EDUCATIONAL DATA MINING (EDM) PARA LA DETERMINACIÓN DE COMPORTAMIENTOS EN ESTUDIANTES DE INGENIERÍA EN LA MODALIDAD VIRTUAL DE UDE@

Francisco Vargas Bonilla*, Lyda Contreras Olivares†, José D. López Hincapié‡ y Adrián Montoya Lince§
Facultad de ingeniería, Universidad de Antioquia (UdeA), Medellín-Colombia.
Emails: *jesus.vargas@udea.edu.co, †lyda.contreras@udea.edu.co, ‡josedavid@udea.edu.co,
§adrian.montoya@udea.edu.co

***Resumen*—En este artículo 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 la plataforma *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 (Learning Management System)*.**

***Palabras claves*—EDM, Learning Analytics, LMS, K-Means, PCA, clustering, Moodle**

I. 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 de Colombia 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 [1] [2] [3]. 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 [4]. 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 [5] [6].

II. Trabajos relacionados

De acuerdo con [2] 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 [7] *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

ISBN: 978-84-608-3145-7

Tabla I
TÉCNICAS DE MINERÍA DE DATOS EN LA EDUACIÓN.

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 de un estudiante con base en lo observado en otros 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 de relaciones.	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.

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 [8]. Estudios previos en [9] [10] [11] [12] [13] [14] 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 presente artículo está enfocado 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 conjunto de propiedades y características similares. Para tal efecto se usó el algoritmo *K-means* [15], ampliamente utilizado por su robustez y eficacia [16]. Para su implementación se escogió *R* (http://www.r-project.org) un software libre que permite una integración con el *LMS*, lo

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 [9]	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 [10]	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 [14]	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 [11]	Utiliza *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 [12]	Aplicación de *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.
Analyzing E-Learning Systems Using Educational Data Mining Techniques	Anduela Lile, 2011 [13]	Analizan 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.

que a su vez, permitirá la automatización del proceso en el servidor, sin ninguna intervención del usuario.

III. METODOLOGÍA

En concordancia con el estándar *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*, por sus siglas en inglés) [17] la metodología usada en el desarrollo del proyecto contempló 4 pasos:

III-A. ***Paso 1. Entendimiento del modelo y datos:***

Los cursos de *Moodle* ofrecidos en la modalidad virtual de la Universidad de Antioquia a través del Programa Eduación Virtual Ude@, responden a los lineamientos metodológicos definidos en el modelo pedagógico de cada programa académico. Como esquema básico, un *Aula semilla* debe

ISBN: 978-84-608-3145-7

Tabla III
TABLAS SELECCIONADAS EN LA BASE DE DATOS DE *Moodle*.

Tablas	Recursos que almacenan
mdl_resource	Imágenes, documentos en PDF, hojas de cálculo, archivos de sonido, archivos de video.
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_forum.
mdl_quiz	Quices, tareas, autoevaluaciones, entro otros.
mdl_log	Todas las interacciones que los usuarios realicen con la plataforma.

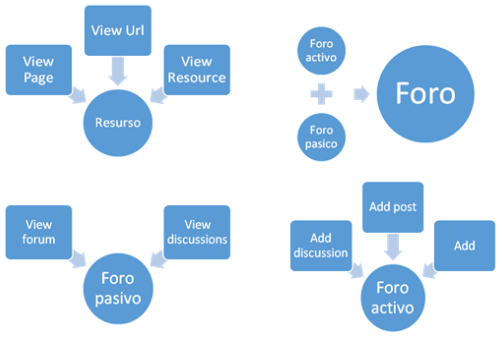

Figura 1. Clasificación de los *Recursos, Foros* y variables implicadas.

contener como mínimo: un material fundamental, una guía de estudio, unas actividades y autoevaluaciones. De esta manera, los contenidos de texto, audio o video de los cursos en *Moodle* pueden corresponderse con cualquiera de estos cuatro esquemas mencionados. De esta manera y dada la diversidad de contenidos en los cursos, nuestro interés está concentrado en medir y clasificar la cantidad de interacciones de los estudiantes con todos esos contenidos definidos en el curso.

Teniendo en cuenta esto, se tomó como caso de estudio un curso en *Moodle* y se realizó la exploración de la base de datos de éste. Teniendo conocimiento de cómo y qué información se almacena en la base de datos, se procedió a seleccionar las tablas y variables a tomar como base para la aplicación de la técnica de *clustering*. La Tabla III resume y describe las tablas seleccionadas de la base de datos.

En cuanto a las variables escogidas se tomaron dos componentes principales: los *recursos* visitados (clasificados como: *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 cinco variables que describen las posibles interacciones se denotan como: VR (Ver Recurso), VF (Ver Foro), VU (Ver Url), VP (Ver Página), AF (Abrir/Modificar Foro).

III-B. 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.

III-C. 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 cuatro clusters con un error tolerable en la suma de los cuadrados de las distancias entre los centroides y una relación de cohesión/dispersión (BSS/TSS) del 91.9 %. Finalmente se hace un análisis de las evaluaciones (Quices) registradas en el curso.

III-D. 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, considerando únicamente los recursos que visitan.

IV. RESULTADOS

Al realizar el primer análisis para todo el grupo de instancias se observó 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 y por lo tanto la variable AF no es significativa en la descripción del comportamiento.

La representación porcentual de las interacciones sobre los recursos clasificados mediante las variables *Página, Archivo, Url* arrojaron como resultado que el 58 % de las interacciones están concentradas en la variable *Página*. En este recurso se pueden almacenar un sinnúmero de actividades de diferente naturaleza (texto, audio, video). Por otro lado las variables *Archivo* y *Url* representan el 34 % y 8 % de las interaccciones medidas en el *LMS*.
Este resultado lleva a la conclusión de que el curso estudiado está construido bajo la estructura de páginas web, donde ha sido subida la mayor parte del contenido de este.

Al aplicar el algoritmo *K-means* sobre el conjunto de

ISBN: 978-84-608-3145-7

Tabla IV
CENTROIDES DE LOS *clusters* EN LOS RECURSOS.

Cluster	VP	VR	VU
1	37.04662	12.308271	0
2	37.16540	1.268492	0
3	37.16587	8.031250	30.3101
4	45.41071	0.100000	0

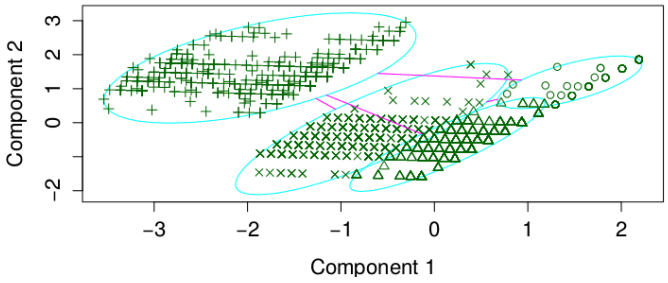

Figura 2. Representación de los *clusters* en los recursos.

Tabla V
PORCENTAJE DE INTERACCIONES EN LOS *clusters*

Cluster	***Foro***	**Sin *Foro***
1	74 %	22 %
2	21 %	14 %
3	4 %	55 %
4	1 %	10 %

interacciones, se obtienen cuatro *clusters* con un BSS/TSS del 91.9 %, los cuales agrupan satisfactoriamente los comportamientos de los estudiantes en el *LMS*. Los centroides de las interacciones en las variables, sin tener en cuenta los foros, se resumen en la Tabla IV. Aquí podemos ver que solo el *cluster* 3 posee interacciones en la variable VU.

Con el fin de visualizar el resultado de los *clusters* hallados en un gráfico 2D, se realizó un análisis de *PCA (Principal Components Analysis)* determinando las variables en función tres componentes: $VP = 0{,}514PC1 + 0{,}762PC2 - 0{,}395PC3$, $VR = -0{,}653PC1 - 0{,}756PC3$ y $VU = -0{,}556PC1 + 0{,}646PC2 + 0{,}522PC3$. Los resultados gráficos se ilustran en la Figura 2 en donde se pueden distinguir los cuatro *clusters* para los componentes $PC1$ y $PC2$ que explican el 76.25 % de la variabilidad del conjunto.

Al observar todas las interacciones del conjunto, se obtienen las agrupaciones resumidas en la Tabla V para los cuatro *clusters* determinados. Aquí se observa que el 74 % de las interacciones han sido agrupadas en el *cluster* 1, lo cual lo hace el *cluster* representativo. Por otro lado, se observa que sin tener en cuenta los foros, el *Cluster* 3 tiene más de la mitad de las interacciones.

Las Figuras 3 y 4, representan, para cada *cluster*, las variables que están por encima de la media en el número de interacciones, con lo cual se logra identificar el comportamiento representativo. Es de notar en la Figura 3,

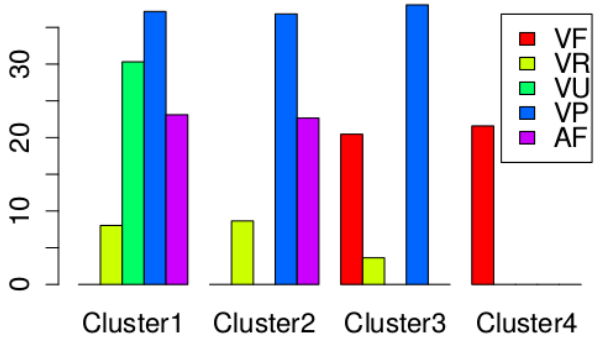

Figura 3. Cantidad de interacciones en cada *cluster* con los recursos por encima de la media, incluyendo los foro.

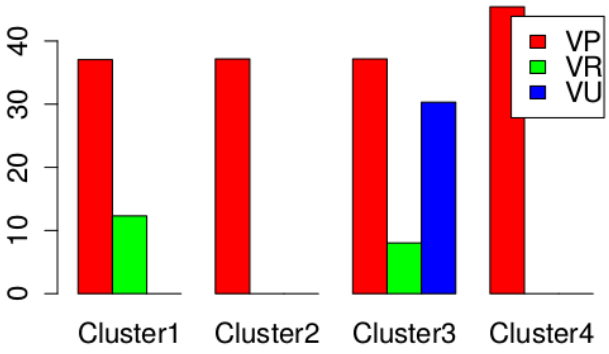

Figura 4. Cantidad de interacciones en cada *cluster* con los recursos por encima de la media, sin incluir los foros.

que el *cluster* 3, con un 4 % de las interacciones registradas, muestra lo que sería el comportamiento cercano al ideal de los estudiantes, ya que dicho grupo se caracteriza por una buena interacción con los foros, páginas y recursos URL, a pesar de tener una muy bajo porcentaje de escritura en foros. Lo mismo ocurre en la figura 4 con el *cluster* 3 que representa el comportamiento cercano al ideal de los estudiantes para los recursos del curso.

Por último, se obtuvo que la nota predominante de los estudiantes obtenidas en los quices o evaluaciones realizadas fue aprobatoria.

V. 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.

Se observó una tendencia de un deseado comportamiento en el uso de los recursos del curso estudiado (55 % de interacciones en el *cluster* 3), en el cual los estudiantes

ISBN: 978-84-608-3145-7

interactúan con las *páginas*, *recursos* y *urls* de los cursos en *Moodle*.

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 de la plataforma *Moodle* que los estudiantes están utilizando y abstraer un comportamiento general para el mismo.

VI. 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 bits, 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.

Referencias

[1] Iqbal AlShammari, Mohammed Aldhafiri, and Zaid Al-Shammari. A meta-analysis of educational data mining on improvements in learning outcomes. *College Student Journal*, 47(2):326–333, 2013.

[2] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135 – 146, 2007.

[3] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, November 2013. The 9th International Conference on Cognitive Science.

[4] Riccardo Mazza and Christian Milani. Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Communication dans: the Workshop on Usage analysis in learning systems, the twelfth International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands*, pages 65–72, July 2005.

[5] Luis Talavera and Elena Gaudioso. 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*, pages 17–23, 2004.

[6] Osmar R Zaïane. Web usage mining for a better web-based learning environment. In *Proceedings of conference on advanced technology for education*, pages 60–64, June 2001.

[7] RSJD Baker et al. Data mining for education. *International encyclopedia of education*, (7):112–118, 2010.

[8] Through Educational Data Mining. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In *Proceedings of conference on advanced technology for education*, 2012.

[9] IP Ratnapala, RG Ragel, and S Deegalla. Students behavioural analysis in an online learning environment using data mining. In *Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on*, pages 1–7. IEEE, December 2014.

[10] Ricardo Timarán Pereira, Andrés Calderón Romero, and Javier Jiménez Toledo. Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Vínculos*, 10(1):373–383, 2013.

[11] Manuel Jesús Espigares Pinazo and Rafael García Pérez. Minería de datos educativos en plataformas virtuales de aprendizaje musical. *Revista electrónica de LEEME*, (27):1–16, 2011.

[12] Juan Ramón Olague Sánchez, Sócrates Torres Ovalle, Felipe Morales Rodríguez, Alicia Guadalupe Valdez Menchaca, and Alicia Elena Silva Ávila. 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, 2010.

[13] Anduela Lile. Analyzing e-learning systems using educational data mining techniques. *MJSS*, 1:2, 2012.

[14] Camilo Ernesto López Guarín. Data mining model to predict academic performance at the universidad nacional de colombia. Master in Systems and Computer Engineering. Research Line: Intelligent Systems, 2013.

[15] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, June 1967.

[16] Navjot Kaur, Jaspreet Kaur Sahiwal, and Navneet Kaur. Efficient k-means clustering algorithm using ranking method in data mining. *International Journal of Advanced Research in Computer Engineering and Technology*, 1(3), May 2012.

[17] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[18] Félix Castro, Alfredo Vellido, Angela Nebot, and Julià Minguillon. Detecting atypical student behaviour on an e-learning system. pages 14–16, 2005.

ISBN: 978-84-608-3145-7

ISBN: 978-84-608-3145-7