Deusto
Universidad de Deusto
Deustuko Unibertsitatea

Berrikuntza gara

Somos innovación

Atendiendo al nuevo perfil de estudiante universitario del siglo XXI.

Experiencias y prácticas universitarias con las que dar respuesta a las demandas, intereses y motivaciones de nuestro alumnado, sus especificidades y expectativas, a la vez que se potencia el logro de los objetivos de aprendizaje.

XXI. mendeko unibertsitateko ikaslearen profil berria.

Gure ikasleen eskaera, interes eta motibazioei, beren berezitasun eta itxaropenei erantzuteko unibertsitateko esperientziak eta praktikak, aldi berean ikaskuntzako helburuak lortzeko asmoz.

Este libro recoge buenas prácticas académicas y de gestión implementadas por el profesorado de la Universidad de Deusto.

Edita: Grupo de Comunicación Loyola-Bilbao

ISBN: 978-84-271-4469-9

BUENAS PRÁCTICAS DE INNOVACIÓN Y CALIDAD

IX Jornada Universitaria de Innovación y Calidad:
"Atendiendo a un nuevo perfil de estudiante"

Título: Fomento de la conciencia de los sesgos introducidos a la hora de desarrollar algoritmos.

Profesorado: Borja Sanz Urquijo e Iker Pastor López

DATOS GENERALES

Nombre de la titulación y asignatura: Grado en Informática. Programación I.

Destinatarios: Alumnado que empieza su formación en el diseño y desarrollo de algoritmos, con el fin de que tomen consciencia del impacto social que pueden tener a la hora de diseñar algoritmos.

DESCRIPCIÓN, OBJETIVOS Y DESARROLLO METODOLÓGICO DE LA PRÁCTICA INNOVADORA

La primera asignatura de programación son el primer punto de contacto que tienen los estudiantes de los distintos grados de la Facultad de Ingeniería de la Universidad de Deusto con el desarrollo de programas informáticos. Durante esta asignatura se desarrollan cuatro competencias específicas, todas ellas centradas en distintas aproximaciones al diseño y desarrollo de soluciones informáticas, como son las estructuras de control, la creación de funciones y el diseño de programas bajo el paradigma de la programación orientada a objetos. A través del diseño y la creación de algoritmos, los alumnos obtienen los programas que dan respuesta a problemas sencillos haciendo uso de lenguajes de programación de alto nivel, trabajando distintos aspectos de la programación.

Se trata, por tanto, de una asignatura eminentemente práctica en la que los alumnos se enfrentan a la resolución de problemas sencillos en clase con el objetivo de ir adquiriendo las competencias en la creación de nuevos programas informáticos. Durante el curso, una breve explicación teórica da paso a la ejecución de programas sencillos que son posteriormente corregidos de forma conjunta. Sin embargo, en esta primera parte de la formación no se trabaja el impacto que tiene el desarrollo de algoritmos y soluciones informáticas en los usuarios y en la sociedad.

El objetivo de la práctica es empezar a trabajar con los alumnos ese sentido crítico a la hora de desarrollar los algoritmos, haciendo que incluyan el impacto social en las métricas de evaluación de los algoritmos que desarrollan desde el principio de su formación. Es por esta razón que esta actividad se lleva a cabo en los primeros cursos de los grados de informática, cuando aún están empezando a conocer las estructuras de programación más elementales.

Para ello, se ha desarrollado una sesión especial, basada en la metodología de Design Thinking o "Pensamiento de Diseño", en la que los alumnos tienen que razonar cómo el desarrollo de algoritmos puede tener impacto tanto en los usuarios como en la sociedad, perpetuando en algunos casos situaciones discriminatorias. Para ello, los alumnos deben analizar los sesgos que impactan a la hora de diseñar y la desarrollar los algoritmos. También evalúan el diseño de algoritmos desde el inicio, incluyendo el proceso de selección de variables, y analizan cómo el diseño de los algoritmos puede introducir soluciones para mitigar este tipo de problemas.

En el comienzo de la sesión se les muestra una base de datos real y pública, que contiene información sobre créditos bancarios. Esta base de datos está compuesta por 1.000 instancias reales, cada una de ellas compuestas por 20 variables, que incluyen por ejemplo los ingresos, las deudas pendientes, el sexo o estado civil de la persona que solicita el crédito. La última de las variables refleja si se concedió el crédito en la entidad bancaria.

Durante la primera parte de la sesión los alumnos tienen que desarrollar algoritmos que decidan si conceden o no créditos bancarios. Para realizarlo, deben combinar las variables disponibles como consideren adecuado (por ejemplo, si es menor de 18 años el crédito debe ser rechazado) con el fin de obtener un algoritmo que determine en qué condiciones es recomendable conceder el crédito. Tras la sorpresa inicial por la tarea, los alumnos por parejas utilizan el conocimiento de las condiciones para la concesión de préstamos para elaborar un algoritmo con las variables que disponen. El hecho de hacer la actividad por parejas favorece la discusión y la reflexión sobre qué variables son las más importantes a la hora de conceder el crédito. Por otro lado, el disponer de la base de datos con la evaluación final de cada caso, les sirve como punto de apoyo para hacer unas pruebas preliminares sobre la efectividad de sus algoritmos.

Posteriormente se hace una puesta en común de los algoritmos desarrollados. Para ello, se analizan no sólo la precisión a la hora de emular el resultado final reflejado en la base de datos, sino también las variables usadas a la hora de desarrollar. Por ejemplo, la diferencia a la hora de determinar si el género y su estado civil afectan en el resultado final para conceder un crédito.

Finalmente, se da una breve explicación sobre el funcionamiento de los algoritmos de aprendizaje automático, de cómo este tipo de algoritmos son sensibles a los distintos sesgos, ocultos en los propios datos, y que pueden perpetuar las discriminaciones existentes. Por ejemplo, el hecho de incluir el sexo o el estado civil de la persona que solicita el crédito como variables a evaluar puede hacer que los algoritmos consideren a estas variables como importantes a la hora de determinar si se concede o no el crédito, pudiendo dar lugar a perpetuar situaciones injustas. Si en la base de datos facilitada a la mayoría de las mujeres divorciadas el crédito les fue rechazado, los algoritmos basados en sistemas de aprendizaje automático pueden determinar que el sexo sea un factor importante a la hora de determinar si el crédito debe de ser concedido o no.

También se mostró como en los últimos años, existen varias investigaciones que buscan el desarrollo de algoritmos que buscan mitigar este tipo de sesgos y el resultado que producen. Para ello se mostró una plataforma de IBM que implementa algunos de estos algoritmos, haciendo uso de la base de datos que los alumnos han trabajado durante toda la sesión.

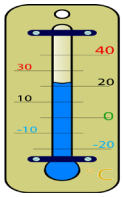
RECURSOS HUMANOS Y MATERIALES

Esta buena práctica se puede llevar a cabo únicamente por un instructor con experiencia en la materia de análisis de datos, con el equipamiento básico (ordenador con un proyector, Excel para hacer operaciones básicas sobre los datos, y un navegador web para mostrar un pequeño vídeo explicativo sobre el funcionamiento de los algoritmos de aprendizaje automático.

El conjunto de datos utilizado se puede descargar de esta URL: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29. En ella se puede obtener el conjunto de datos en distintos formatos, así como la descripción de cada una de las variables.

El vídeo utilizado se encuentra en esta dirección: https://www.youtube.com/watch?v=ukzFI9rgwfU

La demostración sobre los algoritmos para mitigar los sesgos se puede acceder en esta URL: https://aif360.mybluemix.net/data



REFLEXIÓN Y VALORACIÓN

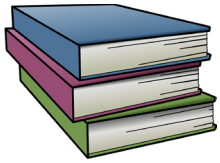
Evaluación de la Buena Práctica y lecciones aprendidas

Para la evaluación de dicha práctica se usaron dos cuestionarios iguales, uno realizado de forma previa a la sesión y otro posterior. El objetivo de estos cuestionarios era analizar si la realización de la sesión tuvo efecto a la hora de cambiar la percepción de los alumnos sobre la parcialidad de este tipo de algoritmos.

Los resultados obtenidos muestran que varios de los alumnos cambiaron de opinión a posiciones más de duda, introduciendo de esta forma un punto de reflexión que puede ser profundizado en próximas sesiones. De estos resultados cabe destacar las respuestas a la pregunta: “Los algoritmos de inteligencia artificial son más justos porque no se basan en las personas”. En el test previo a la sesión, el 46% de los alumnos mostraban una posición neutra (3 sobre 5 en una escala Likert). En el test posterior, las posiciones se distribuyeron entre las distintas posiciones, bajando el porcentaje de alumnos que se ubicaban en una posición neutra en un 26%.

Respecto a las lecciones aprendidas en el desarrollo de la práctica, los alumnos estaban muy centrados en el ámbito más técnico, el desarrollo del algoritmo, pero una

vez abierto el diálogo sobre las consecuencias de este tipo de algoritmos el debate se enriqueció enormemente. Se dio la circunstancia de que muchos de los alumnos hicieron hincapié en el hecho de que la edad era una variable determinante para otorgar un crédito. Al apuntarles que ellos mismos, dado su rango de edad, iban a ser penalizados por sus creaciones, hizo que vieran más claras las consecuencias de sus propios desarrollos. También se puso de relieve la penalización real que tienen las mujeres, en función de su estado civil, en algunas de estas situaciones, desde la propia selección de las variables.



REFERENCIAS

Marco conceptual y Referencia bibliográficas que apoyan esta buena práctica

Con la explosión de las nuevas tecnologías y el uso de los algoritmos en la vida cotidiana, su impacto es cada vez mayor en la población. Por ello, el uso cada vez mayor de sistemas para automatizar la toma de decisiones ha suscitado una gran preocupación por el hecho de que esas elecciones automatizadas puedan producir resultados discriminatorios.

Los sesgos y la parcialidad han estado intrínsecamente arraigados en la cultura y la historia desde el principio de los tiempos. Sin embargo, debido al auge de los datos digitales, ahora pueden difundirse más rápido que nunca y llegar a muchas más personas. Esto ha provocado que el sesgo en los grandes datos, también conocido como Big Data, se haya convertido en un tema de tendencia y controvertido en los últimos años. Las minorías, especialmente, han sentido los efectos dañinos del sesgo de los datos al perseguir los objetivos vitales en situaciones cada vez más gobernadas por este tipo de algoritmos, y que últimamente cubren elementos tan importantes que van desde los préstamos hipotecarios hasta la personalización de la publicidad.

Es común suponer que estos sesgos en los sistemas ocurren ya sea porque aquellos que programan este tipo de algoritmo tienen la intención de discriminar a ciertos colectivos, o bien que posee sesgos inconscientes que afectan a sus desarrollos, o porque el algoritmo mismo aprenderá a estar sesgado en base a los datos que lo alimentan (O'Neil 2016). En los últimos años se ha producido una avalancha de investigaciones sobre la imparcialidad y el sesgo tanto en los propios algoritmos como en los modelos de aprendizaje automático. Esto no es sorprendente, ya que la equidad es un concepto complejo y multifacético que depende del contexto y la cultura. Narayanan describió por lo menos 21 definiciones matemáticas de justicia encontradas en la literatura (Narayanan, 2018). Estas no son sólo diferencias teóricas en cómo medir la equidad y tienen un gran impacto, ya que diferentes las definiciones producen resultados completamente diferentes. Además de la multitud de definiciones de imparcialidad, el manejo de los diferentes sesgos de los algoritmos se aborda desde distintas partes del ciclo de vida del modelo, y la comprensión de cada contribución de la investigación, cómo, cuándo y por qué utilizarlo es un reto incluso para los expertos en la materia. Como resultado, el público en general, la comunidad científica y los profesionales de la IA necesitan claridad sobre cómo proceder.

La situación actual en las carreras STEM, acrónimo que reúne las carreras de ciencia, tecnología, ingeniería y matemáticas), no se encuentra en el mejor momento para poder dar respuesta a estos retos. Los responsables políticos en muchos países están preocupados por la escasez de graduados en el sector STEM, especialmente entre las mujeres. Las distorsiones en la percepción que se tiene de las carreras en STEM, y

más concretamente entre los géneros, pueden explicar en parte por qué las mujeres no se aplican a este tipo de carreras (Diekman et al., 2010).

Las mujeres por tanto forman parte del colectivo afectado en muchos casos por las decisiones de estos algoritmos y no están formándose para incorporarse al equipo de personas que diseñarán e implementarán este tipo de algoritmos en el futuro.

Esta buena práctica busca que los alumnos de las carreras STEAM reflexionen detenidamente sobre este tipo de situaciones y sobre cómo se pueden mitigar estas situaciones desde el propio desarrollo de este tipo de sistemas.

Referencias bibliográficas:

Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. Psychological Science, 21(8), 1051-1057.

Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA.

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.