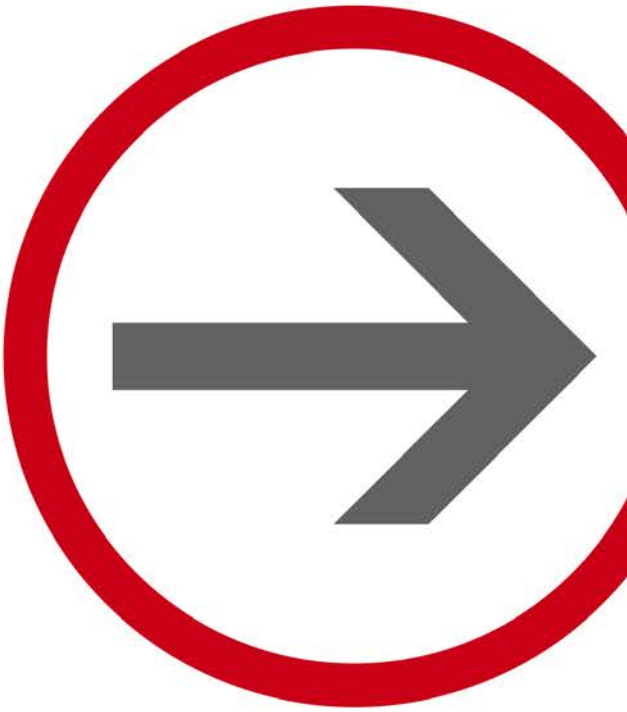
u online



Prácticas de Aprendizaje Automático utilizando las Competiciones de Kaggle

BANCO DE BUENAS PRÁCTICAS DOCENTES

REBECA GOYA ESTEBAN Y OSCAR BAQUERO PÉREZ

UNIVERSIDAD REY JUAN CARLOS online
online.urjc.es

1. La práctica

- **Título:** Metodologías Activas para Educación Artística y Plástica
- Curso Académico: 2016/17
- Asignatura: N/A
- Área/Titulación: Ingeniería
- **Grupo de Estudiantes:** Alumnos de los diferentes grados de Ingeniería de la ETSIT

La ciencia de datos es una de las disciplinas más atractivas en la actualidad tanto desde el punto de vista académico como empresarial. Una de las áreas fundamentales es el aprendizaje automático. La enseñanza de este tópico es, por lo general, muy académica tanto en grado como en máster. Habitualmente, se presentan de forma teórica diferentes algoritmos de aprendizaje automático, acompañados de prácticas, con conjuntos de datos convenientemente preprocesados para la aplicación del algoritmo concreto. Una de las desventajas de esta aproximación es que cuando los alumnos se enfrentan a problemas reales no han desarrollado las herramientas adecuadas para establecer un modelo de datos correcto, elegir el preprocesado adecuado, tratar con los problemas asociados a los datos reales, ni tener que elegir entre los diferentes algoritmos ni sus hiperparámetros, es decir, no han entrenado suficientemente la parte de artesanía del aprendizaje automático. Proponemos en este trabajo utilizar las competiciones de ciencia de datos, estilo Kaggle, en la que los alumnos se enfrentan a un problema real, de forma que tienen que realizar todos los pasos necesarios en un proyecto real de aprendizaje automático. Adicionalmente, los alumnos compiten por quedar los primeros en el ranking, recibiendo un premio los dos equipos ganadores. En este trabajo, participaron 20 alumnos de grado de Ingeniería de Telecomunicación y 4 de máster. Los conocimientos iniciales de los alumnos fueron variados, desde alumnos sin exposición al aprendizaje automático hasta alumnos que habían cursado alguna asignatura con conocimientos introductorios. En todos los casos, los alumnos identificaron un aumento en los conocimientos de aprendizaje automático, sobre todo incidiendo en el aumento de las habilidades prácticas de aplicación a datos reales.

2. Justificación

Una de las disciplinas más atractivas en la actualidad, tanto desde el punto de vista académico, como desde el punto de vista empresarial, es lo que está dado en llamarse ciencia de datos. Este término hace referencia a un nuevo campo multidisciplinar cuyo principal objetivo es utilizar modelos y procesos para extraer conocimiento a partir de datos en diferentes formatos, tanto estructurados como desestructurados.

Actualmente, tanto en el ámbito empresarial como en el de investigación, tenemos acceso a una cantidad ingente de datos: desde las fuentes clásicas de datos, hasta nuestros teléfonos móviles; nosotros mismos nos hemos convertido en una fuente de datos excepcional. Existen estimaciones de que en el mundo se generan 2.5 exabytes de datos cada día, estamos pues ante la era del Big Data (IBM Big Data, 2017). Pero esto no quiere decir que seamos capaces de extraer conocimiento de estos datos. Para ello, deberíamos ser capaces de crear modelos, utilizando estos datos, que nos permitan entender mejor el problema a resolver o/y hacer mejores predicciones. Sin embargo, existe el peligro de que el ritmo al que se adquieren nuevos datos sea muy superior al ritmo al que nuestro entendimiento es capaz de procesarlos, lo cual haría esa información inútil (Silver, 2012).

Uno de los principales problemas se encuentra en la formación de profesionales en este ámbito. Al tratarse de una disciplina eminentemente multidisciplinar es difícil encontrar, actualmente, estudios universitarios de grado que ofrezcan una formación completa en ciencia de datos.

Existen asignaturas en algunos curricula de grados tales como Matemáticas, Estadística, Informática, Telecomunicación, que ofrecen una introducción a este campo de la ciencia. Sin embargo, estas introducciones son, por lo general, insuficientes para proporcionar una formación en ciencia de datos. En la formación académica se introduce a los alumnos en las técnicas, modelos y procedimientos mediante el uso de bases de datos pre-procesadas y perfectamente estructuradas, haciendo poco hincapié en las dificultades que un investigador o desarrollador va a encontrar cuando se enfrente a bases de datos reales. A sabiendas de esta carencia existen plataformas como Kaggle (https://www.kaggle.com/) que fomentan competiciones utilizando datos reales para solventar un problema real propuesto por una empresa.

En este trabajo proponemos un método de enseñanza de aprendizaje automático en el que el elemento principal es una competición estilo Kaggle, con un premio final, en el que se tiene que solucionar un

problema real de datos. Esta competición se acompaña de cuatro seminarios con el objeto de proporcionar a los alumnos las herramientas básicas para poder abordar el problema.

Generalmente, en la formación académica se introduce a los alumnos en las técnicas, modelos y procedimientos mediante el uso de bases de datos pre-procesadas y perfectamente estructuradas, haciendo poco hincapié en las dificultades que un investigador, o desarrollador, va a encontrar cuando se enfrente a bases de datos reales. Habitualmente, se presenta una herramienta de forma teórica y, a continuación, se realiza una práctica de laboratorio en la que se estudia el funcionamiento de ese algoritmo, pidiendo a los alumnos que modifiquen alguno de sus parámetros libres. A sabiendas de esta carencia existen plataformas como Kaggle (https://www.kaggle.com/) que fomentan competiciones utilizando datos reales para solventar un problema real propuesto por una empresa. Sin embargo, no está muy extendida la utilización de tales competiciones como parte de la formación académica.

El método que proponemos, en este trabajo, para completar la formación práctica en aprendizaje automático de los alumnos, presenta un formato de seminarios más competición, y es eminentemente práctico (en la Sección 3 se detalla este método). El objetivo es dotar a los alumnos de una pequeña base teórica y se les muestra a dónde recurrir para profundizar en ella, así como para encontrar las herramientas software que pueden utilizar. A partir de ese momento se convierte en un trabajo de investigación y desarrollo por parte de los alumnos con datos reales. Creemos que esta metodología aporta un valor añadido a la formación de grado y postgrado de los alumnos. El método tiene como audiencia principal los alumnos de grado de las ingenierías de Telecomunicación, grados de estadística y los alumnos de máster relacionados con ciencia de datos.

3. Desarrollo

Objetivos

El modelo de enseñanza del aprendizaje automático que se propone en este trabajo es una competición de datos junto a cuatro seminarios de tres horas cada uno. Ambos, competición y seminarios se realizan de forma simultánea.

La competición de datos está basada en las competiciones de Kaggle (Kaggle, 2017), en las que se propone la resolución de un problema real con datos. Se pidió a los estudiantes que formasen equipos (2-3 miembros) para poder participar en la competición. En este tipo de competiciones se plantea un problema

de aprendizaje automático, en el actual fue un problema de clasificación, en el que se pedía a los contendientes predecir la mortalidad en accidentes de coches. Se puso a disposición de los alumnos un conjunto de datos reales, proporcionado por el Gobierno de la Comunidad Valenciana (España). El conjunto de datos consistía en 9341 accidentes durante un período de al menos 5 años (2005-2010), que se corresponden a 154 carreteras regionales (Figuera, Lillo, MoraJiménez, Rojo-Álvarez, & Caamaño, 2011). En los informes de los accidentes se completan un total de 82 campos relacionados con las circunstancias de los accidentes (hora, día, identificador de la carretera, posición geográfica), información de los conductores y de los pasajeros (edad, sexo, experiencia en conducción, número de pasajeros, años), vehículos involucrados (tipo, marca, modelo, año), carretera (número de carriles, ancho total, radio de curva, presencia de intersecciones, marcas del pavimento, señalización vertical), así como otros posibles factores involucrados en el accidente (velocidad, presencia de alcohol o drogas, condiciones atmosféricas, estado de la calzada). A pesar de que los datos fueron revisados por las autoridades, los datos son reales, lo cual implica la presencia de ruido de diferente tipología: ruido de muestreo, ruido en las mediciones, datos mal introducidos en la base de datos, así como valores perdidos. El objetivo principal es que los equipo emulen el trabajo de un equipo de ciencia de datos en la vida real.

En las competiciones de Kaggle, se utiliza una métrica que permite comparar las soluciones de los diferentes equipos, en el caso concreto de esta competición se utilizó la exactitud (accuracy), que mide el porcentaje de aciertos en la clasificación. El conjunto de datos se divide en dos subconjuntos, uno de entrenamiento (training) y otro de evaluación (test), que es el que se utilizar para establecer el ranking. Adicionalmente, el conjunto de test es dividido de forma aleatoria, en un conjunto público, con el que se va actualizando el ranking durante la vida de la competición, y uno privado, que se utiliza para la clasificación final. Esto se hace para evitar el problema conocido por sobreajuste-al-ranking. Adicionalmente se limita el número de veces que se puede evaluar una solución por un determinado equipo.

La competición estuvo abierta desde el 21 de marzo hasta el 18 de abril de 2017. Para la organización de la competición se utilizó la plataforma de Kaggle, Kaggle in Class, que proporciona todas las herramientas de Kaggle para crear una competición para ámbitos académicos sin coste alguno (Kaggle, Kaggle in Class, 2017). Dicha plataforma se encarga de generar automáticamente los rankings, aceptar la subida de soluciones, proporcionar un foro de intercambio de ideas esencial para el aprendizaje. La dirección web de la competición es: https://inclass.kaggle.com/c/dataton-urjc-2017, ver Figura 2. En la competición se propuso un premio en metálico para los dos primeros equipos clasificados: 100 € y 50 € respectivamente.

El objetivo era mantener lo más realista posible la competición, de forma que los equipos estuviesen lo más motivados posibles. Los premios corren a cargo de una subvención del Vicerrectorado de Extensión Universitaria y Relaciones Internacionales de la Universidad Rey Juan Carlos dentro del Programa de Ayudas a Congresos y Seminarios.

Figura 2. Plataforma kaggle in class, página para la competición Datatón URJC 2017.

El lenguaje de programación propuesto para el desarrollo de la competición es Python (Python.org, 2017) que, junto a R (R Statistical, 2017), es uno de los más extendidos en aprendizaje automático, así como uno de los más recomendados para ciencia de datos. No obstante, y dado que para la solución sólo era necesario subir un archivo de texto plano (.csv) con los resultados de la clasificación, en la competición se permitía utilizar cualquier lenguaje de programación. Python es uno de los lenguajes más empleados para la enseñanza y el desarrollo de modelos de aprendizaje automático, tanto en universidades como en empresas top (Stanford, Google, etc) (Donoho, 2015). Es notorio también el hecho de que utilizando la herramienta de Google Trends, el término Python es el que está más relacionado cuando se realiza una búsqueda con el término machine learning (aprendizaje automático).

Los seminarios se organizaron en cuatro sesiones de tres horas cada uno. Estos seminarios tenían como objetivo presentar a los alumnos las herramientas básicas para poder enfrentarse con el problema de la competición. Fueron seminarios eminentemente prácticos, utilizando como herramienta principal los Notebooks de IPython, que representan un entorno de programación interactivo, en el cual se puede combinar la ejecución de código, texto enriquecido, LaTeX, gráficos, etc. (Perez & Granger, 2007). Los seminarios que se impartieron fueron:

- Seminario 1: introducción al manejo de datos (adquisición, filtrado, etc) y exploración estadística utilizando el módulo de python pandas (pandas, 2017).
- Seminario 2: introducción a los métodos supervisados de clasificación: regresión logística y support vector machines.
- Seminario 3: introducción a los métodos supervisados de regresión: regresión lineal
- Seminario 4: tópicos avanzados en aprendizaje automático: validación cruzada, componentes principales, selección de hiperparámetros.

Se realizó un último seminario (wrap-up seminar) para la entrega de premios, en el que los diferentes equipos presentaron las soluciones propuestas.

4. Resultados

Metodología de análisis

Uno de los principales problemas a los que nos enfrentamos era cómo evaluar el impacto que tiene la competición y los seminarios como herramienta para la enseñanza de aprendizaje automático. Propusimos la utilización de unas encuestas que los alumnos debían realizar tanto a la hora de matricularse en la competición como al finalizar la misma.

Ambos formularios estaban divididos en un conjunto de preguntas subjetivas y un test objetivo. En la parte subjetiva se indicaba a los alumnos que cuantificasen de 0 (desconozco el concepto) a 4 (conozco bien el concepto) sus conocimientos en aprendizaje automático y Python (como herramienta para aprendizaje automático). En concreto se les preguntaba por los conceptos de: entrenamiento supervisado y no supervisado, normalización de datos, conjunto de entrenamiento y de test, sobreajuste, validación cruzada, conocimiento del lenguaje Python, módulos pandas, numpy, scipy, matplotlib y sklearn. La parte subjetiva

se repetía exactamente en la primera y última encuesta. El objetivo era medir la percepción que tenían los alumnos con respecto a los conocimientos adquiridos.

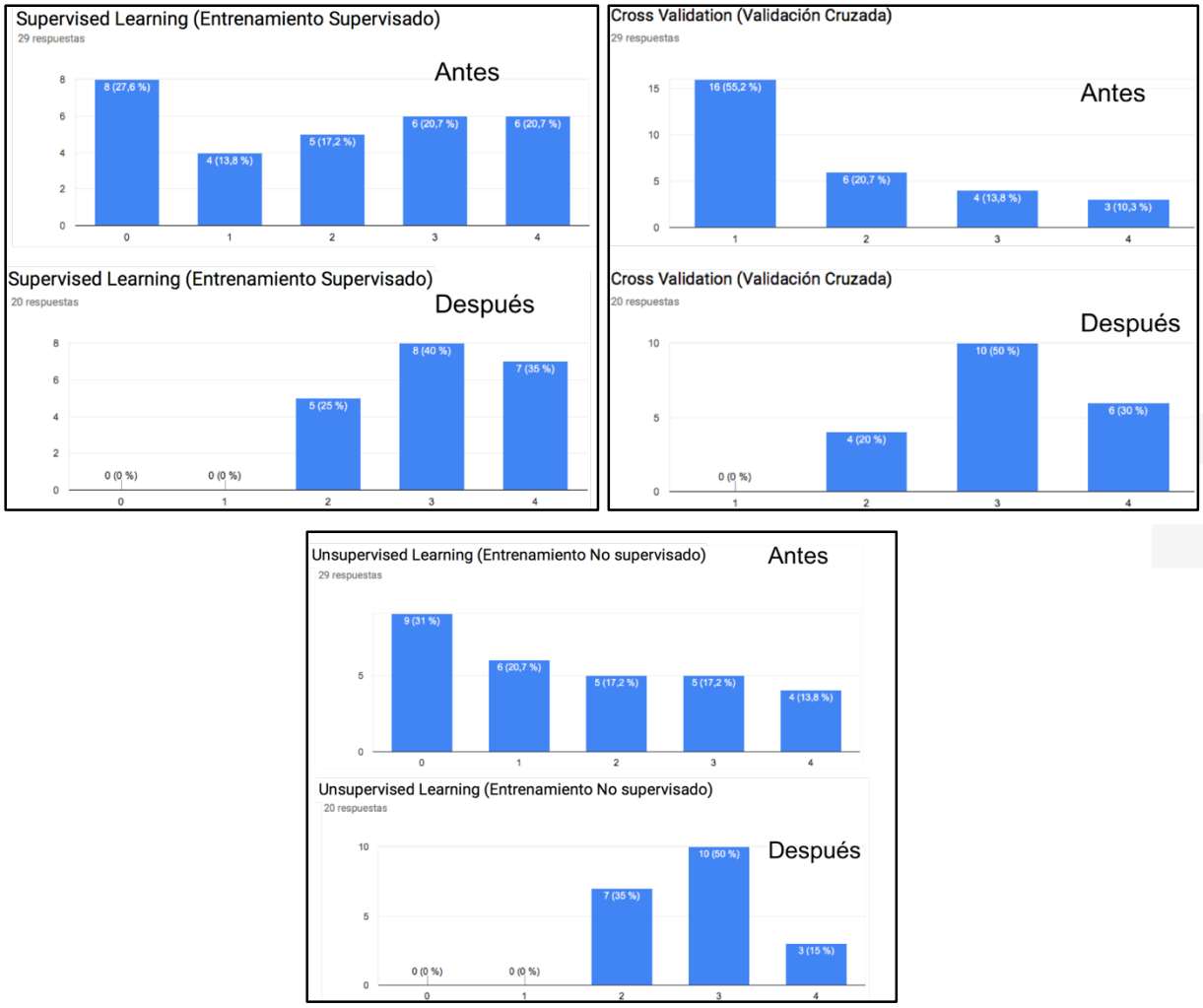


Figura 3. Resultado de las encuestas para evaluación de los conocimientos de aprendizaje automático, antes y después de la competición y seminarios.

En la parte objetiva de los formularios se realizaba un pequeño test para evaluar de forma objetiva los conocimientos de aprendizaje automático de los alumnos. En este caso el test era diferente para la primera y última encuesta. El objetivo era cuantificar los conocimientos adquiridos por los alumnos de forma objetiva.

Los formularios se pueden consultar en los siguientes enlaces: formulario 1: https://drive.google.com/open?id=1CO3hqJKL0ZvvihMd4uL1cuTcquEwQta21LFM8k5JZZw; formulario 2: https://drive.google.com/open?id=19xY6nWwSJEp2qTfwR6G82eoFC9BlOmYjJBwPbJ9TLE.

En la Figura 3 se pueden observar los resultados del formulario para las preguntas sobre el dominio que los alumnos tenían sobre los conceptos de aprendizaje supervisado y no supervisado (supervised and unsupervised learning), así como del concepto de validación cruzada (cross-validation), tanto antes como después de la competición y los seminarios. Se puede observar como el porcentaje de alumnos que percibía haber aumentado el conocimiento de estos conceptos es elevado.

Cabe destacar el manejo del concepto de validación cruzada. Desde la experiencia de los autores, el concepto de validación cruzada es uno de los más complicados de interiorizar por parte del alumno. Por lo general, se entiende la mecánica de esta técnica de forma sencilla (el concepto en sí no es muy complejo), y los alumnos son capaces de resolver prácticas aisladas implementando la técnica. Sin embargo, su utilización práctica de forma adecuada, y la imbricación dentro de un proyecto completo de aprendizaje automático, es una de las partes más complicadas de llegar a realizar correctamente. Esto se puede apreciar en la valoración que los alumnos tienen sobre la compresión de este concepto antes de la competición y los seminarios, pues se trata de uno de los conceptos que un mayor número de alumnos indicó como que no lo conocía suficientemente bien antes de la competición.

En cuanto a las preguntas objetivas, en todas las preguntas se apreció un incremento del porcentaje de respuesta correctas. Por ejemplo, en la pregunta en la que se evaluaba la adquisición del concepto de conjunto de entrenamiento y test, se pasó de un 72% de respuestas correctas a un 90%. En el caso de identificar problemas de clasificación, se pasó de un 50% de aciertos a un 100% de identificación correcta de un problema de clasificación. En general, en todas las preguntas objetivas se pudo constatar una mejora en los conocimientos de aprendizaje automático de los alumnos.

En el formulario final se animó a los alumnos a dejar comentarios o sugerencias sobre la competición y los seminarios. Un sentimiento generalizado fue el de que la competición ha supuesto un factor de motivación muy importante. El hecho de poder constatar cuál era la posición del equipo en el ranking les impelía a buscar por su cuenta, en internet, manuales, foros etc, soluciones nuevas para conseguir mejorar la posición en el ranking. En general, los alumnos interpretaban que esta forma de introducir los conceptos de

aprendizaje automático era muy entretenida y tenía mucho valor, por la implicación que tiene que poner el alumno en la adquisición de los conocimientos.

Los resultados de las encuestas muestran que el método propuesto resulta muy efectivo para que el alumno participe como principal motor en el aprendizaje de los conceptos. Adicionalmente, esta aproximación permite formar en los aspectos prácticos que habitualmente no se cubren en la enseñanza clásica, pero que resultan ser un hecho diferencial.

5. Equipo docente

Rebeca Goya Esteban, Profesora Ayudante Doctor de la Universidad Rey Juan Carlos. Doctora ("Doctor europeus", "cum laude") por la Universidad Rey Juan Carlos de Madrid (2014), Máster en Ingeniería Biomédica por la Universidad de Oporto, Portugal (2008) e Ingeniero Técnico de Telecomunicación por la Universidad Carlos III de Madrid (2006). Áreas principales de investigación: procesamiento digital de señales fisiológicas, análisis de series temporales, estudio de complejidad y dinámicas no lineales en señales y aprendizaje estadístico.

https://scholar.google.es/citations?user=Swn4GeEAAAAJ&hl=es
https://sites.google.com/site/rebecagoyaesteban/home

Óscar Barquero Pérez, Profesor Ayudante Doctor de la Universidad Rey Juan Carlos. Doctor ("Doctor europeus", "cum laude") por la Universidad Rey Juan Carlos (2014), Máster en Ingeniería Biomédica por la Universidade do Porto, Portugal (2008) e Ingeniero Técnico de Telecomunicación por la Universidad Carlos III de Madrid (2005). Áreas principales de investigación: análisis no lineal de series temporales, procesamiento de señales biomédicas y aprendizaje estadístico.

https://scholar.google.com/citations?user=zKdwDv8AAAAJ&hl=es