UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica
http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

CIDC

Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos

Towards the construction of a predictive model dropout academic based data mining techniques

Rumo a construção de um modelo de previsão de deserção acadêmicos com base técnicas de mineração de dados

Jonny Esteban Sotomonte Castro[1]
Cristian Camilo Rodríguez Rodríguez[2]
Carlos Enrique Montenegro Marín[3]
Paulo Alonso Gaona García[4]
John Gabriel Castellanos[5]

Resumen
Existe un problema latente en la educación de nivel superior en Colombia, el cual tiene que ver con los altos índices de deserción académica, adicionalmente son muy pocas las estrategias que se han implementado con el fin de frenar la tasa de deserción, puesto que solo hasta el año 2003, se inician de manera formal los estudios para poder establecer cuáles son las condiciones que propician el abandono de los estudios. Sin embargo, se desconocen las causas que conllevan a que un estudiante abandone su carrera, para ello en este artículo se hará uso de la Minería de Datos, por medio de la cual se pretende generar un modelo de Árbol de Decisión implementando el algoritmo J48 mediante el uso de la herramienta WEKA con el fin de poder identificar estas causas.
Palabras Clave:
Deserción Académica, Minería de Datos, Arboles de Decisión, Algoritmo J48, WEKA

Abstract
There is an imminent problem with the college education in Colombia, and this is related to the high percentage of drop outs,

[1] Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. Contacto: jesotomontec@correo.udistrital.edu.co
[2] Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. Contacto: crcrodriguezr@correo.udistrital.edu.co
[3] Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. Contacto: cemontenegrom@udistrital.edu.co
[4] Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. Contacto: pagaonag@udistrital.edu.co
Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. Contacto: Corio27@gmail.com

on addition there are very few implemented strategies to stop the high percentages of desertion among the college population, due to the fact that only in the year 2003 was formally initiated the studies related to determine which are those conditions that conduct to these drop outs.but nevertheless the reason why the students abandon there education is yet unknown, for it in this article use will be made of data mining, by which is intended to generate a Decision Tree model implementing the algorithm J48 using the tool WEKA in order to identify these causes.
Keywords:
Dropout, Data mining, Decision Trees, J48 algorithm, WEKA

Resumo
Há latente em educação de nível superior na Colômbia, que tem a ver com os elevados níveis de problema do abandono escolar, além disso, existem muito poucas estratégias que foram implementadas, a fim de reduzir a taxa de abandono, uma vez que apenas se 2003 estudos formalmente iniciadas estabelecer quais são as condições propícias para o abandono dos estudos são. No entanto, as causas que levam a um estudante deixa sua carreira, por isso vai fazer uso de mineração de dados neste artigo, através do qual se pretende criar uma árvore modelo de decisão que implementa o algoritmo J48 por desconhecidos a WEKA usando a ferramenta para ser capaz de identificar essas causas.

Palavras-chave:
Deserção Academic, Mineração de dados, Árvores de decisão, algoritmo J48, WEKA

INTRODUCCIÓN

En el entorno de la educación en Colombia se ha evidenciado en los últimos años un crecimiento significativo en los niveles de deserción por parte de los estudiantes de cursos de Educación Superior. (Guzmán & Duran & Franco & Castaño & Gallón & Guzmán & Gómez & Vásquez, 2009, Cuervo y Ballesteros, 2015).

Acerca del estudio de esta problemática se han realizado investigaciones por parte de las Universidades de Antioquia y de los Andes, tomando como referencia la información académica contenida en el Sistema de Información SPADIES que permite establecer diferentes categorías por las cuales se puede presentar la deserción estudiantil. (Guzmán & Duran & Franco & Castaño & Gallón & Guzmán & Gómez & Vásquez, 2009)

Es por esta razón que se quiere realizar un estudio particular para el caso de la Universidad Distrital Francisco José de Caldas con el fin de poder determinar cuáles son las causas que

particularmente en este caso, conllevan a que los estudiantes deserten de la Universidad, para ello se analizarán datos de estudiantes de la Facultad de Ingeniería entre los años 2009 y 2015, para analizar la información recolectada se hará uso de una de la técnica Minería de Datos.

Objetivos

Desarrollar un modelo predictivo mediante el análisis de datos históricos bajo las técnicas de minería de datos para poder determinar las principales causas de deserción de un estudiante.

Para poder cumplir con este objetivo es necesario cumplir con algunos objetivos específicos:

Realizar levantamiento de la información histórica de orden académico y personal de los estudiantes de educación superior, con el fin de generar datos de entrenamiento y datos de testeo para la generación y evaluación del modelo obtenido.

Depurar la información para determinar cuáles son los datos que para el caso de estudio tendrán la mejor calidad y serán más relevantes en el proceso de análisis y desarrollo del modelo (Camargo, et al 2015).

Analizar la información luego de haber sido depurado mediante la aplicación de la técnica de árboles de decisión utilizando como algoritmo base el J48.

MARCO TEÓRICO

Dentro del campo del análisis de datos existen diferentes técnicas que nos pueden ayudar a entender el comportamiento de los datos, dentro de ellas se puede destacar una de las técnicas más utilizadas la cual es conocida como la Minería de Datos, esta técnica consiste en la iteración de una serie de pasos los cuales son: Selección, Limpieza, Transformación, Minería de Datos, Evaluación y Representación del Modelo.

Existen trabajos de minería de datos que se han desarrollado específicamente con el propósito de analizar las causas que dan lugar al fenómeno de la deserción estudiantil en la educación superior en Colombia, en la Figura 1 se muestran estos trabajos.(Amaya et al., 2015; Estrada. 2015)

Figura 1: Tabla comparativo de técnicas de minería de datos,

CUADRO COMPARATIVO DE LAS TÉCNICAS DE MINERÍA DE DATOS Y MODELOS PREDICTIVOS

TABLA 1. TÉCNICAS DE MINERÍA DE DATOS UTILIZADAS EN ESTUDIOS SIMILARES.

N-	PAIS	ESTUDIO	TECNICAS UTILIZADAS
1.	Colombia		árboles de decisión C4.5
		Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. (Timarán P., 2009)	Asociación por medio del algoritmo EquipAsso (Basado en Operadores algebraicos)
2.	Colombia	Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos. (Timarán P., 2010)	TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios KDD del grupo de investigación GRIAS de la Universidad de Nariño.
3.	Colombia	Generación de un modelo predictivo para determinar el desempeño académico en la asignatura fundamentos de programación II del programa de Ingeniería de Sistemas. [4]	ID3
			NAÏVE-BAYES

Fuente: (Amaya et al., 2015)

Dentro de las técnicas predictivas se contemplan los modelos de regresión y la clasificación ad hoc, que a su vez está compuesta por los modelos de Logit, Probit, discriminante, arboles de decisión y redes neuronales, para este caso en particular se hará énfasis en las técnicas de árboles de decisión.(Pérez Marqués, 2014)

METODOLOGÍA

CRISP - DM

Se hará uso de una metodología de Minería de Datos conocida como CRISP – DM la cual consiste en un ciclo de vida iterativo que consta de seis fases donde dichas fases se relacionan entre ellas, En la Figura 2 se muestra el modelo de la metodología con sus fases y relaciones entre sí. (Bawden, 2005).

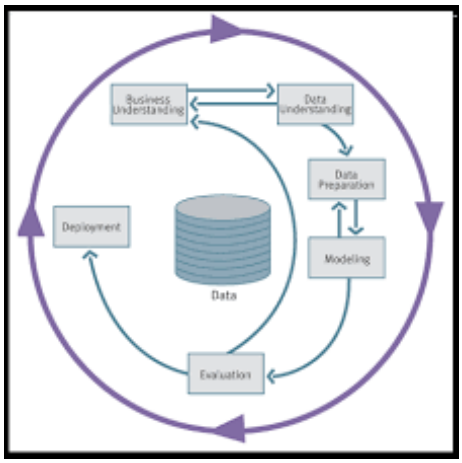

Figura 2. Fases Metodología CRISP - DM, Fuente: (“aníbal goicochea,” n.d.)

Dada la flexibilidad de la metodología es posible agrupar o segregar las fases de tal manera que se ajuste a las fases de la minería de datos tradicional, las cuales son: Fase de Selección, Fase de Limpieza, Fase de codificación o transformación, Fase de análisis, Fase de Interpretación y/o Evaluación y Fase de Despliegue.

RESULTADOS

Para el caso de estudio expuesto se tuvo como base, la Información académica de los estudiantes desde el año 2009 en distintos ámbitos, la cual fue recolectada de las siguientes fuentes de información:

- Estadísticas de deserción de los últimos años en áreas del saber.
- Zonificación y discriminación por Programa, área del Conocimiento, condiciones sociales.
- ICFES.
- Pensum.
- Acuerdo académico en el que se encuentran los estudiantes.

Teniendo esta información se procedió a centralizarla en una bodega de datos, en la cual se generó un datamart que se compone de una tabla de hechos relacionada con varias tablas llamadas, a continuación, se describen brevemente las dimensiones generadas:

Dimensión Plantel: Guarda toda la información relacionada con el colegio del cual proviene el estudiante. Ver Figura 3:

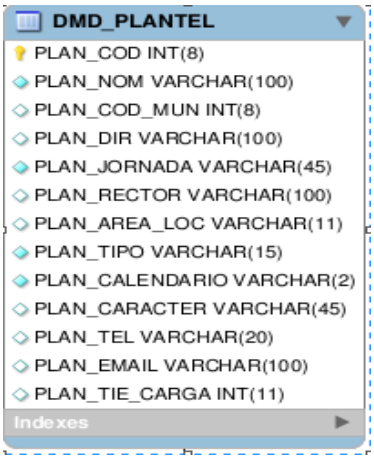

Figura 3. Dimensión Plantel, Fuente: Autor

Dimensión Tiempo: En esta dimensión se almacena la información relacionada con el tiempo que lleva el estudiante en la Universidad desde el momento que ingresó, Ver Figura 4:

Figura 4. Dimensión Tiempo, Fuente: Autor

Dimensión Asignatura: En esta dimensión se puede encontrar la información relacionada con las asignaturas que el estudiante ha cursado, Ver Figura 5:

Figura 5. Dimensión Asignatura, Fuente: Autor

Dimensión Carrera: En esta tabla contiene todo lo referente con el proyecto curricular al cual el estudiante está matriculado, Ver Figura 6:

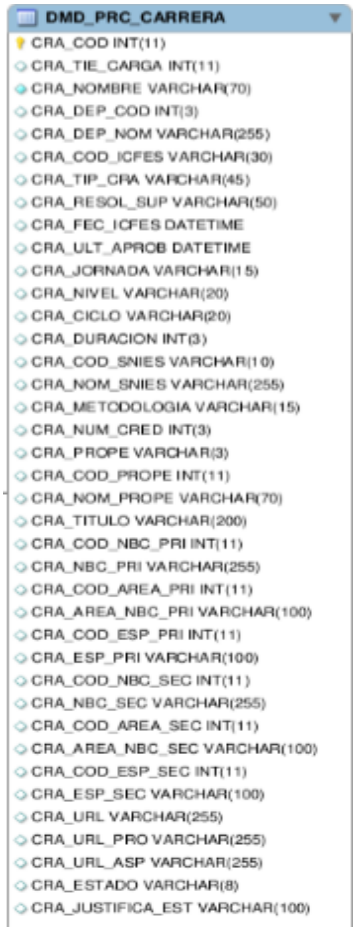

Figura 6. Dimensión Carrera, Fuente: Autor

Dimensión Lugar: En esta tabla fue registrado todo lo que se refiere al lugar geográfico en el cual habita el estudiante, Ver Figura 7.

Dimensión Estudiante: En esta dimensión se almacenó la información básica del alumno.

También se tiene registrada información adicional, dentro de la cual se destacan atributos referentes a:

Nivel de educación, Información Familiar, Información Socioeconómica, entre otras.

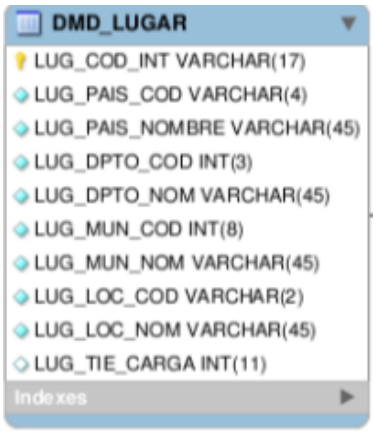

Figura 7. Dimensión Lugar, Fuente: Autor

Después de haber identificado cada una de las dimensiones que tenían relación con la información necesaria, se prosiguió a identificar las dimensiones que tenían relaciones entre sí, de donde se pudo concluir que las dimensiones Plantel y Lugar tenían una relación directa con la dimensión Estudiante, dicha relación se muestra en la Figura 8:

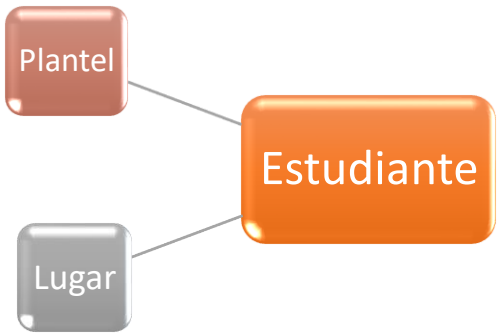

Figura 8. Relación entre dimensiones (Plantel-Lugar) y estudiante, Fuente: Autor

Después de tener las dimensiones y sus relaciones, se registró toda aquella información contenida en las diferentes dimensiones de manera centralizada en la tabla de hechos la cual fue llamada matricula y de esta manera obtener todo el diseño del datamart de la bodega de datos, Ver Figura 9.

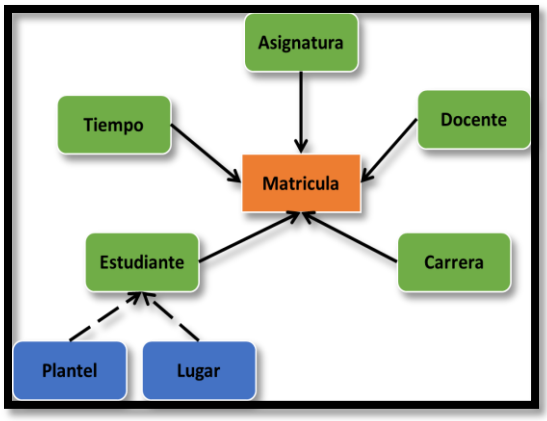

Figura 9. Diseño Bodega de Datos, Fuente: Autor

Después de tener recopilada, centralizada y almacenada toda la información se da por se hace uso de WEKA con el fin de seleccionar de toda aquella información la más relevante y sobre todo la que mejor calidad de información tenga.

A continuación, se muestran algunas de las variables seleccionadas:

Proyecto Curricular: Los proyectos curriculares que aquí se muestran son los ofrecidos por la Universidad Ingeniería, Ver Figura 10.

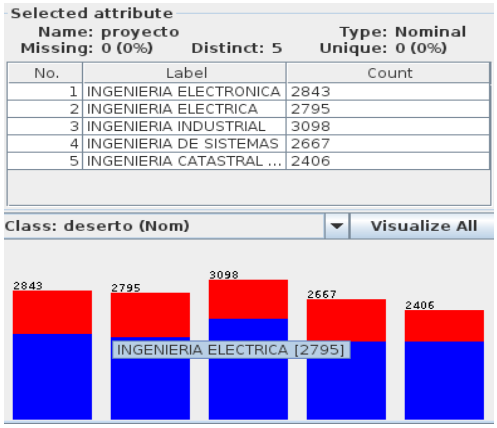

Figura 10. Variable Proyecto Curricular, Fuente: Autor

Género: Esta variable también tiene una buena calidad, al igual que la variable proyecto curricular esta tiene una representación discreta, además es de resaltar que los estudios

realizados muestran que esta variable tiene una fuerte influencia en el problema de la deserción, ver Figura 11.

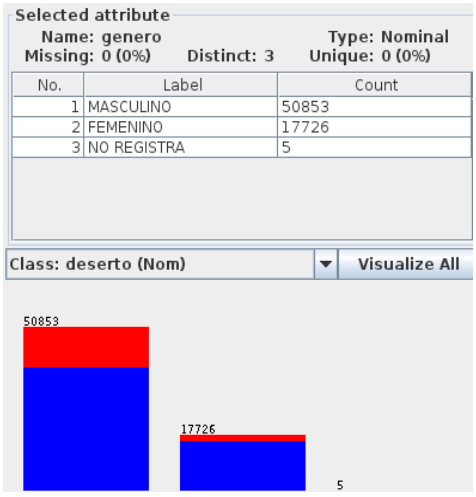

Figura 11. Variable Género, Fuente: Autor

Estrato: Esta variable también presenta una buena calidad de datos, ver Figura 12.

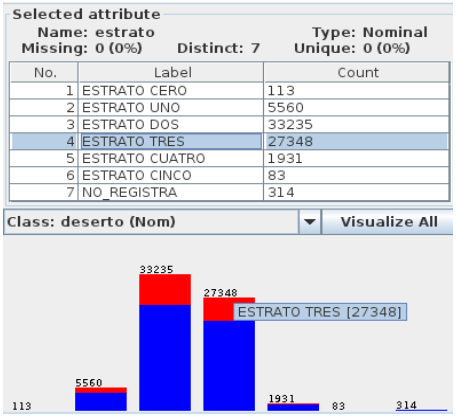

Figura 12. Variable Estrato, Fuente: Autor

Localidad: Esta variable tiene una correlación directa con la variable estrato puesto que el lugar donde se habita determina el estrato socioeconómico, Ver Figura 13.

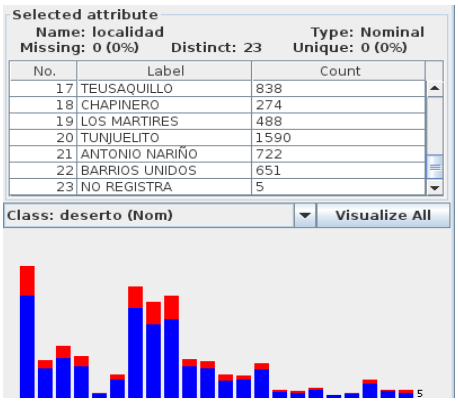

Figura 13. Variable Localidad, Fuente: Autor

Otras de las variables que se tuvieron en cuenta para el trabajo fueron las que están relacionadas con los puntajes y resultados obtenidos por los estudiantes en las pruebas del ICFES.

Acuerdo: Hace referencia a las políticas que rigen al estudiante según el semestre o periodo en el que haya ingresado a estudiar.

Desertó: Esta no será tenida en cuenta como una de las variables que influya en la generación del modelo, pues esta es la variable objetivo.

Asignatura: Es discreta y presenta una buena calidad pues todos los registros tienen valor para ella, Ver Figura 14.

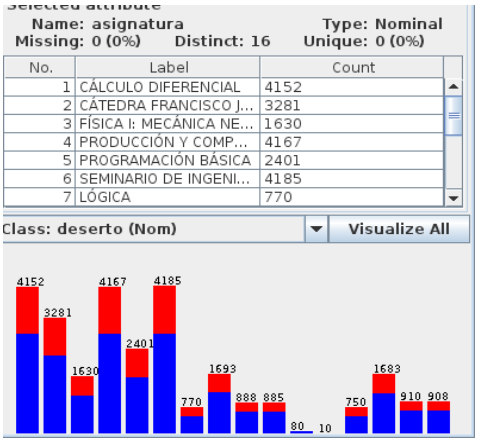

Figura 14. Variable Asignatura, Fuente: Autor

Edad: Esta variable es altamente significativa puesto que ya se ha mencionado que a mayor edad mayor riesgo de deserción, Ver Figura 15.

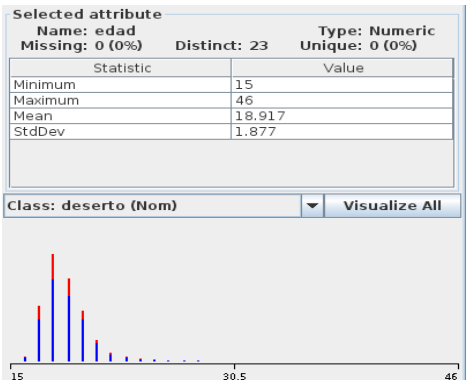

Figura 15. Variable Edad, Fuente: Autor

Aprobó: Toma valores SI o NO. Tiene una relación directa con la variable veces_cursó que hace referencia a la cantidad de veces que un estudiante tuvo que tomar una asignatura, ver Figura 16.

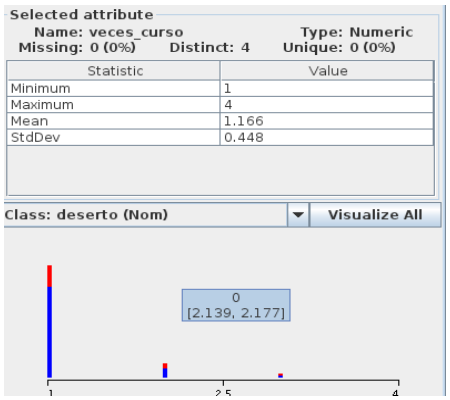

Figura 16. Variable Veces_Cursó, Fuente: Autor

Por último, se tuvo en cuenta la cantidad de materias que un estudiante cursa al mismo tiempo.

Luego de tener la información que será analizada, se procede a realizar limpieza y selección de datos. Para ello se utilizan las técnicas relacionadas en la Tabla 1:

Minería de datos		
Tarea	Técnica	Descripción
Dividir los estudiantes en subgrupos para facilitar su análisis	Cluster Analysis	Por medio de esta técnica se logra la segmentación de los estudiantes de tal forma que se facilita su análisis con mayor nivel de detalle.
Predecir la probabilidad que tienen los estudiantes de perder una materia	Regression analysis	A través de esta técnica se pueden relacionar diversas variables que permitan conocer el comportamiento académico de los estudiantes de la universidad determinar si las variables socioeconómicas influyen en su rendimiento académico
Identificar relaciones entre la deserción y los docentes que dictan las clases	Correlation analysis	Con esta técnica tratamos de ver si existe una relación entre la deserción y los docentes.

Tabla 1. Técnicas de Minería de datos a usar, Fuente: Autor

Usando la herramienta WEKA(“Weka 3 - Data Mining with Open Source Machine Learning Software in Java,” n.d.), la generación del modelo fue realizado usando el árbol de decisión mediante el algoritmo de análisis conocido como Árbol de decisión (j48), dando como resultado la Figura 17:

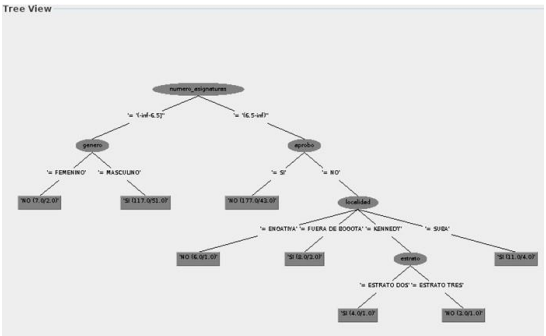

Figura 17. Modelo de árbol de decisión obtenido, Fuente: Autor

Modelo Predictivo para la deserción académica

Después de haber aplicado el algoritmo J48 se obtuvo el modelo predictivo que describe las reglas y o condiciones que son causas de deserción en el caso estudiado, dichas reglas son las que se muestran a continuación:

```
IF Número de Asignaturas<6.5
    IF Genero = Femenino
        No (7.0/2.0)
    ELSE
        SI (117.0/51.0)
ELSE
    IF Aprobó = Si
        No (177.0/43.0)
    ELSE
        IF Localidad = Engativá
                No (6.0/1.0)
        ELSE IF Localidad = Fuera de Bogotá
                Si (8.0/3.0)
        ELSE IF Localidad = Kennedy
                IF Estrato = 2
                        Si (4.0/1.0)
                ELSE IF Estrato = 3
                        No (3.0/1.0)
        ELSE IF Localidad = Suba
                Si (11.0/4.0)
```

Una breve explicación de las reglas encontradas es la siguiente.

Cuando un estudiante cursa menos de 7 asignaturas y es un hombre, tiene mayor probabilidad de desertar dado que hay 117 registros que cumplen esta regla en comparación con los 57 que no cumplen.

Por otra parte, se observa que cuando las asignaturas cursadas son mayores a 7, influye en gran medida el hecho de haberlas o no haberlas aprobado, siendo el caso en que no se aprueban en donde se encuentra la mayor ocurrencia de casos de deserción lo cual también se ve que está ligado al ámbito socioeconómico del estudiante de los cuales se resaltan aspectos tales como la localidad donde vive y el estrato.

CONCLUSIONES

Se concluye que para la aplicación de los diferentes algoritmos de Minería de Datos no siempre la misma técnica de limpieza y transformación prepara la información de manera que se ajusten al algoritmo utilizado, por ello a pesar de haber realizado las fases de limpieza y transformación no se logra la misma efectividad de resultados para el algoritmo A priori como la que se obtuvo en el algoritmo J48.

Se nota que la cantidad de materias vistas es un factor influyente para que los estudiantes tomen la decisión de desertar o pierdan la calidad de estudiantes, esto se debe a que a

mayor cantidad de asignaturas mayor es el esfuerzo que debe realizar un estudiante para aprobarlas todas.

También se nota que el género es otro de los factores que marca una tendencia de deserción, sin embargo, es necesario tener en cuenta que para la muestra analizada en este caso de estudio la mayoría de la población era de género masculino, por ende el hecho de ser un factor influyente en la deserción no aporta mayor información acerca de las causas de deserción puesto que era un resultado esperado, es por esto que en los futuros ajustes al modelo se propone descartar esta variable o tomar la misma proporción de registros tanto masculinos como femeninos y así poder identificar otras causas que no son fácilmente detectadas.

De igual manera se puede identificar al factor socioeconómico como uno de los que promueve la deserción estudiantil, pues vemos que las localidades que más aportan a la deserción son aquellas que se encuentran a una distancia que se puede considerar lejana con respecto a la ubicación en la cual se encuentra la facultad de ingeniería y son estas las que se encuentran en los estratos de donde mayormente proviene la población estudiantil.

REFERENCIAS BIBLIOGRÁFICAS

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, *6*(7), 528–533. http://doi.org/http://dx.doi.org/10.7763/IJIET.2016.V6.745

Amaya, Y., Barrientos, E., & Heredia, D. (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

Bawden, D. (2005). Data Mining and Decision Support: Integration and Collaboration. *Journal of Documentation*, *61*(3), 443–445. Retrieved from http://search.proquest.com/docview/217978314?accountid=34687

Camargo Vega, J., Camargo Ortega, J., & Joyanes Aguilar, L. (2015). Arquitectura Tecnológica Para Big Data. *Revista Científica*, *21*, 7-18. https://doi.org/10.14483/udistrital.jour.RC.2015.21.a1

Cuervo-Gómez, W. O., & Ballesteros-Ricaurte, J. A. (2015). Políticas sobre

aprendizaje móvil y estándares de usabilidad para el desarrollo de aplicaciones educativas móviles. Revista científica, 1(21), 39-52. https://doi.org/10.14483/udistrital.jour.RC.2015.21.a4

Pérez Marqués, M. (2014). *Minería de datos a través de ejemplos*. RC Libros.

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved from http://www.cs.waikato.ac.nz/ml/weka/index.html

Guzmán, C., & Duran, D., & Franco, J., & Castaño, E., & Gallón, S., & Guzmán, C., & Gómez, K., & Vasquez, J. (2009) Deserción Académica en la Educación Superior Colombiana

Sánchez, F., & Márquez, J. (2012) La Deserción en la Educación Superior en Colombia durante la Primera década del Siglo XXI ¿Por qué ha aumentado tanto?

Estrada-Sapuyes, L. O. (2015). -Methodology Proposed for Determining the Curricular Flexibility in Academic Programs Supported by Free Software under the Concept of Viable System Mod. *Revista Científica*, 2(22), 9-30. https://doi.org/10.14483/10.14483/udistrital.jour.RC.2015.22.a2