REVISTA DE INNOVACIÓN Y BUENAS PRÁCTICAS DOCENTES

Potenciando el perfil profesional Científico de Datos mediante dinámicas de competición

Empowering the Data Scientist professional profile through competition dynamics

DAVID GUIJO-RUBIO[1], VÍCTOR VARGAS-YUN[2], ANTONIO M. DURÁN-ROSAL[3], ANTONIO M. GÓMEZ-ORELLANA[4], JAVIER BARBERO-GÓMEZ[5], JUAN CARLOS FERNÁNDEZ-CABALLERO[6], & PEDRO A. GUTIÉRREZ[7]

Fecha de recepción: 16/02/2021; Fecha de revisión: 30/03/2021; Fecha de aceptación: 05/10/2021

Cómo citar este artículo:
Guijo-Rubio, D., Vargas-Yun, V., Durán-Rosal, A.M., Gómez-Orellana, A. M., Barbero-Gómez, J., Fernández-Caballero, J.C., & Gutiérrez, P.A. (2021). Potenciando el perfil profesional Científico de Datos mediante dinámicas de competición. *Revista de Innovación y Buenas Prácticas Docentes, 10*(2), 101-116.

Autor de Correspondencia: dguijo@uco.es

Resumen:
La Ciencia de Datos es el área que comprende el desarrollo de métodos científicos, procesos y sistemas para extraer conocimiento a partir de datos recopilados previamente, con el objetivo de analizar los procedimientos llevados a cabo actualmente. El perfil profesional asociado a este campo es el del Científico de Datos, generalmente llevado a cabo por Ingenieros Informáticos gracias a que las aptitudes y competencias adquiridas durante su formación se ajustan perfectamente a lo requerido en este puesto laboral. Debido a la necesidad de formación de nuevos Científicos de Datos, entre otros fines, surgen plataformas en las que éstos pueden adquirir una amplia experiencia, como es el caso de *Kaggle.* El principal objetivo de esta experiencia docente es proporcionar al alumnado una experiencia práctica con un problema real, así como la posibilidad de cooperar y competir al mismo tiempo. Así, la adquisición y el desarrollo de las competencias necesarias en Ciencia de Datos se realiza en un entorno altamente motivador. La realización de actividades relacionadas con este perfil ha tenido una repercusión directa sobre el alumnado, siendo fundamental la motivación, la capacidad de aprendizaje y el reciclaje continuo de conocimientos a los que se someten los Ingenieros Informáticos.

Palabras clave: competición; experiencia profesional; inteligencia artificial; perfil profesional

Abstract:
Data Science is the area that comprises the development of scientific methods, processes, and systems for extracting knowledge from previously collected data, aiming to analyse the procedures being carried out currently. The professional profile associated with this field is the Data Scientist, generally carried out by Computer Engineers as the skills and competencies acquired during their training are perfectly suited to what this job requires. Due to the need for training new Data Scientists, among other goals, there are different emerging platforms where they can acquire extensive experience, such as Kaggle. The main objective of this teaching experience is to provide students with practical experience on a real problem, as well as the possibility of cooperating and competing at the same time. Thus, the acquisition and development of the necessary competencies in Data Science are carried out in a highly motivating environment. The

[1] Universidad de Córdoba, dguijo@uco.es
[2] Universidad de Córdoba, vvargas@uco.es
[3] Universidad Loyola, amduran@uloyola.es
[4] Universidad de Córdoba, am.gomez@uco
[5] Universidad de Córdoba, jbarbero@uco.es
[6] Universidad de Córdoba, jfcaballero@uco.es
[7] Universidad de Córdoba, pagutierrez@uco.es

development of activities related to this profile has had a direct impact on the students, being fundamental the motivation, the learning capacity and the continuous recycling of knowledge to which Computer Engineers are subjected.

Key Words: artificial intelligence; career profile; competition; job experience

1. INTRODUCCIÓN

La Ciencia de Datos (Van Der Aalst, 2016) es uno de los campos con mayor presencia en la actualidad, ganando especial interés en los últimos años debido a su gran interdisciplinaridad. Son numerosos los campos que se benefician de los avances de la Ciencia de Datos, como por ejemplo la meteorología (Guijo-Rubio *et al*., 2020) o las energías renovables (Dorado-Moreno *et al*., 2020), entre otras áreas científicas. Este novedoso campo de la ciencia se basa en el desarrollo de métodos científicos, procesos y sistemas para extraer conocimiento a partir de conjuntos de datos existentes y así poder identificar, analizar, entender, o incluso, mejorar los procesos actuales llevados a cabo en estos campos.

Como cada vez que surge un área de conocimiento nuevo, hay que definir cuáles son los perfiles profesionales asociados a ella. En este sentido, el perfil profesional atribuido a las personas que trabajan la Ciencia de Datos se conoce como Científico de Datos. Su principal objetivo no es otro que comprender y analizar los fenómenos reales que suceden, empleando técnicas y teorías extraídas de varios campos. Más concretamente, el perfil del Científico de Datos está relacionado con conocimientos en matemáticas, estadística y lenguajes de programación, teniendo una experiencia práctica en el análisis de datos reales y en la elaboración de modelos predictivos. Esta disciplina, hasta el momento, ha sido desarrollada mediante diversos perfiles profesionales, tales como matemáticos, estadísticos, etc., aunque en los últimos años, ha sido el perfil de Ingeniero Informático el que más se ha ajustado, dados los conocimientos y competencias en las áreas anteriormente mencionadas.

Ante esta situación surgen plataformas como *Kaggle* (https://www.kaggle.com), fundada en 2010, con el principal objetivo de poner en contacto a empresas, cuya principal función es proporcionar datos para resolver un determinado problema, con investigadores y Científicos de Datos, los cuales proporcionan soluciones a dichos problemas por medio de técnicas existentes en la Ciencia de Datos. Más concretamente, muchas de estas técnicas están basadas en el aprendizaje automático, el campo de la IA encargado del desarrollo de algoritmos y metodologías capaces de generalizar comportamientos a partir de información previa existente. Por otro lado, *Kaggle* también se utiliza frecuentemente por las empresas para reclutar nuevos empleados atendiendo a su capacidad para resolver problemas reales. El principal formato de participación en esta plataforma es mediante dinámicas de competiciones, en las cuales las soluciones propuestas por los Científicos de Datos al problema que se aborda se evalúan con respecto al rendimiento y la eficacia de las mismas.

Debido a su enorme interdisciplinaridad, el perfil de Científico de Datos es uno de los perfiles más solicitados en la actualidad, por lo que las universidades deben proporcionar competencias no solo de carácter teórico, sino también práctico, lo cual es difícil dada la amplitud de competencias a satisfacer en el Grado en Ingeniería Informática. En este sentido, el principal objetivo que se persigue con el presente proyecto es complementar los conocimientos, los cuales son principalmente teóricos, impartidos en varias asignaturas del Grado de Ingeniería Informática, con una experiencia totalmente práctica, en la que el alumnado de diferentes asignaturas se beneficie de profesionales formados en este tipo de tareas, dado que es el ámbito de investigación de los docentes asociados a estas asignaturas (Kross & Guo, 2019). De esta forma, el alumnado podrá aplicar una gran variedad de técnicas presentadas durante el desarrollo de las asignaturas, y así, enfrentarse a la complejidad que supone un problema real.

Adicionalmente, y de forma transversal a la realización del proyecto, otra competencia a desarrollar es la capacidad de realizar el trabajo de forma remota, también conocido como teletrabajo. Actualmente, debido a la situación sanitaria

producida por la COVID-19 a nivel mundial, la capacidad de trabajar a distancia ha tenido un impacto significativo en muchos puestos de trabajo, teniendo que ser adaptados en gran medida. En Jandrić et al. (2020), se exponen numerosos testimonios que tuvieron lugar durante la pandemia producida por la COVID-19, la cual provocó que muchos docentes se trasladaran a la educación *online*, siendo ésta la única vía posible en muchos de los casos y convirtiéndose en una experiencia sin precedentes en muchas áreas de conocimiento. Sin embargo, en el caso de los Científicos de Datos, esta capacidad debe ser considerada primordial, dado que se disponen de las herramientas necesarias para poder realizar las labores trabajando desde casa, y constituye en sí una importante ventaja competitiva frente a otros puestos de trabajo o perfiles profesionales, donde la presencialidad es requisito *sine qua non*.

Por lo tanto, se ha utilizado *Kaggle* como herramienta TIC en una asignatura del Grado en Ingeniería Informática que oferta la Universidad de Córdoba, de tal forma que el aprendizaje, fundamentalmente teórico, recibido en estas asignaturas, se vea complementado y aplicado en un problema eminentemente práctico. Básicamente, el alumnado se va a enfrentar al flujo de trabajo estándar que se lleva a cabo cuando se aborda un problema de este tipo, estando compuesto por: preprocesamiento de las bases de datos, diseño y entrenamiento de los modelos, validación de los modelos obtenidos, ajuste de parámetros y, finalmente, la evaluación de los modelos predictivos. Esta experiencia de innovación se enmarca en el desarrollo de un proyecto de innovación docente desarrollado por los autores de este trabajo. El proyecto de innovación docente fue otorgado por la Universidad de Córdoba, ajustándose a una de las líneas de acción prioritarias establecidas en la modalidad I del Plan de Innovación y Buenas Prácticas Docentes de la Universidad de Córdoba: transferencia del conocimiento teórico a la práctica.

Previamente a la competición de *Kaggle*, el profesorado asociado a la asignatura, así como compañeros expertos en la materia, tutorizaron, guiaron y prestaron su experiencia en el campo del aprendizaje automático y la Ciencia de Datos a la hora de enfrentarse a un problema real y proporcionaron una guía de cómo afrontar la competición. La principal finalidad de estos talleres fue la de concretar e impartir una formación específica sobre el lenguaje de programación *Python* (https://www.python.org) y la biblioteca *Scikit-Learn* (Pedregosa *et al*., 2011), orientando ambos talleres a la Ciencia de Datos. Es importante resaltar el uso de estas herramientas en el aprendizaje automático, siendo *Python* el lenguaje más utilizado para estas tareas, por su enorme versatilidad, curva de aprendizaje y facilidad de uso, entre otras características (Stack Overflow, 2020). Además, el alumnado utilizó durante todo el proyecto de innovación únicamente *software* libre, anexionándonos de esta forma a la línea de actuación a favor de éste marcada por la Conferencia de Rectores de las Universidades Españolas (CRUE), y contando con el apoyo y la colaboración del Aula de *Software* Libre de la Universidad de Córdoba (https://www.uco.es/aulasoftwarelibre/), incorporando de esta forma las innumerables ventajas otorgadas por el uso de herramientas abiertas (Stallman, 2021). Por otro lado, se utilizó el servicio *InClass Competitions* en *Kaggle*. Este servicio se proporciona de forma gratuita a la comunidad docente e investigadora para uso con estudiantes. Su principal finalidad es facilitar al profesorado en Ciencia de Datos el desarrollo de una competición exclusiva para el alumnado, de forma tal que el nivel esté adaptado a este y no se permita el acceso al resto de la comunidad de *Kaggle*. Los resultados de la competición fueron altamente satisfactorios, no solo por el incremento de conocimientos por parte del alumnado, los cuales aumentaron considerablemente tras la realización de las actividades, sino también por la alta motivación con la que el alumnado acudió a los seminarios y con el nivel y la ratio de participación en las competiciones.

El resto del artículo se organiza de la siguiente forma: la Sección 2 muestra los

antecedentes acerca del perfil profesional en Ciencia de Datos, experiencias docentes previas en dicha área y la finalidad de la utilización de *Kaggle* en este proyecto. La Sección 3 expone las principales actividades llevadas a cabo a lo largo de la ejecución del proyecto. La Sección 4 muestra los resultados conseguidos y, finalmente, la Sección 5 cierra el artículo con las conclusiones extraídas del mismo.

2. ANTECEDENTES

2.1 Ciencia de Datos

La Ciencia de Datos es considerada, frecuentemente, la continuación de otros campos de análisis de datos previos como la estadística, la minería de datos o la analítica predictiva. Este novedoso campo de la ciencia ha sufrido un enorme crecimiento durante la década previa, motivado en gran medida por la puesta en práctica de sus técnicas por numerosas empresas, dotándolas de una interesante ventaja competitiva. En la actualidad, es frecuente encontrarse con procesos organizativos en cualquier sector e industria, que aplican estas técnicas (Brynjolfsson *et al*., 2018). Tamaki (2021) cifró en un 20% extra el beneficio, tanto económico como a nivel de ventajas, que obtenía una empresa que había aplicado estas técnicas frente a sus competidores, los cuales continuaban aplicando métodos tradicionales. Es interesante saber que no solo son las empresas del sector privado quienes buscan una mejora en sus procesos, sino que también los gobiernos han acercado posturas hacia la creación de estrategias gubernamentales acerca de la inteligencia artificial (IA), uno de los campos principales de la Ciencia de Datos. A modo de ejemplo, el Gobierno de España desarrolló en 2019 una estrategia nacional con la finalidad de «[...] alinear las políticas nacionales destinadas a fomentar el desarrollo y el uso de la IA en España, aumentando la inversión, reforzando la excelencia en tecnologías y aplicaciones de IA y fortaleciendo la colaboración entre el sector público y privado, de manera que se produzca un impacto significativo en la sociedad y la economía española.» (Gobierno de España, 2021). Con lo que se denota el interés latente en esta ciencia.

No sorprende que *LinkedIn* denominase la profesión de Científico de Datos como el trabajo que más creció durante 2017 (Columbus, 2017), ni tampoco que *Glassdoor* la catalogase como el mejor trabajo en los Estados Unidos en 2018 (Jackson, 2018). Por lo tanto, se puede decir que el perfil profesional de Científico de Datos ha venido para quedarse, dado que esta tendencia continúa al alza, estimándose que para 2026 serán creados 11,5 millones de empleos según el *U.S. Bureau of Labor Statistics* (Bureau, 2020), lo que resulta impactante dado que en 2020 fueron proyectados 2,7 millones de puestos de trabajo existentes relacionados con la figura de Científico de Datos, solo en Estados Unidos, según IBM (2017). Con estas cifras, se denota la enorme demanda que existe en la actualidad y que está prevista que aumente en los próximos años.

2.2 Docencia en Ciencia de Datos

En los últimos años, son varios los autores que han presentado diversas formas de enseñar Ciencia de Datos a estudiantes del Grado en Ingeniería Informática. Entre otros, Bunner & Kim (2016) presentaron en su artículo una experiencia docente en la que, basándose en el lenguaje de programación *Python*, introducían conceptos relacionados con la Ciencia de Datos a estudiantes sin experiencia previa. Entre otros conceptos hicieron énfasis en el preprocesamiento, la presentación o visualización de los datos y el análisis estadístico y probabilístico, sin profundizar en ninguna de estas áreas. Ramamurthy (2016) detalla una experiencia con el diseño y la implementación del currículum de Ciencia de Datos y la preparación del profesorado en la Universidad de Buffalo (Nueva York, Estados Unidos) con la finalidad de crear una hoja de ruta para los centros que deseen introducir la Ciencia de Datos en su plan de estudios.

Por otro lado, Hicks & Irizarry (2018) analizan su experiencia docente basada principalmente en dos prismas. Por un lado, estudian la importancia del pensamiento estadístico, el cual consideran fundamental y, por otro lado, estudian las tres habilidades más relacionadas con Ciencia de Datos: creación, conexión y computación. Además, en el libro *Data Science for Undergraduates: Opportunities and Options* (National Academies of Sciences, Engineering and Medicine, 2018), se detalla una nueva perspectiva sobre la educación en Ciencia de Datos para estudiantes, enfocada principalmente a la gran demanda de profesionales que se estima para el año 2040 con motivo de la enorme cantidad de actividades que harán uso de metodologías de análisis de datos para la mejora de sus procesos. También especifican cuales deben ser los principales conceptos a impartir para dicha formación, así como la temporalización de los mismos, con la finalidad de guiar y alinear a las diferentes instituciones que educan en Ciencia de Datos.

2.3 *Kaggle*

Como se ha mencionado anteriormente, *Kaggle* es una plataforma que pone en contacto tanto a investigadores como a empresas en el área de la Ciencia de Datos. La principal función de *Kaggle*, como lugar de encuentro, es albergar concursos en los que los Científicos de Datos compiten por proporcionar la mejor solución a un determinado problema, generalmente presentado por empresas privadas, aunque también existen competiciones realizadas por congresos cuya finalidad es dar a conocer nuevas técnicas y proponer novedosas metodologías con respecto a la literatura existente. Como ejemplo de la importancia de *Kaggle* en el mundo de la Ciencia de Datos, cabe señalar que algunas empresas que hacen uso de esta plataforma son *Facebook*, *Netflix*, el banco Santander, *Google* o *Wikipedia*, lo que denota el impacto que tiene esta plataforma a nivel mundial y empresarial.

Kaggle está disponible de forma abierta y distribuida, como herramienta de *cloud computing*, lo que además proporciona una interacción entre usuarios, colaborando y combinando técnicas para alcanzar la mejor solución y proporcionar un análisis detallado y completo sobre el problema tratado, teniendo en cuenta diferentes puntos de vista. Además, enfocándonos en la Ciencia de Datos, *Kaggle* proporciona una serie de foros destinados a consultas sobre lenguajes de programación, técnicas de aprendizaje automático y herramientas de visualización de datos, entre otros.

Desde el punto de vista docente, trabajar con herramientas TIC (Tecnologías de la Información y la Comunicación) que faciliten el aprendizaje del alumnado es de enorme interés (Salinas, 2004). Además de la motivación que supone enfrentarse a un problema real, utilizando la plataforma *Kaggle*, una de las plataformas más novedosas en el campo de la Ciencia de Datos y con tanta repercusión a nivel mundial, nos proporciona un formato de competición (considerado una técnica de gamificación, Lee *et al*., 2011, Caponetto *et al.*, 2014), lo que aumentaría la motivación por alcanzar la mejor solución, trabajando y analizando las ventajas de cada uno de los métodos y visualizando previamente el conjunto de datos, entre otras tareas a considerar por el alumnado. Por otra parte, la colaboración no solo se produce a nivel de alumnado, sino también a nivel de la comunidad de Ciencia de Datos, es decir, con otros usuarios a través de los numerosos foros existentes, comentados previamente, proporcionando también competencias a nivel de trabajo en equipo y autónomo, las cuales son muy importantes en la actualidad. En definitiva, *Kaggle* es una magnífica herramienta para potenciar el desarrollo de competencias en Ciencia de Datos, y, de esta forma, complementar la base teórica impartida en las asignaturas del Grado en Ingeniería Informática.

3. DESARROLLO DE LA EXPERIENCIA DE INNOVACIÓN

El proyecto de innovación docente se ha enfocado a una asignatura concreta.

3.1 Objetivos

En esta sección se especifican los objetivos principales asociados a la realización del proyecto de innovación docente:

- Mejorar la adquisición de aspectos prácticos de las competencias específicas del perfil profesional Científico de Datos, en una asignatura del Grado en Ingeniería Informática, en la mención de Computación.
- Instruir al alumnado en el uso de la plataforma *Kaggle* como un medio para el aprendizaje y motivación para que, de una manera amena y divertida, se compruebe la aplicabilidad que tienen los conceptos de aprendizaje automático que se estudian en varias asignaturas del Grado en Ingeniería Informática, aunque profundizados en la asignatura a la que se aplicó el proyecto de innovación.
- Ampliar los conocimientos sobre los lenguajes de programación y la Ciencia de Datos, dentro del marco de varias sesiones prácticas en la asignatura involucrada. Concretamente, se utilizaron competiciones con fines docentes, organizadas por el profesorado de dicha asignatura, estableciendo pautas básicas para utilizar las herramientas previamente estudiadas en dicha competición. Es decir, la experiencia se realizó en un entorno controlado y limitando el grado de dificultad.
- Fomentar un ambiente tanto de cooperación como de competitividad entre el alumnado, con el objetivo de alcanzar los mejores resultados en la resolución de los problemas. Cooperativo debido a que tienen la opción de trabajar en grupo, y competitivo, ya que se establece una competición para demostrar quién es capaz de alcanzar los mejores resultados, de forma justa y siguiendo los principios de una competitividad sana.
- Enfrentar a los alumnos a la complejidad real de los problemas de aprendizaje automático.
- Cubrir parte de las competencias CTEC4, CTEC5 y CTEC7 de la asignatura del Grado en Ingeniería Informática en la que se aplicó el proyecto de innovación docente. Dichas competencias son:
  - CTEC4: Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación.
  - CTEC5: Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.
  - CTEC7: Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.

3.2 Materiales y métodos

Los objetivos anteriores se llevaron a cabo mediante las actividades que se describen a continuación.

3.2.1 Actividad 1. Búsqueda del problema real.

Con el objetivo de motivar al alumnado, se utilizó un problema real abordado previamente por miembros del equipo de investigación, demostrando el interés que existe por este problema. Concretamente, el problema consiste en la predicción de la altura de ola en una zona del Golfo de Alaska, en Estados Unidos.

El interés por predecir la altura de ola en el océano surge con la finalidad de anticiparse a eventos del medio natural que pueden llegar a ser dañinos, como los tsunamis, los cuales tienen un impacto notable principalmente en las ciudades costeras. Aparte de permitir la anticipación a accidentes, la correcta estimación de la altura de ola también es de interés en el campo de las energías renovables. En la actualidad, una de las fuentes de energía renovable con más futuro es la energía undimotriz, la cual se obtiene a partir del movimiento de las olas generadas por el viento. Dado que tanto el viento como las olas tienen una naturaleza estocástica, la estimación de la energía undimotriz que puede obtenerse en un determinado instante es una tarea compleja y, por tanto, una estimación precisa de la misma repercutiría en un mejor aprovechamiento por parte de los dispositivos mecánicos encargados de convertirla en energía eléctrica, favoreciendo de este modo que sea considerada una fuente de energía estable.

El conjunto de datos se creó a partir de observaciones meteorológicas procedentes de dos fuentes de información. Por un lado, se obtuvieron mediciones registradas por sensores instalados en una boya marítima del Golfo de Alaska. Por otro lado, se utilizaron datos de reanálisis (obtenidos por un modelo de clima matemático que hace estimaciones en localizaciones de todo el mundo), provenientes del nodo de reanálisis más cercano a la localización de la boya. En definitiva, el conjunto de datos estaba formado por 18 columnas o atributos, obtenidos de las dos fuentes de información previas. Para la predicción de la altura de ola, el problema se simplificó por medio de la discretización de la variable continua, en 4 categorías, según la altura de la misma: “baja”, “media”, “moderada” y “muy alta”. De esta forma, el problema se adaptó para que pudiera tener una dificultad asumible por el alumnado de la asignatura.

3.2.2 Actividad 2. Creación de la competición.

Como se ha comentado anteriormente, el principal objetivo del proyecto de innovación es que el alumnado se enfrente a un problema de índole real, susceptible de resolverse mediante técnicas de aprendizaje automático. Para ello, se eligió en la actividad previa la predicción de la altura de ola en una boya ubicada en el Golfo de Alaska.

Dicho problema debe ser adaptado a la plataforma *Kaggle*, de tal forma que la competición sea justa y no pueda ser resuelta por medio de técnicas de fuerza bruta. Para ello, se utilizan dos conjuntos de datos con distinto fin:

- Conjunto de datos de entrenamiento, el cual es el que está disponible para que los alumnos entrenen los modelos. En este conjunto de datos se les proporciona tanto la información de los eventos (información de entrada al modelo, 17 atributos) como la variable objetivo (variable categórica de la altura de ola correspondiente a los anteriores eventos).
- Conjunto de datos de generalización, con el que se valida el modelo entrenado con los datos previos. Para este conjunto de datos únicamente se les proporciona los datos de entrada al modelo y el alumnado debe ser capaz de obtener la salida estimada para cada patrón. Estas salidas se introducen en la plataforma *Kaggle*, la cual es la encargada de ordenar a los competidores de acuerdo a alguna métrica de evaluación, en nuestro caso fue la medida *F-Measure* (Tharwat, 2020), la cual evalúa la calidad de los modelos desarrollados por el alumnado. Una vez *Kaggle* recibe las predicciones obtenidas por el

competidor, realiza dos tipos de clasificaciones:

- Pública: Está formada por el 60% de los datos del conjunto de generalización y es la clasificación o *ranking* a la que se ciñe el alumnado para ver la calidad de su modelo. Es público, por lo que todos los alumnos ven las puntuaciones de todos sus compañeros.
- Privada: Está formada por el restante 40% de los datos del conjunto de generalización. Este *ranking* no lo conoce el alumnado, ni tiene acceso al mismo hasta la finalización de la competición. De esta forma, podemos evitar que las técnicas de fuerza bruta sean exitosas.

Además, se estableció previamente la configuración de la competición, la cual puede ser consultada por el alumnado y de este modo saben cuáles son las reglas de participación en la competición. Entre otros, se configuró la métrica para evaluar los modelos (en este caso la medida *F-Measure*) o el número de entregas máximo a realizar en 24 horas (o número de evaluaciones según el *ranking* público).

3.2.3 Actividad 3. Desarrollo de una práctica para la asignatura.

El profesorado de la asignatura, también participante tanto en el proyecto de innovación de la Universidad de Córdoba, como en el presente artículo, desarrolló una práctica de varias sesiones para la asignatura en cuestión a la que se aplicó el proyecto. El principal objetivo de esta práctica fue el de introducir al alumnado en la plataforma *Kaggle*, explicando tanto el funcionamiento de la plataforma, como el problema real al que se iban a enfrentar. También se ejemplificó el envío de resultados y se proporcionaron indicaciones de cómo se establecen los *rankings* especificados en la Actividad 2.

De esta forma, el equipo de trabajo redactó una práctica de 3 sesiones de duración (6 horas, 2 horas por sesión de clase). Para ello, se hizo uso de las Actividades 1 y 2, detalladas previamente y se crearon varios casos de estudio para proporcionarlos como ejemplos. Además, se facilitaron y realizaron varios códigos de muestra, de forma tal que pudieran probar y practicar con una gran variedad de métodos, siendo así capaces de entender cómo funciona cada uno, sus ventajas e inconvenientes y el por qué unos se adaptan mejor que otros.

3.2.4 Actividad 4. Realización de las sesiones prácticas y exposición del método de evaluación.

El equipo de trabajo llevó a cabo el conjunto de sesiones estipulado para la realización de la práctica, explicando todos los contenidos especificados en la Actividad 3. También se realizó una exposición del método de evaluación seguido para la práctica:

1. 40% de la nota de la práctica se puntúa en función del *ranking* privado, es decir, de esta forma se tiene en cuenta la calidad del modelo desarrollado, así como de todos los pasos de preprocesamiento que se han llevado a cabo. También se evalúa que el código no presente errores de ejecución.

2. 60% de la nota de la práctica se puntúa en función del flujo de trabajo seguido por el alumnado para la obtención de los modelos: preparación y visualización de los datos, eliminación de *outliers* o valores extremos, estudio de correlaciones, transformación o estandarización, selección de características, etc. Por lo tanto, en este bloque se evalúa el análisis realizado por el alumnado, independientemente de la precisión de sus modelos. Así mismo, también se evalúa la limpieza del código utilizado y si existen comentarios que expliquen el flujo de acciones.

La competición se puso en marcha una vez realizada la primera sesión y se cerró la semana siguiente a la finalización de la última sesión de prácticas, de forma tal que

se aprovecharon también las sesiones de prácticas para la resolución de dudas.

3.2.5 Actividad 5. Creación de un equipo de trabajo con el alumnado.

Se creó un equipo de trabajo con el alumnado, el cual estuvo mentorizado por varios de los profesores que participaron en el proyecto de innovación y que son autores de este artículo. Se planificaron numerosas reuniones periódicas en las que se establecieron distintas estrategias (preprocesamiento a aplicar, esquemas de validación de los métodos de aprendizaje automático, métodos de optimización, etc.) para competir en las competiciones reales, de mayor complejidad.

Se formaron equipos de trabajo en los que los miembros del equipo trabajaron tanto de forma cooperativa con el resto de miembros del equipo, como de forma competitiva frente al resto de equipos de trabajo. Se mantuvieron con frecuencia reuniones con los equipos de trabajo en horarios de tutorías, tanto de forma presencial como de forma virtual.

3.2.6 Actividad 6. Impartición de talleres en Ciencia de Datos.

Bajo el auspicio del Aula de Software Libre de la Universidad de Córdoba, se impartieron una serie de talleres sobre Ciencia de Datos. Dichos talleres fueron organizados por los participantes en el proyecto de innovación y autores del presente artículo, ya sea como ponentes o como colaboradores.

El principal objetivo a perseguir con la realización de estos talleres es complementar la formación que recibe el alumnado de la asignatura principal en la que se ha aplicado el proyecto de innovación docente. En dicha asignatura, se enseña al alumnado a crear modelos predictivos utilizando la herramienta *software Weka* (Hall *et al.*, 2009), muy popular en la comunidad científica debido a la facilidad de uso y análisis posterior de los métodos aplicados. Adicionalmente al uso de *Weka*, se quiso añadir la posibilidad de utilizar el lenguaje de programación *Python* junto con la biblioteca *Scikit-learn*, debido a la enorme repercusión y uso a nivel mundial en el ámbito de la investigación en aprendizaje automático. Para ello, se contactó con el Aula de Software Libre de la Universidad de Córdoba, el cual se ofreció a colaborar, preparando los ordenadores de las aulas y difundiendo los talleres entre el alumnado. Concretamente se realizaron dos sesiones, con el siguiente temario:

- Visualización, aprendizaje supervisado y métodos de evaluación.
- Aprendizaje no supervisado.

3.2.7 Actividad 7. Evaluación del proyecto de innovación docente.

Con el objetivo de evaluar el proyecto de innovación docente otorgado por la Universidad de Córdoba, se realizó una encuesta al alumnado antes y después de realizar el conjunto de actividades, tratando de valorar los conocimientos prácticos de que disponían sobre tareas aplicadas de Ciencia de Datos (preprocesamiento, modelado, evaluación de rendimiento, etc.).

Más concretamente, se realizaron encuestas voluntarias de tipo test, una antes de empezar la competición y los talleres, y otra al finalizar el proyecto, tratando de valorar los conocimientos prácticos adquiridos. Ambos formularios incluían las mismas preguntas, siendo 16 de ellas de tipo test para obtener una valoración acerca del nivel del alumnado, y 4 preguntas, en escala de Likert, de autoevaluación sobre la plataforma *Kaggle*, el lenguaje de programación *Python* y la biblioteca *Scikit-learn*, con el objetivo de ratificar si eran tecnologías nuevas y eran conocidas o no por el alumnado participante en el proyecto.

4. RESULTADOS

Los resultados se dividen entre los obtenidos por medio de la Actividad 7 (Sección 3.2.7), en los que se refleja la evaluación del proyecto de innovación docente por medio de los cuestionarios cumplimentados por el alumnado, y los obtenidos de la experiencia de innovación docente, por medio de las consideraciones por parte del equipo de trabajo. Ambos grupos de resultados se detallan en las siguientes subsecciones.

4.1 Resultados obtenidos a partir de los formularios.

Como se especificó anteriormente, se realizaron los formularios en dos fases, antes de realizar el conjunto de actividades y después de su finalización. Además, el formulario estaba dividido en dos partes, una primera parte compuesta por 16 preguntas de tipo test para obtener el nivel del alumnado y 4 preguntas en escala de Likert de autoevaluación sobre las principales herramientas a tratar en el proyecto de innovación.

4.1.1 Resultados de los formularios antes del comienzo de las actividades.

En relación a la primera parte, la realización del formulario de tipo test nos permitió conocer el nivel inicial del alumnado en lo relativo al aprendizaje automático y a las labores que realiza un Científico de Datos. Cada pregunta estaba evaluada con 1 punto, pudiéndose alcanzar de esta forma un máximo de 16 puntos. La Figura 1 expone los resultados obtenidos por los alumnos previo al inicio de las actividades expuestas en el proyecto de innovación. En la Figura 1 se puede ver cómo la media es de 8.42 sobre 16, lo que es un 5.26 sobre 10. También es interesante el intervalo de puntuaciones, entre 4 y 12, lo que indica que hay alumnos cuyos conocimientos son bajos.

Normal	Valor medio	Intervalo
8,42/16 puntos	9/16 puntos	4-12 puntos

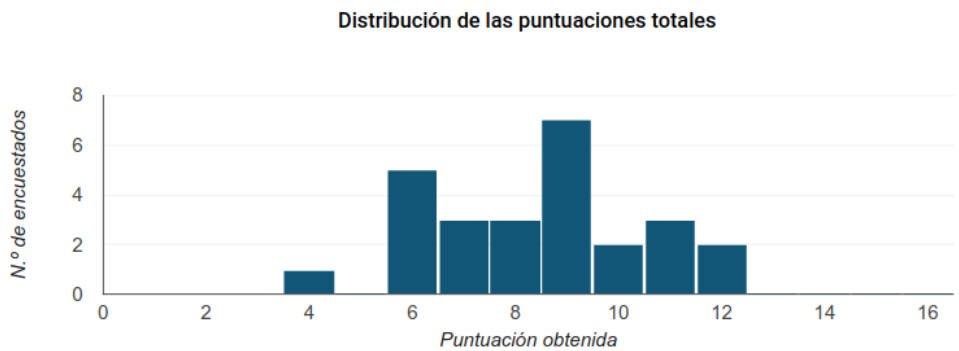

Figura 1. *Distribución obtenida a partir de los resultados de los formularios realizados previamente al inicio de las actividades organizadas por el proyecto de innovación docente. Fuente: elaboración propia.*

En relación a la segunda parte sobre la autoevaluación de las nuevas tecnologías utilizadas, los resultados se muestran en la Figura 2. En este caso, la gran cantidad de respuestas se agrupan en las posiciones 1 y 2 de la escala de Likert, es

decir, que el alumnado tenía conocimientos muy reducidos o bajos sobre este tipo de tecnologías a nivel práctico.

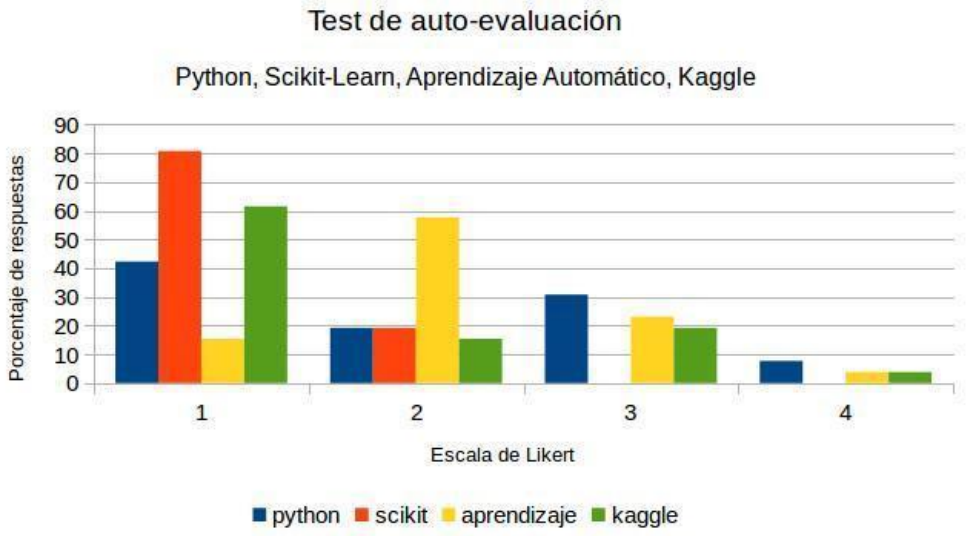

Figura 2. *Distribución obtenida a partir de las autoevaluaciones realizadas previamente al inicio de las actividades organizadas por el proyecto de innovación docente. Fuente: elaboración propia.*

4.1.2 Resultados de los formularios después de la realización de las actividades.

Con el objetivo de evaluar si la realización de las actividades organizadas en el proyecto de innovación docente tuvo un impacto sobre el alumnado, se repitió el formulario anterior (es importante mencionar que el alumnado no tuvo acceso a las respuestas del formulario anterior). Los resultados de este segundo formulario se muestran en la Figura 3. Como se puede apreciar, la media subió a 10.63 sobre 16, lo que es equivalente a 6.64 sobre 10, subiendo casi un punto y medio la calificación media anterior. En cuanto al intervalo, es interesante apreciar que se ha desplazado a un rango superior de puntuaciones, concretamente entre 7 y 14. Con estos resultados se puede concluir que el alumnado tiene mayor formación sobre los contenidos iniciales al terminar la experiencia docente.

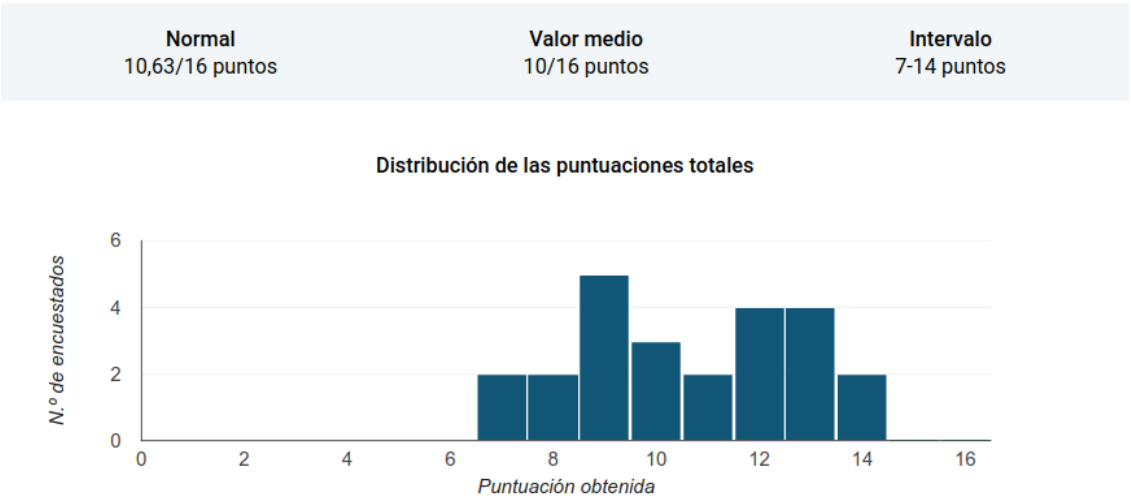

Figura 3: Distribución obtenida a partir de los resultados de los formularios realizados con posterioridad a la finalización de las actividades organizadas por el proyecto de innovación docente. Fuente: elaboración propia.

En relación a la segunda autoevaluación, que fue realizada con posterioridad a la finalización de todas las actividades y cuyos resultados están expuestos en la Figura 4, se puede apreciar cómo el alumnado considera que la realización de las actividades ha surtido efecto en su conocimiento sobre las nuevas herramientas, siendo las posiciones 3 y 4 de la escala de Likert las más consideradas. En resumen, el alumnado considera que sus conocimientos han aumentado hasta un nivel medio-alto sobre este tipo de tecnologías novedosas.

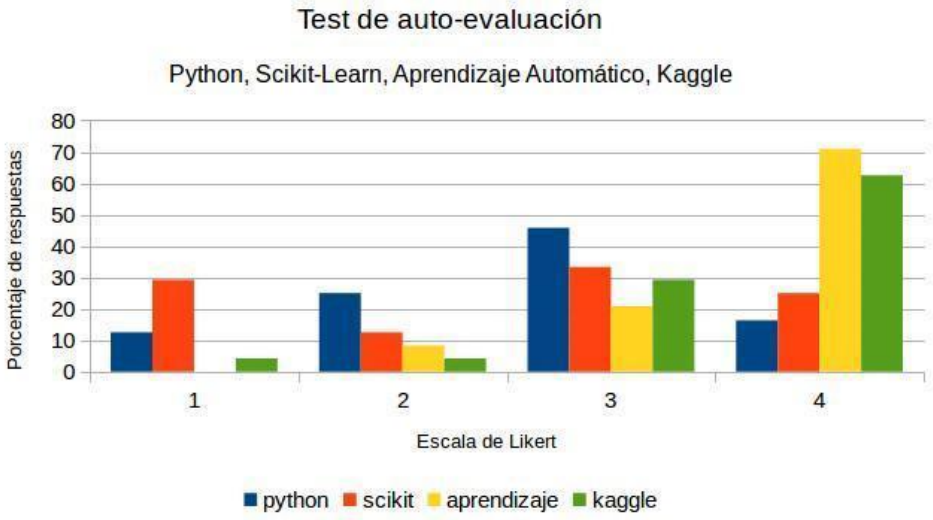

Figura 4. *Distribución obtenida a partir de las autoevaluaciones realizadas con posterioridad a la finalización de las actividades organizadas por el proyecto de innovación docente. Fuente: elaboración propia.*

4.2 Resultados de la experiencia de innovación docente.

Los resultados obtenidos de esta experiencia de innovación docente, por parte de los organizadores de la misma, son los siguientes:

- Se consiguió que el alumnado sea consciente de la importancia y el interés que tiene el perfil profesional de Científico de Datos en la actualidad, así como de la creciente demanda existente alrededor de su figura, convirtiéndose en una exitosa salida profesional para el alumnado del Grado en Ingeniería Informática.
- Se dio a conocer al alumnado las diversas tareas que realiza el Científico de Datos, entre ellas el preprocesado de datos o la visualización, con la finalidad de poder profundizar en sucesivas etapas del flujo de trabajo.
- Se dio a conocer al alumnado una serie de herramientas profesionales y de actualidad para el modelado de problemas reales mediante técnicas de aprendizaje automático, entre las que figuran los lenguajes de programación *Python* y la biblioteca *Scikit-learn*.
- Se consiguió que el alumnado aportase y recibiese nuevos conocimientos prácticos a partir de la gran comunidad existente en la plataforma *Kaggle*, formada por investigadores y Científicos de Datos de numerosos ámbitos, ampliamente consolidados.
- Además de los conocimientos recibidos en materia de lenguajes de programación y en bibliotecas asociadas (como es el caso de *Scikit-learn*) mediante la impartición de los tutoriales por parte del profesorado, se ha instruido en materia relacionada con los métodos en sí, basándonos principalmente en nociones matemáticas y estadísticas, áreas plenamente relacionadas con la Ciencia de Datos al ser precursoras de la misma.

- Se fomentó el uso de las TIC en el aula y se consiguió una buena participación en los talleres.
- Se fomentó el trabajo en grupo promoviendo la cooperatividad entre participantes de un mismo grupo y la sana competitividad entre grupos, haciendo que el *feedback* y las relaciones entre ellos culmine en una mejora de conocimiento para todos.
- Se aumentó la confianza del alumnado en sí mismo para involucrarse en problemas reales donde pueda aplicar sus conocimientos teórico-prácticos para dar una solución tangible.
- Se animó al alumnado a participar en competiciones más complejas y de ámbitos más variados, estando éstas organizadas por empresas en *Kaggle*.
- En definitiva, se mejoraron, a nivel práctico, las competencias a adquirir por el alumnado en la asignatura donde se aplicó el proyecto de innovación docente.

5. CONCLUSIONES

Como conclusión final, ligada a los resultados obtenidos tanto mediante los formularios cumplimentados por el alumnado como por la percepción del profesorado participante en esta experiencia de innovación docente, a nuestro parecer, proyectos de innovación docente como el especificado en este artículo, mejoran de manera sustancial la capacidad del alumnado para enfrentarse a una situación real laboral cuando terminen su titulación. Más concretamente, en esta experiencia de innovación docente, se trabajó con el perfil de Científico de Datos, uno de los perfiles más demandados por la sociedad en la actualidad, teniendo una gran repercusión a nivel mundial. En este sentido, el aplicar técnicas actuales que el equipo de participantes del proyecto usa en su investigación, supone darles la motivación y capacidad para el aprendizaje y reciclaje continuo de conocimientos a los que tienen que someterse los Graduados en Ingeniería Informática.

Adicionalmente, debido a esta experiencia en *Kaggle*, fueron varios los alumnos interesados en realizar Trabajos Fin de Grado relacionados con la investigación y el aprendizaje automático, profundizando, de esta forma, en los trabajos asociados a un Científico de Datos.

Además, sin lugar a dudas, consideramos que el grado de implicación y de interés de muchos alumnos fue muy alto desde el comienzo de las actividades, ocupando bastante tiempo debido a la dificultad y a la exploración de un campo más específico que lo visto hasta el momento en la asignatura.

En relación al esfuerzo del equipo de trabajo del proyecto de innovación docente, hay que reseñar que fue elevado, ya que hubo que adaptar al nivel del alumnado todo el material de los talleres y la práctica docente en la asignatura a la que se aplicó el proyecto. Sin embargo, la satisfacción conseguida hizo que mereciese enormemente la pena. Se espera poder realizar nuevas ediciones con la finalidad de incrementar el conocimiento acerca de este perfil laboral tan demandado en la actualidad y con una tendencia creciente en los próximos años.

REFERENCIAS

Bureau, E. T. 11.5 mn job openings by 2026, sky-high salaries: Why data science is booming. https://bit.ly/3tsDVwZ

Brunner, R.J. and Kim, E.J., 2016. Teaching data science. Procedia Computer Science, 80, pp. 1947-1956. https://bit.ly/3wp6TPM

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy?. In *AEA Papers and Proceedings* (Vol. 108, pp. 43-47). https://bit.ly/3tuE8Qg

Caponetto, I., Earp, J., & Ott, M. (2014). Gamification and education: A literature review. In *European Conference on Games Based Learning* (Vol. 1, p. 50). Academic Conferences International Limited. https://bit.ly/3cLFsIs

Columbus, L. LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning. [Online]. https://bit.ly/2Lt7yNG

Dorado-Moreno, M., Navarin, N., Gutiérrez, P. A., Prieto, L., Sperduti, A., Salcedo-Sanz, S., & Hervás-Martínez, C. (2020). Multi-task learning for the prediction of wind power ramp events with deep neural networks. *Neural Networks*, *123*, 401-411. https://bit.ly/2YRarL3

Gobierno de España. Estrategia Nacional para la Inteligencia Artificial. https://bit.ly/2YUfSsB

Guijo-Rubio, D., Casanova-Mateo, C., Sanz-Justo, J., Gutierrez, P. A., Cornejo-Bueno, S., Hervás, C., & Salcedo-Sanz, S. (2020). Ordinal regression algorithms for the analysis of convective situations over Madrid-Barajas airport. *Atmospheric Research*, *236*, 104798. https://bit.ly/2YVlBhP

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18. https://bit.ly/3pZPHgo

Hicks, S.C. and Irizarry, R.A., 2018. A guide to teaching data science. The American Statistician, 72(4), pp. 382-391. https://bit.ly/3wrbz81

IBM. The Quant Crunch. How the demand for data science skills is disrupting the job market. https://ibm.co/39WX6qU

Jackson, A. E. The 50 Best Jobs in America for 2018. [Online]. Último acceso: febrero 2021. https://bit.ly/3pU2eld

Jandrić, P., Hayes, D., Truelove, I., Levinson, P., Mayo, P., Ryberg, T., ... & Hayes, S. (2020). Teaching in the age of Covid-19. *Postdigital Science and Education*, *2*(3), 1069-1230. Disponible en: https://bit.ly/3fKqR1z

Kross, S., & Guo, P. J. (2019, May). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14). https://bit.ly/3cQlgVI

Lee, J. J. & Hammer, J. (2011). Gamification in Education: What, How, Why Bother? *Academic Exchange Quarterly*, 15(2). https://bit.ly/2Lm8ljc

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. National Academies Press. https://bit.ly/3mqNV70

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research*, *12*, 2825-2830. https://bit.ly/2MuivPB

Ramamurthy, B., 2016, February. A practical and sustainable model for learning and teaching data science. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 169-174). https://bit.ly/3sTpbGF

Salinas, J. (2004). Innovación docente y uso de las TIC en la enseñanza universitaria. *Revista de Universidad y Sociedad del Conocimiento (RUSC), 1*(1). https://bit.ly/2YOchML

Stack Overflow. Stack Overflow Developer Survey 2020. https://bit.ly/3juHrT0

Stallman, R. Por qué las escuelas deberían usar exclusivamente software libre. https://bit.ly/39TuCyx

Tamaki, G. La era del Big Data. https://bit.ly/3cJEA7t

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*. https://bit.ly/2MNXD5L

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg. https://bit.ly/3cR9hrr