Artículo Original

Aplicación de Técnicas de Minería de Datos para el Análisis de la Eficiencia Académica

Application of Data Mining Techniques for the Analysis of Academic Efficiency

Paola K. Grijalva Arriaga, Verónica A. Freire Avilés, Karina P. Real Avilés, Ana M. Arellano Arcentales

Universidad Agraria del Ecuador, Guayaquil, Ecuador, y

Galo Cornejo Gómez

Universidad Católica Santiago de Guayaquil, Guayaquil, Ecuador.

La correspondencia sobre este artículo debe ser dirigida a Paola K. Grijalva Arriaga. Email: pgrijalva@uagraria.edu.ec

Fecha de recepción: 6 de octubre de 2017.

Fecha de aceptación: 15 marzo de 2018.

¿Cómo citar este artículo? (Normas APA): Grijalva Arriaga, P.K., Freire Avilés, V.A., Real Avilés, K.P., Arellano Arcentales, A.M., Cornejo Gómez, G. (2018). Aplicación de Técnicas de Minería de Datos para el Análisis de la Eficiencia Académica. *Revista Científica Hallazgos21,3,*(Suplemento Especial).Recuperado de: http://revistas.pucese.edu.ec/hallazgos21/

Resumen

Las Instituciones de Educación Superior están inmersas en procesos periódicos de evaluación y acreditación que exigen el cumplimiento de estándares mínimos, siendo uno de ellos la Eficiencia Académica, compuesta por el indicador Retención Estudiantil, que se determina por la cantidad de estudiantes matriculados por primera vez al primer año manteniéndose en sus estudios dos años después. El presente estudio se desarrolla en la Universidad Agraria del Ecuador y tiene como objetivo determinar los factores de mayor incidencia en la deserción estudiantil, y se utilizó la Minería de Datos, con las técnicas de árboles de decisión y *clustering* para obtener los patrones de comportamiento que conllevan a la deserción, generando un modelo que permita predecir las deserciones. Se trabajó con los estudiantes de primer año de las diferentes carreras de las sedes de Guayaquil y Milagro, de los períodos lectivos 2014, 2015 y 2016, aplicando la metodología para descubrimiento de conocimiento (KKD,*Knowledge Discovery in Database*) donde se integran los datos iniciales, se seleccionan y se transforman los datos para la aplicación de la técnica escogida y posteriormente se evalúan y se difunden para permitir a las autoridades correspondientes la toma de decisiones, que conlleven a generar planes de acción que incrementen la tasa de retención de los estudiantes en la institución. Con la aplicación de las técnicas citadas se obtuvo que el factor más relevante que afecta a la tasa de retención en la institución, son los bajos promedios obtenidos durante los primeros semestres, así como la cantidad de asignaturas aprobadas, que conllevan a la pérdida de año de los niveles objeto de estudio.

Palabras clave: minería de datos; técnicas de minería de datos; eficiencia académica; árboles de decisión.

Abstract

The Institutions of Higher Education are immersed in periodic processes of evaluation and accreditation that demand the fulfillment of minimum standards, being one of them the Academic Efficiency, integrated by the Student Retention indicator, which is determined by the number of students enrolled for the first time to the first year and are still studying two years later. The present study was carried out at the Agrarian University of Ecuador with the goal of predicting factors of higher incidence that generate student dropout in higher education, using Data Mining, and decision tree techniques and Clustering to obtain desertion behavior patterns. We worked with first year students of different careers at the Guayaquil and Milagro venues, during the academic periods 2014- 2015 and 2016, applying the Knowledge Discovery in Database (KKD) where the initial data is integrated, selected and transformed for the application of the chosen technique and are subsequently evaluated and spread to allow the corresponding authorities to make decisions that lead to the generation of action plans, which increase the student retention rate in the institution. With the application of the mentioned techniques it was concluded that the most relevant factor affecting the retention rate in the institution are the low averages obtained during the first semesters as well as the subjects approved, which lead to the loss of the year of the levels under study.

Keywords: data mining; data mining techniques; academic efficiency; clustering; decision tree.

Aplicación de Técnicas de Minería de Datos para el Análisis de la Eficiencia Académica

Las instituciones de educación superior del Ecuador (IES) están inmersas en procesos de evaluación y acreditación, que

les exigen el cumplimiento de estándares mínimos y las mantienen inmersas en modelos de mejora continua para avalar la calidad de sus servicios educativos, siendo uno de estos, la Eficiencia Académica (EA), que se centra en definir estrategias que permitan a los estudiantes realizar sus estudios en el tiempo adecuado, garantizando su formación académica y preparación profesional en el tiempo mínimo requerido por la institución.

La EA es un indicador evaluativo, cuya dimensión cuantitativa refleja aspectos cualitativos que van desde la calidad del sistema educativo precedente, las políticas de ingreso a la institución, hasta una amplia gama de factores que intervienen en el proceso docente educativo y en las políticas institucionales que garanticen la permanencia de los estudiantes (Rodríguez, Gutiérrez, Wong, & López, 2015), además de la titulación de los mismos.

En el Modelo de Evaluación Institucional de Universidades y Escuelas Politécnicas emitido por el Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES), considera, para el cálculo de la EA, la permanencia o retención de los estudiantes (Patiño Garzón & Cardona Pérez, 2012), es decir, se consideran a aquellos estudiantes que inician su primer año y continúan sus estudios dentro de los siguientes dos períodos lectivos.

Para que una institución pueda obtener una EA elevada, es necesario que se estudien los factores que influyen en la deserción o abandono de la carrera (Octubre, Basilio, Concepción, & Basilio, 2016), entre estos factores los que más se destacan son los de índole académico y socio-económico.

De acuerdo con la información proporcionada por la Universidad Agraria del Ecuador, se puede observar que existen carreras que poseen mayor número de deserciones que otras y al no tener un seguimiento adecuado, se desconocen los factores por los cuales los estudiantes no continúan con sus estudios, y no se han podido tomar las precauciones para disminuir esta deserción. Es por esta razón que se realizó un análisis cuantitativo de la retención de los estudiantes que ingresaron por primera vez al primer año en el 2014 y se mantuvieron en la carrera durante dos períodos consecutivos, aplicado de acuerdo con el estándar del modelo de evaluación. La retención total actualmente en la institución es del 65.39%.

En este artículo se consideraron las técnicas de Minería de Datos y el descubrimiento del conocimiento (Dubey, Dubey, Agarwal, & Khandagre, 2012) para determinar los patrones de comportamiento y relaciones entre los diferentes atributos, que permitan identificar y predecir probabilidades de deserción estudiantil, previendo los factores que puedan influir en su permanencia, generando conocimiento para la toma de decisiones oportunas y dar ventaja competitiva a la institución.

Método

Para determinar los factores que influyen en la eficiencia académica en la Universidad, se utilizaron las técnicas de Minería de Datos directa predictiva que permite obtener resultados futuros basados en datos históricos y actuales, descubriendo patrones y relaciones de captura en los datos, siendo un modelo de clasificación estimado para medir la precisión de los datos iniciales, mediante el uso de un conjunto de pruebas dado un porcentaje de conjunto de tuplas (Assun, 2014).

Una de las técnicas utilizadas son los árboles de decisión (Jankowski & Jackowski, 2014) utilizando el algoritmo C4.5 que permite trabajar con valores continuos, separando los resultados en dos ramas, con el que se agrupó a los estudiantes de

acuerdo a sus atributos o características similares, encontrando patrones en los datos, permitiendo clasificar los factores que afectan a la tasa de retención.

Cada nodo del árbol contiene un punto de división que es una prueba de uno o más atributos y determina cómo se dividen los datos (Yadav, 2012). El algoritmo C4.5 (Pal, 2017), es un clasificador estadístico que utiliza la técnica de post-poda para mejorar la legibilidad del árbol. Se genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, se construye con la estrategia de primero profundidad (*depth-first*), considerando todas las pruebas posibles al dividir los datos en conjuntos y se selecciona la prueba que haya generado el mayor ratio de ganancia (Dai & Ji, 2014).

La técnica de Minería Indirecta aplicada es el clustering que utiliza métodos descriptivos (Berkhin, n.d.), que divide los datos en grupos de objetos similares mediante algoritmos matemáticos, simplificando la información, aunque en el proceso se pierdan algunos detalles de los datos (Sidhu & Kaur, 2013). Entre los algoritmos de clustering más utilizados se encuentran Simple K-Means (Cui, Zhu, Yang, Li, & Ji, 2014), X-Means (Yang, Plant, Shao, He, & Bo, 2013) y Cobweb (Karami, 2013). Una de las ventajas de esta técnica es su naturaleza flexible que le permite combinarse con otras técnicas de minería de datos, generando sistemas híbridos, pero presenta problemas en la selección de factores de tareas de clasificación debido a que no todas las variables tienen la misma importancia a la hora de agrupar los datos (Hossain, Ramakrishnan, Davidson, & Watson, 2013).

Software de Predicción

El software de predicción utilizado en este estudio es Knime, que consiste en una plataforma de exploración de datos modular, desarrollado por el departamento de bioinformática y minería de datos de la Universidad de Konstanz en Alemania (Mihanović, Gabelica, & Krstić, n.d.). Permite crear flujos de datos, ejecutar los pasos de análisis de datos e investigar por medio de vistas interactivas los resultados, incorpora más de 100 nodos de entrada y salida de datos, integrando todos los módulos de análisis de minería de datos e incluye rutinas estadísticas (Lausch et al., 2015).

La metodología que se siguió para obtener el conocimiento que permitió determinar los factores que influyen en la deserción estudiantil y por lo tanto permita mejorar la EA en la Universidad Agraria del Ecuador, se resume en cuatro pasos iniciando con el objeto de estudio determinando la población de estudiantes; para luego establecer el procesamiento y manejo de los datos para la recopilación e integración de los mismos, la selección, limpieza y transformación; aplicar las técnicas de Minería de Datos que permitan generar y validar el modelo para la interpretación de sus resultados; difusión y toma de decisiones. En la Figura 1 se puede observar el proceso utilizado para la generación del conocimiento dentro de este estudio.

Determinación del objeto de estudio

En este estudio se analizó la información personal, socio-económica, académica, entre otros aspectos importantes, de los estudiantes que ingresaron al primer año por primera vez en las diferentes carreras de la Universidad Agraria del Ecuador en el año 2014, para determinar las causas de deserción de los estudiantes que no continuaron en dos períodos lectivos siguientes, es decir en el 2016; y poder realizar un modelo que permita predecir posibles deserciones en las siguientes generaciones. Las carreras que se consideran dentro del estudio corresponden a Ingeniería en Computación e Informática,

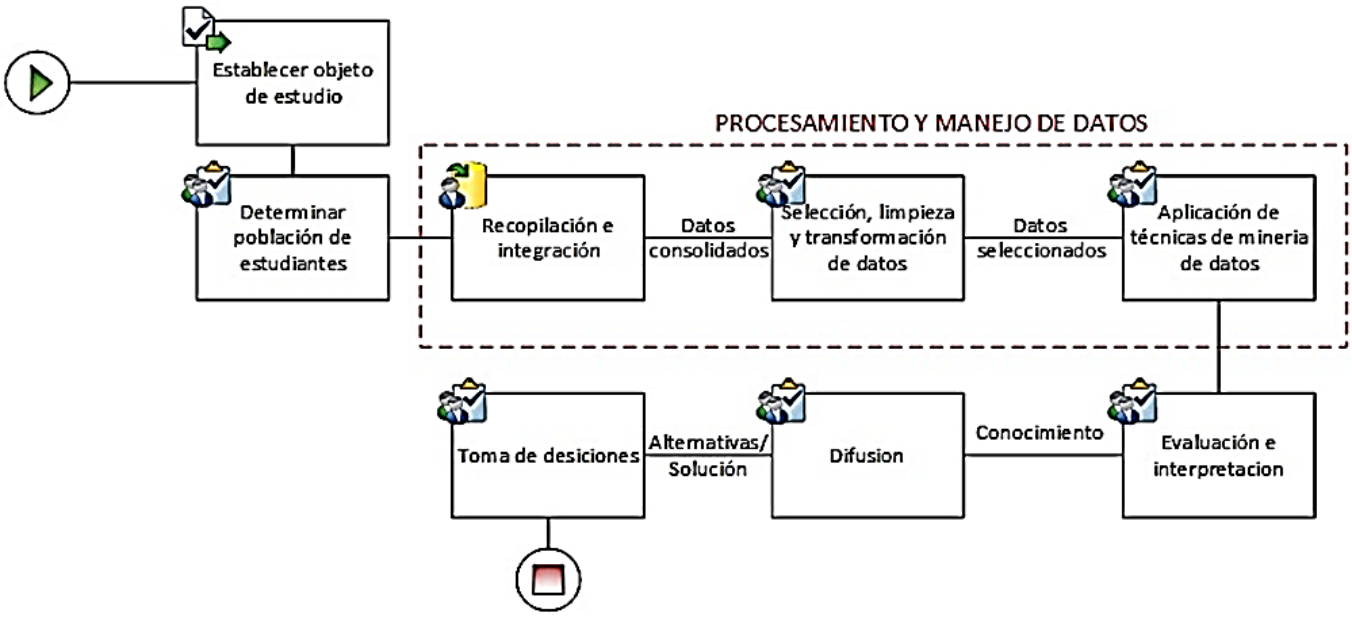

Figura 1. Metodología utilizada para la generación de conocimiento. Fuente: Elaboración de los autores.

Ingeniería Agronómica, Ingeniería Ambiental, Ingeniería Agrícola mención Agroindustrial, Medicina Veterinaria y Economía Agrícola, de las Sedes Guayaquil y Milagro.

Pre - procesamiento y Manejo de Datos

Para obtener el conocimiento utilizando Minería de Datos, no sólo se logra en la consecución de un modelo y patrones, sino en la evaluación e interpretación de los resultados obtenidos, con la finalidad de una toma de decisiones oportuna, aplicando limpieza de datos, integración, transformación, reducción y discretización para completar valores perdidos, eliminar datos redundantes, normalizarlos y agruparlos (Waller & Fawcett, 2013). El pre-procesamiento de datos y su ejecución resulta de vital importancia para obtener patrones de buena calidad, es por ello que se estima que el 50% del tiempo de análisis de datos se destina a esta fase para evitar errores, redundancias e inconsistencias que existen en los datos organizacionales (Kamiran & Calders, 2012).

Recopilación e Integración

La recopilación de los datos (Rubiano & García, 2016) es generar un modelo predictivo de los datos académicos de los estudiantes de la Universidad Agraria del Ecuador, los mismos que se seleccionaron de acuerdo a las fuentes de datos del sistema académico SAIIS, como son las tablas Alumno, Histórico, Calificaciones, Facultad, Carrera, Sede, y archivos adicionales con información personal, recopilada de las diferentes carreras, de los períodos lectivos 2014, 2015 y 2016 que son reportados anualmente a los organismos de control de Educación Superior en el país. Una vez recopilados los datos, fueron cargados en el sistema gestor de base de datos SQL, con la información personal, académica y financiera de los estudiantes de la institución, quedando integrados en un almacenamiento común.

Selección, Limpieza y Transformación

El proceso de limpieza de datos consiste en detectar y corregir errores en estos, donde se pueden aplicar varios tipos de reglas para su calidad, de manera que contribuyan a mejorar la eficiencia y eficacia

Tabla2
Descripción de atributos seleccionados para la creación de la vista minable

Atributo	Descripción
Edad	Generado a partir de la fecha de nacimiento.
Tipo Beca	Tipos de becas que se ofrecen a los estudiantes
Apoyo_econ	Financiamiento de los estudios
Calificaciones	Calificaciones parciales por cada una de las asignaturas
Asignaturas	Asignaturas que debe tomar el estudiante de acuerdo al nivel
Promedio	Promedio obtenido por los estudiantes en el año de estudio
Sexo	Genero del estudiante H(Hombre), M(Mujer)
Discapacidad	S (Si), N (No)
Estado civil	C (Casado), S(soltero), D(Divorciado), V (Viudo)
Sede	Guayaquil, Milagro
Carrera	Ingeniería en Computación e Informática, Ingeniería Agronómica, Ingeniería Ambiental, Ingeniería Agrícola mención Agroindustrial, Medicina Veterinaria y Economía Agrícola

Fuente: Datos obtenidos del Sistema Académico SAIIS.

de los algoritmos de limpieza (Chu, s.f.). En esta fase se estandarizaron los documentos, verificando la existencia de datos anómalos y datos faltantes. En los archivos adicionales se estandarizaron los identificadores de los campos de datos personales de los estudiantes, se completaron aquellos campos nulos o que estaban almacenados con diferentes formatos, además se eliminaron los registros repetidos.

La Tabla 1 presenta los atributos seleccionados para facilitar la fase de minería de datos a partir de los datos personales, financiero y académico de los estudiantes.

Tabla 1
Vista minable generada en SQL

Atributo	Valores posibles
Sexo	hombre (H), mujer (M)
Edad	Numérico
Discapacidad	si (S), no (N)
Estado_civil	casado (C), soltero (S), viudo (V), divorciado (D)
Carrera	Ciencias Económicas, Economía, Ingeniería Agrícola mención Agroindustrial, Ingeniería Agronómica, Ingeniería Ambiental, Ingeniería en Computación e Informática, Medicina Veterinaria
Apoyo_econ.	Si(S), no (N)
Num_mat_aprob	Cantidad de materias que ha aprobado
Num_mat_reprob	Cantidad de materias que ha reprobado
Nivel_aprob_mat	Cantidad de materias aprobadas/total de materias del nivel
Sede	Guayaquil, Milagro
Promedio	Real (Sumatoria Notas / Cantidad de Asignaturas)
Retención	si (S), no (N)

Fuente: Datos del Sistema de Matriculación SAIIS.

Aplicación de las Técnicas de Minería

Para la aplicación de las técnicas citadas anteriormente, árboles de decisión mediante el algoritmo C4.5 y el *clustering*, (k-means), se utilizó el software KNIME que opera en modalidad open source y contiene varias técnicas y algoritmos de KDD. La Tabla 2 muestra la vista minable generada en SQL para ser utilizada en el proceso de minería de datos.

Un aspecto importante para el estudio es la despersonalización de la información, que corresponde a eliminar o sustituir los campos de la cédula de ciudadanía, nombres y apellidos de los estudiantes, es

decir, que no se podrá asociar la información obtenida con un estudiante o grupo de estudiantes de manera particular.

Para encontrar los patrones en base al rendimiento académico se utilizaron los atributos num_mat_aprob, num_mat_reprob, nivel_aprob_mat y promedio, y para la búsqueda de patrones en relación con el ámbito económico se utiliza el atributo apoy_econ el atributo carrera para encontrar patrones de acuerdo con el lugar de estudio.

El atributo retención es el campo Clase, y nos identifica el estado del estudiante, es decir, si ha permanecido durante dos años consecutivos en la institución desde su ingreso por primera vez en primer año.

Generación y Validación del modelo.

Cluster (*K-means*)

Utilizando la técnica de *cluster* se buscó obtener, por medio de agrupaciones, las características comunes que posee el grupo de estudiantes en estudio. Se analiza la retención como un atributo, en función del ámbito académico y socio económico. En el ámbito académico se analiza en función del promedio, el número de materias aprobadas y el nivel de materias aprobadas, que resulta de la cantidad de materias aprobadas en relación con las materias inscritas. Y, en relación con el aspecto socio económico se analizan aspectos como el género y la edad. Para el análisis de la retención utilizando esta técnica, se utilizó el algoritmo k-medias (k-means), proporcionando el número de clusters en los que se segmentó la base de estudiantes. El número de clusters utilizado fue de tres.

Árboles de Decisión (C4.5)

Para la generación del modelo de predicción de la retención de los estudiantes, utilizando árboles de decisión con el algoritmo C4.5, se carga a la herramienta Kmine los datos de los estudiantes del período 2014 de quienes se conoce si continúan o no sus estudios, utilizando como variables independientes el sexo, edad, discapacidad, estado civil, carrera, apoyo económico, número de materias aprobadas, número de materias reprobadas, nivel de materias aprobadas, sede y promedio; como variable dependiente o de predicción, la retención.

Se tomó el 60% de los datos para el entrenamiento y como prueba el 40%, para la generación del modelo se trabajó con un máximo de 2 nodos.

El modelo considera como variable predominante el número de materias aprobadas y el nivel de materias aprobadas. Las reglas generadas y el árbol de decisión obtenido se presentan en las figuras 2 y 3, respectivamente.

```
$NUM_MAT_APROB$ <= 25.0 AND $NUM_MAT_APROB$ <= 35.0 => "N"
$NIVEL_APROB_MAT$ <= 91.17647058823529 AND $NUM_MAT_APROB$ > 25.0 AND $NUM_MAT_APROB$ <= 35.0 => "S"
$NIVEL_APROB_MAT$ > 91.17647058823529 AND $NUM_MAT_APROB$ > 25.0 AND $NUM_MAT_APROB$ <= 35.0 => "N"
$NUM_MAT_APROB$ > 35.0 AND TRUE => "S"
```

Figura 2. Reglas del modelo generadas en KNIME. Fuente: Autores.

Al evaluar el modelo, se obtuvo una Accuracy del 93,90%, Sensitivity de 89,7% para la deserción y 95,8% para la retención; Specificity de 95,8% para la deserción y el 89,7% en los casos de retención y el F-measure fue de 89,2% y 94,5% para cada variable respectivamente. El modelo obtuvo un error de clasificación del 6,098% debido a que hubo 15 errores de clasificación y 231 registros correctamente generados de la

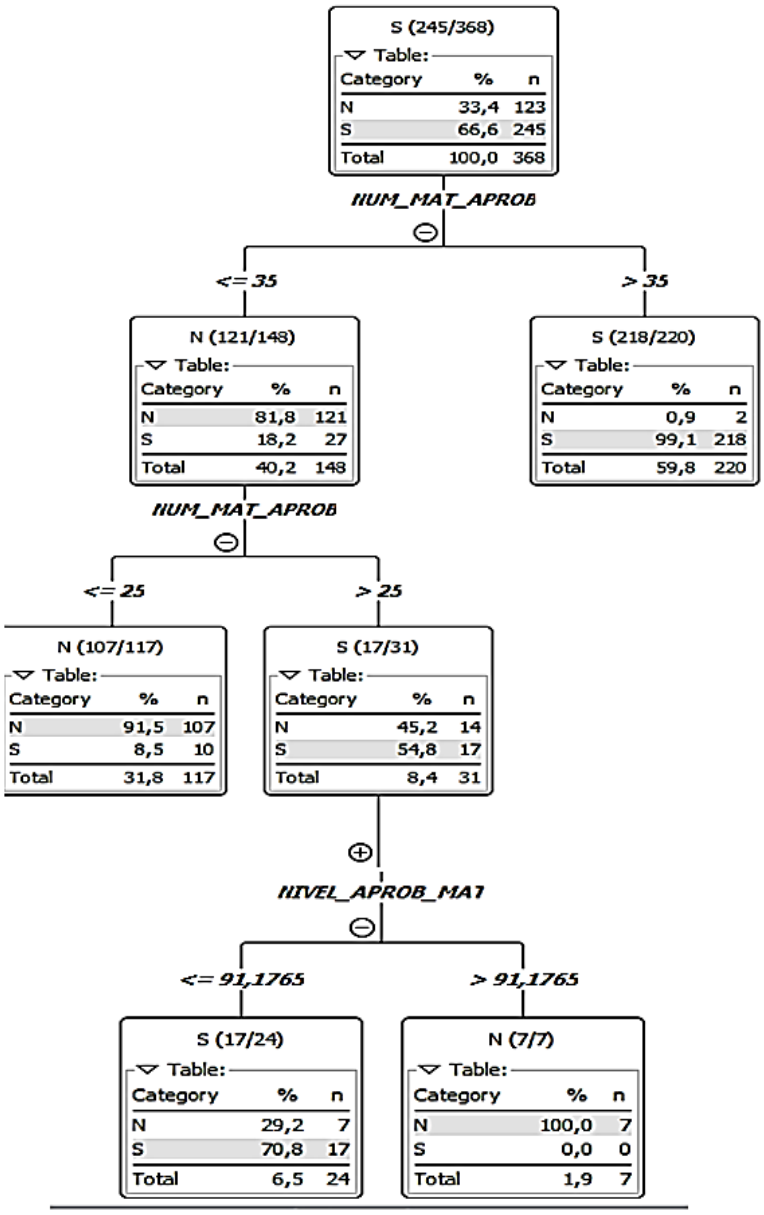

Figura 4. Árbol de decisión obtenido. Fuente: Autores.

muestra, con los que se realizó la prueba. La Accuracy obtenida nos indica que el modelo es confiable.

RETENCION \ Prediction (RETENCION)	N	S
N	70	8
S	7	161

Correct classified: 231 Wrong classified: 15

Accuracy: 93,902 % Error: 6,098 %

Cohen's kappa (κ) 0,859

Figura 3. Matriz de Confusión del modelo obtenido. Fuente: Autores.

En la Figura 4 se puede observar la matriz de confusión generada del modelo obtenido.

Resultados y Discusión

Una vez aplicadas las técnicas y generado el modelo, se analizan los resultados de la aplicación de estas.

Utilizando la técnica de *clustering* y el algoritmo K-means con tres clusters para el análisis de los datos de los estudiantes, se observó que la agrupación se realiza en base a los atributos promedio y el número de materias aprobadas, atributos que predominaron para la formación de estos. Las representaciones de los clusters se detallan en la Tabla 3.

Estos tres grupos conformados se representan por colores, el azul representa los alumnos con mayor número de materias aprobadas y elevado promedio, rojo claro aquellos que poseen mediano promedio y materias aprobadas; y con rojo aquellos estudiantes que tienen promedio cercano a 0 y muy pocas o ninguna materia aprobada.

Se analizaron las diferentes variables en función de la retención, para ver el comportamiento de los grupos y obtener información de relevancia para la institución.

En la Figura 5, se analiza la Retención en función de las diferentes carreras de la Institución.

Se observa que aquellos estudiantes que permanecen en la Institución son aquellos que presentan mayor número de materias aprobadas y un promedio mayor a 6. Las carreras de Ingeniería

Tabla 3
Descripción de Clusters

Cluster	Atributos
Cluster_0	Promedio mayor a 6 con alto número de materias aprobadas
Cluster_1	Promedio entre 4 y 5 con menor número de materias aprobadas
Cluster_2	Promedio cercano a 0 y pocas materias aprobadas

Fuente: Knime.

en Computación e Informática, Ingeniería Agronómica y Economía, tienen estudiantes que no continúan con sus estudios teniendo buen promedio y un elevado número de materias aprobadas. Las carreras de mayor deserción son Ingeniería en Computación e Informática e Ingeniería Agronómica.

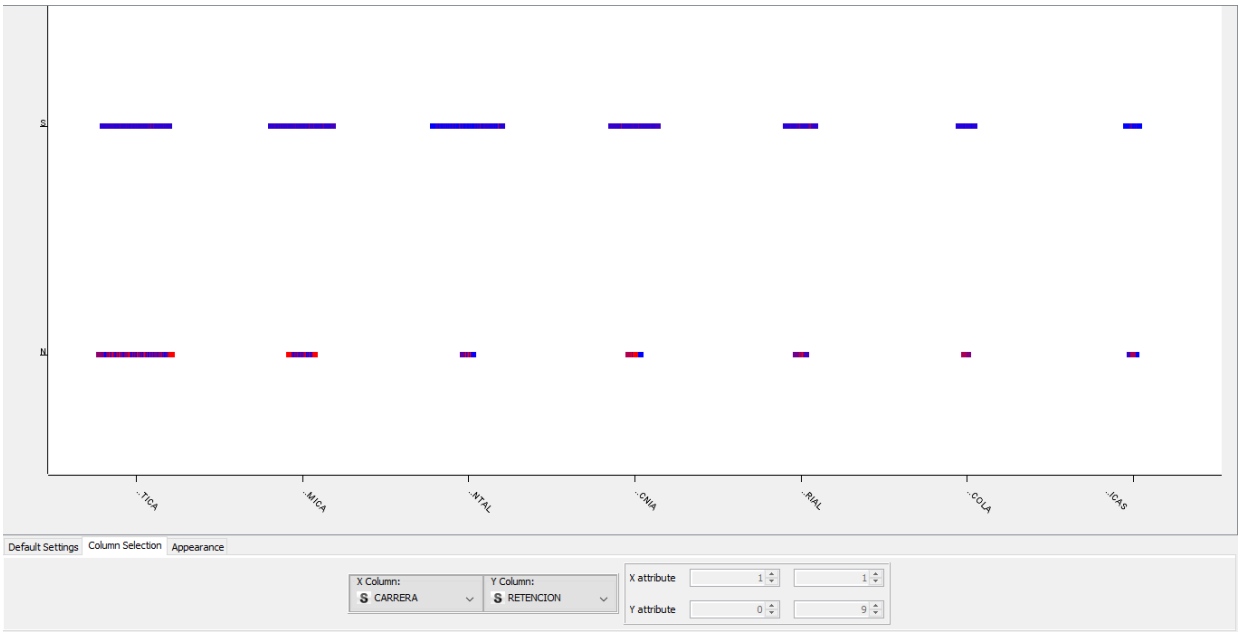

Figura 5. Análisis Retención - Carrera. Fuente: Autores.

Se evidencia también, en la carrera de Ingeniería en Computación e Informática, que el número de estudiantes que ingresan a primer año y continúan sus estudios, es similar al número de los que desertan.

En la Figura 6 se analizan los grupos en función de las variables Retención y el Nivel de Materias aprobadas.

En este análisis, se evidencia que hay un pequeño grupo de estudiantes que teniendo buen nivel de materias aprobadas, no permanecen en la Institución.

En la Figura 7 se analizan los grupos en función de la Retención y el Promedio obtenido. Este análisis permitió observar que el promedio general de los estudiantes en la institución fluctúa entre 6 y 7 sobre 10, y el número de estudiantes con un promedio superior a 9 es reducido. Se evidencia también que existen estudiantes que se retiran a pesar de tener promedios superiores a 7.

En el ámbito social, en la Figura 8, se analiza la Retención en relación al género. De este análisis se observó que en la institución existe equidad de género, evidenciando que el mayor número de deserciones corresponden al género masculino, existiendo inclusive algunos con

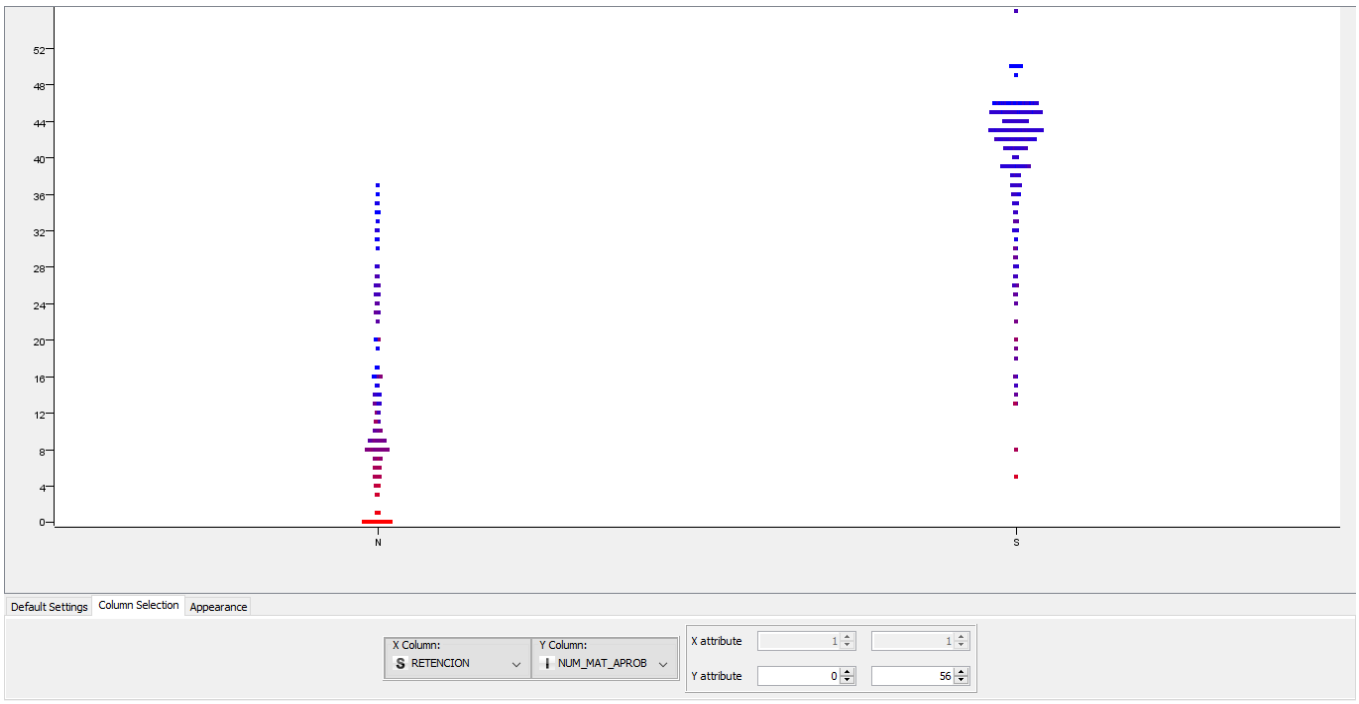

Figura 6. Análisis Retención - Nivel de materias aprobadas. Fuente: Autor.

buen promedio y número de materias aprobadas.

se puede ver en la Figura 9, se analiza la Retención en función de la edad de los estudiantes. Este análisis permitió observar

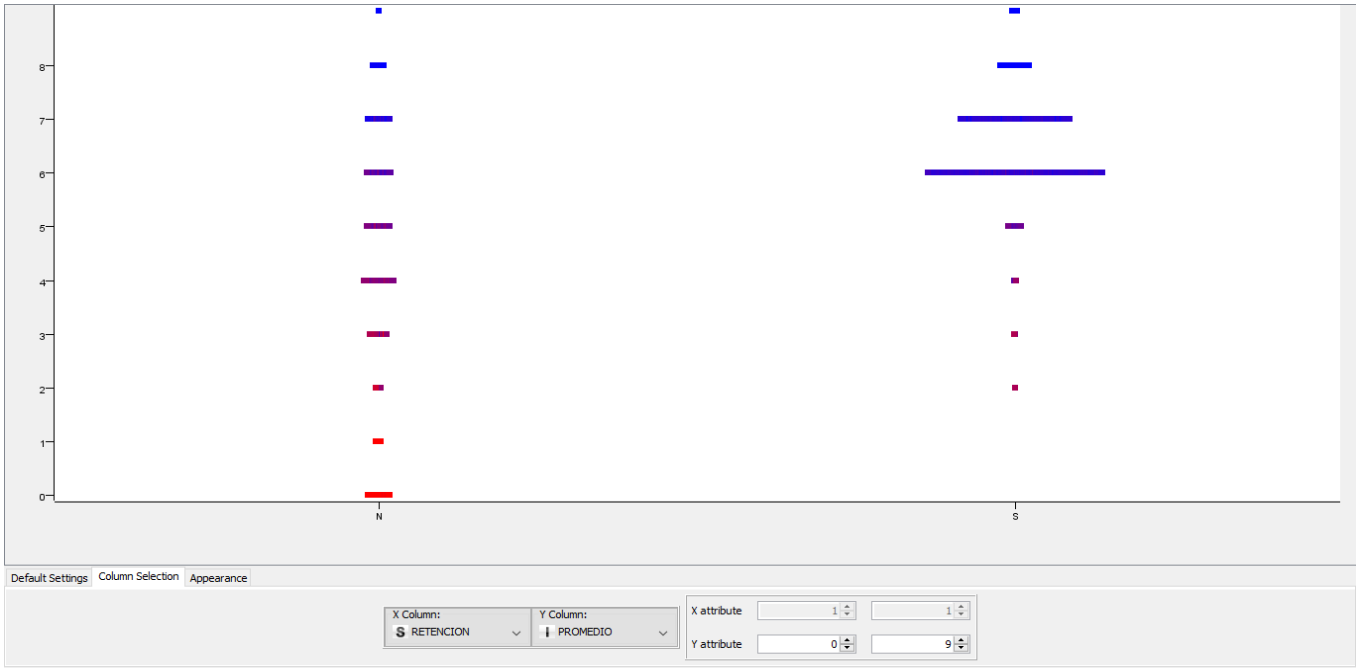

Figura 7. Análisis de la Retención - Promedio. Fuente: Autores.

La edad es otro de los factores sociales que influye en la permanencia de los estudiantes. En otro análisis realizado como

que los estudiantes cuyas edades fluctúan entre los 18 y 20 años son los más vulnerables a desertar en sus carreras.

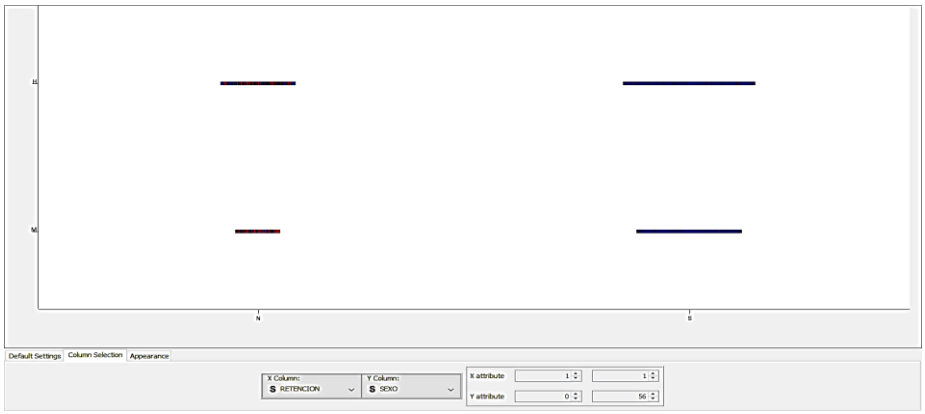

Figura 9. Análisis Retención - Género. Fuente: Autores.

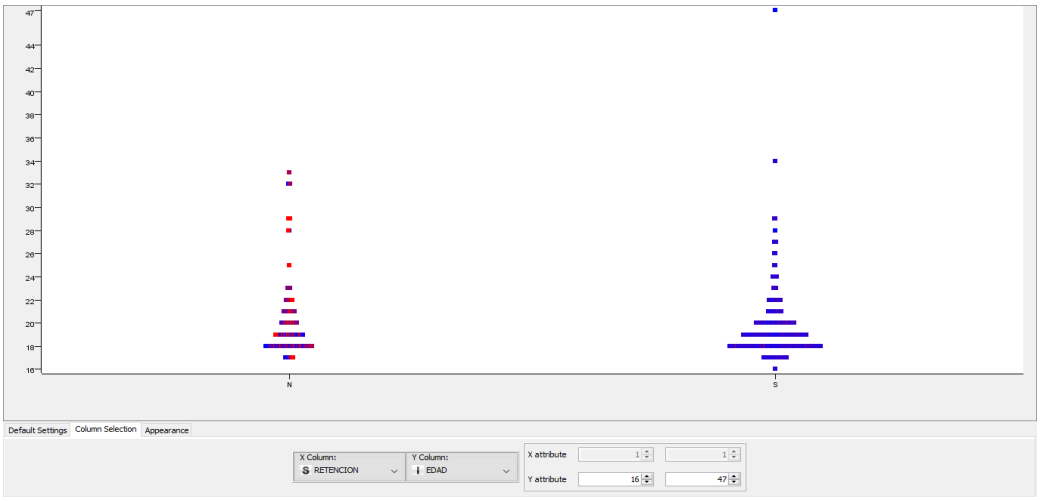

Figura 8. Análisis Retención - Edad. Fuente: Autores.

En la Tabla 4, se muestra un resumen de los diferentes análisis realizados de la información de los estudiantes.

Los resultados producto del análisis utilizando esta técnica, son similares a los obtenidos en Colombia, en la Universidad de Nariño, donde se realizó un proyecto de investigación para detectar patrones de deserción estudiantil utilizando técnicas de Minería de Datos, construyendo un repositorio de datos con información académica, socioeconómica, disciplinar e institucional de los estudiantes, con una proyección a 5 años, utilizando técnicas de clasificación y *clustering*, para definir un patrón común de deserción estudiantil en el cual se destacan los promedios bajos y materias reprobadas en los primeros semestres de la carrera, existiendo también factores socioeconómicos que contribuyeron a la deserción estudiantil. El conocimiento generado permitió la toma de decisiones más certeras por parte de las autoridades para formular políticas y estrategias relacionadas con los programas de deserción estudiantil. (Timar & Jim, 2013)

Al utilizar la técnica de árbol de decisión, se evaluó la información de los estudiantes de primer año que ingresaron en el 2015, considerando todas las sedes de la institución, esto es, Guayaquil, Milagro, El Triunfo y Naranjal, que, de acuerdo con las características que van presentando durante los periodos 2016 y el 2017, podrían no continuar con sus estudios.

Como resultado se obtuvo que el modelo predice que 818 estudiantes, que corresponde al 20% de los ingresos de este año, presentan características que podrían terminar con una deserción. Esta información permitirá a la institución tomar decisiones a nivel académico y administrativo para disminuir la deserción y

Tabla 4
Análisis de Variables realizadas con Knime

Variables	Resultados	Observación
Retención – Carrera (Figura 4)	Permanencia de estudiantes que poseen mayor promedio y mayor número de materias aprobadas,	Carrera con mayor deserción: Ingeniería en Computación e Informática, Ingeniería Agronómica, Economía
Retención - Nivel de materias aprobadas (Figura5)	Permanencia de estudiantes que tienen mayor nivel de materias aprobadas.	
Retención - Promedio (Figura 6)	Permanencia de estudiantes que obtienen mayores promedios.	Existe un pequeño grupo con deserción a pesar de tener la misma característica
Género (Figura 7)	Existe mayor la deserción en los hombres	Este grupo tiene un número considerable de materias aprobadas y promedios superiores.
Edad (Figura 8)	Estudiantes que más desertan son con edades comprendidas entre 18 y 20 años.	Corresponderían a los primeros años de educación superior

Fuente: Resultados de Análisis utilizando Clustering.

mejorar los índices de retención en la Institución, a través de las unidades correspondientes.

Similar resultado se presenta en otras universidades como la Privada César Vallejo de Perú, quien utilizó las técnicas de Minería de Datos para predecir la deserción o el abandono en la educación superior privada, bajo la técnica de árboles de decisión, de la Escuela profesional de Ingeniería de Sistemas, considerando 27 atributos que fueron extraídos del área de registros académicos, Asuntos Estudiantiles y del área de Informática, se hizo el entrenamiento, validación y prueba con 100 datos nuevos en donde se obtuvo una precisión de 89% (Vergaray, 2016). Otro estudio realizado fue el de la Universidad Simón Bolívar de Barranquilla – Colombia, donde se construyó un modelo predictivo para la deserción estudiantil con el objetivo de poder predecir cuál es la probabilidad de que un estudiante abandone sus estudios utilizando árboles de decisión y comparando entre ellos sus resultados, WEKA fue el software utilizado debido a las múltiples herramientas que ofrece (Heredia, Amaya, & Barrientos, 2015).

En la Universidad Nacional de Río Negro (UNRN), a través del uso de las técnicas de Data Mining, se diseñó un modelo de abandono universitario con el fin de identificar las características, tras procesar los datos y analizar las proyecciones de atributos para clases o respuestas esperadas, siendo sus resultados satisfactorios, lo que les permite establecer acciones encaminadas a disminuir el porcentaje de estudiantes que abandonan sus estudios (Delegation, Formia, & Lanzarini, 2013).

Es necesario que en la Institución se impulse más estudios utilizando Minería de datos, que permitan determinar los aspectos sociales y económicos de estos grupos, mejorar la calidad educativa y de servicio al estudiante.

Conclusiones

En muchas universidades de Latinoamérica existe una gran preocupación por la deserción escolar y las altas tasas observadas en los primeros años de estudio, por lo que se han desarrollado trabajos de investigación con respecto a la deserción en la Educación Superior con la aplicación de la Minería de Datos, sus diferentes técnicas, aplicando diversos algoritmos.

A través del procesamiento tradicional de grandes cantidades de datos, es muy difícil visualizar información oculta que puede ser importante para la toma de decisiones, la minería de datos y los algoritmos ID3, C4.5 de árboles de decisión, han resultado eficientes en modelos de predicción.

Se puede concluir que uno de los factores principales en la institución que genera un elevado porcentaje de deserción estudiantil, es que los estudiantes no aprueban las asignaturas cursadas, obteniendo promedios menores a 6 puntos, y durante los dos primeros años pierden aproximadamente entre 3 y 4 asignaturas, lo cual demuestra que llegan con una baja preparación para cumplir con las exigencias del ámbito universitario, y ocasiona que la institución posea más del 34% de deserción. Todas las variables socio económico o personales no fueron considerables en este estudio, debido a la falta de información o porque existen muchos valores erróneos.

Actualmente, los sistemas de información implementados en la institución no contribuyen a la generación del conocimiento necesario para el análisis y la toma de decisiones, por lo cual es necesario que los procesos de registro de datos personales, económicos y académicos de los estudiantes sean obligatorios y actualizados periódicamente, con la finalidad de que se puedan ampliar este tipo de estudios y análisis de la información.

La utilización de técnicas de minería, y especialmente árboles de decisión, nos permitió generar reglas que pueden predecir futuras deserciones, por lo que a los sistemas de información institucionales se les puede añadir el modelo creado, para el análisis de la información institucional. La difusión de los resultados obtenidos, producto de este estudio, se entregaron mediante un informe a las autoridades académicas de la IES para su conocimiento y toma de decisiones de las unidades correspondientes.

Como trabajo futuro al presente estudio, se debe identificar, mediante el uso de Técnicas de Minería, las asignaturas que con mayor frecuencia pierden los estudiantes en las diferentes carreras, durante sus primeros años de estudio, para mejorar el proceso de tutorías o seguimiento académico, así como el análisis de la tasa de titulación institucional.

Referencias

Assun, M. D. (2014). Big Data Computing and Clouds : Trends and Future Directions, 1–44.

Berkhin, P. (n.d.). Survey of Clustering Data Mining Techniques, 1–56.

Chu, X. (n.d.). Holistic Data Cleaning : Putting Violations Into Context, (L).

Cui, X., Zhu, P., Yang, X., Li, K., & Ji, C. (2014). Optimized big data K-means clustering using MapReduce, 1249–1259. https://doi.org/10.1007/s11227-014-1225-7

Dai, W., & Ji, W. (2014). A MapReduce Implementation of C4 . 5 Decision Tree Algorithm, *7*(1), 49–60.

Delegation, A. C., Formia, S., & Lanzarini, L. (2013). Characterization of University Drop-Out at UNRN Using Data Mining . A Study Case, 681–690.

Dubey, A. K., Dubey, A. K., Agarwal, V., & Khandagre, Y. (2012). Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support. *2012 CSI 6th International Conference on Software Engineering, CONSEG 2012*. https://doi.org/10.1109/CONSEG.2012.6349495

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. https://doi.org/10.1109/TLA.2015.7350068

Hossain, M. S., Ramakrishnan, N., Davidson, I., & Watson, L. T. (2013). How to ``alternatize'' a clustering algorithm. *Data Mining and Knowledge Discovery*, *27*(2), 193–224. https://doi.org/10.1007/s10618-012-0288-4

Jankowski, D., & Jackowski, K. (2014). Evolutionary Algorithm for Decision Tree Induction. In

K. Saeed & V. Snášel (Eds.), *Computer Information Systems and Industrial Management: 13th IFIP TC8 International Conference, CISIM 2014, Ho Chi Minh City, Vietnam, November 5-7, 2014. Proceedings* (pp. 23–32). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45237-0_4

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Karami, A. (2013). Data Clustering for Anomaly Detection in Content-Centric Networks, *81*(7), 1–8.

Lausch, A., Schmidt, A., Tischendorf, L., Lausch, A., Schmidt, A., & Tischendorf, L. (2015). Data mining and linked open data – New perspectives for data analysis in environmental in environmental research. *Ecological Modelling*, *295*(October), 5–17. https://doi.org/10.1016/j.ecolmodel.2014.09.018

Mihanović, A., Gabelica, H., & Krstić, Ž. (n.d.). Big Data and Sentiment Analysis using KNIME : Online Reviews vs . Social Media.

Octubre, A. E., Basilio, L., Concepción, N., & Basilio, L. (2016). Causes of the student desertion in the University of San Carlos of Guatemala in New Concepcion, Escuintla, 24–51.

Pal, A. K. (2017). Analysis and Mining of Educational Data for Predicting the Performance of Students Analysis and Mining of Educational Data for Predicting the Performance of Students, (May).

Patiño Garzón, L., & Cardona Pérez, A. M. (2012). REVIEW OF SOME STUDIES ON UNIVERSITY STUDENT DESERTION IN COLOMBIA AND LATIN AMERICA. *Theoría*, *21*(1), 9–20. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=112611591&lang=es&site=ehost-live

Rodríguez, K., Gutiérrez, A., Wong, T., & López, D. (2015). Eficiencia académica: un indicador del que se requiere conocer más. *Edumecentro*, *7*(3), 14.

Rubiano, S. M., & García, J. D. (2016). Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance Methodological Development, 39–48.

Sidhu, N. K., & Kaur, R. (2013). Clustering In Data Mining, *4*(April), 710–714.

Timar, R., & Jim, J. (2013). Determinning school dropout profiles using data analysis, 373–383.

Vergaray, D. (2016). A model based on decision trees to predict student dropout in Private Higher Education, 59–73.

Waller, M. A., & Fawcett, S. E. (2013). Data Science , Predictive Analytics , and Big Data : A Revolution That Will Transform Supply Chain Design and Management, *34*(2), 77–84.

Yadav, S. K. (2012). Mining Education Data to Predict Student ' s Retention : A comparative Study, *10*(2), 113–117.

Yang, Q., Plant, C., Shao, J., He, X., & Bo, C. (2013). Synchronization-Inspired Partitioning and Hierarchical Clustering, *25*(4), 893–905.