Modelo de estimación de riesgo de deserción de los estudiantes de pregrado de la Universidad Autónoma de Bucaramanga con técnicas de minería de datos soportadas en software libre

Miguel E. Ropero y Silvio R. Cuello

Resumen

En este articulo se presentan los resultados obtenidos en la predicción de deserción en los estudiantes de primer ingreso a los programas de pregrado, modalidad presencial, en la Universidad Autónoma de Bucaramanga (UNAB), empleando la técnica de clasificación de Minería de Datos con el algoritmo de arboles de decisión. En la investigación, fueron utilizados los datos académicos, psicológicos y financieros de los estudiantes en los periodos 2014-02, 2015-01, 2015-02 y 2016-01, realizando el proceso de comprensión, análisis, modelado y procesamiento sugerido por la Metodología CRISP-DM. El procesamiento de los datos se realizó utilizando el software Weka a través del algoritmo J48.

Los datos de los estudiantes también fueron evaluados a través del modelo matemático de deserción desarrollado por la UNAB. Los resultados de los dos procesos fueron analizados para determinar cuál de ellos ofrecía un mayor porcentaje de precisión, cruzando los resultados obtenidos en cada uno de los procesos contra la información real de deserción. Al final, se concluye que los porcentajes de precisión del modelo de minería de datos son mayores en comparación con los resultados del modelo matemático.

Abstract

In this article the results obtained in the prediction of the desertion in the students of first entrance to the undergraduate programs are presented, in face-to-face modality, in the Autónoma de Bucaramanga University (UNAB), using the technique of classification of the Mining of Data with the algorithm of decision trees. In the research, the academic, psychological and financial data of the students were used in the periods 2014-02, 2015-01, 2015-02 and 2016-01, performing the process of understanding, analysis, modeling and processing suggested by the Methodology CRISP-DM. The processing of the data was done using the Weka software through the J48 algorithm.

The student data were also evaluated through the mathematical model of desertion developed by the UNAB. The results of the processes were analyzed to determine which was the highest proportion of precision percentages, crossing the results obtained in each of the processes against the actual information of desertion. In the end, it is concluded that the precision percentages of the data mining model are greater in comparison with the results of the mathematical model.

Palabras Clave

Minería de datos, arboles de decisión, técnicas de clasificación, algoritmo J48, Weka

1. Introducción

Las instituciones de educación superior se enfrentan semestre a semestre a sus propios indicadores en cada uno de sus procesos académicos y administrativos. Uno de ellos es la Deserción estudiantil, que según el Ministerio de Educación Nacional, (2009) se define como: “Una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proceso educativo, considerando además como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres consecutivos”.

Así mismo, la deserción es medida desde dos perspectivas diferentes: La deserción por periodo y deserción por cohorte. La deserción por periodo tiene en cuenta la proporción de estudiantes que se matricularon y no reportaron actividad académica por dos semestres consecutivos. Según datos del Sistema para la Prevención de la Deserción en la Educación Superior (SPADIES), para el año 2015 Colombia registró una tasa de deserción por periodo del 9.3% en el nivel de formación universitaria. (Ministerio de Educación Nacional. 2016)

La deserción por cohorte tiene en cuenta la proporción de estudiantes matriculados para un semestre académico en comparación con el semestre inmediatamente anterior. Según el Ministerio de Educación Nacional en este indicador Colombia registra una tasa de deserción del 46.1% en el nivel de formación universitaria. (Ministerio de Educación Nacional., 2016)

Teniendo en cuenta que las IES reciben gran parte de sus ingresos económicos del valor de la matrícula financiera pagada por sus estudiantes, especialmente las IES de carácter privado, la deserción estudiantil merece entonces una atención especial para mantener en las aulas el mayor número de estudiantes a través del desarrollo de programas de retención aplicados en la población vulnerable.

Según Berry y Linoff (2004) “La minería de datos consiste en encontrar relaciones y patrones que existen en la base de datos, pero que están ocultos o a simple vista no son observables dentro de grandes cantidades de datos”. Existen varias técnicas de minería de datos que permiten realizar predicciones a través de un conjunto de datos de entrada. Cada técnica predictiva cuenta con una variedad de algoritmos que permiten obtener diferentes niveles de predicción de acuerdo a la calidad de los datos.

Weka es una herramienta de software de distribución libre, desarrollada por la Universidad de Waikato, la cual, brinda diferentes técnicas y algoritmos de minería de datos, muchos de ellos, orientados a la clasificación, los cuales, pueden ser utilizados para predecir escenarios futuros.

La aplicación de la minería de datos a través del software Weka sobre la información de los estudiantes, permite, de acuerdo a los niveles de predicción obtenidos, apoyar la toma de decisiones para desarrollar y aplicar diferentes planes de retención encaminados a reducir los porcentajes de deserción de los estudiantes de primer ingreso de los programas de pregrado. En este articulo se describe el proceso de minería de datos aplicado en la Universidad Autónoma de Bucaramanga, utilizando la información académica, psicología y financiera de los estudiantes de primer ingreso a los programas de pregrado de los últimos cinco periodos académicos, utilizando el algoritmo de arboles de decisión J48 del software Weka.

2. Minería de datos

De acuerdo con la definición de Witten y Frank, (2005) “la Minería de Datos es la extracción de información implícita, previamente desconocida y potencialmente útil de los datos”, que a través de programas software permiten encontrar de forma automática regularidades o patrones. estos patrones pueden ser utilizados en la aplicación de dos modelos, el modelo descriptivo o el modelo predictivo. El primero explica o resume los datos, es decir, sirven para explorar las propiedades de los datos examinados. El segundo, pretende estimar valores futuros o desconocidos de variables de interés, denominadas variables objetivo o dependientes, usando otras variables, denominadas variables predictivas o independientes. A su vez, cada modelo cuenta con técnicas de minería de datos que tienen diferentes formas de análisis.

2.1 Técnicas de Clasificación

La clasificación es el proceso para encontrar una función que describe y distingue clases o conceptos de datos. El modelo se deriva a partir del análisis de un conjunto de datos de entrenamiento donde el valor de la variable de análisis es conocido. El modelo se utiliza para predecir el valor de la variable que es desconocida, a través de los patrones observados en los datos de entrenamiento. (Han & Kamber & Pei, 2011).

Arboles de Decisión

Son un algoritmo clasificación basado en una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol, donde cada nodo interno (nodo no hoja) denota una prueba en un atributo, cada rama representa un resultado de la prueba, y cada nodo hoja (o nodo terminal) tiene una etiqueta de clase (variable desconocida). El nodo superior es el nodo raíz, (Han & Kamber & Pei, 2011).

2.2 Técnicas de agrupamiento

El agrupamiento consiste en obtener grupos de datos fuertemente relacionados entre si. A diferencia de la clasificación la agrupación no utiliza etiquetas de clase, por el contrario, las crea, ya que cada grupo de datos generado puede considerarse como una etiqueta clase. En este método los datos son agrupados con el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo son muy diferentes a los objetos de otro grupo, (Vercellis, 2009).

2.3 Técnicas de asociación

Se utilizan para identificar asociaciones recurrentes entre grupos de registros en un conjunto de datos. Por ejemplo, es posible determinar qué productos se compran en conjunto en una sola transacción y con qué frecuencia, (Hernández & Ramírez & Ferri, 2004).

3. Antecedentes

La UNAB dentro de sus estrategias para reducir el porcentaje de deserción de sus estudiantes, ideó, a

partir de la experiencia del departamento de Bienestar Universitario, un modelo empírico para calcular el riesgo de deserción asociado a cada uno de los estudiantes que cursan el primer semestre académico en los programas de pregrado.

Este modelo, utiliza la información académica, financiera y psicológica, obtenida como parte del proceso de admisión de los estudiantes, relacionando un conjunto de variables y definiendo los valores de entrada para cada una de ellas (**ver anexos**). La siguiente tabla, presenta el total de variables analizadas en el modelo empírico.

Tipo de Variable	Variables Analizadas
Académicas	8
Financieras	14
Psicológicas	34
Total Variables	56

Tabla 1. Variables analizadas en el modelo empírico

Cada variable tiene asociado un nivel de riesgo, el cual, puede ser alto, medio o bajo. Inicialmente se calcula de forma individual el riesgo académico, el riesgo psicológico y el riesgo financiero, para luego calcular el riesgo general del estudiante, donde cada tipo de riesgo representa un 40%, 30% y 30% respectivamente sobre el riesgo general. La siguiente formula matemática define el calculo del riesgo general.

$$\mathbf{RG} = (\mathbf{RA} * 0{,}4) + (\mathbf{RF} * 0{,}3) + (\mathbf{RP} * 0{,}3)$$

Donde RG representa el riesgo general, RA el riesgo académico, RF el riesgo financiero y RP el riesgo psicológico. Porcentajes superiores al 28% en el riesgo general, son considerados con alto riesgo de deserción, por lo tanto, se les realiza seguimiento dentro de los planes de acompañamiento desarrollados por la UNAB para evitar que al final del semestre, el estudiante abandone sus estudios.

4. Metodología

Para la aplicación del modelo de minería de datos se tuvieron en cuenta las mismas variables utilizadas en el modelo empírico, aplicando la metodología CRISP-DM, la cual, define seis fases. (Chapman et al., 2000)

4.1 Conocimiento del negocio

En esta fase se intenta comprender el problema desde la perspectiva de la minería de datos. Esta etapa, fue abordada desde el inicio mismo del proyecto, al definir que la minería de datos es aplicable sobre la información académica, psicológica y financiera de los estudiantes para abordar el problema de deserción en las instituciones de educación superior.

4.2 Comprensión de los datos

En la fase de comprensión se identifican los posibles problemas de calidad observados en los datos, como campos vacíos, valores nulos o incorrectos. Esta fase incluye la solicitud de los datos a la UNAB y la familiarización con las variables recolectadas.

4.3 Preparación de los datos

Con los problemas identificados en la fase anterior, se seleccionan las variables que serán utilizadas en el análisis, descartando aquellas que no cumplen los criterios de calidad. En esta fase, fueron descartadas algunas variables que no contaban con información suficiente para el análisis.

4.4 Modelado

Para que los datos sean procesados, es necesario que puedan ser interpretados por el software de minería de datos seleccionado. En esta fase, los datos obtenidos se encontraban en formato de archivo Excel (xlsx). Fue necesario realizar proceso de transformación para convertirlos en formato arff, el formato de archivos Weka.

4.5 Evaluación

Los datos previamente preprocesados y modelados, se evalúan a través de la herramienta de minería de datos. De acuerdo a los resultados, puede ser necesario ejecutar nuevamente las fases anteriores, revisando y corrigiendo posibles errores en el proceso. En esta fase, los archivos .arff generados en el modelado, fueron procesados a través de Weka, teniendo en cuenta los resultados, fue necesario definir diferentes estrategias de análisis que permitieran mejorar los porcentajes de predicción. Estas estrategias se describen en la sección 5.

4.6 Despliegue

Esta fase inicia con la evaluación de los resultados y finaliza con la descripción del proceso de implementación, puede incluir también la descripción del proceso de minería de datos realizado y las recomendaciones para trabajos futuros. Esta fase, fue abordada desde el inicio mismo del proceso, documentando cada una de las actividades desarrolladas y finalizando con el análisis de los resultados obtenidos.

5. Análisis

Para aplicar el modelo de minería de datos fueron obtenidos 3370 registros históricos con la información de los estudiantes de primer ingreso a los programas de pregrado desde el año 2014 a 2016. Después de realizar las fases de comprensión y preparación de los datos se seleccionaron 2571 registros, los cuales

se relacionan por periodo académico en la siguiente tabla:

Periodo académico	Desertores	No desertores	Total estudiantes
2014-02	30	242	272
2015-01	71	873	944
2015-02	40	316	356
2016-01	105	894	999
Total	246	2325	2571

Tabla 2. Población distribuida por periodo académico.

Teniendo en cuenta la marcada diferencia entre la cantidad de estudiantes desertores y no desertores y su posible incidencia en los resultados, fue necesario idear diferentes alternativas de análisis que permitiera realizar la evaluación con cantidades mas equilibradas de datos y observar el comportamiento de los porcentajes finales de predicción.

5.1 Análisis General de los Datos

Un primer análisis contempló el procesamiento total de los 2571 registros a través de la opción de testeo Cross validation de Weka, la cual, permite definir un valor porcentual que es utilizado para particionar la totalidad de los datos, definiendo cierta cantidad para el entrenamiento del algoritmo y los restantes para la realización de las pruebas.

Los datos de entrenamiento son utilizados por Weka para identificar patrones en las variables analizadas, con estos patrones, se clasifican los datos de prueba y, de acuerdo a la clasificación realizada, se puede obtener un porcentaje de predicción.

5.2 Análisis Por Periodo Académico

En este análisis, fueron tomados la totalidad de registros de estudiantes desertores por periodo académico, a continuación, se seleccionaron de forma aleatoria un número similar de registros de estudiantes no desertores.

Adicionalmente, se construyeron dos conjuntos de datos, uno de ellos para el entrenamiento del algoritmo y otro para las pruebas. Los datos de entrenamiento y los datos de prueba contaron con el mismo número de registros de estudiantes desertores y no desertores.

La tabla 3 presenta la cantidad de datos seleccionados por periodo académico. Así por ejemplo, en el periodo 2014-02 se tuvieron en cuenta 60 registros en total, 30 de desertores y 30 de no desertores, de los cuales 30 corresponden a datos de entrenamiento y 30 a datos de prueba, cada uno conformado por 15 registros de desertores y 15 registros de no desertores.

Periodo académico	Desertores	No Desertores	Total Periodo
2014-02	30	30	60
2015-01	71	71	142
2015-02	40	40	80
2016-01	105	105	210
Total	246	246	494

Tabla 3. Población analizada por periodo académico

5.3 Análisis Por Periodo Académico Acumulado

A diferencia del análisis anterior, el análisis por periodo acumulado aprovecha el procesamiento histórico de los datos. Como parte de este proceso, se utilizaron los datos históricos como datos de entrenamiento del algoritmo y los datos de prueba fueron tomados del periodo analizado.

Teniendo en cuenta que no se tenían datos anteriores al periodo 2014-02, no fue posible realizar el análisis sobre este periodo académico. En su lugar, sus registros fueron utilizados como datos de entrenamiento para el análisis del periodo 2015-01. A su vez, los registros del periodo 2014-02 y los registros del periodo 2015-01, fueron utilizados como datos de entrenamiento para analizar los registros del periodo 2015-02. Un análisis final adiciona los registros del periodo 2015-02 a los datos de entrenamiento y toma los registros del periodo 2016-01 como datos de prueba.

Periodo académico	Datos de entrenamiento	Datos de Prueba	Total Periodo
2015-01	60	142	202
2015-02	202	80	282
2016-01	282	210	492

Tabla 4. Población seleccionada para el análisis por periodo académico acumulado

De acuerdo con la tabla 4, los 60 registros del periodo 2014-02 son utilizados como datos de entrenamiento para los 142 datos del periodo 2015-01. Luego los 202 registros del periodo 2015-01 son utilizados como datos de entrenamiento para los 80 registros del periodo 2015-02. Finalmente los 282 registros del periodo 2015-02 son utilizados como datos de entrenamiento para los 210 registros del periodo 2016-01.

5.4 Análisis Por Periodo Académico Acumulado sin la Variable Rendimiento Académico

Teniendo en cuenta que la variable rendimiento académico es obtenida por la UNAB una vez el periodo académico ha finalizado y la presunción de la importancia que representa en la deserción, se definió un análisis que no incluyera el rendimiento académico como parte del proceso de predicción.

El objetivo del análisis era permitirle a la UNAB utilizar el modelo en el transcurso del semestre académico, de manera que los planes de retención puedan ser aplicados en tiempo real.

6. Resultados

Teniendo en cuenta los análisis realizados, en este apartado se presentan comparativamente los resultados de predicción obtenidos, teniendo en cuenta no solo el porcentaje general de predicción, sino el porcentaje de estudiantes desertores y no desertores clasificados correcta e incorrectamente.

En el análisis general de los datos, que tuvo en cuenta el total de 2571 registros, Weka no logró reconocer ningún patrón predictivo sobre los datos cargados, a pesar que los registros correctamente clasificados fueron 2325 (90.43%) y los incorrectamente clasificados fueron 246 (9.56%), solo clasificó correctamente los estudiantes no desertores, los estudiantes desertores tuvieron una clasificación nula. La diferencia entre la cantidad de registros de estudiantes no desertores, en relación a la cantidad de registros de estudiantes desertores determinó la calidad de los resultados. Mientras los primeros corresponden a un 90.43%, los segundos solo representan el 9.56% del total de los datos.

Los siguientes análisis solucionaron este inconveniente al utilizar la misma cantidad de registros de estudiantes desertores y no desertores, se esperaba que los porcentajes de predicción fueran mas acertados. La siguiente tabla resume los resultados obtenidos:

Periodo académico	PA	PAA	PAA'
2014-02	96,60	--	--
2015-01	61,11	50,00	48,61
2015-02	65,00	71,25	68,75
2016-01	62,85	69,04	59,52

Tabla 5. Porcentajes de lo resultados generales de clasificación obtenidos en cada análisis.

Donde PA corresponde al análisis por periodo académico individual, PAA al análisis por periodo académico acumulado y PAA' al análisis por periodo académico acumulado sin la variable rendimiento académico.

Se observa que en el periodo 2015-01 el porcentaje de registros clasificados correctamente en el análisis PA(61,11%) fue superior al resultado observado en PAA (50,00%) y PAA' (48,61%). En los periodos 2015-02 y 2016-01 el análisis PAA logró un mejor porcentaje de clasificación, superando hasta por 7 puntos porcentuales los resultados del análisis PA.

Si se tienen en cuenta la cantidad de registros en los datos de entrenamiento en comparación con la cantidad de registros de los datos de prueba en cada uno de los análisis, se observa que se obtienen mejores resultados de clasificación a medida que los primeros son más numerosos que los segundos, esta es la razón por la cual, el análisis por periodo académico acumulado obtiene mejores resultados de clasificación que el análisis por periodo individual en los periodos 2015-02 y 2016-01 donde se cumple esta condición.

Igualmente, se observa que en los resultados obtenidos en los PAA', los porcentajes de clasificación son inferiores en comparación con los resultados del análisis PAA, lo que indica la importancia que tiene esta variable para determinar la deserción de un estudiante.

A continuación, se presenta un segundo análisis de resultados que, a diferencia del análisis anterior, incluye de forma separada los porcentajes de clasificación de registros de estudiantes desertores y no desertores del periodo 2015-01. Hay que tener en cuenta que la suma de los porcentajes individuales no debe ser 100%.

Periodo académico	Tipo de Análisis	No Desertores	Desertores
2015-01	PA	47,22	75,00
	PAA	69,44	30,55
	PAA'	79,16	18,05

Tabla 6. Resultados de clasificación del periodo 2015-01 discriminado por desertores y no desertores.

Los registros de los estudiantes no desertores solo estuvieron por debajo en relación a los registros de estudiantes desertores en el análisis por periodo académico individual (PA). En PAA y PAA', los porcentajes de clasificación de los registros de estudiantes no desertores fueron superiores.
Observando la gráfica 1, que tiene en cuenta el resultado de predicción general del periodo 2015-01 (línea), se observa que el porcentaje de clasificación obtenido en PA es superior a los porcentajes obtenidos en los análisis restantes. Igualmente, se observa que el porcentaje de clasificación de

registros de estudiantes desertores en relación al porcentaje de clasificación de registros de estudiantes no desertores es marcadamente desiguales en cada uno de los análisis realizados.

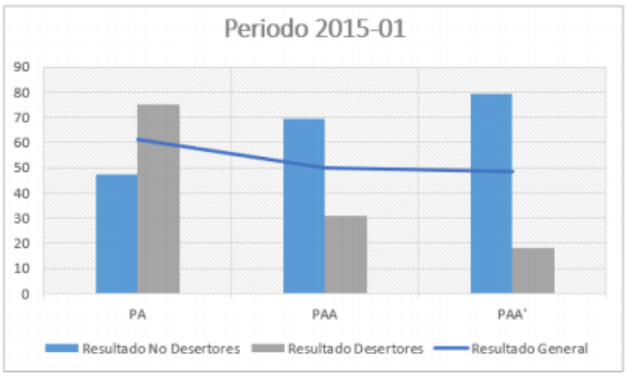

Figura 1. Resultados de clasificación del periodo 2015-01 discriminado por desertores y no desertores. Fuente: elaboración propia

Hay que tener en cuenta que en el periodo 2015-01, solo se tenían los registros del periodo 2014-02 como datos de entrenamiento de los análisis PAA y PAA', por lo tanto, se esperaba que en los siguientes análisis el porcentaje de clasificación fuera más alto en comparación con los resultados de PA. La siguiente tabla presenta la clasificación de los resultados obtenidos en el análisis del periodo 2016-01.

Periodo académico	Tipo de Análisis	No Desertores	Desertores
2016-01	PA	79,24	45,28
	PAA	65,71	72,38
	PAA'	84,76	59,52

Tabla 7. Resultados de clasificación del periodo 2016-01 discriminado por desertores y no desertores.

Se observa que los porcentajes de clasificación, al menos en el análisis PAA, son superiores a los resultados del análisis PA, teniendo en cuenta, que la información histórica en este caso incluye los registros de los periodos 2014-02, 2015-01 y 2015-02. La siguiente gráfica, muestra la tendencia de los resultados:

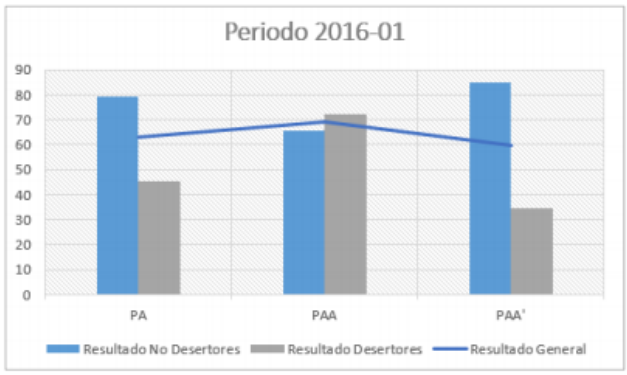

Figura 2. Resultados de clasificación del periodo 2016-01 discriminado por desertores y no desertores. Fuente: elaboración propia

A diferencia de los resultados del análisis PAA, el análisis PAA' marcó una gran diferencia entre el porcentaje de clasificación de los registros de estudiantes desertores y no desertores. Esto confirma, la importancia que tiene la variable rendimiento académico en la determinación de deserción de un estudiante. Igualmente, se mantiene la tendencia de las diferencias marcadas entre los registros de desertores y no desertores clasificados correctamente en los análisis PA y PAA'.

En relación al resultado general (línea), se observa que el porcentaje de clasificación obtenido en PAA es superior a los porcentajes obtenidos en los análisis restantes, siendo el resultado coherente a lo esperado, teniendo que los resultados de los datos de entrenamiento son superiores a los datos de prueba.

Si se tienen en cuenta los resultados de las gráficas 1 y 2, en los análisis PAA, los registros clasificados correctamente para estudiantes no desertores, en el periodo 2015-01, son marcadamente superiores en comparación con los registros clasificados correctamente en los estudiantes desertores, en el periodo 2016-01, donde la cantidad de datos de entrenamiento aumenta, los registros de estudiantes desertores se clasifican porcentualmente mejor que los registros de estudiantes no desertores, lo que indica, que entre mayor sea en número de registros históricos, la clasificación de los estudiantes desertores es mucho más precisa.

El análisis anterior, es igualmente valido si se tienen en cuenta solo los registros de estudiantes desertores en los resultados de los análisis por periodo acumulado (PAA). Entre mayor es el número de registros de entrenamiento, se obtienen mejores resultados de clasificación.

Si se tiene en cuenta solo los resultados de PAA', se observa que el porcentaje de estudiantes desertores clasificados correctamente, desciende drásticamente en comparación con la clasificación realizada en PAA. Contrario a esto, los porcentajes de clasificación de estudiantes no desertores, siempre fue superior en comparación con los demás análisis realizados.

Lo anterior, ratifica que la variable rendimiento académico es la más importante dentro de todas las variables analizadas para determinar la deserción de un estudiante, sin ella, los porcentajes de clasificación de estudiantes desertores desciende drásticamente, sin embargo, la clasificación de estudiantes no desertores, mejora en comparación con los demás análisis.

6.1 Variables determinantes de deserción

Teniendo en cuenta el proceso realizado y los resultados obtenidos a través del procesamiento de los datos en la herramienta Weka con el algoritmo de clasificación J48 en cada uno de los análisis ejecutados, es necesario identificar cuáles de las variables analizadas determinaron en mayor y menor medida la deserción de los estudiantes de primer ingreso a los programas de pregrado. Para ello, se tuvieron en cuenta dos factores:

- Cantidad de veces que la variable fue determinante de deserción en los diferentes análisis realizados.

- Cantidad de estudiantes en los cuales, la variable influyó en su deserción.

Para ello, se analizaron individualmente cada uno de los arboles generados por el algoritmo J48 en cada uno de los análisis realizados. En el árbol, se pueden observar las variables asociadas a la deserción y la cantidad de estudiantes afectados. En la gráfica 3, por ejemplo, se ven involucradas las variables nivel de matemáticas y la variable convive, que indica si el estudiante vive con su familia o fuera del núcleo familiar. A su vez, cada hoja del árbol tiene asociado unos valores que indican el número de desertores y no desertores que fueron influenciados por los valores de la variable.

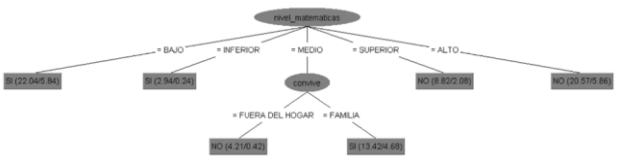

Figura 3. Ejemplo de árbol generado con Weka con el algoritmo J48. Fuente: elaboración propia

De acuerdo a estos criterios y teniendo en cuenta los resultados Weka en los diferentes análisis, se construyó la siguiente tabla:

Nombre Variable	Por Repeticiones				Por Cantidad			
	PA	PAA	PAA'	Total	PA	PAA	PAA'	Total
Rendimiento Académico	**3**	**3**		**6**	**5**	**77**		**82**
Puntaje ICFES	**4**	**2**	**3**	**9**	**40**	**19**	**138**	**197**
Nivel de Matemáticas	**2**	**1**	**2**	**5**	**53**	**7**	**74**	**134**
Nivel Comunicativo		1		1	7			7
¿Con quién vive?	1		2	3	9		37	46
Procedencia	1	1		2	5	9		14
Tipo de Colegio	1			1	6			6
3C	1			1	3			3
Fuente Pago Estudios	1	1		2		5		5
Personas a Cargo		1	1	2		10	5	15
Trabaja		1		1		7		7
Estado Civil			1	1			3	3
12º			1	1			20	20
8I			1	1			26	26
16T			1	1			4	4

Tabla 8. Principales variables asociadas a la deserción

Como se observa, hay variables que, aunque aparecen relacionadas, no representan una fuerte influencia para determinar la deserción de un estudiante, teniendo en cuenta que solo aparecen una sola vez en la totalidad de análisis realizados y además, el número de estudiantes que desertan por su causa es mínimo.

Por otra parte, se tienen las variables de Rendimiento Académico, Puntaje ICFES y Nivel de Matemáticas que son fuertemente identificadas por el análisis de minería de datos como variables altamente determinantes de deserción, teniendo en cuenta, no solo el número de veces que aparecen en los análisis, sino la cantidad de estudiantes que desertaron por su causa, resaltando especialmente el puntaje ICFES, la cual, aparece en 9 del total de 10 análisis individuales realizados y se encuentra asociada con la deserción de 197 estudiantes.

6.2 Análisis comparativo modelo BI vs modelo matemático

Teniendo en cuenta los resultados obtenidos en el Análisis por Periodo Académico Acumulado del periodo 2016-01, se realizó la comparación entre los porcentajes de predicción obtenidos a través del método de minería de datos y el modelo empírico desarrollado por la UNAB para determinar, cuál de ellos presentaba un mayor grado de precisión en relación a los resultados reales deserción.

Para realizar el proceso, se identificaron los registros de estudiantes desertores clasificados por el modelo de minería de datos; estos mismos registros fueron consultados en el modelo empírico y comparados contra los datos reales de deserción. La siguiente tabla resume los resultados obtenidos:

Modelo	TR	CC	CI	PA
BI	210	145	65	69,04
Empírico	210	109	101	51,90

Tabla 9. Comparación de resultados modelo de minería de datos (BI) vs modelo empírico

El modelo de minería de datos supera las predicciones realizadas por el modelo empírico, con 36 registros más acertados sobre una muestra de 210 estudiantes, logrando estar 17.14 puntos porcentuales por encima en el resultado general de deserción. Si se observa el porcentaje de clasificación tanto de desertores como de no desertores, se obtienen los siguientes resultados:

Modelo	DCC	NDCC	PA
BI	72,38	65,71	69,04
Empírico	76,19	27,61	55,71

Tabla 10. Porcentajes de aciertos del modelo de minería de datos (BI) vs modelo empírico

El modelo de minería de datos a diferencia del modelo empírico, realiza una clasificación más equilibrada de registros desertores y no desertores clasificados correctamente, con un 72,38% y un 65,71% respectivamente. La siguiente gráfica, presenta el resultado de la tabla.

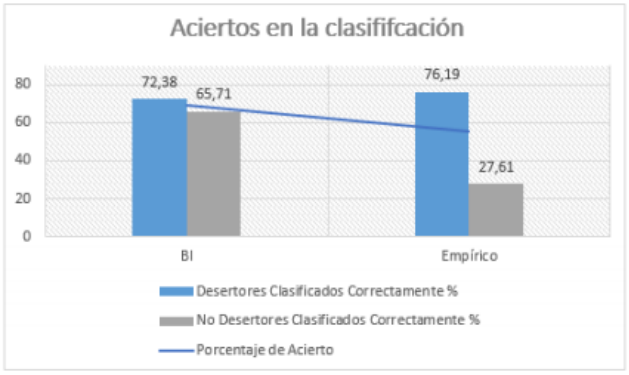

Figura 4. Porcentajes de aciertos del modelo de minería de datos (BI) vs modelo empírico. Fuente: elaboración propia

Si se observa el porcentaje de desertores clasificados correctamente en el modelo empírico (76,19%), se tiene que este es mayor que el porcentaje del modelo de minería de datos (72,38%), sin embargo, si se observa que el porcentaje de no desertores clasificados correctamente (27,61%), se puede ver que en los registros de estudiantes desertores el modelo empírico tiene mayor acierto por que tiende a identificar de forma masiva estudiantes no desertores como desertores, por lo tanto, su porcentaje de error debe ser mucho mayor, esto se comprueba observando los siguientes resultados:

Modelo	DCC	NDCC	PA
BI	34,28	27,61	30,95
Empírico	72,38	23,80	48,09

Tabla 11. Porcentajes de errores del modelo de minería de datos (BI) vs modelo empírico

El resultado de error en los desertores es bastante alto en el modelo empírico (72,38%) en relación al modelo BI (34,28%), ya que clasifica mayormente estudiantes no desertores, como desertores. Las siguiente gráfica muestra la tendencia observada:

Figura 5. Porcentajes de errores del modelo de minería de datos (BI) vs modelo empírico. Fuente: elaboración propia

En los registros de estudiantes no desertores el porcentaje de error es ligeramente superior en el modelo de minería de datos, sin embargo, si se tiene en cuenta que el modelo empírico clasificó acertadamente un número menor de registros de estudiantes no desertores, es normal que tenga un porcentaje de error menor. A nivel general, el modelo de minería de datos obtuvo una tasa de error del 30,95% en relación al 48,09% del modelo empírico, lo que marca una diferencia de 17,14% de diferencia.

7. Conclusiones

En los diferentes análisis de deserción que fueron realizados, se logró atender los diferentes objetivos del proyecto. Iniciando con la implantación de la metodología de minería de datos CRISP-DM y continuando con el primer análisis que incluía el procesamiento masivo de todos los registros de estudiantes caracterizados, en el cual, se obtuvo una clasificación nula de estudiantes desertores, el problema, la desproporción en la cantidad de registros de estudiantes desertores y no desertores, estos últimos, con más del 90% del total de los registros.

Para lograr un porcentaje equilibrado de predicción, fue necesario seleccionar de forma aleatoria, un número de registros de estudiantes no desertores que fuese igual al número de registros de estudiantes desertores, para logra una proporción igual de los datos. Adicionalmente, se definieron registros de entrenamiento y registros de prueba para tener un mayor control sobre las evaluaciones.

Finalmente, se realizaron dos análisis adicionales, uno, con un procesamiento por periodo académico acumulado, que aumentara paulatinamente la cantidad de registros de datos de entrenamiento. En el otro análisis se realizó el mismo proceso, descartando la variable rendimiento académico. De acuerdo a lo anterior, se concluye:

El análisis por periodo académico acumulado, presenta los mejores porcentajes de predicción en relación a los análisis por periodo académico individual y periodo académico acumulado sin la variable rendimiento académico, teniendo en cuenta que, el análisis histórico de los datos brinda mejores oportunidades de predicción y determina la importancia de la variable rendimiento académico como un factor importante dentro del proceso de deserción de un estudiante.

La clasificación de datos de entrenamiento y datos de prueba para realizar los diferentes análisis mejora notablemente los resultados, teniendo en cuenta que, el algoritmo recibe un entrenamiento previo y controlado, antes de realizar el proceso de clasificación.

El análisis histórico de los periodos académicos, permite no solo una predicción más acertada, sino que además, los porcentajes de clasificación de registros de estudiantes desertores y no desertores son más equilibrados.

Las variables “Rendimiento académico”, “Puntaje ICFES” y “Nivel de Matemáticas” son determinantes en el proceso de deserción de los estudiantes, siendo las más repetitivas y con mayor número de desertores en los diferentes análisis realizados.

El análisis por periodo académico acumulado obtiene mejores porcentajes de clasificación en relación al método empírico desarrollado por la UNAB, teniendo en cuenta, no solo el análisis histórico y dinámico de la información, sino la relación que existe entre diferentes grupos de variables. El modelo empírico, está limitado a analizar individualmente las variables a través de un valor porcentual de acuerdo al nivel de riesgo que representan, sin embargo, como se observa en los resultados obtenidos, la deserción no siempre está determinada por una única variable, por el contrario, diferentes grupos de variables asociadas a través de los diferentes valores que estas aceptan, en relación a los patrones históricos, determinan la deserción.

Bibliografía

Berry, M. y Linoff, G. (2004). Data Mining Techniques: For Marketing, Sales and Customer Relationship Management. Indiana, Estados Unidos: Wiley Publishing.

Chapman, P. y Clinton, J. y Kerber, R. y Khabaza, T. y Reinartz, T. y Shearer, C. (2000). CRISP-DM 1.0, International Business Machines IBM. Disponible en ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

Han, J. Y Kamber, M. y Pei, J. (2011). Data mining: Concepts and techniques. Amsterdam, Holanda: Morgan Kaufmann Publishers.

Hernández, J. y Ramírez, M. y Ferri, C. (2004) Introducción a la Minería de Datos. México: Pearson Educación.

Ministerio de Educación Nacional. (2009). Deserción estudiantil en la educación superior colombiana. *Bogotá*, Colombia: Imprenta nacional de Colombia.

Ministerio de Educación Nacional. (2016). Cómo va la graduación y deserción en la educación superior. Recuperado de: https://www.mineducacion.gov.co/1759/articles-359642_recurso.pdf

Vercellis, C. (2009). Business intelligence: Data mining and optimization for decision making. Estados Unidos: Wiley, John & Sons.

Witten, I. y Frank, E. (2005). Data mining: Practical machine learning tools and techniques. San Francisco, Estados Unidos: Morgan Kaufmann Publishers.

ANEXO 1: Variables académicas

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
Tiempo transcurrido entre la última graduación y su ingreso a la UNAB.	3	5	0-2 Años	TTG	0%
			2-4 Años		2%
			4-6 Años		4%
			Más de 6 años		5%
Población Especial	3	5	Discapacidad motriz	PBEM	5%
			Discapacidad cognitiva	PBEC	6%
			Discapacidad sensorial	PBES	5%
			Desplazado	PBED	5%
			Afrodescendiente	PBE	5%
			Indígenas		
			Habitante de Frontera		
			Ser Pilo Paga	PSPP	5%
Estrategias de Aprendizaje	3	24	Metacognitiva	EAP	6%
			Procesamiento de Información		6%
			Manejo de Recursos		6%
			Disposicionales		6%
Nivel Obtenido en la Competencia Comunicativa	3	12	Superior (Nota 4.0 – 5.0)	NCC	0%
			Alto (Nota 3.0 - 3.9)		0%
			Medio (Nota: 2.0 – 2.9)		2%
			Bajo (Nota: 1.0 – 1.9)		4%
			Inferior (Nota: 0.0 – 0.9)		6%
Nivel Obtenido en la Competencia Matemática	3	12	Superior (Nota: 4.0 – 5.0)	NCM	0%
			Alto (Nota: 3.0 - 3.9)		0%
			Medio (Nota: 2.0 – 2.9)		2%
			Bajo (Nota: 1.0 – 1.9)		4%
			Inferior (Nota: 0.0 – 0.9)		6%
Puntaje ICFES	3	6	Más de 50 Puntos	PIC	0%
			40 – 50 Puntos		4%
			0 – 40 Puntos		6%
Porcentaje de Cursos Aprobados	3	5	0% ≤ Cursos Aprobados < 20%	PCA	5%
			20% ≤ Cursos Aprobados < 50%		4%
			50% ≤ Cursos Aprobados < 70%		2%
			70% ≤ Cursos Aprobados < 90%		1%
			90% ≤ Cursos Aprobados < 100%		0%
Porcentaje de insistencia	3	6	0% – 10%	PIN	0%
			11% - 20%		2%
			21% - 50%		4%
			51% - 100%		6%

ANEXO 2: Variables Financieras

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
Estado Civil	1	30	Casado	ESC	30%
			Separado – Divorciado		15%
			Viudo		15%
			Unión Libre		30%
			Soltero		0%
Con Quien Vive	1	30	Familia	CQV	0%
			Fuera del hogar		30%
Tipo de Vinculación Laboral	1	20	Independiente	TVL	20%
			Empleado		20%
Cargo Que Ocupa	1	20	Directivo	COC	0%
			Jefatura		0%
			Coordinador		0%
			Supervisor		20%
			Operario		20%
Lugar de Procedencia	2	20	Área metropolitana de Bucaramanga	PRC	0%
			Fuera del área metropolitana de Bucaramanga		20%
Tiene Personas a Cargo	2	20	Si	TPC	20%
			No		0%
Sector Productivo	2	10	Servicios	SPR	10%
			Industriales		10%
			Manufactura		10%
Salario	2	15	Menos de un salario mínimo	SAL	15%
			Entre 1 y 2 salarios mínimos		10%
			3 salarios mínimos o más		0%
Horario Laboral	2	15	Horario de oficina	HLB	5%
			Turnos		15%
			Nocturno		15%
			Fines de semana		15%
Tipo de Colegio	2	20	Público	TCO	20%
			Privado		5%
Trabaja Actualmente	3	20	Si	TRB	20%
			No		0%
Número de Hijos	3	10	Ninguno	NMH	0%
			Uno		2%
			Dos		5%
			Tres o más		10%
Estrato	3	50	1	EST	50%

			2		50%
			3		30%
			4		0%
			5		0%
			6		0%
Fuente de Pago de Estudios	3	20	Recursos Propios	FPE	0%
			ICETEX		20%
			Otra entidad		20%

ANEXO 3: Variables Psicológicas

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
Rango de Personalidad IC	3	70	Personalidad indicador clínico - 1A	1A	1%
			Personalidad indicador clínico – 2B	2B	7%
			Personalidad indicador clínico - 3C	3C	7%
			Personalidad indicador clínico – 4E	4E	1%
			Personalidad indicador clínico – 5F	5F	1%
			Personalidad indicador clínico – 6G	6G	7%
			Personalidad indicador clínico – 7H	7H	1%
			Personalidad indicador clínico – 8I	8I	7%
			Personalidad indicador clínico – 9L	9L	1%
			Personalidad indicador clínico – 10M	10M	1%
			Personalidad indicador clínico – 11N	11N	1%
			Personalidad indicador clínico – 12O	12O	7%
			Personalidad indicador clínico – 13Q	13Q	1%
			Personalidad indicador clínico – 14R	14R	1%
			Personalidad indicador clínico – 15S	15S	1%
			Personalidad indicador clínico – 16T	16T	7%
			Personalidad indicador clínico – I1	I1	3%
			Personalidad indicador clínico – I2	I2	1%
			Personalidad indicador clínico – I3	I3	3%
			Personalidad indicador clínico – I4	I4	1%
			Personalidad indicador clínico – I5	I5	1%
			Personalidad indicador clínico – NRT	NRT	3%
			Personalidad indicador clínico – PST	PST	3%
			Personalidad indicador clínico - PPA	PPA	3%
Riesgo Psicosocial	3	30	Psico-transtornos del comportamiento	TC	3%
			Psico-riesgos para la salud	RS	3%
			Psico-desordenes psiquiátricos	DP	3%
			Psico-autoafirmación social	AS	3%
			Psico-disfunciones familiares	DF	3%
			Psico-rendimiento escolar	RE	3%
			Psico-desordenes laborales	DL	3%
			Psico-presión del grupo de amigos	PGA	3%
			Psico-uso del tiempo libre	UTL	3%
			Psico-severidad del uso de drogas	SUD	3%