

Metodología de Aprendizaje Automático para la Clasificación y Predicción de Usuarios en Ambientes Virtuales de Educación

Enrique J De La Hoz^{1,2}, Efraín J De La Hoz^{3*} y Tomás J Fontalvo³

¹ Universidad Tecnológica de Bolívar, Facultad de Ingeniería, Programa de Ingeniería Industrial, Campus Ternera, Cartagena. (e-mail: edelahoz@utb.edu.co).

² Universitat Oberta de Catalunya, Rambla del Poblenou, 156, 08018, Barcelona, España

³ Universidad de Cartagena, Facultad de Ciencias Económicas, Programa Administración Industrial. Campus Piedra de Bolívar, Cartagena, Colombia, (e-mail: edelahozg@unicartagena.edu.co; tfontalvo@unicartagena.edu.co).

*Autor a quien debe ser dirigida la correspondencia

Recibido Jun. 25, 2018; Aceptado Ago. 24, 2018; Versión final Oct. 2, 2018, Publicado Feb. 2019

Resumen

Se desarrolla una metodología para clasificar y predecir usuarios en ambientes virtuales de educación, estudiando la interacción de los estudiantes con la plataforma y su desempeño en los exámenes. Para esto se utilizaron las herramientas de aprendizaje automático, componentes principales, clusterización, lógica difusa, y el algoritmo del K vecino más cercano. La metodología relaciona los usuarios según las variables de estudio, para así implementar un análisis de clúster que identifica la formación de grupos. Finalmente utiliza un algoritmo de aprendizaje automático para clasificar los usuarios según su nivel de conocimiento. Los resultados muestran como el tiempo que un estudiante permanece en la plataforma no está relacionado con pertenecer al grupo de conocimiento alto. Se identificaron tres categorías de usuarios, aplicando la metodología Fuzzy K-means para determinar zonas de transición entre niveles de conocimiento. El algoritmo K vecino más cercano presenta los mejores resultados de predicción con un 91%.

Palabras claves: clúster; AVE; aprendizaje automático; KNN; educación

Methodology of Machine Learning for the classification and Prediction of users in Virtual Education Environments

Abstract

A methodology to classify and predict users in virtual education environments, studying the interaction of students with the platform and their performance in exams is proposed. For this, the machine learning tools, main components, clustering, fuzzy and the algorithm of the K nearest neighbor were used. The methodology first relates the users according to the study variables, to then implement a cluster analysis that identifies the formation of groups. Finally uses a machine learning algorithm to classify the users according to their level of knowledge. The results show how the time a student stays in the platform is not related to belonging to the high knowledge group. Three categories of users were identified, applying the Fuzzy K-means methodology to determine transition zones between levels of knowledge. The k nearest neighbor algorithm presents the best prediction results with 91%.

Keywords: cluster; VLE; machine learning; KNN; education

INTRODUCCIÓN

La modelación del comportamiento de los usuarios es uno de las técnicas más usadas por las aplicaciones basadas en la web. La principal motivación para desarrollar el modelado del comportamiento de usuarios es desarrollar contenidos personalizados de acuerdo al nivel de interacción usuario - plataforma. En el ámbito educativo, el objetivo de los Ambientes Virtuales de Educación (AVE) es facilitar las actividades de aprendizaje para el usuario, el acceso a información de manera ágil y la gestión de los recursos de aprendizaje (Bayne, 2008). El mejoramiento de los niveles de personalización de los contenidos asociados a un sistema de información genera una percepción positiva para el usuario (Knijnenburg et al., 2012), haciendo más eficiente el desempeño de la plataforma para gestionar la relación profesor alumno.

En la actualidad el desarrollo de las tecnologías de la información permite almacenar y gestionar grandes cantidades de datos, esto aplicado en los entornos virtuales de educación genera la posibilidad de personalizar los contenidos presentados a los usuarios (Klašnja-Milićević, Ivanović, & Nanopoulos, 2015; Melo-Solarte, Díaz, Melo-Solarte, & Díaz, 2018). El enfoque utilizado en el presente proyecto es la creación de perfiles de usuarios según su nivel de conocimiento, lo cual en el marco de procesos de enseñanza universitaria permite colocar especial atención a usuarios con bajo nivel académico, generando para ellos contenidos personalizados y así plantear estrategias que permitan mejorar su desempeño académico.

Según (Fontalvo et al., 2018), los procesos de enseñanza y aprendizaje de la educación superior deben generar cambios significativos que impacten positivamente los resultados de sus clientes. Sin embargo, la tecnología como propulsor de los procesos de enseñanza en la educación superior genera diferencias conceptuales en los diferentes actores del proceso educativo (Christensen, 2002), por una parte existen profesores que prefieren la interacción personal como base de la relación de aprendizaje, De otra forma los estudiantes no cuentan con una conciencia del nivel de ayuda que un entorno virtual de aprendizaje puede aportar a su formación como profesional, utilizando esta solo para actividades transaccionales como la descarga y carga de archivos. Es así, como en esta investigación se da respuestas a las siguientes preguntas problema. ¿De qué forma clasificar los estudiantes participantes en un ambiente virtual de educación utilizando el aprendizaje automático?, ¿Qué factores asociados a un ambiente virtual de educación determinan los niveles de conocimiento de los estudiantes?, ¿Cómo interpretar los niveles de transición entre los niveles de conocimiento de los estudiantes?, ¿Cómo generar una representación visual de los resultados del proceso de agrupación de estudiantes?, ¿Cómo predecir la pertenencia de un estudiante independiente a un nivel de conocimiento establecido previamente?. Basados en lo anterior, el objetivo principal de la presente investigación es desarrollar una metodología (Ver fig. 1) apoyada en el clustering, en lógica fuzzy y el vecino más cercano KNN para clasificar estudiantes participantes en ambientes virtuales de aprendizaje identificando las relaciones entre la intensidad de uso de la plataforma y los resultados en los exámenes semestrales. Así mismo la metodología permite predecir la pertenencia de un estudiante independiente a un conglomerado, cuando se cuenta con la información y variables del estudiante utilizadas en esta investigación.

Los modelos de clasificación para usuarios de sistemas web se apoyan en la definición de grupos a priori para que a través de un modelo de aprendizaje automático y con entradas de información conocidas se pueda clasificar la unidad de estudio en un grupo determinado. En esta técnica se definen variables dependientes e independientes que permiten disminuir la varianza dentro de los grupos y maximizar la varianza entre grupos, con lo cual se logran grupos lo más homogéneos intra-grupos y heterogéneos extra-grupos.

Diferentes investigaciones desarrollaron importantes contribuciones a la literatura utilizando la minería de datos en la clasificación de individuos en categorías. Por ejemplo, (Carpaneto, Chicco, Napoli, & Scutariu, 2006) identificaron y pronosticaron usuarios del servicio eléctrico según sus niveles de consumo. Ye, Zhang, & Law (2009) clasificaron los usuarios de servicios turísticos utilizando los comentarios de estos sobre los lugares visitados. A su vez, en otros trabajos posteriores, Clavero, Salicrú, & Turbón (2015) pronosticaban el sexo de una persona fallecida utilizando la medida del fémur. De La Hoz Domínguez, Mendoza Mendoza, & Ojeda De La Hoz (2017) plantean un modelo para identificar usuarios de un periódico digital usando minería de datos, (Arista-Jalife, Calderón-Auza, Fierro-Radilla, & Nakano, 2017) clasifican imágenes satelitales de ciudades usando redes neuronales. En temas educativos Fontalvo-Herrera et al (2018) proponen una clasificación de programas de ingeniería industrial usando aprendizaje automático, Siirtola et al. (2013) estudian las relaciones entre asignaturas de un plan de estudio a través de nodos interrelacionados. Morsi et al. (2007) identifica las relaciones entre asignaturas y competencias a través de mapas conceptuales.

Pardos et al. (2012), resaltan que no existen estudios suficientes sobre la estimación del conocimiento inicial de los usuarios así que sugiere utilizar una red bayesiana en orden de aprender los parámetros previos al modelo de usuario y medir su desempeño. En la literatura, aunque existen muchos enfoques desarrollados para la modelar el conocimiento del usuario web, se evidencia la ausencia aplicaciones en el contexto educativo. Las razones básicas para esto son la dependencia del modelo desarrollado al campo de aplicación,

la complejidad y el enfoque en los modelos teóricos y también solo unos pocos modelos han sido probados en conjuntos de datos educativos (Liao et al., 2014).

Análisis de Clúster: El análisis de clúster es una técnica de aprendizaje no supervisado en el área del aprendizaje automático, capaz de crear grupos de tal forma que las observaciones pertenecientes a un grupo están muy cercanas entre ellas y apartadas de las observaciones ubicadas en otro clúster. Existen 4 categorías de algoritmos de clustering: 1) de particionamiento; 2) basados en densidad; 3) basados en redes; y 4) jerárquicos. Los algoritmos de partición tal como k-means y de partición alrededor de las medianas iterativamente refinan un conjunto de k clústeres y por lo general no escalan bien para grandes conjuntos de datos. Los algoritmos basados en densidad, son capaces de generar clústeres de tamaño arbitrario y de lidiar con valores extremos. Los algoritmos basados en red reducen el espacio de agrupamiento a celdas dentro de una red, permitiendo un agrupamiento eficiente de conjuntos de datos de gran tamaño. Los algoritmos jerárquicos pueden ser de aglomeración o divisivos, en la forma aglomerada repetidamente se fusionan dos clústeres, mientras en la forma divisiva repetidamente se divide un clúster en dos.

Cálculo de Distancia: En el proceso del análisis de clúster es importante establecer una distancia o disimilaridad de los elementos de estudio, en este caso los estudiantes usuarios de la plataforma. La interacción de los alumnos con la plataforma se desarrolla por un factor de motivación que es responsabilidad del profesor al programar las actividades y recursos de la plataforma, así pues, el agrupamiento deseado debe seguir el siguiente enfoque: Cuanto más tiempo participen dos estudiantes de la plataforma y obtengan notas similares, más cerca estarán de pertenecer al mismo grupo.

$$\text{Distancia Bray - Curtis} = \frac{\sum_{t=1}^T |x_{it} - x_{jt}|}{\sum_{t=1}^T (x_{it} + x_{jt})} \quad (1)$$

Teniendo en cuenta que el objetivo del estudio es identificar perfiles de usuarios en cuanto a su conocimiento e interacción con la plataforma de aprendizaje, se necesita una distancia que tenga en cuenta el ángulo y la magnitud. La distancia elegida para el desarrollo del estudio es la Bray - Curtis, ya que nos permitirá identificar perfiles en los cuales factores como la intensidad de uso sean tenidos en cuenta.

Algoritmo KNN: El KNN es uno de los clasificadores basados en vecindad más populares en el aprendizaje automático (Kataria y Singh, 2013), dado su simplicidad y eficiencia para detectar y clasificar elementos en categorías. El parámetro k en KNN hace referencia al número de vecinos con el cual se define la pertenencia a una categoría, este parámetro usualmente se determina empíricamente, dependiendo del problema se prueba con diferentes valores de K, eligiendo el parámetro con el mejor desempeño en precisión. El funcionamiento del algoritmo se basa en calcular una matriz de distancia entre todos los puntos del dataset de entrenamiento, usando la ecuación (1). Al ingresar un elemento de prueba se le asigna una categoría utilizando la ecuación (2).

$$\text{Pertenencia de categoría} = L_c = \sum_{i=1}^{L_c} \frac{1}{i} \quad (2)$$

Donde L_c , es el número de puntos de la categoría c , i es el orden del punto el conjunto de datos de entrenamiento después de ordenar las distancias.

METODOLOGÍA

Para el desarrollo de esta investigación inicialmente se realizó un análisis de clustering, el cual se desarrolló en dos fases la primera un clúster jerárquico para generar una idea inicial de la formación de grupos. La segunda fase desarrolló un algoritmo de clúster no jerárquico basado en medianas. Se introduce en este paso la metodología de lógica Fuzzy para dar interpretación al proceso de transición de los estudiantes entre niveles de conocimiento.

En tercer lugar, se aplicó la técnica de aprendizaje automático K-vecinos cercanos (KNN) para clasificar a los estudiantes de la plataforma, adicionalmente para verificar la consistencia y pertenencia de los estudiantes a los perfiles o clúster definidos previamente se utilizó los modelos discriminante lineal y cuadrático, para comparar los resultados de clasificación, validando así el contexto del estudio y promoviendo la generación de futuras investigaciones en el tema. Por último, se generó el proceso de visualización de los resultados implementando la técnica del escalado multidimensional, generando un esquema para el proceso de agrupamiento general y otro para el agrupamiento basado en lógica Fuzzy. Como se muestra en la Fig. 1.

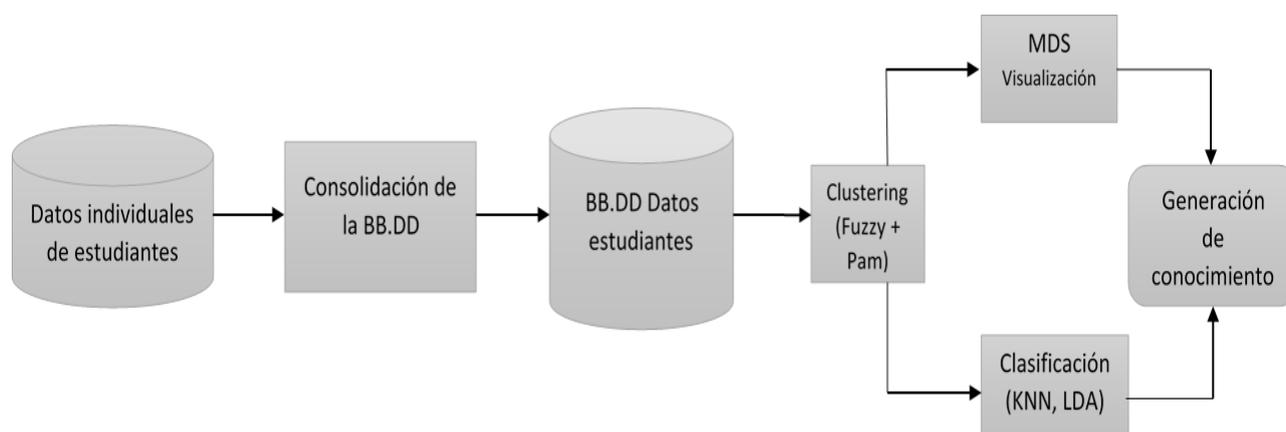


Fig. 1: Metodología de aprendizaje automático para clasificación de estudiantes en entornos virtuales de educación

Datos del estudio: Para la realización de esta investigación se tomó información de una base de datos con un total de 403 usuarios del Sistema de Aprendizaje Virtual Interactivo (SAVIO) de la Universidad Tecnológica de Bolívar, plataforma soportada por Moodle que sirve como soporte al proceso de enseñanza universitaria. Los usuarios son estudiantes del programa de Ingeniería Industrial y las estadísticas asociadas a la presente investigación hacen referencia a la materia Investigación de Operaciones.

Las variables de estudio representan actividades desarrolladas por el estudiante en la plataforma de aprendizaje. Estas son frecuencia de visitas, tiempo activo en la web en temas enmarcados como claves de estudio y desempeño en exámenes semestrales. Los datos se extraen de la interacción plataforma – estudiante, a través de un sistema UMS (User Modelling System) (Kobsa, 2001) el cual rastrea actividades de aprendizaje, rutas de navegación, clics y tiempos usados para cada actividad. El proceso de conformación del conjunto de datos ha sido una fase clave para el desarrollo de la presente investigación, dado que la estructura de almacenamiento de la información de la plataforma Moodle no permite realizar procesos de barrido de la información generales o Web Scrapping, por lo tanto, se recogió la información individual por cada estudiante, depurando observaciones y valores atípicos para finalmente consolidar un conjunto de datos apto para modelar. El análisis se desarrolló en R (Team, 2013) y el paquete dendextend (Galili, 2015).

Variables del modelo:

TRO: Tiempo de estudio de los recursos enmarcados como obligatorios. (Minutos)

NIP: Número de veces que ingresó a la plataforma e hizo uso de los materiales obligatorios. (Frecuencia)

TRC: Tiempo de estudio del usuario para los recursos relacionados como complementarios. (Minutos)

REP: El desempeño del estudiante en el examen parcial de la asignatura Investigación de Operaciones (Escala de 0 -5).

REF: El desempeño del usuario en el examen final de la asignatura Investigación de Operaciones. (Escala de 0 -5).

NCS: Nivel de Conocimiento del Usuario (Variable de salida). (Variable categórica)

Para esta investigación el origen de la ciencia se generó de un análisis empírico, es decir la información generada por la plataforma virtual y el análisis racional de los investigadores al articular los diferentes modelos y herramientas del aprendizaje automático para a través de un análisis estadístico probabilístico determinar la pertenencia de los estudiantes a los grupos establecidos a través del análisis de conglomerado. De igual forma la esencia de la ciencia tiene un enfoque combinado toda vez que se aborda desde la relación del sujeto o investigadores quienes racionalizan las diferentes teorías y plantean una construcción de objeto y el análisis empírico que se genera de la información de la plataforma y su análisis a través de las diferentes herramientas de aprendizaje automático utilizadas en esta investigación. Estas herramientas son pertinentes en primer lugar por la cantidad de datos estudiados, en segundo por la capacidad de generar un modelo reproducible y replicable y tercero por la necesidad de garantizar la privacidad y sensibilidad de los datos utilizados.

RESULTADOS

En primera instancia se buscó a través del análisis de clúster, identificar grupos de niveles de usuario, interpretar sus características comunes y representar los elementos patrón de cada grupo.

Clúster Jerárquico: Para generar una idea del número de grupos presentes en los datos se aplicó un análisis de clúster jerárquico, procediendo a realizar una inspección visual del dendograma y luego aplicar un análisis de clúster no jerárquico basado en partición, tal como se explica en Punj y Stewart, (1983). Al analizar el

dendograma para el clúster jerárquico se aprecia la existencia de 2 ó 3 grupos claramente marcados (Ver Fig. 2). Como se explica en (De La Hoz & López Polo, 2017), en los procesos de clustering el conocimiento del fenómeno en estudio es importante a la hora de decidir el número de grupos a elegir, por lo cual basados en la experiencia docente de los investigadores se decidió trabajar con 3 grupos para el resto de la investigación, esto para generar un nivel de información más específico sobre la clasificación de los estudiantes analizados.

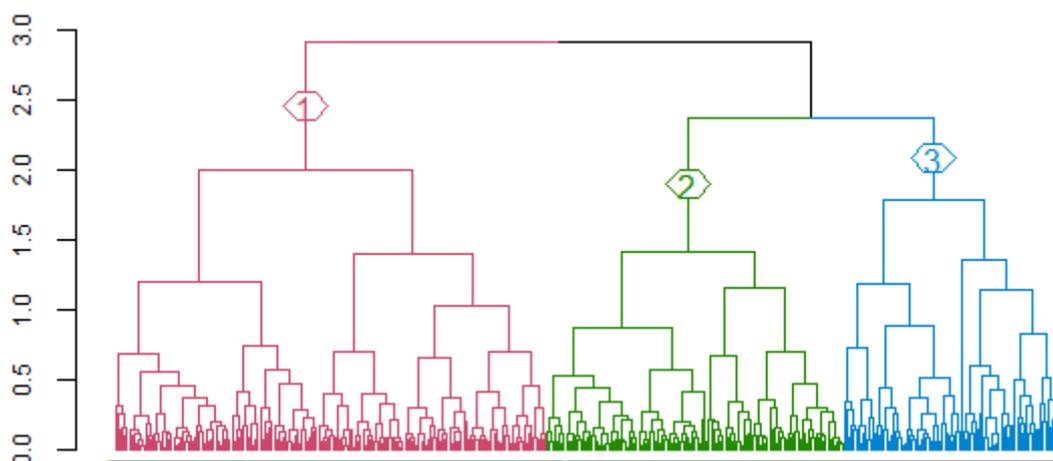


Fig. 2: Dendograma para 3 grupos con el método Ward.

El estadístico Silhouette es utilizado para verificar el número de grupos escogido, en la Fig. 3 se puede observar como los máximos valores del estadístico se obtienen en los valores $k=2$ y $k=3$.

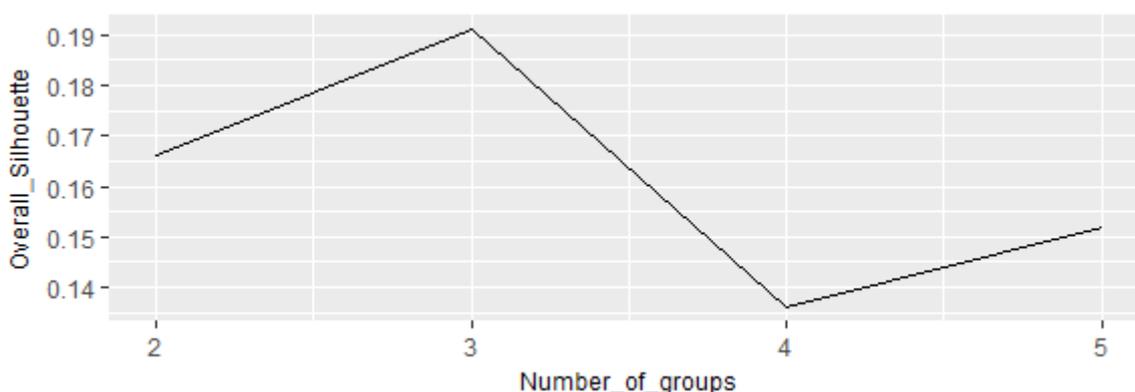


Fig. 3: Silhouette test para el clúster jerárquico

Resultados del Clúster No Jerárquico: En el análisis, se utilizó el método PAM (Partición Alrededor de Medoides), donde un medoide se define como el elemento que minimiza la suma de la disimilaridad media entre los puntos de un grupo. Utilizando tres grupos, se obtuvieron los elementos característicos de cada grupo con lo cual se realiza la representación en dos dimensiones. La identificación de los elementos representativos por cada grupo a partir de los medoides (Ver Tabla 1), sumado a la interpretación de la representación gráfica en 2D, permitió desarrollar los siguientes perfiles de Usuario. 1) Perfil Grupo 1: Usuarios con resultados medios en los exámenes que dedican poco tiempo de estudios a las áreas relacionadas con los objetivos globales. Este perfil se rotulará según el conocimiento obtenido como “Nivel Medio”; 2); Perfil Grupo 2: Usuarios con un alto resultado en el examen que dedican un tiempo medio a las áreas de estudio relacionadas con los objetivos globales. Este perfil se rotulará como “Nivel Alto”; 3); Perfil Grupo 3. Usuarios caracterizados por notas bajas en los exámenes, los cuales dedican mucho tiempo a estudiar los contenidos definidos como obligatorios en el AVE. Este perfil se rotulará como “Nivel bajo”.

Tabla 1: Elementos representativos por cada perfil

	TRO	NIP	TRC	REP	REF
Perfil 1	0.32	0.32	0.29	0.48	0.26
Perfil 2	0.29	0.3	0.56	0.78	0.56
Perfil 3	0.475	0.33	0.79	0.31	0.25

La variabilidad explicada por el eje horizontal, está definida mayoritariamente por las variables desempeño del estudiante en el examen parcial (REP) y desempeño del estudiante en el examen final (REF), la correlación proyectada de estas sobre este eje es 0.91 y 0.92 respectivamente, por esto se rotulará al eje horizontal como calificaciones. De manera similar, las variables asociadas al tiempo de uso de la plataforma (TRO y TRC) y a la frecuencia de visita (NIP) generan una alta correlación con el eje vertical, con valores de 0.84 ,0.82 y 0.74 respectivamente, por esto el eje vertical se rotulará como Intensidad de estudio. Estas denominaciones permiten una mejor interpretación de las gráficas en función de la toma de decisiones.

Analisis Fuzzy K-means: Con el análisis Fuzzy K-means se pudo identificar cómo los perfiles nivel de usuario presentan una etapa de transición, en el cual los usuarios van adquiriendo un nivel más alto de conocimiento requiriendo menos tiempo de estudio y a su vez usuarios que incrementan su tiempo de estudio presentan mejores resultados (ver Fig. 4). La Tabla 2 muestra baja probabilidad de transición de un clúster a otro, con lo cual se observa la tendencia de los estudiantes analizados a mantenerse dentro del grupo de clasificación inicial. En todo caso, se infiere en los resultados que existe una mayor probabilidad de transición del perfil 1 al 3, es decir estudiantes con notas medias transitando al perfil de notas bajas.

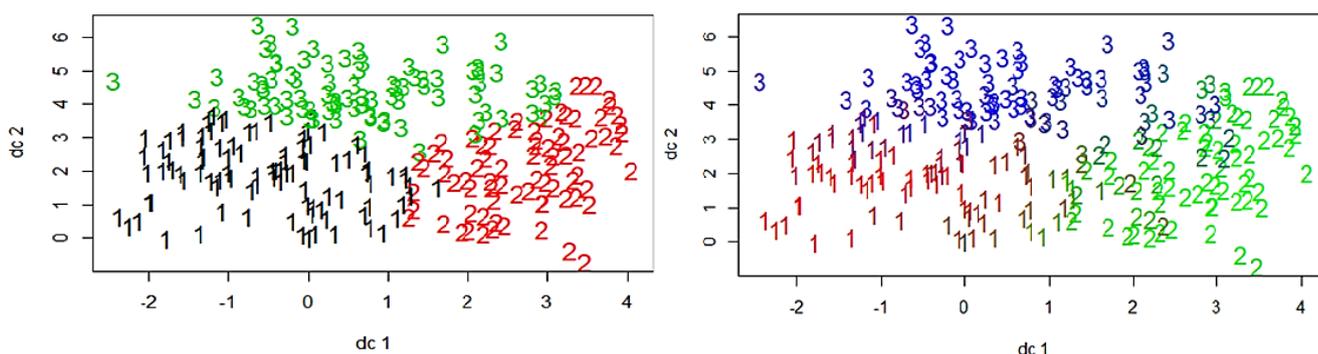


Fig. 4: Representación gráfica del análisis de clúster No Jerárquico vs Fuzzy Clúster

Lo cual es importante para la toma de decisiones por parte de las autoridades académicas, considerando que esta investigación aporta criterios que permiten la toma de decisiones administrativas y académicas para que lo estudiantes de bajo o medio rendimiento mejoren su desempeño académico.

Tabla 2: Probabilidades de transición Fuzzy entre grupos

Clúster	Vecino	Probabilidad de transición (Fuzzy)
(1) Medio	(3) Bajo	0.20
(3) Bajo	(2) Alto	0.18
(3) Bajo	(1) Medio	0.11
(2) Alto	(1) Medio	0.09
(1) Medio	(2) Alto	0.06

Modelo de clasificación: El proceso educativo busca mejorar las condiciones de aprendizaje para los futuros estudiantes, por lo tanto, la idea es que una vez establecidas las características de los estudiantes según su nivel de interacción con la plataforma se pueda identificar a que grupo pertenecería. Del total de 403 observaciones se tomaron 258 como datos de entrenamiento y 145 para evaluación. Los resultados del proceso de clasificación (Ver Tabla 3) muestran como el clasificador basado en el algoritmo KNN presenta los mejores resultados de predicción con un 91% de precisión, mejorando a los clasificadores discriminante lineal y cuadrático, estos con una precisión de 85% y 84% respectivamente.

Tabla 3: Matriz de confusión de los valores reales versus valores predichos

	KNN (91%)			Lineal (85%)			Cuadrático (84%)		
	(1) Medio	(2) Alto	(3) Bajo	(1) Medio	(2) Alto	(3) Bajo	(1) Medio	(2) Alto	(3) Bajo
(1) Medio	46	2	0	45	5	1	40	2	0
(2) Alto	1	29	6	3	24	9	1	22	4
(3) Bajo	3	0	58	2	2	54	9	7	60

DISCUSIÓN FINAL

Es importante resaltar el resultado obtenido en el análisis de clúster, evidenciando que los alumnos que pasan un mayor tiempo en la plataforma no obtienen las mejores notas, refrendando los resultados en Britton y Tesser, (1991); Nonis y Hudson, (2006); Plant, Ericsson, Hill, y Asberg (2005); Rau y Durand (2000) desarrollados en ambientes tradicionales de educación. El análisis se puede apreciar en la tabla 4. En las investigaciones mencionadas se utilizan métodos estadísticos basados en datos recolectados por medios de encuestas, al contrario de esta investigación donde la metodología utiliza información objetiva del uso del AVE por parte del alumno, además de articular una fase de aprendizaje no supervisado en el análisis de clúster con un algoritmo supervisado como el KNN.

Tabla 4: Comparación de resultados de otras investigaciones

<i>Autor (año)</i>	<i>Objeto</i>	<i>Metodología</i>	<i>Resultados</i>
(Rau & Durand, 2000)	Estudiantes de la Universidad de Michigan	Análisis de correlación y regresión lineal	No se evidencia relación entre las horas de estudio y las notas. La disciplina y la metodología si están relacionadas con las notas.
(Plant et al., 2005)	88 voluntarios de Florida State University	Encuesta y regresión múltiple moderada	La cantidad de tiempo estudiado es un mal predictor del desempeño académico. Recomienda involucrar la calidad del material de estudio.
(Nonis & Hudson, 2006)	276,449 estudiantes de 413 universidades en EE. UU.	Encuesta y análisis de correlación	El tiempo de estudio no tiene influencia directa en el promedio acumulado. Es significativo solo para estudiantes que trabajan.
(Salamonson & Andrew, 2006)	267 estudiantes de enfermería en Australia	Encuesta y análisis de regresión	Trabajar en temas relacionados con la carrera arroja peores resultados académicos. Los estudiantes que no reciben un pago por su trabajo tienen un mejor rendimiento

Se infiere que los alumnos con mayor grado de conocimiento de la materia hacen uso de la plataforma de manera moderada, mientras aquellos con malas notas en el primer examen hacen un uso exhaustivo de está intentando nivelar las notas del segundo examen, creando así un efecto de sobredimensionamiento de esta variable en los resultados. Los resultados del proceso de clasificación y pronostico muestran buenos resultados en términos de precisión al utilizar el algoritmo de KNN, representando beneficios en la facilidad de la interpretación y comparabilidad de los resultados, para este caso el KNN con un 91% de precisión supera en 7.1% el desempeño del clasificador lineal y cuadrático. Contextualizando este resultado en el ámbito universitario, para un profesor es importante conocer las diferentes formas como un alumno se relaciona con el proceso educativo y mucho mejor si puede tomar acciones para facilitar el aprendizaje de los estudiantes, re-conceptualizando la función del profesor en cuanto a la gestión de los entornos virtuales de aprendizaje.

CONCLUSIONES

Como valor agregado de esta investigación se aporta a la comunidad científica y académica una metodología para clasificar estudiantes participantes en ambientes virtuales de aprendizaje, identificando las relaciones entre la intensidad de uso de la plataforma y los resultados en los exámenes semestrales. Así mismo se establecen criterios para el estudio y análisis de los procesos de transición de los perfiles de los estudiantes de acuerdo al rendimiento alcanzado previamente. De igual forma con la metodología propuesta se puede pronosticar la pertenencia a un conglomerado de un estudiante independiente cuando se tienen las variables inherentes utilizadas en esta investigación.

REFERENCIAS

- Arista-Jalife, A., G. Calderón-Auza, A. Fierro-Radilla y M. Nakano, Clasificación de Imágenes Urbanas Aéreas: Comparación entre Descriptores de Bajo Nivel y Aprendizaje Profundo, doi: 10.4067/S0718-07642017000300021, Información Tecnológica, 28(3), 209-224 (2017)
- Bayne, S., Higher education as a visual practice: seeing through the virtual learning environment, doi: 10.1080/13562510802169665, Teaching in Higher Education, 13(4), 395-410 (2008)

- Britton, B.K. y A. Tesser, Effects of time-management practices on college grades, *Journal of educational psychology*, 83(3), 405 (1991)
- Carpaneto, E., G. Chicco, R. Napoli y M. Scutariu, Electricity customer classification using frequency–domain load pattern data, doi: 10.1016/j.ijepes.2005.08.017, *International Journal of Electrical Power & Energy Systems*, 28(1), 13-20 (2006)
- Christensen, R., Effects of technology integration education on the attitudes of teachers and students, doi: 10.1080/15391523.2002.10782359, *Journal of Research on technology in Education*, 34(4), 411–433 (2002)
- Clavero, A., M. Salicrú y D. Turbón, Sex prediction from the femur and hip bone using a sample of CT images from a Spanish population, doi: 10.1007/s00414-014-1069-y, *International Journal of Legal Medicine*, 129(2), 373-383 (2015)
- De La Hoz, E., A. Mendoza y H. Ojeda, Clasificación de perfiles de lectores de un periódico digital, *Revista U.D.C.A., Actualidad y Divulgación Científica*, 20(2), 469-478 (2017)
- De La Hoz, E. y L. Polo, Aplicación de Técnicas de Análisis de Conglomerados y Redes Neuronales Artificiales en la Evaluación del Potencial Exportador de una Empresa, doi:10.4067/S0718-07642017000400009, *Información Tecnológica*, 28(4), 67-74 (2017)
- Fontalvo, T.J., E. De La Hoz, Diseño e Implementación de un Sistema de Gestión de la Calidad ISO 9001:2015 en una Universidad Colombiana, doi: 10.4067/S0718-5006201800010003, *Formación Universitaria*, 11(1), 35-44 (2018)
- Fontalvo, T., E. De La Hoz-Domínguez y A. Mendoza- Mendoza, Aplicación de Minería de Datos para la Clasificación de programas universitarios de Ingeniería Industrial Acreditados en alta calidad en Colombia, doi:10.4067/S0718-07642018000300089, *Información Tecnológica*, 29(3) (2018)
- Gallili, T., Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering, doi.org/10.1093/bioinformatics/btv428, *Bioinformatics*, 31(22), 3718-3720 (2015)
- Kataria, A. y M.D. Singh, A review of data classification using k-nearest neighbour algorithm, *International Journal of Emerging Technology and Advanced Engineering*, 3(6), 354–360 (2013)
- Klašnja-Milićević, A., M. Ivanović y A. Nanopoulos, Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions, doi.org/10.1007/s104, *Artificial Intelligence Review*, 44(4), 571–604 (2015)
- Knijnenburg, B.P., M.C. Willemsen, Z. Gantner, H. Soncu y C. Newell, Explaining the user experience of recommender systems, doi: 10.1007/s11257-011-9118-4, *User Modeling and User-Adapted Interaction*, 22(4-5), 441-504 (2012)
- Kobsa, A., Generic user modeling systems. *User modeling and user-adapted interaction*, doi: 10.1023/A:1011187500863, 11(1-2), 49–63 (2001)
- Liao, S.-H., P.Y. Hsian y G.L. Wu, Mining user knowledge for investigating the facebook business model: the case of Taiwan users, doi: 10.1080/08839514.2014.927695, *Journal Applied Artificial Intelligence*, 28(7), 712–736 (2014)
- Melo-Solarte, D.S. y P. Díaz, El Aprendizaje Afectivo y la Gamificación en Escenarios de Educación Virtual, doi: 10.4067/S0718-07642018000300237, *Información Tecnológica*, 29(3), 237-248 (2018)
- Morsi, R., W. Ibrahim y F. Williams, Concept maps: Development and validation of engineering curricula, doi: 10.1109/FIE.2007.4418180, *Frontiers in Education Conference-Global Engineering*. (p. T3H–18). IEEE. (2007)
- Nonis, S.A. y G.I. Hudson, Academic performance of college students: Influence of time spent studying and working, doi: 10.3200/JOEB.81.3.151-159, *Journal of Education for Business*, 81(3), 151–159 (2006)
- Pardos, Z. A., S.M. Gowda, R.S. Baker y N.T. Heffernan, The Sum is Greater Than the Parts: Ensembling Models of Student Knowledge in Educational Software, doi: 10.1145/2207243.2207249, *SIGKDD Explor. Newsl.*, 13(2), 37–44 (2012)
- Plant, E.A., K.A. Ericsson, L. Hill y K. Asberg, Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance, doi:10.1016/j.cedpsych.2004.06.001, *Contemporary Educational Psychology*, 30(1), 96-116 (2005)
- Punj, G. y D.W. Stewart, Cluster analysis in marketing research: Review and suggestions for application, doi: 10.2307/3151680, *Journal of marketing research*, 20(2), 134–148 (1983)
- Rau, W. y A. Durand, The academic ethic and college grades: Does hard work help students to "make the grade"? doi: 10.2307/2673197, *Sociology of Education*, 73(1), 19–38 (2000)
- Salamonson, Y. y S. Andrew, Academic performance in nursing students: Influence of part-time employment, age and ethnicity, *Journal of Advanced Nursing*, 55(3), 342–349 (2006)
- Siirtola, H., K.J. Raiha y V. Surakka, Interactive curriculum visualization, doi: 10.1109/IV.2013.13, *Information Visualisation (IV)*, 2013 17th International Conference, 108–117, IEEE (2013)
- Team, R.C. R: A language and environment for statistical computing (2013)
- Ye, Q., Z. Zhang y R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, doi: 10.1016/j.eswa.2008.07.035, *Expert Systems with Applications*, 36 (3, Part 2), 6527-6535 (2009)