting culture* • *Patenting culture*	• *No patenting culture* • *Patenting culture*
PASA1 (2)			
• *No IP production* • *IP production*			

4.2.2.1 Discretización de datos.

En el Anexo 2 se explica con detalle cómo se ha llevado a cabo del proceso de discretización. Se han preparado un conjunto de 35 juegos de datos cuyas variables ANA, PENA y PASA han sido segmentadas en intervalos similares, es decir, por frecuencia. A estos conjuntos de datos se les ha aplicado un conjunto de 24 algoritmos cuyos resultados, correspondientes a las métricas descritas en el punto 1.6.2, se presentan también en dicho Anexo 1. Dichas métricas son (i) el porcentaje de precisión, (ii) el error absoluto medio, MAE, (iii) la raíz del error cuadrado medio, RMSE, (iv) la precisión, (v) la exhaustividad y (vi) el valor-F.

4.2.3 Resultados.

Si analizamos las diferentes tablas del Anexo 2, que contienen los resultados de las métricas utilizadas en esta investigación podemos observar que, por ejemplo, en la **Tabla 19** que muestra el porcentaje de precisión o correctitud, los mejores resultados se obtienen para la discretización D7-5 con el algoritmo C4.5 con un valor de 90,00. Teniendo en cuenta solamente los valores de MAE y RMSE, **Tabla**

20 y **Tabla 21**, la mejor discretización sería las D6-5 con los algoritmos TAN, MLP y MLP CS en el primer caso con un valor de 0,07 y con TAN y *Random Forest* con un valor 0,20 en el segundo. Si miramos las tablas de precisión, exhaustividad y valor-F, **Tabla 22**, **Tabla 23** y **Tabla 24**, lo que obtenemos es que, para la precisión, las discretizaciones D7-1-5 con el valor 0,92 en todos los algoritmos utilizados presentan los mejores resultados. Analizando la exhaustividad, las discretizaciones D2-5 y D3-5 obtienen el mejor valor 1,00 para el algoritmo SVM (RBF) y, asimismo, el mismo valor con el algoritmo SVM (*Normalized Polynomial*) en el caso de la D3-5. Para El valor-F, se alcanza el valor de 0,94 en las discretizaciones D7-1-5 en todos los algoritmos excepto K2, Hill Climber y Naïve Bayes con un valor ligeramente inferior de 0,92.

Para la explicación del modelo, que se presenta a continuación, se ha tomado como referencia la discretización D5-1, es decir, la que presenta un porcentaje de 82,49, entre los más elevados, para poder analizar la influencia de todas las variables. Se han construido sobre ella, por su fácil representación e interpretabilidad, los modelos de red Bayesiana y árbol de decisión que se exponen a continuación.

4.2.3.1 Red Bayesiana.

La red bayesiana[3] para PASA1 como variable objetivo ha alcanzado una precisión del 89,01%.

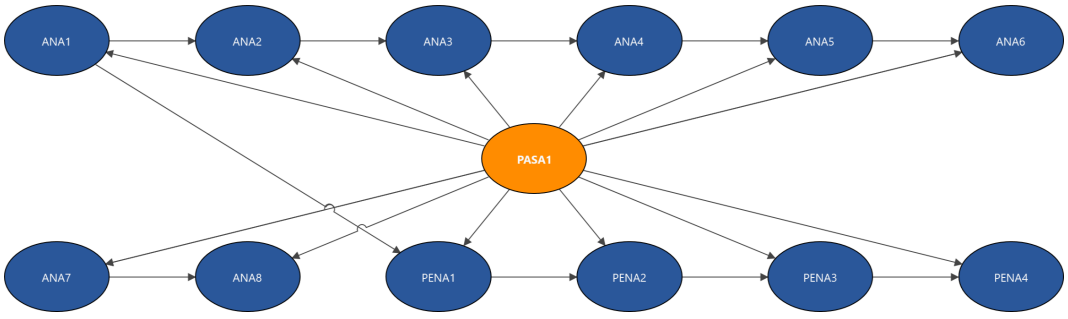

Figura 42: Modelo de red bayesiana obtenido para la variable objetivo PASA1 para la tecnología de hierro fundido. La figura muestra las relaciones entre variables que permiten la generación real de propiedad intelectual para un inventor.

[3] La red Bayesiana se ha calculado con un algoritmo anterior al utilizado en las tablas del Anexo 1 por lo que los valores son ligeramente diferentes.

Los valores de MAE, que muestra un valor de 0,16, y RMSE, con un valor de 0,28.

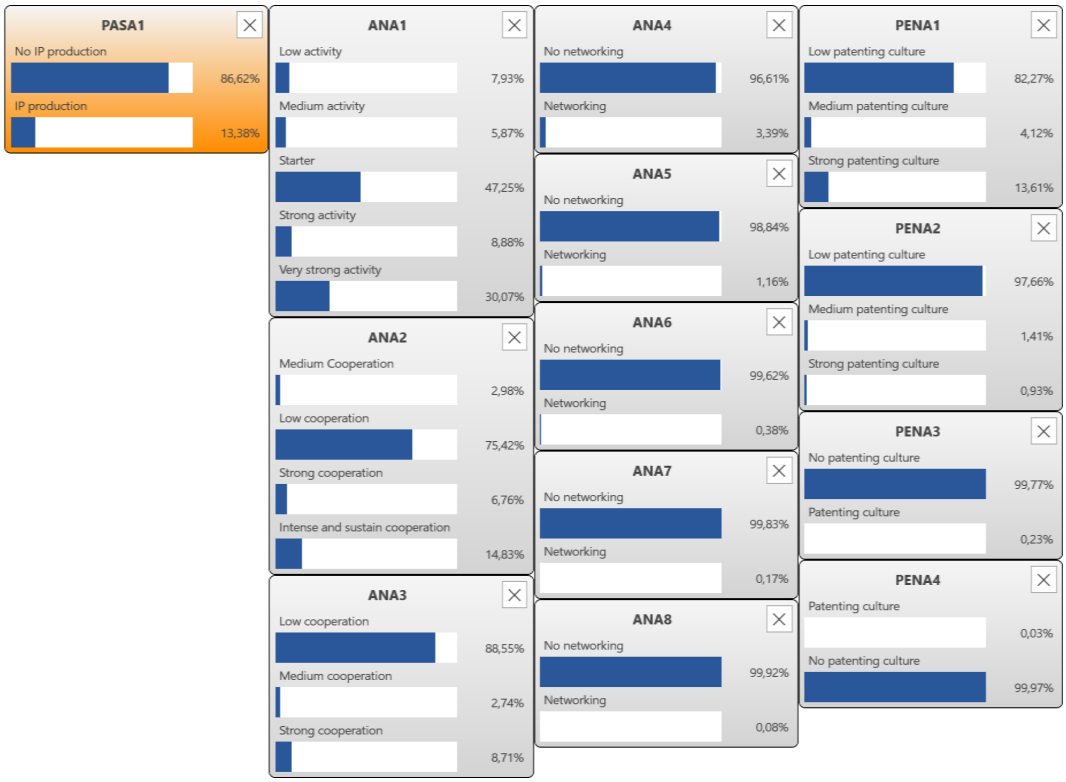

Figura 43: Monitores utilizados para comprobar no sólo el comportamiento de la variable objetivo, en este caso, PASA1, sino también grupos de variables no dependientes (ANA y PENA) en el caso de la tecnología de fundición de hierro.

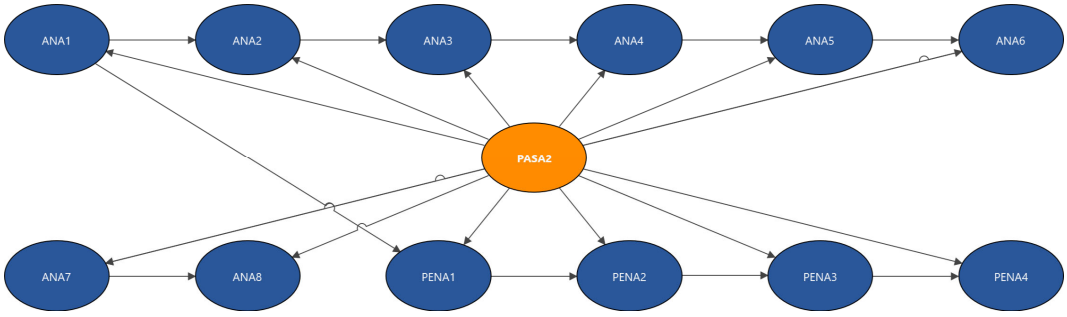

Figura 44: Modelo de red bayesiana para la variable objetivo PASA2 para la tecnología de hierro fundido. La figura revela la combinación de variables que generan propiedad intelectual real considerando pares de inventores.

La **Figura 44** muestra el modelo de red bayesiana para la variable objetivo PASA2 con una precisión del 97,57%. Los valores de MAE y RMSE son 0,04 y 0,15

respectivamente. A pesar de la elevada precisión, como sucedía también cuando se analizó la tecnología de estampación en caliente, el modelo no es representativo. La razón principal es que el número de ceros es muy elevado para esta variable, 9290 de 9847 valores, es decir, un 94,34%.

4.2.3.2 Árbol de decisión.

En la **Figura *45*** se muestra el modelo de árbol de decisión que se obtiene cuando se establece como variable objetivo la variable PASA1 que tendencia a patentar de forma individual.

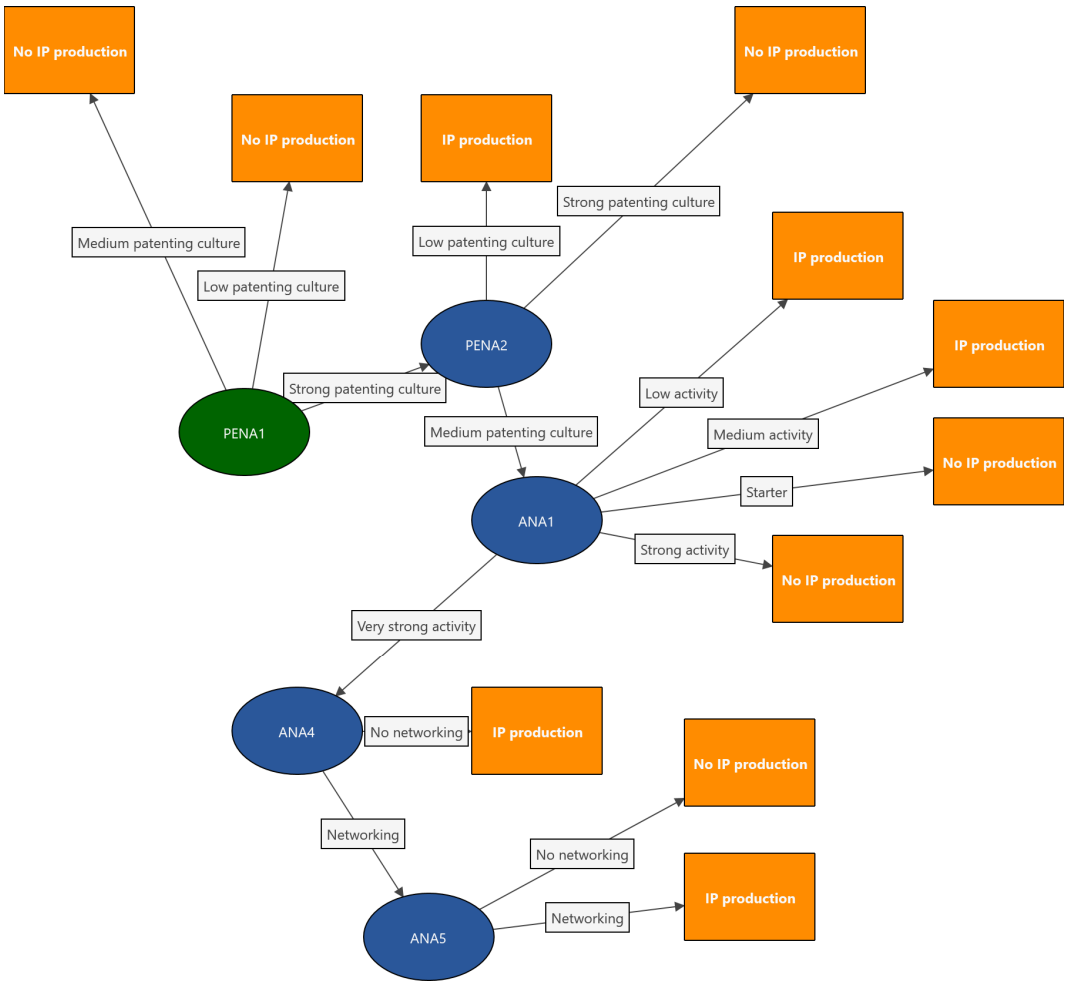

Figura 45: Modelo de árbol de decisión para la variable objetivo PASA1. La figura muestra las condiciones en las que se genera la propiedad intelectual real.

Como resultado, se han se han obtenido los siguientes valores para las métricas definidas en la sección 1.6.2 Métricas utilizadas., que se corresponden con el porcentaje de precisión o correctitud, y las tasas de error, siendo estas, el error medio absoluto y la raíz del error cuadrado medio.

- **Correctitud**: 89,13% frente al 89,01% al modelo de red Bayesiana.
- **MAE**: 0,16 en ambos modelos.
- **RMSE**: 0,28 en ambos modelos.

Cuando se intenta generar un modelo de árbol de decisión para la variable objetivo PASA2, no se obtiene un modelo. Si unimos esto a que la red Bayesiana para esta variable tampoco sea representativa, implica que no podemos obtener información cuando seleccionamos la variable PASA2 como objetivo.

Analizando el árbol de decisión[4], observamos una serie de diferentes escenarios que representaremos y analizaremos en la **Tabla 10**. Se ha estimado, también, el porcentaje de producción de propiedad intelectual correspondiente a cada uno de los escenarios en los que se produce generación real de propiedad intelectual. El resto de las situaciones se descarta por ser irrelevante en términos de futura producción de propiedad intelectual.

Tabla 10: Escenarios con producción real de propiedad intelectual para la tecnología de fundición de hierro considerando como variable objetivo la variable PASA1, es decir, la generación de patentes.

ID	PENA1	PENA2	ANA1	ANA4	ANA5	% IP Production
1	Strong patenting culture	Medium patenting culture	Very stong activity	Networking	Networking	87,66
2	Strong patenting culture	Low patenting culture				60,42
3	Strong patenting culture	Medium patenting culture	Medium activity			53,62
4	Strong patenting culture	Medium patenting culture	Very stong activity	No networking		51,00
5	Strong patenting culture	Medium patenting culture	Low activity			48,54

[4] El árbol de decisión se ha calculado con un algoritmo anterior al utilizado en las tablas del Anexo 2 por lo que los valores son ligeramente diferentes.

El 5º escenario, fila 1 de la **Tabla 10** con valor ID igual a 5, se caracteriza porque no tiene una componente que haga referencia a la publicación de artículos científicos en revistas indexadas. Se produce cuando los inventores tienen una fuerte cultura para publicar sus patentes, con sus inventos, de forma individual y solo ocasionalmente desarrollan esta actividad en parejas. Para este 5º escenario se alcanza una producción de propiedad intelectual del 60,42%.

Los escenarios 1, 3 y 4 son similares en cuanto a la componente que hace referencia a la proclividad a patentar por parte de los inventores. Se caracterizan por una fuerte cultura para patentar de forma individual y una cultura media para hacerlo en parejas. En cuanto a la componente relativa a la publicación de artículos científicos vemos que cubre todo el rango. De todos ellos, el escenario 1 presenta una baja actividad relacionada con la publicación de artículos científicos alcanzándose un 48,54% de generación de PI, el escenario 3 que muestra una actividad media en cuanto a la producción de literatura científica llega hasta un porcentaje del 53,62% de producción de PI. Finalmente, el escenario 4 que manifiesta una muy fuerte actividad concerniente a la publicación de artículos, pero no presenta *networking* relativo a la variable ANA4 cae hasta un 51% en cuanto a generación de PI.

Por otro lado, el escenario 1 se caracteriza por una alta cultura para patentar de forma individual y una cultura media para hacerlo por parejas. Así mismo presenta una muy fuerte actividad relativa a la publicación de artículos científicos junto con *networking* en las variables ANA4 y ANA5. Esta componente de *networking* hace que la generación de propiedad intelectual alcance el máximo valor de todos los escenarios con un 87,66%.

4.2.3.3 Evolución en el tiempo.

Los resultados obtenidos cuando realizamos un análisis temporal para la tecnología de hierro fundido teniendo en cuenta los 2 años de vigencia de las patentes se presentan en la siguiente gráfica.

Si observamos la gráfica de la **Figura 46** vemos una evolución similar entre ambas modelos, es decir, a lo largo de los años la precisión de los estos va decreciendo, aunque podemos apreciar, no obstante, una ligera diferencia.

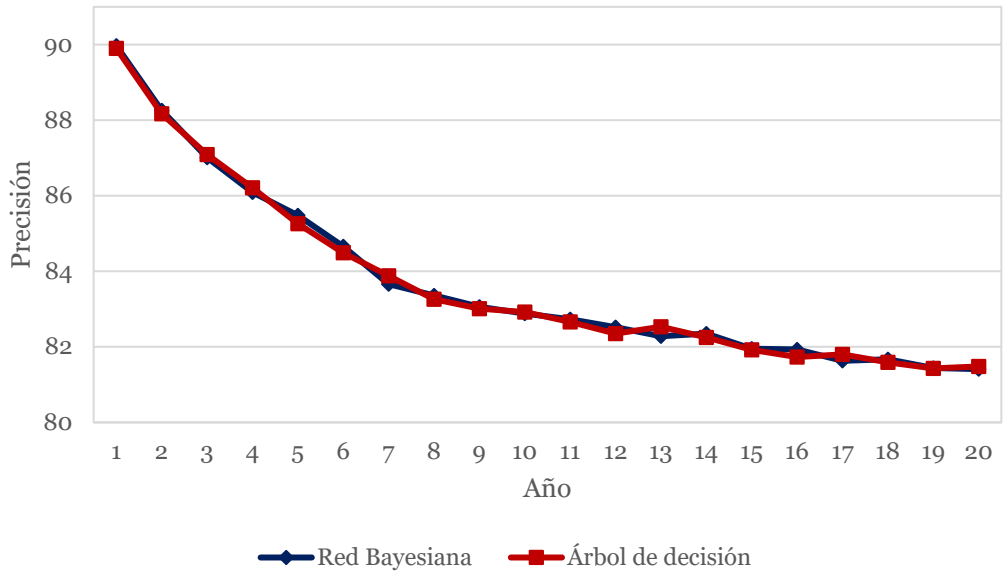

Figura 46: Evolución de la precisión para los modelos de red Bayesiana y árbol de decisión para la tecnología de fundición de hierro durante los 20 años de vigencia de las patentes.

Si consideramos por un lado que la evolución de las tecnologías analizadas en este trabajo de investigación, tanto estampación en caliente como fundición de hierro, presentan una rápida evolución y por otro lado que los planes estratégicos en las empresas se elaboran para un periodo que comprende una media de 4 años, estaríamos hablando de un mínimo de un 85% de precisión.

4.2.4 Discusión y conclusiones.

En esta investigación, se ha puesto el foco en la validación de la metodología establecida para la tecnología de estampación en caliente y confirmar que es factible su aplicación en otras tecnologías como puede ser, por ejemplo, la de fundición de hierro.

Se han utilizado las mismas técnicas que en el caso de la tecnología de estampación en caliente, tanto estadísticas como de inteligencia artificial. Las fuentes de datos utilizadas han sido una vez más WoS y PatBase®, empleándose también los nombres del autor del artículo o inventor de la patente, y el año de publicación de estos para desarrollar los modelos.

En cuanto al tema de la heterogeneidad de los nombres de autores e inventores, al proceder los datos de las mismas fuentes, el problema vuelve a presentarse. No se trata solamente en la forma en que aparecen escritos, es decir el orden de nombre, primer apellido y segundo apellido, sino también a que incorporan, como ya se explicó, otros datos como el nombre de la empresa a la que pertenecen, o su titulación académica, dentro de lo que sería el valor del campo nombre.

En lo esencial, se procedió tal y como se explica en el punto 3.4.2 calculando la inversión acumulada, la proclividad a la generación de PI y la generación real de PI a partir de los datos adquiridos de WoS y PatBase®. Se ha vuelto a generas un nuevo conjunto de datos relativos a la tecnología de fundición de hierro que contiene las variables ANA1 a ANA8, PENA1 a PENA8 y PASA1 a PASA8 ya explicadas. En relación con las técnicas de aprendizaje automático aplicadas se ha creado una red bayesiana y un árbol de decisión que nos han permitido, como en el caso anterior extraer ciertas reglas. La precisión de estos modelos se mantiene entre el 89 y el 90%.

A diferencia de lo que sucedía cuando analizábamos la tecnología de estampación en caliente, el número de escenarios resultantes que producen la generación de PI utilizando PASA1 como variable objetivo, es mucho más reducido. Estamos hablando de un conjunto de 5 escenarios para la tecnología de hierro fundido frente a los 15 que se producían cuando analizamos el caso de estampación en caliente. Es importante, también, tener en cuenta que en el primer estudio realizado conseguíamos 2 escenarios adicionales analizando modelos que tenían como objetivo la variable PASA2, que no se da en el análisis del hierro fundido.

Los principales escenarios utilizando PASA1 como variable objetivo indican que una cultura de patentes es esencial para continuar generando PI. Además, cuando este factor se combina con una alta producción de literatura científica y añadiendo *networking* en las variables ANA4 y ANA5 la generación de PI aumenta considerablemente. No se han podido generar modelos que aporten más información utilizando PASA2 como variable objetivo.

Estas diferencias en cuanto a los escenarios que se producen mediante esta técnica pueden explicarse en base a la madurez de la tecnología. Mientras que en el caso

del hierro fundido tenemos una tecnología madura, para el caso de estampación en caliente esta es relativamente joven.

Como en el caso anterior, se han obtenido algunos modelos, pero aún hay margen de mejora como, por ejemplo, introducir más metadatos en los análisis como la entidad a la que pertenece el autor y/o inventor, la combinación autor-institución e inventor-institución entre otras posibles u obtener conjuntos de datos más específicos.

Se podría también aumentar la precisión de los modelos de aprendizaje automático probando otro tipo de algoritmos más potentes que podrían conseguir resultados diferentes. Finalmente, podríamos considerar otras fuentes de datos adicionales como conferencias, patentes, informes técnicos o proyectos de I+D.

4.3 Sumario.

En este capítulo se valida la metodología propuesta para la predicción de la innovación basada en la utilización conjunta de métodos estadísticos y aprendizaje automático, que se desarrolló en el capítulo 3. Se han utilizado, como en el caso anterior, datos relativos a publicaciones científicas, concretamente artículos publicados en revistas indexadas, y datos referentes a patentes para la generación de modelos. Estos datos se han preprocesado obteniéndose un conjunto de datos que pueda ser procesado mediante técnicas estadísticas, cuya salida será la entrada al proceso de aprendizaje automático. Se generan varios modelos que alcanzan precisiones en el rango del 89% al 90%, que han dado lugar a una serie de patrones de comportamiento que establecen cuando va a producirse la generación de propiedad intelectual. Se ha analizado, por último, la evolución en el tiempo de estas predicciones ante un posible futuro en constante cambio, estimándose su validez en un periodo alrededor de 4 años, la misma duración que los planes estratégicos.

En conclusión, se valida nuevamente la hipótesis de esta tesis doctoral, en este caso para la tecnología de fundición de hierro obteniéndose un nuevo modelo que difiere del anterior al ser tecnologías con distinto grado de madurez.

5
Conclusiones

Una vez llegados a este punto, en el cual se ha descrito y detallado todo el proceso de investigación realizado, es momento de mirar hacia atrás, ver cuál ha sido el camino recorrido y valorar los resultados obtenidos. Para ello, en este capítulo, se resumirán las contribuciones de esta tesis doctoral por comparación con los objetivos marcados en el inicio de la misma, así como con la hipótesis de partida.

Con respecto a la innovación, se ha constatado la importancia que ha tenido históricamente y aún hoy en día sigue teniendo. Asimismo, se ha hecho constar el valor de la realización de las predicciones, en un entorno empresarial altamente competitivo, que brindan una ventaja estratégica a las organizaciones.

En consecuencia, tras un estudio de las técnicas utilizadas en este ámbito, se ha planteado la creación de un modelo de predicción de la innovación. Se ha utilizado, por un lado, información relativa a las investigaciones publicadas en revistas científicas indexadas y, por otro lado, información referente a las patentes que han publicado los mismos inventores que han creado la mencionada literatura científica.

El presente capítulo se encuentra organizado de la siguiente forma. Comenzaremos con la sección 5.1 donde se presenta un resumen del proceso llevado a cabo. Se incluye, no solo la metodología aplicada, sino también los resultados alcanzados en las diferentes etapas del proceso. A continuación, en la sección 5.2 se enumeran aquellas áreas de aplicación para las cuales esta investigación puede ser relevante. Posteriormente, en la sección 5.3 se exponen las limitaciones que tiene la solución propuesta. Por último, en la sección 5.4 se recogen las futuras líneas de investigación que han surgido tras la finalización de esta tesis doctoral.

5.1 Resumen y resultados de la investigación.

Considerando el marco temporal, este trabajo de investigación comenzó con un análisis bibliométrico que se ha llevado a cabo sobre la tecnología de estampación en caliente. Se analizó en este primer abordaje el periodo comprendido entre los años 2009 y 2019, es decir, una década. Los datos de partida para este estudio se han obtenido de la base de datos Scopus, que, tras haber sido analizados, han presentado gran cantidad de información relevante.

Podemos incluir, en el conjunto de resultados obtenidos, los siguientes indicadores.

1. **Número de publicaciones científicas en revistas indexadas por año**. El número de publicaciones científicas ha tenido un crecimiento exponencial durante estos años, **Figura *16***, pasando de 16 publicaciones en 2009 a 139 en 2019, es decir, un incremento del 869%. Esto indica que se trata de una tecnología en pleno auge durante ese periodo.
2. **Número de artículos científicos publicados en las diez revistas indexadas más influyentes**. Destaca entre todas las revistas *Journal of Materials Processing Technology* con 60 artículos publicados en dicho periodo, que han sido citados un total de 3095, veces lo que la convierte así mismo en una revista de referencia, **Figura *17***. En este mismo contexto, la revista *CIRP Annals – Manufacturing technology* con solo 17 artículos, pero con 1359 citaciones confirma su importancia.
3. **Evolución del número de artículos científicos publicados por año en las cinco revistas indexadas más influyentes**. Destaca, sobre

el conjunto, **Figura *18***, la evolución de la revista *International journal of advanced manufacturing technology*, una de las últimas en incorporarse en este sector pero que está adquiriendo una gran relevancia.

4. **Número de artículos científicos publicados por los diez autores más relevantes**. El investigador que destaca sobre el resto es, sin duda, J. Lin con 31 publicaciones, **Figura *19***. Investigadores más relevantes, como Merklein, M., Bruschi, S. y Ghiotti, A. son también los autores de las revisiones más relevantes en este campo, lo que confirma que, en la tecnología de estampación en caliente, los autores con mayor producción literaria son también los referentes científicos.
5. **Número de artículos científicos publicados por las diez instituciones/empresas más relevantes**. Las instituciones más relevantes son en su mayoría universidades, entre las que se encuentran las universidades chinas, **Figura *20***. Llama especialmente la atención que la Universidad de Lulea, la ciudad donde nació el proceso de estampación en caliente haya mantenido sus vínculos con esta tecnología durante más de 40 años. Se constata también la presencia de Arcelor Mittal, un agente industrial, entre las instituciones o entidades que más publican. Esto subraya el carácter aplicado que posee la investigación sobre la estampación en caliente.
6. **Número de citaciones por año**. El número de citaciones por año ha tenido un incremento sustancial en este periodo pasando de 130 citaciones en 2009 a 2063 en 2019, **Figura *21***. Este hecho ratifica el interés creciente mostrado por la tecnología de estampación en caliente.
7. **Valor del índice h**. El valor del índice h se ha establecido en 53, **Figura *22***, lo que indica que 53 publicaciones han sido citadas al menos 53 veces. Comparando este valor con otra tecnología de referencia como es la estampación en frío, cuyo índice h es de 30 en el mismo periodo, se denota un gran interés, por parte de la comunidad científica, para esta tecnología.
8. **Número de citaciones por autores**. H. Karbasian y A.E. Tekkaya son los autores más citados con su revisión en el año 2010, seguidos de M. Merklein y J. Lechler con una segunda revisión publicada en 2016. Así

mismo, otros autores relevantes en este campo, como A. Ghiotti y S. Bruschi, también figuran entre los autores de mayor influencia, **Figura *23***.

9. **Número de citaciones por revista**. *Journal of Materials Processing Technology* y *CIRP Annals - Manufacturing Technology*, **Figura *24***, se constituyen como las revistas más influyentes en el aspecto referente a la difusión científica de la tecnología de estampación en caliente.
10. **Número de publicaciones por país**. La mayoría de las publicaciones proceden de China con 249 publicaciones, **Figura *25***, seguida de lejos por Alemania con 105 y Corea del Sur con 78. Estados Unidos, junto con Canadá, desarrollan también una importante actividad investigadora. Estos datos concuerdan con los datos anteriores que indican que J. Lin es el autor con más publicaciones y las instituciones chinas se encuentran entre las más productivas.
11. **Red de palabras clave**. Podemos observar un total de 207 artículos agrupados en 6 clústeres diferentes, **Figura *26***. Las palabras clave más relevantes se encuentran en el clúster verde con términos como estampación en caliente, estampación y máquinas de forja entre otras. Presenta, también, gran relevancia el clúster rojo con términos como microestructura, propiedades mecánicas, acero de alta resistencia o temple.
12. **Red de autores**. Encontramos un total de 316 autores, **Figura *27***, agrupados en 23 clústeres que especifican los autores que colaboran más estrechamente. J. Lin, el investigador con más publicaciones en esta década, aparece en el clúster rosa. S. Brushi, el segundo investigador relevante aparece en el clúster amarillo.

Teniendo en cuenta los resultados de los indicadores utilizados en este análisis bibliométrico se puede concluir que:

- La estampación en caliente se ha convertido en una tecnología que evoluciona de forma rápida, con una producción científica creciente y un alto nivel de impacto medido en términos de índice h.
- La mayor parte de la producción científica se concentra en un conjunto establecido de fuentes en lo que se refiere tanto a la producción como a las citaciones. Las revistas más prolíficas tienen también el mayor número de

citaciones, pero en cuanto al número de citaciones por número de documentos totales publicados, *CIRP Annals - Manufacturing Technology* destaca como fuente de alta relevancia.

- Los autores más relevantes son también referentes en las principales revistas del sector.
- Es reseñable que la investigación industrial en esta área aparece entre las diez primeras instituciones en términos de producción científica. Esto constata la importancia de la industria como motor importante en la generación de conocimiento.
- Geográficamente, la producción científica se concentra en China, con otros dos focos en Europa (liderada por Alemania) y América del norte (liderada por EEUU).
- El gráfico de red de palabras clave muestra que existen seis grupos temáticos diferentes relativos a la investigación de la estampación en caliente, aquellos relacionados con los parámetros del proceso, las propiedades de los materiales, las tecnologías de ensamblaje, el propio proceso de temple, las alternativas a la estampación en caliente del acero y la refrigeración de las matrices.
- En cuanto a las redes de investigadores, se identifican varios grupos de estables, con una gran concentración en torno a los investigadores chinos y otros grupos a su alrededor en países como Suecia, Italia, Alemania, Reino Unido y Canadá.

Sin duda alguna, se puede afirmar que estamos ante una tecnología prometedora y en constante desarrollo y con una actividad muy intensa, tanto industrial como académica.

En conclusión, vemos que esta primera aproximación nos permite obtener información valiosa, pero no nos permite predecir tendencias de una forma clara u sencilla, no siendo suficiente para abordar el planteamiento realizado al plantear esta investigación.

Posteriormente, el enfoque del primer análisis bibliométrico basado en la tecnología de estampación en caliente fue ampliado mediante la utilización de mapas evolutivos. Con la introducción de este nuevo concepto se permite la

realización de un análisis diferente de la esta tecnología, para detectar y visualizar los temas o áreas conceptuales que engloba, así como su evolución en el tiempo.

En esta ocasión, por similitud con abordajes similares en este tipo de estudios se ha utilizado la base de datos de literatura científica WoS, concretamente la *Web of Science Core collection*. En periodo de tiempo para el cual se ha realizado la búsqueda abarca desde sus inicios hasta el año 2019. El desarrollo de este proceso se basa en la técnica de co-word, que se centra en la co-ocurrencia de las palabras clave de los documentos científicos. Dicha técnica es utilizada para identificar patrones y temas emergentes a partir de la literatura científica, explorando las relaciones y conexiones entre diferentes áreas sujetas a investigación. Se utilizan el título, las palabras clave y el resumen de los artículos seleccionados para desarrollar este análisis bibliométrico.

Como resultado, se definieron tres periodos claramente diferenciados.

1. **De 1950 hasta 2009**. Corresponde con los inicios de la tecnología de estampación en caliente y se caracteriza por su baja producción científica.
2. **De 2010 a 2015**. Comienza con la revisión de Karbasian, H. sobre esta tecnología. Comprende el 36.8% de los artículos.
3. **De 2016 a 2019**. Comienza con la segunda revisión realizada por Merklein, M. en 2016. Incluye el 57.4% del total de artículos.

Una vez parametrizado y realizado el análisis de co-word, se detectaron 46 diferentes temas que se agruparon en 8 familias tecnológicas diferentes: (i) metalurgia del acero, (ii) propiedades de los materiales, (iii) tecnología de recubrimiento, (iv) modelización y simulación, (v) diseño de piezas, (vi) tecnología de fabricación, (vii) aleaciones de aluminio y (viii) tecnología de moldes, lo que ha permitido realizar un análisis de la evolución de la estampación en caliente.

En estas circunstancias, se han determinado los siguientes resultados para cada uno de los periodos.

1. **De 1950 hasta 2009**. **Figura *30***.
   - El grupo de metalurgia del acero refleja la exploración de alternativas a los aceros al boro en los primeros años.
   - El grupo de propiedades materiales se constituye como un campo de investigación muy activo, con temas motores emergentes y en declive.
   - El grupo de tecnología de recubrimiento es un elemento aislado, sin industrializar, pero se convierte en un tema motor debido a la patente de uso del material recubierto de AlSi de ArcelorMittal.
   - El grupo de modelización y simulación se encuentra en la zona de muy baja densidad, ya que no se disponía de ningún software comercial que pudiera realizar una simulación termo-mecánica-microestructural.
   - El grupo de diseño de piezas es un grupo emergente impulsado por la demanda del mercado, diseño de piezas monolíticas (no adaptadas).
   - El grupo de tecnología de fabricación. es un tema básico/transversal que coincide con el hecho de que la tecnología aún no estaba establecida como campo de investigación. Pasará a convertirse en un tema motor.
2. **De 2010 a 2015**. **Figura *31***.
   - La posición del grupo metalúrgico del acero se ajusta a la consolidación industrial del 22MnB5 como el principal grado de acero para la estampación en caliente.
   - El grupo de propiedades de los materiales se sitúa en los cuatro cuadrantes del diagrama debido a su gran actividad.
   - El grupo de tecnología de recubrimiento se ha convertido en un tema motor de acuerdo con las demandas reales recibidas de la industria.
   - El grupo de modelado y simulación muestra un cambio que coincide con el desarrollo de paquetes de software para resolver escenarios de estampación en caliente.
   - El grupo de diseño de piezas se interpreta como una producción baja pero constante en nuevas aplicaciones tecnológicas. El creciente número de piezas, BIW, estampadas en caliente apoya esta interpretación.
   - Las aleaciones de aluminio constituyen un nuevo grupo emergente que coincide con la popularidad de los desarrollos de BIW de aluminio de

Jaguar-Land Rover y los primeros estudios de la industria aeroespacial sobre la estampación en caliente de aluminio 2024.

- La tecnología de las matrices es un nuevo elemento debido a un creciente interés por la optimización de los tiempos de ciclo, la productividad de las matrices y la geometría de las piezas.
- El grupo de tecnologías de fabricación se extiende durante este periodo en los diferentes cuadrantes. Los términos "soldadura láser" y "optimización" son temas muy desarrollados, justificándose por el aumento del número de líneas de producción de calidad para la automoción y de instalaciones de corte por láser para la estampación en caliente.

3. **De 2016 a 2019**. **Figura *32***.
   - El grupo de metalurgia del acero se adentra en la posición de los temas en declive. Vuelven a aparecer temas como las microaleaciones, pero esto sólo refleja que se han hecho pocos esfuerzos de publicación, no implica que la actividad haya sido escasa. Han aparecido en el mercado productos de ArcelorMittal y de Kobe Steel que indican que las innovaciones patentadas se mantuvieron en secreto hasta su completo desarrollo.
   - El grupo de propiedades de los materiales gana relevancia en este periodo posicionándose como tema motor.
   - El grupo de tecnología de recubrimientos mantiene su papel de motor en este campo, debido a los revestimientos de AlSi y a la investigación en los revestimientos de Zn y ZnNi como posibles sustitutos.
   - El grupo de modelización y simulación se mantiene como tema motor, coincidiendo con el uso de software comercial para la estampación en caliente y el desarrollo de modelos de daños para la evaluación de fallos.
   - El grupo de diseño de piezas se sitúa en el cuadrante inferior izquierdo, que se explica por la actividad permanente en el desarrollo de nuevos componentes y aplicaciones.
   - La tecnología de matrices podría ser un tema motor potencial para el futuro próximo, en consonancia con el creciente interés por la

tecnología de fabricación aditiva y su potencial aplicación a la refrigeración en matrices de estampación en caliente.

- El grupo de tecnología de fabricación es un gran representante de los grupos motores "estampación en caliente", rodeado de grupos de densidad media, lo que implica su reconocimiento como una disciplina en sí misma con tecnologías satélite relativas a proceso.

En base a los 3 periodos establecidos se ha creado un mapa de evolución conceptual, **Figura *33***, que nos permite mostrar la evolución de los conceptos clave. Se han obtenido las siguientes conclusiones.

- La «estampación en caliente» está presente en los tres periodos adquiriendo relevancia como estado del arte.
- Existe una intensa relación en la «estampación en caliente» como tema clave a lo largo de los tres periodos, consolidándose como una tecnología con buenas perspectivas de futuro.
- Las relaciones entre el «grupo de propiedades de los materiales» y el «grupo de tecnologías de fabricación» son un caso similar, que muestra la relevancia de los diferentes tipos de aceros y aleaciones en el proceso industrial.
- El «Grupo de modelización y simulación» también muestra la evolución en términos de capacidad de cálculo y se puede observar cómo este factor ha afectado al proceso de optimización.

Para dar por concluido este segundo abordaje bibliométrico, hay que constatar que, la información extraída es relevante y nos muestra no solo la evolución de esta tecnología a lo largo del tiempo, sino también indicios del camino que va a tomar el futuro. No obstante, aunque puede considerarse una primera aproximación hacia la predicción de la innovación no es suficiente para lograr los objetivos de esta tesis doctoral.

Para el siguiente paso, se ha optado por un enfoque diferente. Se ha planteado una metodología que combina, por un lado, métodos estadísticos y, por otro lado,

técnicas o modelos de aprendizaje automático que se han aplicado a un conjunto de datos procedentes de WoS y de PatBase®, es decir, artículos científicos publicados en revistas indexadas y patentes respectivamente.

Se ha partido de la premisa de que la aparición de artículo constituye un evento estadístico. A partir de dicho evento se ha calculado la inversión acumulada para la generación de nuevo conocimiento mediante la investigación y se ha obtenido un conjunto de 8 variables ANA, de 1 a 8, cuyo valor contiene el sumatorio de todas las veces que un autor aparece en los artículos anteriores al año del evento estadístico para la variable 1. El resto de las variables se han calculado mediante combinaciones no repetitivas del número de autores que participan, es decir, de 2 a 8.

Se ha considerado también la publicación de una patente como una propiedad asociada a un evento estadístico. De forma similar al cálculo de la variable ANA, se han calculado las variables PENA y PASA que representan, respectivamente, la proclividad de los inventores a patentar sus descubrimientos y la generación real de propiedad intelectual en forma de patente. Se ha tenido en cuenta en este cálculo la fecha de prioridad de las patentes.

Una vez que los datos fueron procesados y se dispuso del conjunto de datos relativos a las variables ANA, PENA y PASA, se preparó un conjunto de 35 diferentes discretizaciones a las que se aplicó el siguiente grupo de algoritmos de aprendizaje automático.

1. Red bayesiana (K2)
2. Red bayesiana (Hill Climber)
3. Red bayesiana (TAN)
4. Naïve Bayes
5. ANN (MLP)
6. ANN (MLP CS)
7. SVM (Polynomial)
8. SVM (Normalized Polynomial)
9. SVM (Pearson VII)
10. SVM (RBF)
11. KNN (K=1)
12. KNN (K=2)
13. KNN (K=3)
14. KNN (K=4)
15. KNN (K=5)
16. KNN (K=6)
17. C4.5
18. Random Forest (i=50)
19. Random Forest (i=100)
20. Random Forest (i=150)
21. Random Forest (i=200)
22. Random Forest (i=250)
23. Random Forest (i=300)
24. Random Forest (i=350)

En términos de correctitud, los mejores resultados se obtienen para las discretizaciones D7-3 t D7-5 con el algoritmo C4.5 con un valor de 82,50. Teniendo en cuenta solamente los valores de MAE y RMSE, la mejor discretización sería las D6-4 con los algoritmos ANN en el primer caso con un valor de 0,09 y con TAN, KNN, C4.5 y *Random Forest* con un valor 0,23 en el segundo. Si miramos las tablas de precisión, exhaustividad y valor-F, lo que obtenemos es que, para la precisión, la discretización D1-1-5 con el valor 0,84 y con los algoritmos SVM (*Polynomial*), SVM (*Normalized Polynomial*) y SVM (RFB) es la que presenta mejores resultados. Analizando la exhaustividad, la discretización D3-1-5 obtiene el mejor valor 0.99 para el algoritmo SVM (RBF) y, asimismo, la discretización D6-4 obtiene el mismo valor con los algoritmos SVM (*Normalized Polynomial*) y SVM (RBF). Para El valor-F, se alcanza el valor de 0,89 en las discretizaciones D4-4 y D6-4. La primera de ellas lo obtiene para los algoritmos ANN (MLP), ANN (MLP CS), SVM (*Normalized Polynomial*), SVM (Pearson VII), KNN, C4.5 y *Random Forest*, mientras que para la segunda se obtiene dicho valor con los algoritmos ANN (MLP), ANN (MLP CS) y *Random Forest*.

No obstante, para lograr una mejor compresión de los resultados se han analizado los modelos de red Bayesiana y árbol de decisión que han presentado los siguientes valores en cuanto a las métricas utilizadas cuando la variable objetivo fue PASA1.

- **Correctitud**: 85,26% en el modelo de árbol de decisión frente al 84,97% del modelo de red Bayesiana.
- **MAE**: 0,22 en el modelo de árbol de decisión frente al 0,21, en el modelo de red Bayesiana.
- **RMSE**: 0,33 en ambos modelos.

La interpretación de estos modelos ha permitido el establecimiento de una serie de escenarios en los que se produce la generación de propiedad intelectual en forma de patente. Los porcentajes de producción intelectual obtenidos varían entre un 34,97% y un 70,81% en función de la proclividad de los inventores a patentar y de la cooperación en lo que respecta a la publicación de artículos científicos. La variedad de escenarios se explica con más detalle en la **Tabla *6***.

Además, cuando se ha cambiado la variable objetivo a PASA2 se ha obtenido un porcentaje de correctitud del 98,55% y del 98,52% para la red Bayesiana y el árbol de decisión respectivamente y un valor de MAE de 0,03 y de RMSE de 0,12 en ambos casos. Aunque estos modelos no son muy representativos debido al elevado número de ceros en la variable, si se pudieron deducir varias reglas. Se han alcanzado porcentajes de producción intelectual ente el 76,30% y el 81,71% en función de, únicamente, la proclividad a patentar al cual se añade la colaboración para la publicación de artículos científicos.

Se ha establecido también la evolución de la precisión en el tiempo de ambos modelos que, para un plazo aproximado de 4 años, duración estimada de un plan estratégico, que tiene en cuenta, además, que se está hablando de una tecnología en continua evolución está en el entorno del 86,5%.

Finalmente, y viendo los resultados obtenidos en el análisis de la tecnología de estampación en caliente, se ha decidido aplicar la misma metodología a la tecnología de hierro fundido a fin de comprobar la transversalidad de la misma. El proceso se ha desarrollado de forma similar y ha comenzado con la adquisición de los datos que nos permiten realizar el cálculo de las variables ANA, PENA y PASA. Una vez obtenidas se volvió a preparar un conjunto similar de 35 discretizaciones a las cuales se aplicaron los mismos algoritmos de aprendizaje automático.

En términos de precisión o correctitud, los mejores resultados se obtienen para la discretización D7-5 con el algoritmo C4.5 con un valor de 90,00. Teniendo en cuenta solamente los valores de MAE y RMSE, la mejor discretización sería las D6-5 con los algoritmos TAN, MLP y MLP CS en el primer caso con un valor de 0,07 y con TAN y *Random Forest* con un valor 0,20 en el segundo. Si miramos las tablas de precisión, exhaustividad y valor-F, lo que obtenemos es que, para la precisión, las discretizaciones D7-1-5 con el valor 0,92 en todos los algoritmos utilizados presentan los mejores resultados. Analizando la exhaustividad, las discretizaciones D2-5 y D3-5 obtienen el mejor valor 1,00 para el algoritmo SVM (RBF) y, asimismo, el mismo valor con el algoritmo SVM (*Normalized Polynomial*) en el caso de la D3-5. Para El valor-F, se alcanza el valor de 0,94 en las discretizaciones D7-1-5 en

todos los algoritmos excepto K2, Hill Climber y Naïve Bayes con un valor ligeramente inferior de 0,92.

Del mismo modo, para tener una mejor legibilidad de los resultados, se prepararon y analizaron los modelos de red Bayesiana y árbol de decisión. Los valores obtenidos y que se corresponden con las métricas definidas, siendo la variable objetivo PASA1, son las siguientes.

- **Correctitud**: 89,13% en el modelo de árbol de decisión frente al 89,01% del modelo de red Bayesiana.
- **MAE**: 0,16 en ambos modelos.
- **RMSE**: 0,28 en ambos modelos.

En igual forma, se ha generado un conjunto de escenarios cuyos porcentajes de generación de propiedad intelectual varían entre un 51,00% y un 87,66%, más influenciados en este caso por la proclividad a patentar que por la publicación de artículos de forma individual o en colaboración, aunque este hecho también hace que el porcentaje aumente.

No ha sido posible, en este caso, la obtención de resultados para la variable objetivo PASA2.

En el mismo orden de ideas, la evolución de la precisión en el tiempo en un marco de referencia de 4 años nos sitúa en torno al 86%.

La aplicación de la misma metodología a dos tecnologías, una madura, fundición de hierro, y otra relativamente joven, estampación en caliente, indica que dicha metodología puede ser aplicada a ámbitos diferentes.

5.2 Validación del modelo.

Una vez resumido todo el trabajo de investigación realizado y los logros obtenidos en este, cabe reseñar que nos encontramos ante un modelo de predicción de la innovación que presenta ciertas ventajas respecto a los modelos analizados en el apartado 2.2.

En primer lugar, en la mayor parte de la literatura analizada que hacía uso de los métodos mencionados en dicho apartado utilizaban solamente datos relativos a las patentes como datos de análisis, bien a través de los metadatos de estas o bien mediante el uso del texto completo de la misma. En este modelo se utilizan dos fuentes distintas de información, artículos científicos indexados y patentes, estableciéndose una correlación entre ellos que asegura la calidad de los resultados.

En segundo lugar, a partir de los datos obtenidos, una vez preprocesados, se calculan tres diferentes parámetros:

- La inversión acumulada o esfuerzo realizado por los investigadores para difundir su investigación a través de artículos científicos indexados.
- La proclividad o tendencia de los investigadores a patentar sus inventos, de decir, su propiedad intelectual.
- La generación real de propiedad intelectual.

De esta forma se consigue representar, de una forma sencilla, lo que ha ocurrido y está sucediendo en el ámbito de la tecnología analizada a partir de estos conceptos elementales.

En tercer lugar, el análisis de la red Bayesiana obtenida junto con el árbol de decisión, nos permiten inferir un conjunto de reglas de comportamiento referentes a la tecnología estudiada. Este conjunto de reglas, de fácil interpretación, permiten de una forma inequívoca, no solo clasificar los nuevos sucesos producidos, es decir, las nuevas publicaciones realizadas, sino deducir con un alto grado de precisión si va a surgir una nueva patente relativa a dichos sucesos.

En cuarto lugar, se han alcanzado unos porcentajes de precisión en torno al 85%, lo cual se considera un buen porcentaje que asegura la validez del modelo.

En quinto y último lugar, se ha comprobado la viabilidad del modelo con tecnologías diferentes, lo que abre unas amplias perspectivas para su uso.

Para dar por concluida esta exposición, es necesario señalar que se han podido cumplir todos los objetivos iniciales planteados en esta tesis doctoral, es decir:

- Se ha validado la existencia de la correlación existente entre autores de artículos científicos en revistas indexadas e inventores que protegen su propiedad intelectual en forma de patentes.
- Se ha confirmado, en ambos casos, que existe la masa crítica suficiente susceptible de ser analizada en función del ámbito científico-tecnológico examinado, es decir, estampación en caliente y hierro fundido.
- Se ha demostrado que la correlación existente puede ser modelada en base a métodos estadísticos y técnicas de aprendizaje automático, dando lugar a diferentes modelos, habiéndose utilizado un modelo de red Bayesiana y otro modelo de árbol de decisión para facilitar la comprensión de ambas tecnologías.
- Se ha verificado que los modelos alcanzan la precisión necesaria para predecir la aparición de innovaciones, así como para justificar su protección intelectual en forma de patente.
- Se han identificado patrones de comportamiento de la actividad científica susceptibles de producir nueva propiedad intelectual, que pueden ser utilizados para tomar decisiones en el ámbito de la inteligencia competitiva.

Habiéndose superado todos y cada uno de los objetivos planteados al comienzo de esta investigación, se considera superado el objetivo principal de la misma recogido en su hipótesis de investigación, que no es otra sino,

> Es posible generar un modelo predictivo de innovación, que varía en función del ámbito científico-tecnológico analizado, basado en la correlación existente entre el histórico de producción científica elaborado en base a artículos publicados en revistas indexadas, y la subsiguiente generación de propiedad intelectual que debe ser protegida

5.3 Limitaciones del modelo.

En primer lugar, una de las limitaciones que nos afecta es el preprocesamiento de datos una vez adquiridos utilizando las palabras clave en las búsquedas realizadas en las diferentes bases de datos utilizadas, es decir, WoS y PatBase®. Estamos

hablando de sistemas heterogéneos en cuanto a la presentación de los datos. Esto implica que hay que hacer un trabajo de homogeneización del dato. Para este proyecto de investigación se ha desarrollado de forma manual, lo cual conlleva un gran consumo de tiempo.

Esto se debe, a que los datos del campo autor referentes a los artículos publicados e inventor cuando hacemos referencia a las patentes no aparecen de la misma forma, ni existe forma de identificarlo de forma biunívoca en ambos conjuntos de datos.

Cuando hablamos de artículos científicos, la nomenclatura de los nombres de los autores es «Apellido», «Inicial_del_Nombre». No sucede lo mismo en los nombres de los inventores dónde se ha encontrado una estructura «Nombre» «Apellido_1» «Apellido_2», lo que implica un proceso de homogeneización de estos de forma que la relación quede establecida. Esto requiere no solo la pericia del investigador, sino también una gran cantidad de tiempo hasta optimizar el conjunto inicial de datos. Si unimos esto a que es frecuente que la mencionada estructura en patentes aparezca incluyendo los títulos de las personas como, por ejemplo, Dr., Ing., Lcdo., o el país o empresa de estos, la tarea de limpieza y adaptación de los datos de partida se hace aún más compleja.

Surge, además, en este contexto, otra fuente de incertidumbre inherente a la forma establecida de almacenar dicha información, especialmente cuando hablamos de patentes. Cuando tenemos dos o más autores con el mismo apellido y nombre que comienzan por la misma inicial no podemos diferenciarlos y se convierten en una persona única, lo cual se convierte en una fuente de error que, aunque es mínimo, habría que ver si es necesario tenerlo en cuenta.

En cuanto a las fechas tanto de artículos como de patentes, se ha detectado que se pierden datos ya que carecen de fecha y en ocasiones esta aparece escrita de diferente forma, es decir, día-mes-año, mes-año e incluso solo año. Es por esto por lo que se optó por utilizar solamente el año, no pudiendo ajustar exactamente los 18 meses de prioridad de una patente y teniendo que utilizar como parámetro 2 años.

Otra de las limitaciones importantes que presentan los modelos es que únicamente se recogen datos de dos fuentes diferentes, artículos científicos y patentes, que si bien muestran información importante podrían ser complementados con otras fuentes para tratar de ajustarlos más a la realidad.

Así mismo, cabe reseñar que solo se han utilizado los nombres de los autores/inventores como datos de entrada a los procesos, tanto estadístico como de aprendizaje automático, lo cual limita el análisis y la obtención de resultados.

De la misma forma, es necesario tener en cuenta las limitaciones de la inteligencia artificial en el contexto del aprendizaje automático. La interpretabilidad de los modelos es uno de los puntos para tener en cuenta. Algoritmos como las redes neuronales o los modelos de ensamblaje, son inherentemente complejos a la hora de realizar su interpretación. Este aspecto dificulta la comprensión desarrollada en la toma de decisiones limitando su aplicación en entornos críticos. Además, podemos hablar de una generalización limitada, es decir, los modelos de inteligencia artificial pueden presentar problemas a la hora de generalizar los patrones ocultos en conjuntos de datos nuevos o poco comunes, no contemplados en el conjunto de datos de entrenamiento. Esto puede llevar al modelo a presentar rendimientos deficientes. En lo referente a los conceptos de sesgo y equidad, en necesario tener en cuenta que los algoritmos de IA pueden reflejar y amplificar los sesgos existentes en el conjunto de datos de entrenamiento proporcionados. Estos errores sistemáticos en los que una IA puede incurrir al realizar los muestreos o ensayos hacen que unas respuestas sean seleccionadas de forma preferente contra otras, conduciendo a decisiones injustas o discriminatorias si hablamos de aplicaciones en el mundo real. De igual forma, estos algoritmos de aprendizaje automático requieren de grandes cantidades de datos etiquetados y limpios para realizar, con precisión, el entrenamiento de los modelos.

Es conveniente la selección óptima o adecuada del modelo, aunque esto puede representar un desafío en sí mismo ya que existe una amplia variedad de estos con capacidades y complejidades diferentes. Al mismo tiempo, la correcta selección de las métricas de evaluación constituye un punto esencial a la hora de comparar y seleccionar modelos.

Con respecto a la representación y ajustes de los modelos es necesario una correcta elección de las características que van a incluir, considerando el impacto que esto va a suponer en su rendimiento. Puede requerir un conocimiento experto del dominio estudiado. A menudo, es también necesario llevar a cabo una transformación en los datos previo a su incorporación al modelo. En ciertas ocasiones, puede ser también fundamental la creación de nuevas características a partir de las existentes para establecer mejor la relación entre las variables de entrada y la variable objetivo. Es imprescindible, también, realizar un ajuste adecuado de los hiperparámetros inherentes al modelo a ensayar en aras de optimizar el rendimiento del modelo. Esto puede constituir un proceso complejo y exigente desde un enfoque computacional, en especial con modelos más avanzados como pueden ser las redes neuronales.

De la misma forma, hay que prestar un cuidado especial al sobreajuste (del inglés *Overfitting*) o subajuste (del inglés *Underfitting*) de los modelos ya que,

- un sobreajuste en los datos de entrenamiento puede capturar ruido de los patrones subyacentes lo que produce un rendimiento deficiente en los nuevos datos y,
- un subajuste puede no capturar íntegramente la estructura subyacente de los datos produciendo igualmente un rendimiento deficiente.

5.4 Aplicaciones de la investigación.

Las aportaciones de esta tesis doctoral tienen una doble vertiente. El hecho de disponer de nuevos modelos predictivos de innovación representa, para el campo de la investigación, una oportunidad para explorar, comprender y predecir los cambios. En este sentido, este tipo de modelos son importantes en varios aspectos.

1. Ayudan en la identificación de nuevas oportunidades.
    a. Exploran nuevas ideas.
    b. Identifican áreas de oportunidad.
    c. Detectan las tecnologías emergentes.
    d. Identifican patrones emergentes de innovación.

2. Mejoran la eficiencia de los recursos disponibles mediante su optimización.
    a. Recursos humanos.
    b. Recursos financieros.
    c. Gestionan el tiempo.
3. Apoyan con la toma de decisiones la planificación estratégica.
    a. Enfoque de la investigación.
    b. Planificación de políticas y estrategias organizativas.
4. Contribuyen al avance del conocimiento.
    a. Ciencia de datos.
    b. Estadística.
    c. Inteligencia artificial.
5. Ayudan a comprender la dinámica de la innovación mediante el análisis de factores clave.
    a. Generación y desarrollo de ideas innovadoras.
    b. Aplicación en las diferentes líneas de investigación.

Es evidente, no obstante, que esta investigación tiene también un enfoque claramente orientado a la empresa facilitando la toma de decisiones en ámbito de la inteligencia competitiva. El desarrollo de nuevos modelos predictivos de innovación posibilita no solo un mayor conocimiento del comportamiento de la tecnología o ámbito analizado, sino también un análisis de las tendencias futuras que presenta.

Si se tiene en cuenta que una vez que hemos hecho el primer análisis, hacer un seguimiento, cada vez que incorporamos tanto nuevos artículos científicos como nuevas patentes a nuestro sistema, es una tarea relativamente sencilla, estos modelos son de gran interés para el ámbito empresarial.

Ya se ha explicado con anterioridad que la generación de los modelos nos muestra una serie de patrones de comportamiento referentes a la tecnología analizada. Cuando añadimos de forma periódica tanto nuevos artículos como nuevas patentes, sería suficiente realizar un análisis similar al realizado inicialmente, pero con un conjunto de datos de entrada diferentes, es decir, un subconjunto reducido

de estos, que analice solamente el comportamiento de los nuevos autores/inventores incorporados y nos proporcione los nuevos valores de las variables ANA, PENA y PASA para dicho evento. En función de los valores obtenidos sabremos si cumplen o no alguna de las reglas establecidas y si van a dar lugar o no a la generación de nueva propiedad intelectual en forma de patente, es decir, innovación.

Así mismo, es necesario revisar periódicamente estas reglas ya que las tecnologías pueden presentar cambios de comportamiento, sobre todo en aquellos casos en que la tecnología no está madura.

La aplicación de estos modelos puede ser de especial interés para despachos de patentes ya que pueden ordenar su actividad comercial anticipándose a la generación de las patentes. De la misma forma, cualquier empresa que tenga una base tecnológica y necesite conocer el estado de esta, podría estar interesada en la información proporcionada por los modelos generados, conociendo las líneas de desarrollo tanto actuales como futuras. Este conocimiento posibilita, llegado el caso, la reorientación de sus líneas de investigación, si fuese necesario, o incluso el abandono de estas si entienden que están en clara desventaja frente a sus competidores.

En definitiva, estamos hablando de modelos que facilitan la toma de decisiones de alto nivel empresarial. Teniendo en cuenta que estamos en un mercado globalizado y altamente competitivo, un posicionamiento preventivo ante los cambios que se están produciendo de una forma constante e inminente resulta de un gran interés.

5.5 Trabajo futuro.

Una vez generados estos modelos predictivos de innovación y vista su viabilidad, se plantean los siguientes retos de futuro.

El primero de ellos consiste el disminuir el incluso eliminar el preprocesamiento manual de los datos procedentes tanto de WoS como de PatBase® de forma que quede automatizado. Esto implicaría una disminución de tiempo considerable en el proceso.

Además de esto, para mejorar, la precisión de los modelos, así como el conjunto de reglas de comportamiento se plantea añadir nuevas fuentes de información. Estamos hablando de incorporar datos de conferencias o congresos, informes técnicos, proyectos de I+D, y cualquier otra fuente de datos que se considere que posee información relevante. Esto incrementará notablemente la heterogeneidad de los datos de partida, pero enriquecerá los modelos generados.

Se abre, también, una nueva línea que puede mejorar los modelos conseguidos mediante la inclusión de más metadatos en el proceso de cálculo como pueden ser las entidades donde tanto autores como inventores trabajan e incluso la combinación de estas variables para generar, por ejemplo, la combinación autor/inventor-entidad_a_la_que_pertenece como una nueva variable.

Del mismo modo, una línea paralela a la anterior sería utilizar las palabras clave que aparecen tanto en los artículos científicos como en las patentes de forma que se establezca de una forma más eficiente la concordancia entre estos, pudiendo descartar alguna temática dentro de la tecnología que no se corresponda exactamente con el análisis realizado.

Se ha considerado también la posibilidad de utilizar algoritmos de clasificación más modernos y sofisticados que los utilizados para esta investigación en la que se han utilizado dos algoritmos básicos como son la red Bayesiana y el árbol de decisión. Se ha pensado también la utilización de metaclasificadores para ver si la relación entre el grado de precisión alcanzada y el tiempo de procesado pueden verse compensados por los resultados.

Así mismo, se ha pensado en la automatización de la adquisición de los datos iniciales, por medio de APIs o algún procedimiento similar, de forma que podamos recoger esa información en tiempo real, analizarla de forma inmediata y comprobar los resultados.

Por último, pero no por ello menos importante, se plantea la posibilidad de utilizar modelos de lenguaje de gran tamaño (del inglés *Large Lenguaje Models* (LLMs)), es decir, modelos de IA entrenados con grandes conjuntos de datos de texto sin etiquetar, de forma que pueden tanto comprender como generar lenguaje natural. Otra posibilidad que se abre es el uso de procesos de Extracción (*Extract*),

Transformación (*Transform*) y Carga (*Load*), procesos ETL que, a partir de datos no estructurados procedentes de diversas fuentes los transforma a su formato final y los carga en un destino que puede estar representado por un almacén o base de datos.

En conclusión, se ha conseguido la modelización de la predicción de la innovación, es decir, la creación de modelos que permiten prever la generación de nuevas patentes abriéndose de forma simultánea líneas de desarrollo futuro que presentan un gran interés.

5.6 Reflexión final.

Cerraremos este trabajo de la misma forma que lo comenzamos: afirmando que la capacidad de predecir o prever el futuro ha sido una búsqueda constante a lo largo de la historia de la humanidad.

En este mundo en el que vivimos, en cambio constante y de una forma cada vez más rápida, la posibilidad misma de predecir, con precisión y fiabilidad, que se genere propiedad intelectual, léase innovación, nos da una ventaja importante de cara a la optimización de recursos si hablamos en nombre de la ciencia y frente a nuestros competidores si hablamos desde el punto de vista empresarial.

ANEXO 1

El presente anexo recoge los resultados obtenidos para a las métricas descritas en la sección 1.6.2 Métricas utilizadas., en el análisis realizado sobre la tecnología de estampado en caliente, que no son otras que (i) el porcentaje de precisión, (ii) el error absoluto medio, MAE, (iii) la raíz del error cuadrado medio, RMSE, (iv) la precisión, (v) la exhaustividad y (vi) el valor-F. Los resultados corresponden al conjunto de 35 discretizaciones analizadas mediante la aplicación de 24 algoritmos que se enumeran a continuación.

1. Red bayesiana (K2)
2. Red bayesiana (Hill Climber)
3. Red bayesiana (TAN)
4. Naïve Bayes
5. ANN (MLP)
6. ANN (MLP CS)
7. SVM (Polynomial)
8. SVM (Normalized Polynomial)
9. SVM (Pearson VII)
10. SVM (RBF)
11. KNN (K=1)
12. KNN (K=2)

13. KNN (K=3)
14. KNN (K=4)
15. KNN (K=5)
16. KNN (K=6)
17. C4.5
18. Random Forest (i=50)
19. Random Forest (i=100)
20. Random Forest (i=150)
21. Random Forest (i=200)
22. Random Forest (i=250)
23. Random Forest (i=300)
24. Random Forest (i=350)

Tabla 11: Nº de intervalos de discretización en los que se han dividido las variables ANA, PENA y PASA para los 35 juegos de datos que se han analizado.

	Nº de intervalos de discretización						
Variable	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
ANA1	5	5	6	5	4	7	5
ANA2	4	5	6	4	5	7	4
ANA3	3	5	6	3	5	6	3
ANA4	2	5	6	3	4	6	2
ANA5	2	4	5	2	3	6	2
ANA6	2	4	5	2	3	5	2
ANA7	2	3	4	2	2	2	2
ANA8	2	3	4	2	2	2	2
PENA1	4	5	6	5	4	7	3
PENA2	4	3	4	4	2	4	3
PENA3	2	2	3	3	2	3	2
PASA1	3	5	6	5	4	7	2
PASA2	2	2	2	2	2	2	2
PASA3	2	2	2	2	2	2	2
PASA4	2	2	2	2	2	2	2
PASA5	2	2	2	2	2	2	2

Las discretizaciones realizadas a las diferentes variables ANA, PENA y PASA para configurar los 35 juegos de datos ensayados se resumen en la **Tabla *11***. Se prepararon 7 discretizaciones diferentes, D1 a D7, cuya segmentación para las diferentes variables fue realizada utilizando intervalos iguales, es decir, por frecuencias. Cada uno de estos intervalos fue, asimismo, etiquetado. Para cada uno de estos 7 conjuntos de datos, en el caso de estampado en caliente, la discretización Dx-1 tenía en cuenta todas las variables que no fuesen 0, la Dx-2 descartaba las variables ANA6, ANA7 y ANA 8, la Dx-3 descartaba además la variable PENA3, la Dx-4 descartaba solamente las variables PASA3, PASA4 y PASA5 y finalmente las Dx-5 descartaba todas las variables mencionadas. La x representa los 7 conjuntos de discretizaciones diferentes que se prepararon. La razón de ir descartando diferentes conjuntos de variables se debe a la poca representatividad que podían tener debido a sus valores.

El etiquetado realizado en las variables depende del número de intervalos definidos. Se especifican a continuación los intervalos para el caso D6 que contiene el mayor número de segmentaciones. Para las diferentes variables se definieron las escalas que se muestran a continuación.

ANA1

No activity – Starters – Low activity – Low-Medium activity – Medium activity – Medium-Strong activity – Strong activity

ANA2

No cooperation – Weak cooperation - Low cooperation – Low-Medium cooperation – Medium cooperation – Medium-Strong cooperation – Strong cooperation

ANA3 – ANA4 – ANA 5

No networking – Low networking – Low-Medium networking – Medium networking – Medium-Strong networking – Strong networking

ANA6

No networking – Weak networking - Low networking – Medium networking – Strong networking

ANA7 – ANA8

No networking – Networking

PENA1

No patenting culture – Weak patenting culture – Low patenting culture – Low-Medium patenting culture – Medium patenting culture – Medium-Strong patenting culture – Strong patenting culture

PENA2

No patenting culture –Low patenting culture – Medium patenting culture – Strong patenting culture

PENA3

Low patenting culture – Medium patenting culture – Strong patenting culture

PASA1

No IP production – Weak IP production – Low IP production – Low-Medium IP production – Medium IP production – Medium-Strong IP production – Strong IP production

PASA2 – PASA3

No IP production – IP production

Tabla 12: Porcentaje de precisión (o acierto), correctitud, obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	71,00	71,03	74,02	70,99	73,87	73,87	73,25	73,38	73,67	73,94	73,55	73,47	73,41	73,45	73,42	73,39	73,92	73,73	73,69	73,68	73,69	73,66	73,67	73,66
D1-2	71,08	71,10	73,99	71,07	73,76	73,76	73,25	73,37	73,68	73,93	73,55	73,51	73,50	73,49	73,47	73,47	73,91	73,71	73,70	73,67	73,70	73,68	73,67	73,66
D1-3	71,09	71,10	73,99	71,07	73,84	73,84	73,26	73,39	73,72	73,93	73,57	73,51	73,50	73,49	73,48	73,49	73,86	73,74	73,71	73,68	73,71	73,69	73,68	73,68
D1-4	71,05	71,05	74,03	71,03	73,88	73,88	73,25	73,40	73,64	73,94	73,54	73,47	73,41	73,44	73,43	73,41	73,92	73,70	73,67	73,66	73,66	73,63	73,64	73,63
D1-5	71,14	71,14	74,00	71,10	73,83	73,83	73,25	73,40	73,70	73,94	73,57	73,52	73,52	73,50	73,49	73,50	73,89	73,72	73,70	73,67	73,68	73,66	73,66	73,65
D2-1	66,25	66,27	68,38	66,24	68,19	68,19	68,69	68,63	68,72	68,67	68,09	68,03	67,92	67,88	67,81	67,83	68,96	68,56	68,60	68,61	68,61	68,62	68,62	68,61
D2-2	66,30	66,29	68,45	66,30	68,34	68,34	68,69	68,67	68,77	68,67	68,23	68,09	67,97	67,93	67,91	67,97	68,95	68,69	68,71	68,73	68,72	68,76	68,74	68,75
D2-3	66,30	66,29	68,43	66,30	68,30	68,30	68,69	68,67	68,77	68,67	68,21	68,08	67,97	67,90	67,90	67,97	68,95	68,70	68,72	68,71	68,73	68,73	68,72	68,72
D2-4	66,24	66,27	68,35	66,22	68,27	68,27	68,68	68,61	68,74	68,67	68,10	67,99	67,93	67,89	67,80	67,83	68,95	68,58	68,60	68,64	68,63	68,64	68,64	68,62
D2-5	66,29	66,29	68,41	66,27	68,37	68,37	68,72	68,65	68,81	68,67	68,22	68,07	68,00	67,92	67,91	67,98	68,92	68,68	68,71	68,72	68,70	68,72	68,73	68,73
D3-1	65,58	65,54	67,65	65,59	67,58	67,58	67,66	67,66	68,19	67,49	67,15	66,72	66,76	66,81	66,76	66,66	68,09	67,83	67,86	67,89	67,88	67,88	67,89	67,91
D3-2	65,56	65,54	67,65	65,55	67,48	67,48	67,67	67,69	68,10	67,52	67,10	66,71	66,76	66,81	66,76	66,65	68,25	67,79	67,82	67,85	67,83	67,83	67,83	67,85
D3-3	65,55	65,54	67,64	65,54	67,43	67,43	67,69	67,72	68,09	67,51	67,12	66,69	66,75	66,80	66,75	66,63	68,26	67,76	67,80	67,82	67,82	67,82	67,82	67,83
D3-4	65,56	65,54	67,56	65,55	67,48	67,48	67,63	67,69	68,19	67,49	67,15	66,71	66,77	66,81	66,76	66,66	68,08	67,83	67,88	67,90	67,88	67,87	67,89	67,91
D3-5	65,54	65,54	67,56	65,49	67,44	67,44	67,65	67,68	68,08	67,51	67,10	66,67	66,73	66,78	66,73	66,63	68,24	67,77	67,78	67,81	67,77	67,79	67,79	67,82
D4-1	66,74	66,73	68,67	66,75	68,69	68,69	68,98	68,99	68,81	68,85	68,33	68,13	68,04	67,92	67,90	67,86	68,70	68,75	68,75	68,75	68,76	68,75	68,74	68,75
D4-2	66,85	66,85	68,72	66,86	68,52	68,52	69,02	68,99	68,79	68,85	68,34	68,15	68,06	67,95	67,93	67,86	68,68	68,69	68,67	68,66	68,66	68,66	68,65	68,64
D4-3	66,85	66,85	68,72	66,86	68,67	68,67	69,02	68,98	68,78	68,85	68,34	68,15	68,08	67,97	67,94	67,86	68,68	68,66	68,64	68,66	68,65	68,63	68,62	68,63
D4-4	69,89	69,89	75,22	69,86	74,65	74,65	74,49	74,34	75,23	73,79	74,75	74,56	74,47	74,47	74,44	74,43	75,08	75,03	75,06	75,10	75,10	75,08	75,08	75,11
D4-5	66,84	66,84	68,70	66,82	68,68	68,68	69,01	68,98	68,81	68,85	68,34	68,14	68,08	67,98	67,93	67,83	68,70	68,69	68,67	68,67	68,67	68,66	68,65	68,64
D5-1	67,98	67,99	70,13	67,97	70,20	70,20	69,81	69,95	70,64	69,96	70,33	70,28	70,16	70,10	70,04	69,98	70,36	70,52	70,52	70,49	70,51	70,54	70,54	70,54
D5-2	67,97	67,97	70,17	67,98	70,26	70,26	69,79	69,94	70,57	69,97	70,30	70,23	70,10	70,10	70,08	70,02	70,46	70,48	70,46	70,43	70,43	70,48	70,46	70,47
D5-3	67,97	67,97	70,16	67,97	70,15	70,15	69,81	69,95	70,57	69,97	70,27	70,21	70,08	70,09	70,06	70,01	70,44	70,45	70,45	70,42	70,41	70,46	70,45	70,46
D5-4	67,98	67,98	70,14	67,99	70,21	70,21	69,78	69,95	70,64	69,96	70,33	70,26	70,18	70,11	70,06	69,99	70,31	70,49	70,50	70,47	70,49	70,53	70,52	70,52
D5-5	67,97	67,97	70,16	67,97	70,11	70,11	69,79	69,95	70,58	69,97	70,23	70,19	70,11	70,09	70,06	70,00	70,39	70,43	70,42	70,38	70,38	70,44	70,43	70,43
D6-1	64,96	64,96	67,22	64,95	66,74	66,74	67,47	67,39	67,09	67,35	66,27	66,11	65,88	65,86	65,86	65,87	67,21	66,68	66,67	66,71	66,74	66,71	66,73	66,76
D6-2	64,99	64,96	67,28	64,97	66,78	66,78	67,48	67,44	67,11	67,37	66,29	66,10	65,86	65,80	65,77	65,78	67,31	66,68	66,70	66,73	66,73	66,71	66,70	66,74
D6-3	64,99	64,96	67,28	64,97	66,81	66,81	67,50	67,46	67,12	67,36	66,29	66,11	65,87	65,82	65,78	65,79	67,33	66,66	66,67	66,72	66,73	66,73	66,72	66,75
D6-4	69,67	69,63	73,56	69,68	72,74	72,74	73,51	73,16	73,51	72,49	72,94	72,63	72,55	72,53	72,44	72,44	73,55	73,05	73,11	73,09	73,14	73,14	73,13	73,15
D6-5	64,96	64,96	67,19	64,91	66,70	66,70	67,43	67,46	67,12	67,36	66,30	66,10	65,88	65,82	65,79	65,81	67,30	66,60	66,61	66,67	66,68	66,66	66,66	66,69
D7-1	78,42	78,39	82,40	78,42	82,23	82,23	82,45	82,45	82,28	82,15	82,18	82,17	82,16	82,16	82,18	82,19	82,49	82,25	82,26	82,25	82,26	82,26	82,26	82,25
D7-2	78,09	78,08	82,40	78,10	82,30	82,30	82,45	82,45	82,36	82,15	82,27	82,24	82,22	82,21	82,21	82,21	82,49	82,33	82,33	82,32	82,33	82,33	82,34	82,33
D7-3	78,06	78,06	82,41	78,07	82,32	82,32	82,45	82,45	82,37	82,15	82,29	82,26	82,23	82,21	82,22	82,22	82,50	82,36	82,36	82,35	82,37	82,36	82,37	82,36
D7-4	78,32	78,30	82,40	78,33	82,25	82,25	82,45	82,45	82,28	82,15	82,19	82,16	82,15	82,16	82,18	82,19	82,49	82,25	82,25	82,24	82,25	82,26	82,26	82,25
D7-5	77,98	77,98	82,41	77,99	82,33	82,33	82,45	82,45	82,37	82,15	82,30	82,25	82,22	82,21	82,22	82,22	82,50	82,36	82,36	82,35	82,37	82,37	82,37	82,36

Tabla 13: Error absoluto medio (*Mean Absolute Error*, MAE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,24	0,24	0,24	0,24	0,24	0,24	0,30	0,30	0,30	0,30	0,24	0,24	0,24	0,24	0,25	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
D1-2	0,24	0,24	0,24	0,24	0,24	0,24	0,30	0,30	0,30	0,30	0,24	0,24	0,24	0,24	0,25	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
D1-3	0,24	0,24	0,24	0,24	0,24	0,24	0,30	0,30	0,30	0,30	0,24	0,24	0,24	0,24	0,24	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
D1-4	0,24	0,24	0,24	0,24	0,24	0,24	0,30	0,30	0,30	0,30	0,24	0,24	0,24	0,24	0,25	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
D1-5	0,24	0,24	0,24	0,24	0,24	0,24	0,30	0,30	0,30	0,30	0,24	0,24	0,24	0,24	0,24	0,25	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
D2-1	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D2-2	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D2-3	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D2-4	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D2-5	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D3-1	0,14	0,14	0,14	0,14	0,14	0,14	0,24	0,24	0,24	0,24	0,14	0,14	0,14	0,14	0,15	0,15	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
D3-2	0,14	0,14	0,14	0,14	0,14	0,14	0,24	0,24	0,24	0,24	0,14	0,14	0,14	0,14	0,14	0,15	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
D3-3	0,14	0,14	0,14	0,14	0,14	0,14	0,24	0,24	0,24	0,24	0,14	0,14	0,14	0,14	0,14	0,15	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
D3-4	0,14	0,14	0,14	0,14	0,14	0,14	0,24	0,24	0,24	0,24	0,14	0,14	0,14	0,14	0,15	0,15	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
D3-5	0,14	0,14	0,14	0,14	0,14	0,14	0,24	0,24	0,24	0,24	0,14	0,14	0,14	0,14	0,14	0,15	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
D4-1	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D4-2	0,17	0,17	0,16	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D4-3	0,17	0,17	0,17	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D4-4	0,14	0,14	0,13	0,14	0,13	0,13	0,26	0,26	0,26	0,26	0,13	0,13	0,14	0,14	0,14	0,14	0,14	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D4-5	0,17	0,17	0,17	0,17	0,16	0,16	0,27	0,27	0,27	0,27	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17	0,17
D5-1	0,20	0,20	0,20	0,20	0,20	0,20	0,29	0,29	0,29	0,29	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D5-2	0,20	0,20	0,20	0,20	0,20	0,20	0,29	0,29	0,29	0,29	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D5-3	0,20	0,20	0,20	0,20	0,20	0,20	0,29	0,29	0,29	0,29	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D5-4	0,20	0,20	0,20	0,20	0,20	0,20	0,29	0,29	0,29	0,29	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D5-5	0,20	0,20	0,20	0,20	0,20	0,20	0,29	0,29	0,29	0,29	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D6-1	0,12	0,13	0,12	0,12	0,12	0,12	0,22	0,22	0,22	0,22	0,12	0,12	0,13	0,13	0,13	0,13	0,13	0,12	0,12	0,12	0,12	0,12	0,12	0,12
D6-2	0,12	0,13	0,12	0,12	0,12	0,12	0,22	0,22	0,22	0,22	0,12	0,12	0,13	0,13	0,13	0,13	0,13	0,12	0,12	0,12	0,12	0,12	0,12	0,12
D6-3	0,12	0,13	0,12	0,12	0,12	0,12	0,22	0,22	0,22	0,22	0,12	0,12	0,13	0,13	0,13	0,13	0,13	0,12	0,12	0,12	0,12	0,12	0,12	0,12
D6-4	0,11	0,11	0,10	0,11	0,09	0,09	0,21	0,21	0,21	0,21	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D6-5	0,13	0,13	0,12	0,12	0,12	0,12	0,22	0,22	0,22	0,22	0,12	0,12	0,13	0,13	0,13	0,13	0,13	0,12	0,12	0,12	0,12	0,12	0,12	0,12
D7-1	0,27	0,27	0,27	0,27	0,27	0,27	0,18	0,18	0,18	0,18	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-2	0,27	0,27	0,27	0,27	0,27	0,27	0,18	0,18	0,18	0,18	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-3	0,27	0,27	0,27	0,27	0,27	0,27	0,18	0,18	0,18	0,18	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-4	0,27	0,27	0,27	0,27	0,27	0,27	0,18	0,18	0,18	0,18	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-5	0,27	0,27	0,27	0,27	0,27	0,27	0,18	0,18	0,18	0,18	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28

Tabla 14: Raíz del error cuadrado medio (*Root Mean Square Error*, RMSE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,38	0,38	0,35	0,38	0,35	0,35	0,39	0,39	0,39	0,39	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
D1-2	0,37	0,37	0,35	0,37	0,35	0,35	0,39	0,39	0,39	0,39	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
D1-3	0,37	0,37	0,35	0,37	0,35	0,35	0,39	0,39	0,39	0,39	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
D1-4	0,38	0,38	0,35	0,38	0,35	0,35	0,39	0,39	0,39	0,39	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
D1-5	0,37	0,37	0,35	0,37	0,35	0,35	0,39	0,39	0,39	0,39	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
D2-1	0,31	0,31	0,29	0,31	0,30	0,30	0,36	0,36	0,36	0,36	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D2-2	0,31	0,31	0,29	0,31	0,29	0,29	0,36	0,36	0,36	0,36	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D2-3	0,31	0,31	0,29	0,31	0,29	0,29	0,36	0,36	0,36	0,36	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D2-4	0,31	0,31	0,29	0,31	0,30	0,30	0,36	0,36	0,36	0,36	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D2-5	0,31	0,31	0,29	0,31	0,29	0,29	0,36	0,36	0,36	0,36	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D3-1	0,28	0,28	0,27	0,28	0,28	0,28	0,34	0,34	0,34	0,34	0,28	0,28	0,28	0,28	0,28	0,28	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
D3-2	0,28	0,28	0,27	0,28	0,28	0,28	0,34	0,34	0,34	0,34	0,28	0,28	0,28	0,28	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
D3-3	0,28	0,28	0,27	0,28	0,28	0,28	0,34	0,34	0,34	0,34	0,28	0,28	0,28	0,28	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
D3-4	0,28	0,28	0,27	0,28	0,28	0,28	0,34	0,34	0,34	0,34	0,28	0,28	0,28	0,28	0,28	0,28	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
D3-5	0,28	0,28	0,27	0,28	0,27	0,27	0,34	0,34	0,34	0,34	0,28	0,28	0,28	0,28	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
D4-1	0,31	0,30	0,29	0,31	0,29	0,29	0,35	0,35	0,35	0,35	0,30	0,30	0,30	0,30	0,30	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D4-2	0,30	0,30	0,29	0,30	0,29	0,29	0,35	0,35	0,35	0,35	0,30	0,29	0,29	0,29	0,29	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D4-3	0,30	0,30	0,29	0,30	0,29	0,29	0,35	0,35	0,35	0,35	0,30	0,29	0,29	0,29	0,29	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D4-4	0,29	0,29	0,26	0,29	0,26	0,26	0,34	0,35	0,34	0,35	0,27	0,27	0,27	0,27	0,27	0,26	0,27	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D4-5	0,30	0,30	0,29	0,30	0,29	0,29	0,35	0,35	0,35	0,35	0,30	0,29	0,29	0,29	0,29	0,30	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
D5-1	0,34	0,34	0,32	0,34	0,32	0,32	0,37	0,37	0,38	0,37	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
D5-2	0,33	0,33	0,32	0,33	0,32	0,32	0,37	0,37	0,38	0,37	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
D5-3	0,33	0,33	0,32	0,33	0,32	0,32	0,37	0,37	0,38	0,37	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
D5-4	0,34	0,34	0,32	0,34	0,32	0,32	0,37	0,37	0,38	0,37	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
D5-5	0,33	0,33	0,32	0,33	0,32	0,32	0,37	0,37	0,38	0,37	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
D6-1	0,26	0,26	0,25	0,26	0,26	0,26	0,32	0,32	0,32	0,32	0,27	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D6-2	0,26	0,26	0,25	0,26	0,26	0,26	0,32	0,32	0,32	0,32	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D6-3	0,26	0,26	0,25	0,26	0,26	0,26	0,32	0,32	0,32	0,32	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D6-4	0,25	0,25	0,23	0,25	0,24	0,24	0,32	0,32	0,32	0,32	0,24	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D6-5	0,26	0,26	0,25	0,26	0,26	0,26	0,32	0,32	0,32	0,32	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D7-1	0,41	0,41	0,38	0,41	0,38	0,38	0,42	0,42	0,42	0,42	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38
D7-2	0,41	0,41	0,38	0,41	0,38	0,38	0,42	0,42	0,42	0,42	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,37	0,37	0,37	0,37	0,37	0,37	0,37
D7-3	0,41	0,41	0,38	0,41	0,38	0,38	0,42	0,42	0,42	0,42	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,37	0,37	0,37	0,37	0,37	0,37	0,37
D7-4	0,41	0,41	0,38	0,41	0,38	0,38	0,42	0,42	0,42	0,42	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38
D7-5	0,41	0,41	0,38	0,41	0,38	0,38	0,42	0,42	0,42	0,42	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,37	0,37	0,37	0,37	0,37	0,37	0,37

Tabla 15: Precisión obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,81	0,81	0,83	0,81	0,83	0,83	0,84	0,84	0,83	0,84	0,82	0,82	0,82	0,82	0,81	0,81	0,83	0,82	0,82	0,82	0,82	0,82	0,82	0,82
D1-2	0,81	0,81	0,83	0,81	0,83	0,83	0,84	0,84	0,83	0,84	0,82	0,82	0,82	0,82	0,82	0,81	0,83	0,82	0,82	0,82	0,82	0,82	0,82	0,82
D1-3	0,81	0,81	0,83	0,81	0,83	0,83	0,84	0,84	0,83	0,84	0,82	0,82	0,82	0,82	0,82	0,81	0,83	0,82	0,82	0,82	0,82	0,82	0,82	0,82
D1-4	0,81	0,81	0,83	0,81	0,83	0,83	0,84	0,84	0,83	0,84	0,82	0,82	0,82	0,82	0,81	0,81	0,83	0,82	0,82	0,82	0,82	0,82	0,82	0,82
D1-5	0,81	0,81	0,83	0,81	0,83	0,83	0,84	0,84	0,83	0,84	0,82	0,82	0,82	0,82	0,82	0,81	0,83	0,82	0,82	0,82	0,82	0,82	0,82	0,82
D2-1	0,77	0,78	0,77	0,77	0,78	0,78	0,78	0,77	0,78	0,74	0,77	0,77	0,76	0,76	0,76	0,75	0,77	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D2-2	0,78	0,78	0,77	0,78	0,79	0,79	0,78	0,77	0,78	0,74	0,77	0,77	0,77	0,76	0,76	0,76	0,77	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D2-3	0,78	0,78	0,77	0,78	0,79	0,79	0,77	0,77	0,78	0,74	0,77	0,77	0,77	0,76	0,76	0,76	0,77	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D2-4	0,77	0,78	0,77	0,77	0,79	0,79	0,78	0,77	0,78	0,74	0,77	0,77	0,76	0,76	0,76	0,75	0,77	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D2-5	0,78	0,78	0,77	0,78	0,79	0,79	0,78	0,77	0,78	0,74	0,77	0,77	0,77	0,76	0,76	0,76	0,77	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D3-1	0,77	0,77	0,76	0,77	0,77	0,77	0,74	0,73	0,77	0,70	0,76	0,75	0,75	0,74	0,74	0,73	0,76	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D3-2	0,77	0,77	0,76	0,77	0,77	0,77	0,74	0,73	0,77	0,70	0,76	0,75	0,75	0,75	0,74	0,73	0,76	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D3-3	0,77	0,77	0,76	0,77	0,77	0,77	0,74	0,73	0,77	0,70	0,76	0,75	0,75	0,75	0,74	0,73	0,76	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D3-4	0,77	0,77	0,76	0,77	0,77	0,77	0,74	0,73	0,77	0,70	0,76	0,75	0,75	0,74	0,74	0,73	0,76	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D3-5	0,77	0,77	0,76	0,77	0,77	0,77	0,74	0,73	0,77	0,70	0,76	0,75	0,75	0,75	0,74	0,73	0,76	0,78	0,78	0,78	0,78	0,78	0,78	0,78
D4-1	0,78	0,78	0,78	0,78	0,80	0,80	0,78	0,78	0,79	0,77	0,79	0,79	0,78	0,78	0,78	0,78	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D4-2	0,78	0,78	0,78	0,78	0,80	0,80	0,78	0,78	0,80	0,77	0,79	0,79	0,78	0,78	0,78	0,78	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80
D4-3	0,78	0,78	0,78	0,78	0,80	0,80	0,78	0,78	0,79	0,77	0,79	0,79	0,78	0,78	0,78	0,78	0,79	0,79	0,80	0,79	0,80	0,80	0,80	0,80
D4-4	0,80	0,80	0,82	0,80	0,83	0,83	0,81	0,82	0,83	0,79	0,82	0,82	0,82	0,82	0,81	0,81	0,82	0,83	0,83	0,83	0,83	0,83	0,83	0,83
D4-5	0,78	0,78	0,78	0,78	0,80	0,80	0,78	0,78	0,79	0,77	0,79	0,79	0,78	0,78	0,78	0,78	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80
D5-1	0,79	0,79	0,80	0,79	0,81	0,81	0,77	0,77	0,80	0,77	0,79	0,79	0,79	0,79	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D5-2	0,79	0,79	0,80	0,79	0,81	0,81	0,77	0,77	0,80	0,77	0,79	0,79	0,79	0,79	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D5-3	0,79	0,79	0,80	0,79	0,81	0,81	0,77	0,77	0,80	0,77	0,79	0,79	0,79	0,79	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D5-4	0,79	0,79	0,80	0,79	0,81	0,81	0,77	0,77	0,80	0,77	0,79	0,79	0,79	0,79	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D5-5	0,79	0,79	0,80	0,79	0,81	0,81	0,77	0,77	0,80	0,77	0,79	0,79	0,79	0,79	0,79	0,79	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80
D6-1	0,76	0,76	0,76	0,76	0,76	0,76	0,72	0,71	0,75	0,69	0,75	0,74	0,73	0,72	0,71	0,71	0,75	0,77	0,77	0,77	0,77	0,77	0,77	0,77
D6-2	0,76	0,76	0,76	0,76	0,76	0,76	0,72	0,71	0,75	0,69	0,75	0,74	0,73	0,72	0,71	0,71	0,75	0,77	0,77	0,77	0,77	0,77	0,77	0,77
D6-3	0,76	0,76	0,76	0,76	0,76	0,76	0,72	0,71	0,75	0,69	0,75	0,74	0,73	0,72	0,71	0,71	0,75	0,77	0,77	0,77	0,77	0,77	0,77	0,77
D6-4	0,80	0,80	0,82	0,80	0,83	0,83	0,80	0,78	0,82	0,76	0,81	0,80	0,79	0,79	0,78	0,78	0,81	0,83	0,83	0,83	0,83	0,83	0,83	0,83
D6-5	0,76	0,76	0,76	0,76	0,77	0,77	0,72	0,72	0,75	0,69	0,75	0,74	0,73	0,72	0,71	0,71	0,75	0,77	0,77	0,77	0,77	0,77	0,77	0,77
D7-1	0,71	0,71	0,79	0,71	0,79	0,79	0,78	0,78	0,79	0,78	0,78	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79
D7-2	0,70	0,70	0,79	0,70	0,79	0,79	0,78	0,78	0,79	0,78	0,79	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79
D7-3	0,70	0,70	0,79	0,70	0,79	0,79	0,78	0,78	0,79	0,78	0,79	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79
D7-4	0,71	0,71	0,79	0,71	0,79	0,79	0,78	0,78	0,79	0,78	0,78	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79
D7-5	0,70	0,70	0,79	0,70	0,79	0,79	0,78	0,78	0,79	0,78	0,79	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79

Tabla 16: Exhaustividad (del vocablo inglés *Recall*) obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,88	0,88	0,91	0,88	0,91	0,91	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D1-2	0,88	0,88	0,91	0,88	0,91	0,91	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D1-3	0,88	0,88	0,91	0,88	0,91	0,91	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D1-4	0,88	0,88	0,91	0,88	0,91	0,91	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D1-5	0,88	0,88	0,91	0,88	0,91	0,91	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D2-1	0,91	0,91	0,94	0,91	0,94	0,94	0,94	0,95	0,94	0,96	0,94	0,94	0,94	0,94	0,94	0,95	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D2-2	0,91	0,91	0,94	0,91	0,94	0,94	0,94	0,95	0,94	0,96	0,94	0,94	0,94	0,94	0,94	0,94	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D2-3	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,95	0,94	0,96	0,94	0,94	0,94	0,94	0,94	0,94	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D2-4	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,95	0,94	0,96	0,94	0,94	0,94	0,94	0,94	0,95	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D2-5	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,95	0,94	0,96	0,94	0,94	0,94	0,94	0,94	0,94	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D3-1	0,91	0,91	0,95	0,91	0,95	0,95	0,96	0,96	0,95	0,99	0,95	0,96	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D3-2	0,91	0,91	0,95	0,91	0,95	0,95	0,96	0,96	0,95	0,99	0,95	0,95	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D3-3	0,91	0,91	0,95	0,91	0,95	0,95	0,96	0,96	0,95	0,99	0,95	0,95	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D3-4	0,91	0,91	0,95	0,91	0,95	0,95	0,96	0,96	0,95	0,99	0,95	0,96	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D3-5	0,91	0,91	0,95	0,91	0,94	0,94	0,96	0,96	0,95	0,99	0,95	0,95	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D4-1	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,94	0,93	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D4-2	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,94	0,93	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D4-3	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,94	0,93	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D4-4	0,92	0,92	0,97	0,92	0,96	0,96	0,97	0,97	0,96	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D4-5	0,91	0,91	0,94	0,91	0,93	0,93	0,94	0,94	0,93	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D5-1	0,90	0,90	0,93	0,90	0,92	0,92	0,95	0,95	0,93	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D5-2	0,90	0,90	0,93	0,90	0,92	0,92	0,95	0,95	0,93	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,92	0,92	0,93	0,93	0,93	0,93
D5-3	0,90	0,90	0,93	0,90	0,92	0,92	0,95	0,95	0,93	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,92	0,93	0,93	0,93	0,93
D5-4	0,90	0,90	0,93	0,90	0,92	0,92	0,95	0,95	0,93	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D5-5	0,90	0,90	0,93	0,90	0,92	0,92	0,95	0,95	0,93	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,92	0,92	0,93	0,93	0,93	0,93
D6-1	0,92	0,92	0,95	0,92	0,95	0,95	0,98	0,98	0,96	0,99	0,96	0,96	0,96	0,97	0,97	0,97	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D6-2	0,92	0,92	0,95	0,92	0,95	0,95	0,98	0,98	0,96	0,99	0,96	0,96	0,96	0,97	0,97	0,97	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D6-3	0,92	0,92	0,95	0,92	0,95	0,95	0,98	0,98	0,96	0,99	0,96	0,96	0,96	0,97	0,97	0,97	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
D6-4	0,93	0,93	0,97	0,93	0,96	0,96	0,98	0,99	0,96	0,99	0,96	0,97	0,97	0,98	0,98	0,98	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D6-5	0,92	0,92	0,95	0,92	0,95	0,95	0,98	0,98	0,96	0,99	0,96	0,96	0,96	0,97	0,97	0,97	0,96	0,94	0,94	0,95	0,95	0,95	0,95	0,95
D7-1	0,66	0,66	0,68	0,66	0,68	0,68	0,69	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68
D7-2	0,66	0,66	0,68	0,66	0,68	0,68	0,69	0,69	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68
D7-3	0,66	0,66	0,68	0,66	0,68	0,68	0,69	0,69	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68
D7-4	0,66	0,66	0,68	0,66	0,68	0,68	0,69	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68
D7-5	0,66	0,66	0,68	0,66	0,68	0,68	0,69	0,69	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68	0,69	0,68	0,68	0,68	0,68	0,68	0,68	0,68

Tabla 17: Valor-F obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,84	0,84	0,87	0,84	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D1-2	0,84	0,84	0,87	0,84	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D1-3	0,84	0,84	0,87	0,84	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D1-4	0,84	0,84	0,87	0,84	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D1-5	0,84	0,84	0,87	0,84	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D2-1	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,85	0,85	0,84	0,84	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D2-2	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,85	0,85	0,84	0,84	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D2-3	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,85	0,85	0,84	0,84	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D2-4	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,85	0,85	0,84	0,84	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D2-5	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,85	0,85	0,84	0,84	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D3-1	0,83	0,84	0,85	0,83	0,85	0,85	0,83	0,83	0,85	0,82	0,85	0,84	0,84	0,84	0,83	0,83	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D3-2	0,84	0,84	0,85	0,83	0,85	0,85	0,83	0,83	0,85	0,82	0,85	0,84	0,84	0,84	0,83	0,83	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D3-3	0,84	0,84	0,85	0,83	0,85	0,85	0,83	0,83	0,85	0,82	0,85	0,84	0,84	0,84	0,83	0,83	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D3-4	0,84	0,84	0,85	0,83	0,85	0,85	0,83	0,83	0,85	0,82	0,85	0,84	0,84	0,84	0,83	0,83	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D3-5	0,84	0,84	0,85	0,84	0,85	0,85	0,83	0,83	0,85	0,82	0,85	0,84	0,84	0,84	0,83	0,83	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D4-1	0,84	0,84	0,85	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-2	0,84	0,84	0,85	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-3	0,84	0,84	0,85	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-4	0,86	0,86	0,89	0,86	0,89	0,89	0,88	0,89	0,89	0,87	0,89	0,89	0,89	0,89	0,88	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D4-5	0,84	0,84	0,85	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D5-1	0,84	0,84	0,86	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,86	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D5-2	0,84	0,84	0,86	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D5-3	0,84	0,84	0,86	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D5-4	0,84	0,84	0,86	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,86	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D5-5	0,84	0,84	0,86	0,84	0,86	0,86	0,85	0,85	0,86	0,85	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D6-1	0,83	0,83	0,84	0,83	0,85	0,85	0,83	0,82	0,84	0,82	0,84	0,83	0,83	0,83	0,82	0,82	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D6-2	0,83	0,83	0,85	0,83	0,85	0,85	0,83	0,83	0,84	0,82	0,84	0,83	0,83	0,83	0,82	0,82	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D6-3	0,83	0,83	0,85	0,83	0,85	0,85	0,83	0,83	0,84	0,82	0,84	0,83	0,83	0,83	0,82	0,82	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D6-4	0,86	0,86	0,89	0,86	0,89	0,89	0,88	0,87	0,88	0,86	0,88	0,88	0,87	0,87	0,87	0,87	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D6-5	0,83	0,83	0,85	0,83	0,85	0,85	0,83	0,83	0,84	0,82	0,84	0,83	0,83	0,83	0,82	0,82	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D7-1	0,68	0,68	0,73	0,68	0,73	0,73	0,74	0,74	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73
D7-2	0,68	0,68	0,73	0,68	0,73	0,73	0,74	0,74	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73
D7-3	0,68	0,68	0,73	0,68	0,73	0,73	0,74	0,74	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73
D7-4	0,68	0,68	0,73	0,68	0,73	0,73	0,74	0,74	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73
D7-5	0,68	0,68	0,73	0,68	0,73	0,73	0,74	0,74	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73	0,73

ANEXO 2

El presente anexo recoge los resultados obtenidos para a las métricas descritas en la sección 1.6.2 Métricas utilizadas., en el análisis realizado sobre la tecnología de hierro fundido, que no son otras que (i) el porcentaje de precisión, (ii) el error absoluto medio, MAE, (iii) la raíz del error cuadrado medio, RMSE, (iv) la precisión, (v) la exhaustividad y (vi) el valor-F. Los resultados corresponden al conjunto de 35 discretizaciones analizadas mediante la aplicación de 24 algoritmos que se enumeran a continuación.

1. Red bayesiana (K2)
2. Red bayesiana (Hill Climber)
3. Red bayesiana (TAN)
4. Naïve Bayes
5. ANN (MLP)
6. ANN (MLP CS)
7. SVM (Polynomial)
8. SVM (Normalized Polynomial)
9. SVM (Pearson VII)
10. SVM (RBF)
11. KNN (K=1)
12. KNN (K=2)

13. KNN (K=3)
14. KNN (K=4)
15. KNN (K=5)
16. KNN (K=6)
17. C4.5
18. Random Forest (i=50)
19. Random Forest (i=100)
20. Random Forest (i=150)
21. Random Forest (i=200)
22. Random Forest (i=250)
23. Random Forest (i=300)
24. Random Forest (i=350)

Tabla 18: Nº de intervalos de discretización en los que se han dividido las variables ANA, PENA y PASA para los 35 juegos de datos que se han analizado.

	Nº de intervalos de discretización						
Variable	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
ANA1	5	5	6	5	4	7	5
ANA2	4	5	6	4	5	7	4
ANA3	3	5	6	3	5	6	3
ANA4	2	5	6	3	4	6	2
ANA5	2	4	5	2	3	6	2
ANA6	2	4	5	2	3	5	2
ANA7	2	3	4	2	2	2	2
ANA8	2	3	4	2	2	2	2
PENA1	4	5	6	5	4	7	3
PENA2	4	3	4	4	2	4	3
PENA3	2	2	3	3	2	3	2
PENA4	2	2	2	2	2	2	2
PASA1	3	5	6	5	4	7	2
PASA2	2	2	2	2	2	2	2
PASA3	2	2	2	2	2	2	2
PASA4	2	2	2	2	2	2	2
PASA5	2	2	2	2	2	2	2

Las discretizaciones realizadas a las diferentes variables ANA, PENA y PASA para configurar los 35 juegos de datos ensayados se resumen en la **Tabla *12***. Se prepararon 7 discretizaciones diferentes, D1 a D7, cuya segmentación para las diferentes variables fue realizada utilizando intervalos iguales, es decir, por frecuencias. Cada uno de estos intervalos fue, asimismo, etiquetado. Para cada uno de estos 7 conjuntos de datos, en el caso de fundición de hierro, la discretización Dx-1 tenía en cuenta todas las variables que no fuesen 0, la Dx-2 descartaba las variables ANA6, ANA7 y ANA 8, la Dx-3 descartaban además las variables PENA3 y PENA4, la Dx-4 descartaba solamente las variables PASA3, PASA4 y PASA5 y finalmente las Dx-5 descartaba todas las variables mencionadas. La x representa los 7 conjuntos de discretizaciones diferentes que se prepararon. La razón de ir descartando diferentes conjuntos de variables se debe a la poca representatividad que podían tener debido a sus valores.

El etiquetado realizado en las variables depende del número de intervalos definidos. Se especifican a continuación los intervalos para el caso D6 que contiene el mayor número de segmentaciones. Para las diferentes variables se definieron las escalas que se muestran a continuación.

ANA1

No activity – Starters – Low activity – Low-Medium activity – Medium activity – Medium-Strong activity – Strong activity

ANA2

No cooperation – Weak cooperation - Low cooperation – Low-Medium cooperation – Medium cooperation – Medium-Strong cooperation – Strong cooperation

ANA3 – ANA4 – ANA 5

No networking – Low networking – Low-Medium networking – Medium networking – Medium-Strong networking – Strong networking

ANA6

No networking – Weak networking - Low networking – Medium networking – Strong networking

ANA7 – ANA8

No networking – Networking

PENA1

No patenting culture – Weak patenting culture – Low patenting culture – Low-Medium patenting culture – Medium patenting culture – Medium-Strong patenting culture – Strong patenting culture

PENA2

No patenting culture –Low patenting culture – Medium patenting culture – Strong patenting culture

PENA3

Low patenting culture – Medium patenting culture – Strong patenting culture

PENA4

No patenting culture – Patenting culture

PASA1

No IP production – Weak IP production – Low IP production – Low-Medium IP production – Medium IP production – Medium-Strong IP production – Strong IP production

PASA2 – PASA3

No IP production – IP production

Tabla 19: Porcentaje de precisión (o acierto), correctitud, obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	82,27	82,27	84,24	82,27	84,13	84,13	84,80	84,83	84,25	84,84	84,23	84,33	84,34	84,38	84,38	84,40	84,73	84,15	84,13	84,14	84,15	84,16	84,15	84,15
D1-2	82,27	82,27	84,23	82,27	84,23	84,23	84,79	84,83	84,29	84,84	84,26	84,34	84,36	84,38	84,37	84,39	84,76	84,17	84,15	84,16	84,17	84,17	84,17	84,18
D1-3	82,27	82,27	84,18	82,27	84,08	84,08	84,76	84,83	84,23	84,84	84,18	84,30	84,32	84,36	84,38	84,40	84,73	84,10	84,09	84,11	84,11	84,12	84,11	84,11
D1-4	82,27	82,27	84,17	82,27	84,20	84,20	84,75	84,82	84,31	84,84	84,23	84,32	84,35	84,37	84,37	84,39	84,76	84,15	84,12	84,13	84,14	84,15	84,14	84,16
D1-5	83,47	83,45	85,95	83,46	85,97	85,97	86,43	86,43	86,14	86,04	85,77	85,68	85,51	85,47	85,42	85,37	86,37	85,98	86,02	86,00	86,00	86,01	86,02	86,02
D2-1	79,61	79,98	81,84	79,63	81,32	81,32	81,92	81,93	81,76	81,93	81,63	81,66	81,60	81,54	81,52	81,50	81,96	81,44	81,45	81,49	81,50	81,50	81,48	81,50
D2-2	79,61	79,98	81,86	79,61	81,51	81,51	81,89	81,93	81,69	81,91	81,66	81,74	81,67	81,61	81,58	81,56	81,93	81,44	81,45	81,48	81,49	81,49	81,49	81,50
D2-3	79,61	79,98	81,79	79,62	81,44	81,44	81,92	81,93	81,76	81,92	81,64	81,66	81,60	81,54	81,52	81,50	81,95	81,44	81,47	81,49	81,48	81,49	81,49	81,50
D2-4	79,61	79,98	81,80	79,61	81,46	81,46	81,89	81,92	81,68	81,93	81,66	81,73	81,68	81,60	81,58	81,56	81,90	81,44	81,46	81,49	81,51	81,49	81,49	81,50
D2-5	80,64	81,12	83,37	80,64	82,96	82,96	83,50	83,28	83,25	83,32	82,76	82,64	82,52	82,52	82,46	82,41	83,43	83,11	83,21	83,22	83,23	83,23	83,25	83,25
D3-1	75,95	76,50	78,48	75,97	78,17	78,17	78,84	78,95	78,42	78,84	78,23	78,18	78,19	78,24	78,33	78,39	78,99	78,23	78,23	78,25	78,28	78,25	78,28	78,26
D3-2	75,95	76,50	78,48	75,92	78,22	78,22	78,89	78,95	78,42	78,82	78,27	78,21	78,23	78,25	78,36	78,40	79,02	78,23	78,24	78,24	78,26	78,25	78,29	78,28
D3-3	75,92	76,50	78,49	75,95	78,12	78,12	78,85	78,95	78,40	78,81	78,24	78,21	78,22	78,28	78,37	78,42	78,99	78,22	78,22	78,24	78,27	78,21	78,26	78,26
D3-4	75,92	76,50	78,49	75,89	78,18	78,18	78,91	78,94	78,41	78,80	78,28	78,25	78,27	78,30	78,40	78,44	79,01	78,23	78,25	78,26	78,29	78,25	78,28	78,29
D3-5	78,22	78,64	81,37	78,18	81,00	81,00	81,62	81,66	80,81	81,49	80,36	79,93	79,83	79,65	79,52	79,47	81,72	80,81	80,85	80,86	80,88	80,89	80,90	80,91
D4-1	79,84	79,97	81,69	79,82	81,51	81,51	81,87	81,93	81,86	81,93	81,71	81,73	81,74	81,72	81,72	81,70	81,91	81,70	81,70	81,70	81,72	81,72	81,71	81,72
D4-2	79,83	79,97	81,70	79,82	81,64	81,64	81,86	81,93	81,83	81,92	81,71	81,69	81,71	81,70	81,70	81,69	81,92	81,73	81,72	81,71	81,73	81,72	81,72	81,73
D4-3	79,84	79,97	81,65	79,82	81,61	81,61	81,86	81,91	81,80	81,90	81,69	81,72	81,73	81,72	81,73	81,71	81,91	81,68	81,69	81,68	81,68	81,69	81,69	81,70
D4-4	79,83	79,97	81,65	79,82	81,61	81,61	81,86	81,91	81,80	81,93	81,70	81,71	81,71	81,70	81,70	81,70	81,92	81,70	81,70	81,69	81,71	81,72	81,71	81,72
D4-5	80,77	80,89	83,27	80,73	82,99	82,99	83,41	83,20	83,01	82,98	82,72	82,62	82,59	82,55	82,45	82,39	83,16	83,03	83,05	83,00	83,02	83,02	83,02	83,04
D5-1	79,92	79,95	82,46	79,94	82,10	82,10	82,66	82,65	82,48	82,65	82,35	82,33	82,38	82,38	82,38	82,35	82,58	82,27	82,30	82,28	82,31	82,32	82,33	82,31
D5-2	79,91	79,95	82,45	79,91	82,16	82,16	82,66	82,65	82,51	82,65	82,36	82,32	82,37	82,39	82,41	82,38	82,58	82,29	82,32	82,31	82,35	82,34	82,35	82,34
D5-3	79,92	79,95	82,40	79,94	82,08	82,08	82,65	82,65	82,43	82,65	82,33	82,29	82,36	82,36	82,36	82,34	82,59	82,26	82,29	82,28	82,30	82,30	82,31	82,30
D5-4	79,91	79,95	82,40	79,92	82,14	82,14	82,65	82,65	82,46	82,65	82,34	82,28	82,35	82,35	82,37	82,36	82,59	82,29	82,33	82,33	82,36	82,35	82,35	82,34
D5-5	81,27	81,36	83,97	81,26	83,56	83,56	84,15	83,93	83,76	83,67	83,62	83,41	83,35	83,32	83,25	83,22	84,01	83,69	83,73	83,72	83,76	83,76	83,78	83,77
D6-1	75,63	76,69	78,34	75,64	77,98	77,98	78,80	78,86	78,29	78,15	78,09	78,03	78,14	78,29	78,26	78,21	78,56	78,09	78,11	78,12	78,11	78,10	78,11	78,11
D6-2	75,65	76,69	78,33	75,62	78,07	78,07	78,80	78,86	78,29	78,14	78,18	78,13	78,23	78,34	78,29	78,25	78,57	78,02	78,02	78,05	78,06	78,06	78,08	78,07
D6-3	75,62	76,69	78,35	75,63	78,04	78,04	78,82	78,85	78,28	78,15	78,10	78,03	78,15	78,29	78,27	78,22	78,58	78,03	78,03	78,07	78,07	78,07	78,07	78,07
D6-4	75,63	76,69	78,34	75,61	78,09	78,09	78,82	78,84	78,28	78,14	78,19	78,14	78,23	78,35	78,30	78,25	78,59	78,03	78,05	78,08	78,08	78,09	78,11	78,10
D6-5	77,80	78,70	81,04	77,80	80,87	80,87	81,08	81,08	80,69	80,98	79,97	79,58	79,68	79,61	79,47	79,43	81,55	80,73	80,76	80,75	80,77	80,77	80,76	80,77
D7-1	86,88	86,88	89,18	86,88	89,17	89,17	89,51	89,51	89,46	89,25	89,44	89,47	89,45	89,38	89,33	89,25	89,46	89,41	89,42	89,42	89,42	89,42	89,42	89,43
D7-2	86,87	86,88	89,20	86,87	89,21	89,21	89,51	89,51	89,54	89,25	89,53	89,52	89,52	89,46	89,41	89,33	89,46	89,48	89,49	89,48	89,47	89,48	89,48	89,48
D7-3	86,88	86,88	89,18	86,88	89,09	89,09	89,51	89,51	89,43	89,25	89,39	89,42	89,42	89,35	89,32	89,24	89,47	89,39	89,41	89,39	89,40	89,39	89,40	89,40
D7-4	86,88	86,88	89,19	86,88	89,20	89,20	89,51	89,51	89,51	89,25	89,48	89,48	89,48	89,42	89,40	89,32	89,47	89,46	89,47	89,47	89,46	89,46	89,47	89,47
D7-5	86,80	86,82	89,76	86,80	89,53	89,53	89,94	89,92	89,82	89,29	89,73	89,68	89,69	89,58	89,56	89,49	90,00	89,80	89,82	89,82	89,81	89,81	89,82	89,81

Tabla 20: Error absoluto medio (*Mean Absolute Error*, MAE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,15	0,15	0,15	0,15	0,15	0,15	0,27	0,27	0,27	0,27	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
D1-2	0,15	0,15	0,15	0,15	0,15	0,15	0,27	0,27	0,27	0,27	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
D1-3	0,15	0,15	0,15	0,15	0,15	0,15	0,27	0,27	0,27	0,27	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
D1-4	0,15	0,15	0,15	0,15	0,15	0,15	0,27	0,27	0,27	0,27	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
D1-5	0,14	0,14	0,14	0,14	0,13	0,13	0,26	0,26	0,26	0,27	0,13	0,14	0,14	0,14	0,14	0,14	0,14	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D2-1	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D2-2	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D2-3	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D2-4	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D2-5	0,10	0,10	0,09	0,10	0,09	0,09	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D3-1	0,10	0,10	0,10	0,10	0,10	0,10	0,23	0,23	0,23	0,23	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D3-2	0,10	0,10	0,10	0,10	0,10	0,10	0,23	0,23	0,23	0,23	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D3-3	0,10	0,10	0,10	0,10	0,10	0,10	0,23	0,23	0,23	0,23	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D3-4	0,10	0,10	0,10	0,10	0,10	0,10	0,23	0,23	0,23	0,23	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D3-5	0,09	0,09	0,09	0,09	0,08	0,08	0,23	0,23	0,23	0,23	0,09	0,09	0,09	0,09	0,09	0,09	0,10	0,09	0,09	0,09	0,09	0,09	0,09	0,09
D4-1	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,26	0,26	0,10	0,10	0,10	0,11	0,11	0,11	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D4-2	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,26	0,26	0,10	0,10	0,10	0,11	0,11	0,11	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D4-3	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,26	0,26	0,10	0,10	0,10	0,11	0,11	0,11	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D4-4	0,11	0,11	0,10	0,11	0,10	0,10	0,25	0,25	0,26	0,26	0,10	0,10	0,10	0,11	0,11	0,11	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D4-5	0,10	0,10	0,10	0,10	0,09	0,09	0,25	0,25	0,25	0,25	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
D5-1	0,13	0,13	0,13	0,13	0,13	0,13	0,27	0,27	0,27	0,27	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D5-2	0,13	0,13	0,13	0,13	0,13	0,13	0,27	0,27	0,27	0,27	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D5-3	0,13	0,13	0,13	0,13	0,13	0,13	0,27	0,27	0,27	0,27	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D5-4	0,13	0,13	0,13	0,13	0,13	0,13	0,27	0,27	0,27	0,27	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13
D5-5	0,12	0,12	0,12	0,12	0,11	0,11	0,27	0,27	0,27	0,27	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12	0,12
D6-1	0,09	0,09	0,08	0,09	0,08	0,08	0,21	0,21	0,21	0,21	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
D6-2	0,09	0,09	0,09	0,09	0,08	0,08	0,21	0,21	0,21	0,21	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
D6-3	0,09	0,09	0,08	0,09	0,08	0,08	0,21	0,21	0,21	0,21	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
D6-4	0,09	0,09	0,09	0,09	0,08	0,08	0,21	0,21	0,21	0,21	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
D6-5	0,08	0,08	0,07	0,08	0,07	0,07	0,21	0,21	0,21	0,21	0,08	0,08	0,08	0,08	0,08	0,08	0,09	0,08	0,08	0,08	0,08	0,08	0,08	0,08
D7-1	0,17	0,17	0,16	0,17	0,15	0,15	0,10	0,10	0,11	0,11	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
D7-2	0,17	0,17	0,16	0,17	0,15	0,15	0,10	0,10	0,10	0,11	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
D7-3	0,17	0,17	0,16	0,17	0,15	0,15	0,10	0,10	0,11	0,11	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
D7-4	0,17	0,17	0,16	0,17	0,15	0,15	0,10	0,10	0,10	0,11	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
D7-5	0,16	0,16	0,15	0,16	0,14	0,14	0,10	0,10	0,10	0,11	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15

Tabla 21: Raíz del error cuadrado medio (*Root Mean Square Error*, RMSE) obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,30	0,30	0,28	0,30	0,28	0,28	0,35	0,35	0,35	0,35	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D1-2	0,30	0,30	0,28	0,30	0,28	0,28	0,35	0,35	0,35	0,35	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D1-3	0,30	0,30	0,28	0,30	0,28	0,28	0,35	0,35	0,35	0,35	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D1-4	0,30	0,30	0,28	0,30	0,28	0,28	0,35	0,35	0,35	0,35	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D1-5	0,29	0,29	0,26	0,29	0,26	0,26	0,34	0,34	0,34	0,34	0,27	0,27	0,27	0,27	0,27	0,27	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D2-1	0,25	0,25	0,23	0,25	0,24	0,24	0,34	0,34	0,34	0,34	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D2-2	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D2-3	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D2-4	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D2-5	0,24	0,24	0,22	0,24	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D3-1	0,24	0,24	0,23	0,24	0,23	0,23	0,33	0,33	0,33	0,33	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D3-2	0,24	0,24	0,23	0,24	0,23	0,23	0,33	0,33	0,33	0,33	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D3-3	0,24	0,24	0,23	0,24	0,23	0,23	0,33	0,33	0,33	0,33	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D3-4	0,24	0,24	0,23	0,24	0,23	0,23	0,33	0,33	0,33	0,33	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D3-5	0,23	0,23	0,21	0,23	0,22	0,22	0,33	0,33	0,33	0,33	0,22	0,22	0,22	0,22	0,22	0,22	0,23	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D4-1	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D4-2	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D4-3	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D4-4	0,25	0,25	0,23	0,25	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
D4-5	0,24	0,24	0,22	0,24	0,23	0,23	0,34	0,34	0,34	0,34	0,23	0,23	0,23	0,23	0,23	0,23	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D5-1	0,28	0,28	0,26	0,28	0,26	0,26	0,35	0,35	0,35	0,35	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D5-2	0,28	0,28	0,26	0,28	0,26	0,26	0,35	0,35	0,35	0,35	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D5-3	0,28	0,28	0,26	0,28	0,26	0,26	0,35	0,35	0,35	0,35	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D5-4	0,28	0,28	0,26	0,28	0,26	0,26	0,35	0,35	0,35	0,35	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
D5-5	0,27	0,27	0,25	0,27	0,25	0,25	0,35	0,35	0,35	0,35	0,25	0,25	0,25	0,25	0,25	0,25	0,24	0,25	0,25	0,25	0,25	0,25	0,25	0,25
D6-1	0,23	0,22	0,21	0,23	0,22	0,22	0,31	0,31	0,31	0,31	0,22	0,22	0,22	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D6-2	0,23	0,22	0,21	0,23	0,22	0,22	0,31	0,31	0,31	0,31	0,22	0,22	0,22	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D6-3	0,23	0,22	0,21	0,23	0,22	0,22	0,31	0,31	0,31	0,31	0,22	0,22	0,22	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D6-4	0,23	0,22	0,21	0,23	0,22	0,22	0,31	0,31	0,31	0,31	0,22	0,22	0,22	0,21	0,21	0,21	0,21	0,22	0,22	0,22	0,22	0,22	0,22	0,22
D6-5	0,22	0,21	0,20	0,22	0,21	0,21	0,31	0,31	0,31	0,31	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,20	0,20	0,20	0,20	0,20	0,20	0,20
D7-1	0,32	0,32	0,28	0,32	0,29	0,29	0,32	0,32	0,32	0,33	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-2	0,32	0,32	0,28	0,32	0,28	0,28	0,32	0,32	0,32	0,33	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-3	0,32	0,32	0,28	0,32	0,29	0,29	0,32	0,32	0,32	0,33	0,29	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-4	0,32	0,32	0,28	0,32	0,28	0,28	0,32	0,32	0,32	0,33	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
D7-5	0,32	0,32	0,27	0,32	0,28	0,28	0,32	0,32	0,32	0,33	0,28	0,28	0,28	0,28	0,28	0,28	0,27	0,28	0,28	0,28	0,28	0,28	0,28	0,28

Tabla 22: Precisión obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,88	0,88	0,88	0,88	0,88	0,88	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D1-2	0,88	0,88	0,88	0,88	0,88	0,88	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D1-3	0,88	0,88	0,88	0,88	0,88	0,88	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D1-4	0,89	0,89	0,88	0,89	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,87	0,87	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88
D1-5	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D2-1	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D2-2	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D2-3	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,85	0,85	0,85	0,85	0,85	0,84	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D2-4	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,85	0,87	0,84	0,86	0,86	0,86	0,85	0,85	0,85	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D2-5	0,83	0,83	0,82	0,83	0,83	0,83	0,81	0,81	0,82	0,80	0,83	0,82	0,82	0,82	0,82	0,81	0,82	0,83	0,83	0,83	0,83	0,83	0,83	0,83
D3-1	0,83	0,83	0,82	0,83	0,83	0,83	0,81	0,81	0,82	0,80	0,83	0,82	0,82	0,82	0,82	0,81	0,82	0,84	0,83	0,83	0,83	0,83	0,83	0,83
D3-2	0,83	0,83	0,82	0,83	0,83	0,83	0,81	0,81	0,82	0,80	0,83	0,82	0,82	0,82	0,82	0,81	0,82	0,83	0,83	0,83	0,83	0,83	0,83	0,83
D3-3	0,88	0,88	0,88	0,88	0,87	0,87	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D3-4	0,83	0,83	0,82	0,83	0,83	0,83	0,81	0,81	0,82	0,80	0,83	0,82	0,82	0,82	0,82	0,81	0,82	0,83	0,83	0,83	0,83	0,83	0,83	0,83
D3-5	0,85	0,86	0,85	0,85	0,86	0,86	0,83	0,83	0,85	0,83	0,84	0,83	0,83	0,83	0,82	0,82	0,83	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-1	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-2	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,86	0,84	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-3	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,85	0,84	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-4	0,86	0,86	0,85	0,86	0,86	0,86	0,84	0,84	0,86	0,84	0,86	0,85	0,85	0,85	0,85	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86
D4-5	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,87	0,84	0,86	0,86	0,86	0,86	0,85	0,85	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D5-1	0,86	0,86	0,85	0,86	0,86	0,86	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D5-2	0,86	0,86	0,85	0,86	0,86	0,86	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D5-3	0,86	0,86	0,85	0,86	0,86	0,86	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D5-4	0,86	0,86	0,85	0,86	0,86	0,86	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,85
D5-5	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,85	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,87
D6-1	0,01	0,00	0,02	0,01	0,10	0,10			0,18		0,12	0,12	0,06	0,03	0,04	0,04	0,12	0,13	0,14	0,13	0,11	0,14	0,13	0,12
D6-2	0,01	0,00	0,03	0,01	0,11	0,11			0,23		0,12	0,11	0,06	0,03	0,04	0,04	0,21	0,15	0,15	0,18	0,14	0,14	0,14	0,13
D6-3	0,01	0,00	0,03	0,01	0,13	0,13			0,18		0,12	0,12	0,06	0,03	0,04	0,04	0,13	0,16	0,16	0,17	0,16	0,18	0,13	0,12
D6-4	0,01	0,00	0,03	0,01	0,12	0,12			0,18		0,12	0,12	0,05	0,03	0,04	0,04	0,25	0,10	0,13	0,15	0,12	0,13	0,11	0,11
D6-5	0,49	0,50	0,53	0,49	0,49	0,49	0,51	0,47	0,58		0,38	0,37	0,40	0,41	0,44	0,45	0,55	0,50	0,49	0,49	0,49	0,50	0,50	0,49
D7-1	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D7-2	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D7-3	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D7-4	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D7-5	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92

Tabla 23: Exhaustividad (del vocablo inglés *Recall*) obtenida aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,94	0,94	0,96	0,94	0,96	0,96	0,98	0,98	0,97	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D1-2	0,94	0,94	0,96	0,94	0,96	0,96	0,98	0,98	0,97	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D1-3	0,94	0,94	0,96	0,94	0,96	0,96	0,98	0,98	0,97	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D1-4	0,94	0,94	0,96	0,94	0,96	0,96	0,98	0,98	0,97	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D1-5	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D2-1	0,94	0,95	0,98	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,97	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D2-2	0,94	0,95	0,98	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D2-3	0,94	0,95	0,98	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,97	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D2-4	0,94	0,95	0,98	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D2-5	0,95	0,96	0,98	0,95	0,98	0,98	0,99	0,99	0,98	1,00	0,98	0,98	0,98	0,98	0,98	0,98	0,99	0,98	0,98	0,98	0,98	0,98	0,98	0,98
D3-1	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,97	0,99	0,97	0,97	0,97	0,98	0,98	0,98	0,98	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D3-2	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,97	0,99	0,97	0,97	0,97	0,98	0,98	0,98	0,98	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D3-3	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,97	0,99	0,97	0,97	0,97	0,98	0,98	0,98	0,98	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D3-4	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,97	0,99	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D3-5	0,95	0,95	0,99	0,95	0,98	0,98	0,99	1,00	0,98	1,00	0,98	0,98	0,98	0,99	0,99	0,99	1,00	0,97	0,97	0,97	0,98	0,98	0,98	0,98
D4-1	0,95	0,95	0,98	0,95	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D4-2	0,95	0,95	0,98	0,95	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D4-3	0,95	0,95	0,98	0,95	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D4-4	0,95	0,95	0,98	0,95	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D4-5	0,95	0,95	0,98	0,95	0,98	0,98	0,99	0,99	0,98	1,00	0,98	0,98	0,98	0,98	0,98	0,98	0,99	0,98	0,98	0,98	0,98	0,98	0,98	0,98
D5-1	0,94	0,94	0,98	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D5-2	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D5-3	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D5-4	0,94	0,94	0,97	0,94	0,97	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,97	0,97	0,97	0,97	0,97
D5-5	0,95	0,95	0,98	0,95	0,98	0,98	0,98	0,98	0,98	0,99	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98
D6-1	0,00	0,00	0,00	0,00	0,02	0,02	0,00	0,00	0,01	0,00	0,03	0,03	0,01	0,00	0,00	0,00	0,01	0,02	0,02	0,02	0,02	0,02	0,02	0,02
D6-2	0,00	0,00	0,00	0,00	0,02	0,02	0,00	0,00	0,01	0,00	0,03	0,03	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,02	0,02	0,02	0,02	0,02
D6-3	0,00	0,00	0,00	0,00	0,02	0,02	0,00	0,00	0,01	0,00	0,03	0,03	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,02	0,02	0,02	0,02	0,02
D6-4	0,00	0,00	0,00	0,00	0,02	0,02	0,00	0,00	0,01	0,00	0,03	0,03	0,01	0,00	0,00	0,00	0,01	0,02	0,02	0,02	0,02	0,02	0,02	0,02
D6-5	0,16	0,16	0,15	0,16	0,17	0,17	0,08	0,00	0,15	0,00	0,17	0,14	0,11	0,09	0,09	0,09	0,20	0,18	0,17	0,17	0,17	0,17	0,17	0,17
D7-1	0,92	0,92	0,96	0,92	0,96	0,96	0,97	0,97	0,97	0,96	0,97	0,97	0,97	0,96	0,96	0,97	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D7-2	0,92	0,92	0,96	0,92	0,96	0,96	0,97	0,97	0,97	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,96	0,96
D7-3	0,92	0,92	0,96	0,92	0,96	0,96	0,97	0,97	0,97	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96
D7-4	0,92	0,92	0,96	0,92	0,96	0,96	0,97	0,97	0,97	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,97	0,97	0,96	0,96	0,96	0,96
D7-5	0,93	0,93	0,96	0,93	0,96	0,96	0,97	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96

Tabla 24: Valor-F obtenido aplicando el conjunto de clasificadores a los conjuntos de datos discretizados de forma diferente.

Dataset	Red bayesiana (K2)	Red bayesiana (Hill Climber)	Red bayesiana (TAN)	Naïve Bayes	ANN (MLP)	ANN (MLP CS)	SVM (Polynomial)	SVM (Normalized Polynomial)	SVM (Pearson VII)	SVM (RBF)	KNN (K=1)	KNN (K=2)	KNN (K=3)	KNN (K=4)	KNN (K=5)	KNN (K=6)	C4.5	Random Forest (i=50)	Random Forest (i=100)	Random Forest (i=150)	Random Forest (i=200)	Random Forest (i=250)	Random Forest (i=300)	Random Forest (i=350)
D1-1	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D1-2	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D1-3	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D1-4	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D1-5	0,91	0,91	0,93	0,91	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,92	0,92	0,92	0,92	0,93	0,93	0,93	0,93	0,93	0,93	0,93	0,93
D2-1	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D2-2	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D2-3	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D2-4	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D2-5	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,91	0,91	0,91	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D3-1	0,88	0,88	0,89	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D3-2	0,88	0,88	0,89	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D3-3	0,88	0,88	0,89	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D3-4	0,88	0,88	0,89	0,88	0,89	0,89	0,89	0,89	0,89	0,88	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89	0,89
D3-5	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,90	0,90	0,90	0,90	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D4-1	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D4-2	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D4-3	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D4-4	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D4-5	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,91	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D5-1	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D5-2	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D5-3	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D5-4	0,90	0,90	0,91	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,91
D5-5	0,91	0,91	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,91	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
D6-1	0,00	0,00	0,00	0,00	0,02	0,02			0,04		0,05	0,04	0,02	0,01	0,01	0,01	0,02	0,03	0,04	0,03	0,03	0,04	0,03	0,03
D6-2	0,00	0,00	0,00	0,00	0,03	0,03			0,04		0,05	0,04	0,02	0,01	0,01	0,01	0,02	0,04	0,04	0,04	0,03	0,04	0,04	0,04
D6-3	0,00	0,00	0,01	0,00	0,03	0,03			0,03		0,04	0,04	0,02	0,01	0,01	0,01	0,02	0,03	0,03	0,03	0,03	0,03	0,03	0,03
D6-4	0,00	0,00	0,01	0,00	0,02	0,02			0,03		0,04	0,04	0,02	0,01	0,01	0,01	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
D6-5	0,24	0,24	0,24	0,24	0,25	0,25	0,13	0,04	0,23		0,23	0,20	0,17	0,15	0,15	0,15	0,28	0,25	0,25	0,25	0,25	0,25	0,25	0,25
D7-1	0,92	0,92	0,94	0,92	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94
D7-2	0,92	0,92	0,94	0,92	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94
D7-3	0,92	0,92	0,94	0,92	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94
D7-4	0,92	0,92	0,94	0,92	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94
D7-5	0,92	0,92	0,94	0,92	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94

Bibliografía

[AAOBC92] Langley, P., Iba, W., & Thompson, K. (1992, July). An analysis of Bayesian classifiers. In *Aaai* (Vol. 90, pp. 223-228).

[ABIIS13] Radicchi, F., & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627-637. DOI: 10.1007/s11192-013-1027-3.

[ACADIM22] de la Torre, U. (2022). As-cast ausferritic ductile iron materials obtained by a controlled cooling process (Doctoral dissertation, Universidad de Deusto).

[ACIED16] Biswas, S., Monroe, C., & Prucha, T. (2016). Analysis of published cast iron experimental data. In Proceedings of the 3rd World Congress on Integrated Computational Materials Engineering (ICME 2015) (pp. 293-303). Springer International Publishing.

[ACOC04] Bilgi, Ö. (Ed.). (2004). Anatolia: Cradle of Castings. Döktas.

[ADPIC22] OMC | ADPIC | ¿qué se entiende por ADPIC? (2022). Disponible online https://www.wto.org/spanish/tratop_s/trips_s/intel1_s.htm (accedido el 02 de agosto de 2022)

[ADQVRF11a] Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy sets Theory

field. Journal of Informetrics. 2011a, 5(1), 146–166. https://doi.org/10.1016/j.joi.2010.10.002.

[ADTO09] Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation, 29(6-7), 481-497.

[AEKMCA02] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.

[AESLC72] Piñero, J. M. L. (1972). *El análisis estadístico y sociométrico de la literatura científica*. Centro de Documentación e Informática Médica, Facultad de Medicina.

[AGBMT20] Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*, 6363-6381.

[AH20AIF22] Acuerdo casi final sobre Horizon 2020 - ARISTOS Innovation Funding SL (2022) Disponible online https://www.aristos.cat/acuerdo-casi-final-sobre-horizon-2020/ (Accedido el 18 de mayo de 2022)

[1] Curso de capacitación de gestores de proyectos europeos (2022) Disponible online https://ope.ciemat.es/OPEportal/portal.do?TR=A&IDR=1&identificador=350 (Accedido el 18 de mayo de 2022)

[AIFCA10] Poole, D. L., & Mackworth, A. K. (2010). *Artificial Intelligence: foundations of computational agents*. Cambridge University Press.

[AIID96] Koza, J.R., Bennett, F.H., Andre, D., Keane, M.A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero, J.S., Sudweeks, F. (eds) Artificial Intelligence in Design '96. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-0279-4_9

[AIMA10] Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach*. London.

[AIMLA22] Tyagi, A. K., & Chahal, P. (2022). Artificial intelligence and machine learning algorithms. In Research Anthology on Machine Learning Techniques, Methods, and Applications (pp. 421-446). IGI Global.

[AITDS98] Lin, D. (1998, July). An information-theoretic definition of similarity. In Icml (Vol. 98, No. 1998, pp. 296-304).

[AMLVRR23] Bai, R., Chen, X., Chen, Z. L., Cui, T., Gong, S., He, W., ... & Zhang, H. (2023). Analytics and machine learning in vehicle routing research. International Journal of Production Research, 61(1), 4-30.

[AMSVEP17] González, R., Barrientos, A., Toapanta, M., & Cerro, J. D. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. *Revista Iberoamericana de Automática e Informática industrial*, *14*(4), 394-405.

[ANCE22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Agenda de nuevas cualificaciones y empleos: una contribución europea de cara al pleno empleo. COM(2010) 682 final (23.11.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/COM_x2010x_682_final_x23.11.2010x.pdf/51263f5b-3d66-42fd-8f2d-1d2ee2792800 (accedido el 12 de mayo de 2022)

[AOASLT99] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), 988-999.

[APDWAR15] Altuntas, S., Dereli, T., & Kusiak, A. (2015). Analysis of patent documents with weighted association rules. Technological Forecasting and Social Change, 92, 249-262.

[APP50] Hebb, D. O. (1950). Animal and physiological psychology. *Annual review of psychology*, *1*(1), 173-188.

[AROHS10] Karbasian, H., & Tekkaya, A. E. (2010). A review on hot stamping. *Journal of Materials Processing Technology*, *210*(15), 2103-2118.

[ARWARD21] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

[ASAOTMA12] Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. Expert Systems with Applications, 39(13), 11443-11455.

[ASDMRD86] Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.

In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26).

[ASICW98] Coulter, N., Monarch, I., & Konda, S. Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science. 1998, 49, 1206–1223. DOI: 10.1002/(SICI)1097-4571(1998)49:133.3.CO;2-6

[AT14] Strawn, G. (2014). Alan Turing. *IT Professional*, *16*(1), 5-7.

[ATSHSCC12] R. Hund. Advanced tools and systems for hot stamping of components with complex shape. *CHS, Hanover* (2012)

[AUP03] Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In Proceedings of the 41st annual meeting of the association for computational linguistics (pp. 423-430).

[BA20YRS16] Heradio, R., Perez-Morago, H., Fernandez-Amoros, D., Cabrerizo, F. J., & Herrera-Viedma, E. (2016). A bibliometric analysis of 20 years of research on software product lines. *Information and Software Technology*, *72*, 1-15. DOI: 10.1016/j.infsof.2015.11.004

[BAITSR13] Cobo, M. J., Chiclana, F., Collop, A., de Ona, J., & Herrera-Viedma, E. (2013). A bibliometric analysis of the intelligent transportation systems research based on science mapping. *IEEE transactions on intelligent transportation systems*, *15*(2), 901-908. DOI: 10.1109/TITS.2013.2284756.

[BCIRCWA01] Ding, Y., Chowdhury, G.G. & Foo, S. Bibliometric cartography of information retrieval research by using co-word analysis. Information processing & management. 2001, 37(6), 817–842.

[BDAMLMI19] Reiz, A. N., de la Hoz, M. A., & García, M. S. (2019). Big data analysis y machine learning en medicina intensiva. *Medicina Intensiva*, *43*(7), 416-426.

[BDHGN16] Ophir, S. (2016). Big data for the humanities using Google Ngrams: Discovering hidden patterns of conceptual trends. First Monday.

[BFPIDI12] Back to the Future - Prediction of Incremental and Disruptive Innovations. Arianfar, Somaya & Kallenbach, Jan & Mitts, Håkan & Mäkinen, Olli & Reya, Lionel & Ahmed, Abu Shohel & Grišakov, Kristi. (2012).

[BIB12] Ardanuy, J. (2012). Breve introducción a la bibliometría. *La base de datos scopus y otros e-recursos del CBUES como instrumento de gestión de la actividad investigadora; 1*.

[BIOAL15] Belter, C. W. (2015). Bibliometric indicators: opportunities and limits. *Journal of the Medical Library Association: JMLA*, *103*(4), 219. DOI: 10.3163/1536-5050.103.4.014.

[BIQMSP10] Durieux, V. & Gevenois, P.A. Bibliometric Indicators: Quality Measurements of Scientific Publication 1. Radiology. 2010, 255(2), 342. DOI: 10.1148/radiol.09090626

[BLD14] Garner, B. A. (2014). Black's Law Dictionary, 10. Eagan, MN: Thomson West.

[BNC97] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*, 131-163.

[BRDCPA04] Dou, H. J. M. (2004). Benchmarking R&D and companies through patent analysis using free databases and special software: a tool to improve innovative thinking. World Patent Information, 26(4), 297-309.

[BTCPCMA06] Yu, W. D., Cheng, S. T., Shie, Y. L., & Lo, S. S. (2006). Benchmarking technological competitiveness of precast construction through patent map analysis. In Proceedings of International Symposium on Automation and Robotics in Construction 2006 (ISARC 2006) (pp. 3-5).

[CAAP07] Cascini, G., & Russo, D. (2007). Computer-aided analysis of patents and search for TRIZ contradictions. International Journal of Product Development, 4(1-2), 52-67.

[CAFIMLR18] Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., ... & Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. Radiology, 288(2), 318-328.

[CASIT22] Katuri, A., Salugu, S., Tharuni, G., & Gouri, C. S. (2022). Conversion of Acoustic Signal (Speech) Into Text By Digital Filter using Natural Language Processing. arXiv preprint arXiv:2209.04189.

[CBKBM06] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Aaai (Vol. 6, No. 2006, pp. 775-780).

[CBMCI90] Stefanescu, D. M. (1990). Classification and basic metallurgy of cast iron. ASM International, Metals Handbook. Tenth Edition., 1, 3-11.

[CBNC13] Cheng, J., & Greiner, R. (2013). Comparing Bayesian network classifiers. arXiv preprint arXiv:1301.6684.

[CCFD14] Mandhare, A., Banerjee, P., Bhutkar, S., & Hirwani, R. (2014). 'Click chemistry'for diagnosis: a patent review on exploitation of its emerging trends. Expert Opinion on Therapeutic Patents, 24(12), 1287-1310.

[CCGPE22] Curso de capacitación de gestores de proyectos europeos (2022) Disponible online https://ope.ciemat.es/OPEportal/portal.do?TR=A&IDR=1&identificador=350 (Accedido el 18 de mayo de 2022)

[CDCPPO17] Shen, Y. C., Lin, G. T., Lin, J. R., & Wang, C. H. (2017). A cross-database comparison to discover potential product opportunities using text mining and cosine similarity.

[CEITP68] Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. Journal of the ACM (JACM), 15(1), 8-36.

[CESIP84] Altshuller, G. S. (1984). Creativity as an exact science: the theory of the solution of inventive problems. Gordon and Breach Science Publishers.

[CGFSGCI20] Baer, W. (2020). Chunky graphite in ferritic spheroidal graphite cast iron: formation, prevention, characterization, impact on properties: an overview. International Journal of Metalcasting, 14(2), 454-488.

[CIFS95] Garfield, E. Citation Indexes for Science. Science. 1995, 122 (3159), 108-111.

[CIHFT11] Hund, R., & Braun, M. (2011, July). Continuous improvement of hot forming technology. In *3rd International Conference on Hot Sheet Metal Forming of High Performance Steel, CHS2, Kassel, Germany* (pp. 189-200).

[CIT88] Elliott, R. (1988). Cast iron technology.

[CLCWSS98] Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet sense similarity for word sense identification. wordnet, an electronic lexical database. The MIT Press.

[CMEIA18] The Challenges While Measuring Enterprise Innovative Activities - the Case from a Developing Country. Indira POPADIĆ, Jelena BOROCKI, Mladen RADIŠIĆ, Ivan ŠTEFANIĆ, Lena DUSPARA.
Tehnički vjesnik, Vol. 25 No. Supplement 2, 2018.
https://doi.org/10.17559/TV-20180507100421

[CMPTD22] Phuong, V. L. Q., Dong, N. V., Thu, T. N. M., & Khang, P. N. (2022, November). Combine Clasification Algorithm and Centernet Model to Predict Trafic Density. In International Conference on Future Data and Security Engineering (pp. 588-600). Singapore: Springer Nature Singapore.

[CMPUBM05] Van Raan, A.F.J. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics. 2005, 62(1), 133–143. DOI: 10.1007/s11192-005-0008-6.

[CNNAAP21] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

[COTEC06] Communication from the commission to the council, the European parliament, the European economic and social Committee and the committee of the regions. Putting knowledge into practice: A broad-based innovation strategy for the EU (2006) Disponible online http://archive.erisee.org/downloads/triangle/com2006_0502en01.pdf (accedido el 09 de mayo de 2022)

[CPL01] Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: a window on competition. RAND journal of economics, 129-151.

[CRAHSSS18] T. Taylor. A critical review of automotive hot stamped sheet steel from an industrial perspective. Material Science and Technology 2018, Volume 34, Pages 809-861.

[CSGGCI18] Stefanescu, D. M., Alonso, G., Larrañaga, P., De la Fuente, E., & Suarez, R. (2018). A comparative study of graphite growth in cast iron and in analogous systems. International Journal of Metalcasting, 12, 722-752.

[CSHNDT02] Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications*, *23*(3), 245-254.

[CTCOC18] Bornmann, L., Haunschild, R., & Hug, S. E. Visualizing the context of citations referencing papers published by Eugene Garfield: A new type of keyword co-occurrence analysis. Scientometrics. 2018, 114, 427–437. DOI: 10.1007/s11192-017-2591-8.

[CWATDN91] Callon, M., Courtial, J., & Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research—The case of polymer chemistry. Scientometrics. 1991, 22(1), 155– 205.

[CZEAFC21] CZ. La evolución del acero en la fabricación de carrocerías - CZ Revista técnica de Centro Zaragoza. 2021. Disponible online: https://revistacentrozaragoza.com/wp-content/uploads/2018/05/aceros_0-800x445.jpg (accedido el 12 de febrero de 2021)

[DABMP17] J. Fekete, R. Hall. Design of auto body: materials perspective. In Rana R, Singh SBBT-AS (eds). Woodhead publishing, 2017, Pages 1-18

[DABTIF22] Abuein, Q., Shatnawi, M. Q., & Ghazalat, L. (2022). Detection of Americans' Behavior toward Islam on Facebook. *Journal of ICT Research & Applications*, *16*(3).

[DAESKI13] Mousavi, H., Gao, S., & Zaniolo, C. (2013, August). Discovering attribute and entity synonyms for knowledge integration and semantic web search. In Proceedings of the 3rd International Workshop on Semantic Search Over the Web (pp. 1-4).

[DANDSSP52] Fix, E., & Hodges, J. L. (1952). Discriminatory analysis: Nonparametric discrimination: Small sample performance.

[DBDBBSD20] Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies, 1(1), 387-395.

[DCIMPT10] Shih, M. J., Liu, D. R., & Hsu, M. L. (2010). Discovering competitive intelligence by mining changes in patent trends. Expert Systems with Applications, 37(4), 2882-2890.

[DDSMLA22] Chang, V., Kandadai, K., Xu, Q. A., & Guan, S. (2022). Development of a Diabetes Diagnosis System Using Machine Learning Algorithms.

International Journal of Distributed Systems and Technologies (IJDST), *13*(1), 1-22.

[DEBCI96] Spinak, E. (1996). *Diccionario enciclopédico de bibliometría, cienciometría e informetría* (p. 245). Caracas: Unesco.

[DI21] Derwent innovation (2021). Derwent Innovations Index - Web of Science platform - LibGuides at Clarivate Analytics. Disponible online: https://clarivate.libguides.com/webofscienceplatform/dii (accedido el 26 de septiembre de 2021)

[DIM21] Dimensions (2021). Why did we build Dimensions|Dimensions. Disponible online: https://www.dimensions.ai/why-dimensions/ (accedido el 25 de septiembre de 2021)

[DL15] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[DMADK06] Yen, S. J., & Lee, Y. S. (2006). An efficient data mining approach for discovering interesting knowledge from customer transactions. Expert Systems with Applications, 30(4), 650-657.

[DMGFPPS15] Comins, J. A. (2015). Data-mining the technological importance of government-funded patents in the private sector. Scientometrics, 104(2), 425-435.

[DMIS04] Chen, S. Y., & Liu, X. (2004). The contribution of data mining to information science. Journal of Information Science, 30(6), 550-558.

[DMPMLT00] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2000). Output: knowledge representation. Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, San Francisco: Morgan Kaufmann Publishers Inc, 72.

[DMSLTIP17] Roh, T., Jeong, Y., & Yoon, B. (2017). Developing a methodology of structuring and layering technological information in patent documents through natural language processing. Sustainability, 9(11), 2117.

[DMTCRM09] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

[DNTO09] Lee, S., Yoon, B. & Park, Y. An approach to discovering new technology opportunities: Keyword-based patent map approach.

Technovation. 2009, 29 (6-7), 481–497. DOI: 10.1016/j.technovation.2008.10.006

[DNTOBP17] Song, K., Kim, K. S., & Lee, S. (2017). Discovering new technology opportunities based on patents: Text-mining and F-term analysis. Technovation, 60, 1-14.

[DNTOP17] Song, K., Kim, K. S., & Lee, S. (2017). Discovering new technology opportunities based on patents: Text-mining and F-term analysis. *Technovation*, *60*, 1-14.

[DPCC00] Oppenheim, C. (2000). Do patent citations count. The web of knowledge: A festschrift in honor of Eugene Garfield, 405-432.

[DPISME15] Agostini, L., Caviggioli, F., Filippini, R., & Nosella, A. (2015). Does patenting influence SME sales performance? A quantity and quality analysis of patents in Northern Italy. European Journal of Innovation Management, 18(2), 238-257.

[DRDSTAA86] Aude, J. S., & Kahn, H. J. (1986, June). A design rule database system to support technology-adaptable applications. In 23rd ACM/IEEE Design Automation Conference (pp. 510-516). IEEE.

[DSFPCA08] Neuhaus, C., & Daniel, H. D. (2008). Data sources for performing citation analysis: An overview. Journal of Documentation, 64(2), 193–210.

[DSFRC01] Liu, B., Hsu, W., & Ma, Y. (2001, August). Discovering the set of fundamental rule changes. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 335-340).

[DSNTO12] Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. Scientometrics, 90(2), 445-461.

[DSWOS18] Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science?. Journal of informetrics, 12(2), 430-435.

[DTEPRAI23] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, *6*.

[DTIITO08] Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. Expert Systems with Applications, 35(1-2), 124-135.

[DTT92] Durand, T. (1992). Dual technological trees: Assessing the intensity and strategic significance of technological change. Research policy, 21(4), 361-380.

[DVHIS98] Lenard, M. J., Madey, G. R., & Alam, P. (1998). The design and validation of a hybrid information system for the auditor's going concern decision. *Journal of Management Information Systems*, *14*(4), 219-237.

[DWNP19] Gupta, S., & Gupta, A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science, 161, 466-474.

[E202022] Europe 2020 (2022) Disponible online https://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf (accedido el 12 de mayo de 2022)

[ECDGRI21] European Commission, Directorate-General for Research and Innovation, (2021). Horizon Europe, the EU research and innovation programme (2021-27): for a green, healthy, digital and inclusive Europe, Publications Office. https://data.europa.eu/doi/10.2777/052084

[ECSWSD98] Burgess, C., Livesay, K. & Lung, K. (1998). Explorations in context space: Words, sentences, discourse. Disc. Proc. 25, 2–3, 211– 257.

[EDL24] Noticias - Archivos - Temas destacados - Estrategia de Lisboa (2024) Disponible online https://www.europarl.europa.eu/highlights/es/1001.html (accedido el 19 de febrero de 2024)

[EFGAIS21] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2021). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Ethics, governance, and policies in artificial intelligence*, 19-39.

[EHSTBR22] Arias, L. M., Artola, G., Gaviria-de-la-Puerta, J., & Porto-Gomez, I. (2022). Evolution of hot stamping technology: a bibliometric review. DYNA, 97(3), 308-314.

[EIETMLPI18] Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple

patent indicators. Technological Forecasting and Social Change, 127, 291-303.

[EKBFTM17] Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. Technological Forecasting and Social Change, 116, 208-215.

[EMEA08] Campbell, F. C. (Ed.). (2008). Elements of metallurgy and engineering alloys. ASM international.

[EN156111] European Committee for Standardization, Founding. Grey Graphite Cast Irons, EN 1561, 2011.

[EN156318] European Committee for Standardization, Founding. Spheroidal Graphite Cast Irons, EN 1563, 2018.

[EN156411] European Committee for Standardization, Founding. Ausferritic Spheroidal Graphite Cast Irons, EN 1564, 2011.

[ENET21] Espacenet (2021). Espacenet – patent search. Disponible online: https://worldwide.espacenet.com/ (accedido el 26 de septiembre de 2021)

[EOMEA08] Campbell, F. C. (Ed.). (2008). *Elements of metallurgy and engineering alloys*. ASM international.

[EPI22] EPI Una introducción a las patentes en Europa (2022). Disponible online: http://www.ub.edu/centredepatents/pdf/material_referencia/EPI_Una_introduccion_a_las_patentes_en_Europa.pdf (accedido el 03 de agosto de 2022)

[EPO21] EPO (2021). EPO - 40 years Timeline. Disponible online: https://www.epo.org/about-us/timeline.html (accedido el 26 de septiembre de 2021)

[ERPISPA08] Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips. R&d Management, 38(5), 550-562.

[ESAARRD08] Visentin, G. (2008). The ESA A&R technology R&D plan 2007-2009: serving European future missions. In 9th International Symposium on Artificial Intelligence, Robotics and Automation in Space.

[ESDHML19] Subasi, A., Kevric, J., & Abdullah Canbaz, M. (2019). Epileptic seizure detection using hybrid machine learning methods. *Neural Computing and Applications*, *31*, 317-325.

[ESLDMIP01] Hastie, T., Friedman, J. & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York. (pp. 437-444).

[ESLDMIP09] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[ETOPITM18] Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. Technological Forecasting and Social Change, 132, 105-117.

[ETSC95] Schmoch, U. (1995). Evaluation of technological strategies of companies by means of MDS maps. International journal of technology Management, 10(4-6), 426-440.

[ETSPSC91] Bayes, T. (1991). An essay towards solving a problem in the doctrine of chances. 1763. *MD computing: computers in medical practice*, *8*(3), 157-171.

[EUER22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Una Europa que utilice eficazmente los recursos – Iniciativa emblemática con arreglo a la Estrategia Europa 2020. COM (2011) 21 final (26.01.2011) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_4-7_COM_x2011x_21_Final_x26.01.2011x.pdf/daeb8b85-67e7-48ee-8d43-634c1e373d69 (accedido el 12 de mayo de 2022)

[EUJHE22] The EU and Japan open Horizon Europe association talks | European Commission (2022) Disponible online https://ec.europa.eu/info/news/eu-and-japan-open-horizon-europe-association-talks-2022-may-12_en (Accedido el 23 de mayo de 2022)

[FETBPA06] Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological forecasting and social change, 73(8), 981-1012.

[FETDL17] Zhou, Y., Dong, F., Li, Z., Du, J., Liu, Y., & Zhang, L. (2017). Forecasting emerging technologies with deep learning and data augmentation: convergence emerging technologies vs non-convergence emerging technologies.

[FETSLPA17] Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. Technological Forecasting and Social Change, 125, 236-244.

[FFPH21] Forming the future. Press hardening with pch flex – fast, flexible cost-effective. Disponible online: https://www.schulergroup.com/major/us/technologien/produkte/formhaerteanlagen/index.html (accedido el 21 de enero de 2024)

[FMSCI08] Berns, H., & Theisen, W. (2008). Ferrous materials: steel and cast iron. Springer Science & Business Media.

[FMSCI08] Berns, H., & Theisen, W. (2008). Ferrous materials: steel and cast iron. Springer Science & Business Media.

[FOTA08] Cagnin, C., Keenan, M., Johnston, R., Scapolo, F., & Barré, R. (2008). Future-oriented technology analysis. Strategic Intelligence for an Innovative Economy, 170.

[FPSOM06] Huysmans, J., Baesens, B., Vanthienen, J., & Van Gestel, T. (2006). Failure prediction with self organizing maps. *Expert Systems with Applications*, *30*(3), 479-487.

[FRT21] Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. Scientific reports, 11(1), 100.

[GA21] Google Académico (2021). About Google Scholar. Disponible online: https://scholar.google.com/scholar/about.html (accedido el 25 de septiembre de 2021)

[GACPRS11] Matuszek, C., Mayton, B., Aimi, R., Deisenroth, M. P., Bo, L., Chu, R., ... & Fox, D. (2011, May). Gambit: An autonomous chess-playing robotic system. In 2011 IEEE International Conference on Robotics and Automation (pp. 4291-4297). IEEE.

[GBD21] Global Brand Database (2021). WIPO Global Brand Database. Disponible online: https://www3.wipo.int/branddb/en/ (accedido el 26 de septiembre de 2021)

[GBH202022] Guía básica Horizon 2020 (2022) Disponible online http://www.caminos.upm.es/Documentos/Guia_basica_HORIZON_2020_UPM-V1.2.pdf (accedido el 16 de mayo de 2022)

[GCBCNFS02] Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European journal of operational research*, *136*(1), 190-211.

[GDD21] Global Design Database (2021). WIPO Global Design Database. Disponible online: https://www3.wipo.int/designdb/en/index.jsp (accedido el 26 de septiembre de 2021)

[GNM64] Goffman, W., & Newill, V. A. (1964). Methodology for test and evaluation of information retrieval systems (p. 0022). Center for Documentation and Communication Research, School of Library Science, Western Reserve University.

[GSSW0S16] Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics, 106, 787-804.

[GTDP06] De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In Lrec (Vol. 6, pp. 449-454).

[HAEP22] EPO - How to apply for a European patent (2022). Disponible online: https://www.epo.org/applying/basics.html (accedido el 03 de agosto de 2022)

[HANNGA07] Kim, H. J., & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing*, *7*(2), 569-576.

[HAPIR13] Lee, C., Song, B., & Park, Y. (2013). How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships. Technology analysis & strategic management, 25(1), 23-38.

[HATMPN19] Zhou, X., Huang, L., Zhang, Y., & Yu, M. (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. Scientometrics, 121(2), 699-737.

[HC16] Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

[HCL14] Rojo, G. (2014). Hispanic corpus linguistics. *The Routledge handbook of Hispanic applied linguistics*, 371-387.

[HDPHML21] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[HE24] Horizon Europe the EU's funding programme for research and innovation (2024). Disponible online: https://commission.europa.eu/funding-tenders/find-funding/eu-funding-programmes/horizon-europe_en (accedido el 19 de febrero de 2024)

[HFOHS19] Billur, E., Berglund, G., & Gustafsson, T. (2019). History and future outlook of hot stamping. Hot Stamping of Ultra High-Strength Steels: From a Technological and Business Perspective, 31-44.

[HHBSNS08] Berglund, G. (2008, October). The history of hardening of boron steel in northern Sweden. In *1st international conference on hot sheet metal forming of high-performance steel, Kassel, Germany* (pp. 175-177).

[HMACSM05] Hsieh, N. C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert systems with applications*, *28*(4), 655-665.

[HNLP18] Martin, R. L., Iraola, D. M., Louie, E., Pierce, D., Tagtow, B. A., Labrie, J. J., & Abrahamson, P. G. (2018). Hybrid natural language processing for high-performance patent and literature mining in IBM Watson for Drug Discovery. IBM Journal of Research and Development, 62(6), 8-1.

[HSAAA22] Atxaga, G., Arroyo, A., & Canflanca, B. (2022). Hot stamping of aerospace aluminium alloys: Automotive technologies for the aeronautics industry. Journal of Manufacturing Processes, 81, 817-827.

[HSBSSTP16] Merklein, M., Wieland, M., Lechner, M., Bruschi, S., & Ghiotti, A. (2016). Hot stamping of boron steel sheets with tailored properties: A review. *Journal of materials processing technology*, *228*, 11-24.

[HSMF11] Neugebauer, R., Schieck, F., Rautenstrauch, A., & Bach, M. (2011). Hot sheet metal forming: The formulation of graded component characteristics based on strategic temperature management for tool-

based and incremental forming operations. CIRP Journal of Manufacturing Science and Technology, 4(2), 180-188.

[HSMMPA16] Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis. Journal of Informetrics, 10(4), 1108-1130.

[HSUHSSP17] Mori, K.; Bariani, P.F.; Behrens, B.A.; Brosius, A.; Bruschi, S.; Maeno, T.; Merklein, M.; Yanagimoto, J. Hot stamping of ultra-high strength steel parts. CIRP Annual Manufacturing Technology. 2017, 66, pp 681-686

[IAHSS04] Schaeffler, D. (2004). Introduction to advanced high-strength steels-Part I. *Stamping journal*, *16*, 22-28.

[IBC08] Introducing Booz & Company [J]. Strategy+Business. Banerji, Shumeet. (2008).

[IBIPA24] Inicio - Balder IP Abogados. Disponible online https://balderip.com/es/ (accedido el 13 de febrero de 2024)

[ICDCNN12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

[ICH81] Walton, C. F., & Opar, T. J. (1981). Iron castings handbook: covering data on gray, malleable, ductile, white, alloy, and compacted graphite irons. (No Title).

[ICIIS00] Norton, M. Introductory concepts in information science. Asis Monograph Series. 2000. Medford, NJ: Information Today.

[ICPD19] Caesar, A.G., English: Iron-carbon phase diagram under atmospheric pressure, 2019. Wikimedia Commons. Disponible online https://commons.wikimedia.org/wiki/File:Iron_carbon_phase_diagram.svg (accedido el 06 de marzo de 2024).

[ICPM19] Govindarajan, U. H., Trappey, A. J., & Trappey, C. V. (2019). Intelligent collaborative patent mining using excessive topic generation. Advanced Engineering Informatics, 42, 100955.

[IETT11] Yoon, J., & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics, 88(1), 213-228.

[IFPA91] Brockhoff, K. K. (1991, October). Indicators of firm patent activities. In Technology management: the new international language (pp. 476-481). IEEE.

[IFSSP94] John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994* (pp. 121-129). Morgan Kaufmann.

[IILRHR99] Schaal, S. (1999). Is imitation learning the route to humanoid robots?. *Trends in cognitive sciences*, *3*(6), 233-242.

[IIS97] Gallouj, F., & Weinstein, O. (1997). Innovation in services. Research policy, 26(4-5), 537-556.

[ILSA98] Landauer, T., Foltz, P. & Laham, D. (1998). Introduction to latent semantic analysis. Dis. Proc. 25, 2–3, 259–284.

[IMDTET19] Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. Technological Forecasting and Social Change, 146, 687-705.

[IMIR83] Salton, G., McGill, M.J. (1983). Introduction to Modern Information Retrieval. New York: McGraw Hill.

[IML20] Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

[IMLNNDL20] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology*, *9*(2), 14-14.

[INNSM20] Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2020). An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, *76*, 2098-2118.

[IPCPD11] Jun, S. (2011). IPC code analysis of patent documents using association rules and maps–patent analysis of database technology. In Database Theory and Application, Bio-Science and Bio-Technology (pp. 21-30). Springer, Berlin, Heidelberg.

[IPFNAP11] Yoon, J., Choi, S., & Kim, K. (2011). Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. Scientometrics, 86(3), 687-703.

[IPISAO12] Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. Scientometrics, 90(2), 515-529.

[IPVMP04] Reitzig, M. (2004). Improving patent valuations for management purposes—validating new indicators by analyzing application rationales. Research policy, 33(6-7), 939-957.

[IQISRO05] Hirsch, J. An index to quantify an individuals scientific research output. Proceedings of the National Academy of Sciences. 2005, 102, 16569–16572. DOI: 10.1073/pnas.0507655102

[IRDSAA92] Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). Information retrieval: data structures and algorithms. Prentice-Hall, Inc..

[IRSS63] Swets, J. A. (1963) Information retrieval Systems Science, 141 (3577): p. 245-250

[ISBC94] Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Uncertainty Proceedings 1994* (pp. 399-406). Morgan Kaufmann.

[ISBC94] Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Uncertainty Proceedings 1994* (pp. 399-406). Morgan Kaufmann.

[ISLWAR13] Witten, D., & James, G. (2013). *An introduction to statistical learning with applications in R*. springer publication.

[ISO1611216] International Organization for Standardization, Compacted (Vermicular) Cast Irons-Classification, ISO Standard 16112, 2016.

[ITMT13] Mousavi, H., Gao, S., & Zaniolo, C. (2013). Ibminer: A text mining tool for constructing and populating infobox databases and knowledge bases. Proceedings of the VLDB Endowment, 6(12), 1330-1333.

[IUPI03] Jung, S. (2003). Importance of using patent information. In WIPO—Most intermediate training course on practical intellectual property issues in business, organized by the World Intellectual Property Organization (WIPO), Geneva, November 10–14.

[JEM22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Juventud en Movimiento. Una iniciativa destinada a impulsar el potencial de los jóvenes para lograr un crecimiento inteligente, sostenible e integrador en la Unión Europea. COM (2010) 477 final

(15.09.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_2-7_COM_x2010x_477_Final_x15.09.2010x.pdf/887043ce-0b83-4a08-a263-2ff98e69a431 (accedido el 12 de mayo de 2022)

[KDBNC20] Ren, H., & Wang, X. (2020). Scalable Structure Learning of K-Dependence Bayesian Network Classifier. *IEEE Access*, *8*, 200005-200020.

[KIPRIS21] KIPRIS (2021). KIPRIS (Korea Intellectual Property Rights Information Service), Free Patent Information Search Service. Disponible online: http://eng.kipris.or.kr/enghome/main.jsp (accedido el 26 de septiembre de 2021)

[KIPRISS21] KIPRIS Search (2021). KIPRIS Brochure (EN). Disponible online: http://file.kipris.or.kr/pr/pr_new/KIPRIS_brochure(en).zip (accedido el 26 de septiembre de 2021)

[KSTMPA15] Noh, H., Jo, Y., & Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. Expert Systems with Applications, 42(9), 4348-4360.

[LHKSVD98] Laham, T. K. L. D., & Foltz, P. (1998). Learning human-like knowledge by singular value decomposition: A progress report. Advances in neural information processing systems, 10, 45.

[LPANSE03] Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z. K., & Roco, M. C. (2003). Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field. Journal of nanoparticle research, 5(3), 333-363.

[LPCTCS22] List of PCT Contracting States (2022). Disponible online: http://www.wipo.int/pct/en/list_states.pdf (accedido el 03 de agosto de 2022)

[LPMTD88] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*, 9-44.

[LRBPE86] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

[LRPA14] Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information, 37, 3-13.

[LSDFLW21] IMECHE. Light, Strong, and Defect-Free Laser Welding: Perfecting the Process for the Automotive Industry. 2021. Disponible online: https://www.imeche.org/images/default-source/oscar/industry-sectors/automotive/arcelormittal_fig_01.png?sfvrsn=650ab212_2&size=705 (accedido el 12 de febrero de 2021)

[MAIPM13] Bezrukova, T. L., Morkovina, S. S., Shanin, I. I., & Popkova, E. G. (2013). Methodological approach to the identification of predictive models of socio-economic processes for investment and innovative development of enterprises. World Applied Sciences Journal, 27(11), 1443-1449.

[MAPA02] Breitzman, A. F., & Mogee, M. E. (2002). The many applications of patent analysis. Journal of information science, 28(3), 187-205.

[MARCA07] Kuo, R. J., Lin, S. Y., & Shih, C. W. (2007). Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. Expert Systems with Applications, 33(3), 794-808.

[MARLD01] Han, J., & Kamber, M. (2001). Mining Association Rules in Large Databases. Data Mining.

[MATRTM08] Yoon, B., Phaal, R., & Probert, D. (2008). Morphology analysis for technology roadmapping: application of text mining. R&d Management, 38(1), 51-68.

[MBP80] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

[MCCBRM05] Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. Expert Systems with Applications, 28(4), 773-781.

[MCCBSM01] Song, H. S., kyeong Kim, J., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. Expert systems with applications, 21(3), 157-168.

[MDM83] Sobol, I. M., & Vega, C. (1983). *Método de Montecarlo* (pp. 55-62). Mir.

[MDOC18] Kaloshin, D. N. (2018). Method for Determining the Operating Characteristics of Frequency Converter Using Interphase Power Controller Technology.

[MDUML21] Bhavsar, K. A., Singla, J., Al-Otaibi, Y. D., Song, O. Y., Zikria, Y. B., & Bashir, A. K. (2021). Medical diagnosis using machine learning: a statistical review. Computers, Materials and Continua, 67(1), 107-125.

[MEPO22] EPO - Member states of the European Patent Organisation (2022). Disponible online: https://www.epo.org/about-us/foundation/member-states.html (accedido el 03 de agosto de 2022)

[METKBA17] Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change, 114, 281-292.

[MGCGUB15] Torok, L., Pelegrino, M., Trevisan, D. G., Clua, E., & Montenegro, A. (2015). A mobile game controller adapted to the gameplay and user's behavior using machine learning. In Entertainment Computing-ICEC 2015: 14th International Conference, ICEC 2015, Trondheim, Norway, September 29-Ocotober 2, 2015, Proceedings 14 (pp. 3-16). Springer International Publishing.

[MLAI23] Hain, D., Jurowetzki, R., Lee, S., & Zhou, Y. (2023). Machine learning and artificial intelligence for science, technology, innovation mapping and forecasting: Review, synthesis, and applications. Scientometrics, 128(3), 1465-1472.

[MLAOFD19] Minastireanu, E. A., & Mesnita, G. (2019). An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection. Informatica Economica, 23(1).

[MLFMI17] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505-515.

[MLIM15] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920-1930.

[MLIM19] Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. BMC medical research methodology, 19, 1-18.

[MLIRC19] Lozano Colomer, C., & Martínez de Ibarreta Zorita, C. (2019). Machine Learning I: Regresión y clasificación/Machine Learning I: Regression and classification.

[MLMNID18] Alkasassbeh, M., & Almseidin, M. (2018). Machine learning methods for network intrusion detection. arXiv preprint arXiv:1809.02610.

[MLMSTSF19] Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

[MLS55] Allen, K., Berry, M. M., Luehrs Jr, F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. American Documentation (pre-1986), 6(2), 93.

[MLTPP15] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

[MOTSMS06] Cheong, S. H. (2006, July). MOT by using scientific methodology in Samsung R&D. In 2006 Technology Management for the Global Future-PICMET 2006 Conference (Vol. 1, pp. xlv-lxxxi). IEEE.

[MPEC20] Gonzalez Ciordia, B. (2020). *Mejora de la productividad en estampación en caliente mediante un nuevo método de calentamiento* (Tesis doctoral, Universidad del País Vasco).

[MPITE15] Corredoira, R. A., & Banerjee, P. M. (2015). Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity. Research Policy, 44(2), 508-521.

[MPSCIFT08] Cascini, G., & Zini, M. (2008). Measuring patent similarity by comparing inventions functional trees. In Computer-aided innovation (CAI) (pp. 31-42). Springer, Boston, MA.

[MSIPPP15] Borner, K., Theriault, T. N., & Boyack, K. W. Mapping science introduction: Past, present and future. Bulletin of the Association for Information Science and Technology. 2015, 41(2), 12–16. DOI: 10.1002/bult.2015.1720410205.

[MTICRM09] Chang, C. W., Lin, C. T., & Wang, L. Q. (2009). Mining the text information to optimizing the customer relationship management. Expert Systems with applications, 36(2), 1433-1443.

[MTPS10] Moehrle, M. (2010). Measures for textual patent similarities: a guided way to select appropriate approaches. Scientometrics, 85(1), 95-109.

[MTSME13] Knowledge trends in the subfields of manufacturing engineering at the platform of ISO/IEC standardization (2013). Živadin MICIĆ, Milica TUFEGDŽIĆ. Metalurgia internacional.

[MTSMFA13] Rahn, R., & Schruff, I. (2013). Modern tool steels for hot sheet metal forming applications. In *Proceedings of the 4th International Conference on Hot Sheet Metal Forming of High-Performance Steel, CHS2* (pp. 489-496).

[MTTOS94] Garfield, E. Scientography: Mapping the tracks of science. Current Contents: Social & Behavioral Sciences. 1994, 7, 5–10.

[MWCGCI09] Kim, S., Cockcroft, S. L., Omran, A. M., & Hwang, H. (2009). Mechanical, wear and heat exposure properties of compacted graphite cast iron at elevated temperatures. Journal of Alloys and Compounds, 487(1-2), 253-257.

[NAFTPA17] Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. Technological Forecasting and Social Change, 117, 228-237.

[NAKNN10] Sarkar, K., Nasipuri, M., & Ghose, S. (2010). A new approach to keyphrase extraction using neural networks. arXiv preprint arXiv:1004.3274.

[NAPC18] New Approaches to Predicting Cluster (2018). Gerd Meier zu Köcker, Matthias Künzel, Michael Nerger. iit perspective.

[NBSVU24] NVIDIA Blog: Supervised Vs. Unsupervised Learning. The Official NVIDIA Blog. Disponible online https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/. (Accedido el 15 de enero de 2024)

[NDAHSA18] J. Schmitt, T. Lung. New developments of advanced high strengths steels for automotive applications. *Comptes rendus physique* (2018)

[NDPMA19] Wang, J., & Chen, Y. J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. Advanced Engineering Informatics, 42, 100941.

[NGUSBS14] Taylor, T., Fourlaris, G., Evans, P., & Bright, G. (2014). New generation ultrahigh strength boron steel for automotive hot stamping technologies. Materials Science and Technology, 30(7), 818-826.

[NGVPA18] Kepuska, V., & Bohouta, G. (2018, January). Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In 2018 IEEE 8th annual computing and communication workshop and conference (CCWC) (pp. 99-103). IEEE.

[NLP01] Liddy, E. D. (2001). Natural language processing.

[NLP18] Jain, A., Kulkarni, G., & Shah, V. (2018). Natural language processing. International Journal of Computer Sciences and Engineering, 6(1), 161-167.

[NLPPP10] Bitter, C., Elizondo, D. A., & Yang, Y. (2010). Natural language processing: a prolog perspective. Artificial Intelligence Review, 33(1), 151-173.

[NLPPTD04] Cascini, G., Fantechi, A., & Spinicci, E. (2004, September). Natural language processing of patents and technical documentation. In International Workshop on Document Analysis Systems (pp. 508-520). Springer, Berlin, Heidelberg.

[NNFLCDS91] Lin, C. T., & Lee, C. S. G. (1991). Neural-network-based fuzzy logic control and decision system. IEEE Transactions on computers, 40(12), 1320-1336.

[NNFPR95] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

[NPAARM13] Jun, S. (2013). A new patent analysis using association rule mining and Box-Jenkins modeling for technology forecasting. Information-An International Interdisciplinary Journal, 16(1 B), 555-562.

[NTPI2011] Callejón, M. & García-Quevedo, J. (2011). Nuevas tendencias en políticas de innovación. Papeles de Economía Española. 176:194.

[NVACCS08] Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, *59*(1), 77-85. DOI: 10.1002/asi.20732.

[OEPM22] Oficina Española de Patentes y Marcas (2022). Disponible online: https://www.oepm.es/es/ (accedido el 03 de agosto de 2022)

[OFMTR11] Yoon, J., Lim, J., Choi, S., Kim, K., & Kim, C. H. (2011). Ontological functional modeling of technology for reusability. Expert Systems with Applications, 38(8), 10484-10492.

[OM18] Oslo manual 2018 © OECD/European Union 2018

[ONU15] Organización de las Naciones Unidas (2015) Objetivos de desarrollo sostenible. Disponible online https://www.un.org/sustainabledevelopment/es/2015/09/la-

asamblea-general-adopta-la-agenda-2030-para-el-desarrollo-sostenible/# (Accedido el 23 de mayo de 2022)

[OUOG13] Mousavi, H., Kerr, D., Iseli, M., & Zaniolo, C. (2013). Ontoharvester: An unsupervised ontology generator from free text. UCLA.

[PACTIIT05] Dou, H., Leveillé, V., Manullang, S., & JM Jr, D. (2005). Patent analysis for competitive technical intelligence and innovative thinking. Data science journal, 4, 209-236.

[PAKT02] Jung, S., & Imm, K. Y. (2002). The patent activities of Korea and Taiwan: a comparative case study of patent statistics. World Patent Information, 24(4), 303-311.

[PATB23] PatBase® (2023). Available online: https://www.patbase.com/ (Accessed on 21st May, 2022)

[PBIP05] Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005). Patent-based inventor profiles as a basis for human resource decisions in research and development. R&d Management, 35(5), 513-524.

[PBPI16] Principios básicos de la propiedad industrial. Organización Mundial De La Propiedad Intelectual (OMPI) (2016). Disponible online: https://tind.wipo.int/record/35228. DOI: 10.34667/tind.35228. ISBN 9789280525908. (accedido el 02 de agosto de 2022).

[PBUSE20] dos Santos Brito, K., & Adeodato, P. J. L. (2020, July). Predicting Brazilian and US elections with machine learning and social media data. In 2020 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

[PCAIIS18] Tefas, A., & Pitas, I. (2018). Principal component analysis. In Intelligent Systems (pp. 16-1). CRC Press.

[PCIPV07] Tuomo, N., Hermans, R., & Kulvik, M. (2007). Patent citations indicating present value of the biotechnology business.

[PCSA73] Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis* (Vol. 3, pp. 731-739). New York: Wiley.

[PCTSIP22] PCT - El sistema internacional de patentes (2022). Disponible online: https://www.wipo.int/pct/es/index.html (accedido el 03 de agosto de 2022)

[PDTF11] Trappey, C. V., Wu, H. Y., Taghaboni-Dutta, F., & Trappey, A. J. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics*, *25*(1), 53-64.

[PECP22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. La Plataforma Europea contra la Pobreza y la Exclusión Social: Un marco europeo para la cohesión social y territorial COM (2010) 758 final (16.12.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_7-7_COM_x2010x_758_Final_x16.12.2010x.pdf/46b57914-596f-48de-ac28-e2299775379f (accedido el 12 de mayo de 2022)

[PEDP21] Shcherbatov, I., Lisin, E., Rogalev, A., Tsurikov, G., Dvořák, M., & Strielkowski, W. (2021). Power equipment defects prediction based on the joint solution of classification and regression problems using machine learning methods. Electronics, 10(24), 3145.

[PHICT17] R. Neugebauer, F. Schieck, S. Polster, A. Mosel, A. Rautenstrauch, J. Schönherr, N. Pierschel. Press hardening – An innovative and challenging technology. *Arch. Civil. Mech. Eng*. 2017, *Volume 12*, Pages 113-118

[PIGBP03] Propiedad intelectual. Guía de buenas prácticas (2003). Disponible online: https://www.oepm.es/export/sites/oepm/comun/documentos_relacionados/Publicaciones/Folletos/Guia_Buenas_practicas.pdf (accedido el 02 de agosto de 2022)

[PIIC22] La propiedad intelectual, industrial y comercial | Fichas temáticas sobre la Unión Europea | Parlamento Europeo (2022). Disponible online https://www.europarl.europa.eu/factsheets/es/sheet/36/intellectual-industrial-and-commercial-property (accedido el 02 de agosto de 2022)

[PIIG22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Una política industrial integrada para la era de la globalización: poner la competitividad y la sostenibilidad en el punto de mira. COM (2010) 614 final (28.10.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_5-7_COM_x2010x_614_Final_x28.10.2010x.pdf/cedcd4de-1a21-4cc5-8bd8-c0721b1dbb19 (accedido el 12 de mayo de 2022)

[PKESTM18] Kim, J., Choi, J., Park, S., & Jang, D. (2018). Patent keyword extraction for sustainable technology management. Sustainability, 10(4), 1287.

[PKNA14] Choi, J., & Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change, 83, 170-182.

[PMLNA18] Pereira, C. G., Picanco-Castro, V., Covas, D. T., & Porto, G. S. (2018). Patent mining and landscaping of emerging recombinant factor VIII through network analysis. Nature biotechnology, 36(7), 585-590.

[PMTF12] Jun, S., Park, S. S., & Jang, D. S. (2012). Patent management for technology forecasting: A case study of the bio-industry.

[PMTT15] A Predictive Model of Technology Transfer Using Patent Analysis (2015). Jaehyun Choi, Dongsik Jang, Sunghae Jun and Sangsung Park. Sustainability — Open Access Journal

[PNPTBM62] Orbach, J. (1962). Principles of neurodynamics. Perceptrons and the theory of brain mechanisms. *Archives of General Psychiatry*, *7*(3), 218-219.

[PRML6] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

[PS21] Patentscope (2021). OMPI – Búsqueda en las colecciones de patentes nacionales e internacionales. Disponible online: https://patentscope.wipo.int/search/es/search.jsf (accedido el 26 de septiembre de 2021)

[PSIRCT94] Falconer, J. A., Naughton, B. J., Dunlop, D. D., Roth, E. J., Strasser, D. C., & Sinacore, J. M. (1994). Predicting stroke inpatient rehabilitation outcome using a classification tree approach. *Archives of Physical Medicine and Rehabilitation*, *75*(6), 619-625.

[PSN90] Heckerman, D. (1990). Probabilistic similarity networks. *Networks*, *20*(5), 607-636.

[PTSP12] ¿Qué hay que saber sobre la presentación y tramitación de las solicitudes de patente? (2012). Disponible online: https://www.oepm.es/export/sites/oepm/comun/documentos_relacionados/PDF/Manual_Solic_Patentes_Actualizado_FEB2012.pdf (accedido el 03 de agosto de 2022)

[PTTFT83] Campbell, R. S. (1983). Patent trends as a technological forecasting tool. World Patent Information, 5(3), 137-143.

[QELPI22] ¿Qué es la propiedad intelectual? (2022). Disponible online: https://www.wipo.int/about-ip/es/index.html (accedido el 02 de agosto de 2022)

[QEPI20] ¿Qué es la propiedad intelectual? Organización Mundial De La Propiedad Intelectual (OMPI) (2020). Disponible online: https://tind.wipo.int/record/44180. DOI: 10.34667/tind.44180. ISBN 9789280532241. (accedido el 02 de agosto de 2022).

[RBMSASM14] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[RDDNN06] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504-507.

[RIAP21] Banerjee Chattapadhyay, D., Putta, J., & Rao P, R. M. (2021). Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model. Buildings, 11(4), 172.

[RICCA20] Distanont, A., & Khongmalai, O. (2020). The role of innovation in creating a competitive advantage. Kasetsart Journal of Social Sciences, 41(1), 15-21.

[RLIRS13] Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238-1274.

[RLPCP05] Langford, J., & Zadrozny, B. (2005, August). Relating reinforcement learning performance to classification performance. In *Proceedings of the 22nd international conference on Machine learning* (pp. 473-480).

[RMSE22] Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development, 15(14), 5481-5487.

[RNRDS21] Arana, C. (2021). *Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales* (No. 797). Serie Documentos de Trabajo.

[RRATF03] Martino, J. P. (2003). A review of selected recent advances in technological forecasting. Technological forecasting and social change, 70(8), 719-733.

[RRFKS98] Galunic, D. C., & Rodan, S. (1998). Resource recombinations in the firm: Knowledge structures and the potential for Schumpeterian innovation. Strategic management journal, 19(12), 1193-1201.

[RSME18] Rabbi, F. (2018). A review of the use of machine learning techniques by social media enterprises. Journal of Contemporary Scientific Research (ISSN (Online) 2209-0142), 2(4).

[SACATM15] No, H. J., An, Y., & Park, Y. (2015). A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. Technological Forecasting and Social Change, 97, 181-192.

[SADTC20] AL-Sharuee, M. T., Liu, F., & Pratama, M. Sentiment analysis: dynamic and temporal clustering of product reviews. Applied Intelligence. 2020, 1-20. DOI: 10.1007/s10489-020-01668-6.

[SAEOCS15] Stefanescu, D. M. (2015). Science and engineering of casting solidification. Springer.

[SAONAP11] Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C. H. (2011). SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics, 88(3), 863-883.

[SAONAP11] Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C. H. (2011). SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics, 88(3), 863-883.

[SciMAT12] Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology. 2012, 63(8), 1609–1630. https://doi.org/10.1002/asi.22688.

[SCO21] Scopus (2021). Content - How Scopus Works - Scopus - | Elsevier solutions. Disponible online:

https://www.elsevier.com/solutions/scopus/how-scopus-works/content (accedido el 30 de marzo 2021)

[SCPD08] Huang, S. H., Ke, H. R., & Yang, W. P. (2008). Structure clustering for Chinese patent documents. Expert Systems with Applications, 34(4), 2290-2297.

[SEM98] Mokyr, J., & Strotz, R. H. (1998). The second industrial revolution, 1870-1914. Storia dell'economia Mondiale, 21945(1).

[SFDBC95] Pazzani, M. J. (1995, January). Searching for dependencies in Bayesian classifiers. In *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics* (pp. 424-429). PMLR.

[SHFGTD21] SAAB History - From Glory to Decay (Part Two). Disponible online: https://aoonuauto.com/blogs/news/saab-history-from-glory-to-decay-part-two/ 6_61a2b6dc-efb0-491d-8177-c19e751e0fc3_large.jpg (accedido el 28 de enero de 2021)

[SHSU13] Yao, Y., Meng, J. P., Ma, L. Y., Zhao, G. Q., & Wang, L. R. (2013). Study on hot stamping and usibor 1500P. Applied Mechanics and Materials, 320, 419-425.

[SLRCDP22] Sobreiro, P., Martinho, D. D. S., Alonso, J. G., & Berrocal, J. (2022). A slr on customer dropout prediction. IEEE Access, 10, 14529-14547.

[SMDM09] Sternitzke, C., & Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. Scientometrics, 78(1), 113-130.

[SNBC91] Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal, March 6–8, 1991 Proceedings 5* (pp. 206-219). Springer Berlin Heidelberg.

[SOAM12] Kohonen, T. (2012). Self-organization and associative memory (Vol. 8). Springer Science & Business Media.

[SRLD09] Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and autonomous systems*, *57*(5), 469-483.

[SRSUML20] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised

machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

[SSCSLT97] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

[SSEGAIS23] Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., ... & Isaac, W. (2023). Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

[SSIMH19] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, *62*(1), 15-25.

[STN21] STN (2021). Home | STN International. Disponible online: https://www.stn-international.com/ (accedido el 26 de septiembre de 2021)

[STSCBWS08] Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 1-25.

[SUANBC96] Kohavi, R. (1996, August). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd* (Vol. 96, pp. 202-207).

[SWLNNF18] Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, *107*, 3-11.

[TARTGT20] Woloszko, N. (2020). Tracking activity in real time with Google Trends.

[TBOBB17] Walter, L., Radauer, A., & Moehrle, M. G. (2017). The beauty of brimstone butterfly: novelty of patents identified by near environment analysis based on text mining. Scientometrics, 111(1), 103-115.

[TCOPDP08] Wanner, L., Baeza-Yates, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., ... & Zervaki, V. (2008). Towards content-oriented patent document processing. World Patent Information, 30(1), 21-33.

[TENCAP16] Van Ratingen, M., Williams, A., Anders, L., Seeck, A., Castaing, P., Kolke, R., ... & Miller, A. (2016). The european new car assessment programme: A historical review. Chinese journal of traumatology, 19(02), 63-69.

[TESS21] TESS (2021). Trademark Electronic Search System (TESS). Disponible online: https://tmsearch.uspto.gov/bin/gate.exe?f=searchss&state=4810:3nkwh5.1.1 (accedido el 26 de septiembre de 2021)

[TFATIFM04] Technology Futures Analysis Methods Working Group. (2004). Technology futures analysis: Toward integration of the field and new methods. Technological Forecasting and Social Change, 71(3), 287-303.

[TFEE01] Slocum, M. S., & Lundberg, C. O. (2001). Technology forecasting: from emotional to empirical. Creativity and Innovation Management, 10(2), 139-152.

[TFIAPD16] Caviggioli, F. (2016). Technology fusion: Identification and analysis of the drivers of technology convergence using patent data. Technovation, 55, 22-32.

[TFMMPC12] Jun, S., Park, S. S., & Jang, D. S. (2012). Technology forecasting using matrix map and patent clustering. Industrial Management & Data Systems.

[TFTBPA15] Kim, G. J., Park, S. S., & Jang, D. S. (2015). Technology forecasting using topic-based patent analysis.

[TGNLGCI09] Sommerfeld, A., & Tonn, B. (2009). Theory of graphite nucleation in lamellar graphite cast iron. International Journal of Metalcasting, 3, 39-47.

[ThIROM10] Bornmann, L., Mutz, R., & Daniel, H. D. (2010). The h index research output measurement: Two approaches to enhance its accuracy. *Journal of Informetrics*, *4*(3), 407-414. DOI: 10.1016/j.joi.2010.03.005.

[TIM19] Doloreux, D., de la Puerta, J. G., Pastor-López, I., Porto Gómez, I., Sanz, B., & Zabala-Iturriagagoitia, J. M. Territorial innovation models: to be or not to be, that's the question. Scientometrics. 2019, 120(3), 1163-1191. DOI: 10.1007/s11192-019-03181-1

[TIMM05] Lichtenthaler, E. (2005). The choice of technology intelligence methods in multinationals: towards a contingency approach. International Journal of Technology Management, 32(3-4), 388-407.

[TIRS00] Meadow, C., Boyce, B., and Kraft, D. 2000. Text Information Retrieval Systems, second ed. Academic Press.

[TLS22] The Lisbon Strategy in short (2022) Disponible online https://portal.cor.europa.eu/europe2020/Profiles/Pages/TheLisbonStrategyinshort.aspx (accedido el 09 de mayo de 2022)

[TMAPD09] Xu, Y. (2009, November). Apply text mining in analysis of patent document. In 2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design (pp. 2350-2352). IEEE.

[TMGPD19] Choi, W., Ahn, J., & Shin, D. (2019). Text mining geo-visualization of patent documents on geo-spatial big-data industry. Spatial Information Research, 27(1), 109-120.

[TMPET19] Kim, K. H., Han, Y. J., Lee, S., Cho, S. W., & Lee, C. (2019). Text mining for patent analysis to forecast emerging technologies in wireless power transfer. Sustainability, 11(22), 6240.

[TMPN04] Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research, 15(1), 37-50.

[TMTIRP17] Alves, T., Rodrigues, R., Costa, H., & Rocha, M. (2017, June). Development of text mining tools for information retrieval from patents. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 66-73). Springer, Cham.

[TNPESA10] Guglielmi, M., Williams, E., Groepper, P., & Lascar, S. (2010). The technology management process at the European space agency. Acta Astronautica, 66(5-6), 883-889.

[TODDC19] Choi, J., Jeong, B., & Yoon, J. (2019). Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications. Technological Forecasting and Social Change, 148, 119737.

[TOFCT06] Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomedical Digital Libraries, 7.

[TOI14] Lee, Y., Kim, S., Shin, J. (2014) Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis. Scientometrics 100(1), 227–244

[TPgI06] Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131-152.

[TPPQHF11] Banik, J., Lenze, F. J., Sikora, S., & Laurenz, R. (2011, June). Tailored properties–a pivotal question for hot forming. In *3rd International conference on hot sheet metal forming of high-performance steel* (pp. 13-20).

[TPRI16] Guan, J. C., & Yan, Y. (2016). Technological proximity and recombinative innovation in the alternative energy field. Research Policy, 45(7), 1460-1473.

[TPSAM07] Lo, S.S. (2007). A study on the technological positioning and strategy analysis model (TPSAM)—a case study of precast concrete technology, Dissertation in partial fulfilment of degree of Ph.D., Department of Technological Management, Chung Hua University, Hsinchu, Taiwan.

[TPTTFP12] Yoon, J., Kim, K. (2012) TrendPerceptor: a property function based technology intelligence system for identifying technology trends from patents. Expert Syst. Appl. 39(3), 2927–2938

[TRNMLR07] Lin, C. J., Weng, R. C., & Keerthi, S. S. (2007, June). Trust region newton methods for large-scale logistic regression. In *Proceedings of the 24th international conference on Machine learning* (pp. 561-568).

[TTDVND09] Choudhury, P., & Fallah, M. H. (2009, December). A technology tree based VND model for identifying the top technologies in the US renewable energy industry. In 2009 IEEE International Conference on Industrial Engineering and Engineering Management (pp. 21-25). IEEE.

[TTHRPA20] Stoffels, M. A., Klauck, F. J., Hamadi, T., Glorius, F., & Leker, J. (2020). Technology trends of catalysts in hydrogenation reactions: a patent landscape analysis. Advanced synthesis & catalysis, 362(6), 1258-1274.

[UADE22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Una Agenda Digital para Europa. COM (2010) 245 final/2 (26.08.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_3-7_COM_x2010x_245_Final-2_x26.08.2010x.pdf/44f8494c-f6fe-4d31-9cbb-8b577df7cb2e (accedido el 12 de mayo de 2022)

[UCRR16] Understanding Copyright and Related Rights (2016). Disponible online:

https://www.wipo.int/edocs/pubdocs/en/wipo_pub_909_2016.pdf (accedido el 02 de agosto de 2022)

[UCWQ06] Beitzel., Steven M. (2006). On Understanding and Classifying Web Queries (Ph.D. thesis). IIT. CiteSeerX: 10.1.1.127.634.

[UICESS95] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

[UIP16] Understanding Industrial Property (2016). Disponible online: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_895_2016.pdf (accedido el 02 de agosto de 2022)

[UKMCA20] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

[UKMCA20] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

[ULDA19] Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W., & Wang, R. (2019). Unsupervised linear discriminant analysis for jointly clustering and subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, *33*(3), 1276-1290.

[UMSRWD03] Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. In Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 Proceedings 4 (pp. 241-257). Springer Berlin Heidelberg.

[UPDFTF11] Trappey, C. V., Wu, H. Y., Taghaboni-Dutta, F., & Trappey, A. J. (2011). Using patent data for technology forecasting: China RFID patent analysis. Advanced Engineering Informatics, 25(1), 53-64.

[UPLI22] Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Iniciativa emblemática de Europa 2020. Unión por la innovación. COM (2010) 546 final (06.10.2010) (2022) Disponible online https://www.conselleriadefacenda.gal/documents/10433/33435/IE_1-7_COM_x2010x_546_Final_x06.10.2010.pdf/c5c4f38e-07e6-41cb-a643-5abb3bc21a38 (accedido el 12 de mayo de 2022)

[UPSTO21] UPSTO (2021). About us | UPSTO. Disponible online: https://www.uspto.gov/about-us (accedido el 26 de septiembre de 2021)

[UPSTOPF21] UPSTO Patent full – Text and image database (2021). US Patent Full-Text Database Boolean Search. Disponible online: https://patft.uspto.gov/netahtml/PTO/search-bool.html (accedido el 26 de septiembre de 2021)

[USPN00] Tsourikov, V. M., Batchilo, L. S., & Sovpel, I. V. (2000). U.S. Patent No. 6,167,370. Washington, DC: U.S. Patent and Trademark Office.

[VPAET08] Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. Expert systems with applications, 34(3), 1804-1812.

[VSALA94] Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033.

[VSMAI75] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[VTFBPC14] Choi, S., & Jun, S. (2014). Vacant technology forecasting using new Bayesian patent clustering. Technology Analysis & Strategic Management, 26(3), 241-251.

[VVVSL21] VOSviewer - Visualizing scientific landscapes. Disponible online: https://www.vosviewer.com/ (accedido el 28 de enero de 2021)

[WDSG13] Aschhoff, B., Licht, G., & Schliessler, P. (2013). Who drives smart growth? The contribution of small and young firms to inventions in sustainable technologies (No. 47). WWWforEurope Working Paper.

[WFRCCF13] Carnabuci, G., & Operti, E. (2013). Where do firms' recombinant capabilities come from? Intraorganizational networks, knowledge, and firms' ability to innovate through technological recombination. Strategic management journal, 34(13), 1591-1613.

[WIAI21] Bartneck, C., Lütge, C., Wagner, A., Welsh, S. (2021). What Is AI?. In: An Introduction to Ethics in Robotics and AI. SpringerBriefs in Ethics. Springer, Cham. https://doi.org/10.1007/978-3-030-51110-4_2

[WIPO21] WIPO (2021). Descubra la nueva versión del WIPO IP Portal. Disponible online: https://www.wipo.int/news/es/ipportal/2021/news_0002.html (accedido el 26 de septiembre de 2021)

[WIPO99] World Intellectual Property Organization. (1999). International Patent Classification: Guide, survey of classes, and summary of main groups (Vol. 9). World Intellectual Property Organization.

[WIPOIP21] WIPO IP portal (2021). WIPO IP Portal. Disponible online: https://ipportal.wipo.int/ (accedido el 26 de septiembre de 2021)

[WIRRA20] Porto-Gomez, I., Larreina, M., & Gaviria-de-la-Puerta, J. Does wine innovation research require ageing? A bibliometric review. Profesional De La Información. 2020, 29(6). DOI: 10.3145/epi.2020.nov.15

[WoS21] WoS (2021). Web of Science: Summary of Coverage - Web of Science platform - LibGuides at Clarivate Analytics. Disponible online: https://clarivate.libguides.com/webofscienceplatform/coverage (accedido el 30 de marzo 2021)

[WoSvsS10] Escalona, M. I., Lagar, P., & Pulgarín, A. (2010). Web of Science vs. Scopus:un estudio cuantitativo en Ingeniería Química. Anales de Documentación, 13,159–175.

[WPMAF18] Sofean, M., Aras, H., & Alrifai, A. (2018, October). A workflow-based large-scale patent mining and analytics framework. In International Conference on Information and Software Technologies (pp. 210-223). Springer, Cham.

[X24] X (2024) Disponible online https://twitter.com/?lang=es (accedido el 07 de enero de 2024)