ANEXO III. MEMORIA FINAL DE PROYECTO

USO DE LA PLATAFORMA DE SIMULACIONES PREDICTIVAS KAGGLE PARA LA ADQUISICIÓN DE COMPETENCIAS RELACIONADAS CON EL PERFIL PROFESIONAL CIENTÍFICO DE DATOS EN ASIGNATURAS DEL GRADO EN INGENIERÍA INFORMÁTICA. CÓDIGO DEL PROYECTO: 2017-1-5008

Pedro Antonio Gutiérrez Peña *, Juan Carlos Fernández Caballero *, César Hervás Martínez, Manuel Dorado Moreno, Antonio Manuel Durán Rosal, David Guijo Rubio, Julio Camacho Cañamón, Javier Sánchez Monedero, María Pérez Ortiz, Antonio Manuel Gómez Orellana

* {pagutierrez@uco.es, jfcaballero@uco.es} **(Departamento de Informática y Análisis Numérico, Campus de Rabanales, Edificio Einstein, 3ª planta, 14071 Córdoba, España)**

Universidad de Córdoba

Resumen

La Ciencia de Datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento de datos en sus diferentes formas. Los investigadores se apoyan en modelos, ecuaciones y algoritmos, así como en la evaluación e interpretación de los resultados. Dentro de la Ciencia de Datos, el Aprendizaje Automático (campo de la Inteligencia Artificial) se encarga de crear modelos capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

A la persona que trabaja en la Ciencia de Datos se le conoce como Científico de Datos, siendo uno de los trabajos más solicitados en la actualidad. Ante esta figura emergente surgen plataformas como Kaggle, que es una plataforma para modelos predictivos y competiciones analíticas en Aprendizaje Automático. En Kaggle, empresas, investigadores y científicos de datos de todo el mundo publican sus datos para resolver un determinado problema, así como los resultados y las estadísticas conseguidas a partir de los mejores modelos de predicción creados para ello.

Dentro del mundo profesional de la Ingeniería Informática, el rol de Científico de Datos es una figura que está siendo imprescindible, tanto para empresas públicas y privadas como para instituciones gubernamentales, de forma que el tipo de conocimientos que se requiere para ello es cada vez más demandado a las Universidades. Aunque a nivel teórico sí que se cubren en las Universidades algunas competencias básicas necesarias para este perfil, consideramos que es necesario complementar esos conocimientos con una experiencia práctica real que enfrente al alumnado a su complejidad.

El objetivo de este proyecto docente sería el usar la plataforma Kaggle como un medio TIC de aprendizaje y motivación para el alumnado, de forma que de una manera amena y divertida sirva de apoyo a los conocimientos específicos que se imparten en la asignatura de Introducción al Aprendizaje Automático de 3º de Grado en Ingeniería Informática. Concretamente se lleva a cabo una competición en la que se tiene que obtener un mejor modelo de predicción en un problema real de predicción de altura de olas en el Golfo de Alaska, únicamente utilizando esta plataforma y las herramientas de modelado abordadas en la asignatura. El objetivo de esta actividad será el de que los alumnos adquieran competencias para el perfil profesional Científico de Datos, complementando de esta forma la formación teórica.

Palabras clave: Inteligencia artificial; Aprendizaje automático; Ciencia de datos; Kaggle; Scikit-learn

1. INTRODUCCIÓN

La **Ciencia de Datos** [1] es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas. Esta ciencia es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva. Dentro de la Ciencia de Datos, el **Aprendizaje Automático** se encarga de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos.

A la persona que trabaja en la Ciencia de Datos se le conoce como **Científico de Datos**, y es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y en la elaboración de modelos predictivos, siendo uno de los trabajos más solicitados en la actualidad (en 2015, 4.4 millones de trabajos se crearon solo para soportar tareas de **Big Data**). Es por tanto, que el **Ingeniero Informático** tiene un perfil que se ajusta perfectamente a esta disciplina.

Ante esta figura emergente surgen plataformas como **Kaggle** [2] para modelos predictivos y competiciones analíticas en Aprendizaje Automático, donde empresas, investigadores y científicos de datos de todo el mundo publican sus datos para resolver un determinado problema. Los usuarios registrados pertenecen a todo tipo de disciplinas profesionales, lo que enriquece las soluciones y permite resolver problemas de todo tipo.

El enfoque de Kaggle es abierto y distribuido, ya que está en la nube como herramienta de **Cloud Computing**, y su éxito está basado en la colaboración entre participantes y sus técnicas. De esta manera la sinergia de estrategias y técnicas utilizadas para modelar un problema del mundo real ayuda a obtener mejores resultados, a la vez que se produce una función de difusión fundamental para la comunidad científica. Kaggle organiza concursos en los que los compiten por la oportunidad

de conseguir una entrevista en empresas líderes en ciencia de datos como Facebook, Winton Capital y Walmart, por un premio en metálico, o simplemente por darse a conocer y aportar soluciones a la ciencia.

Dicho esto, existen en titulaciones como la **Ingeniería Informática**, algunas asignaturas que imparten en sus contenidos temas relacionados con la Ciencia de los Datos. Así, aunque a nivel teórico se cubren algunas competencias básicas necesarias para este perfil, consideramos que es necesario complementar esos conocimientos con una experiencia práctica que enfrente al alumnado a la complejidad que suponen los problemas reales.

Este proyecto pretende utilizar Kaggle como herramienta **TIC** (Tecnologías de la Información y la Comunicación) en algunas asignaturas del Grado de Ingeniería Informática que oferta la Universidad de Córdoba, de manera que permita al alumnado aplicar y comprobar en la práctica [3][4] los conocimientos teóricos adquiridos dentro de los programas de las asignaturas. Desde el punto de vista docente, la introducción de TICs en el aula facilitan el proceso de aprendizaje del alumnado, estimulan la motivación y proveen un ambiente abierto de colaboración entre los estudiantes.

Muchas de las competencias que se pretenden cubrir en estas asignaturas requieren que el alumnado adquiera conocimientos de preprocesamiento de bases de datos, entrenamiento, validación y evaluación de modelos predictivos. En la actualidad, estas competencias se aplican diariamente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la ingeniería, bien en la empresa privada, o bien en la pública, y por supuesto, también en el caso de que se quiera continuar una carrera docente e investigadora en la Universidad.

Debido a los hechos anteriores, creemos que la participación del alumnado en plataformas como Kaggle reportaría una gran ventaja: la colaboración entre usuarios de distintas partes del mundo. Los problemas reales a los que se enfrentarán los estudiantes en un futuro necesitan de un equipo de trabajo para abordar los objetivos propuestos, lo que hace que la colaboración con los demás usuarios sea más que necesaria para poder resolver los problemas propuestos. Además de la colaboración, existe una competitividad sana entre los investigadores por el hecho de conseguir la mejor predicción para un problema determinado.

Para crear una experiencia completa para el alumnado como participante de una competición en Kaggle, se solicitó, como parte del presupuesto del proyecto, un premio simbólico en material de reprografía para los grupos de alumnos que consigan las puntuaciones más altas. Este proyecto además supondría ventajas para el alumnado y para el profesorado. Para los primeros, les ayudaría a reforzar y concretar conocimientos teóricos y a resolver problemas reales, más cercanos al mundo profesional de la Ingeniería, cubriendo competencias relacionadas con este aspecto. Por otro lado, al profesorado le permitiría validar el grado de comprensión de las explicaciones teóricas por parte del alumnado y le ayudaría, también, a actualizar los conocimientos sobre problemas reales sobre los que se pueden aplicar las asignaturas.

En concreto, las asignaturas implicadas en el proyecto serían:

- **Introducción al Aprendizaje Automático (IAA)**, tercer curso del Grado en Ingeniería Informática, especialidad en Computación, segundo cuatrimestre.
- **Introducción a los Modelos Computacionales (IMC)**, cuarto curso del Grado en Ingeniería Informática, especialidad en Computación, primer cuatrimestre. En esta asignatura no se ha podido aplicar el proyecto ya que la resolución provisional del mismo se obtuvo a finales del primer cuatrimestre, con lo que no era viable en tiempo su preparación.

Como en la segunda de las asignaturas citadas no se ha podido realizar el proyecto, el equipo de trabajo ha optado por realizar una serie de talleres-seminarios de programación en **Python** usando la biblioteca ***Scikit-Learn*** [5] para Aprendizaje Automático, la cual se puede utilizar en Kaggle. Estos talleres-seminarios se han hecho en colaboración con el **Aula de Software Libre de la Universidad de Córdoba** [6] y se ha dejado abierta su participación a cualquier alumno que forme parte de ella (además de los alumnos implicados en la asignatura de IAA). El esfuerzo que ha tenido que realizar el equipo de trabajo y el Aula de Software Libre para poder llevar a cabo dichos talleres-seminarios ha sido muy alto, pero estamos muy satisfechos de la experiencia, que ha superado con creces nuestras expectativas iniciales, ya que la asistencia e implicación del alumnado ha sido considerable.

2. OBJETIVOS

A continuación se indican los objetivos planteados y conseguidos con este proyecto, los cuales han variado levemente con respecto a la asignatura de Introducción a los Modelos Computacionales, que al ser en el primer cuatrimestre no fue factible la aplicación de la presente experiencia docente en tiempos, pero que se han reforzado por otro lado con la creación de los talleres-seminarios en el Aula de Software Libre, tanto para alumnos de la asignatura de IAA como para cualquier otro alumnado que quisiera asistir:

- Usar herramientas TIC en la nube de manera docente para:
    - Las prácticas de la asignatura de IAA, tercer curso de Grado en Ingeniería Informática, especialidad en Computación.
    - Para los alumnos que quieran asistir al Aula de software libre de la Universidad de Córdoba.
- Crear talleres-seminarios del uso de Python y *Scikit-Learn* como herramientas en Aprendizaje Automático a poder usar en la plataforma Kaggle:
    - Para las prácticas de la asignatura de IAA, tercer curso de Grado en Ingeniería Informática, especialidad en Computación.
    - Para los alumnos que quieran asistir al Aula de software libre de la Universidad de Córdoba.

- Desarrollar aspectos prácticos del perfil profesional asociado a la titulación en la que se aplica el proyecto, en concreto, aspectos relacionados con la Ciencia de Datos.
- Usar Kaggle como un medio de aprendizaje y motivación para el alumnado de una manera amena y divertida, y que sirva de apoyo y validación sobre los conocimientos específicos que se imparten en la asignatura de IAA.
- Realizar una competición en Kaggle:
    - Incluir tres sesiones prácticas en la asignatura involucrada. Concretamente se ha llevado a cabo una competición entre el alumnado, en la que se ha tenido que obtener un mejor modelo en un problema real de predicción de altura de ola significativa en el Golfo de Alaska [7].
    - La misma competición se ha abierto entre los alumnos que han asistido a los seminarios impartidos en el Aula de Software Libre. A diferencia de los alumnos de IAA no han tenido la explicación del profesorado en horas de clase, pero con el material proporcionado en los seminarios han suplido esta carencia.
- Fomentar un ambiente tanto de cooperación como de competitividad entre el alumnado, con el objetivo de alcanzar los mejores resultados en los problemas.
- 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 IAA, tercer curso de Grado en Ingeniería Informática, especialidad en Computación.

3. MATERIAL Y MÉTODOS

Los objetivos anteriormente mencionados se han llevado mediante la realización de las siguientes actividades:

- **Actividad 1**: Búsqueda y adaptación de un problema a ser susceptible de resolverse mediante técnicas de aprendizaje automático. Para ello se ha recurrido a un problema real abordado por el grupo de investigación del que forman parte los participantes de este proyecto (Grupo AYRNA - http://www.uco.es/ayrna/), concretamente un problema de predicción de altura de ola en una zona del Golfo de Alaska [7]. El problema se ha adaptado a datos más actuales y se ha incluido mucha más información en cuanto a las características a partir de las cuales se podría obtener un modelo de predicción. Ello obliga a que para obtener buenos resultados el alumnado tenga que emplear técnicas de selección de características y de pre-procesado de datos. Toda la información proporcionada al alumnado se ha adaptado al formato que exige la plataforma Kaggle y proviene de una boya real situada en el océano.
- **Actividad 2.** Creación de una competición y una práctica evaluable en la nota final en la asignatura de IAA. Para ello, el equipo de trabajo se ha encargado de redactar una práctica de 3 sesiones de duración (6 horas, 2 horas por sesión de clase). También ha sido necesario crear la competición de Kaggle en la nube, disponer la base de datos con la información sobre la predicción de altura de ola, crear las normas de competición, configurar el entorno de trabajo en la nube y todo lo necesario para que la plataforma esté lista para subir resultados por parte de los alumnos. Se formaron equipos de trabajo, donde cada equipo cooperó internamente y compitió con el resto de equipos, siempre mentorizados por el profesorado, que acompañó durante toda la competición a los participantes de los equipos. Se mantuvieron con frecuencia reuniones con los equipos en horarios de tutorías y también fuera de esos horarios oficiales (a petición de los alumnos), tanto en forma presencial como virtual.
- **Actividad 3.** Adecuación de la misma competición que se ha hecho para los alumnos de IAA, pero abierta también a los alumnos que hayan asistido al Aula de Software Libre de la UCO. Para ello, en la propia plataforma Kaggle se indica cómo deben registrarse en la competición los alumnos de IAA y los asistentes al aula de Software Libre, de forma que el profesorado pueda tener constancia de cómo se va desarrollando la competición y quién consigue cada posición del ranking.
- **Actividad 4.** Creación e impartición de talleres-seminarios en el Aula de Software Libre de la UCO para aprender programación en Python y usar la biblioteca para aprendizaje automático *Scikit-Learn* [5] y otras asociadas a facilitar el uso de datos, como son *numpy* y *pandas*. En la asignatura IAA se enseña al alumnado a crear modelos de predicción utilizando la herramienta software *Weka* [8], muy usada por la comunidad científica como entorno de trabajo para preprocesar datos y con multitud de algoritmos que crean modelos predictivos, pero también se ha querido añadir la posibilidad de usar Python y *Scikit-Learn* por ser otra opción que actualmente está teniendo una gran repercusión y uso a nivel mundial. Se contactó entonces con el **Aula de Software Libre de la UCO, que se brindó a colaborar, preparar los ordenadores de las aulas y difundir los seminarios entre el alumnado.**

  Concretamente se realizaron las siguientes 4 sesiones:

  1. *Introducción práctica a la ciencia de datos y al aprendizaje automático con Python. Realizada el 09 marzo 2018, de 16:00 – 19:00, Aula S1, Ed. Ramón y Cajal, Campus de Rabanales.* En la Figura 1 se muestra un ejemplo de cómo se han ido publicitando los talleres-seminarios por parte del Aula de Software libre. También se ha dado publicidad a través de la plataforma Moodle de la UCO en las asignaturas que imparten algunos de los participantes de este proyecto. En https://github.com/ayrna/tutorial-scikit-learn-asl se puede ver y descargar el contenido de este primer taller, preparado para que se pueda ejecutar en las aulas mediante cuadernos lanzados con Jupyter (al igual que el resto). Además se ha ofrecido al alumnado una parte teórica que acompañe a las prácticas y que se puede descargar en https://github.com/javism/introduccion-cd-ml. Concretamente en este primer contacto en cuanto a los talleres-seminarios, se enseñó al alumnado a configurar el entorno de trabajo con Python y a utilizar de manera básica las bibliotecas *numpy*, *pandas* y

Scikit-Learn, de forma que ya estuvieran preparados para practicar en las siguientes sesiones a más profundidad. Los siguientes 3 talleres-seminario al Taller de Introducción se dividen en puntos o secciones claramente diferenciadas en https://github.com/ayrna/tutorial-sklearn, concretamente para la primera sesión los puntos que van del 1.01 al 1.09, la segunda sesión los puntos que van del 2.01 al 2.05, y por último una tercera sesión que va desde los puntos 3.01 al 3.06. .

Figura 1 Difusión de seminarios por el Aula de Software Libre de la UCO

2. *Talleres de ciencia de datos y aprendizaje automático. Sesión 1: "Visualización, aprendizaje supervisado y métodos de evaluación". Realizado el 06 de abril 2018, des 16:00 – 20:00, Aula S1, Ed. Ramón y Cajal, Campus de Rabanales.* En https://github.com/ayrna/tutorial-sklearn se puede ver y descargar el contenido de la primera Sesión (y posteriores) después del Taller de Introducción. Esta sesión comprende visualización, aprendizaje supervisado y métodos de evaluación.
3. *Talleres de ciencia de datos y aprendizaje automático. Sesión 2: "Segunda sesión del taller de ciencia de datos y aprendizaje automático". Realizado el 20 abril 2018 de 16:00 – 19:00, aula S1 del edificio Ramón y Cajal, Campus de Rabanales.* Esta sesión comprende Aprendizaje no supervisado.
4. *Talleres de ciencia de datos y aprendizaje automático. Sesión 3: "Árboles de decisión, procesamiento de texto y caso práctico de detección de spam".* Esta sesión comprende árboles de decisión, procesamiento de texto, un caso práctico de detección de spam y selección de características.

- **Actividad 5.** Para el proyecto se realizaron encuestas voluntarias de tipo test al alumnado, una antes de empezar la competición y los talleres-seminarios, y otra al finalizar el proyecto, tratando de valorar los conocimientos prácticos adquiridos. Para este fin el equipo de trabajo utilizó una herramienta que ofrece Google, https://docs.google.com/forms/, concretamente se usaron "test de autoevaluación", accesibles en https://docs.google.com/forms/d/e/1FAIpQLScT7m4dsOcuG5RWHAyAfcA6w3AEWYvlLnqcJ6bTvn3KK84xaw/viewform
  Se incluyeron un total de 16 preguntas de tipo test para obtener un sondeo del nivel inicial de los alumnos y que se listan a continuación:
  1. **Señale cuáles de las siguientes afirmaciones son correctas:**
     Puede haber una o más de una respuestas correctas, aunque la pregunta siempre se formulará en plural
  - ❑ El aprendizaje automático extrae conocimiento de los datos de forma automática, con el objetivo de realizar predicciones con nuevos datos no conocidos o de describir los datos disponibles.
  - ❑ En aprendizaje supervisado no se conoce la clase de pertenencia de los patrones.
  - ❑ El conjunto de 'test' se utiliza para validar la capacidad de generalización del modelo obtenido en la fase de entrenamiento.

  2. **Señale cuáles de las siguientes afirmaciones son correctas:**
     Puede haber una o más de una respuestas correctas, aunque la pregunta siempre se formulará en plural
  - ❑ Python facilita la representación y el manejo de los conjuntos de datos mediante el uso de matrices.
  - ❑ Python dispone de paquetes que permiten representar gráficamente los datos, facilitando el análisis de los mismos.
  - ❑ Scikit-learn es una biblioteca para aprendizaje automático que incorpora, entre otros, algoritmos de clasificación, regresión y procedimientos para el preprocesamiento de los datos.

  3. **La estandarización es una técnica de reescalado, tras aplicarla sobre un conjunto de datos todos sus atributos tendrán media 0 y desviación típica 1.**
  - ❑ Verdadero

❑ Falso

4. **Se dispone de un conjunto de datos con el cual abordar una tarea de clasificación, la opción más recomendable es:**

❑ Utilizarlo en su totalidad para realizar el entrenamiento. Luego emplear el 25% de los patrones, elegidos aleatoriamente, para realizar la generalización.
❑ Emplear un 75% de datos para entrenar y el 25% restante para generalizar, obviando la distribución de patrones por clase.
❑ Emplear un 75% de datos para entrenar y el 25% restante para generalizar, manteniendo la distribución de patrones por clase.

5. **Los algoritmos de 'clustering' agrupan los patrones en función de una etiqueta que ha establecido un supervisor y de una medida de distancia que indica lo parecidos o diferentes que son unos de los otros.**

❑ Verdadero
❑ Falso

6. **¿Cuál es la principal desventaja que presenta el algoritmo k-NN?**

❑ Es un modelo lineal.
❑ Su rendimiento depende del correcto ajuste del parámetro k.
❑ No tiene buen rendimiento.

7. **¿Qué es esencial para una comparación honesta entre modelos?**

❑ Utilizar el mismo software.
❑ Utilizar el mismo conjunto de datos de entrenamiento y test.
❑ Todas las respuestas son correctas.

8. **El Análisis de Componentes Principales (PCA) es:**

❑ Una técnica para la extracción de características y reducción de la dimensionalidad.
❑ Un algoritmo no supervisado de agrupamiento o clustering.
❑ Ninguna respuesta es totalmente correcta.

9. **¿El porcentaje de patrones correctamente clasificados es una buena métrica de rendimiento?**

❑ Sí.
❑ No.
❑ Depende del problema.

10. **La diagonal principal de una matriz de confusión expresa:**

❑ Los patrones bien clasificados.
❑ Los patrones mal clasificados.
❑ Ninguna respuesta es totalmente correcta

11. **Señale cuáles de las siguientes afirmaciones son correctas sobre los árboles de decisión: Puede haber una o más de una respuestas correctas, aunque la pregunta siempre se formulará en plural**

❑ Requieren la transformación de las variables categóricas.
❑ Son modelos muy interpretables e intuitivos.
❑ Requieren poda para no incurrir en sobre-entrenamiento.

12. **Con las técnicas de selección de características:**

❑ Se reduce la dimensionalidad de los datos, y en ocasiones, se obtienen modelos con mayor capacidad de generalización.
❑ Se reduce la dimensionalidad de los datos y siempre se mejora la capacidad de generalización de los modelos.
❑ Ninguna respuesta es totalmente correcta.

13. **Señale cuáles de las siguientes afirmaciones son correctas sobre el algoritmo DBSCAN: Puede haber una o más de una respuestas correctas, aunque la pregunta siempre se formulará en plural**

❑ No necesita que se le indique el número de 'clusters' a localizar.
❑ Es más robusto al ruido y 'outliers'.
❑ Puede detectar 'clusters' con formas geométricas arbitrarias.

14. **¿Cuál de las siguientes metodologías no pertenece al aprendizaje automático?**

❑ Agrupamiento.
❑ Métodos de clasificación/regresión.
❑ Inversión de matrices.
❑ Extracción de características.

15. **Utilizar un conjunto de validación nos puede ayudar a detectar la aparición del sobre-entrenamiento.**

❑ Verdadero
❑ Falso

16. **¿Cuáles de las siguientes afirmaciones son correctas?**
    Puede haber una o más de una respuestas correctas, aunque la pregunta siempre se formulará en plural

❑ El conjunto de validación es otro nombre que se le da al conjunto de test.
❑ El conjunto de test se puede utilizar para decidir el valor de los parámetros de un modelo.
❑ El conjunto de validación se puede utilizar para decidir el valor de los parámetros de un modelo.

Por otro lado, se incluyen 4 preguntas para **auto-evaluar los conocimientos sobre la plataforma Kaggle y las bibliotecas de Python para aprendizaje automático**, de forma que el equipo de trabajo de este proyecto pueda ratificar si son tecnologías nuevas y conocidas por el alumnado. Las preguntas siguen una escala de Likert y se listan a continuación:

1. **Valora tus conocimientos sobre Python:**
   1 Muy reducidos, 4 Lo domino en profundidad
   □1 □ 2 □ 3 □ 4
2. **Valora tus conocimientos sobre scikit-learn**
   1 Muy reducidos, 4 Lo domino en profundidad
   □1 □ 2 □ 3 □ 4
3. **Valora tus conocimientos sobre aprendizaje automático**
   1 Muy reducidos, 4 Lo domino en profundidad
   □1 □ 2 □ 3 □ 4
4. **Valora tu capacidad para abordar una competición de Kaggle**
   1 Muy reducidos, 4 Lo domino en profundidad
   □1 □ 2 □ 3 □ 4

Una vez indicadas las preguntas de tipo test y las preguntas del test de auto-evaluación, estos son los resultados que se han obtenido, descritos de manera breve por no alargar en demasía el contenido de la presente memoria.

Resultados antes de realizar el proyecto:

- Para las 16 preguntas tipo test:
  Cada respuesta acertada vale 1 punto, con lo que se pueden obtener un máximo de 16. Dicho esto se puede observar en la Figura 2 que el promedio es de 7,9 sobre 16, más o menos un 5 sobre 10. La mediana se encuentra en 8, con lo que refuerza la idea de que más o menos la mitad de los encuestados obtendrían un 5 sobre 10, y por último el intervalo de puntuaciones está entre 3 y 12, con lo cual indica que hay alumnos que tienen conocimientos muy bajos. Nótese que las preguntas no dejan de ser de tipo teórico y generales en cuanto a Aprendizaje Automático.

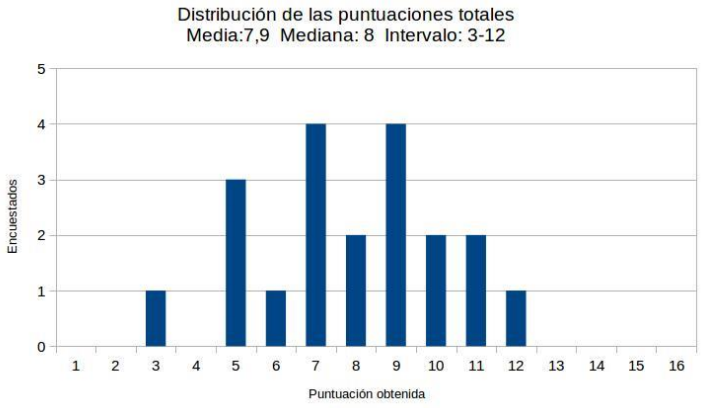


Figura 2 Test en encuestas previas al proyecto docente

- Para las 4 preguntas sobre auto-evaluación de la nueva tecnología utilizada:
  En este caso los resultados son mucho más abrumadores, ya que en todas las preguntas el mayor porcentaje de contestación se encuentra en las posiciones 1 y 2 de la escala de Likert, es decir, que el alumnado tiene conocimientos muy reducidos o bajos sobre este tipo de tecnologías a nivel práctico.

1. **Valora tus conocimientos sobre Python:**
   □1 - 40% □ 2 - 35% □ 3 - 15% □ 4 - 10%
2. **Valora tus conocimientos sobre scikit-learn:**
   □1 - 95% □ 2 - 5% □ 3 - 0% □ 4 - 0%
3. **Valora tus conocimientos sobre aprendizaje automático:**
   □1 - 40% □ 2 - 55% □ 3 - 5% □ 4 - 0%
4. **Valora tu capacidad para abordar una competición de Kaggle:**
   □1 - 95% □ 2 - 5% □ 3 - 0% □ 4 - 0%

La Figura 3 visualiza los resultados anteriores, mostrandose los valores de Likert más usados el 1 y el 2:

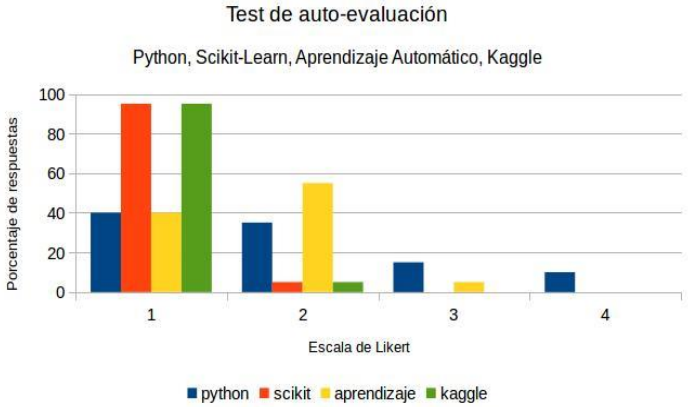


Figura 3 Auto-evaluación de tecnologías a usar previas al proyecto docente

Resultados después de realizar el proyecto:

- Para las 16 preguntas tipo test:
  El test posterior lo realizaron pocos alumnos, pero se puede observar en la Figura 4 que el promedio de los que han contestado es de 14,5 sobre 16, casi un 10 en puntuación. La mediana se encuentra también en el mismo valor, y el valor mínimo se encuentra en 13 puntos sobre 16. Esto quiere decir que el alumnado tiene mayor formación sobre los contenidos iniciales al terminar la experiencia docente.

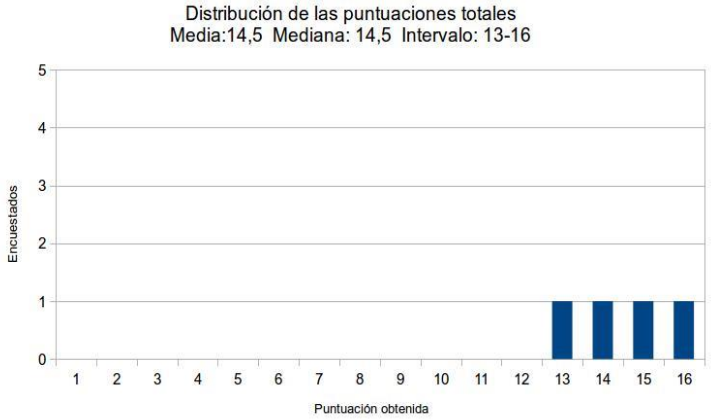


Figura 4 Test en encuestas posteriores al proyecto docente

- Para las 4 preguntas sobre auto-evaluación de la nueva tecnología utilizada:
  En este caso los resultados afirman lo positivo de la formación después del proyecto con respecto a las tecnologías utilizadas, siendo las posiciones 3 y 4 de la escala de Likert las más usadas, es decir, que el alumnado tiene conocimientos de medios a altos sobre este tipo de tecnologías a nivel práctico.

1. **Valora tus conocimientos sobre Python:**
   □1 - 25% □ 2 - 50% □ 3 - 25% □ 4 - 0%
2. **Valora tus conocimientos sobre scikit-learn:**
   □1 - 25% □ 2 - 25% □ 3 - 50% □ 4 - 0%
3. **Valora tus conocimientos sobre aprendizaje automático:**
   □1 - 0% □ 2 - 0% □ 3 - 50% □ 4 - 50%
4. **Valora tu capacidad para abordar una competición de Kaggle:**
   □1 - 0% □ 2 - 0% □ 3 - 75% □ 4 - 25%

La Figura 5 visualiza los resultados anteriores:

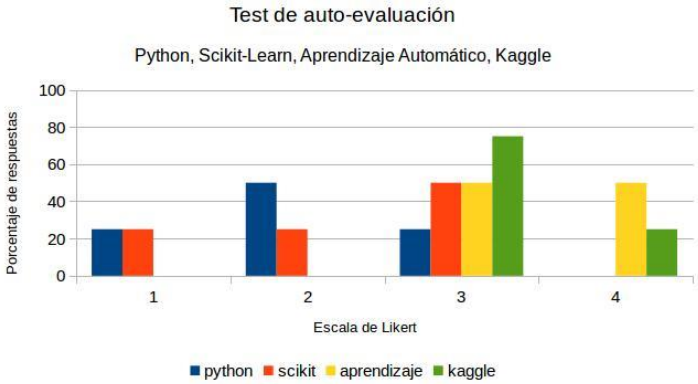


Figura 5 Auto-evaluación de tecnologías a usar posterior al proyecto docente

3.1 Presupuesto para el proyecto

El presupuesto que se solicitó para este proyecto fue de 500 euros a usar en los siguientes conceptos:

- Realización de las encuestas e impresión de tickets para retirada de material de reprografía en Don Folio e impresión de diplomas: 300 euros.
- Premios a los alumnos ganadores y mejores participantes de la competición: 200 euros.

De los 500 euros solicitados se concedieron 409.66 euros, cuyo gasto finalmente se ha repartido de la siguiente manera:

- 9.66 euros en concepto de impresión de tickets y diplomas.
- 400 euros en concepto de premios a los alumnos concursantes de ambas competiciones.

Se ha optado por esta división final porque Copisterías Don Folio de Córdoba nos ha permitido repartir todo el dinero otorgado al proyecto entre los participantes de las competiciones, para que lo puedan gastar en tickets de 5 euros en material de reprografía. La Figura 6 y Figura 7 muestran una copia escaneada de un ticket original, indicándose en la parte trasera que se enmarca dentro del presente proyecto de innovación docente de la Universidad de Córdoba.

Figura 6 Parte delantera de ticket para Don Folio

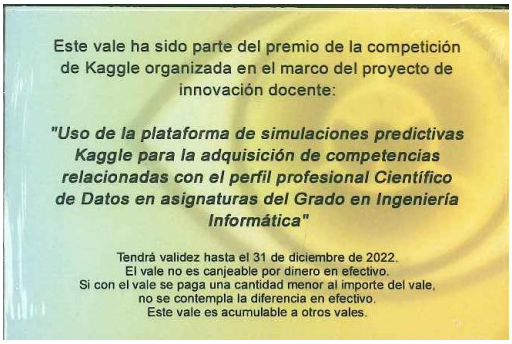


Figura 7 Parte trasera de ticket para Don Folio

3.2 Ranking en la competición

En cuanto a diplomas y rankings, la Figura 8 muestra un ejemplo de diploma para uno de los alumnos de la competición y la Figura 9 muestra el ranking de los 10 primeros equipos de trabajo, medido en función del valor de la métrica *F-Measure*, muy usada en Aprendizaje Automático para obtener la bondad de un modelo en cuanto a su capacidad de predicción. En https://www.kaggle.com/c/competition-iaa-1819 se puede ver la competición y sus datos públicos y privados. Se formaron un total de 23 equipos de trabajo en la competición, y el reparto en tickets para los ganadores se ha realizado entre los 7 primeros de la siguiente manera:

- **1er premio: 100 €** - Villalon Vaquero, Pedro José; Montenegro Alcántara, Jose Antonio; Quesada Sanchezbarba, Alejandro.
- **2o premio: 75 €** - Luque Reigal, Alejandro.
- **3er premio: 75 €** - Peñas Pastilla, Javier.
- **4o premio: 50 €** - Cubero Torres, Nicolás; Pérez Polo, Antonio Jesús.
- **5o premio: 50 €** - Delgado Zamorano, Jose Manuel; Luque Moreno, Pablo.
- **6o premio: 25 €** - Almeda Luna, Eduardo.
- **7o premio: 25 €** - Rioja Bravo, Jesús; Navarro Fuentes, Manuel Rafael.

En cuanto a los diplomas se han otorgado a los 3 primeros equipos (al resto se entregan certificados de participación):

- **3 Diplomas de 1er puesto:** Villalon Vaquero, Pedro José; Montenegro Alcántara, Jose Antonio; Quesada Sanchezbarba, Alejandro.
- **1 Diploma de 2o puesto:** Luque Reigal, Alejandro.
- **1 Diploma de 3er puesto:** Peñas Pastilla, Javier.


UNIVERSIDAD D CORDOBA

PRIMER PUESTO

en la Competición de Kaggle en el marco del proyecto de innovación docente de la Universidad de Córdoba

"Uso de la plataforma de simulaciones predictivas Kaggle para la adquisición de competencias relacionadas con el perfil profesional Científico de Datos en asignaturas del Grado en Ingeniería Informática"

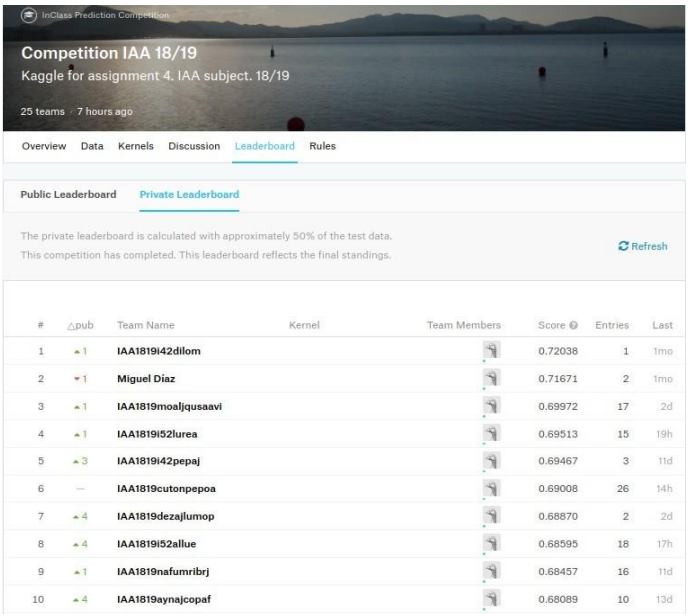
Villalón Vaquero, Pedro José

23 de mayo de 2018

Pedro Antonio Gutiérrez Peña
Responsable del Proyecto de Innovación Docente

Juan Carlos Fernández Caballero
Responsable del Proyecto de Innovación Docente

Figura 8 Diploma para uno de los alumnos participantes


InClass Prediction Competition

Competition IAA 18/19

Kaggle for assignment 4. IAA subject. 18/19

25 teams · 7 hours ago

Overview Data Kernels Discussion Leaderboard Rules

Public Leaderboard Private Leaderboard

The private leaderboard is calculated with approximately 50% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	IAA1819i42dilom			0.72038	1	1mo
2	▼1	Miguel Diaz			0.71671	2	1mo
3	▲1	IAA1819moaljqusaavi			0.69972	17	2d
4	▲1	IAA1819i52lurea			0.69513	15	19h
5	▲3	IAA1819i42pepaj			0.69467	3	11d
6	—	IAA1819cutonpepoa			0.69008	26	14h
7	▲4	IAA1819dezajlumop			0.68870	2	2d
8	▲4	IAA1819i52allue			0.68595	18	17h
9	▲1	IAA1819nafumribrj			0.68457	16	11d
10	▲4	IAA1819aynajcopaf			0.68089	10	13d

Figura 9 Ranking final de la competición

4. Resultados obtenidos y Discusión

Los resultados que se han obtenido en esta experiencia docente, al entender de los participantes de este proyecto de innovación, son los siguientes:

- Se ha dado a conocer al alumnado una serie de herramientas profesionales y de actualidad para el modelado de problemas reales mediante técnicas de Aprendizaje Automático.
- Se ha conseguido que el alumnado se cerciore de que el perfil profesional Científico de Datos es cada vez más demandado, y que puede ser una exitosa salida profesional. Muestra de ello es que la plataforma Kaggle informa que tiene más de 1.000.000 de usuarios registrados (http://blog.kaggle.com/2017/06/06/weve-passed-1-million-members/), y que en ella participan las principales empresas de Ingeniería Informática del mundo.
- Se ha dado a conocer al alumnado las diversas tareas que realiza el científico de datos, entre ellas el preprocesado de datos, con la finalidad de poder profundizar en un futuro.
- Se han mejorado a nivel práctico las competencias a adquirir por el alumnado en la asignatura de IAA.
- Se ha conseguido que el alumnado aporte y reciba nuevos conocimientos prácticos a partir de la gran comunidad existente en la plataforma Kaggle, formada por investigadores y Científicos de Datos de diversos ámbitos, ámpliamente consolidados.
- Se ha fomentado el uso de TICs en el aula y se ha conseguido una buena participación en seminarios (Figura 10).

- Se ha fomentado el trabajo en grupo, ya sea entre participantes de un mismo grupo de competición como entre grupos, haciendo que el *feedback* y la relaciones entre ellos culmine en una mejora de conocimiento para todos.
- Se ha aumentado la confianza del alumnado en sí mismo para involucrarse motu proprio en problemas reales donde pueda aplicar sus conocimientos teórico-prácticos para dar una solución tangible.

Figura 10 Uno de los seminarios organizados

5. CONCLUSIONES

Como conclusión final, ligada a los resultados obtenidos y la discusión realizada en la sección anterior, a nuestro parecer este tipo de proyectos de innovación docente mejoran de manera notable la capacidad del alumnado para enfrentarse a un problema real cuando terminen su titulación. 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 Ingenieros en Informática.

Como experiencia ha sido muy satisfactoria, no esperábamos que el grado de implicación y de interés de muchos de los alumnos fuera tan alto con la cantidad de asignaturas y esfuerzo que tienen que realizar a lo largo del curso académico. Esta implicación ha sido así desde el inicio, cuando ni siquiera se comentó que la competición estaba ligada a premios en tickets canjeables por material de reprografía, con lo que descartamos que el móvil económico en material haya sido el que haya suscitado tanto interés.

Con respecto al esfuerzo del equipo de trabajo hay que reseñar que ha sido elevado, ya que ha habido que adaptar al nivel del alumnado todo el material de los seminarios y la práctica docente en IAA, pero la satisfacción conseguida hace que haya merecido la pena. Se espera poder realizar nuevas ediciones en cursos académicos futuros.

AGRADECIMIENTOS

Agradecemos al Aula de Software Libre de la Universidad de Córdoba y a la Copistería Don Folio su colaboración en la realización de este proyecto de innovación docente.

BIBLIOGRAFÍA

[1] LOUKIDES, M. *"What Is Data Science?"*, O'Reilly Media Inc, 2011.

[2] KAGGLE. *Your Home for Data Science,* [Online] Disponible en: https://www.kaggle.com/, 2017.

[3] ANECA, *Los procesos de inserción laboral de los titulados universitarios en España. Factores de facilitación y de obstaculización*, [Online]. Disponible en: http://www.aneca.es/media/308144/publi_procesosil.pdf, 2009.

[4] ROJAS, F. *Enfoques sobre el aprendizaje humano*, [Online] Disponible en: http://ares.unimet.edu.ve/programacion/psfase3/modII/biblio/Enfoques_sobre_el_aprendizaje1.pdf, 2011.

[5] SCIKIT-LEARN, *Machine Learning in Python,* [Online] Disponible en: http://scikit-learn.org/stable/, 2018.

[6] AULA DE SOFTWARE LIBRE DE LA UCO, [Online] Disponible en: https://www.uco.es/aulasoftwarelibre/, 2018.

[7] FERNANDEZ, J. C., SALCEDO-SANZ, S., GUTIÉRREZ, P. A., ALEXANDRE, E., HERVÁS-MARTÍNEZ, C., *Significant wave height and energy flux range forecast with machine learning classifiers*, Engineering Applications of Artificial Intelligence, vol. 43, pags 44–53, 2015.

[8] FRANK, E., HALL, M. A., WITTEN, I. H., *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.