Encuentro Internacional de Educación en Ingeniería ACOFI
Una formación de calidad en ingeniería para el futuro
Centro de Convenciones Cartagena de Indias
15 al 18 de Septiembre de 2015

EDUCATIONAL DATA MINING (*EDM*) PARA LA DETERMINACIÓN DE COMPORTAMIENTOS EN ESTUDIANTES DE INGENIERÍA DEL MODELO UDE@

Adrián Montoya Lince, Jesús Francisco Vargas Bonilla, Lyda Yaneth Contreras Olivares

Universidad de Antioquia
Medellín, Colombia

Resumen

Se describe la experiencia de la aplicación de técnicas de *EDM* (*clustering)* a un curso disponible en la plataforma Ude@ de la Universidad de Antioquia. El objetivo es clasificar los patrones de interacción de los estudiantes a partir de la información almacenada en la base de datos de *Moodle*. Para ello, se generan informes sobre el uso de los recursos y la autoevaluación que permiten analizar el comportamiento y los patrones de navegación de los estudiantes durante el uso del *LMS*.

Palabras clave: *EDM, Learning Analytics, LMS, K-Means, clustering, Moodle*

Abstract

We describe the experience of the application of a EDM technique(clustering) in a course of the University of the Antioquia at Ude@ e-learning platform. The goal is to realize a classification of patterns behavior of students by using the information saved in database of an LMS (Moodle). Finally, reports about the use of resources and autoevaluation are generated which allow to analyze the behavior and navigation patterns of the students during the use of LMS.

Keywords: *EDM, Learning Analytics, LMS, K-Means, clustering, Moodle*

1. Introducción

La era del conocimiento, la ciencia y la incorporación de las tecnologías de la información y las comunicaciones (TIC) dentro de la sociedad han transformado los esquemas de producción de contenidos, almacenamiento y disposición de la información y, por tanto, los conceptos de enseñanza y aprendizaje.

El Ministerio de Educación Nacional ha promovido políticas, como el Plan Decenal de Educación (2006-2016), para resaltar la necesidad de establecer compromisos con el fin de promover, desarrollar y fomentar el uso de las TIC en el entorno educativo, contribuyendo al fortalecimiento de la capacidad de innovación en la educación colombiana.

En este sentido, la Unidad de Virtualidad Ude@ de la Universidad de Antioquia ofrece programas de pregrado virtuales en ingeniería bajo un modelo educativo centrado en el estudiante, donde el docente-tutor lo acompaña y estimula al análisis y la reflexión conjunta para aprender, reconocer la realidad y reconstruirla, teniendo presente el logro de los objetivos propuestos. Para que esto suceda, es primordial la interacción continua y la comunicación sincrónica y asincrónica entre docentes-tutores, compañeros (pares) y monitores, así como el uso del amplio abanico de recursos y ayudas educativas que se ponen a disposición, a través de la plataforma LMS-Moodle y WizIQ.

Los sistemas de enseñanza virtual han empezado a aplicar técnicas de minería de datos como herramienta para mejorar el aprendizaje de los estudiantes demostrando su alta efectividad (AlShammari, et al., 2013 & C. Romero, et al., 2007 & Siti Khadijah, et al., 2013). Desde un punto de vista tecnológico, la educación virtual exige de los servidores que soportan las plataformas de contenido (*LMS y LCMS*) robustez y mayor capacidad de almacenamiento, permitiendo así el resguardo de todas las interacciones y modificaciones que se realicen en la plataforma (Mazza & Milani, 2005). Esta información es valiosa para las instituciones ya que al ser analizada, puede ayudar a mejorar aspectos de esta modalidad de estudio, tanto en diseño y contenido de la plataforma virtual, como el acceso de los estudiantes, buscando favorecer los métodos de estudio y, en consecuencia, el rendimiento en los cursos (Talavera & Gaudioso, 2004 , Zaiane & Luo, 2001).

2. Trabajos relacionados

De acuerdo a (C. Romero & S. Ventura, 2007) la minería de datos en educación (*EDM*) permite responder preguntas sobre qué sabe realmente un estudiante y cómo está aprendiendo. De esta manera, *EDM* permite descubrir información útil que ayuda a los profesores y coordinadores de las instituciones interesadas en determinar la manera más pertinente para guiar a sus estudiantes, maximizando su aprendizaje.

Según (Ryan S.J.d. Baker, 2011) *EDM* involucra cinco métodos: predicción,

agrupamiento, minería de relación, destilación y descubrimiento de modelos. Cada uno de ellos con un objetivo y aplicación diferente como se resume en la tabla I.

Realmente, todos los procesos y técnicas involucradas en las actividades descritas en la tabla I se suelen denominar de diversas formas según el objeto de estudio. Por ejemplo: *EDM, Learning Analytics (LA), Big Data, Text Mining, Knowledge Discovery in Databases* (*KKD*), entre otros. Sin embargo, en el presente trabajo usaremos el término *EDM* de una forma genérica para denominar todas estas actividades. Cabe mencionar que realmente *EDM* se enfoca en el desarrollo de nuevas técnicas y herramientas para el descubrimiento de patrones en los datos involucrados en el aprendizaje, mientras que *LA* aplica dichas técnicas y herramientas para analizar los datos recolectados y crear aplicaciones que tienen una influencia directa sobre el proceso de enseñanza-aprendizaje (Mining, T. E. D., 2012).

TABLA I. TÉCNICAS DE MINERÍA DE DATOS EN EDUCACIÓN

Técnica	**Objetivos**	**Aplicaciones**
Predicción	Desarrollo de un modelo que pueda inferir una variable a partir de la combinación de los datos disponibles	Detección del comportamiento del estudiante con base en lo observado en otros alumnos con características similares. Predicción y entendimiento de los resultados académicos de un estudiante.
Agrupamiento (*Clustering*)	Encontrar conjuntos de datos que se agrupen naturalmente, separando el conjunto completo en una serie de categorías.	Agrupar a los usuarios de acuerdo a su comportamiento de navegación. Agrupar páginas web por su contenido, tipo o acceso. Identificar grupos de estudiantes con base en sus estilos cognitivos.
Minería de relaciones	Modelado de un fenómeno mediante predicción, agrupamiento o ingeniería del conocimiento, es usado como componente en una futura predicción o minería	Descubrimiento de asociaciones entre cursos ofrecidos según sus contenidos. Descubrimiento de estrategias pedagógicas que guíen en un proceso más efectivo de aprendizaje. Descubrir relaciones o asociaciones entre distintas páginas Web visitadas.
Descubrimiento mediante modelos	Modelado de un fenómeno mediante predicción, agrupamiento o ingeniería del conocimiento. Es usado como componente en una futura predicción o minería de relaciones	Descubrimiento de relaciones entre el comportamiento de los estudiantes y sus características. Análisis de parámetros de investigación para una amplia variedad de contextos
Destilado de datos	Los datos son destilados para permitir a un humano identificar o clasificar rápidamente propiedades de los datos	Identificación humana de patrones en el aprendizaje de los estudiantes. Etiquetado de datos para su uso en desarrollos posteriores de modelos predictivos.

Estudios previos en (Ratnapala, I. P., et al, 2014, Pereira, R. T., et al, (2013), Espigares P, M. J et at, 2011, Olague S, Juan et at, 2010 , Anduela Lile, 2011, López Guarín Camilo Ernesto, 2013, Timarán P, R et al 2013) resumidos en la tabla II, muestran la efectividad de la aplicación de técnicas de *EDM* para la clasificación y agrupación de estudiantes con objetivos que van desde la monitorización de comportamientos, predicción de la deserción hasta desarrollo de modelos y

descubrimiento de nuevo conocimiento.

El objetivo del presente estudio está centrado en el descubrimiento de comportamientos comunes de los estudiantes en el uso de la plataforma Moodle, por lo tanto se requiere aplicar una técnica de agrupamiento (*clustering* en inglés) para segmentar el espacio dado en un número adecuado de grupos homogéneos que comparten un número de propiedades y características similares. Para tal efecto se usó el algoritmo *K-means* (MacQueen, 1966), ampliamente utilizado por su robustez y eficacia (Kaur, N. et al, 2012). Para su implementación se escogió *R*[1] un software libre que permite una integración con el *LMS*, lo que a su vez permitirá la automatización del proceso en el servidor, sin ninguna intervención del usuario.

TABLA II. RESUMEN DE ARTÍCULOS *EDM*.

Nombre	***Autores & Año***	***Descripción***
Students behavioural analysis in an online learning environment using data mining	Ratnapala, I. P., et al, 2014	Aplicación de clustering K-means en varios cursos Moodle usando Weka.
Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos	Pereira, R. T., et al, 2013	Usa técnicas de clasificación (árboles de decisión J48) y clustering K-means para descubrir perfiles socioeconómicos y académicos de los estudiantes que desertan, usando software Weka.
Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia	López Guarín Camilo Ernesto, 2013	Aplicación de técnicas de agrupamiento y clasificación para el análisis de datos académicos de estudiantes de Ingeniería Agrícola e Ingeniería de Sistemas.
Minería de datos educativos en plataformas virtuales de aprendizaje musical	Espigares Pinazo, M. J., & García Pérez, R. (2011).	Utilizó *EDM* aplicada a Moodle para el aprendizaje musical online, con la técnica de *Clustering,* para observar el comportamiento de las actividades. Se obtuvo que las herramientas más utilizadas son los foros y los chats.
Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales: Un caso de estudio en el norte de Coahuila	Olague S, Juan et at, 2010	Se aplicó *EDM* a pruebas VARK a un curso de programación de computadores en *Moodle,* concluyendo que el estilo de aprendizaje de los estudiantes se describe dentro de las categorías: kinestésico-auditivo, visual-kinestésico-lectoescritura y kinestésico-auditivo-visual-lectoescritura.

[1] http://www.r-project.org

Analyzing E-Learning Systems Using Educational Data Mining Techniques	Anduela Lile, 2011	Se analiza un curso de programación de C en Moodle, usando varias técnicas de *EDM* para identificar los procesos de enseñanza más eficaces que se puede utilizar para mejorar el proceso educativo, usando RapidMiner y Weka.

3. Metodología

En concordancia con el estándar *CRISP-DM* (*Cross-Industry Standard Process for Data Mining,* por sus siglas en inglés) (Chapman P, et al., 2000) la metodología usada en el desarrollo del proyecto contempló 4 pasos:

Paso 1. Entendimiento del modelo y datos: se realizó una exploración de la base de datos de un curso creado en *Moodle*. Teniendo conocimiento de cómo y qué información se almacena, se procedió a seleccionar las tablas y variables que se tomarán como base para la aplicación de la técnica de *clustering*. La tabla III resume y describe las tablas seleccionadas.

TABLA III. TABLAS SELECCIONADAS DE LA BASE DE DATOS DE *MOODLE*

Tablas	Recursos que almacenan
mdl_resource	Imágenes, documentos en PDF, hojas de cálculo, archivos de sonido, archivos de
mdl_url	Enlaces web.
mdl_page	Páginas que son creadas por los profesores en *HTML*.
mdl_forum	Información de los foros que crean usuarios.
mdl_forum_discussions	Interacciones que los usuarios tienen con los diferentes foros que están almacenados en
mdl_quiz	Quices, tareas, autoevaluaciones, entro otros.
mdl_log	Todas las interacciones que los usuarios realicen con la plataforma.

En cuanto a las variables escogidas se tomaron dos componentes principales: los recursos visitados (*Página, Archivo, url*) y los foros que fueron vistos y modificados, de esta manera se describe el comportamiento del curso al agrupar las interacciones como se muestra en la figura
1. Las variables escogidas se denotan como: ***VerR*** (Ver Recurso), ***VerF*** (Ver Foro), ***VerU*** (Ver Url), ***VerP*** (Ver Página), ***EscF*** (Modificar Foro).

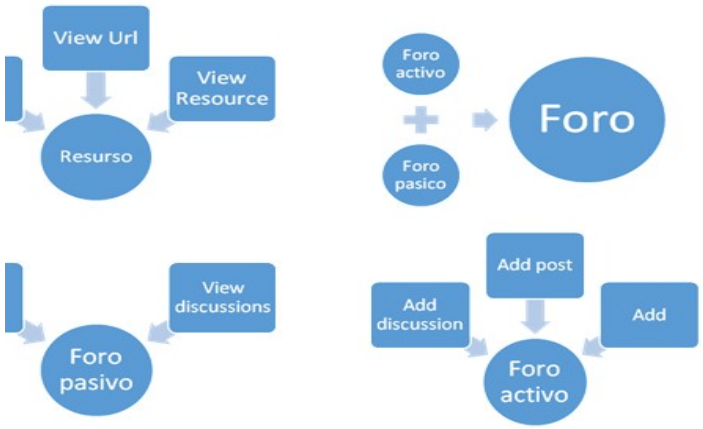

Figura 1. Clasificación de los recursos

Paso 2. Preparación de datos: los datos de dichas variables se recopilan a través de una consulta *SQL* a la base de datos de un curso real de ude@ que es cargado a la versión de *Moodle* instalada en el *PC* de escritorio y son exportados en un archivo con extensión *CSV*. Debido a la gran carga computacional que esto requiere por la cantidad de información en la base de datos, fue necesario realizar consultas de manera separada, exportando al final cinco tablas que contienen toda la información. Ya que los datos almacenados en el archivo CSV son de tipo alfanumérico, fue necesario programar un *script* en *R* para preprocesar los datos.

Paso 3. Modelado (*Clustering*):
Con la tabla de datos depurada se procedió primero a un análisis estadístico de las interacciones más frecuentes, de los recursos definidos en el curso. Luego se procedió a la aplicación del algoritmo *K-means* en el software *R* arrojando como resultado 4 *clusters* con un error tolerable en la suma de los cuadrados de las distancias entre los centroides. Finalmente se hace un análisis de las evaluaciones (Quices) registradas en el curso.

Paso 4. Visualización y análisis de resultados
Al realizar el análisis sobre los resultados, se determinó que la variable *Foro* es altamente significativa para la descripción del comportamiento del curso, por lo que se realizó la minería en primer lugar para todo el grupo de instancias y luego se hizo sin tener en cuenta dicha variable, pudiendo de esta manera analizar el comportamiento de los estudiantes, solo teniendo en cuenta los recursos que visitan.

4. Resultados.
En la figura 2 se observa un 95% de interacciones relacionadas con la visita a los foros del curso analizado, mientras que solo el 5% corresponden a operaciones de escritura de comentarios en ellos. Esto implica que el comportamiento de los estudiantes es mayoritariamente pasivo.

Figura 2. Representación de las interacciones del foro acuerdo al tipo de participación

La figura 3 hace referencia a la participación que tienen los estudiantes sobre los Recursos clasificados mediante las variables Página, Archivo, Url. Se observa que el 58% de las interacciones lo tiene la variable Página. En este recurso se pueden

almacenar un sinnúmero de actividades de diferente naturaleza (texto, audio, video). Este resultado lleva a la conclusión de que el curso está construido bajo la estructura de páginas web, donde ha sido subida la mayor parte del contenido de este.

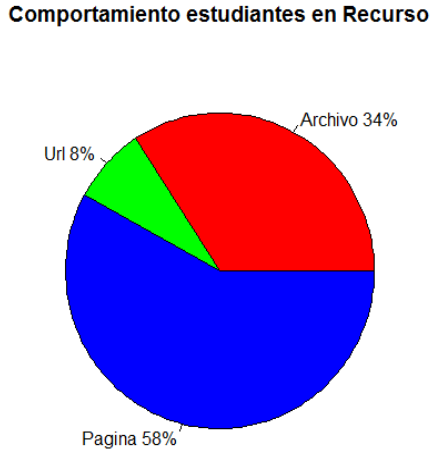

Figura 3. Representación porcentual de las interacciones con los recursos utilizados

Al aplicar el algoritmo *K-means*, se obtienen las agrupaciones se muestra en las figuras 4 (Foro incluido) y 5 (sin Foro incluido).

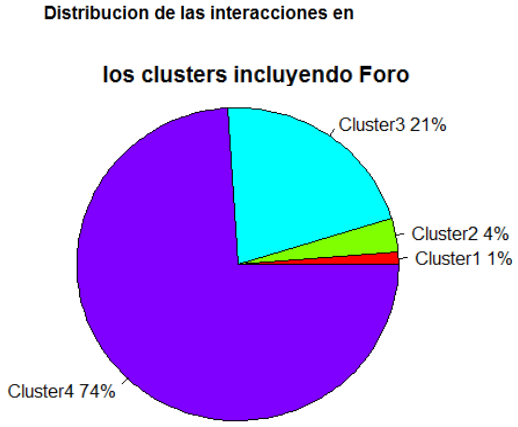

Figura 4. Distribución porcentual de las interacciones en los *clusters* teniendo en cuenta Foro

De la figura 4 se observa que el 74% de las interacciones han sido agrupadas en el *Cluster 4*, lo cual lo hace el *cluster* representativo. De la Figura 5 se observa que el *Cluster 2* tiene aproximadamente la mitad de las interacciones.

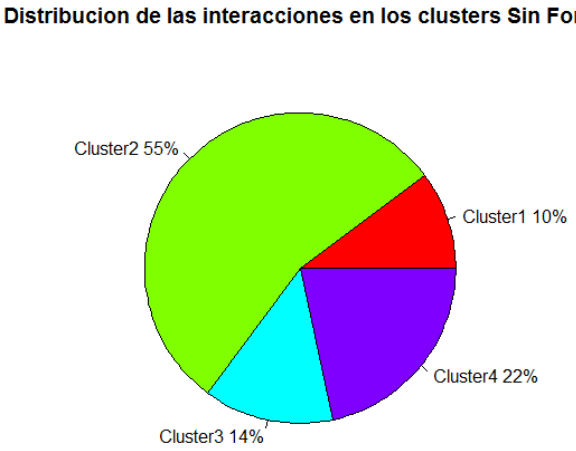

Figura 5. Distribución porcentual de las interacciones en los *clusters* sin incluir Foro

Las figuras 6 y 7, representan para cada *cluster* cuáles recursos están por encima de la media de cada recurso, con lo cual se logra identificar el comportamiento representativo de cada *cluster* sobre los recursos. Es de notar que en la Figura 6, el *Cluster 2* que cuenta solo con un 4% de las interacciones muestra lo que sería el comportamiento esperado de los alumnos.

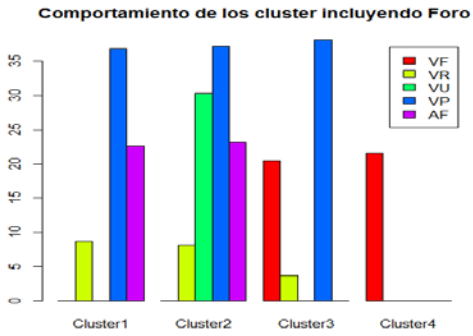

Figura 6. Comportamiento representativo de cada *cluster* con los recursos por encima de la media, incluyendo Foro

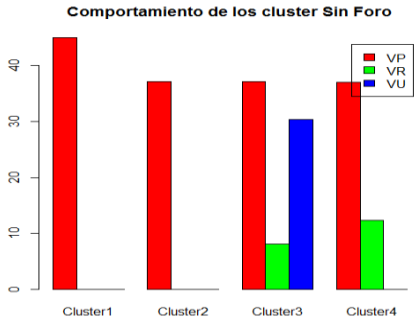

Figura 7. Comportamiento representativo de cada *cluster* con los recursos por encima de la media, sin incluir Foro

Por último se analiza aparte el comportamiento de los estudiantes respecto a las notas obtenidas en los quices o evaluaciones realizadas. En la figura 8, se observa la agrupación realizada por *cluster* y en la figura 9, se observa que la nota

predominante en el curso fue 5, entonces respecto a la media, la mayoría de los estudiantes aprobaron el curso.

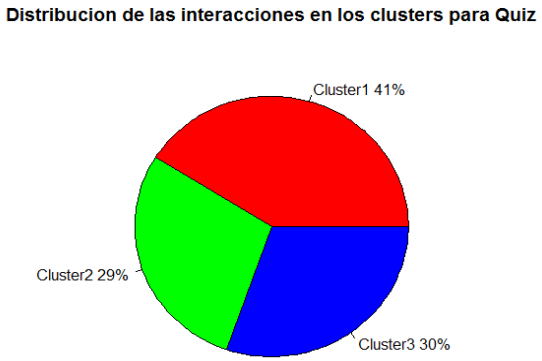

Figura 8. Agrupaciones de las interacciones para la tabla Quiz

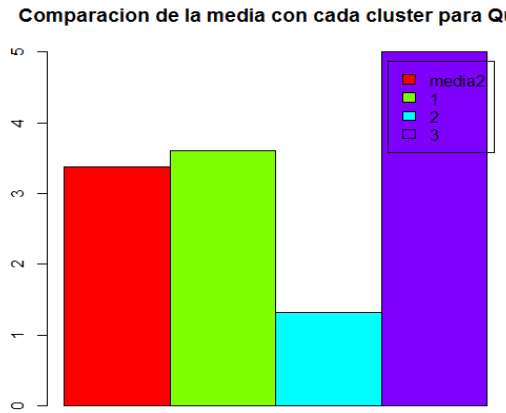

Figura 9. Comparación de la media de las interacciones con Quiz con su representación en cada *cluster*

5. Conclusiones

Con la aplicación del *clustering* sobre la plataforma Ude@ se identificó que los comportamientos de los estudiantes tienden a ser pasivos en los foros y de mucha interacción con los recursos de páginas web *HTML*. Esto podría sugerir que el curso estudiado no posee contenidos multimedia atractivos para los estudiantes o que carece de ellos y que en el proceso de enseñanza no se estimula la participación en los foros de discusión o que existen muy pocos.

Si bien, se ha mostrado que la técnica *EDM* usada permite la clasificación de los estudiantes y el análisis de comportamientos para un curso en particular, este proceso puede ser aplicado a cualquier curso de Ude@, de manera que permite el análisis de los recursos que los estudiantes están utilizando y abstraer un comportamiento general para el mismo.

6. Trabajo futuro

Para lograr una integración del proceso de minería en el *LMS* y que éste sea lo más transparente posible para el usuario que usará los resultados de la minería, se identificaron una serie de recomendaciones que se describen a continuación:

- La asignación de códigos a los materiales contenidos en un recurso, facilitaría el pre- procesamiento de los datos y se lograría tener un mayor control sobre los resultados obtenidos, pudiendo identificar en cada curso, no solo el tipo de recurso visitado, sino también el material dentro de ese recurso con el que se tuvo mayor interacción.
- Si se quiere realizar un informe, donde se presente el rendimiento de los estudiantes al interactuar con ciertos contenidos del curso, será necesario relacionar los contenidos, a criterio del profesor, con una evaluación determinada. Esto puede ser posible creando una etiqueta, con la cual se identifiquen los contenidos y la evaluación relacionada con esos recursos.
- En los experimentos realizados, se identificó que dentro del proceso de minería, la etapa que toma más tiempo en su ejecución, es el filtrado en la base de datos, tardando en promedio cuatro días en un equipo con un procesador *Intel Core* i5 a 64 bytes, con 6 *GB* de memoria *RAM*, un disco duro de 500 *GB,* con una carga de procesamiento del 95% en el procesador. Es por ello que se recomienda que esta etapa, no se ejecute al mismo tiempo que se invoque la función en *R*, sino, que se realice previamente. Esta actividad se debe realizar constantemente en un servidor que almacene las tablas, de esta manera se podrán invocar las tablas ya almacenadas y se podrán obtener los informes requeridos en tiempo real.

7. Referencias

- AlShammari, I., Aldhafiri, M., & Al-Shammari, Z. (2013).A Meta-Analysis of Educational Data Mining on Improvements in Learning Outcomes. *College Student Journal*, *47*(2), 326-333.
- Baker, R. S. 1. (2011). Data mining for education. In International encyclopedia of education. 3rd ed. Oxford, UK: Elsevier.
- Castro, F., Vellido, A., Nebot, A., & Minguillon, J. (2005). Detecting atypical student behaviour on an e-learning system. In I Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educación, Granada (pp. 153–160).
- Chapman P, Clinton J, Kerber R, Khabaza,T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.
- C. Romero, S. Ventura (2007).Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications, Volume 33, Issue 1, July 2007, Pages 135-146, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2006.04.005.
- Espigares Pinazo, M. J., & García Pérez, R. (2011). *Minería de datos educativos en plataformas virtuales de aprendizaje musical*. Universitat de València: Servei de Publicacions.
- Kaur, N., Sahiwal, J. K., & Kaur, N. (2012). Efficient k-means clustering algorithm

using ranking method in data mining. *International Journal of Advanced Research in Computer Engineering & Technology*, *1*(3), 85-91 MacQueen, J. B. (1966). *Some methods for classification and analysis of multivariate observations*. Ft. Belvoir: Defense Technical Information Center.

- Mazza, R., & Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education
- Mining, T. E. D. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief.
- Lile, A. (2011). Analyzing E-Learning Systems Using Educational Data Mining Techniques. *Mediterranean Journal Of Social Sciences MJSS*. http://doi.org/10.5901/mjss.2011.v2n3p403
- López Guarín, C. E. *Data mining model to predict academic performance at the Universidad Nacional de Colombia* (Doctoral dissertation, Universidad Nacional de Colombia).
- Olague Sánchez, Juan Ramón, Torres Ovalle, Sócrates, Morales Rodríguez, Felipe, Valdez Menchaca, Alicia Guadalupe, & Silva Ávila, Alicia Elena. (2010). Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales: un caso de estudio en el norte de Coahuila. *Revista mexicana de investigación educativa*, *15*(45), 391-421. Recuperado en 09 de junio de 2015, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662010000200004&lng=es&tlng=es. Pereira, R. T., Romero, A. C., & Toledo, J. J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Vínculos*, *10*(1), 373-383.
- Ratnapala, I. P., Ragel, R. G., & Deegalla, S. (2014, December). Students behavioural analysis in an online learning environment using data mining. In *Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on* (pp. 1-7). IEEE.
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications*, *33*(1), 135–146. http://doi.org/10.1016/j.eswa.2006.04.005
- Siti Khadijah Mohamad, Zaidatun Tasir (2013).Educational Data Mining: A Review, Procedia - Social and Behavioral Sciences, Volume 97, 6 November 2013, Pages 320-324, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2013.10.240.
- Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence (pp. 17–23).
- Zaiane, O., & Luo, J. (2001). Web usage mining for a better learning environment. In Proceedings of conference on advanced technology for education, Banff, Alberta (pp. 60–64).

Sobre los autores

- **Adrián Montoya Lince**: Ingeniero Electrónico, MSc Ingeniería, Universidad de Antioquia. Profesor asistente. adrian.montoya@udea.edu.co
- **Jesús Francisco Vargas Bonilla**, Ingeniero Electrónico, MSc, PhD, Profesor de la Facultad de Ingeniería. Universidad de Antioquia. jesus.vargas@udea.edu.co
- **Lyda Yaneth Contreras Olivares**: Administradora, Especialista en Gerencia de Proyectos, Departamento de Recursos de Apoyo e Informática, Universidad de Antioquia. lyda.contreras@udea.edu.co

---