Asociación Colombiana de Facultades de ingeniería

Revista Educación en Ingeniería
http://www.educacioneningenieria.org

Proceso KDD como apoyo a las estrategias del proyecto SARA (Sistema de Acompañamiento para el Rendimiento Académico)

Leidy Carolina Calvache-Fernández, Valentina Álvarez-Vallejo & Jorge Iván Triviño-Arbeláez

Facultad de Ingeniería, Universidad del Quindío, Armenia, Colombia. lccalvachef@uqvirtual.edu.co, valvarezv@uqvirtual.edu.co, jitrivino@uniquindio.edu.co

Resumen— Uno de los principales problemas que enfrenta Colombia en cuanto a la educación, concierne a los altos índices de deserción estudiantil en la Educación Superior, según fuentes del Ministerio de Educación Nacional, de cada cien estudiantes que ingresan a la educación superior cerca de la mitad no logran concluir su propósito educativo [1]. En este trabajo se presenta el proyecto SARA (Sistema de Acompañamiento para el Rendimiento Académico), proyecto creado por el programa de Ingeniería de Sistemas y Computación de la Universidad del Quindío, además de un análisis y propuesta de incluir el proceso denominado KDD (Knowledge Discovery in Databases), como un soporte de análisis de datos, consiguiendo así definir estrategias que ayuden en la intervención de la vida académica de los estudiantes, a través de la inclusión de técnicas de minería de datos para identificar patrones que permitan caracterizar o predecir posibles casos de deserción dentro del programa.

Palabras Clave — bodega de datos; deserción estudiantil; KDD (knowledge discovery in databases); minería de datos; SARA (sistema de acompañamiento para el rendimiento académico).

Recibido para revisar: Febrero 16 de 2018, aceptado: Mayo 16 de 2018, versión final: Mayo 28 de 2018

KDD (Knowledge Discovery in Databases) process as support of the SARA (Accompaniment System for the Academic Performance) project strategies

Abstract— One of the most important problems which is facing our country about education, it is regarding the high indicators of students dropout in universities. According to the sources of the Ministry of National Education, almost fifty percent of the students who enter to university don't manage to finish their studies. In this study, it is showing the SARA project (Accompaniment System for the Academic Performance), this project was created by the computer science and system engineering program at the Quindío University as well as a proposal to include the process called KDD (Knowledge Discovery in Databases) to support the data analysis, therefore, strategies are defined in order to help to intervene in the academic life of students throughout the use of data mining techniques to identify patterns which allow to profile or predict different cases of dropout inside of the program.

Keywords— data warehouse; dropout; KDD (knowledge discovery in databases); data mining; SARA project (accompaniment system for the academic performance).

1. Introducción

La deserción estudiantil en los programas de pregrado a nivel nacional tiene un impacto negativo en el desarrollo económico y social de un país, ya que las pérdidas financieras y sociales que representan los estudiantes desertores son altas para la sociedad [2]. Existen muchas situaciones que encierran el contexto de la deserción estudiantil como el perfil vocacional que define los intereses, aptitudes, personalidad y capacidades que tiene una persona con respecto a la elección de una carrera universitaria [3], otros factores como los económicos, familiares, sociales y/o personales hacen parte de este contexto. Se puede definir entonces la deserción estudiantil como el fenómeno que comprende a quienes no siguieron la trayectoria esperada de su programa académico, es decir, un estudiante que no se matricula en el mismo programa académico durante dos o más períodos consecutivos y no se encuentra como graduado o retirado por motivos disciplinarios [4].

El Ministerio de educación nacional realiza un seguimiento a la deserción estudiantil mediante el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior (SPADIES), en el informe de Estadísticas Deserción y Graduación del 2015 presentado por el Ministerio de Educación Nacional, se registra una tasa de deserción del 9.3% en formación Universitaria. Por su parte la Universidad del Quindío como lo registra el SPADIES tiene una tasa de deserción del 8.6% al finalizar el año 2015, una tasa superior a la Universidad Nacional de Colombia y la Universidad de Nariño con un índice de deserción del 5.92% y 7.11% respectivamente en el mismo año.

El programa de Ingeniería de Sistemas y Computación como un programa adscrito a la Universidad del Quindío no se encuentra al margen de esta gran problemática, al cierre del año 2016 el SPADIES registró para el Programa de Ingeniería de Sistemas y Computación una tasa de deserción del 12.5 %.

Este documento está estructurado de la siguiente manera: la sección 2 está dedicada al estado del arte donde se evidencian trabajos relacionados y los resultados obtenidos, seguidamente esta la sección 3 donde se contextualiza las estrategias manejadas por el proyecto SARA, la sección 4 muestra la metodología propuesta y cada una de las etapas del proceso KDD, en la sección 5 se detallan los resultados y la discusión, finalmente la sección 6 muestra las conclusiones y trabajos futuros del artículo.

2. Trabajos relacionados

El proceso KDD en la educación no es un término nuevo y su estudio y aplicación ha sido muy importante y relevante en los últimos años, el uso de este proceso permite analizar grandes volúmenes de

Como citar este artículo: Calvache-Fernández, Proceso KDD como apoyo a las estrategias del proyecto SARA (Sistema de Acompañamiento para el Rendimiento Académico). Educación en Ingeniería, 13(26), pp. 82-89, Julio, 2018.

datos encontrando relaciones y patrones no triviales [5].

Tanto a nivel Nacional e Internacional, muchas Universidades han desarrollado proyectos de investigación con respecto a la deserción estudiantil guiados en la aplicación de KDD. En Colombia, Universidades como la Universidad de Nariño y la Institución Universitaria CESMAG llevaron a cabo un proyecto de investigación el cual tenía como objetivo detectar patrones de deserción estudiantil en los programas de pregrado de estas dos Universidades, se aplicaron diferentes tareas y técnicas de minería de datos obteniendo perfiles socioeconómicos y académicos. Este proyecto unificó la información de las dos Universidades encontrando un patrón general que contiene notas, materias perdidas y resultados de las Pruebas Saber 11, permitiendo soportar la toma de decisiones y motivar estrategias en los programas de retención estudiantil que actualmente se encuentran establecidos dentro de las Universidades. Los perfiles arrojados por medio de técnicas de Minería de datos indican que se generaron modelos consistentes con la realidad observada [6].

En Argentina, la Universidad Gastón Dachary realizó una investigación para analizar el fenómeno de la deserción estudiantil, que tuvo en cuenta la información personal y los antecedentes académicos para identificar factores que influyen en la deserción de los estudiantes de la carrera de Ingeniería informática, se utilizaron múltiples algoritmos de clasificación como J48, BayesNet y OneR. Se identificó que las variables más influyentes son las asignaturas aprobadas, procedencia y edad de ingreso. El mayor porcentaje de deserción según los resultados se da en el primer año de la carrera [7].

Un grupo de estudiantes de la Universidad de Misiones en Argentina enfocaron su tesis en la utilización de técnicas de minería de datos para clasificar y agrupar a los estudiantes de acuerdo a sus características académicas, factores sociales y demográficos, todo esto con el objetivo de reducir el porcentaje de deserción en los programas académicos de esta Universidad. Dentro de esta tesis se informa que los mejores resultados en cuanto al análisis de la deserción se obtuvieron bajo la técnica de árboles de decisión y han planeado la posibilidad de contemplar más variables socioeconómicas [8].

Otro trabajo relacionado con esta temática se llevó acabo en la Universidad de Oklahoma en los Estados Unidos, el principal objetivo de esta investigación era predecir estudiantes en riesgo de deserción e intervenir de forma apropiada, este estudio utilizó datos recopilados durante cinco años al igual que una variedad de técnicas de minería de datos para caracterizar el perfil de deserción. Dentro de los resultados se pudo analizar que el conjunto de datos equilibrado obtiene mejores resultados que los no equilibrados; además, las variables educativas y financieras se encuentran como los factores más importantes dentro de este fenómeno [9].

Finalmente, la Universidad de Purvanchal de la India, realizó una investigación para obtener un modelo predictivo y generar una lista que incluya aquellos estudiantes que podrían necesitar de apoyo académico, el modelo predictivo se realizó por medio un proceso KDD. Los patrones de deserción resultantes permitieron realizar una comparación entre los múltiples algoritmos de árboles de decisión, entre los cuales están ID3, C4.5, CART y ADT, el mejor resultado se obtuvo con el algoritmo ID3 con el 90.9091% de instancias correctas. Los atributos más relevantes de la investigación fueron los ingresos y la ocupación de la madre [10].

Los trabajos mencionados son un ejemplo de las numerosas investigaciones donde se implementó un proceso KDD para buscar soluciones o estrategias encaminadas en la retención estudiantil.

3. Proyecto SARA

El Sistema de Acompañamiento para el Rendimiento Académico SARA, es un proyecto constituido a mediados del año 2014, se crea con el propósito de ayudar y acompañar académicamente a los estudiantes del Programa de Ingeniería de Sistemas y Computación de la Universidad del Quindío, en los primeros semestres o aquellos que tiene bajo rendimiento académico, brindando apoyo en temas o en áreas que presentan mayor dificultad de aprendizaje [11].

SARA ha venido consolidando un modelo de estrategias de acompañamiento estudiantil permanente que incentiva un mejor desarrollo personal, académico y vocacional de la mano de Bienestar Institucional, buscando incrementar la motivación y la posibilidad de permanencia de los estudiantes que ingresan al programa de Ingeniería de Sistemas y Computación. Se estudian y se ponen en práctica una variedad de estrategias que buscan no solo mejorar el nivel educativo de los estudiantes, sino que también se brindan herramientas que permitan afrontar de una manera adecuada sus estudios.

Entre las estrategias manejadas por el proyecto SARA se encuentran la inducción a estudiantes nuevos, inducción a biblioteca y bases de datos virtuales, se realiza un seguimiento continuo de las Pruebas BADyG (Batería de Aptitudes Diferenciales y Generales), se coordinan asesorías con docentes y estudiantes, finalmente se realiza un seguimiento y acompañamiento a estudiantes en situación condicional, este tipo de condición permite dar continuidad a los estudiantes que han demostrado algunas deficiencias como la perdida consecutiva de espacios académicos

4. Metodología

Tomando como metodología el Proceso de descubrimiento de conocimiento en base de datos o KDD, se inicia con un estudio de la problemática desde un enfoque holístico partiendo desde una revisión literaria de la deserción estudiantil en diferentes IES y finalizando concretamente en el Programa de Ingeniería de Sistemas y Computación de la Universidad del Quindío. Se seleccionaron de las bases de datos de la Universidad del Quindío los datos que relacionan características personales, socioeconómicas y académicas de los estudiantes. Con los datos recopilados para el análisis se construyó un repositorio que fue procesado y transformado con el propósito de obtener un conjunto de datos depurado y listo para la aplicación de algoritmos de Minería de Datos. Se descubrieron reglas y perfiles personales, socioeconómicos y académicos de los estudiantes utilizando la técnica de Árboles de decisión y Orange como herramienta de implementación. Los patrones resultantes fueron interpretados, evaluados, y finalmente usados para soportar la toma de decisiones en el proyecto SARA, permitiendo fortalecer la retención estudiantil en el programa académico. La Fig. 1 detalla el proceso metodológico para el desarrollo del trabajo.

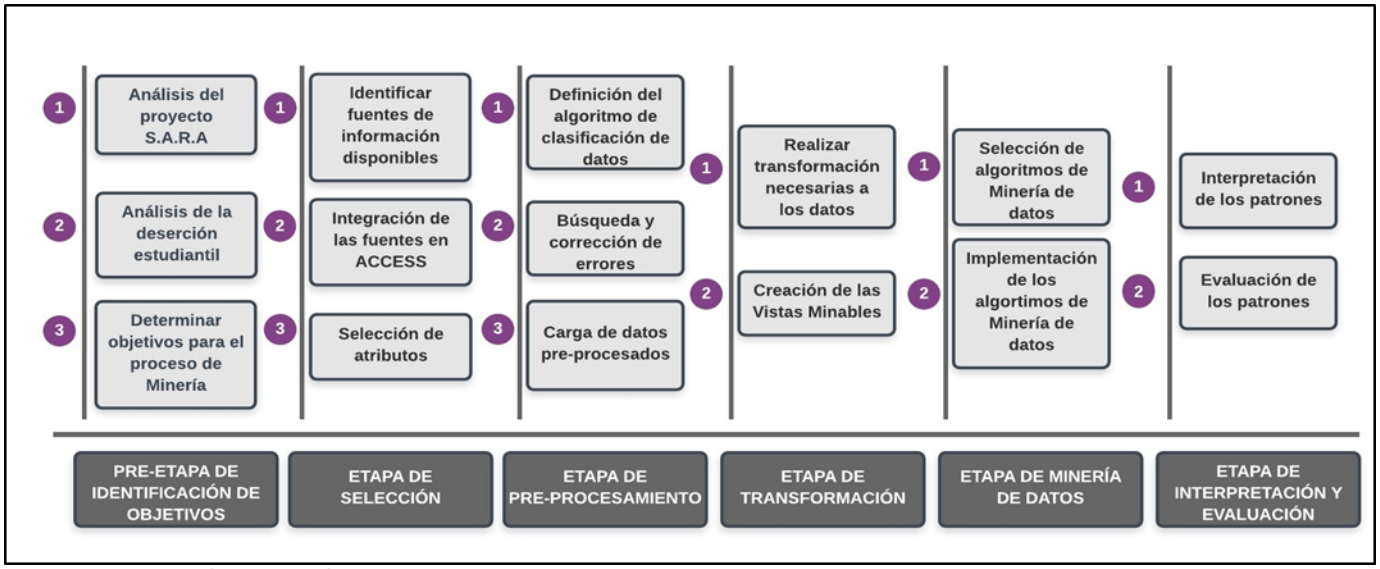

Figura 1. Implementación Metodología proceso KDD.
Fuente: Los autores

4.1. Pre - etapa de identificación de objetivos

Según [12] esta pre-etapa consiste en el conocimiento y la identificación del objetivo KDD desde el punto de vista del cliente. Esta etapa es de vital importancia ya que se debe de conocer previamente los objetivos, procesos y actividades de la organización donde se desea intervenir, dado que tener un conocimiento previo permitirá encaminar el proceso de minería de manera que se pueda obtener resultados óptimos y de calidad, se decidió como primera actividad realizar un análisis del proyecto SARA y de cada una de sus estrategias manejadas como se muestra en la sección 3, seguidamente se incorpora a este trabajo un análisis de la deserción para identificar variables y su comportamiento en relación a este fenómeno. El análisis de la deserción se realizó de forma Dinámica y por Cohorte; la primera hace referencia a la medición del fenómeno durante un espacio de cinco años que comprenden los periodos 2013-1 hasta 2017-2, sin tomar ninguna cohorte en especial, sino que se tomó la población total, es decir, los estudiantes que registran matrícula académica durante este lapso de tiempo. En cuanto al análisis de la deserción por cohorte, se toma la definición del SPADIES que hace referencia al Porcentaje acumulado de estudiantes de una cohorte que no ha registrado matricula por dos o más períodos consecutivos en un programa académico.

Se considera que esta pre-etapa es crucial dentro del proceso, ya que dependiendo de los objetivos definidos en esta instancia se determinó los datos que han de usarse dentro del proceso de extracción de conocimiento, así mismo, si los datos con los que se dispone permiten alcanzar o no los objetivos previstos. Para este trabajo es importante analizar e identificar los factores de deserción que más relevancia tienen tanto en la fase de inscripción como en el transcurso de la vida académica de los estudiantes. La fase de inscripción que hace referencia al proceso de admisión que el estudiante realiza para ingresar al programa académico, esta fase recopila información personal y socioeconómica del estudiante. Para el transcurso de la vida académica se tiene como referente el desarrollo de la vida universitaria, donde se analizó su desempeño académico que involucra resultados de parciales, registro de asistencia en asesorías e información de estudiantes que se encuentran en situación condicional.

Tabla 1.
Objetivos de Minería de Datos

Objetivos de Minería de Datos
Objetivo 1. Identificar características comunes en el ámbito personal y socioeconómico que permitan predecir si un estudiante que va a ingresar al programa está en riesgo o no de desertar.
Objetivo 2. Determinar patrones de los estudiantes con referencia a resultados de parciales, acuerdos y monitorías, para establecer el comportamiento de dichas variables en el fenómeno de la deserción.

Fuente: Los autores

A continuación, se detalla cada uno de los objetivos definidos para el proceso de minería en concreto. Ver Tabla 1.

4.2. Etapa de selección

El objetivo de esta etapa es determinar las fuentes de datos y el tipo de información a utilizar, es la etapa donde los datos relevantes para el análisis son extraídos desde las fuentes de datos, para el caso de estudio se seleccionaron las bases de datos Matriculados Totales, Colegios, Encuesta, y Primíparos de la Oficina de Planeación y Desarrollo. Adicionalmente las bases de datos Resultados de Parciales, Acuerdos y Monitorías suministradas por el proyecto SARA.

De las fuentes de datos Matriculados Totales, Colegios, Encuesta, y Primíparos se seleccionó la población total entre 2013-1 y 2017-2, es decir, todos los estudiantes que han registrado matricula académica entre estos periodos inclusive, sin tomar una Cohorte en específico, entendiéndose esto como Deserción Dinámica, se obtuvo como resultado un repositorio con información personal y socioeconómica de los estudiantes con un total de 1644 registros y 23 atributos, el cual será identificado como REP01 a lo largo de este documento.

Tabla 2.
Descripción de los atributos del repositorio: REP01

N°	Atributo	Descripción
1	Edad	Edad del estudiante
2	Jornada	Jornada del programa al que pertenece el estudiante (Diurna/ Nocturna)
3	Ciudad _Residencia	Ciudad de residencia del estudiante
4	Genero	Género del estudiante (Femenino/Masculino)
5	Estado_Civil	Estado civil del estudiante (Solter@, Casad@, Viudo, Unión Libre, Madre Soltera y Separad@)
6	Grupo_Etnico	Grupo étnico al que pertenece el estudiante si aplica (Negritudes, Pastos, Achagua, Awa, Embera, Pastos, Pijaos, Wayuu, Yanacona)
7	Victima_Conflicto	Si el estudiante es víctima del conflicto armado (Si/No)
8	Es desplazado	Si el estudiante es desplazado (Si/ No)
9	Es Discapacitado	Si el estudiante es discapacitado (Si/No)
10	Nombre_Discapacidad	Nombre de la discapacidad del estudiante si aplica
11	Estrato_Socioeconómico	Estrato socioeconómico al que pertenece el estudiante
12	Régimen_salud	Régimen de salud del estudiante (Contributo o subsidiado)
13	Categoría_Sisbén	Categoría del Sisbén del estudiante
14	Cant_grupo_familiar	Número de personas por el que está conformado la familia del estudiante
15	Nucleo_familiar	Núcleo por el que está conformado su familia (Alguno de sus padres, Dos padres y hermanos, Hermanos o familiares, Solo, Solo sus dos padres, Sus hijos, Espos@ o compañer@, Sus hijos)
16	Aporta_Economicamente	Si aporta económicamente a la familia (Si/No)
17	Cantidad_Ingresos	Cantidad de ingresos del estudiante en salarios mínimos
18	Fuente_financiación	Fuente por la cual financia sus estudiaos (Beca, Padres, Ingresos Personales, Entidades Financieras, Icetex, Espos@ o compañer@)
19	Labora	Si el estudiante labora o no (Si /No)
20	Nombre_Institucion	Nombre de la institución donde termino sus estudios de bachiller
21	Inst_Naturaleza	Naturaleza de la institución donde termino sus estudios de bachiller (Publica / Privada)
22	Municipio_Institucion	Municipio donde está ubicada la institución donde termino sus estudios de bachiller
23	Departamente_Institucion	Departamento donde está ubicada la institución donde termino sus estudios de bachiller

Fuente: Los autores

De las fuentes de datos Resultados de Parciales, Acuerdos y Monitorias se seleccionó la Cohorte del 2016-1, entendiéndose esto como Deserción por Cohorte, teniendo como resultado un repositorio con información académica de los estudiantes, con un total de 143 registros y 14 atributos, el cual será identificado como REP02 a lo largo de este documento. Finalmente, las fuentes de datos Graduados y Estado Estudiante suministradas por la Oficina de Calidad, permitió descartar estudiantes que no hacen parte del análisis de la deserción estudiantil como: estudiantes graduados, en continuidad académica, retirados por motivos disciplinarios y estudiantes de intercambio.

En la Tabla 2 se detallan los atributos que se seleccionaron para el análisis en referencia al repositorio REP01:

En la Tabla 3 se detallan los atributos que se seleccionaron para el análisis en referencia al repositorio REP02:

Como resultado se obtuvo un repositorio en Microsoft Access que integró la información de REP01 y REP02. Como actividad final de esta etapa se realizó un reconocimiento por tabla y gráfico que permitió identificar el estado en el que se encuentran las fuentes de datos incorporadas en el estudio, ver Fig. 2. El reconocimiento incluye una descripción de las variables de cada fuente, formato y tipo de dato, moda y media, cantidad de valores nulos y la cantidad de valores diferentes que puede tomar el atributo, todo esto con el fin de identificar el estado de cada uno de los repositorios y usar este análisis para la posterior etapa.

4.3. Etapa de pre-procesamiento

Esta etapa consiste en un análisis intensivo del conjunto de datos seleccionado en la etapa anterior, donde se ponen en práctica operaciones y técnicas necesarias para eliminar ruido, inconsistencias o redundancias con las que puedan venir los datos. El principal objetivo de esta etapa es preparar los datos seleccionados para la posterior aplicación de los algoritmos. Como primera actividad se implementó un algoritmo desarrollado bajo el lenguaje Java, dicho algoritmo permitió automatizar el proceso de clasificación (Desertor y No Desertor).

Tabla 3.
Descripción de los atributos del repositorio: REP02

N°	Atributo	Descripción
1	Acuerdo	0 si la persona no ha estado en acuerdo y 1 en caso contrario
2	P1POO	Resultado o nota del primer parcial de Paradigma Orientada a Objetos
3	P2POO	Resultado o nota del segundo parcial de Paradigma Orientada a Objetos
4	P3POO	Resultado o nota del tercer parcial de Paradigma Orientada a Objetos
5	P1CALCULO_DIF	Resultado o nota del primer parcial de Cálculo Diferencial
6	P2CALCULO_DIF	Resultado o nota del segundo parcial de Cálculo Diferencial
7	P3CALCULO_DIF	Resultado o nota del tercer parcial de Cálculo Diferencial
8	P4CALCULO_DIF	Resultado o nota del cuarto parcial de Cálculo Diferencial
9	P1GEOMATRIA	Resultado o nota del primer parcial de Geometría Analítica
10	P2GEOMATRIA	Resultado o nota del segundo parcial de Geometría Analítica
11	P3GEOMATRIA	Resultado o nota del tercer parcial de Geometría Analítica
12	P4GEOMATRIA	Resultado o nota del cuarto parcial de Geometría Analítica
13	CALCULO	0 si no ha asistido a asesorías en el Área de Cálculo y 1 en el caso contrario
14	PROGRAMACION	0 si no ha asistido a asesorías en el Área de programación y 1 en el caso contrario

Fuente: Los autores

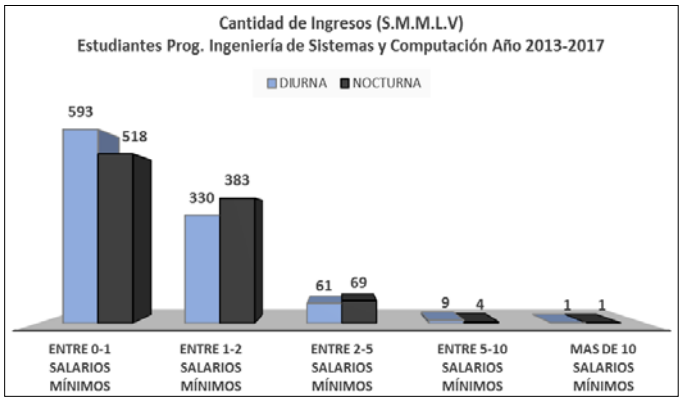

Figura 2. Cantidad de Ingresos (S.M.M.L.V). Prog. Ingeniería de Sistemas y Computación. Año 2013-2017.
Fuente: Cálculo de los autores

Por medio del reconocimiento realizado en la etapa anterior identificaron aquellas inconsistencias o errores con los que venían las fuentes permitiendo analizar de esta manera la calidad de los datos. Teniendo en cuenta la importancia de los atributos seleccionados, los valores nulos encontrados fueron actualizados con la moda o media dependiendo de la naturaleza del atributo, y así las inconsistencias encontradas fueron corregidas con el valor más aproximado al valor real.

4.4. Etapa de transformación

La etapa de transformación de los datos es la etapa que proporciona la vista minable, es decir, los datos listos para ser analizados por los algoritmos de minería. La transformación de los datos es la construcción de atributos, la cual consiste en construir automáticamente nuevos atributos aplicando alguna operación o función a los atributos originales con el objeto de que estos nuevos atributos hagan más fácil el proceso de minería [13]. Para facilitar la extracción de patrones se discretizaron los valores numéricos a valor nominales, como ejemplo se tiene la edad del estudiante, otros atributos como el régimen de seguridad, nombre de discapacidad, estado civil entre otros se numerización para favorecer el proceso de minería.

Tabla 4.
Rangos de discretización de la edad del Estudiante

Edad	Discretización
Edad >=15 & Edad <=20	1
Edad >=21 & Edad <=23	2
Edad >=24 & Edad <=28	3
Edad >=29 & Edad <=56	4

Fuente: Los autores

Tabla 5.
Selección del algoritmo de minería de datos

Objetivos	Tarea	Algoritmo
1. Identificar características comunes en el ámbito personal y socioeconómico que permitan predecir si un estudiante que va a ingresar al programa está en riesgo o no de desertar.	Predecir un valor discreto.	Algoritmo de árboles de decisión
2. Determinar patrones de los estudiantes con referencia a resultados de parciales, acuerdos y monitorias, para establecer el comportamiento de dichas variables en el fenómeno de la deserción.	Predecir un valor discreto.	Algoritmo de árboles de decisión

Fuente: Los autores

Para discretizar la edad se tomó en cuenta la discretización simple o simple binning mencionada el libro (Hernández et al., 2004). La discretización más sencilla también llamada simple binning es aquella discretización que realiza intervalos del mismo tamaño y utiliza el mínimo y máximo valor como referente, para ello se resta el máximo y el mínimo valor, el valor resultante de la resta se divide por la cantidad de intervalos deseados, puede resultar más adecuado aquellos intervalos que tienen un número constante de individuos o aquella discretización que mantiene una distribución similar a la normal. Para el caso de estudio se tomó el mínimo valor de la edad (15 años) y máximo valor (56 años) como referente, creando cuatro rangos que mantengan

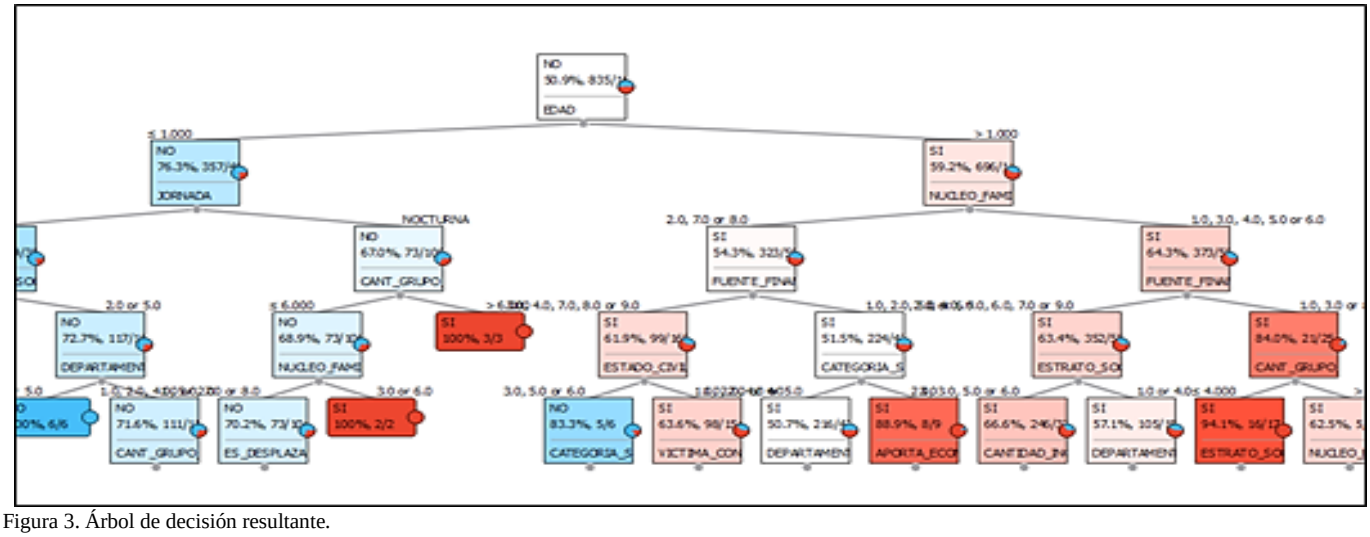

Figura 3. Árbol de decisión resultante.
Fuente: Los autores

la proporción de los datos, puede notarse que los intervalos no poseen un tamaño fijo ya que era necesario asegurar una distribución proporcional de los individuos.

Una vez realizada cada una de las trasformaciones requeridas, la última actividad de esta etapa es la creación de las vistas minables, dichas vistas fueron usadas para la aplicación de los algoritmos de minería de datos.

4.5. Etapa de minería de datos

Minería de Datos es la siguiente etapa del proceso KDD, cuyo objetivo es producir conocimiento, es decir, encontrar relaciones entre los datos que no han sido detectadas a simple vista, para ser usado por el cliente o la organización como apoyo a sus decisiones. Es importante mencionar que para la aplicación de la

Tabla 6.
Interpretación de reglas Objetivo 1

Regla	Interpretación	Clase	Probabilidad
1	Estudiantes cuya edad es mayor de 20 años Que no vivan con alguno de los siguientes miembros: ➢ Dos padres y hermanos, ➢ Solo sus dos padres ➢ Sus hijos Que sus gastos no sean financiados por: ➢ Becas de la Universidad ➢ Entidades financieras ➢ Usted y su esposa Que su estrato Socioeconómico sea diferente de 1 y 4 Que la cantidad de ingresos sea alguno de los siguientes: ➢ Entre 2 y 5 (S.M.M.L.V) ➢ Entre 5 y 10 (S.M.M.L.V) ➢ Mayores a 10 (S.M.M.L.V) Que su categoría del Sisbén sea; 3,4 o 5 Que la cantidad de personas que conformen su núcleo sea mayor a 3 personas.	Deserta	100%
2	Estudiantes cuya edad es mayor de 20 años Que no vivan con algunos de los siguientes miembros: ➢ Dos padres y hermanos, ➢ Solo sus dos padres ➢ Sus hijos Que sus gastos sean financiados por: ➢ Becas de la Universidad, ➢ Entidades financieras ➢ Usted y su esposa Que la cantidad de integrantes de su familia sea mayor e igual a 4 personas.	Deserta	94.1%
3	Estudiantes cuya edad sea mayor a 20 años Que su núcleo familiar este conformado por alguno de los siguientes miembros: ➢ Dos padres y hermanos ➢ Solo sus dos padres ➢ Sus hijos Que sus gastos sean financiados por: ➢ Becas de la Universidad ➢ Becas Externas ➢ Icetex ➢ Ingreso de padres y/o familia Que su categoría del Sisbén sea igual a 3	Deserta	83.3%
4	Estudiantes cuya edad es menor e igual a 20 años Que pertenezcan a la jornada diurna Que su estrato socioeconómico sea diferente de 2 y 5 Que su grupo familiar este compuesto por menos de 7 personas.	No Deserta	83.6%
5	Estudiantes cuya edad sea mayor a 20 años Que su núcleo familiar este compuesto por alguno de los siguientes miembros: ➢ Dos padres y hermanos ➢ Solo sus dos padres ➢ Sus hijos Que su fuente de financiación sea diferente de: ➢ Becas de la Universidad ➢ Becas Externas ➢ Icetex ➢ Ingreso de padres y/o familia Que su estado civil sea alguno de los siguientes estados: ➢ Separado ➢ Unión Libre ➢ Viudo(a)	No Deserta	83.3%

Fuente: Los autores

Tabla 7.
Interpretación de reglas Objetivo 2

Regla	Interpretación	Clase	Probabilidad
1	Aquellos estudiantes que al menos presenten el tercer parcial del área de Paradigma Orientada a Objetos, además una nota mayor o igual a 3,4 en el último parcial de Geometría Analítica.	No Deserta	100%
2	Aquellos estudiantes que presenten el tercer parcial del área de Paradigma Orientada a Objetos. Que la nota del cuarto parcial de Geometría sea menor o igual a 3,4. Que sus notas en el tercer parcial de Cálculo Diferencial sea mayor o igual 2,7. El tercer parcial de Geometría sea mayor o igual a 2,7. Que el segundo parcial de Paradigma Orientada a Objetos sea mayor a 3,4. Y que asista a las monitorías del área de Cálculo en los siguientes temas: ➢ Circunferencias y parábolas ➢ Derivadas implícitas ➢ Derivadas.	No Deserta	100%
3	Aquellos estudiantes que no asisten al tercer parcial del área de Paradigma Orientada a Objetos.	Deserta	92.3%
4	Aquellos estudiantes que no presenten el tercer parcial del área de Paradigma Orienta a Objetos, además que la nota del segundo parcial del área de Cálculo Diferencial sea menor o igual a 3,4 o simplemente no lo presente.	Deserta	98%

Fuente: Los autores

etapa de Minería de datos se debe de tener en cuenta la tarea apropiada para identificar factores influyentes en la deserción, la tarea de minería de datos aplicada para el caso de estudio es la Clasificación, una vez identificada la tarea se eligió la técnica de minería de datos que resuelva esta necesidad, la técnica usada para los objetivos propuestos es Árboles de Decisión, ver Tabla 5.

Los árboles de decisión fueron implementados en la herramienta Orange, las reglas más relevantes se muestran en la sección 5 de Resultados las cuales superan una probabilidad de ocurrencia del 80%.

4.6. Etapa de interpretación e evaluación

Continuando con las etapas del proceso KDD se procede a la siguiente etapa denominada Interpretación e Evaluación, en este punto se verifica la calidad de los patrones resultantes. Idealmente, los patrones descubiertos deben de tener tres cualidades: ser precisos, comprensibles e interesantes, es decir, útiles y novedosos

La evaluación de los modelos se realizó por medio de componentes brindados por Orange, obteniendo una precisión del 70.2% y 80% para el objetivo 1 y 2 respectivamente. Adicionalmente se descartaron los patrones irrelevantes, dejando únicamente los de mayor ocurrencia. Finalmente se tradujo cada uno de los patrones en términos entendibles para el proyecto SARA. Como se puede visualizar en la Fig. 3 los resultados son presentados en forma de árbol n-arios en donde a partir de un nodo padre o raíz se desprenden n cantidad de hijos.

Cabe resaltar que se maneja una escala e intensidad de colores donde cada color representa los posibles valores que puede tomar la variable, para este caso la variable clase (Desertor o No desertor).

5. Resultados y discusión

Como resultado de interpretar los Árboles de decisión generados por Orange a partir de los repositorios *REP01* y *REP02*, se obtuvieron las reglas de clasificación más relevantes con una probabilidad mínima del 80% como se muestra en la Tabla 6 para el Objetivo 1 y en la Tabla 6 para el Objetivo 2.

Como se puede observar en la Tabla 6 los factores de deserción más relevantes en cuanto al ámbito personal y socioeconómico hacen referencia al núcleo familiar, la cantidad de integrantes que lo conforman y la fuente de financiación. Particularmente la probabilidad de deserción es más alta cuando no se vive con; Dos padres y hermanos, Solo sus dos padres y Sus hijos, o cuando su fuente de financiación involucra Becas de la Universidad, Entidades financieras o Usted y su esposa, adicionalmente que el grupo familiar este compuesto por más de 3 personas.

En referencia a la Tabla 7, en aspectos académicos los factores de más relevancia tienen que ver con un promedio de notas bajo, particularmente no presentar los parciales del área de Paradigma Orientada a Objetos se ha convertido en el factor de mayor impacto. Adicionalmente cabe resaltar que uno de los factores que garantiza mayor índice de supervivencia es la asistencia a asesorías, un estudiante que asiste al menos a una asesoría de programación garantiza su supervivencia durante su carrera.

6. Conclusiones y trabajo futuros

Para la ejecución de este trabajo se aplicaron una serie de conocimientos que se adquirieron a lo largo de nuestra vida académica, los conocimientos plasmados en este documento abarcan conceptos de bases de datos, Inteligencia de Negocios y Minería de datos, siendo esta ultima el área la de mayor aplicabilidad.

Finalizando cada una de las etapas de proceso KDD y habiendo cumplido con todos los objetivos en su totalidad, puede concluirse de este trabajo lo siguiente:

- ✓ La implementación del proceso KDD basado en minería de datos permite encontrar conocimiento útil y novedoso que no es perceptible a simple vista. La Minería de datos para el contexto de la deserción estudiantil permite al programa y especialmente al proyecto S.A.R.A, conocer cuales variables deberán tenerse en cuanta con mayor prioridad para intervenir en la retención de los estudiantes.

- ✓ Al arrojar patrones, un proceso de Minería de datos permite el manejo eficiente de la información y soportar la toma decisiones dentro de una organización o empresa, brindando la posibilidad de mejorar aspectos o procesos de su entorno. Al detectar información relevante y no trivial, un proceso de minería permitirá dar un valor diferenciador para la organización.
- ✓ La etapa de pre-procesamiento es la etapa que consume gran parte del tiempo, requiere de un análisis detallado ya que asegurar la calidad de los datos nunca será una tarea fácil.
- ✓ El conocimiento producto de los patrones permitirá generar e incluir estrategias alineadas con la visión y los objetivos que persigue SARA, las cuales podrían estar encaminadas en la intervención de la vida académica de los estudiantes que se perfilan como potenciales desertores, algunas de estas estrategias podrían estar relacionadas con analizar la distribución horaria de asesorías para mejorar ambientes de estudio y crear estrategias motivacionales para que los estudiantes se incorporen con mayor frecuencia en ellas.
- ✓ En cuanto a la aplicación de técnicas de minería de datos en el ámbito educativo, muchos de los patrones encontrados en este trabajo se pueden convertir en un punto de partida para motivar estrategias tempranas de retención estudiantil, relacionadas con la flexibilidad horaria, asesorías académicas, ayudas psicológicas y acompañamiento académico que permitan a los estudiantes continuar con su ciclo educativo.

En cuanto a trabajos futuros se sugiere que:

- ✓ Se establezcan revisiones de las fuentes de datos, principalmente cuando se analiza las Cohortes que serán objeto de estudio, ya que por medio del reconocimiento realizado en la etapa de Selección se pudo identificar que existen estudiantes que aparecen como nuevos en distintas Cohortes, lo cual implica una alteración en los índices de Deserción estudiantil, se recomienda entonces un análisis a mayor profundidad de las fuentes con las que se está trabajando a nivel Institucional
- ✓ El surgimiento de nuevas tecnologías ha hecho que el mundo actual genere datos de forma masiva, las bases de datos no relacionales por su alta escalabilidad han permitido adaptarse de forma adecuada a un crecimiento continuo de datos; las redes sociales, aplicaciones, páginas web, entre otras, son ejemplo de tecnologías que usan este tipo de bases de datos no estructurada. En futuros trabajos se piensa en incluir redes sociales con el propósito de conocer la opinión de los estudiantes del Programa de Ingeniería de Sistemas y Computación acerca de las estrategias enmarcadas por SARA, de esta manera se contará con un elemento crucial que permitirá conocer y evaluar la efectividad de las estrategias e incluir en las estrategias las recomendaciones más apropiadas.
- ✓ La aplicación de nuevas técnicas de minería de datos como Clúster o Red neuronal se pueden convertir en una alternativa para descubrir segmentos de poblaciones o relaciones que no se pudieron encontrar mediante la aplicación de Árboles de decisión.

Referencias

[1] Guzmán-Ruiz, C., Muriel-Durán, D. y Franco-Gallego, J., Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención, Ministerio de Educación Nacional, 2009.

[2] Cepero-González, A., Las preferencias profesionales y vocacionales del alumnado de secundaria y formación profesional específica, Tesis, Facultad de Ciencias de la Educación, Universidad de Granada, Granada, España, 2010.

[3] Universidad Pedagógica Nacional, La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN, 10 P, 2008.

[4] Ministerior de Educación, SPADIES, Sistema para la Prevención de la Deserción de la Educación Superior, 2002. [en línea]. Disponible en: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254707.html.

[5] Pereira, R.T., Romero, A.C. y Toledo, J.J., La minería de datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en instituciones de educación superior, ACOFI, IFEES, 9 P, 2013.

[6] Timaran, R., Jimenez, J., Pereira, R.T. y Toledo, J.J., Detección de patrones de deserción estudiantil en programas de pregrado de instituciones de educación superior con CRISP-DM, en Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación, Buenos Aires, Argentina, 2014, pp. 1-19.

[7] Eckert, K.B. y Suénaga, R., Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos, Form. Univ., 8(5), pp. 3-12, 2015, DOI: 10.4067/S0718-50062015000500002.

[8] Pautsh, J.G.A., Minería de datos aplicada al análisis de la deserción en la carrera de analista en sistemas de computación, Tesis de grado, Facultad de Ciencias de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones, Posadas, Argentina, 2009, 111 P.

[9] Delen, D., A comparative analysis of machine learning techniques for student retention management, Decis. Support Syst., 49(4), pp. 498-506, 2010. DOI: 10.1016/j.dss.2010.06.003.

[10] Head, S.P., Mining educational data to reduce dropout rates of engineering students, Inf. Eng. Electron. Bus., 2(2), pp. 1-7, 2012. DOI: 10.5815/ijieeb.2012.02.01.

[11] Quiceno, C. and Pulgarón, R., S.A.R.A Sistema de acompañamiento para el rendimiento académico, ACOFI, 2016, 8 P.

[12] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., From data mining to knowledge discovery in databases, AI Mag., pp. 37-54, 1996, DOI: 10.1609/aimag.v17i3.1230.

[13] Hernández, J., Ramirez, M. and Ferri, C., Introducción a la minería de datos. España, 2004.

L.C. Calvache-Fernández, recibió el título de Ing. Sistemas en la Universidad del Quindío en el año 2018, Armenia, Colombia. Desde el 2016 trabajó como auxiliar de bases de datos de la vicerrectoría de investigaciones de esta misma Universidad hasta diciembre de 2017. Actualmente labora como analista de inteligencia comercial en Avianca Holdings.
ORCID: 0000-0003-1726-1023

V. Álvarez-Vallejo, recibió título de Ing. de Sistemas y Computación en la Universidad del Quindío, Armenia, Colombia. Desde el 2016 al 2017 trabajó como auxiliar de labor de asesoría, vigilancia, supervisión y control del Laboratorio de Ingeniería de Sistemas y Computación de esta misma Universidad. Actualmente es la administradora COAVI del Grupo Empresarial Don Pollo.
ORCID: 0000-0002-4325-4401.

J.I. Triviño-Arbeláez, recibió el título de Ing. Sistemas en la Universidad del Quindío en el año 2004, Armenia, Colombia. MSc. en Ingeniería con la Universidad de Eafit en el año 2016. Cursos de corta duración en la Universidad La Gran Colombia - Seccional Armenia – UGCA Diplomado en Docencia Universitaria en 2007. Académicamente ha laborado en la Escuela de Administración y Mercadotecnia del Quindío, Fundación Universitaria San Martin y en la Universidad del Quindío desde 2007. Es coautor de los libros: Fundamentos de bases de datos (2009) y Aprendiendo a Programar en Java.
ORCID: 0000-0002-1264-3519