Estimación del riesgo de no superar una asignatura de evaluación continua mediante aprendizaje automático
Estimating the risk of not passing a continuous evaluation module by machine learning

Emilio Serrano, José Mario López, Damiano Zanardini
emilioserra@fi.upm.es, jose.lleiva@alumnos.upm.es, damiano@fi.upm.es

Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
Madrid. España

Resumen- Este trabajo aplica técnicas de minería de datos para estimar el riesgo de suspender la asignatura de Lógica impartida en la ETSIINF UPM (Ingeniería Informática), partiendo de las calificaciones obtenidas por el alumno dentro del proceso de evaluación continua. Lo que se pretende es cuantificar la probabilidad de un alumno tiene de suspender conociendo las calificaciones obtenidas, por ejemplo, durante el primer mes de curso. El conjunto de datos estudiado son las notas (parciales y finales) de los alumnos en los años anteriores. Se ha desarrollado una aplicación web para que el alumno pueda ingresar las calificaciones obtenidas hasta el momento y saber qué probabilidad tiene de aprobar finalmente la asignatura.

Palabras clave: minería de datos, predicción de resultados académicos, evaluación continua

Abstract- The present work uses data mining in order to estimate how likely it is that a student will fail the exam of Logic in the Computer Science Degree at the ETSIINF, UPM. This is done starting from his or her previous grades during the semester (continuous evaluation is used in this course). Previous knowledge used in the learning process comes in the form of grades obtained by students in previous years: based on this, data mining techniques extract relevant patterns and predict the probability for the current student to pass or fail. A web application has been developed, which allows a student to insert grades obtained so far (for example, during the first month) and see the probability to finally pass or fail the course according to the results of previous years.

Keywords: data mining, prediction of academic results, continuous evaluation

1. INTRODUCCIÓN

Este trabajo presenta un análisis de minería de datos realizado sobre la colección de calificaciones de la asignatura de Lógica impartida en la ETSI de Ingenieros Informáticos de la UPM. Se ha desarrollado una aplicación web que se vale de métodos predictivos utilizando el aprendizaje automático (la aplicación debe aprender del conjunto de notas) para formular un modelo que permita estimar el riesgo o probabilidad de que un alumno suspenda (o apruebe) la asignatura en base a las calificaciones obtenidas en actividades evaluables dentro del proceso de evaluación continua.

Se ha elegido la *regresión logística* como método de predicción en el contexto de la minería de datos debido a que es un método que calcula el grado de pertenencia a una clase. Sin embargo, también se han realizado pruebas con metaclasificadores más potentes como el "*random forest*". Previamente a la aplicación de estos métodos, se analizó el conjunto de datos y sus particularidades para proceder a un preproceso de los datos (parte del proceso de descubrimiento de conocimiento en bases de datos). Después de este preprocesamiento, se efectuó un filtrado de variables para proceder con el entrenamiento del modelo y los correspondientes análisis de bondad, que consisten en comprobar lo bueno y fiable que es el predictor obtenido.

Una vez realizada la parte de minería de datos, se ha elaborado una aplicación web en la que el alumno puede ingresar las calificaciones obtenidas en sus evaluaciones para conocer el riesgo de suspender.

Más allá de la aplicación concreta de la asignatura de lógica, este trabajo es extrapolable a otras asignaturas de evaluación continua. Estas calculadoras de riesgo son particularmente útiles en asignaturas que, como lógica, se encuentran planificadas en un primer curso universitario y en las que el estudiante suele ser sobre-optimista en la preparación que necesita para aprobar.

2. CONTEXTO

El uso de técnicas de minería de datos y la realización de una aplicación web se enmarca en el contexto de una asignatura de primero en un Grado en Ingeniería Informática. La asignatura de Lógica cuenta cada año con un gran número de alumnos (más de seiscientos en el último curso), lo que permite disponer de una gran cantidad de datos. A lo largo del curso los alumnos que optan por la modalidad "evaluación continua" realizar una serie de pruebas individuales o grupales.

A. Objetivos

Los objetivos principales de este trabajo son

- Realizar un seguimiento de las notas obtenidas por los alumnos, identificando las pruebas de mayor impacto de la tasa de fracaso.

DOI: 10.26754/CINAIC.2017.000001_118

- Favorecer que los alumnos tomen conciencia lo antes posible de lo que significa estudiar una asignatura de nivel universitario, y que puedan tomarse en serio el estudio después de darse cuenta de que no es fácil aprobar si no se empieza pronto a estudiar.

3. DESCRIPCIÓN

Este trabajo, basado en un conjunto de datos con las calificaciones de estudiantes en los últimos tres años, aplica paradigmas de aprendizaje supervisado, analizar sus salidas, y formula conclusiones de las mismas. Estas conclusiones se realimentan en la preparación de la asignatura para cursos posteriores y permite la creación de una calculadora del riesgo de no superar la asignatura.

B. Plan de Trabajo

El trabajo ha consistido en dos fases:

- Usar técnicas de minería de datos, en concreto el paradigma de la regresión logística, para estimar la probabilidad de que una variable aleatoria asuma cierto valor. En este caso, la regresión logística se ajusta muy bien al objetivo porque la variable estudiada sólo tiene dos posibles valores: "aprobado" (clase positiva) y "suspenso" (clase negativa); e interesa no una predicción categórica sino un la probabilidad de pertenencia a una u otra clase.
- Realizar una aplicación web para que los alumnos puedan hacer preguntas al sistema sobre la probabilidad de aprobar dados los resultados obtenidos hasta el momento.

C. Análisis de los datos: introducción

La minería de datos (Witten, 2011) se define como el proceso de encontrar conocimiento de interés (patrones, anomalías, clasificaciones y estructuras de datos) dentro de grandes colecciones de datos. La necesidad de estas técnicas surge de la existencia de volúmenes de datos cada vez más grandes que no pueden ser procesados manualmente. De manera genérica, un proceso de descubrimiento de conocimiento se desglosa en los siguientes pasos (Fayyad, Piatetsky-Shapiro, & Smyth , 1996):

Limpieza de datos: consiste en la eliminación de datos faltantes, corruptos, que añaden ruido innecesario o simplemente no son relevantes para el análisis que se busca llevar a cabo. La limpieza debe garantizar cierta calidad de los datos: su exactitud, integridad, entereza, validez, consistencia y uniformidad, densidad y unicidad.

Integración de datos: es el proceso en que datos provenientes de fuentes de información heterogénea se integran en un único lugar.

Selección de datos: es el proceso en el que se obtiene de la base de datos la información realmente relevante.

Transformación de datos: se trata de dar a los datos un formato apropiado para el proceso de minería de datos.

Minería de datos: es el proceso donde se aplican métodos inteligentes para extraer los patrones de datos.

Evaluación de patrones: se busca encontrar los patrones relevantes para el análisis deseado, identificando además las variables dentro del conjunto de datos que son más importantes y cómo se mide la importancia de esas variables.

Presentación de la información: se trata de presentar el conocimiento obtenido de manera coherente y contrastada.

Es primordial realizar un estudio minucioso del conjunto de datos en sí: saber de dónde provienen, cómo han sido tomados y almacenados, determinar las variables existentes, así como tratar las particularidades de instancias concretas (variables faltantes, significado de la falta de las mismas), que permitan decidir acerca de la elección de una técnica que permita realizar un proceso de minería de datos exitoso. En definitiva, lo que se busca es determinar cómo se comportan las variables para calcular la variable de clase, es decir, aquella que recoge la clasificación de una instancia en el contexto del conjunto de datos que se está analizando.

D. Metodología y fundamentos teóricos

En este proyecto se ha elegido una metodología de trabajo iterativa e incremental que permita seguir los pasos ordenados del proceso de Knowledge Discovery from Data (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth , 1996) descrito anteriormente, para interpretar la salida y los resultados obtenidos en cada etapa y al final, formular conclusiones generales sobre todo el trabajo.

Ha sido necesario un entorno que permitiera realizar operaciones relacionadas con la minería de datos. Este entorno es *RStudio*[1]**,** un entorno de programación de código abierto para el desarrollo de proyectos en lenguaje R. El lenguaje R es uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo también de amplia difusión en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras. La encuesta anual de KDNuggets sobre software de análisis de datos actualizada recientemente (Piatetsky, 2017), muestra que Python y R son las herramientas más empleadas en análisis de datos (52,6%, y 52,1%, respectivamente).

El algoritmo de aprendizaje automático empleado ha sido la *regresión logística* (Hosmer & Lemeshow, 1989), un modelo de regresión utilizado para la clasificación binaria.

Una *regresión lineal* es un modelo matemático utilizado para encontrar la relación de dependencia entre una variable dependiente Y y las variables independientes X_1,.., X_n. Su entrenamiento supone el cálculo de los *coeficientes de regresión* β_0, …, β_p tales que

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

donde ε representa las perturbaciones aleatorias o factores no controlables en los datos. Estos coeficientes identifican una función lineal que aproxima los datos minimizando el error obtenido al comparar la predicción con la etiqueta numérica de cada caso. Como se puede ver en la Figura 1, el modelo constituido por una línea recta intenta acercarse lo más posible a los puntos que son los datos. Es decir, se predicen valores de la variable dependiente Y en relación con la única variable independiente X.

[1] https://www.rstudio.com/

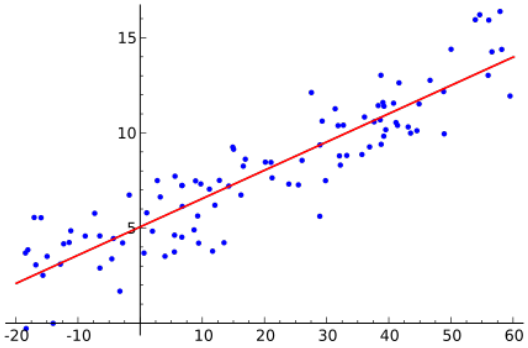

Figura 1. Ejemplo de regresión lineal.

La *regresión logística* esencialmente formula un problema de clasificación binaria (predecir si un caso pertenece a la clase positiva o negativa) como un problema de regresión (Rohrer, 2016). Para ello, se calcula como de probable es pertenecer a la clase positiva. Muchas de las calculadoras de riesgo que se utilizan en campos como la medicina o las aseguradoras utilizan la regresión logística porque ofrece una solución intuitiva a una clasificación donde se pretende no sólo dar una respuesta blanco/negro a la predicción sino también modelar áreas grises. En el trabajo que nos ocupa, se realizará una clasificación binaria siendo las dos clases "Aprobado" y "Suspenso".

El entrenamiento de una regresión logística puede reducirse al cálculo para de una regresión lineal para la función logit de la probabilidad como muestra la siguiente ecuación:

$$Y = 1/(1 + e^{-(\beta 0 + \beta 1 X1 + \ldots + \beta n Xn)})$$

La predicción del modelo Y está acotada entre 0, predicción de clase negativa, y 1, predicción de clase positiva. Como muestra la Figura 2, la regresión logística no sólo da las predicciones extremas sino todos los valores intermedios sin cambios abruptos en la clasificación.

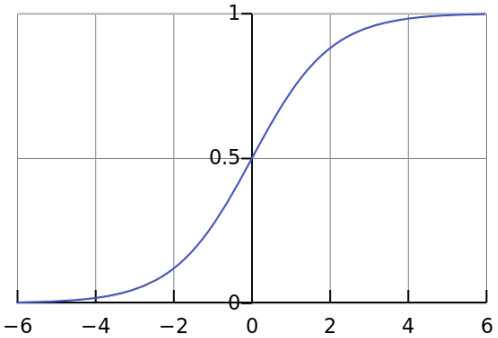

Figura 2. Ejemplo de regresión logística.

E. Preprocesiamiento de los datos

Los datos han sido tomados por los profesores de la asignatura de Lógica durante un periodo de tres años, y han sido tabulados en una hoja de cálculo donde cada columna representa una de las variables (en este caso, las variables independientes corresponden a actividades evaluables dentro de la asignatura), y cada una de las líneas representa una instancia (observación) anonimizada. La modalidad de evaluación dentro de la asignatura de refleja en las variables. La asignatura está dividida en dos bloques: Lógica Proposicional y Lógica de Primer Orden, y en cada bloque se realizan actividades de evaluación tanto individuales como grupales.

LPindividual: Nota individual (NI) obtenida en el bloque temático de Lógica Proposicional (LP). Tiene un peso de al menos 35% sobre la nota final (existe una fórmula para calcular los pesos de cada actividad de evaluación, cuya explicación no es necesaria para comprender este trabajo; en todo caso, se trata de modificar los pesos de cada nota para evitar que un alumno se aproveche del trabajo de sus compañeros de grupo) (Hernández, 2016).

LPgrupo: Nota obtenida en las actividades grupales (trabajos en grupo, ejercicios grupales en clase, etc.) del bloque de LP. Tiene un peso de hasta el 15%.

LPrepesca: Nota obtenida en el examen de recuperación del bloque de LP; este nota sustituye la nota correspondiente a LPindividual. El examen no es obligatorio y excluyente con LPOrepesca, por lo que un valor desconocido puede indicar: que no había necesidad porque ya se tenía esa parte aprobada, o un no presentado porque se ha dejado la asignatura, o que se ha decidido optar por la repesca de LPO.

LPnota: nota final del bloque de LP. Redundante con las anteriores y por tanto omitida en el entrenamiento del modelo.

LPOindividual: Nota individual obtenida en el bloque temático de Lógica de Primer Orden (LPO). Tiene un peso de al menos 35% sobre la nota final.

LPOgrupo: Nota obtenida en las actividades grupales del bloque de LPO. Tiene un peso de hasta el 15%.

LPOrepesca: Nota obtenida en el examen de recuperación del bloque de LPO. El examen no es obligatorio y excluyente con LPrepesca, por lo que un valor desconocido puede indicar: que no había necesidad porque ya se tenía esa parte aprobada, o un no presentado porque se ha dejado la asignatura, o que se ha decidido optar por la repesca de LP.

LPOnota: nota final del bloque de LPO. Redundante con las anteriores y por tanto omitida en el entrenamiento del modelo.

NotaFinal: nota final de la asignatura, que viene dada por la media aritmética entre LPnota y LPO nota. Redundante con las anteriores y por tanto omitida en el entrenamiento del modelo.

De cara a una mejor comprensión de los resultados para su transferencia a otros docentes que imparten la materia lógica en primer curso, se puede encontrar un ejemplo de prueba de LP y de LPO en el siguiente enlace: https://goo.gl/fcfVV7.

Las notas de cada bloque se calculan como media pesada entre las notas individuales (o las repescas) y las grupales. Los dos bloques tienen el mismo peso para la nota final.

F. Limpieza y transformación

Dentro del conjunto de datos existen campos vacíos, sin que se pueda por esto clasificar directamente el dato como "actividad no realizada". Por ejemplo, la falta de un dato en la columna LPrepesca puede significar que el alumno no ha necesitado mejorar la nota LPindividual presentándose al examen de repesca, o no ha podido presentarse (de hecho, sólo

se permite realizar una de las repescas). Una fase de limpieza de los datos se encarga de mejorar los datos para que puedan ser usados en la fase de análisis propiamente dicha. Por ejemplo, si un alumno tiene un suspenso en ambas pruebas individuales (LPindividual y LPOindividual) y aun así no se ha presentado a ninguna de las dos repesca, esto significa que *ha decidido* no presentarse, por lo que lo más razonable es sustituir un 0 a las notas faltantes en ambas repescas. Por otro lado, si se hubiera presentado a la repesca del primer bloque (LPrepesca), sería más apropiado sustituir la nota faltante en LPOrepesca con un valor que no significase necesariamente una nota mala (por ejemplo, la media obtenida por los demás alumnos en LPOrepesca).

G. Selección de variables

Normalmente es necesario hace un filtrado de variables para que la solución implementada sea computacionalmente eficiente. Lo que se busca es un conjunto óptimo de variables predictoras, es decir, identificar y eliminar las variables redundantes o irrelevantes. En este trabajo, dada la naturaleza de los datos, sería contraproducente eliminar variables demasiado pronto ya que, en principio, todas las notas son relevantes para calcular la nota final de la asignatura. Aún así, se ha realizado un estudio de filtrado (algo que proporcionan las herramientas de minería de datos utilizadas) para conocer detalles que puedan ser interesantes para una mejor interpretación de los resultados. Este estudio ha detectado que las variables que, como ya sabíamos, algunas notas que son agregados de otras deben ser omitidas por ser redundantes (nota bloque LP, bloque LPO, asignatura). Por tanto, el conjunto de las variables independientes pasa a tener 6 predictores.

H. Modelo predictivo

Se ha utilizado el algoritmo de aprendizaje automático elegido, la regresión logística, para entrenar un modelo que prediga la clase (es decir, el valor de la variable dependiente Y) asociando una cierta probabilidad.

Para asegurar que las métricas de calidad del modelo sean realistas, este se debe evaluar con casos que no hayan sido considerados en el entrenamiento. En este trabajo se ha usado una *validación cruzada repetida* de 10 pliegues y con tres repeticiones. En la *K-fold cross validation*: se particiona el conjunto de datos en k segmentos iguales llamados pliegues (folds), y uno de los pliegues se usa para validar mientras que los otros k-1 se usan para entrenar el modelo. El proceso se repite k veces y se calcula una traza de rendimiento (precisión) del modelo para cada subconjunto de manera que se agrega el error cometido. En la variante con repeticiones, este proceso es repetido un número determinado de veces para mitigar los efectos de la aleatoriedad inherente en el proceso de segmentación de la validación cruzada.

Tras el entrenamiento de la regresión logística, la probabilidad de que una instancia pertenezca a una clase positiva (es decir, que el resultado final sea “aprobado” es dada por la siguiente fórmula:

$$P(Y = "aprobado" \,|X_j) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{j1} + \cdots + \beta_k X_{jk})}}$$

Se observa que esta probabilidad depende de los coeficientes obtenidos en la fase de entrenamiento, y además de los valores de las variables independientes (es decir, las notas obtenidas previamente). Como explicado con anterioridad, las variables X_1,.., X_6 corresponden a las seis calificaciones que constituyen el conjunto de las actividades de evaluación dentro de la asignatura de Lógica omitiendo agregados de estas.

Además de con la regresión logística, se ha experimentado con un random forest o selva aleatoria de 500 árboles. Este meta-clasificador calcula numerosos árboles de decisión que observan un subconjunto de los datos de entrenamiento (tanto en casos como en variables) y que votan sobre la clase predicha. Los random forests son una las técnicas de aprendizaje automático que mejor resultado dan en muchas de las competiciones de análisis de datos realizadas en Kaggle[2] junto al Boosting. La contrapartida respecto a la regresión logística es que se pierde mucha interpretabilidad: es más difícil revisar 500 árboles de decisión que una sencilla formula con pesos por cada variable predictora.

Los experimentos alcanzan una precisión (accuracy) del 92% utilizando validación cruzada repetida con 10 pliegues y 3 repeticiones. Los coeficientes obtenidos se muestran en la Tabla 1.

Tabla 1. Coeficientes de la regresión logística de mayor a menor.

LPOindividual	1.3965
LPgrupo	0.8929
LPindividual	0.7101
LPOgrupo	0.5206
LPOrepesca	0.2153
LPrepesca	0.2134

Estos datos indican que la segunda prueba individual (correspondiente a la variable LPOindividual) es la actividad de evaluación que más impacto tiene sobre la nota final. Aunque tenga el mismo peso en la fórmula para calcular la nota, LPindividual es mucho menos relevante. Esto parece sorprendente porque, al fin y al cabo, la primera prueba parcial es la que casi todo el mundo hace, esté preparado o no, por lo que sería de esperar que la nota obtenida fuera muy relevante para la nota final. Es decir, debería haber una correlación muy fuerte entre una mala nota en la primera prueba parcial y el suspenso final, porque una parte de los alumnos simplemente habrán dejado la asignatura. En cambio, la gran relevancia de la variable LPOindividual parece más bien indicar que los alumnos que reciben una calificación baja en el primer parcial se dan, por así decirlo, una segunda posibilidad, y es el segundo parcial que tiene un gran impacto en el resultado final.

El modelo de random forest supera el porcentaje de predicciones correcta y alcanza el 95%. Además, la Tabla 2 muestra el impacto de las variables en la predicción.

[2] https://www.kaggle.com/

Tabla 2. Impacto de variable en predicción para random forest.

	Mean Decrease GINI
LPOindividual	35.57
LPOrepesca	22.04
LPindividual	11.69
LPOgrupo	10.67
LPrepesca	9.38
LPgrupo	5.77

Al igual que en la regresión logística, LPOindividual es la prueba que más impacto tiene para decidir el aprobado final. Sin embargo, en este segundo modelo el examen de repesca de LPO aparece como segundo factor de mayor importancia subiendo desde el quinto puesto en la regresión logística. Además, la nota de grupo de LP tiene un impacto casi nulo. Si bien estos resultados son más intuitivos y consiguen mayor acierto en las predicciones, una vez más, la contrapartida es un modelo más complejo y menos interpretable que la regresión logística.

I. Aplicación web

Se ha implementado como aplicación web una calculadora que, dadas las calificaciones obtenidas en un subconjunto de las pruebas de evaluación, estima la nota final prevista y la probabilidad de suspender. Se trata principalmente de una herramienta sencilla para aplicar los datos obtenidos en el análisis. La Figura 3 muestra la interfaz de la calculadora de riesgo.

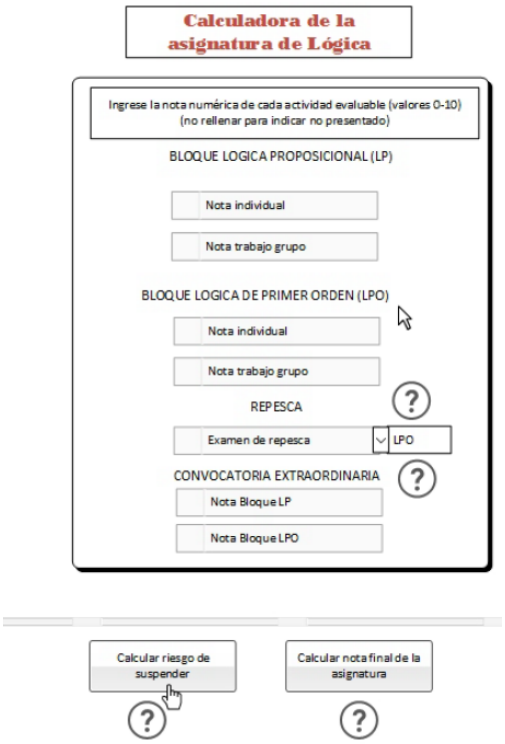

Figura 3. Interfaz de la calculadora.

La Figura 4 muestra un ejemplo de salida en la que se estima el riesgo de no superar la asignatura en un 60%, es decir, un 10% por encima de la predicción de aprobado. Por ello, se anima al estudiante a ponerse al día.

Gracias a esta aplicación web, los estudiantes pueden comprobar desde las primeras pruebas, el riesgo que sufren de no pasar la asignatura con las calificaciones conocidas. Además, también se da la opción de calcular la nota conocidas todas las calificaciones y de consultar estadísticas descriptivas de años anteriores.

Figura 4. Respuesta de la calculadora.

4. CONCLUSIONES

Este trabajo ha mostrado un proceso de descubrimiento de conocimiento en bases de datos (KDD) aplicado a la generación de una herramienta web que prediga el riesgo de no superar una asignatura de evaluación continua. Esta herramienta se basa en paradigmas de aprendizaje automático como la regresión logística o los random forests. La motivación de una herramienta de estas características es que los estudiantes, y particularmente aquellos que accedan a los primeros cursos universitarios, puedan cuantificar el riesgo que asumen al dejar de seguir la evaluación continua.

En el caso concreto de la asignatura analizada, Lógica, se consigue entre un 92% y un 95% de predicciones correctas (*accuracy*). La validación cruzada con repetición asegura que ninguna muestra usada en el entrenamiento del modelo ha sido usada para su evaluación. El factor más importante en el cálculo de riesgo ha sido el segundo examen parcial de la asignatura (nota individual de lógica de primer orden).

Si bien estos resultados son de gran interés para estudiantes y profesores de la asignatura analizada, para el público docente general el interés de este trabajo radica en que los resultados aquí expuestos son extrapolables de manera directa a otras asignaturas una vez se cambien los datos de entrenamiento y se realice un preproceso similar.

Tanto los datos, como el análisis realizado en lenguaje R, y el código fuente de la aplicación web; están disponibles para el lector interesado bajo petición a los autores.

AGRADECIMIENTOS

Este trabajo de investigación es financiado por la Universidad Politécnica de Madrid bajo el proyecto de innovación educativa: "Métodos, experiencias y herramientas para el aprendizaje experiencial de la Ciencia de Datos"; y por el Ministerio de Economía, Industria y Competitividad en el ámbito del proyecto "Datos 4.0: Retos y soluciones – UPM" (TIN2016-78011-C4-4-R, AEI/FEDER, UE).

REFERENCIAS

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth , P. (1996). From Data Mining to Knowledge Discovery: An Overview. En *Advances in Knowledge Discovery and Data Mining.* American Association for Artificial Intelligence.

Hernández, P. (2016). *Guía de la Asignatura de Lógica. https://www.upm.es//comun_gauss/publico/guias/2016-17/1S/GA_10II_105000002_1S_2016-17.pdf.*

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* Wiley.

Piatetsky, G. (Mayo de 2017). *New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll.* Obtenido de http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Rohrer, B. (Enero de 2016). *What questions can data science answer?* Obtenido de http://www.kdnuggets.com/2016/01/questions-data-science-answer.html

Witten, I. F. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier Science.