REXE. Revista de Estudios y Experiencias en Educación
ISSN: 0717-6945
ISSN: 0718-5162
rexe@ucsc.cl
Universidad Católica de la Santísima Concepción
Chile

Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento

Martínez-Abad, Fernando; Hernández-Ramos, Juan Pablo
Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento
REXE. Revista de Estudios y Experiencias en Educación, vol. 2, núm. Esp.2, 2018
Universidad Católica de la Santísima Concepción, Chile
Disponible en: https://www.redalyc.org/articulo.oa?id=243156768012
DOI: https://doi.org/10.21703/rexe.Especial3201812514512

Investigación

Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento

Data mining techniques with free software for the detection of factors associated with performance

Fernando Martínez-Abad fma@usal.es
Universitad de Salamanca, España
Juan Pablo Hernández-Ramos juanpablo@usal.es
Universitad de Salamanca, España

REXE. Revista de Estudios y Experiencias en Educación, vol. 2, núm. Esp.2, 2018

Universidad Católica de la Santísima Concepción, Chile

Recepción: 12 Julio 2017
Aprobación: 29 Julio 2017

DOI: https://doi.org/10.21703/rexe.Especial3201812514512

Redalyc: https://www.redalyc.org/articulo.oa?id=243156768012

Resumen: La potencia de cómputo que permiten los equipos informáticos en la actualidad, unido a la existencia de información y datos masivos en todos los ámbitos sociales, incluido el educativo, exige el desarrollo y aplicación de técnicas estadísticas y software informáticos que faciliten la obtención de información significativa en estos universos de datos y su transformación en conocimiento útil para la sociedad. Partiendo de un proyecto de investigación en desarrollo actualmente, este trabajo presenta el potencial del software estadístico Weka para desarrollar análisis estadísticos de información masiva a partir de bases de datos de evaluaciones a gran escala, que permite aplicar técnicas de Minería de Datos, consideradas como parte de las técnicas del denominado Big Data. Así, se muestra una propuesta para el aprovechamiento de software informático en el análisis y detección de información no trivial entre la inmensidad de los datos disponibles. De esta manera, se presenta a la comunidad científica una serie de procedimientos y técnicas estadísticas que pueden ser valiosas y replicables en otros ámbitos educativos y/o sociales, concluyendo el trabajo con una propuesta de transferencia del conocimiento generado a la sociedad en general y a los agentes educativos en particular

Palabras clave: Big data, minería de datos, evaluación, software libre, valor añadido en educación.

Abstract: The computational capacities allowed by current computer equipment, coupled with the availability of mass data in all areas, including Education Sciences, demand the development and application of statistical techniques and software that help in obtaining meaningful information on these mass data and that facilitate the data transformation into useful knowledge for society. The collection of meaningful information in these data universes and its transformation into useful knowledge for society. Based on a research project under development, this paper presents the potential of the Weka statistical software to develop statistical analysis from massive large-scale evaluation databases. In this context, Weka allows researchers to apply techniques of Data Mining, considered within the techniques of the so-called Big Data. Thus, this work shows a proposal for the use of Weka software in the analysis and detection of nontrivial information between the immensity of the available data. In this way, this study presents to the scientiic com- munity a set of statistical procedures and techniques that can be valuable and replicable in mul- tiple educational and social ields. The conclusions relect on the possibilities of the transference of the knowledge generated to the society in general and to the educational agents in particular.

Keywords: Big data, data mining, assessment, freeware, added value in education.

1. INTRODUCCIÓN

Este trabajo se enmarca dentro del proyecto 'Detección de buenas prácticas educativas en escuelas de alto valor añadido mediante técnicas de Big Data', incidiendo en el valor del software libre Weka, disponible de manera libre y gratuita para usuarios de sistemas Windows, Mac y Linux, para la aplicación de técnicas estadísticas de minería de datos (Data Mining). Este proyecto, actualmente en fase de desarrollo, pretende aprovechar el potencial de estas técnicas de minería de datos, incluidas dentro del actualmente conocido como Big Data, para obtener información que pueda ser considerada no trivial a partir de bases de datos masivas en educación (Han, Pei y Kamber, 2011), esto es, evaluaciones a gran escala como las pruebas PISA (OECD, 2013; OECD, 2016), las evaluaciones PIRLS (Mullis, Martin, Kennedy, Trong y Sainsbury, 2009) o los exámenes TIMSS (Educational Resources Information Center & National Science Foundation, 1996). Des- de una perspectiva global, el proyecto en el que se enmarca este estudio pretende detectar aquellos factores de proceso, también llamados no contextuales (Martínez- Abad, Chaparro Caso y Lizasoain Hernández, 2014), que se muestren de manera generalizada en centros educativos de alto valor añadido, esto es, centros cuyo rendimiento escolar en las áreas curriculares evaluadas sea superior al esperable en función de los factores de entrada o contextuales (Joaristi Olariaga, Lizasoain Hernández y Azpillaga Larrea, 2014), en relación con los centros en los que el rendimiento escolar sea inferior al esperable en función de estos factores (centros de bajo valor añadido).

1.1 Software Weka

Weka (Wakaito Environment for Knowledge Analysis) es un entorno informático desarrollado por la Universidad de Wakaito, que está ideado para la aplicación y evaluación de técnicas de las conocidas comúnmente como de minería de datos (Data Mining). Más específicamente, el software de libre distribución Weka, que está construido en código abierto bajo un lenguaje Java, permite trabajar en el preprocesado, clasificación, agrupación, asociación, predicción y visualización de las bases de datos, incorporando múltiples técnicas y algoritmos de análisis que no facilitan otros software comerciales y libres más convencionales como el SAS, SPSS, Winstat, etc.

A pesar de que la entrada de datos en Weka no resulta intuitiva, este software se destaca de entre otros proyectos de software de libre disposición como R, en que posee una interfaz gráfica por defecto que permite la navegación por ventanas para configurar las herramientas y acceder a los análisis disponibles. Por esta razón, Weka se postula como una herramienta informática que puede ser empleada de manera simple para el análisis bases de datos educativas masivas, o evaluaciones a gran escala, como las provistas por la Organización para la Cooperación y el Desarrollo Económico (OCDE), o la propia Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA).

Weka dispone principalmente de 3 grupos de algoritmos para el análisis de datos:

• Reglas de asociación: Empleadas para buscar relaciones entre sucesos nominales, habitualmente cuando el interés es realizar un estudio exploratorio de un conjunto muy amplio de variables. Los algoritmos de asociación tratan de identificar la ocurrencia conjunta de varios sucesos, permitiendo extraer información acerca de cómo la ocurrencia o no ocurrencia de algunos sucesos puede inducir la aparición de otros.

• Algoritmos de clustering: Empleados para analizar tendencias de respuesta o comportamientos comunes en grupos de sujetos. Así, estos algoritmos agrupan sujetos maximizando la homogeneidad entre sí y la heterogeneidad con respecto al resto, a partir generalmente de una variable criterio de agrupación.

• Algoritmos de clasificación: Estos son los algoritmos de minería de datos más frecuentemente empleados en el ámbito de las Ciencias Sociales y de la Educación (Martínez Abad y Chaparro Caso, 2017). Se aplican con el objetivo de construir modelos predictivos, por lo que es necesario definir una variable criterio categórica o categorizada y una o más variables predictoras nominales, ordinales o cuantitativas.

1.2 Minería de datos

Como expresábamos anteriormente, el Data Mining es un conjunto de técnicas estadísticas que permiten discriminar entre variables predictoras más o menos importantes en función de una variable criterio (Nghe, Janecek y Haddawy, 2007), en este caso rendimiento. Así, la minería se datos dirige su objetivo principal a extraer información sobre la relación entre variables a partir de grandes cantidades de datos de diversa naturaleza. Dicho de manera simplificada, estas técnicas tratan de explotar datos masivos para extraer la información más valiosa de los mismos.

Mientras que a partir de finales de los años 80 del siglo XX, y principalmente durante los años 90, se comienzan a asentar estas técnicas en el ámbito de las ingenierías, matemáticas o la economía (Carter y Hamilton, 1995; Houtsma y Swami, 1995; Marquez, Shack-Marquez y Wascher, 1985), la minería de datos no empieza a suscitar interés en el ámbito de las Ciencias de la Educación hasta bien entrado el siglo XXI (Alcover et al. 2007; Hsieh, 2013; Ma, 2005; Nghe et al., 2007; Schumacher, Olinsky, Quinn y Smith, 2010). Lo que es más, surge durante estos años el término Minería de Datos Educativos (MDE), para hacer referencia al empleo de estas técnicas y conjunto de algoritmos para la búsqueda, análisis y extracción de información valiosa para la definición de modelos predictivos en el ámbito educativo (Ballesteros, Sánchez-Guzmán y García, 2014).

A pesar de que la MDE se ha extendido en el ámbito científico de manera considerable, y numerosos investigadores defienden su potencial para la detección de factores asociados al rendimiento en estudios a gran escala en los cuales el volumen de los datos enturbia y dificulta todo el proceso de localización de información no trivial (Castro y Lizasoain Hernández, 2012), fundamentalmente con técnicas inferenciales tradicionales, estas técnicas se han desarrollado de manera marginal en comparación con las tradicionales (Alcover et al. 2007; Nghe et al., 2007; Osmanbegoic y Suljic, 2012; Schumacher et al., 2010,

Tekin, 2014). De hecho, la aplicación de estas técnicas en el ámbito de la educación obligatoria es prácticamente nula (Hsieh, 2013; Kiray, Gok y Bozkir, 2015; Ma, 2005).

1.3 Objetivos del estudio

En este punto debemos distinguir entre los objetivos del proyecto de investigación en el que se enmarca este trabajo y los objetivos de este estudio en concreto.

En cuanto al proyecto de investigación marco, podemos plantear como objetivo principal la detección de factores asociados al rendimiento a partir de la aplicación de técnicas de minería de datos con sotware libre Weka en escuelas de alto valor añadido para la elaboración y difusión a la comunidad de un catálogo de buenas prácticas educativas.

Derivados de este objetivo general, podemos hablar de 3 objetivos clave:

1. Aplicación de modelos jerárquicos lineales para la detección de escuelas de alto y bajo valor añadido en base a evaluaciones a gran escala.

2. Aplicación de técnicas de Big Data para la detección de factores asociados al rendimiento en las escuelas de alto valor añadido.

3. Diseño de un catálogo de buenas prácticas educativas y difusión de resultados a la comunidad educativa y científica.

En cuanto al objetivo concreto de este trabajo, lo podemos definir como mostrar el potencial de las técnicas de Data Mining en general, y del sotware libre especializado Weka en particular, para la detección de factores asociados al rendimiento a partir de la información disponible en las evaluaciones educativas a gran escala. Así, en las siguientes líneas se expondrán las características y posibilidades del software Weka, intercalando la exposición con algunos ejemplos prácticos.

2. METODOLOGÍA

El proyecto en el que se enmarca este estudio se desarrolla bajo una perspectiva puramente cuantitativa, con interés exploratorio y correlacional, a partir de un diseño no experimental o ex-postfacto. No se incluye dentro de los objetivos de la investigación, por tanto, manipular las variables estudiadas, sino analizarlas directamente en su contexto natural, de cara a detectar las relaciones significativas que puedan aportar pistas sobre los principales factores asociados a la eficacia escolar.

Tomando como referencia la población de estudiantes de 15 años, cursando en el momento de la recogida de información el segundo ciclo de Educación Secundaria Obligatoria en España, se parte de la muestra de estudiantes (n1=37.205), profesores (n2=4.286) y centros educativos españoles (m=980) incluida en programa de evaluación a gran escala PISA 2015.

Las variables estudiadas son, como variable criterio la identificación del centro como de alto o bajo valor añadido y como variables predictoras las variables de proceso igualmente incluidas en PISA, tanto a nivel de estudiante, como a nivel de profesor y de escuela. Por ejemplo, a nivel de escuela, se trabaja con algunas variables predictoras como la autonomía en la toma de decisiones de director, equipo directivo o

profesores, la existencia de clases extraescolares en el centro, el liderazgo del equipo directivo en cuanto al desarrollo curricular o instruccional, la participación de los profesores y familias en la vida y organización del centro, las evaluaciones internas y ex- ternas existentes, etc. A nivel de profesor se emplean algunas variables como la satisfacción del profesor con su labor profesional y con la escuela, las actividades de formación continua en las que participa, su formación de base, los problemas de material y de formación que detecta en el profesorado, la coordinación entre los profesores del claustro, las técnicas de evaluación que emplea en el aula, etc. Por último, a nivel de estudiante existen numerosas variables predictoras, como el apoyo parental a los estudios, la motivación hacia el aprendizaje, la ansiedad ante las actividades académicas, el sentimiento de pertenencia al centro, el gusto por la cooperación en el aula y el valor que el estudiante le otorga, las actividades de ocio y de manejo de TIC que el estudiante realiza en el centro y fuera del centro, actividades extraescolares y deberes realizados, etc.

En cuanto al análisis de datos, merece la pena remarcar que la aplicación de las técnicas de Minería de Datos se lleva a cabo en la segunda fase del proyecto, tras la obtención del indicador de eficacia de los centros. Este indicador resulta de la aplicación de modelos jerárquicos lineales, o modelos multinivel, en los que, a partir de la incorporación de variables de entrada o contextuales como variables predictoras en el modelo, y de las variables de rendimiento en lectura, matemáticas y ciencias como variables dependientes (Y), se obtienen modelos estadísticos multivariantes (Gamazo, Martínez-Abad, Olmos-Migueláñez y Rodríguez-Conde, 2018). Estos modelos facilitan la predicción del rendimiento medio de un centro educativo a partir de sus características contextuales o de entrada concretas (Nivel socio-económico de las familias, índice de estudiantes repetidores, multiculturalidad, recursos del centro, etc.). De esta manera, es posible determinar, a partir del residuo de cada centro (rendimiento medio del centro menos puntuación de la predicción para el mismo), un índice que informa de si el rendimiento medio del centro está por encima o por debajo de lo que sería esperable dadas sus características contextuales, y cuánto por encima y por debajo se sitúa de esta predicción. A este residuo es a lo que denominamos eficacia, y es lo que se toma como variable dependiente o criterio en el estudio de minería de datos posterior.

3. RESULTADOS

3.1 Reglas de asociación

Como ya se ha indicado previamente, las reglas de asociación se emplean para buscar relaciones o asociaciones entre un conjunto muy numeroso de sucesos, normalmente dicotómicos (el suceso ocurre o no ocurre). El algoritmo fundamental que incorpora Weka al respecto es el algoritmo A priori, que establece las reglas teniendo en cuenta el soporte de los datos a la regla y la confianza de la propia regla. El soporte se refiere al número de sujetos que están incluidos en la regla, y la confianza al

porcentaje de los mismos, de entre el total del soporte, que cumplen la regla:

$$Sop\left(A \Rightarrow B\right) = P\left(A \cap B\right)$$

$$Conf\left(A \Rightarrow B\right) = P\left(B \mid A\right) = \frac{P\left(A \cap B\right)}{P\left(A\right)}$$

A nivel general Weka entiende que, bajo un nivel de confianza dado, una regla será más interesante cuanto mayor sea el soporte bajo el que esté generada. Así, el algoritmo comienza buscando las reglas que alcancen una mayor confianza cuyo soporte sea superior, bajando dentro de un mismo nivel de confianza el nivel de soporte hasta el límite fijado. Cuando se alcanza el límite mínimo de soporte, se vuelven a generar normas para el siguiente nivel de confianza inferior.

Este tipo de algoritmos son empleados habitualmente, por ejemplo, en la determinación si un cliente de una empresa (por ejemplo, un banco) tiene altas probabilidades de adquirir un producto nuevo que la empresa ha sacado al mercado. Otro ámbito en el que se emplea habitualmente es en establecer relaciones entre la compra de ciertos productos en un supermercado, estableciendo qué productos es más probable que estén presentes en la lista de la compra si la persona ha adquirido otros; de este modo, el supermercado puede establecer un protocolo a la hora de ordenar y juntar los productos en los estantes del local.

3.2 Algoritmos de clustering

Recordemos que el objetivo de las reglas de clustering no era otro que establecer agrupaciones de sujetos en función de su afinidad en las tendencias de las puntuaciones de un grupo de variables. Existen gran cantidad de algoritmos de clustering, siendo los 3 más habitualmente empleados en Weka los siguientes:

- Clustering Numérico (k-medias): Empleado con variables estrictamente cuantitativas, ya que asigna al sujeto al clúster correspondiente empleando la distancia euclidea al centroide del grupo como medida de agrupación. El proceso se itera, sujeto a sujeto, hasta que todos los sujetos se mantienen en el mismo centroide (figura 1).

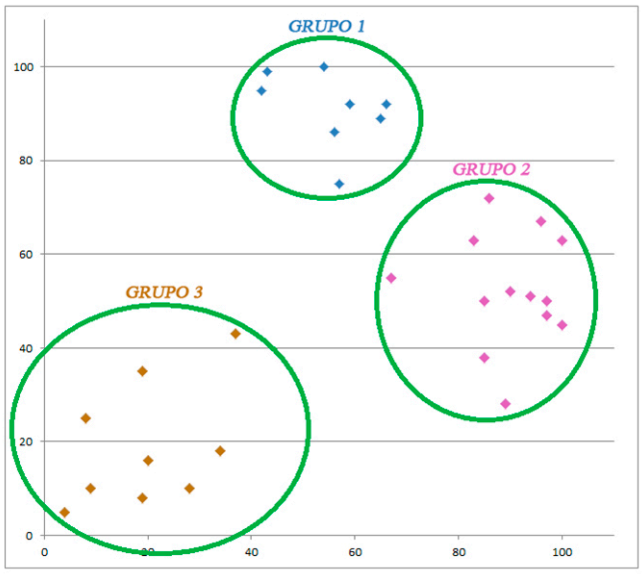

Figura 1.
Técnica clustering k-medias para 2 variables predictoras.

Clustering Conceptual (COBWEB): Algoritmo empleado cuando disponemos de variables predictoras cualitativas, ya que entiende los clústeres como distribuciones de probabilidad sobre el espacio de las puntuaciones de los sujetos, calculando los puntos de corte en las variables (ya sean categóricas o numéricas) maximizando la distancia entre grupos. En realidad, COBWEB genera árboles de clasificación.

• Clustering Probabilístico (EM): Algoritmo en el que, a diferencia de los dos anteriores modelos, el resultado no depende del orden en el que estén presentados los sujetos en la base de datos ni tiende a sobreajustar los clústeres obtenidos en las muestras de entrenamiento. En lugar de buscar sujetos parecidos entre sí de manera iterativa, lo que intenta EM es buscar el grupo de clústeres más probables dado un conjunto de puntuaciones.

Estos algoritmos de agrupación permiten dos tipos de análisis:

• Exploración descriptiva: testeo de las agrupaciones que aseguren una mayor homogeneidad intragrupo y una mayor heterogeneidad intergrupo a partir de una serie de variables de entrada.

• Evaluación a partir de una variable criterio (Classes to cluster evaluation): probar si un conjunto de variables de entrada puede emplearse como predictor de una variable de clase o variable criterio, generando unas agrupaciones conforme al criterio en base a las puntuaciones obtenidas en las variables predictoras.

En ambos casos, se pueden generar modelos empleando la muestra completa como muestra de entrenamiento, o estableciendo algún tipo de control del sobreajuste a partir de validaciones. La validación de los datos a partir de submuestras es altamente recomendable, ya que los procedimientos de Clustering (al igual que las técnicas de clasificación) tienden a generar modelos que sobreestiman la verdadera relación entre la variable criterio y las variables predictoras. Las principales técnicas para la validación de datos que nos ofrece Weka son:

• División de la muestra (Supplied test set o Percentage split): Se establece de antemano una submuestra que será considerada muestra de entrenamiento (muestra a partir de la que se genera el modelo con las normas de agrupación), a partir de la que se genera el modelo principal. Ese modelo es contrastado a partir de la otra submuestra, que es considerada simplemente para esta validación.

• Validación cruzada (Cross-validation; sólo disponible en las técnicas de clasificación): Se genera en primer lugar el modelo de clasificación, contando con la muestra completa como muestra de entrenamiento. Posteriormente, se divide a la muestra en k submuestras y el modelo es testeado en todas ellas. Los resultados de la validación se muestran indicando una media aritmética de los índices de ajuste obtenidos en cada una de las k submuestras.

En la figura 2 se puede observar la interfaz del menú disponible para obtener las anteriores especificaciones.

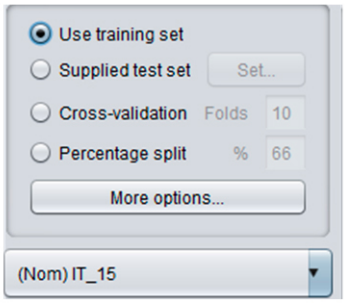

Figura 2.
Validaciones y selección de variable criterio.

3.3 Algoritmos de clasificación

El objetivo de estas técnicas es construir un modelo predictivo, que sea capaz de establecer con la mayor precisión posible en qué valor se encontrará el sujeto en la variable criterio a partir de la información obtenida con otras variables que podrían ser consideradas como predictoras. Entre otros, Weka permite la utilización del algoritmo J48, cuyo empleo está muy extendido en los estudios que incluyen este tipo de técnicas. Entre los parámetros estimados bajo este procedimiento, destaca el nivel de confianza establecido para la poda del árbol generado, confidence level, puesto que influye notoriamente en el tamaño y capacidad de predicción del árbol construido.

Se podría explicar el clasificador J48 de la siguiente manera: para tomar la decisión sobre el corte realizado en la iteración 'n', se busca la variable predictora y el punto de corte exacto en el que el error cometido es más bajo (tomando como criterio una variable preestablecida), siempre y cuando nos encontremos en niveles de confianza superiores a los establecidos previamente. Una vez realizado el corte, el algoritmo vuelve a repetirse, hasta que ninguna de las variables predictoras alcance un nivel de confianza superior al establecido. Se destaca la importancia de trabajar con el nivel de confianza, ya que, en caso de tener un gran número de sujetos y variables, este árbol puede resultar demasiado grande. Otra forma

de limitar el tamaño del árbol es especificando el mínimo número de instancias por nodo.

La figura 3 muestra un ejemplo de árbol de decisión obtenido a partir de bases de datos masivas, en concreto en el contexto mexicano. En el marco de este estudio, se aplicó este conjunto de técnicas para analizar la importancia de factores asociados al rendimiento, incorporando como variable criterio el rendimiento medio de las escuelas, categorizado en bajo, medio-bajo, medio, alto y excelente.

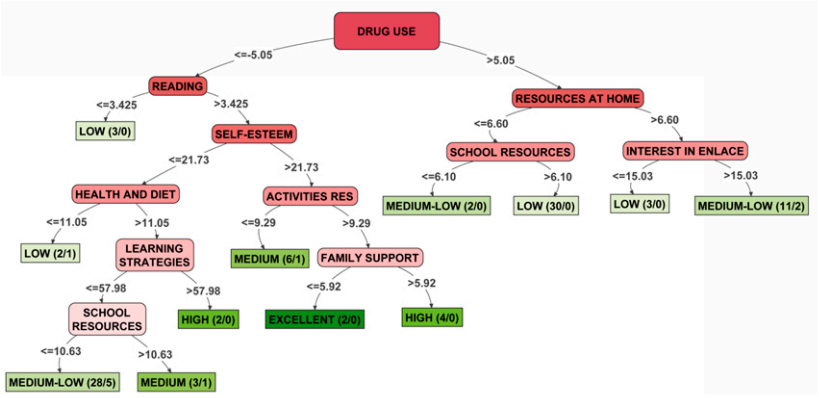

Figura 3.
Ejemplo de árbol de decisión obtenido con el algoritmo J48 (Martínez Abad y Chaparro Caso López, 2017, p. 49).

En cuanto a los resultados obtenidos a partir de este procedimiento, Weka provee varios indicadores de la bondad de ajuste del modelo cuyo análisis resulta de gran interés:

- Porcentaje de instancias bien clasificadas a nivel global.
- Índice Kappa de Cohen: Índice de bondad de ajuste del modelo completo. Se entiende que el modelo es apropiado si este valor supera el valor 0.7.
- Error absoluto medio: Estadístico que indica el error de estimación cometido. Valores más bajos indican que el modelo es más apropiado.
- Porcentaje del área bajo la curva ROC: La curva ROC indica, en un eje de abscisas y ordenadas, la relación entre la sensibilidad (verdaderos positivos entre el total de positivos) y especificidad (verdaderos negativos entre el total de negativos) del modelo de clasificación establecido. Se establece una curva para cada categoría, en relación con el resto de categorías, indicando la capacidad del modelo para detectar casos pertenecientes a esa categoría de la variable criterio. Este porcentaje nos indica, por lo tanto, la precisión que tiene el modelo para identificar correctamente a los sujetos de un grupo, teniendo en cuenta tanto el porcentaje de acierto en la detección de esa categoría y el porcentaje de desaciertos al identificar a sujetos de esa categoría.
- Matriz de confusión: Tabla de contingencia que relaciona la clasificación dada por el modelo con el valor real que alcanza el sujeto en la variable criterio.
- Precisión: Porcentaje de instancias bien clasificadas de entre todas las seleccionadas como positivas.

4. DISCUSIÓN Y CONCLUSIONES

Los paquetes informáticos estadísticos para el análisis de datos han ido evolucionando de forma paralela a la creciente capacidad de procesamiento de los equipos informáticos, adaptándose a las necesidades emergentes de la sociedad actual y a la capacidad de obtención y análisis de grandes bases de datos. Así, surgen y evolucionan las técnicas de Minería de Datos, que permiten obtener y detectar información no trivial a partir de datos masivos (Nghe et al., 2007). Este estudio presenta el software Weka como una alternativa sencilla para el investigador aplicado de las Ciencias Sociales, mostrando un ejemplo de aplicación a través de un proyecto de investigación en el ámbito educativo, en el que puede ser provechosa.

Así, con la descripción llevada a cabo en este trabajo del potencial de las técnicas y algoritmos principales que incluye Weka, parece que se ha dado una respuesta adecuada al objetivo planteado inicialmente, cuyo interés no era otro que el de proveer al investigador aplicado en Ciencias de la Educación de una herramienta sencilla de análisis de datos que le permita agregar valor añadido a los análisis aplicados tradicionalmente, principalmente cuando se enfrenta a bases de datos a gran escala.

En este sentido, dadas las posibilidades que ofrecen estas técnicas (Martínez Abad y Chaparro Caso López, 2017), y su emergencia en numerosas áreas afines, se espera un crecimiento importante de estudios dentro del ámbito de Ciencias de la Educación que empleen este conjunto de técnicas para extraer información significativa a partir de datos extensos (Ballesteros et al., 2014).

Parece que estas técnicas no sólo complementan las tradicionales, sino que ofrecen un abanico de posibilidades añadidas que debe ser explorado y aprovechado. Por ejemplo, ya se ha evidenciado la capacidad de los árboles de decisión para detectar efectos de interacción en subgrupos de sujetos, en capas o niveles profundos del árbol o en ciertas regiones del mismo (Castro y Lizasoain, 2012), cuestión que puede abrir la puerta a la detección más nítida de factores asociados al rendimiento. Por su parte, Baradwaj y Pal (2011) detectaron que a través de las técnicas de clustering se hace posible identificar regiones densas y esparcidas del objeto de análisis, descubriendo patrones de distribución y correlaciones sobre los atributos de los datos difícilmente localizables con los índices más habituales de relación.

En suma, el presente trabajo se ha forjado con la pretensión de aportar información valiosa que apoye el conocimiento y la utilización de la minería de datos en la investigación aplicada, principalmente en el ámbito de las Ciencias de la Educación. Cabe señalar al respecto algunos puntos débiles del estudio, principalmente en lo relacionado con su aplicabilidad directa, ya que se presenta una reflexión meramente teórica y técnica. No obstante, la fortaleza principal de este trabajo reside en su disposición simple como manual aplicado, ofreciendo soluciones sin establecer reflexiones o análisis muy profundos sobre los fundamentos matemáticos e informáticos de la aplicación analizada.

Cabe, por tanto, de cara a futuros estudios, desarrollar análisis más prácticos de la herramienta Weka, comparando su potencial en la detección de factores asociados al rendimiento con otras técnicas clásicas o más habituales. De esta manera, se podrá analizar de manera más objetiva el valor añadido que puede suponer su integración entre los análisis estadísticos habituales dentro de este ámbito de conocimiento.

AGRADECIMIENTOS

Proyecto realizado con la Beca Leonardo a Investigadores y Creadores Culturales 2017, Fundación BBVA (la Fundación BBVA no se responsabiliza de las opiniones, comentarios y contenidos incluidos en este documento).

Referencias

Alcover, R., Benlloch, J., Blesa, P., Calduch, M. A., Celma, M., Ferri, C., y Zúnica, L. R. (2007, July). Análisis del rendimiento académico en los estudios de informática de la Universidad Poli- técnica de Valencia aplicando técnicas de minería de datos (pp. 163-170). Paper presented at XIII Jornadas de Enseñanza Universitaria de la Informatica.

Ballesteros, A., Sánchez-Guzmán, D., & García, R. (2014). Minería de datos educativa: Una he- rramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. Latin-American Journal of Physics Education, 17, 662-668.

Baradwaj, B. K., & Pal, S. (2011). Mining Educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications, 2(6), 63-69.

Carter, C. L., & Hamilton, H. J. (1995). A fast, on-line generalization algorithm for knowledge discovery. Applied Mathematics Letters, 8(2), 5-11. Doi: 10.1016/0893-9659(95)00002-8

Castro, M., & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos lineales. Revista Española de Pedagogía, 70, 131-148.

Educational Resources Information Center, N. C. for E. S., & National Science Foundation. (1996). Third International Mathematics and Science Study (TIMSS). Washington D.C.: U.S. Dept. of Education.

Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., & Rodríguez-Conde, M. J. (2018). Evaluación de factores relacionados con la eicacia escolar en PISA 2015. Un análisis multinivel. Revista de educación, (379), 56-84.

Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed). Amsterdam: Elsevier.

Houtsma, M., & Swami, A. (1995). Set-oriented data mining in relational databases. Data and Knowledge Engineering, 17, 245-262. Doi: 10.1016/0169-023X(95)00024-M

Hsieh, M. (2013). Data mining from education databases examine the factors impacting the school performance in the United States. International

Journal of Intelligent Technologies and Ap- plied Statistics, 6, 135-143. Doi: 10.6148/IJITAS.2013.0602.03

Joaristi Olariaga, L., Lizasoain Hernández, L., & Azpillaga Larrea, V. (2014). Detección y caracterización de los centros escolares de alta eicacia de la Comunidad Autónoma del País Vasco mediante Modelos Transversales Contextualizados y Modelos Jerárquicos Lineales. ESE : Estudios Sobre Educación, (27), 37–61.

Kiray, S. A., Gok, B., & Bozkir, A. S. (2015). Identifying the factors afecting science and mathematics achievement using data mining methods. Journal of Education in Science, Environment and Health (JESEH), 1, 28-48.

Ma, X. (2005). Growth in Mathematics achievement: Analysis with classiication and regression trees. The Journal of Educational Research, 99, 78-86.

Marquez, J., Shack-Marquez, J., & Wascher, W. L. (1985). Statistical inference, model selection and research experience. A multinomial model of data mining. Economics Letters, 18, 39-44. Doi: 10.1016/0165-1765(85)90075-8

Martínez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. School Efectiveness and School Improvement, 28(1), 39– 55. Doi: 10.1080/09243453.2016.1235591

Martinez-Abad, F., Chaparro Caso López, A. A., & Lizasoain Hernández, L. (2014). The socioeconomic index in the analysis of large-scale assessments: Case study in Baja California (Mexico). In Proceedings TEEM' 14. Technological Ecosystems for Enhancing Multiculturality (pp. 461–467). Salamanca: ACM.

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). PIRLS 2011 Assessment Framework. Amsterdam: International Association for the Evaluation of Educational Achievement.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In 37th ASEE/IEEE Frontiers in Education Conference (pp. T2G-7–T2G- 12).

OECD. (2013). PISA 2012 results. París: OECD Publishing OECD. (2016). PISA 2015 results. Paris: OECD Publishing.

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 10 (1), 3-14.

Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. (2010). A comparison of logistic regression, neural networks, and classiFication trees predicting sucess of actuarial students. Journal of Education for Business, 85, 258-263.

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. Eurasian Journal of Educational Research, 54, 207-226.