Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos

Yegny Karina Amaya Torrado[a], Edwin Barrientos Avendaño[a], Diana Judith Heredia Vizcaíno[b]

[a] Universidad Francisco de Paula Santander, Ocaña, Colombia.
ykamayat@ufpso.edu.co
ebarrientos@ufpso.edu.co

[b] Universidad Simón Bolívar, Barranquilla, Colombia.
dianahv@unisimonbolivar.edu.co

Resumen. La minería de datos o *Data Mining* permite descubrir información oculta en grandes cantidades de datos, información que por procedimientos tradicionales es muy difícil de visualizar. Esta rama de la computación permite manejar y clasificar grandes cantidades de datos, para lo cual se utilizan una gran variedad de técnicas, entre las que se encuentran los arboles de decisión C4.5 y el ID3 que han demostrado ser muy eficientes para casos específicos de predicción; este tipo de técnicas generan árboles que, de acuerdo a la complejidad del tema de estudio, pueden ser muy variables: se pueden obtener arboles con muchos nodos y hojas en el caso de ID3 y arboles más pequeños si utilizamos C4.5. Este artículo muestra la construcción de un modelo predictivo de deserción estudiantil, caracterizando a los estudiantes de la Universidad Simón Bolívar con el objetivo de poder predecir la probabilidad de deserción de los estudiantes; dicho modelo demostró el desempeño de los algoritmos presentados para clasificar datos bajo contextos variables y la precisión de uno con respecto al otro. Para la creación del modelo se utilizó la herramienta WEKA que permite de forma muy eficiente el procesamiento y clasificación de los datos con resultados satisfactorios.

Palabras claves. Deserción estudiantil, Estudiantes, Minería de datos, Modelo predictivo.

1. Introducción

Las instituciones educativas de educación superior cuentan con sistemas de información académicos los cuales registran, entre otras cosas, los datos personales, socio-económicos y los derivados del desempeño académico de los estudiantes antes y durante su permanencia en la institución. Este tipo de sistemas brinda una serie de reportes o informes para sus usuarios, pero son académicos y con información general. El objetivo de este trabajo es aplicar diferentes técnicas de *Data Mining* sobre estos datos, con el fin de obtener un modelo predictivo que permita conocer de antemano qué estudiantes están en riesgo de abandonar sus estudios. El alto nivel de deserción estudiantil es uno de los problemas principales que enfrentan las instituciones de educación superior; según

estadísticas del Ministerio de Educación Nacional, de cada cien estudiantes que ingresan, cerca de la mitad no logra culminar su ciclo académico y obtener la graduación [1].

Con la creación del modelo predictivo de deserción estudiantil se busca determinar la probabilidad de que un estudiante abandone la universidad, teniendo en cuenta las reglas de conducta y el entorno del estudiante, las cuales pueden afectar las variables primarias que inciden directamente en la deserción.

2. ¿Qué es la minería de datos?

La minería de datos se refiere a la extracción "o la minería" del conocimiento de los grandes volúmenes de datos. El término es en realidad un nombre poco apropiado, la minería de datos de una manera más apropiada debería haber sido llamada "la minería de conocimiento de datos", que es lamentablemente largo. "La minería de conocimiento", un término más corto, no puede reflejar el énfasis en la minería de los volúmenes de datos. Muchas personas tratan la minería de datos como un sinónimo de otro término utilizado popularmente, el Descubrimiento de Conocimiento de Datos, o KDD, consiste en una secuencia iterativa de los siguientes pasos:

1. Datos de limpieza, 2. La integración de datos, 3. La selección de datos, 4. Transformación de datos, 5. La minería de datos, 6. Modelo de evaluación, 7. La presentación del conocimiento [2].

Las técnicas que conforman el campo de la Minería de Datos buscan descubrir, en forma automática, el conocimiento contenido en la información almacenada en las bases de datos de las organizaciones. Por medio del análisis de datos, se pretende descubrir patrones, perfiles y tendencias. Es importante que estas técnicas sean las adecuadas al problema abordado. En este sentido, se pueden establecer dos grandes grupos de técnicas o métodos analíticos: los métodos simbólicos y los métodos estadísticos [3]. Entre los métodos simbólicos se incluyen a las Redes Neuronales, Algoritmos Genéticos, Reglas de Asociación, Lógica Difusa, entre otros. Estos derivan del campo de la Inteligencia Artificial.

3. Análisis de técnicas de minería de datos utilizadas en estudios sobre deserción estudiantil

En las instituciones de educación superior se ve la necesidad de contar con mecanismos que ayuden a disminuir la deserción de los estudiantes en las diferentes carreras que se ofrecen. Es por esto que este estudio se basa en el análisis de técnicas de minería de datos que utilizan los diferentes modelos predictivos con la finalidad de buscar la que mejor se comporte para predecir la probabilidad de que un estudiante deserte, comprobándose con

el menor margen de error. En el siguiente cuadro se muestran varios estudios realizados al respecto y las técnicas utilizadas.

Tabla 1. Técnicas de Minería de Datos utilizadas en estudios similares (Fuente: Autores del Proyecto)

N-	**PAIS**	**ESTUDIO**	**TECNICAS UTILIZADAS**
1.	Colombia	Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. (Timarán P., 2009)	árboles de decisión C4.5
			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
4.	Tailandia	*A Comparative Analysis of Techniques for Predicting Academic Performance.* [5]	Árboles de Decisión (J48)
			Redes Bayesianas
5.	Athens, Ohio, USA	*A Model to Predict Ohio University Student Retention From Admissions and Involvement Data.* [6]	Regresión lineal la regresión logística, árboles de decisión
6.	Estados Unidos	*A Comparison of Logistic Regression, Neural Networks, and Classification Trees Predicting Success of Actuarial Students Phyllis Schumacher.* [7]	Regresión logística, redes neuronales, árboles de clasificación
7.	Argentina	Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. [8]	Técnica de Regresión Logística,
8.	México	Minería de datos: predicción de la deserción escolar mediante el algoritmo de	Árboles de decisión C4.5

		árboles de decisión y el algoritmo de los k vecinos más cercanos. [9]	Técnica de los k vecinos más cercanos
9.	México	Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos. [10]	árboles de decisión mediante el algoritmo C4.5,
10.	Estados Unidos	*New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention.* [11]	árboles de clasificación y regresión logística
11.	Nueva Zelanda	*Predicting student success by mining enrolment data.* [12]	árboles de clasificación) y regresión logística
12.	Croatia,	*Student dropout analysis with application of data mining methods.* [13]	regresión logística, árboles de decisión y redes neuronales
13.	Estados Unidos	*Learning Patterns of University Student Retention.* [14]	Uno-R, C4.5, ADTrees, Naive Bayes, Bayes redes y redes radiales sesgo
14.	Estados unidos	*Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks.* [15]	Redes neuronales (*backpropagation feed-forward*)
15.	India	*Data Mining: A prediction for performance improvement using classification.* [16]	Algoritmo de clasificación Bayesiano (Naïve Bayes).
16.	India	*Mining Education Data to Predict Student's Retention: A comparative Study.* [17]	Árbol de decisión ID3
			Árbol de decisión C4.5
			Árbol de decisión ADT
17.	México	Modelos predictivos y explicativos del rendimiento académico universitario: caso de una institución privada en México. [18]	
18.	México	Creación de un modelo de predicción del desempeño académico de los alumnos de la Facultad de Ingeniería de la UNAM en el primer semestre. [19]	Naive-Bayes
19.	España	Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. [20]	Árboles de Decisión CART
			Regresión multivariante.

4. Metodología

Para el desarrollo de este estudio, se llevaron a cabo las siguientes fases:

- Recopilación información de las técnicas de minería de datos para elegir la adecuada de acuerdo al problema planteado, analizar los modelos predictivos existentes de deserción estudiantil en educación superior.
- Caracterización los datos personales y académicos de los estudiantes de educación superior. Esta incluye: Preparación, Preprocesamiento y Transformación de datos.
- Construcción y prueba del modelo de deserción en la educación superior.
- Validación del modelo de deserción y análisis de los resultados arrojados en su despliegue.

4.1. Selección de las técnicas a utilizar en el desarrollo del trabajo

Se ha optado por la inducción de árboles de decisión porque además de ser la técnica más común dentro las técnicas de clasificación de datos, representa una gran ventaja con respecto a las demás técnicas de clasificación: poder representar el conocimiento extraído en un conjunto de reglas de decisión de fácil entendimiento; además en el cuadro comparativo podemos observar que son los que tienen mayor precisión (Heredia, d. & Nieto, w. 2011; Jadric, M. & Otros 2010; Kumar Y, S. & Otros 2012; Nandeshwara, A & Otros 2011; Nghe, N. & Otros 2007; Roth, S. y Koonce, D. 2008; Rodallegas R. & Otros 2010; Schumacher, P. & Otros 2010, Timarán P., R. O. 2009; Valero Orea, & Otros 2010; Winstead, D. 2010).

Árboles de decisión. Un árbol de decisión es un diagrama de flujo, con estructura de árbol, en donde los nodos internos representan validaciones sobre los atributos, las ramas representan las salidas de las validaciones, y los "nodos hoja" representan las clases. El nodo en la parte superior del árbol se le conoce como nodo raíz. Para clasificar una instancia "desconocida", se sigue el flujo del árbol hacia abajo, de acuerdo a los valores que tengan los atributos para cada nodo, y cuando se llega a un "nodo hoja", la instancia se clasifica de acuerdo a la clase asignada por dicho nodo **[21]**. Existen diversos métodos para la inducción de árboles de decisión (ID3, C4, C4.5, Bayesiano, CART, etc.), cada uno de ellos ofrece diferentes capacidades, pero en general, dichos algoritmos son apropiados para solucionar «problemas de clasificación».

El algoritmo ID3 [22]. La idea básica del algoritmo ID3 tiene su fundamento en la iteración. Un subconjunto del conjunto total de datos de entrenamiento, al cual se le conoce como *window*, es elegido de manera aleatoria para formar un árbol de decisión; este árbol clasifica de manera correcta todos los objetos que pertenecen a *window*. El resto de los objetos, dentro del conjunto de datos de entrenamiento, es clasificado utilizando dicho árbol. Si el árbol da una respuesta correcta para todos estos objetos, entonces también es correcto para el conjunto total de datos de entrenamiento, terminando así el proceso. Si no, una selección de objetos clasificados incorrectamente es adicionada al subconjunto *window*, y el proceso continúa.

El algoritmo C4.5. El algoritmo forma parte de la familia de los «Árboles Inducidos de Arriba hacia Abajo» (*Top Down Induction of Decision Trees*, TDIDT por sus siglas en inglés). Pertenece a los métodos inductivos del *Machine Learning*, los cuales aprenden a

partir de ejemplos preclasificados. Tanto el algoritmo ID3 como el C4.5 fueron propuestos por Ross Quinlan [22]. El algoritmo C4.5 es una extensión del algoritmo ID3, el cual trabaja únicamente con valores discretos en los atributos. En cambio, el algoritmo C4.5 permite trabajar con valores continuos, separando los posibles resultados en dos ramas: una para aquellos $A_i <= N$ y otra para todos los $A_i > N$. De esta forma, C4.5 genera un árbol de decisión a partir de datos mediante particiones realizadas de manera recursiva.

4.2. Caracterización de los datos personales y académicos de los estudiantes de educación superior

Fuente de datos. Para el desarrollo del estudio del presente trabajo, la Universidad Simon Bolivar proporcionó datos socio económicos y académicos de los estudiantes del programa de Ingenieria de Sistemas, correspondientes a los últimos 6 años. Para la elaboracion del modelo se utilizaron dos algoritmos de mineria de datos, ID3 y C4.5 (En Weka, éste último está implementado como J4.8), los cuales inducen árboles de decisión. Se ha optado por la inducción de árboles de decisión porque además de ser la técnica más común dentro las técnicas de clasificación de datos, representan una gran ventaja con respecto a las demás técnicas del mismo tipo, dado que esta puede representar el conocimiento extraído en un conjunto de reglas de decisión de fácil entendimiento.

Muestra inicial. Como primer paso, se extrajo una muestra de los datos de estudiantes. La información se generó en un archivo de excel separados por comas (CSV) para tener mayor flexibilidad al momento de exportar los datos. Cabe señalar que esta primera muestra contiene información restringida de los formularios que el estudiante diligencia al momento de inscribirse a la universidad y algunas actualizaciones que realiza durante su permanencia en la misma, junto con otra que se genera durante el proceso académico, por lo que en todo momento se respetó la confidencialidad de los datos de los estudiantes. Esta muestra inicial de datos constó de 201 instancias con 40 atributos, los cuales corresponden a informacion que registra el estudiantes al momento de matricularse y alguna que se genera durante su permanencia en la universidad como promedio, asignaturas reprobadas, entre otras.

El atributo *Desertado*, nos identifica el estado actual del estudiante en la universidad, teniendo en cuenta que para el Ministerio de Educación un estudiante se considera desertor de un programa, una institución o del sistema de educación superior si abandona sus estudios durante dos períodos consecutivos y no registra matrícula.

Una vez se tienen los datos cargados en WEKA se procedió a estudiar los datos seleccionados para entender el significado de los atributos, detectar errores de integración, estandarizar datos, hacer agrupaciones, etc. Se aplicó el filtro Remove de WEKA, para los atributos que no se consideran relevantes para el desarrollo del modelo; se pasó de tener 40 a 16 atributos considerados como relevantes para crear el modelo predictivo. Para tener una mejor percepción de los atributos finales se aplicó el filtro de Discretize con el cual

se dividieron en 2 particiones (bins=2) los atributos de los estudiantes que tienen la posibilidad de desertar de la universidad

Atributos finales. Una vez que la muestra inicial de datos fue preprocesada se obtuvieron los atributos finales para la construcción del modelo: Semestre (en el cual está matriculado el estudiante), Edad actual, Ciudad de procedencia, Estrato, Jornada, Sexo, Valor de la matrícula, Ocupación, Materias cursadas, Materias perdidas, Promedio, Estado Civil, Nivel de estudios del Padre, Nivel de estudios de la Madre, Ingresos y Desertado.

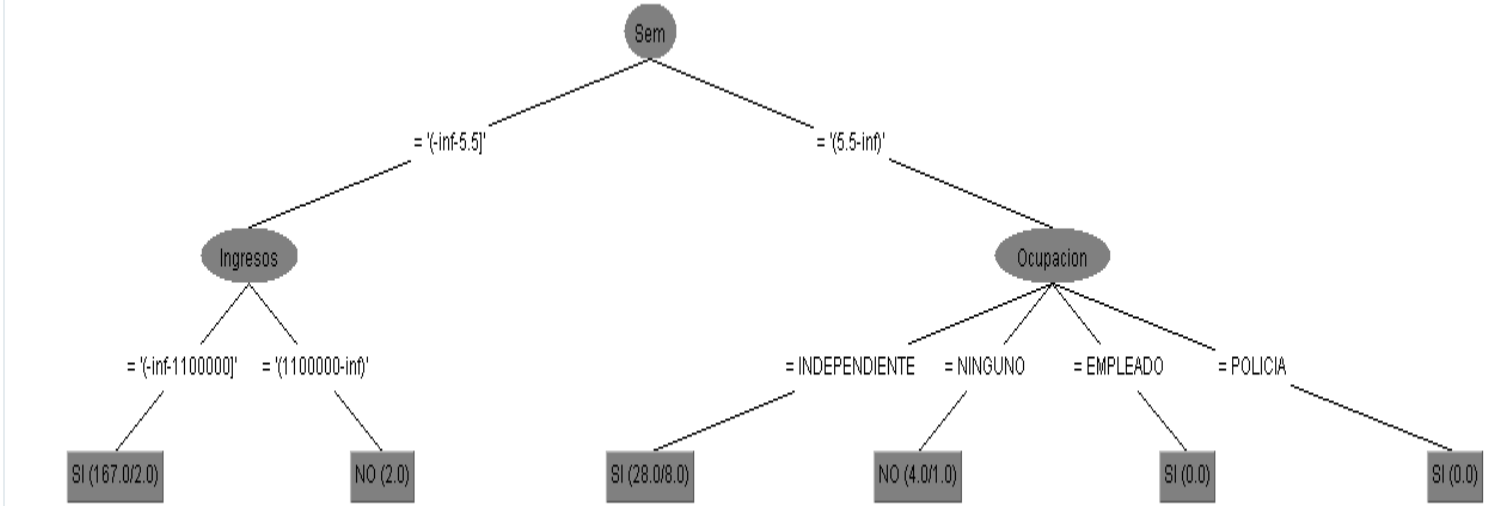
5. Resultados

5.1. Construcción y pruebas del modelo de deserción en la educación superior.

Interpretación de resultados

Árbol de decisión obtenido J4.8. El modelo se construyó a partir de 201 instancias, las cuales fueron utilizadas como conjunto de entrenamiento y para probar el modelo obtenido. Una vez clasificada la informacion con el algoritmo J4.8 de Weka, se obtuvo el siguiente arbol de decision como salida.

Figura 1. Árbol de decisión del modelo obtenido (Fuente: Autores del Proyecto)



5.2. Validación del modelo de deserción estudiantil

Una vez que se tiene el modelo se procede a probarlo con datos nuevos, con el fin de visualizar qué estudiantes tienen posibilidad de desertar de la Universidad.

Luego de realizar varias pruebas se cambió el factor de confianza para analizar los resultados: Colocando el factor en 0.25 se puede observar que solo muestra una rama y la posibilidad de que los estudiantes deserten es cero. Con el factor de confianza en 0.75 y 1.0 muestra que varios estudiantes tienen posibilidades de desertar. En la reevaluación de los modelos se comparan los resultados arrojados utilizando el árbol de decisión J48 y IDE3. La tabla 2 muestra los resultados de esta reevaluación (Se eliminaron líneas irrelevantes): En la primera columna los estudiantes resaltados con color amarillo son los que desertaron según la información del segundo semestre del año 2011 representados en el Spadies (Sistema para la Prevención de la Deserción de la Educación Superior, Colombia) en el año 2012-I y las demás columnas muestra en color rojo los estudiantes que tienen la posibilidad de desertar según el modelo.

Tabla 2. Reevaluación de los Modelos (Fuente: Autores del Proyecto)

MODELO GENERADO CON J48 FACTOR EN: 0.25	**MODELO GENERADO CON J48 FACTOR EN: 0.75**	**MODELO GENERADO CON J48 FACTOR EN: 1.0**	**MODELO GENERADO CON IDE3**
=== Re-evaluation on test set === User supplied test set Relation: DATOS_VALI-weka.filters.unsupervised.attribute.MakeIndicator-C3-V1-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-Rfirst-last Instances: unknown (yet). Reading incrementally Attributes: 16 === Predictions on test set === inst#, actual, predicted, error, probability distribution 1 ? 1:NO	== Re-evaluation on test set === User supplied test set Relation: DATOS_VALI-weka.filters.unsupervised.attribute.MakeIndicator-C3-V1-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-Rfirst-last Instances: unknown (yet). Reading incrementally Attributes: 16 === Predictions on test set === inst#, actual, predicted, error, probability distribution 1 ? 1:NO + *1 0	=== Re-evaluation on test set === User supplied test set Relation: DATOS_VALI-weka.filters.unsupervised.attribute.MakeIndicator-C3-V1-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-Rfirst-last Instances: unknown (yet). Reading incrementally Attributes: 16 === Predictions on test set === inst#, actual, predicted, error, probability distribution 1 ? 1:NO + *1 0	=== Re-evaluation on test set === User supplied test set Relation: DATOS_VALI-weka.filters.unsupervised.attribute.MakeIndicator-C3-V1-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-Rfirst-last Instances: unknown (yet). Reading incrementally Attributes: 16 === Predictions on test set === inst#, actual, predicted, error, probability distribution 1 ? 1:NO + *1 0

+ *0.841 0.159	2 ? 1:NO	2 ? 1:NO	2 ? 1:NO
2 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	3 ? 1:NO	3 ? 1:NO	3 ? 1:NO
3 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	4 ? 1:NO	4 ? 1:NO	4 ? 1:NO
4 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	5 ? 1:NO	5 ? 1:NO	5 ? 1:NO
5 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	6 ? 1:NO	6 ? 1:NO	6 ? 1:NO
6 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	7 ? 1:NO	7 ? 1:NO	7 ? 1:NO
7 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	8 ? 1:NO	8 ? 1:NO	8 ? 1:NO
8 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	9 ? 1:NO	9 ? 1:NO	9 ? 1:NO
9 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	10 ? 1:NO	10 ? 1:NO	10 ? 1:NO
10 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	**11 ? 2:SI**	**11 ? 2:SI**	**11 ? 2:SI**
11 ? 1:NO	**+ 0.25 *0.75**	**+ 0.25 *0.75**	**+ 0 *1**
+ *0.841 0.159	12 ? 1:NO	12 ? 1:NO	12 ? 1:NO
12 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	13 ? 1:NO	13 ? 1:NO	13 ? 1:NO
13 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	**14 ? 2:SI**	**14 ? 2:SI**	**14 ? 2:SI**
14 ? 1:NO	**+ 0.25 *0.75**	**+ 0.25 *0.75**	**+ 0 *1**
+ *0.841 0.159	15 ? 1:NO	15 ? 1:NO	15 ? 1:NO
15 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	**16 ? 2:SI**	**16 ? 2:SI**	16 ? 1:NO
16 ? 1:NO	**+ 0.25 *0.75**	**+ 0.25 *0.75**	+ *1 0
+ *0.841 0.159	17 ? 1:NO	17 ? 1:NO	17 ? 1:NO
17 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	18 ? 1:NO	18 ? 1:NO	18 ? 1:NO
18 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	19 ? 1:NO	19 ? 1:NO	19 ? 1:NO
19 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	36 ? 1:NO	36 ? 1:NO	36 ? 1:NO
36 ? 1:NO	+ *0.95 0.05	+ *0.95 0.05	+ *1 0
+ *0.841 0.159	37 ? 1:NO	37 ? 1:NO	**37 ? 2:SI**
37 ? 1:NO	+ *0.931 0.069	+ *0.931 0.069	**+ 0 *1**
+ *0.841 0.159	**38 ? 2:SI**	**38 ? 2:SI**	**38 ? 2:SI**
38 ? 1:NO	**+ 0.25 *0.75**	**+ 0.25 *0.75**	**+ 0.333 *0.667**
+ *0.841 0.159	39 ? 1:NO	39 ? 1:NO	39 ? 1:NO
39 ? 1:NO	+ *0.931 0.069	+ *0.931 0.069	+ *1 0
+ *0.841 0.159	40 ? 1:NO	40 ? 1:NO	40 ? 1:NO
40 ? 1:NO	+ *0.931 0.069	+ *0.931 0.069	+ *1 0
+ *0.841 0.159	41 ? 1:NO	41 ? 1:NO	41 ? 1:NO
41 ? 1:NO	+ *1 0	+ *1 0	+ *1 0
+ *0.841 0.159	42 ? 1:NO	42 ? 1:NO	42 ? 1:NO
42 ? 1:NO	+ *0.931 0.069	+ *0.931 0.069	+ *1 0
+ *0.841 0.159	43 ? 1:NO	43 ? 1:NO	**43 ? 2:SI**
43 ? 1:NO	+ *0.667 0.333	+ *0.667 0.333	**+ 0 *1**
+ *0.841 0.159	44 ? 1:NO	44 ? 1:NO	**44 ? 2:SI**

44 ? 1:NO + *0.841 0.159 45 ? 1:NO + *0.841 0.159 46 ? 1:NO + *0.841 0.159 47 ? 1:NO + *0.841 0.159 48 ? 1:NO + *0.841 0.159 49 ? 1:NO + *0.841 0.159 50 ? 1:NO + *0.841 0.159	+ *0.8 0.2 45 ? 1:NO + *0.857 0.143 **46 ? 2:SI** **+ 0.4 *0.6** **47 ? 2:SI** **+ 0.4 *0.6** 48 ? 1:NO + *0.8 0.2 49 ? 1:NO + *0.931 0.069 50 ? 1:NO + *0.95 0.05	+ *0.8 0.2 45 ? 1:NO + *0.857 0.143 **46 ? 2:SI** **+ 0.4 *0.6** **47 ? 2:SI** **+ 0.4 *0.6** 48 ? 1:NO + *0.8 0.2 49 ? 1:NO + *0.931 0.069 50 ? 1:NO + *0.95 0.05	**+ 0 *1** **45 ? 2:SI** **+ 0 *1** **46 ? 2:SI** **+ 0 *1** **47 ? 2:SI** **+ 0 *1** 48 ? 1:NO + *1 0 49 ? 1:NO + *1 0 50 ? 1:NO + *1 0

En la Tabla 2 se observa que de 5 estudiantes que desertaron según Spadies en el año 2012-I, los modelos generados coinciden con más de la mitad de los estudiantes que tienen la posibilidad de desertar.

6. Conclusiones

Al analizar las diferentes técnicas que se utilizan en la minería de datos específicamente para la predicción, se observa que los arboles de decisión resultan ser buenos clasificadores, según los resultados obtenidos.

Para la creación del modelo predictivo se tuvieron en cuenta variables de tipo personal como edad actual, ciudad de procedencia, estrato, sexo, ocupación, estado civil, nivel de estudios del padre y de la madre; variables económicas como valor de la matricula e ingresos y de carácter académico como semestre, jornada, materias cursadas, materias perdidas y promedio. Estas variables se discretizaron en 2 binas para obtener un mejor análisis en el momento de clasificar los estudiantes que pueden desertar.

De las 16 variables que se utilizaron para la construcción del modelo se observó que 4 fueron descartadas por el algoritmo J48 entre los que se encuentran sexo, ocupación, materias perdidas y nivel de estudios de la madre.

Luego de realizar varias pruebas con los algoritmos IDE3 y J48, se observa que el modelo varía considerablemente debido a las características que tiene cada algoritmo y a su vez las características de los datos. Por ejemplo, con el algoritmo J48 se produce un árbol de decisión de 6 reglas y con IDE3, uno de 22 reglas; lo que muestra que el árbol con J48 es pequeño, pero al momento de observar su capacidad de predecir se observa que de los 201 estudiantes matriculados en el segundo semestre del año 2011 desertaron 14 estudiantes identificados en Spadies en el año 2012-I. Al ejecutar los datos en el modelo muestra que 21 estudiantes pueden desertar, de los cuales 9 son acertados por la información suministrada de los desertores y se debe tener en cuenta que hay 12 estudiantes que son tomados por el modelo como posibles desertores.

Al cambiar los datos de entrenamiento teniendo en cuenta diferentes semestres se observó que el modelo cambia sustancialmente; por ejemplo el primero solo toma 3 atributos (semestre, ingreso y ocupación) descartando los 12 restantes y del semestre 2011-2 toma 11 atributos descartando cuatro, lo que muestra que el modelo es especifico a la población objeto de estudio y a las características de los atributos. El descarte de los atributos, el modelo lo realiza sobre las variables que no generan relevancia para el mismo, ya que por presentar algún grado de similitud no se pueden identificar características que reflejen un patrón que evidencie cuando el estudiante puede desertar.

Para este caso de estudio y de acuerdo con los resultados obtenidos en la validación del modelo, se puede decir que el algoritmo ID3 presenta mayor precisión al momento de predecir qué estudiantes tenían la probabilidad de desertar de la Universidad. Esto se ve reflejado por el mayor número de reglas que se generan en el modelo utilizando ID3, lo que tiene mayores probabilidades de clasificación.

Teniendo un modelo construido se puede pensar en usar las reglas generadas en proyectos de desarrollo de software que permita que la minería de datos sea un poco más dinámica, ya que se apreció en el desarrollo del trabajo que al realizar el proceso utilizando la herramienta Weka se hace tedioso la puesta en marcha del modelo, dado que el volumen de datos sobre el cual se va a predecir debe cumplir con unas características que Weka requiere; en este caso, por ejemplo, se requieren 15 atributos a los cuales primero se les debe aplicar los filtro *MakerIndicator* y de discretización en 2 binas, para que el modelo pueda hacer el proceso de predicción.

7. Referencias bibliográficas

[1] C. Guzmán Ruíz, D. Durán Muriel, J. Franco Gallego, E. Castaño Vélez, S. Gallón Gómez, K. Gómez Portilla y J. Vásquez Velásquez, Deserción estudiantil en la educación superior colombiana, Bogota. Colombia: Imprenta Nacional de Colombia, 2009.

[2] H. Jiawei y K. Micheline, Data Mining, Concepts and Techiniques, Elsevier Inc., 2006.

[3] P. Britos, Minería de Datos, Buenos Aires: Nueva Librería, 2005.

[4] D. Heredia y W. Nieto, «Generación de un modelo predictivo para determinar el desempeño académico en la asignatura fundamentos de programación II del programa de Ingeniería de Sistemas,» Colombia, 2011.

[5] N. T. Nghe, P. Janecek y P. Haddawy, «A comparative analysis of techniques for predicting academic performance. Asian Institute of Technology 37th ASEE/IEEE Frontiers in Education Conference T2G-7,» 2007.

[6] S. Roth y D. Koonce, «A Model to Predict Ohio University Student Retention From Admissions and Involvement Data,» USA, 2008.

[7] P. Schumacher, A. Olinsky, R. Quinn y R. Bryant, «A Comparison of Logistic Regression, Neural Networks, and Classification Trees Predicting Success of Actuarial Students,» Estados Unidos, 2010.

[8] E. A. Porcel, G. N. Dapozo y M. V. López, «Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa,» *Revista Electronica de Investigaciones Educativas,* 2010.

[9] S. Valero Oreal, A. Salvador Vargas y M. García Alonso, «Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos,» 2010.

[10] E. Rodallegas R, G. Torres y B. Gaona C, «Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos,» Mexico, 2010.

[11] D. Kelley-Winstead, «New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention,» 2010.

[12] Z. J. Kovacic, «Predicting student success by mining enrolment data,» 2010.

[13] M. Jadric, Z. Garaca y M. Cukušc, «Student dropout analysis with application of data mining methods,» Croatia, 2010.

[14] A. Nandeshwara, T. Menziesb y C. Nelson, «ALearning Patterns of University Student Retention,» Estados Unidos, 2011.

[15] R. Alkhasawneh y R. Hobson, «Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks,» Estados unidos, 2011.

[16] B. Kumar B y S. Pal, «Data Mining: A prediction for performance improvement using classification,» India, 2011.

[17] S. Kumar Y, B. Bharadwaj y S. Pal, «Mining Education Data to Predict Student's Retention: A comparative Study,» India, 2012.

[18] M. Guzmán B, «Modelos predictivos y explicativos del rendimiento académico universitario: caso de una institución privada en México,» Madrid, 2012.

[19] E. P. Ibarra García y P. M. Mora E., «Creación de un Modelo de Predicción del desempeño académico de los alumnos de la Facultad de Ingeniería de la UNAM en el primer semestre.,» Mexico, 2010.

[20] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Mas, M. J. Ramirez Quintana, A. Robles, J. M. Valiente, M. J. Vicent y L. R. Zúnica, «Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.,» España, 2007.

[21] I. H. Witten y E. Frank, Data Mining: Practical machine learning tools and techniques with java implementations, San Francisco: Morgan Kaufmann Publishers,

2000.

[22] J. R. Quinlan, Induction of Decision Trees, 1986.

[23] J. Han y M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers, 2000.

[24] R. Timarán P., «Una lectura sobre deserción universitaria en estudiantes de pregrado desde: la perspectiva de la Minería de Datos,» Colombia, 2010.

[25] R. Timarán P., «Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos.,» de *Octava Conferencia Iberoamericana*, Colombia, 2009.

[26] D. Winstead, «New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention,» 2010.