MEJORAR EL MODELO DE ESTIMACIÓN DE RIESGO DE DESERCIÓN DE LOS ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA EMPLEANDO HERRAMIENTAS BUSINESS INTELLIGENCE SOPORTADAS EN SOFTWARE LIBRE

MIGUEL EDUARDO ROPERO SILVA

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA – UNAB
FACULTAD DE INGENIERÍA
GRUPO DE INVESTIGACIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN
MAESTRÍA EN SOFTWARE LIBRE
BUCARAMANGA
2018

MEJORAR EL MODELO DE ESTIMACIÓN DE RIESGO DE DESERCIÓN DE LOS ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA EMPLEANDO HERRAMIENTAS BUSINESS INTELLIGENCE SOPORTADAS EN SOFTWARE LIBRE

MIGUEL EDUARDO ROPERO SILVA

Proyecto de grado presentado como requisito para optar al título de Magíster en Software Libre

Director:
SILVIO RAFAEL CUELLO DE ÁVILA

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA – UNAB
FACULTAD DE INGENIERÍA
GRUPO DE INVESTIGACIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN
MAESTRÍA EN SOFTWARE LIBRE
BUCARAMANGA
2018

RESUMEN

El documento “MEJORAR EL MODELO DE ESTIMACIÓN DE RIESGO DE DESERCIÓN DE LOS ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA EMPLEANDO HERRAMIENTAS BUSINESS INTELLIGENCE SOPORTADAS EN SOFTWARE LIBRE” creado por Miguel Eduardo Ropero Silva para obtener el título de Magister en SOFTWARE LIBRE, presenta los resultados de la aplicación de técnicas de minería de datos para predecir la deserción académica con el uso del software Weka y el algoritmo de clasificación J48.

En la investigación se obtuvieron los datos académicos, financieros y psicológicos de los estudiantes de primer ingreso en los programas de pregrado de los periodos 2014-02, 2015-01, 2015-02 y 2016-01, los cuales, fueron analizados, modelados y evaluados a través de la metodología de minería de datos CRISP-DM con la ayuda del software de distribución libre Weka y el algoritmo de clasificación de Arboles de Decisión J48.

Los registros obtenidos también fueron evaluados a través del modelo matemático (empírico) de deserción desarrollado por la UNAB. Los resultados de deserción del modelo de minería de datos y del modelo empírico, fueron cruzados contra los resultados reales de deserción para determinar cuál de ellos ofrecía un mayor porcentaje de precisión. Finalmente, fueron identificadas las principales variables involucradas en la deserción de un estudiante de acuerdo a la información previamente analizada.

Palabras clave: Minería de datos, algoritmo de clasificación J48, Arboles de decisión, Weka.

CONTENIDO

Pág.

LISTA DE GRÁFICAS

Pág.

LISTA DE TABLAS

Pág.

INTRODUCCIÓN

La Universidad Autónoma de Bucaramanga dentro de las estrategias desarrolladas por el departamento de Bienestar Universitario para la reducción de las tasas de deserción, desarrolló un modelo matemático que calcula el nivel de riego asociado a sus estudiantes de primer semestre, en base a la información académica, psicológica y financiera almacenada en sus sistemas de información institucionales. Cada de una de las variables tiene asignado un peso dependiendo del nivel de riesgo que representa.

Teniendo en cuenta el notable crecimiento de las estrategias y herramientas para análisis de grandes volúmenes de datos y su aplicación en la identificación de patrones y tendencias útiles en la predicción de eventos futuros, se identificó una oportunidad para hacer uso de las herramientas existentes aprovechando la información histórica de los estudiantes de primer ingreso a los programas de pregrado con el objetivo de mejorar los porcentajes de predicción del modelo matemático existente.

De acuerdo a lo anterior, el presente documento desarrolla un modelo de minería de datos que permite clasificar a un estudiante como desertor o no desertor en base a su información académica, psicológica y financiera y a los registros históricos de estudiantes de primer semestre de los periodos 2014-02 al 2016-01.

Para la aplicación del modelo, se desarrolló el proceso sugerido por la metodología CRISP-DM, la cual, define el proceso para el desarrollo de proyectos de minería de datos. Un primer paso permite comprender del modelo matemático de la UNAB, a través del conocimiento de las variables involucradas en el análisis de deserción y las fórmulas matemáticas empleadas (Sección 5).

Más adelante, en el marco teórico, se presentan los conceptos asociados a la minería de datos y las diferentes metodologías existentes para su aplicación. Se describen los diferentes modelos, métodos, técnicas, algoritmos y herramientas software, incluyendo Weka, el cual fue utilizado para el procesamiento de la información a través del algoritmo J48. En la sección 7 (Marco metodológico) se describen los diferentes análisis definidos para la aplicación del modelo, de acuerdo a la información de los estudiantes obtenida.

Habiendo definido los análisis de la información, en la sección 8 (Resultados), se describe paso a paso el proceso desarrollado, desde la elección del software de minería de datos, su instalación en sistemas Windows y Linux, la presentación de la interfaz gráfica y el detallado de los pasos de la metodología CRISP-DM (Conocimiento del negocio, comprensión de los datos, preparación de los datos, modelado y análisis).

Dentro del proceso de análisis descrito en el documento, se procesa la información y se obtienen los resultados en el análisis general de los datos, análisis por periodo académico, análisis por periodo académico acumulado (aprovecha la información histórica) y análisis por periodo académico acumulado sin la variable rendimiento académico.

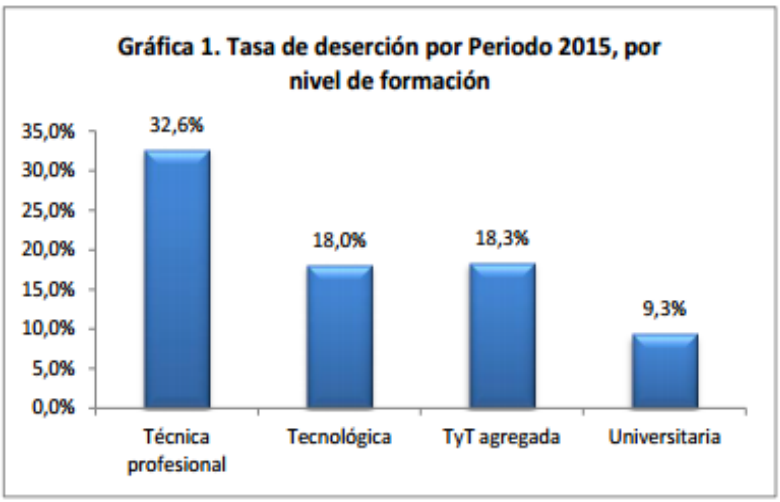
Finalmente, se presentan los resultados de predicción obtenidos en cada uno de los análisis realizados. Estos resultados, son comparados con los resultados del modelo matemático, identificando los porcentajes de precisión de cada uno de ellos y las principales variables asociadas a la deserción de los estudiantes. De acuerdo con los resultados, se presentan las respectivas conclusiones del proceso de minería de datos.

1. PLANTEAMIENTO DEL PROBLEMA

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 Académica, que según el Ministerio de Educación Nacional (2010), 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.

La deserción en Colombia es medida desde dos perspectivas diferentes: La deserción por periodo y deserción por cohorte (Ministerio de Educación Nacional, 2010). 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, como se observa en la figura 1.

Figura 1. Tasa de deserción por periodo, año 2015, por niveles de formación



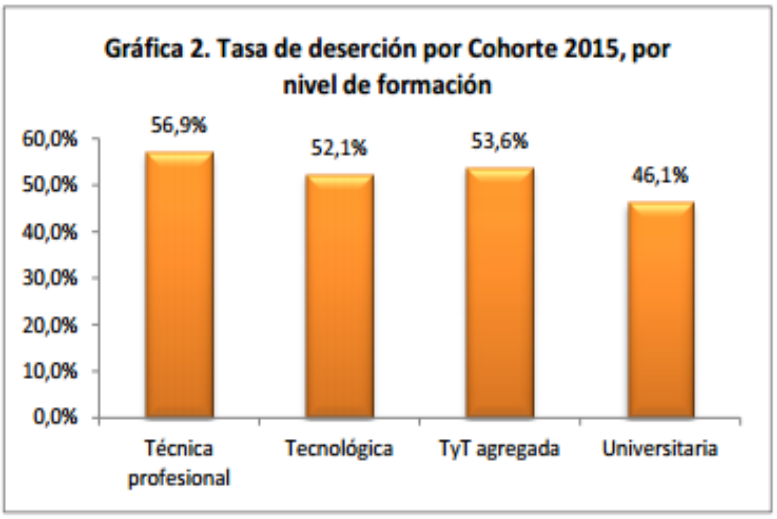
Fuente Ministerio de Educación Nacional (2016). Disponible en: http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-350629_estadisticas_pdf2015.pdf

La deserción por periodo tiene en cuenta la proporción de estudiantes que se matricularon en una institución de educación superior, y no se reportaron

matriculados para el inicio del siguiente año (reporta dos semestres sin actividad académica).

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, (2016) en este indicador Colombia registra una tasa de deserción del 46.1% en el nivel de formación universitaria, como se observa en la figura 2.

Figura 2. Tasa de deserción por cohorte, año 2015, por niveles de formación



Fuente: Ministerio de Educación Nacional (2016). Disponible en: http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-350629_estadisticas_ pdf2015.pdf

De acuerdo con la información, en Colombia por cada 100 estudiantes que se matriculan en una institución de Educación Superior (IES) para el nivel de formación universitaria 54, de ellos continúan con su formación, mientras que 46 de ellos abandonan sus estudios. Desde una perspectiva general, 1 de cada dos estudiantes abandonan las IES.

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 de diferentes programas de retención.

La Universidad Autónoma de Bucaramanga UNAB, es una institución dedicada al servicio de la Educación Superior, de carácter privado, con acreditación institucional de alta calidad, que se consolida hoy en día como la primera universidad del nororiente colombiano (Universidad Autónoma de Bucaramanga, 2016a).

La UNAB dentro de los servicios de bienestar dirigidos a la comunidad estudiantil que permiten ofrecer una formación integral, mejoramiento de la calidad de vida y construcción de comunidad, a cargo de la dependencia de Bienestar Universitario, ha desarrollado diferentes estrategias que le permiten elaborar y aplicar planes encaminados a identificar y retener a aquellos estudiantes con riesgo de deserción a través del análisis de su información académica, financiera y psicológica (Universidad Autónoma de Bucaramanga, 2016b).

Dentro de las últimas estrategias aplicadas para facilitar la retención de los estudiantes, se encuentra la realización de un modelo matemático desarrollado por el departamento de bienestar universitario, en base la experiencia adquirida por su personal. Este modelo, obtiene la información académica, psicológica y financiera de los estudiantes, la cual, clasifica en diferentes variables, cada una de ellas con un peso determinado. Al final, aplica una operación matemática para obtener el nivel de riesgo de cada estudiante.

Otro esfuerzo destacable es la investigación “Incorporación de elementos de Inteligencia de Negocios al análisis de deserción estudiantil en la Universidad Autónoma de Bucaramanga UNAB” realizada en el Año 2015, en la cual, se incorporaron elementos de inteligencia de negocios en bases de datos estatales y propias de la UNAB, para genera un Datamart cuyo propósito fue servir como repositorio de datos al cual se le puedan aplicar técnicas OLAP y de minería de datos para generar conocimiento estadístico, con el objetivo de orientar estrategias para disminuir la deserción de estudiantes en la Institución (Castellanos Guarín, 2015).

Las técnicas OLAP facilitan el procesamiento de grandes volúmenes de información de forma ágil y permiten generar informes que muestran generalmente patrones, tendencias o comportamientos que sirven para apoyar la toma decisiones.

La Inteligencia de Negocios abarca un gran abanico de oportunidades a través de la aplicación de diferentes herramientas y procedimientos, uno de ellos es la minería de datos, que cuenta con diferentes técnicas, cada una de ellas con diferentes algoritmos que permiten, más allá de definir patrones y tendencias, predecir futuros comportamientos, entre ellos, la oportunidad de identificar de forma temprana aquellos estudiantes con un alto riesgo de deserción.

2. JUSTIFICACIÓN

El Ministerio de Educación Nacional, (2015b) en su “Guía para la implementación del modelo de gestión de permanencia y graduación estudiantil en instituciones de educación superior”, define como uno de sus componentes la Cultura de la Información, en la cual, resalta la importancia del acceso, disponibilidad, calidad y uso de la información para la toma de decisiones en diferentes ámbitos operativos, tácticos y estratégicos en las instituciones de educación superior.

La inteligencia de Negocios desde hace varios años ha sido aplicada con éxito en diferentes campos como la economía de valores, las ventas y el sector bancario. La minería de datos como parte de las estrategias de la inteligencia de negocios recientemente viene aplicándose en el sector educativo con el nombre de *Educational Data Mining* (EDM) o Minería de Datos para la Educación. A pesar de ser un campo de aplicación bastante joven, ha presentado importantes avances en los últimos años, a través de las diferentes herramientas y tecnologías que permiten obtener resultados positivos.

La minería de datos cuenta con diferentes técnicas que permiten realizar predicciones a través de un conjunto de datos de entrada. Cada técnica predictiva cuenta con una variedad de algoritmos que de acuerdo a los objetivos que se persiguen y la calidad de los datos, permiten obtener diferentes niveles de predicción.

Grandes empresas del sector tecnológico, como IBM, Oracle y Microsoft han apostado por desarrollar sus propias suites de procesamiento de datos atendiendo la demanda actual de diferentes organizaciones para aprovechar los grandes volúmenes información que almacenan sus bases de datos, de manera que se puedan explotar de forma eficiente para la obtención de conocimiento que apoye la toma de decisiones administrativas.

Desde la perspectiva del Software Libre, son varias las herramientas tecnológicas disponibles como WEKA, RapidMiner y KNIME desarrolladas y apoyadas por destacadas universidades u organizaciones y mantenidas por grandes comunidades de usuarios. Estas herramientas permiten aplicar de forma eficiente diferentes algoritmos de predicción para cada una de las técnicas que son soportadas.

La aplicación de la Minería de Datos en conjunto con las diferentes herramientas de software libre disponibles y los datos financieros, económicos y psicológicos

de los estudiantes, permiten, 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 en los estudiantes de primer ingreso de los programas de pregrado en la UNAB.

La dependencia de Bienestar Universitario de acuerdo a los resultados obtenidos, tendría la posibilidad de identificar de forma temprana, aquellos estudiantes en riesgo de desertar, de manera que estas estrategias de acompañamiento puedan ser aplicadas en tiempo real, en el transcurso mismo del semestre académico, contribuyendo no solo a reducir los porcentajes de deserción sino a disminuir los esfuerzos en el desarrollo de estrategias de apoyo para aumentar el nivel de retención de estudiantes.

De aplicarse con éxito en la identificación de estudiantes con riesgo de deserción, la minería de datos puede ser utilizada en diferentes procesos administrativos y académicos, en los cuales se cuente con información almacenada en las bases de datos institucionales.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Evaluar y mejorar el modelo estimación de riesgo de deserción de los estudiantes de pregrado modalidad presencial de la Universidad Autónoma de Bucaramanga empleando herramientas de *Business Intelligence* (BI) que empleen modelos estadísticos y predictivos.

3.2 OBJETIVOS ESPECÍFICOS

Analizar el modelo y las variables de riesgos del modelo actual (académicas, financieras y psicológicas) que afectan la permanencia de los estudiantes universitarios de primer ingreso de pregrado en sus respectivos programas académicos.

Proponer un modelo ajustado a partir de la extracción de conocimiento de la información de tres semestres académicos empleando herramientas de BI fundamentadas a través de técnicas de minería de datos que empleen algoritmos de árboles de decisión.

Comparar la precisión del modelo empírico desarrollado por la universidad y el modelo desarrollado con las herramientas de BI empleando algoritmos de predicción con la información real de deserción de un periodo diferente.

4. RESULTADOS ESPERADOS

Con el desarrollo del proyecto se espera identificar el nivel de precisión de las herramientas software utilizadas y sus algoritmos de predicción, cruzando la información de deserción obtenida en los respectivos análisis contra la información real de deserción para un determinado periodo académico.

Se espera identificar aquellas variables que de acuerdo a los análisis realizados son determinantes en la deserción de un estudiante y la relación que estas tienen con otras variables.

Obtener un modelo de predicción que funcione como una herramienta de apoyo que permita a la universidad adelantar programas de retención estudiantil de forma temprana sobre aquellos estudiantes que presentan un alto riesgo de deserción.

5. ANTECEDENTES

La Minería de Datos en la actualidad no deja de ser aún un campo de estudio en desarrollo, sin embargo, ya se observan grandes avances en diferentes áreas de conocimiento, tales como la medicina, las finanzas, el mercadeo, comportamiento humano o la educación. Este último conocido como Minería de Datos para la Educación o *Educational Data Mining* (EDM), que es descrito por la *International Educational Data Mining Society* (s.f.) como: "Una disciplina emergente, preocupada por el desarrollo de métodos para la exploración de tipos de datos únicos que provienen de entornos educativos, y el uso de estos métodos para entender mejor a los estudiantes, y los escenarios en los cuales ellos aprenden".

Esta organización desde el año 2004 inició una serie de talleres que tenían como tema principal el análisis de datos en la educación. Debido a su éxito a partir del año 2008 se creó la Conferencia Internacional en Minería de Datos para la Educación, que hasta el día de hoy se realiza con éxito cada año alrededor del mundo.

Sin embargo, no es la única organización que promueve el *Data Mining* en la educación. *Society for Learning Analytics Research* – SoLAR (2014) "es una red interdisciplinaria que reúne los principales investigadores internacionales que exploran el papel y el impacto de la analítica en la enseñanza, el aprendizaje la formación y el desarrollo". SOLAR desarrolla anualmente y desde el año 2011 la Conferencia Internacional de Analítica de Aprendizaje y Conocimiento LAK (*Learning Analitycs and Knowledge*) en colaboración con la *Association For Computing Machinery* (ACM).

El esfuerzo realizado hasta ahora por estas organizaciones y por diferentes investigadores en todo el mundo ha permitido la realización de diferentes estudios que abordan este tema, sin embargo, EDM avanza a un ritmo lento en comparación con otros campos de aplicación de la Minería de Datos, debido a que hasta ahora las instituciones educativas de diferentes niveles de formación como escuelas, colegios y universidades empiezan a conocer la importancia que tienen sus bases de datos para soportar la toma de decisiones en sus procesos administrativos.

Es una realidad que en la actualidad las instituciones educativas, en especial las instituciones de educación superior, apoyan sus procesos académicos y administrativos en diferentes herramientas software que les permiten almacenar grandes cantidades de información histórica del comportamiento académico de

sus estudiantes, desde el momento de ingreso hasta la desvinculación de la institución, ya sea por la terminación total de sus estudios o por algún factor de deserción académica.

Uno de los procesos en los cuales la minería de datos puede brindar apoyo a las instituciones educativas es en la identificación temprana de aquellos estudiantes con riesgo de deserción. Como lo describe Timaran Pereira, (2009) en su estudio Una Lectura Sobre Deserción Universitaria en Estudiantes de Pregrado Desde la Perspectiva de la Minería de Datos: “A través de técnicas de minería de datos aplicadas a los datos históricos almacenados en las bases de datos de una institución de educación superior (IES), es posible predecir las características del estudiante que va a abandonar sus estudios o predecir quiénes están propensos a desertar”. El determinar perfiles de deserción permite predecir qué estudiantes son los que desertan. Teniendo estos perfiles la IES puede tomar acciones anticipadas que le permitan disminuir el índice de deserción.

Este estudio fue una de las primeras apuestas por la minería de datos en el sector educativo en el ámbito Nacional, desarrollado en la Universidad de Nariño de la ciudad de Pasto, el cual, más que predecir posibles deserciones, buscaba identificar las características comunes o patrones de aquellos estudiantes que desertaron de sus programas haciendo uso de sus datos académicos y personales. Un factor interesante es la utilización del software de distribución libre TariyKDD desarrollado por la Facultad de Ingeniería de la misma Universidad, aplicando una metodología que define las siguientes etapas:

- Etapa de selección
  Donde además de la selección de la base de datos interna de la universidad, hicieron uso de una base de datos proporcionada por el Ministerio de Educación con la información de los diferentes colegios de los cuales eran egresados sus estudiantes universitarios.

- Etapa de Preprocesamiento de datos
  Donde elimina aquellos datos con valores nulos o anómalos y aplican diferentes técnicas para obtener aquellos datos faltantes, como la media o la moda.

- Etapa de transformación de datos
  Transforman los datos de acuerdo a sus necesidades, ya sea en numéricos o caracteres y seleccionan aquellas variables de mayor importancia.

- Etapa de minería de datos
  Donde generan las reglas de clasificación a través del algoritmo C4.5 y reglas de asociación con el algoritmo EquipAsso, disponibles en la herramienta TariyKDD.

- Etapa de interpretación y evaluación de resultados:
  De acuerdo a las técnicas de asociación y clasificación empleadas se identifican los patrones de deserción y bajo rendimientos de los estudiantes.

Para la realización de estudio fueron procesados datos históricos de los últimos 18 años, factor que jugó en contra de sus resultados, debido a varios factores que ellos mismos concluyen, como la mala calidad de los datos, teniendo en cuenta que para una base de datos tan histórica es difícil controlar valores nulos, faltantes o redundantes, además, la información académica y personal de los estudiante que se captura actualmente no es la misma que se tuvo en cuenta en un principio por la institución.

El estudio de Azoumana (2013) llamado “Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos” mejora ciertos aspectos del estudio anterior teniendo en cuenta que usa un histórico de datos mucho menor que resulta más fácil de controlar y utiliza diversos algoritmos para corregir aquellos valores nulos, erróneos o desfasados de sus datos de entrada. Azoumana desde una misma perspectiva intenta solucionar un mismo problema de identificación de patrones de deserción en un contexto de espacio tiempo diferente, a partir de la información personal y académica de los estudiantes de Ingeniería de Sistemas de los periodos comprendidos entre el año 2007 y 2012, con la herramienta software de distribución libre WEKA y la aplicación del módulo de Clasificación el cual permite seleccionar diferentes algoritmos como el ID3 y C4.5 de la técnica de árboles decisión, sin embargo, el estudio no especifica el algoritmo utilizado para la obtención de sus resultados. En cuanto a la metodología utilizada el estudio se basa en la propuesta de Cabena (1998) describiendo las siguientes etapas:

- Selección: Permite seleccionar las fuentes de información

- Reprocesamiento: Para realizar limpieza de datos (Valores nulos, erróneos o desfasados)

- Conversión de datos en un modelo analítico: Para que los datos puedan ser procesados por las herramientas y algoritmos seleccionados

- Interpretación de los resultados obtenidos: De acuerdo a la salida de las herramientas utilizadas

- Aplicación del conocimiento descubierto

Aunque este estudio controla en gran medida la información de entrada que será procesada a través de las diferentes técnicas de minería de datos falla en un aspecto muy importante, no clasifica adecuadamente las variables definidas como posibles causas de deserción. Las variables seleccionadas fueron las siguientes:

- Pérdida de semestre.
- Dificultad financiera
- Ingreso al mercado laboral.
- Otros intereses atraen al estudiante.
- Indeterminado

Se observa que la última variable seleccionada como parte del estudio es "deserción por causa indeterminada", representa un valor ambiguo que finalmente no puede aportar información precisa de una causa observable de deserción. Desafortunadamente para la investigación, fue esta variable la que los algoritmos aplicados determinaron como la principal causa de deserción de los estudiantes durante los periodos evaluados. A pesar de los resultados es importante destacar el proceso aplicado por Azoumana en función de la metodología aplicada, la utilización del software de distribución libre Weka y los criterios de selección de la técnica de minería de datos.

En el ámbito local uno de los estudios más reciente en este tema es el de Castellanos Guarín. (2015) desarrollado en la Universidad Autónoma de Bucaramanga con el título "Incorporación de Elementos de Inteligencia de Negocios al Análisis de Deserción Estudiantil de la universidad Autónoma de Bucaramanga", a través de información de diferentes bases de datos externas a la institución como la BD del SNIES (Sistema nacional de información de la educación superior) con la información de los estudiantes matriculados de pregrado, la BD del ICFES (Instituto Colombiano de Fomento a la Educación Superior) con la información de los resultados de aplicación de los exámenes de estado para el acceso a la educación superior y la BD del SPADIES (Sistema para la prevención de la deserción de la educación superior) con la información de los promedios de deserción de las diferentes instituciones educativas del país.

La investigación utilizó la información académica, financiera y psicológica de los estudiantes de primer ingreso de los programas de pregrado de la institución. Esta información en conjunto fue utilizada para la creación de un Datamart que luego fue procesado por la herramienta de distribución libre PENTAHO, permitiendo la generación de informes de acuerdo a las variables de deserción seleccionadas.

El estudio, aplica la metodología CRISP-DM como parte de su proceso de generación de conocimiento y confronta los resultados obtenidos con los resultados del modelo matemático que aplica la institución, concluyendo que las políticas y estrategias que utiliza la UNAB han sido efectivos hasta ahora para lograr la máxima retención de estudiantes en sus respectivos programas académicos.

La UNAB dentro de sus estrategias para mitigar el porcentaje de deserción de sus estudiantes ha ideado, a partir de la experiencia del personal del departamento de recurso humano, su propio método matemático para identificar aquellos estudiantes con riesgo deserción.

A partir de la información almacenada en su Sistema de Gestión de Información Académica SIGA, procesa diferentes variables pertenecientes al ámbito Académico, Financiero y Psicológico de sus estudiantes.

Inicialmente, para cada una de las variables identificadas, asocia cierto nivel de riesgo, el cual, puede ser alto, medio o bajo, como se observa en la siguiente imagen:

Figura 3. Niveles de riesgo, Modelo matemático UNAB

Fuente: Castellanos Guarín (2015)

Variables Riesgo Académico

Dentro del ámbito académico se identifican las ocho variables descritas en la tabla 1 y tabla 2.

Tabla 1. Variables Académicas Parte 1 – Modelo empírico UNAB

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	PBPP	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%

Fuente: Castellanos Guarín (2015)

Tabla 2. Variables Académicas Parte 2 – Modelo empírico UNAB

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
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%

Fuente: Castellanos Guarín (2015)

Variables Riesgo Financiero

Para el ámbito financiero se definen las siguientes variables:

Tabla 3. Variables Financieras Parte 1 – Modelo empírico UNAB

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%

Fuente: Castellanos Guarín (2015)

Tabla 4. Variables Financieras Parte 2 – Modelo empírico UNAB

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
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%

Fuente: Castellanos Guarín (2015)

Tabla 4. Variables Financieras Parte 3 – Modelo empírico UNAB (continuación)

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
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%

Fuente: Castellanos Guarín (2015)

Variables Riesgo Psicológico

Finalmente, para el ámbito psicológico se definieron las variables descritas en las tablas 5 y 6:

Tabla 5. Variables Psicológicas Parte 1 – Modelo empírico UNAB

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
Rango de Personalidad IC	3	70	Personalidad indicador clínico - 1A	TTG	1%
			Personalidad indicador clínico – 2B		7%
			Personalidad indicador clínico - 3C		7%
			Personalidad indicador clínico – 4E		1%
			Personalidad indicador clínico – 5F		1%
			Personalidad indicador clínico – 6G		7%
			Personalidad indicador clínico – 7H		1%
			Personalidad indicador clínico – 8I		7%

Fuente: Castellanos Guarín (2015)

Tabla 6. Variables Psicológicas Parte 2 – Modelo empírico UNAB

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
Rango de Personalidad IC	3	70	Personalidad indicador clínico – 9L		1%
			Personalidad indicador clínico – 10M		1%
			Personalidad indicador clínico – 11N		1%
			Personalidad indicador clínico – 12O		7%
			Personalidad indicador clínico – 13Q		1%
			Personalidad indicador clínico – 14R		1%
			Personalidad indicador clínico – 15S		1%
			Personalidad indicador clínico – 16T		7%
			Personalidad indicador clínico – I1		3%
			Personalidad indicador clínico – I2		1%
			Personalidad indicador clínico – I3		3%
			Personalidad indicador clínico – I4		1%
			Personalidad indicador clínico – I5		1%
			Personalidad indicador clínico – NRT		3%
			Personalidad indicador clínico – PST		3%
			Personalidad indicador clínico - PPA		3%
Riesgo Psicosocial	3	30	Psico-transtornos del comportamiento		3%
			Psico-riesgos para la salud		3%
			Psico-desordenes psiquiátricos		3%
			Psico-autoafirmación social		3%
			Psico-disfunciones familiares		3%

Fuente: Castellanos Guarín (2015)

Tabla 6. Variables Psicológicas Parte 3 – Modelo empírico UNAB (continuación)

VARIABLE	NIVEL RIESGO	PESO MÁXIMO	OPCIONES	CÓDIGO	PESO
			Psico-rendimiento escolar		3%
			Psico-desordenes laborales		3%
			Psico-presión del grupo de amigos		3%
			Psico-uso del tiempo libre		3%
			Psico-severidad del uso de drogas		3%

Fuente: Castellanos Guarín (2015)

Ahora bien, para realizar el cálculo en cada uno de los tipos de variables, realiza el cálculo del riesgo asociado a cada tipo en general. Así, por ejemplo, para calcular cada riesgo aplica las siguientes formulas:

Riesgo Financiero:

(([**ESC**] + [**CQV**] + [**TVL**] + [**COC**]) * 0.1) + (([**PRC**] + [**SPR**] + [**SAL**] + [**HLB**] + [**TCO**] + [**TPC**]) * 0.4) + (([**FPE**] + [**EST**] + [**NMH**] + [**TRB**]) * 0.5)

Riesgo Académico:

(([**EAP**] + [**PIN**] + [**TTG**] + [**PBE**] + [**PBED**] + [**PBEM**] + [**PBEC**] + [**PBES**] + [**NCC**] + [**NCM**] + [**PIC**] + [**RAC**] + [**PSPP**]) * 1)

Riesgo Psicológico:

(([**EAP**] + [**PIN**] + [**TTG**] + [**PBE**] + [**PBED**] + [**PBEM**] + [**PBEC**] + [**PBES**] + [**NCC**] + [**NCM**] + [**PIC**] + [**RAC**] + [**PSPP**]) * 1)

Una vez obtenido el valor para cada uno de los riesgos presentados, se calcula el porcentaje de riesgo total, donde, el riesgo financiero representa un peso del 30%, el riesgo académico 40% y el riesgo psicológico el 30% restante:

Riesgo General:

(**Riesgo Financiero** * 0,3) + (**Riesgo Académico** * 0,4) + (**Riesgo Psicológico** * 0.3)

De esta manera, se puede asociar para cada uno de los estudiantes procesados un nivel de riesgo en base a su información. Aquellos estudiantes con un nivel de riesgo mayor, son los que se consideran como estudiantes con riesgo de deserción.

Ahora, si revisamos el ámbito internacional son varios los estudios realizados al respecto, muchos de ellos enfocados no solo en la identificación de estudiantes con riesgo de deserción, sino en otros aspectos importantes para la educación como: la identificación de estrategias de aprendizaje aplicables de forma particular a ciertos grupos de estudiantes con características similares, la identificación de estudiantes con problemas de aprendizaje, el mejoramiento de los currículos académicos, entre otros.

Uno de los estudios más importantes en este campo es Kumar y Pal, (2012) en la investigación titulada "*Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*", el cual, en base a la información registrada por los estudiantes de Ingeniería al momento de sus admisión y los datos académicos de las asignaturas que actualmente se encontraban cursando, aplicó técnicas de minería de datos de clasificación con árboles de decisión, haciendo uso de los algoritmos ID3, C4.5 y CART a través de la herramienta de software libre WEKA, para determinar cuáles de estos estudiantes se encontraban en riesgo de reprobar sus próximos exámenes. Para ello, el estudio inicialmente capturó los datos de las notas obtenidas por los estudiantes para una primera sesión de exámenes, luego con esta información se buscaba predecir los resultados de la siguiente sesión.

De la investigación de Kumar y Pal (2012), se destaca especialmente un factor no mencionado en las investigaciones previamente analizadas, la medición de precisión de los algoritmos utilizados, haciendo uso de datos de entrenamiento y la técnica de validación cruzada, la cual permite obtener un porcentaje de precisión del algoritmo e identificar cuál de ellos brinda mayor exactitud en sus resultados, así por ejemplo, se logró identificar que el algoritmo C4.5 obtuvo un porcentaje mayor de precisión (67.7%) en comparación con los algoritmos ID3 (62.2%) y CART (62.2%).

Teniendo en cuenta lo anterior, el estudio concluye que el algoritmo de C4.5 que es un algoritmo de aprendizaje automático, puede brindar resultados más exactos si se entrena con datos históricos, llegando incluso a predecir a corto

plazo los estudiantes de primer ingreso con riesgo de reprobar y aquellos que requieren un acompañamiento especial.

Si observamos de forma general las investigaciones analizadas observamos ciertos aspectos comunes en el proceso de minería de datos:

- El uso de herramientas de software libre, especialmente el software WEKA como principal herramienta de procesamiento de datos.

- Clasificación, como principal método de identificación de patrones que permitan identificar estudiantes con riesgo de deserción.

- EL uso de la técnica de árboles de decisión, especialmente el uso de los algoritmos ID3 y C4.5.

Se podría pensar que estos factores comunes pueden ser una coincidencia, teniendo en cuenta la cantidad de investigaciones analizadas, sin embargo, si observamos el estudio de Amaya, Barrientos y Heredia (2015) titulado “Student Dropout Predictive Model Using Data Mining Techniques”, el cual, realiza un breve análisis de las técnicas de minería de datos y modelos predictivos utilizados en estudios similares alrededor del mundo (ver tabla 7), se puede identificar que los Árboles de Decisión son la principal técnica utilizada para encontrar patrones que permitan identificar alumnos con riesgo de deserción haciendo uso de los diferentes algoritmos de los cuales disponen.

Tabla 7. Técnicas de Minería de Datos utilizadas en estudios similares.

N°	PAÍS	ESTUDIO	TÉCNICA
1	Colombia	Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos.	**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.	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.

Fuente: Amaya, Barrientos y Heredia (2015)

Tabla 7. Técnicas de Minería de Datos utilizadas en estudios similares. (continuación)

N°	PAÍS	ESTUDIO	TÉCNICA
4	Tailandia	A Comparative Analysis of Techniques for Predicting Academic Performance.	**Árboles de Decisión (J48)** Redes Bayesianas
5	Estados Unidos	A Comparison of Logistic Regression, Neural Networks, and Classification Trees Predicting Success of Actuarial Students Phyllis Schumacher.	Regresión logística, Redes neuronales, Árboles de clasificación
6	Argentina	Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa.	Técnica de Regresión Logística
7	México	Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos.	**Árboles de decisión C4.5** Técnica de los k vecinos más cercanos
8	México	Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos.	**Árboles de decisión mediante el algoritmo C4.5**
9	Estados Unidos	New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention.	Árboles de clasificación Regresión logística
10	Nueva Zelanda	Predicting student success by mining enrolment data.	Árboles de clasificación Regresión logística
11	Croacia	Student Dropout Analysis With Application of Data Mining Methods.	Regresión logística **Árboles de decisión** Redes neuronales
12	Estados Unidos	Learning Patterns of University Student Retention.	Uno-R, **C4.5**, ADTrees, Naive Bayes, Bayes redes y redes radiales sesgo
13	Estados Unidos	Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks.	Redes neuronales (red backpropagation feed-forward)
14	India	Data Mining: A prediction for performance improvement using classification.	Algoritmo de clasificación Bayesiano (Naïve Bayes).
15	India	Mining Education Data to Predict Student's Retention: A comparative Study.	**Árbol de decisión ID3** **Árbol de decisión C4.5** **Árbol de decisión ADT**

Fuente: Amaya, Barrientos y Heredia (2015)

6. MARCO TEÓRICO

Desde una perspectiva general, el mejoramiento de la estimación de riesgo de deserción de los estudiantes de pregrado de la Universidad Autónoma de Bucaramanga empleando herramientas Business Intelligence soportadas en Software Libre debe ser abordado desde tres bases teóricas que incluyen, la deserción en la educación superior, la inteligencia de negocios (*Business Intelligence*) y la Minería de Datos (*Data Mining*)

6.1 DESERCIÓN EN LA EDUCACIÓN SUPERIOR

La deserción es un término del que hoy en día no se tiene un claro consenso acerca de su definición, teniendo en cuenta las diferentes características que pueden definir a un estudiante como desertor y los diferentes factores que se deben tener en cuenta. Para el Ministerio de Educación Nacional (2010), la deserción puede ser definida desde dos puntos de vista diferentes:

- Desde el punto de vista institucional todos los estudiantes que abandonan una IES pueden ser clasificados como desertores, teniendo en cuenta que cada estudiante que abandona la institución crea un lugar vacante que pudo haber sido ocupado por otro alumno persistente en los estudios.

- Desde el punto de vista del Sistema de Educación Superior se considera la deserción como la situación en la cual un estudiante abandona totalmente el sistema educativo. No se considera como deserción en el sentido riguroso del término a aquellos estudiantes que realizan transferencias entre instituciones de educación superior, los cuales son clasificados como migrantes dentro del sistema.

De acuerdo con las perspectivas anteriores, el Ministerio de Educación Nacional, (2010) adopta para la medición y seguimiento de esta problemática la siguiente definición: “La deserción se puede entender 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”.

Para realizar la medición de la deserción se utilizan dos indicadores principales que evidencian su magnitud. Las tasas de deserción anual y por cohorte. La primera mide el porcentaje de estudiantes que estaban matriculados un año antes y que figuran como desertores un año después. En este indicador, Colombia registra una tasa de 10,3% al cierre de 2014, cifra superior a la del Reino Unido, (8,6%) e inferior a la de Brasil (18%) y Estados Unidos (18.3%). Por su parte, la tasa de deserción por cohorte muestra el porcentaje de no culminación de estudios, en tanto ilustra la cantidad de estudiantes que desertan de cada 100 que ingresan a algún programa universitario. En este indicador, Colombia alcanza una cifra de 45,8%, lo que la deja en una posición intermedia entre México con 42%, Argentina con 43%, Venezuela con 52% y Chile con 54% (Ministerio de Educación Nacional, 2015a).

De acuerdo con los porcentajes anteriores en Colombia, por cada 100 estudiantes que se matriculan en una institución de educación superior, solo 54 continúan sus estudios y 48 estudiantes abandonan el sistema. Es tal la preocupación existente alrededor de esta problemática que dentro del documento Visión Colombia II Centenario, el Gobierno Nacional espera para el año 2019 haber reducido al 25% el índice de deserción en la educación superior (Departamento Nacional de Planeación, 2016).

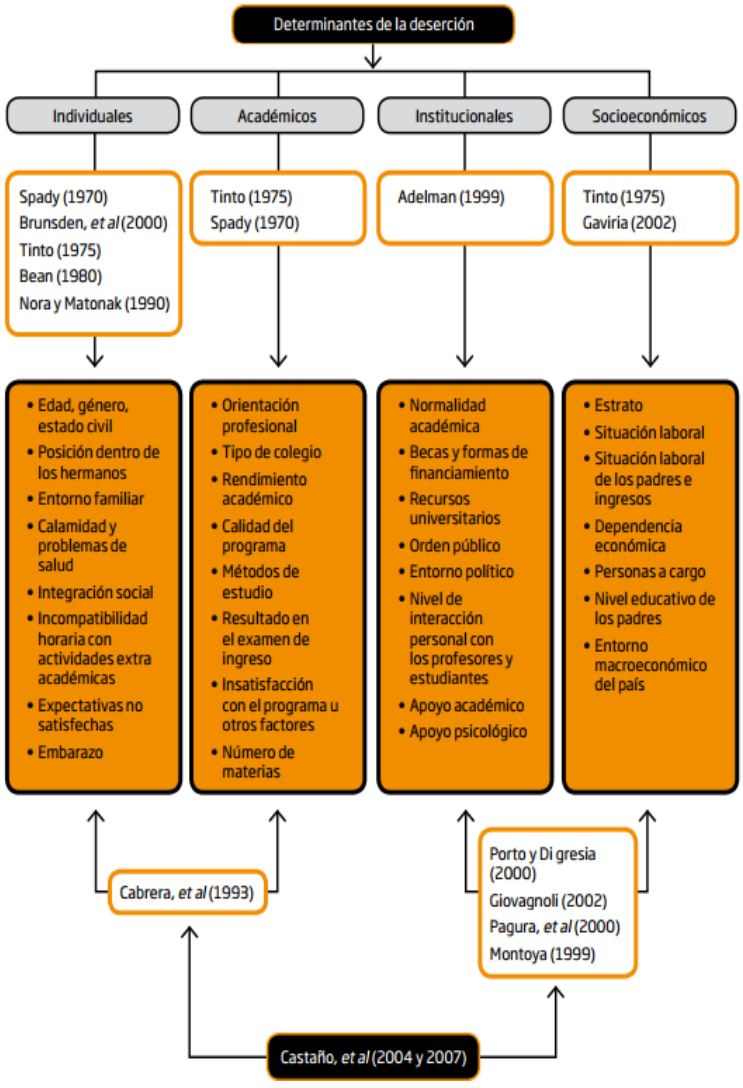
Pero, ¿por qué son tan altos los niveles de deserción de los estudiantes en la educación superior? Son varios los motivos que llevan a los estudiantes a abandonar sus estudios, algunos propios de la persona y otros debido a factores externos. El Ministerio de Educación Nacional (2010), en su estudio Deserción Estudiantil en la Educación Superior Colombiana, destaca los factores comunes de la literatura que son considerados determinantes de esta problemática (Ver Figura 4), identificando los aspectos individuales, institucionales, académicos y socioeconómicos del estudiante.

Para cada uno de los factores identificados se presentan ciertas características que afectan la situación del estudiante, como la edad, el género, el estado civil y números de hermanos en el factor individual o el tipo de colegio (privado u oficial), resultado de exámenes de ingreso y número de asignaturas en el factor académico o el estrato, la situación laboral, personas a cargo y el nivel educativo de los padres en el factor socioeconómico.

Teniendo en cuenta que gran parte de esta información es registrada por las instituciones de educación superior durante el proceso de admisión de los estudiantes o durante la permanencia de los mismos en su programa académico ¿es posible a través de esta información identificar de forma temprana, aquellos estudiantes con riesgo de deserción? Pues bien, esta es una de las preguntas que las instituciones educativas empiezan a hacerse, gracias a la gran evolución tecnológica de los últimos años en relación al procesamiento masivo de

información y su utilización para transformar datos en conocimiento, de manera que este conocimiento apoye la toma de sus decisiones, lo que se conoce como Inteligencia de Negocios o Business Intelligence.

Figura 4. Determinantes de la deserción en la educación superior



Fuente: Ministerio de Educación Nacional (2010)

El objetivo de la presente investigación se encuentra orientado a la aplicación de métodos estadísticos y predictivos que permitan identificar de forma temprana los estudiantes con riesgo de deserción, evaluando las características académicas, financieras y psicológicas de los estudiantes de primer ingreso en los programas de pregrado. Estos métodos estadísticos y predictivos se aplican a través de los que se conoce como Minería de Datos, que a su vez forma parte fundamental del proceso general de la Inteligencia de Negocios o BI por sus siglas en Ingles (*Business Intelligence*).

A continuación, se presenta el concepto general de la BI, la Minería de Datos y las diferentes metodologías que describen el proceso de extracción de conocimiento a través de las principales técnicas de minería existentes.

6.2 *BUSINESS INTELLIGENCE* (BI)

Gracias a las facilidades de almacenamiento de datos, los individuos y las organizaciones pueden llegar a almacenar grandes volúmenes de información, por ejemplo, el registro de clientes, ventas, proveedores y reclamos si se habla de una empresa comercial, o el comportamiento financiero de sus clientes si se tiene en cuenta una empresa del sector bancario. Esta gran cantidad de información almacenada representa una gran dificultad en el proceso de toma de decisiones dentro de las organizaciones, haciendo demasiado complejo y dinámico procesar eficazmente los datos a través de un enfoque intuitivo, por lo que requieren métodos más rigurosos basados en metodologías de análisis y modelos matemáticos (Vercellis, 2011).

La inteligencia de negocios puede definirse como un conjunto de modelos matemáticos y metodologías de análisis que explotan sistemáticamente los datos disponibles para recuperar la información y conocimiento útil para apoyar los procesos de toma de decisiones complejas, Vercellis, (2011). BI combina productos, tecnologías y métodos para organizar la información clave que se necesita para mejorar la gestión de los beneficios y el rendimiento (Howson, 2007).

Según Parr Rud (2000), desde el punto de vista comercial BI brinda la capacidad de tomar mejores y más rápidas decisiones que permitan aumentar los ingresos y reducir los costos, ya que aporta los medios necesarios para facilitar la comprensión del contexto comercial de las organizaciones, como los clientes, el mercado, la oferta, la demanda, los recursos y los competidores, de acuerdo con análisis históricos, actuales y predictivos de las operaciones comerciales.

De acuerdo con Han, Kamber y Pei (2001), gran parte de los análisis realizados en BI se basan en diferentes tecnologías que aplican técnicas de Minería de Datos, así por ejemplo, las técnicas de clasificación son utilizadas para realizar análisis predictivos, ya sea en el comportamiento de los mercados o las ventas; y las técnicas de agrupación, que permiten por ejemplo apoyar los procesos de gestión de las relaciones con los clientes, permitiendo entender mejor las características de grupos de clientes y desarrollar programas de fidelización para clientes de alto valor.

Es así como la Minería de Datos se convierte en una base fundamental dentro de los procesos de la BI, permitiendo obtener el conocimiento a partir de los datos de las organizaciones, conocimiento que es utilizado para orientar la toma de decisiones en pro de obtener mejores resultados. A continuación, se presenta un análisis conceptual de la Minería de Datos y sus procesos asociados.

6.3 *DATA MINING*

La Minería de datos es un término que en muchos casos suele ser mal definido o confundido con otros conceptos. Han, Kamber y Pei (2001), definen la minería de datos como el proceso de descubrimiento de patrones y conocimientos interesantes a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, almacenes de datos, la web y otros repositorios de información o datos que se transmiten al sistema de forma dinámica.

Han, Kamber y Pei (2001) realizan un breve análisis de los conceptos “Minería” y “Datos” por separado en relación a otros contextos. Por ejemplo, el proceso de extracción de oro se realiza sobre la tierra o sobre la roca, y esta actividad se conoce como Minería de Oro, y no como minería de roca, análogamente el término Minería de datos debería llamarse de forma apropiada Minería de Conocimiento, sin embargo, este término se queda corto al no reflejar el énfasis de la minería sobre grandes volúmenes de datos.

Otros autores como Witten, Frank y Hall, (2005), exponen: “La Minería de Datos es la extracción de la información implícita, previamente desconocida y potencialmente útil de los datos, que a través de programas software permiten encontrar de forma automática regularidades o patrones. Patrones, que una vez identificados permiten hacer predicciones exactas sobre datos futuros”.

Según Fayyad, Piatetsky-Shapiro y Smyth, (1996), “La Minería de Datos es solo un paso dentro del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD por sus siglas en inglés) que consiste en la aplicación de análisis de datos y algoritmos de descubrimiento, en virtud de las limitaciones de eficiencia computacionales aceptables que producen una enumeración particular de patrones (o modelos) en los datos”.

De acuerdo a los autores presentados se puede definir la minería de datos como el proceso de extracción de conocimiento a partir de grandes volúmenes de datos, a través de herramientas software que aplican diferentes algoritmos, para intentar encontrar patrones que a simple vista no son observables por el ser humano. Estos patrones permiten a su vez realizar tareas de predicción y análisis sobre un determinado contexto.

Dentro de un contexto general la Minería de Datos puede ser usado por las empresas y organizaciones para explotar sus datos de manera que permita: encontrar evidencias de robo o fraude, por ejemplo, identificando compras sospechosas a través de tarjetas de crédito de acuerdo al historial de compra del cliente y declinar la transacción; identificar patrones de compra, permitiendo ofrecer a los clientes durante su proceso de compra nuevos productos o servicios de acuerdo a los intereses identificados; o evidenciar la necesidad de nuevos productos, nuevos mercados o nuevas fuentes de ingresos (Sauter, 2011).

Turban, Sharda, Denle y King, (2013), en su libro “Business Intelligence, A Managerial Approach” definen las principales áreas de aplicación que puede tener la Minería de Datos ya sea para resolver problemas urgentes de las organizaciones o simplemente para explorar nuevas oportunidades de negocio emergentes:

Tabla 8. Algunas aplicaciones de la Minería de Datos

Sector	Forma de Aplicación
Bancario	Automatización del proceso de solicitud de préstamos al predecir con precisión los morosos más probables. Detección de transacciones fraudulentas con tarjetas de crédito Identificación de formas de maximizar el valor del cliente mediante la venta de productos y servicios que son más propensos a comprar Optimizar el rendimiento para pronosticar con precisión el flujo de dinero en efectivo en las entidades bancarias (por ejemplo, cajeros automáticos, sucursales bancarias)
Ventas	Predecir los volúmenes de ventas precisas en las tiendas con el fin de determinar los niveles de inventario correctos Identificar los niveles de consumo de diferentes tipos de productos para optimizar la logística y por lo tanto maximizar las ventas
Manufactura y Producción	Predecir la avería de maquinaria antes de que ocurran a través de la utilización de los datos sensoriales Identificar las anomalías y los puntos comunes en los sistemas de producción para optimizar la capacidad de fabricación Descubrir nuevos patrones para identificar y mejorar la calidad del producto.
Mercado de Valores	Predecir cuándo y cuánto los precios de los bonos van a cambiar; Pronosticar el alcance y la dirección de las fluctuaciones de valores. Identificar y prevenir actividades fraudulentas en el comercio de valores

Fuente: Turban, Sharda, Denle y King (2013)

Tabla 8. Algunas aplicaciones de la Minería de Datos (continuación)

Sector	Forma de Aplicación
Seguros	Determinar planes de tarifas óptimas basadas en el análisis de las reclamaciones y los datos del cliente. Predecir qué clientes son más propensos a comprar nuevos productos con características especiales. Prevenir la reclamación de pagos incorrectos y actividades fraudulentas.
Hardware y Software	Predecir fallos de unidades de disco mucho antes de producirse Identificar y filtrar el contenido web y mensajes de correo electrónico no deseados Identificar los productos de software potencialmente no seguros.
Turismo	Predecir las ventas de los diferentes servicios (tipos de asientos en los aviones, los tipos de habitaciones en hoteles, tipos de automóviles en las empresas de alquiler) La demanda prevista en diferentes lugares para asignar mejor los recursos limitados de la organización Identificar los clientes más rentables para proporcionarles servicios personalizados
Salud	Identificar nuevas relaciones de costes y beneficios entre los diferentes tratamientos para desarrollar estrategias más efectivas Prever el nivel de demanda en diferentes centros de salud para asignar de forma óptima los recursos de la organización
Medicina	Descubrir las relaciones entre los síntomas y enfermedades (así como las enfermedades y tratamientos exitosos) para ayudar a los profesionales médicos a tomar decisiones informadas y correctas en el momento oportuno.
Industria del Entretenimiento	Analizar los datos de visualización para decidir qué programas mostrar durante un determinado horario y cómo maximizar el rendimiento de los anuncios publicitarios. Predecir el éxito financiero de las películas antes de que se produzcan para tomar decisiones de inversión y para optimizar los rendimientos.
Seguridad Nacional	Identificar los patrones de conductas terroristas Descubrir patrones de criminalidad (por ejemplo, lugares, tiempos, conductas delictivas, y otros atributos relacionados) para ayudar a resolver casos criminales en el momento oportuno.

Fuente: Turban, Sharda, Denle y King (2013)

Gran parte de la adopción de la Minería de Datos se debe a dos factores: la gran cantidad de datos que deben ser procesados y la no posibilidad del ser humano de procesar estos datos de forma manual, por lo tanto, se hace necesario contar con software que permita realizar el procesamiento y generar información útil que resulte fácilmente analizable para la generación de conocimiento, sin embargo, esta tarea no resulta tan sencilla, el software no puede procesar los datos si estos no son adecuados, y cuando decimos adecuados nos referimos a que mucha veces estos no contienen formatos válidos, poseen valores nulos y erróneos que afectan enormemente el proceso de minería, por lo tanto, se hace necesario realizar ciertas tareas previas que permitan realizar un procesamiento adecuado para conseguir resultados precisos para su posterior análisis.

A nivel general la Minería de Datos no define un proceso para obtener el conocimiento a partir de los datos, incluso, en la comunidad en general no existe un consenso claro que identifique cada una de las etapas que debe seguir el proceso, sin embargo, a partir de las diferentes fuentes bibliográficas consultadas se pueden identificar dos metodologías aplicables y un proceso: Las metodologías SEMMA (*Sample, Explore, Modify, Model, Assess*) y CRISP-DM (*Cross- Industry Standard Process for Data Mining*) y el proceso KDD (*Knowledge Discovery in Databases*).

6.4 METODOLOGÍAS DE MINERÍA DE DATOS

Las propuestas de metodologías para implementar proyectos de minería de datos no resultan tan abundantes como las técnicas de extracción de conocimiento. Incluso, el día de hoy no existe un proceso general definido que indique la manera en la cual deben ser explotados los datos para llegar al conocimiento. De acuerdo con Chapman et al., (2000), las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos.

En el mismo estudio, los autores hacen referencia a la diferencia que existe entre lo que se define como una metodología y lo que se define como un modelo de proceso, donde este último establece qué hacer, y la metodología especifica cómo hacerlo. Una metodología no solo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas (Chapman et al., 2000). De acuerdo con lo anterior, a continuación, se definen CRISP-DM y SEMMA como metodologías de extracción de conocimiento y KDD como modelo de proceso de extracción de conocimiento.

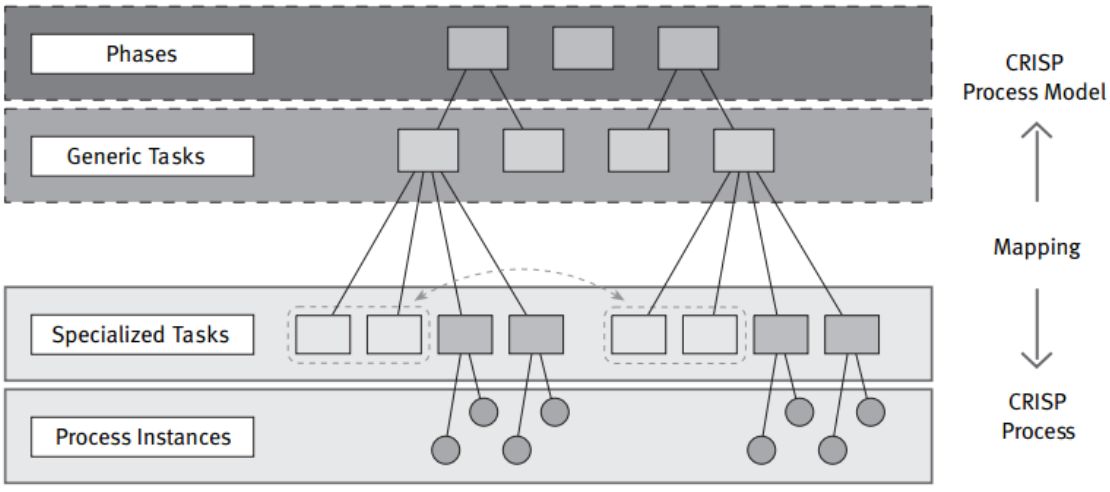
6.4.1 Metodología CRISP-DM

CRoss Industry Standard Process for Data Mining es una metodología que nace en al año 1996 gracias a la iniciativa de tres empresas que en su momento se encontraban dedicando gran parte de sus operaciones al nuevo mercado de la Minería de Datos, por lo tanto, tenían la necesidad de idear una metodología de trabajo estándar, no propietaria, de libre acceso, que les permitiera generar confianza a sus clientes. Fue así como DaimlerChrysler (entonces Daimler-Benz), SPSS y NCR unieron esfuerzos y presentaron esta metodología que definen como:

CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consiste en un conjunto de tareas que se describen en cuatro niveles de abstracción (de lo general a lo específico): fase, tareas genéricas, tareas especializadas e instancia de procesos (Chapman et al., 2000) (ver figura 5).

En el nivel superior, el proceso de minería de datos se organiza en una serie de fases; cada fase se compone de varias tareas genéricas de segundo nivel. Este segundo nivel se llama genérico, ya que se pretende que sea suficiente para cubrir todas las posibles situaciones de minería de datos en general. El tercer nivel, Tareas especializadas, es el lugar para describir cómo las acciones en las tareas genéricas deben llevarse a cabo en ciertas situaciones específicas. El cuarto nivel, Instancia de procesos, es un registro de las acciones, decisiones y resultados de un trabajo real de la minería de datos.

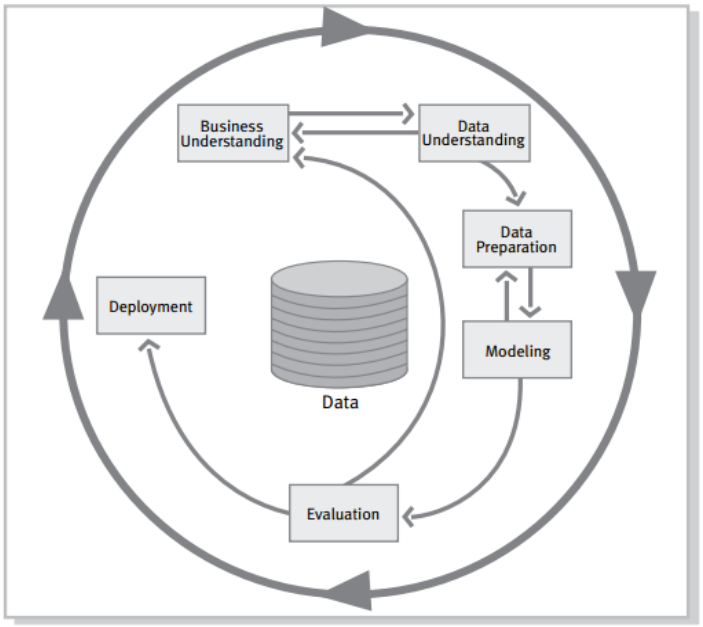
Figura 5. Niveles de Abstracción de la Metodología CRISP:DM



Fuente: Chapman et al., (2000)

Según Chapman et al., (2000) CRISP-DM a su vez define un ciclo de vida de un proyecto de minería de datos que contempla 6 fases fundamentales (Ver figura 6), donde el direccionamiento de las fases indica que su secuencia no es necesariamente estricta, teniendo en cuenta que es posible retomar fases anteriores si el proyecto así lo requiere, sin embargo, el resultado obtenido en cada fase determina el éxito de las fases posteriores. A continuación, se presenta una breve descripción de cada una de las fases definidas:

Figura 6. Ciclo de vida CRISP-DM de un proyecto de minería de datos



Fuente: Chapman et al., (2000)

Conocimiento del negocio (*Business Understanding*)
Implica la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, permitiendo comprender el proyecto como un problema de minería de datos para definir una planificación preliminar que permita alcanzar los objetivos.

Comprensión de los datos (*Data Understanding*)
Fase que inicia con la recolección de los datos, para luego realizar un proceso de familiarización, donde se identifica problemas de calidad y se formulan hipótesis preliminares sobre la información que a simple vista no es observable.

Preparación de datos (*Data Preparation*)
Fase que incluye la selección, transformación y limpieza de los datos de manera que puedan ser procesados por las herramientas de modelado utilizadas. Esta fase es susceptible a ser realizada varias veces y en cualquier orden prescrito.

Modelado (*Modeling*)
Fase donde se seleccionan y aplican diferentes técnicas de modelado, las cuales son definidas de acuerdo a las necesidades mismas del proyecto. La selección de la técnica depende en gran medida del proceso de preparación realizado previamente. En algunos casos la técnica requiere características específicas en los datos, lo que hace necesario volver a la fase de preparación.

Evaluación (*Evaluation*)
Fase donde se evalúa el modelo aplicado de acuerdo a las diferentes técnicas seleccionadas, de manera que se identifique claramente si se han cumplido los objetivos inicialmente planteados o si se ha abordado totalmente el problema. Esta fase permite identificar posibles errores que se hayan cometido en las fases previas y corregirlos antes de continuar con la fase de despliegue.

Despliegue (*Deployment*)
Comprende la fase final del proceso, pero no con ello su terminación, ya que es necesario presentar documentos que describan de forma clara el conocimiento adquirido, de manera que este pueda ser utilizado y aplicado por el cliente.

6.4.2 Metodología SEMMA

Es una metodología desarrollada por la empresa SAS, la cual ofrece servicios de gestión de datos y procesamiento masivo alrededor del mundo e incorpora esta metodología como parte de su servicio SAS Enterprise Miner. SEMMA es el acrónimo de *Sample, Explore, Modify, Model y Assess*, que definen las fases de un proyecto de minería de datos. A continuación, se presenta una breve descripción de cada una de ellas (SAS Institute Inc, 2003).

Muestra (*Sample*)
Fase que identifica uno o más conjuntos de datos de entrada, los cuales deben ser los suficientemente grandes para contener información significativa. Los conjuntos de datos deben representar valores de entrenamiento y validación. En esta fase se incluyen herramientas de preparación e importación de datos, así como técnicas de muestreo estadístico.

Explorar (*Explore*)
Fase que explora los datos en busca de relaciones, tendencias y anomalías con el fin de obtener un grado de comprensión de los mismos. Esta fase incluye el uso de herramientas para la obtención de la información estadística y la exploración gráfica que permitan identificar variables importantes para un posterior proceso de asociación.

Modificar (*Modify*)
Fase que prepara los datos para el análisis, mediante la creación y transformación de las variables existentes e incluye herramientas para definir transformaciones de datos, identificar valores atípicos y reemplazar valores perdidos.

Modelar (*Model*)
Fase que permite modelar los datos mediante el uso de herramientas analíticas como árboles de decisión, redes neuronales, regresiones lineales, regresiones logísticas o cualquier otra técnica de minería de datos. Los algoritmos utilizados son entrenados a través de datos de prueba que permiten mejorar la exactitud de los resultados obtenidos.

Evaluar (*Assess*)
Fase que permite evaluar la utilidad y fiabilidad de los resultados del proceso de minería de datos. Este paso incluye el uso de herramientas de comparación, soporte de decisiones y generación de informes.

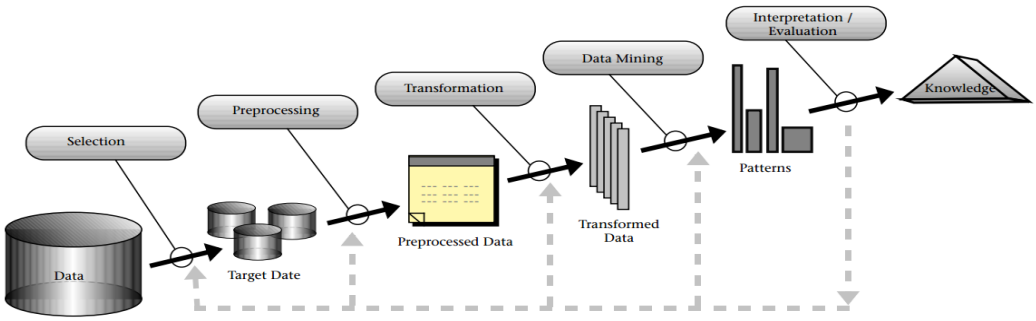
Aunque la metodología SEMMA contiene algunos de los elementos esenciales de cualquier proyecto de minería de datos, sólo se refiere a la estadística, la modelización y la manipulación de datos dentro del proceso de minería de datos. Además, SEMMA está diseñado específicamente para trabajar con el software Enterprise Miner, el software de minería de datos del SAS Institute, por lo tanto, no puede aplicarse fuera de las limitaciones de dicho sistema (Rohanizadeh y Moghadam, 2010).

6.4.3 Proceso KDD - *Knowledge Discovery in Databases*

Según Moine, Haedo y Gordillo, (2001), “KDD no es propiamente una metodología de minería de datos, sin embargo, en los inicios del año 1996, constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información”, sin embargo, hoy día, el término KDD es frecuentemente confundido con el concepto de Minería de Datos, e incluso son utilizados indistintamente para referirse al proceso de descubrimiento de conocimiento. Fayyad, Piatetsky-Shapiro y Smyth,

(1996), definen KDD como el proceso global de descubrimiento de conocimiento útil a partir de los datos y enfatiza en que la minería de datos es solo un determinado paso en este proceso (Ver figura 7).

Figura 7. Modelo de proceso KDD para la extracción de conocimiento



Fuente: Fayyad, Piatetsky-Shapiro y Smyth (1996)

De acuerdo con la gráfica, El proceso de KDD es interactivo e iterativo e implica numerosos pasos, muchos de ellos apoyados por decisiones tomadas por el usuario. A continuación, se describen los pasos del proceso:

Un primer paso incluye la selección de las fuentes de datos, haciendo referencia a bases de datos, almacenes de datos y la web. Estas fuentes de datos pueden ser internas (de la misma organización que implementa el proyecto) o externas.

A continuación, la etapa limpieza y preprocesamiento, que según Fayyad, Piatetsky-Shapiro y Smyth, (1996), realiza operaciones básicas sobre los datos que incluyen la eliminación del ruido y la determinación de las estrategias para el manejo de campos faltantes.

En tercer lugar, se encuentra la etapa de transformación y reducción, que de acuerdo a lo definido por Fayyad, Piatetsky-Shapiro y Smyth, (1996), se centra especialmente en la búsqueda de características útiles para representar los datos en función de los objetivos del proyecto de minería de datos. Con la reducción o métodos de transformación, el número de variables en consideración puede ser reducido.

EL cuarto paso define y aplica las diferentes técnicas existentes para la extracción del conocimiento, tales como la clasificación por medio de árboles de decisión, técnicas de regresión o clustering.

Un quinto paso incluye la interpretación y evaluación de los resultados obtenidos de acuerdo a las técnicas de minería de datos aplicadas. Se analizan los patrones identificados (si se utilizaron técnicas de clasificación) o se interpretan los agrupamientos (si se aplicó la técnica de clustering).

Finalmente, una última etapa permite la presentación de informes y la documentación del conocimiento extraído. Informes que son presentados a los clientes para la toma decisiones.

6.5 MODELOS DE MINERÍA DE DATOS

Hasta ahora se ha presentado el concepto de Minería de Datos y las diferentes metodologías y procesos para la extracción de conocimiento. En este apartado se exponen los conceptos relacionados con los Modelos, Métodos y las principales Técnicas de Minería. La Figura 8 presenta una visión general de los temas a tratar:

Figura 8. Visión general de la minería de dato

Fuente: Elaboración propia

Según Hernández Orallo, Ramírez Quintana y Ferri Ramírez, (2004) el conocimiento en la Minería de Datos puede ser identificado en forma de relaciones, patrones o reglas inferidas de los datos, o bien en forma de una descripción más concisa. Existen varias formas de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos, estos modelos son el Descriptivo y el Predictivo.

6.5.1 Modelo Descriptivo

Identifica patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados más no para predecir nuevos datos (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

6.5.2 Modelo Predictivo

Pretenden estimar valores futuros o desconocidos de variables de interés, denominadas variables objetivo o dependientes, usando otras variables o campos de las bases de datos denominadas variables predictivas o independientes (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

Para cada uno de los modelos definidos existen una serie de métodos, que permiten conseguir ya sea los objetivos de descripción o los objetivos de agrupamiento, así, por ejemplo, se consideran dentro del modelo descriptivo los métodos de agrupamiento (Clustering) y asociación, mientras que los métodos de clasificación y regresión se clasifican dentro del modelo predictivo.

6.6 MÉTODOS DE MINERÍA DE DATOS

A continuación, se presenta una perspectiva general de los principales métodos de minería de datos de acuerdo a las fuentes consultadas.

6.6.1 Método de Clasificación

Según Han, Kamber y Pei, (2001), “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 (es decir, los objetos de datos para los que se conocen las etiquetas de clase). El modelo se utiliza para predecir la etiqueta de clase de objetos que son desconocidas”. Turban, Sharda, Denle y King, (2013), complementan diciendo que “la Clasificación aprende los patrones de los datos del pasado con el fin de identificar nuevas instancias (etiquetas desconocidas) en sus respectivos grupos o clases”.

La clasificación identifica los valores de las variables predictivas de acuerdo a unos valores de entrada que se conocen como datos de entrenamiento, estos datos, permiten al modelo entrenarse para conseguir los valores predictivos más exactos posibles.

Después de que el modelo ha sido entrenado, el modelo se prueba con una muestra de datos reservados para la evaluación de la precisión. Según Turban, Sharda, Denle y King, (2013), hay varios factores que se consideran en la evaluación del modelo, incluyendo los siguientes:

- **La exactitud predictiva**. Es la capacidad del modelo para predecir correctamente la etiqueta de clase de datos nuevos o previamente no conocidos. Para calcular esta medida, las etiquetas de un conjunto de clases

de datos de prueba se comparan con las etiquetas de clase predichas por el modelo. La precisión puede entonces ser calculada como un porcentaje de exactitud, que es el porcentaje de muestras de datos de prueba correctamente clasificados por el modelo.

- **Velocidad**. Tiene en cuenta los costes computacionales involucrados en la generación y uso del modelo, en donde más rápido se considera mejor.

- **Robustez**. Es la capacidad del modelo para hacer predicciones razonablemente exactas, teniendo en cuenta los datos con ruido o datos con valores perdidos y erróneos.

- **Escalabilidad**. Define la capacidad de construir un modelo de predicción de manera eficiente con una cantidad grande de datos.

- **Interpretabilidad**. Tiene en cuenta el nivel de comprensión y conocimiento proporcionado por el modelo.

Dentro del método de clasificación podemos encontrar las técnicas tales como reglas de clasificación, árboles de decisión, fórmulas matemáticas, redes neuronales y clasificadores bayesianos.

6.6.2 Método de Regresión

Según Han, Kamber y Pei (2001), “la regresión se utiliza para predecir los valores que faltan o los datos numéricos no disponibles. Esta es la principal diferencia respecto a la clasificación; el valor a predecir es numérico”. Turban, Sharda, Denle y King, (2013), complementan el concepto afirmando que “el objetivo en este caso es minimizar el error entre el valor predicho y el valor real”.

Turban, Sharda, Denle y King (2013), afirman que los métodos de regresión se prueban mediante el cálculo de varias estadísticas que miden la diferencia entre los valores predichos y los valores esperados. Los datos históricos para un proyecto de regresión normalmente se dividen en dos conjuntos de datos: una para la construcción del modelo y otro para probarlo.

6.6.3 Método de Agrupamiento (Clustering)

El agrupamiento consiste en obtener grupos de datos fuertemente relacionados entre ellos. A diferencia de la regresión y 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 (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

6.6.4 Método de Asociación

La regresión tiene como principal objetivo identificar relaciones entre los elementos en grandes cantidades de datos que a simple vista no son observables. En el comercio, por ejemplo, la asociación es utilizada para identificar aquellos productos que se compran juntos con mayor frecuencia, de esta manera los comercios pueden organizar estos productos de forma consecutiva para facilitar la compra por parte de sus clientes.

6.7 TÉCNICAS DE MINERÍA DE DATOS

En la actualidad existen una gran cantidad de técnicas para cada uno de los métodos presentados, cada una de ellas con sus respectivos algoritmos. A continuación, se exponen las principales técnicas que apoyan el proceso de Minería de Datos:

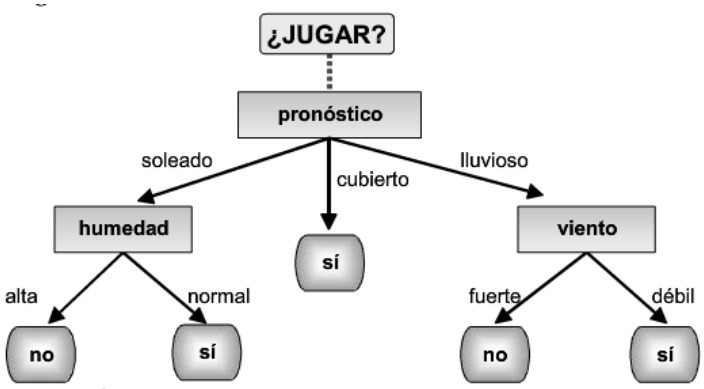
6.7.1 Arboles de Decisión

Según Han, Kamber y Pei, (2001), “son 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. El nodo superior en un árbol es el nodo raíz”.

Esta técnica se usa en tareas de clasificación, agrupamiento y regresión. Los árboles de decisión usados para predecir variables categóricas reciben el nombre de árboles de clasificación, ya que distribuyen las instancias en clases. Cuando los arboles de decisión se usan para predecir variables continuas se llaman árboles de regresión (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

La figura 9 muestra un ejemplo de árbol de decisión para un problema ficticio que muestra las condiciones climatológicas (pronóstico, humedad y viento) adecuadas para jugar cierto deporte. El árbol trata de decidir si se debe jugar o no.

Figura 9. Ejemplo de árbol de decisión



Fuente: Hernández Orallo, Ramírez Quintana y Ferri Ramírez (2004)

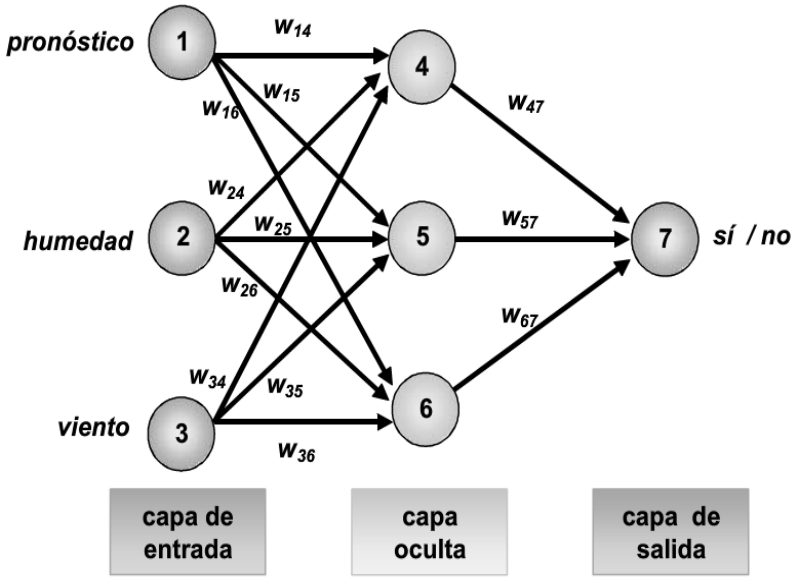
Dentro de los algoritmos principales que pueden ser aplicados bajo la técnica de árboles de decisión se destacan el ID3, C4.5 y CART. Los algoritmos ID3 y CART se desarrollaron de forma independiente, mientras que C4.5 es una mejora del algoritmo ID3.

6.7.2 Redes Neuronales Artificiales

Las redes neuronales pueden verse como un grafo dirigido con muchos nodos (elementos del proceso) y arcos entre ellos (sus interconexiones). Cada uno de estos elementos funciona independiente de los demás, usando datos locales para dirigir su procesamiento (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

Según Hernández Orallo, Ramírez Quintana, y Ferri Ramírez, (2004), “una red neuronal se organiza comúnmente con una capa de entrada donde cada nodo se corresponde con una variable independiente a examinar, unos nodos internos organizados en una o varias capas ocultas y una capa de salida con los nodos salida (los posibles valores de la variable objetivo). Cada nodo de la capa de entrada está conectado a cada nodo de la capa oculta. Los nodos de la capa oculta pueden estar conectados a otros nodos de otra capa oculta o a los nodos de la capa de salida”, como se observa en la figura 10.

Figura 10. Ejemplo de la representación gráfica de una red neuronal artificial



Fuente: Hernández Orallo, Ramírez Quintana y Ferri Ramírez (2004)

Según Turban, Sharda, Denle y King (2013), la técnica de redes neuronales trata de asemejar de alguna forma la capacidad de procesamiento del cerebro humano, incluso poseen algunos rasgos deseables similares a las de las redes neuronales biológicas, tales como las habilidades para aprender, auto-organizarse y la tolerancia a fallos.

6.7.3 Clasificadores bayesianos

Los métodos bayesianos pertenecen a la familia de modelos probabilísticos de clasificación. Calculan explícitamente la probabilidad de que una observación dada pertenezca a una clase específica mediante el teorema de Bayes, Vercellis, (2011).

A diferencia de las demás técnicas presentadas hasta ahora, la cuales no se basan en suposiciones probabilísticas, la técnica de clasificadores bayesianos requiere que el usuario estime la probabilidad que puede tener una observación dada de pertenecer a una clase específica.

Dentro de los algoritmos más destacados de esta técnica se encuentran los algoritmos *Naive Bayes*, Redes Bayesianas y el algoritmo de clasificadores basados en redes bayesianas.

6.7.4 Reglas de Clasificación

Esta técnica utiliza un conjunto de reglas IF-THEN (Si-Entonces) para realizar el proceso de clasificación. Una regla IF-THEN es una expresión que se representa de la siguiente forma:

IF condición THEN conclusión.

Un ejemplo de una regla sería el siguiente:

R1: IF edad = Joven AND estudiante = si THEN comprar_computador = sí.

Donde la parte "IF" (o izquierda) se conoce como el antecedente de la regla o condición previa. La parte "THEN" (o derecho) es el consecuente de la regla. En R1, por ejemplo, el antecedente viene dado por dos condiciones unidas por el conector lógico AND. En este caso la regla contiene una predicción que indica que, si un cliente es joven y a su vez es estudiante, entonces comprará un computador. Han, Kamber y Pei, (2001)

6.7.5 Regresión Lineal

De acuerdo con Oracle Help Center (2016), “el análisis de regresión busca determinar los valores de los parámetros de una función que mejor se ajusten a un conjunto de observaciones proporcionadas”. Por ejemplo, de acuerdo con la siguiente ecuación:

$$Y = F(\mathbf{x}, \theta) + e$$

La regresión es el proceso que permite determinar el valor de un objetivo, en este caso la Variable (y) como una función (F) de uno o más predictores (x1, x2, ..., xn), un conjunto de parámetros (θ1, θ2, ..., θn), y una medida de error (e). Los predictores pueden entenderse como variables independientes y el objetivo como una variable dependiente. El error, también llamado el residual, es la diferencia entre el valor esperado y pronosticado (Oracle Help Center, 2016).

Otras técnicas de regresión son la **regresión lineal multivariable,** que hace referencia a la regresión lineal con dos o más variables predictoras (**x**1, **x**2...**x**n); **la regresión no lineal** que aplica cuando la relación entre (x) y (y) no genera una línea recta y la **regresión no lineal multivariable**, donde se tienen dos o más variables predictoras en una regresión no lineal (Oracle Help Center, 2016).

6.7.6 Reglas de Asociación

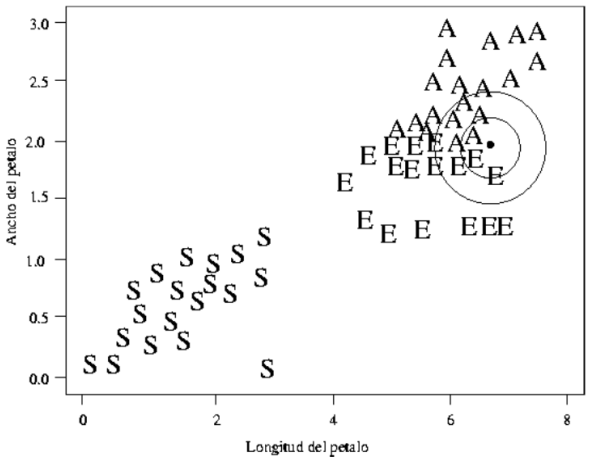
Las reglas de asociación o grupos de afinidad, se utilizan para identificar asociaciones recurrentes entre grupos de registros en un conjunto de datos. Por ejemplo, es posible determinar qué productos se compran en conjunto en una sola transacción y con qué frecuencia. (Vercellis, 2011)

Las reglas de asociación en general gozan de una especial popularidad debido a su fácil nivel de interpretabilidad. Por ejemplo, dadas dos proposiciones **Y** y **Z**, que pueden ser verdaderas o falsas, se puede afirmar de manera general que una regla es una implicación del tipo **Y ⇒ Z** con el siguiente significado: si **Y** es cierto, entonces **Z** también es cierto. Una regla se llama probabilística si la validez de Z se asocia con una probabilidad **p**: si **Y** es cierto, entonces **Z** también es cierto con probabilidad **p**. **(**Vercellis, 2011).

6.7.7 K-Vecinos

Algoritmo que determina para cada región del espacio la probabilidad de que un elemento que este situado en ella pertenezca a cada una de las reglas existentes. No existen reglas predefinidas, por lo tanto, la clasificación se va realizando a medida que aparecen los casos. Cuando un nuevo caso aparece, se genera un círculo con centro en dicho punto y un radio prefijado como parámetro del sistema. Se calcula el número de ejemplos que caen dentro del círculo y se etiqueta al nuevo caso como perteneciente a la clase más numerosa dentro del círculo (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

Figura 11. Representación gráfica de la técnica K-vecinos



Fuente: Hernández Orallo, Ramírez Quintana y Ferri Ramírez (2004)

Al analizar la gráfica, se puede identificar, teniendo en cuenta la circunferencia de mayor radio, que 6 ejemplos de A y 5 ejemplos de B son alcanzados, por lo tanto, el nuevo caso se etiqueta como perteneciente a la clase A.

6.7.8 K-Medias

Del inglés *K-means*, es una técnica de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos sin etiquetar. La idea de K-medias es situar a los prototipos o centros, en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares (Hernández Orallo, Ramírez Quintana y Ferri Ramírez, 2004).

Desde una perspectiva general, la figura 12 nos presenta diferentes algoritmos de las diferentes técnicas presentadas, asociando para cada uno de estos algoritmos los modelos y métodos dentro de los cuales son aplicables.

Figura 12. Técnicas de minería de datos y sus posibilidades de aplicación

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones / Factorizaciones
Redes neuronales	✓	✓	✓		
Árboles de decisión ID3, C4.5, C5.0	✓				
Árboles de decisión CART	✓	✓			
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		
Regresión lineal y logarítmica		✓			✓
Regresión logística	✓			✓	
Kmeans			✓		
Apriori				✓	
Naive Bayes	✓				
Vecinos más próximos	✓	✓	✓		
Análisis factorial y de comp. ppales.					✓
Twostep, Cobweb			✓		
Algoritmos genéticos y evolutivos	✓	✓	✓	✓	✓
Máquinas de vectores soporte	✓	✓	✓		
CN2 rules (cobertura)	✓			✓	
Análisis discriminante multivariante	✓				

Fuente: Hernández Orallo, Ramírez Quintana y Ferri Ramírez (2004)

Así, por ejemplo, las redes neuronales, los algoritmos genéticos y evolutivos, vecinos más próximos y algunos árboles de decisión presentan mayor aplicabilidad en los diferentes métodos y modelos de minería de datos.

6.8 HERRAMIENTAS SOFTWARE PARA LA MINERÍA DE DATOS

Las herramientas software dentro de la minería de datos facilitan la aplicación de las diferentes técnicas presentadas en el apartado anterior, no solo en el procesamiento de grandes cantidades de datos, sino en todo el proceso de minería, teniendo en cuenta que algunos de ellos permiten tareas de preparación, preprocesamiento y visualización de los resultados.

Dentro de las alternativas se encuentran diferentes herramientas software tanto libres como privativas. Weka, RapidMiner y Orange son los más destacados dentro de la comunidad del código abierto gracias a su funcionalidad extendida en comparación con otras herramientas. Por su parte Microsoft SQL Server, SAS Enterprise Miner y Oracle DM sobresalen en las herramientas software privativas.

6.8.1 Weka

The *Waikato Environment for Knowledge Analysis* (Weka) es un conjunto de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos o bien se pueden aplicar directamente sobre un conjunto de datos o a través de llamados propios desde código Java. Weka contiene herramientas para el procesamiento previo de datos, clasificación, regresión, clustering, reglas de asociación, y la visualización. También es muy adecuado para el desarrollo de nuevos sistemas de máquinas de aprendizaje (The University Of Waikato, s.f.).

El nombre WEKA proviene de un ave no voladora originaria de Nueva Zelanda. Dentro de sus principales características se encuentran las tres opciones gráficas ofrece, las cuales apoyan el preprocesamiento de datos, la selección de atributos, el aprendizaje, la visualización y un entorno experimental que permite probar y evaluar algoritmos. A continuación, se presentan las principales características de esta herramienta (Bouckaert, 2010):

- Preprocesamiento de datos. Así como un formato de archivo nativo (ARFF), WEKA es compatible con varios otros formatos (por ejemplo, CSV, archivos ASCII de Matlab), y conectividad con bases de datos a través de JDBC. Los datos pueden ser filtrados por un gran número de métodos (más de 75), que van desde la eliminación de atributos particulares a funcionalidades más avanzadas tales como el análisis de componentes principales.

- Clasificación. Los clasificadores de WEKA se dividen en métodos bayesianos (Naïve Bayes, redes bayesianas, etc.), los métodos de cercanías (vecino más cercano y variantes), los métodos basados en reglas (tablas de decisión), los

árboles de decisión (C4.5, árboles Naive Bayes, M5), las decisiones basadas en funciones (regresión lineal, SVMs, procesos de Gauss) y métodos diversos. Por otra parte, WEKA incluye interfaces para clasificadores implementados en Groovy y Jython.

- La agrupación. Weka implementa dentro de sus funciones el aprendizaje no supervisado, soportado por varios esquemas de conglomeración, como modelos EMbased, k-medias y varios algoritmos de agrupamiento jerárquico.

- Visualización de datos. Los datos pueden ser inspeccionados visualmente representando los valores de atributos en contra de la clase, o en contra de otros valores de atributos. Las salidas de las técnicas de clasificación pueden ser comparadas con datos de entrenamiento para detectar valores atípicos y observar las características. Para algunos métodos específicos existen herramientas especializadas de visualización, tales como un visor de árboles de clasificación, un visualizador de red de Bayes con diseño automático y un visor de agrupamiento jerárquico.

6.8.2 RapidMiner

Es una plataforma de software desarrollado por la compañía del mismo nombre que proporciona un entorno integrado para el aprendizaje automático, la minería de datos, la minería de texto, análisis predictivo y análisis de negocios. Se utiliza para aplicaciones comerciales e industriales, así como para la investigación, la educación, la formación, la creación rápida de prototipos y desarrollo de aplicaciones. Es compatible con todos los pasos del proceso de minería de datos (Bouckaert, 2010). Dentro de sus principales características se encuentran:

- Importación de datos. RapidMiner proporciona diferentes formatos de importación como Excel, CSV y XML.

- Enlace de datos. RapidMiner permite combinar diferentes formatos de datos, a través de plugins que facilitan el procesamiento

- Generación de funciones. Permite la creación de nuevas características de minería de datos.

- La utilización de XML para describir el proceso de descubrimiento de conocimiento

- Soporta algoritmos nativos del software Weka

6.8.3 Orange

Orange es una suite completa de software, basado en componentes para máquinas de aprendizaje y minería de datos, desarrollado en el Laboratorio de Bioinformática, Facultad de Informática y Ciencias de la Información de Universidad de Ljubljana en Eslovenia en apoyo con la comunidad de código abierto (University of Ljubljana, 2016).

Permite tanto la programación a través de secuencias de comandos en lenguaje Python, así como en formato gráfico, por lo tanto, no requiere un nivel avanzado de programación para su utilización. Su funcionalidad incluye los siguientes componentes (Demšar, 2016):

- Gestión y procesamiento previo de datos de entrada y salida, filtrado, toma de muestras, imputación y la selección de características

- Clasificación, con la implementación de varios algoritmos de aprendizaje automático supervisado como árboles de decisión, modelos bayesianos y reglas de inducción.

- Regresión, que incluye las técnicas de regresión multivariable, regresión lineal, regresión de mínimos cuadrados parciales y árboles de regresión

- Asociación, principalmente con las técnicas de reglas de asociación y conjuntos de elementos frecuentes

- Clustering, con la técnica de k-medias y el enfoque de agrupación jerárquica,

- Evaluación con validación cruzada y otros procedimientos basados en muestreo, funciones para calificar la calidad de los métodos de predicción y los procedimientos para la estimación de la fiabilidad.

Proyecciones con implementación de análisis de componente, escalamiento multidimensional y mapas de auto-organización.

7. MARCO METODOLÓGICO

7.1 METODOLOGÍAS IMPLEMENTADAS

Teniendo en cuenta la naturaleza estadística del proyecto, la investigación se define como Cuantitativa, particularmente como una investigación Ex-post-facto también llamada No Experimental, la cual, tiene como objetivo hallar posibles explicaciones de una situación en la que no se manipulan las variables, por lo tanto, son naturales y cercanas a la realidad cotidiana (Valenzuela y Flores, 2014).

Teniendo en cuenta los objetivos del proyecto, es necesario, identificar, modelar y procesar las diferentes variables académicas, financieras y psicológicas de los estudiantes de pregrado de primer ingreso a la UNAB, sin embargo, no se contempla un proceso de manipulación o alteración de las mismas, entendiéndose que a pesar que las variables serán formateadas para permitir su procesamiento, los valores que estas registran no serán modificados para la obtención de resultados experimentales.

Dos características de las investigaciones ex-post-facto son: el análisis del pasado poco remoto y la explicación causal de las variables identificadas. El análisis de los periodos académicos 2014-02, 2015-01, 2015-02 y 2016-01 al igual que la identificación de las causas de deserción de los estudiantes en base a esta información, hace que el proyecto cumpla con las dos características descritas (Valenzuela y Flores, 2014).

Dentro de los diseños de las investigaciones ex-post-facto se encuentra los longitudinales, los cuales, realizan observaciones en dos o más momentos en el tiempo, con el fin de establecer tendencias o el desarrollo evolutivo de un grupo (Valenzuela y Flores, 2014). De acuerdo a lo anterior, la metodología Ex-post-facto con diseño longitudinal fue seleccionada para el desarrollo del proyecto.

Desde la perspectiva de la Minería de Datos las metodologías CRISP-DM y SEMMA se imponen como las más representativas, siendo CRISP-DM la seleccionada, teniendo en cuenta que el proceso que describe en las fases de identificación del problema, compresión de los datos, preparación de los datos, modelado, evaluación y despliegue es más detallado y aplicable en diferentes contextos. SEMMA por su parte, resulta ser más restrictiva, ya que fue desarrollada por la empresa SAS y se enfoca en servir especialmente a sus productos.

7.2 HIPÓTESIS

Empleando métodos de Business Intelligence es posible aumentar la precisión en la identificación de estudiantes con riesgo de deserción en los programas de pregrado en comparación con el método empírico utilizado por la Universidad Autónoma de Bucaramanga.

7.3 POBLACIÓN Y MUESTRA

La población seleccionada es de tipo finita, conformada por los registros de estudiantes de primer ingreso de los programas de pregrado en los periodos 2014-02, 2015-01, 2015-02 y 2016-01. En total, se tienen 3370 registros, los cuales se encuentran discriminados por periodo académico en la siguiente tabla.

Tabla 9. Población de estudiantes por periodo académico

Periodo Académico	Total Estudiantes
2014-02	468
2015-01	1203
2015-02	501
2016-01	1198
TOTAL	**3370**

Fuente: Elaboración propia

La muestra se define de acuerdo a los diferentes análisis que serán realizados. Un primer análisis incluye el total de registros de estudiantes caracterizados, lo que la convierte en una muestra no aleatoria, teniendo en cuenta que los registros seleccionados tienen una característica común que los hace seleccionables. Los estudiantes caracterizados, son aquellos que cuentan con información en las dimensiones académicas, psicológicas y financieras. En general se obtiene una muestra de 2571 registros.

Los siguientes análisis incluyen la selección de igual número de estudiantes desertores y no desertores. Teniendo en cuenta que la cantidad de estudiantes desertores es mucho menor que la cantidad de estudiantes no desertores, serán seleccionados de forma aleatoria los registros de estudiantes no desertores. A diferencia de la muestra anterior, cada registro tiene la misma probabilidad de ser seleccionado.

7.4 RECOLECCIÓN DE LA INFORMACIÓN

Los datos necesarios para realizar en análisis predictivo fueron solicitados a la UNAB cumpliendo las respectivas normas internas que exige la universidad. Para ello, fue necesario solicitar por escrito la información, a través de un formato de acta de compromiso frente al uso responsable de la información.

Se solicitaron los registros de primer ingreso a los programas de pregrado de los periodos 2014-02, 2015-01, 2015-02 y 2016-01 con la información académica, psicológica y financiera de los estudiantes. La UNAB recolecta la información financiera y parte de la información académica a través de un formulario que debe ser diligenciado por los estudiantes como parte de su proceso de admisión, esta información, es almacena en el sistema de información Banner.

La información académica también es obtenida a través del Sistema de Gestión Académica SIGA, que aporta el rendimiento académico de los estudiantes. Finalmente, la información Psicológica es obtenida por medio de un test que identifica la personalidad del estudiante y su ambiente psicosocial.

7.5 PLAN DE INVESTIGACIÓN

El plan de investigación está dividido en base a las diferentes fases que describe la metodología de minería de datos CRISP-DM. Dentro de cada fase, se aborda en detalle cada una de las actividades que se desarrollan a lo largo de proceso investigativo, iniciando con la identificación del problema hasta el análisis y la documentación de los resultados obtenidos, sin embargo, en el proceso investigativo en general se desarrollaron una serie de actividades previas que orientaron el desarrollo adecuado del proyecto y plantea en conjunto las siguientes actividades para la obtención de los resultados:

- Anteproyecto
- Identificación del problema
- Desarrollo del marco teórico
- Marco metodológico
- Aprendizaje de las herramientas de minería de datos
- Implementación de la metodología CRISP-DM
- Solicitud de los datos
- Análisis de la información
- Procesamiento
- Análisis de resultados
- Identificación de las variables que determinan la deserción

- Comparación de resultados del modelo BI contra el modelo empírico de la UNAB
- Conclusiones

7.5.1 Aplicación de la metodología CRISP-DM

La metodología CRISP-DM, define seis fases que inician con el conocimiento del negocio y finaliza con el despliegue y presentación de los documentos que describen de forma clara el conocimiento adquirido. Cada una de estas fases es abordad como parte del proceso de minería de datos desarrollado, siendo la evaluación de los datos y el análisis de resultados las de mayor detalle en el proceso.

Conocimiento del Negocio
La compresión del problema desde el punto de vista de la minería de datos se ha realizado a lo largo de la investigación. Independientemente de los resultados finales, es posible la aplicación de técnicas de clasificación que permitan identificar de forma temprana el riesgo de deserción de un estudiante en base a la información académica, psicológica y financiera que registra.

Compresión de los Datos
Esta fase incluye la solicitud de los datos a la UNAB y la familiarización con las variables recolectadas. Se identifican los posibles problemas de calidad de los datos, como campos vacíos, valores nulos o incorrectos y se definen estrategias para mitigar su impacto en los resultados finales.

Preparación de los Datos
Con los problemas identificados en la fase anterior, se deben limpiar y seleccionar aquellos registros que cuenten con información valida de análisis. Esta fase, en conjunto con la anterior, permiten que los resultados de las herramientas de procesamiento no se vean afectados por la calidad de los datos.

Modelado
Es importante no confundir esta fase con la fase anterior. La fase de modelado depende del software de minería de datos seleccionado, la fase de preparación depende netamente de los datos obtenidos. Para que los datos sean procesados, es necesario que puedan ser interpretados por el software de minería de datos. En esta fase, los datos recibidos que se encuentran en formato de archivo Excel (xlsx), serán exportados inicialmente a formato csv y luego al formato arff, formato de archivos Weka.

Evaluación
En esta fase, los datos previamente preprocesados y modelados, se evalúan a través de la herramienta Weka. De acuerdo a los resultados, se define si es necesario realizar un nuevo procesamiento, con ajustes en la selección de las muestras procesadas.

Despliegue
Esta fase es abordada en el desarrollo mismo del proyecto. Incluye la documentación de las fases previas y el análisis de los resultados obtenidos.

7.5.2 Descripción de los mecanismos de análisis de datos

Con el proceso de minería de datos definido, es necesario establecer diferentes mecanismos de evaluación de los datos, a través de los cuales se pueda identificar, cuál de ellos arroja mejores resultados de predicción en base a una comparación de los porcentajes de aciertos obtenidos con el algoritmo de clasificación. A continuación, se describen los análisis definidos:

Análisis General de los Datos: Un primer análisis contempla el procesamiento de todos los 2571 registros de estudiantes caracterizados en un solo archivo a través de la opción de testeo Cross validation de Weka, la cual, permite definir un valor numérico que es utilizado para particionar la totalidad de los datos, definiendo cierta cantidad para datos de entrenamiento y los restantes como datos de prueba.

Análisis Por Periodo Académico: En este análisis, es necesario identificar la cantidad de registros de estudiantes desertores por periodo académico del total de 2571 datos obtenidos. Con esta cantidad, es necesario seleccionar una cantidad igual, de registros de estudiantes no desertores. La selección de los datos se realiza teniendo en cuenta los siguientes aspectos:

- La cantidad de registros de estudiantes desertores siempre es menor que la cantidad de registros de estudiantes no desertores.

- Los registros de estudiantes no desertores se seleccionan de forma aleatoria para no alterar la imparcialidad de los resultados.

La decisión de seleccionar igual cantidad de registros tanto de desertores como de no desertores, se hace con el objetivo de comparar los resultados obtenidos con los resultados del análisis general e identificar si esta estrategia es efectiva

para el mejoramiento de los porcentajes de predicción, teniendo en cuenta, que la cantidad desproporcionada entre ambos tipos de registros, puede afectar la calidad del procesamiento.

Adicionalmente, se deben construir dos conjuntos de datos, uno de ellos para el entrenamiento del algoritmo y otro para 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 acertada de los datos, se puede obtener un porcentaje de predicción. Los datos de entrenamiento y los datos de prueba deben contar con el mismo número de registros de estudiantes desertores y no desertores.

Análisis Por Periodo Académico Acumulado: A diferencia del análisis anterior, el análisis por periodo acumulado aprovecha el procesamiento histórico de los datos para intentar obtener porcentajes aceptables de predicción. Como parte de este proceso, se utilizan los datos históricos como datos de entrenamiento del algoritmo y los datos de prueba son seleccionados del periodo analizado.

Teniendo en cuenta que no se tienen datos anteriores al periodo 2014-02, no es posible realizar el análisis sobre este periodo académico. En su lugar, sus registros serán 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, serán 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.

Análisis Por Periodo Académico Acumulado sin la Variable Rendimiento Académico: La variable rendimiento académico es obtenida una vez el periodo de estudio ha finalizado. Esta variable representa un factor importante para analizar históricamente los factores de deserción, sin embargo, cuando se quiere identificar de forma temprana esta condición, no es posible contar con valores analizables desde el inicio del periodo académico.

De acuerdo a lo anterior, es necesario realizar un análisis que no incluya la variable rendimiento académico como parte del proceso predictivo, de manera que los resultados obtenidos brinden un beneficio en tiempo real, para la toma de decisiones de forma inmediata.

El proceso realizado en este caso, toma los mismos datos que se utilizan en el análisis anterior, eliminando la variable Rendimiento Académico de procesamiento. Al final, las variables que determinan la deserción en los dos análisis serán comparadas para identificar diferencias y porcentajes de predicción.

Análisis de las Variables Determinantes de Deserción: De acuerdo a los resultados obtenidos en los análisis descritos en los ítems anteriores, es necesario identificar cuáles de las variables procesadas son determinantes en la deserción de un estudiante. Para ello, las variables serán analizadas teniendo en cuenta dos factores: 1. El número de veces que la variable es identificada como variable determinante de deserción en los diferentes análisis realizados y 2. Por la cantidad de estudiantes que desertan por su causa.

Análisis Comparativo del Moldeo BI vs Resultados Empíricos: Este análisis contempla identificar los registros de los estudiantes evaluados en el análisis por periodo académico acumulado 2016-02 e identificar estos mismos registros en el análisis del modelo empírico de la UNAB. Con la misma población de registros en ambos modelos, se calcula el porcentaje de predicciones correctamente realizadas en base a la información real de deserción. Este proceso nos permite identificar cuál de los modelos es más acertado en sus predicciones.

8. RESULTADOS

En este apartado se presenta el proceso de minería de datos realizado a través de las fases que describe la metodología CRISP-DM, como el preprocesamiento de los datos, el modelado y su posterior evaluación y análisis, sin embargo, una etapa previa incluye el análisis y la elección del software de minería de datos y el algoritmo de clasificación de árboles de decisión. Al final, los resultados obtenidos en el proceso, son analizados y comparados contra los resultados del modelo empírico de la UNAB para determinar cuál de los dos modelos predictivos es más acertado.

8.1 SOFTWARE DE MINERÍA DE DATOS

El creciente interés de empresas, universidades e investigadores en general por el análisis de datos ha generado también el crecimiento de diferentes herramientas de análisis avanzado, también llamadas herramientas para la ciencia de datos, análisis estadístico, análisis predictivo etc. Si se tienen en cuenta los objetivos del proyecto, entre las herramientas más aplicables se encuentran Weka, RapidMiner y Knime, las cuales, comparten las siguientes características:

- Ofrecen una interfaz gráfica que facilita el flujo de trabajo en procesos de análisis de datos.

- Disponen de excelente documentación y grandes comunidades de usuarios.

- Están soportados por grandes empresas y universidades, por lo tanto, sus lanzamientos y actualizaciones son constantes.

- Son herramientas *Open Source*.

El análisis de Muenchen, (2017) publicado en 4stats, muestra la popularidad de diferentes herramientas de software de análisis de datos en diferentes contextos, entre ellos, la popularidad entre anuncios de empleo, artículos académicos publicaciones en revistas científicas, empresas de investigación de TI, encuestas de uso y foros de discusión.

Dentro de las categorías analizadas, Weka se destaca en anuncios de empleo, artículos académicos y encuestas de uso, por su parte, RapidMiner y Knime se destacan principalmente por su uso en empresas del sector de TI. Weka es especialmente popular en proyecto de ámbito académico, mientras que, RapidMiner y Knime son mayormente utilizados en ambientes comerciales, en el que además brindan versiones empresariales.

Teniendo en cuenta el ámbito académico del proyecto, Weka fue seleccionado como herramienta de Minería de Datos para el análisis de predicción, teniendo en cuenta, además que, a nivel de funcionalidad, en comparación con RapidMiner y Knime, las diferencias no son significativas, rescatando si, la posibilidad que tiene Weka para la integración con diferentes herramientas de análisis de datos, especialmente con el lenguaje R, que permite extender la funcionalidad natural que ofrece el software.

8.2 ALGORITMO DE CLASIFICACIÓN

Dentro de las diferentes técnicas de clasificación que ofrece Weka, se encuentra los arboles de decisión, reglas de asociación, funciones, bayes, etc. Según Han, Kamber y Pei, (2001), “los Arboles de decisión, son 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. El nodo superior en un árbol es el nodo raíz”.

El conjunto de algoritmos de árboles de decisión de Weka incluye los arboles *LMT, Decisión Stump, Random Forest, Random Tree, ReP Tree y J48*, este último, es la implementación del popular algoritmo ID3. Ekkachai, Jatsada, y Nittaya, (2012) en su artículo *Classification Model Indcution for Studing Recruiting* (Inducción del modelo de clasificación para el reclutamiento de estudiantes), el cual tiene como objetivo identificar cuáles son las características de deserción de los estudiantes admitidos en base a la información de ingreso, hace una comparación de diferentes algoritmos de árboles de decisión (J48, Naive Bayes, OneR) y concluye que los mejores porcentajes de clasificación son obtenidos por el algoritmo J48.

En base a los resultados del estudio de Ekkachai, Jatsada, y Nittaya, (2012) y teniendo en cuenta su similitud con los objetivos del actual proyecto, se selecciona el algoritmo J48 para realizar el proceso de clasificación de deserción de los estudiantes de primer ingreso a los programas de pregrado en la UNAB.

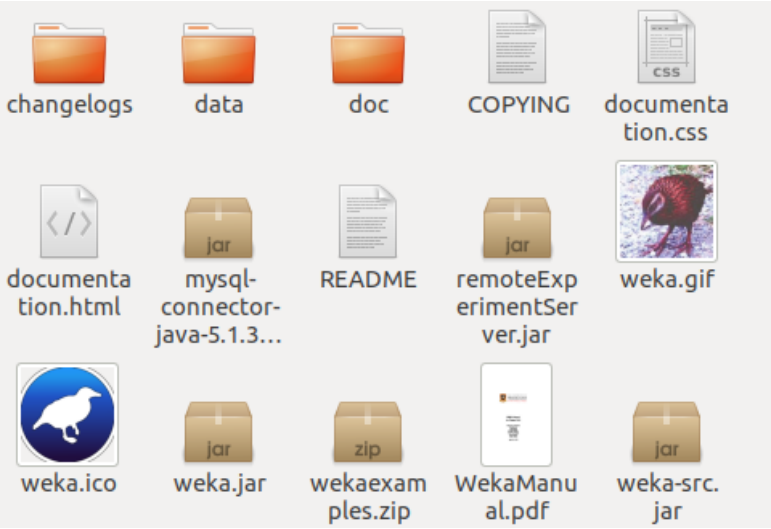
8.3 INTRODUCCIÓN A WEKA

Weka es un software que ofrece un conjunto de algoritmos de aprendizaje automático para tareas de minería de datos desarrollado por la universidad de Waikato. Su nombre proviene de un ave no voladora originaria de Nueva Zelanda (The University Of Waikato, s.f.). Actualmente puede ser descargado en su última versión estable 3.8 desde la url oficial: http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

8.3.1 Instalación

En sistemas Windows, puede ser instalado de forma tradicional, en sistemas Linux se descarga un directorio comprimido que contiene los archivos y carpetas que se observan en la siguiente imagen.

Figura 13. Estructura de directorios y archivos Weka



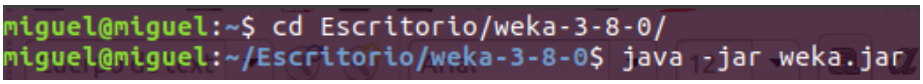
Fuente: Elaboración propia

En la imagen, se destacan las carpetas doc, que contiene la documentación del software, data, almacena los resultados de los análisis. El archivo Weka.jar es el ejecutable para iniciar la aplicación. El conector de base de datos de MySql es opcional, en caso de requerir conexiones a base de datos.

Para ejecutar Weka en sistemas Windows, solo se debe buscar el ejecutable a través del menú de aplicaciones. En sistemas Linux, es necesario acceder al

directorio de archivos descrito en la imagen anterior y ejecutar el siguiente comando:

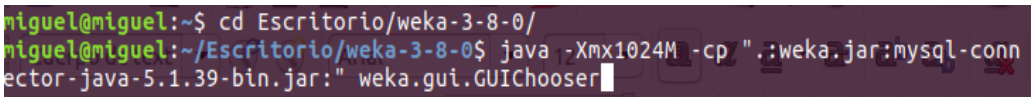
Figura 14. Comando de ejecución de Weka en sistemas Linux



Fuente: Elaboración propia

Si se requiere una conexión a bases de datos MySql, es necesario que el conector jdbc se encuentre dentro del directorio de archivos de Weka. El siguiente comando, habilita Weka para conexiones a bases de datos. La opción Xmx permite definir la cantidad máxima de memoria RAM que se puede usar para el procesamiento de los datos a través de las diferentes técnicas disponibles.

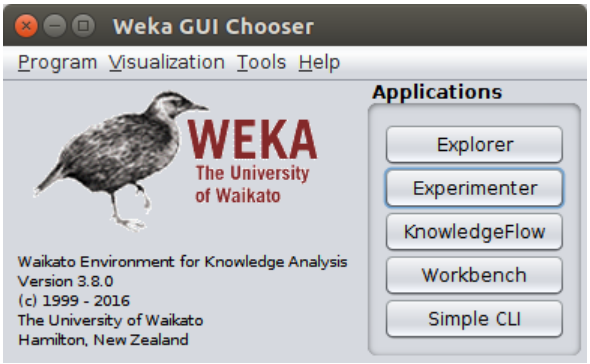
Figura 15. Comando de ejecución de Weka en sistemas Linux con conexión a base de datos MySql



Fuente: Elaboración propia

Una de las principales características de Weka, son las diferentes opciones gráficas que ofrece: *Explorer, Experimenter, KnowledgeFlow y WorkBench*. La siguiente imagen muestra la pantalla inicial de Weka.

Figura 16. Pantalla inicial de Weka con sus diferentes opciones gráficas. Fuente: Elaboración propia

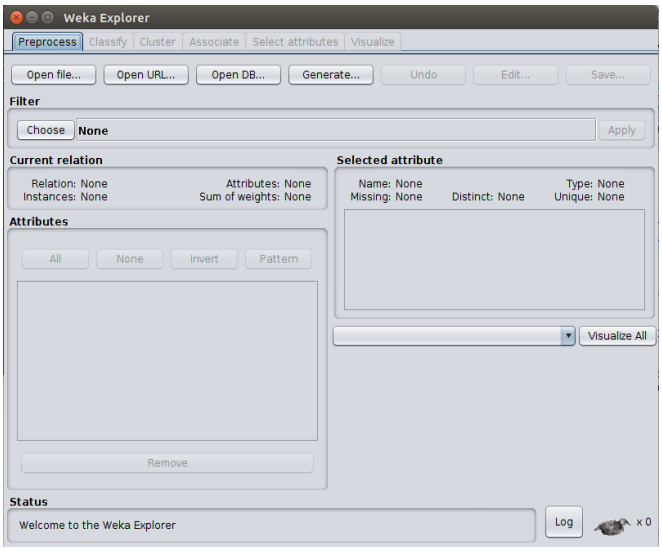


Fuente: Elaboración propia

8.3.2 Interfaz de Preprocesamiento

Las opciones de clasificación se encuentran en el botón *Explorer*. Además, esta opción incluye métodos de agrupamiento, asociación, selección de atributos, visualización y preprocesamiento de los datos. La siguiente imagen muestra la opción gráfica descrita.

Figura 17. Pantalla de exploración de Weka



Fuente: Elaboración propia

La primera pestaña permite realizar todo lo relacionado al preprocesamiento de los datos que serán analizados. La fila de botones de la parte superior facilita las opciones de importación de datos. Weka ofrece cuatro mecanismos:

- ***Open File***: Los datos son seleccionados desde el directorio de archivos local. Weka soporta diferentes formatos, entre los más destacados se encuentran el formato arff (formato por defecto de Weka), csv y json.

- ***Open Url***: Carga los datos desde una url externa.

- ***Open BD***: Consulta los datos desde una base de datos a través de un driver jdbc.

- ***Generate***: Permite generar los datos directamente desde la aplicación, estos datos vienen preestablecidos por la herramienta.

La opción Open File, soporta varios formatos de archivo, entre ellos, el formato arff. Este tipo de formato es propio de Weka, con una estructura dividida en dos bloques: el encabezado y los datos. El encabezado, describe los atributos que serán procesados y los datos son representados en una lista separados por comas. Un ejemplo de la estructura de un archivo arff se presenta a continuación:

```
@relation analisis_de_deserción_unab

@attribute genero {MASCULINO, FEMENINO}
@attribute tiene_hijos {SI, NO}
@attribute puntaje_icfes real
@attribute deserto {SI, NO}

@data
MASCULINO,NO,74,NO
FEMENINO,NO,85,NO
FEMENINO,SI,45,SI
MASCULINO,SI,75,SI
```

La anotación **@relation** es utilizada para dar nombre a la relación, puede ser cualquier texto, sin espacios; es solo un campo informativo. Los atributos pueden ser de varios tipos, a continuación, se describen los más representativos:

- **Nominal**: Son atributos a los cuales se les definen los diferentes valores que puede asumir la variable.

@attribute genero {MASCULINO, FEMENINO}

- **Real**: Son atributos que toman valores reales

@attribute puntaje_icfes real

- **Numeric**: Atributos que toman valores numéricos enteros

@attribute edad numeric

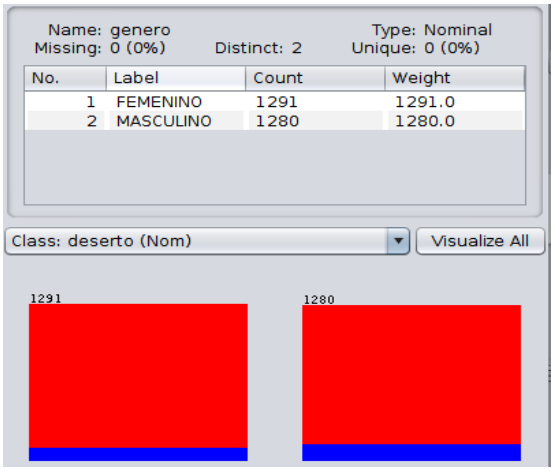
- **String**: Define valores de variable que son de tipo texto, pero no pueden ser predichas.

@attribute nombre_estudiante string

En el anexo A se puede consultar el formato de un archivo arff real, utilizado en el proyecto. Para importar los datos en Weka, es necesario hacer clic sobre el botón *Open File*, seleccionar el archivo arff y finalmente se da clic al botón Abrir.

Una vez cargado el archivo, Weka presenta un resumen visual de la información. Para cada variable, identifica el total de valores presentes y los muestra en una gráfica de barras, adicionando, la relación de los valores con la variable de deserción. Así, por ejemplo, en la gráfica siguiente se puede observar el número de registros por género, donde el color azul representa el número de estudiantes desertores en cada género y el color rojo los registros de no desertores.

Figura 18. Ejemplo de análisis gráfico de la variable género en la pestaña de preprocesamiento

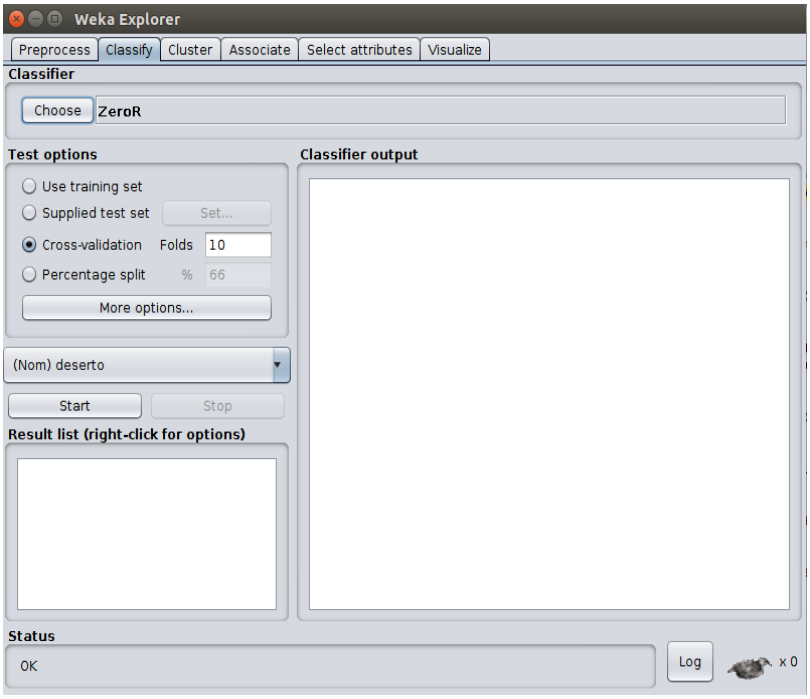


Fuente: Elaboración propia

8.3.3 Pestaña de Clasificación

En la pestaña clasificación Weka dispone de diferentes algoritmos y estrategias de testeo de los datos cargados previamente en la pestaña de preprocesamiento. La siguiente imagen muestra las opciones gráficas de la interfaz.

Figura 19. Pestaña de clasificación de Weka



Fuente: Elaboración propia

En la opción ***Classifier***, Weka permite seleccionar el algoritmo de clasificación dentro de un grupo de categorías que se describen a continuación (Remco y Eibe, 2016):

- ***Bayes***: Contiene el conjunto de clasificadores bayesianos, por ejemplo, Naive-Bayes.

- ***Functions***: Ofrece algoritmos de predicción con soporte de máquinas vectoriales, regresión y redes neuronales artificiales.

- ***Lazy***: Los algoritmos de este tipo, realizan el aprendizaje en el mismo momento en el que realizan el proceso de clasificación; por ejemplo, el algoritmo K-vecinos más cercanos.

- ***Meta***: Ofrece meta-clasificadores que utilizan uno o más clasificadores como entrada.

- ***Misc***: Ofrece una serie de algoritmos que no se ajustan a ninguna otra categoría.

- ***Rules***: Contiene una serie de algoritmos basados en reglas de clasificación, por ejemplo, ZeroZ.

- ***Trees***: Contiene un conjunto de algoritmos de árboles de decisión, por ejemplo, el algoritmo J48.

La opción ***Test Option*** permite seleccionar el modo de prueba que se quiere aplicar al proceso de clasificación. Weka ofrece cuatro modos de prueba:

- ***Use Training Set*****:** Permite que el algoritmo seleccionado sea clasificado de acuerdo a la totalidad de los datos cargados, es decir, los mismos datos se utilizan como datos de entrenamiento y como datos de prueba, por lo tanto, el resultado de clasificación es bastante optimista, sin embargo, no es recomendable su uso en entornos reales.

- ***Supplied test set*****:** Esta opción permite cargar de forma separada los datos de entrenamiento y los datos de prueba. Los datos de entrenamiento se cargan previamente en la pestaña Preprocess**,** mientras que los datos de prueba se cargan a través del botón "set" que se habilita una vez que se selecciona la opción.

- ***Cross-validation*****:** En este caso, el algoritmo de clasificación se evalúa por medio de validación cruzada, utilizando el valor del campo **fold**, como el número de pliegues.

- ***Percentage Split*****:** El algoritmo de clasificación es evaluado con un porcentaje de datos para el entrenamiento y otro porcentaje para las pruebas. El valor numérico que es ingresado, representa el porcentaje de los datos que serán utilizados para las pruebas.

Variable de Clasificación

Bajo la opción "*Test Option*" Weka permite seleccionar cuál de las variables cargadas será utilizada como variable de clasificación. Por defecto, es seleccionada la última de las variables que aparece en el archivo cargado previamente, esta puede ser cambiada haciendo clic sobre el elemento

desplegable. Hay que resaltar que algunos algoritmos requieren que la variable de clasificación sea de tipo nominal, otros solo aceptan variables numéricas y algunos aceptan los dos tipos de variable.

La opción ***Result List*** muestra el listado de las evaluaciones realizadas. A través del clic derecho, ofrece un menú de opciones que incluye:

- *View in main window* (Ver los resultados en la misma ventana)
- *View in separate window* (Ver los resultados en una ventana nueva)
- *Save result buffer* (Permite guardar los resultados en un archivo de texto)
- *Load model* (Permite cargar resultados guardados previamente)
- *Save model* (Permite guardar los resultados en un archivo binario, utilizando el Serialized Object de Java)
- *Visualize tree* (Permite ver gráficamente el árbol generado en el proceso de clasificación)

8.3.4 Resultados Weka

Finalmente, el botón ***start*** permite iniciar el proceso de clasificación de acuerdo a los parámetros previamente configurados. Los resultados se analizan en detalle en la sección de análisis de resultados del proyecto.

8.4 CONOCIMIENTO DEL NEGOCIO

La fase inicial del proceso de Minería de Datos inicia con la **comprensión del negocio**, de tal manera que este pueda ser entendido como un problema de minería. En este punto, esta fase ha sido abordad teniendo en cuenta el proceso hasta ahora realizado.

8.5 COMPRENSIÓN DE LOS DATOS

La compresión de los datos, es definida por CRISP-DM como la fase de recolección de los datos, para luego realizar un proceso de familiarización. Como parte de este proceso, fueron solicitados a la UNAB los registros de primer ingreso en los programas de pregrado de los periodos 2014-02, 2015-01, 2015-02 y 2016-01 con la información académica, psicológica y financiera de los estudiantes.

Para cada una de las dimensiones analizadas fue necesario conocer las variables asociadas, los diferentes valores aceptados e identificar cuáles de ellas contaban con información valida de análisis. Se inició el proceso con las variables académicas, las cuales se describen en la siguiente tabla.

Tabla 10. Variables Académicas Seleccionadas. Parte 1

Tipo de Variable	Variable	Valores Aceptados
Académicas	Genero	Masculino
		Femenino
	Población Especial	Discapacidad Motriz
		Discapacidad Cognitiva
		Discapacidad Sensorial
		Desplazado
		Afrodescendiente
		Indígenas
		Habitante de Frontera
	Nivel Obtenido en la Competencia Comunicativa	Superior
		Alto
		Medio
		Bajo
		Inferior
	Nivel Obtenido en la Competencia Matemática	Superior
		Alto
		Medio
		Bajo
		Inferior

Fuente: Elaboración propia

Tabla 11. Variables Académicas Seleccionadas Parte 2

Tipo de Variable	Variable	Valores Aceptados
Académicas	Puntaje ICFES	Más de 50 Puntos
		40 – 50 Puntos
		0 – 40 Puntos
	Tiempo transcurrido entre la última graduación y su ingreso a la UNAB.	0-2 Años
		2-4 Años
		4-6 Años
		Más de 6 años
	Porcentaje de Cursos Aprobados (Rendimiento Académico)	0% ≤ Cursos Aprobados < 20%
		20% ≤ Cursos Aprobados < 50%
		50% ≤ Cursos Aprobados < 70%
		70% ≤ Cursos Aprobados < 90%
		90% ≤ Cursos Aprobados < 100%
	Ser Pilo Paga	P (SI)
		Vacío (NO)

Fuente: Elaboración propia

Teniendo en cuenta que no todas las variables contaban con información o la información que registraban no aportaba ningún valor predictivo, algunas de ellas fueran descartadas del análisis de Minería de Datos. Estas variables se describen a continuación en la tabla 12.

En el caso de la variable **Porcentaje de Insistencia**, los datos recibidos no registraban ningún valor en los semestres analizados. En el caso de la variable **Estrategias de Aprendizaje**, todos los registros tenían asignado el mismo valor, por lo tanto, deja de ser considerada una variable. Seguidamente, fueron analizadas las variables financieras (ver tabla 13).

Tabla 12. Variables Académicas Descartadas

Tipo de Variable	Variable	Valores Aceptados
Académicas	Porcentaje de Inasistencia	0% – 10%
		11% - 20%
		21% - 50%
		51% - 100%
	Estrategias de Aprendizaje	Metacognitiva
		Procesamiento de Información
		Manejo de Recursos
		Disposicionales

Fuente: Elaboración propia

Tabla 13. Variables Financieras Seleccionadas Parte 1

Tipo de Variable	Variable	Valores Aceptados
Financieras	Lugar de Procedencia	Área metropolitana de Bucaramanga
		Fuera del área metropolitana de Bucaramanga
	Estado Civil	Casado
		Separado – Divorciado
		Viudo
		Unión Libre
		Soltero
	Número de Hijos	Ninguno
		Uno
		Dos
		Tres o más
	Tiene Personas a Cargo	Si
		No
	Estrato	1
		2
		3
		4
		5
		6

Fuente: Elaboración propia

Tabla 14. Variables Financieras Seleccionadas Parte 2

Tipo de Variable	Variable	Valores Aceptados
Financieras	Con Quien Vive	Familia
		Fuera del hogar
	Trabaja Actualmente	Si
		No
	Tipo de Vinculación Laboral	Independiente
		Empleado
	Sector Productivo	Servicios
		Industriales
		Manufactura
	Cargo Que Ocupa	Directivo
		Jefatura
		Coordinador
		Supervisor
		Operario
	Salario	Menos de un salario mínimo
		Entre 1 y 2 salarios mínimos
		3 salarios mínimos o más
	Horario Laboral	Horario de oficina
		Turnos
		Nocturno
		Fines de semana
	Fuente de Pago de Estudios	Recursos Propios
		ICETEX
		Otra entidad
	Tipo de Colegio	Público
		Privado

Fuente: Elaboración propia

En este caso, los datos obtenidos contaban con información valida en su totalidad, a pesar de que no todos los registros se encontraban diligenciados, se consideran variables analizables aquellas que reporten información en al menos un porcentaje, por lo tanto, todas las variables fueron aceptadas para el proceso. Finalmente, se analizaron las variables psicológicas, las cuales se describen a continuación:

Tabla 15. Variables Psicológicas Descartadas Parte 1

Tipo de Variable	Variable	Valores Aceptados
Psicológicas	Personalidad indicador clínico - **1ª**	1
		0
	Personalidad indicador clínico - **2B**	1
		0
	Personalidad indicador clínico – **3C**	1
		0
	Personalidad indicador clínico – **4E**	1
		0
	Personalidad indicador clínico – **5F**	1
		0
	Personalidad indicador clínico – **6G**	1
		0
	Personalidad indicador clínico – **7H**	1
		0
	Personalidad indicador clínico – **8I**	1
		0
	Personalidad indicador clínico – **9L**	1
		0
	Personalidad indicador clínico – **10M**	1
		0
	Personalidad indicador clínico – **11N**	1
		0
	Personalidad indicador clínico – **12º**	1
		0
	Personalidad indicador clínico – **13Q**	1
		0
	Personalidad indicador clínico – **14R**	1
		0
	Personalidad indicador clínico – **15S**	1
		0
	Personalidad indicador clínico – **16T**	1
		0

Fuente: Elaboración propia

Tabla 16. Variables Psicológicas Seleccionadas Parte 2

Tipo de Variable	Variable	Valores Aceptados
	Psico-transtornos del comportamiento (**TC**)	1
		0
	Psico-riesgos para la salud (**RS**)	1
		0
	Psico-desordenes psiquiátricos (**DPS**)	1
		0
	Psico-autoafirmación social (**AUS**)	1
		0
	Psico-disfunciones familiares (**DF**)	1
		0
	Psico-rendimiento escolar (**RE**)	1
		0
	Psico-desordenes laborales (**DL**)	1
		0
	Psico-presión del grupo de amigos (**PGA**)	1
		0
	Psico-uso del tiempo libre (**UTL**)	1
		0
	Psico-severidad del uso de drogas (**SUD**)	1
		0

Fuente: Elaboración propia

En la dimensión psicológica, fue necesario descartar las variables que no registraban ningún valor valido en los diferentes semestres analizados. Todos sus valores de respuesta se encontraban vacíos, algunos de ellos con valores no validos o con información nula. Estas variables se describen a continuación:

Tabla 17. Variables Psicológicas Descartadas

Tipo de Variable	Variable	Valores Aceptados
Psicológicas	Personalidad indicador clínico – I1	1
		0
	Personalidad indicador clínico – I2	1
		0
	Personalidad indicador clínico – I3	1
		0
	Personalidad indicador clínico – I4	1
		0
	Personalidad indicador clínico – I5	1
		0
	Personalidad indicador clínico – NRT	1
		0
	Personalidad indicador clínico – PST	1
		0
	Personalidad indicador clínico – PAA	1
		0

Fuente: Elaboración propia

De los registros obtenidos, se descartaron aquellos que no correspondían a programas de pregrado. Se consideraron solo los registros de los siguientes programas académicos:

Tabla 18. Programas Académicos de Pregrado seleccionados para el Análisis

Programa Académico	Código
Artes Audiovisuales	AAU
Administración Hotelera	ADH
Administración de Empresas	ADM
Comunicación Social	COM
Derecho	DER
Economía	ECO

Fuente: Elaboración propia

Tabla 18. Programas Académicos de Pregrado seleccionados para el Análisis (continuación)

Programa Académico	Código
Enfermería	ENF
Ingeniería en Energía	IEE
Ingeniería Biomédica	IBM
Ingeniería Financiera	IFI
Ingeniería Mecatrónica	IMK
Ingeniería de Mercados	IMR
Ingeniería de Sistemas	ISI
Medicina	MED
Música	MUS
Negocios Internacionales	NGI
Psicología	PSI
Gestión de Sistemas de Información	GSI
Contaduría Pública	COP
Licenciatura en educación preescolar	EPR
Ingeniería Biomédica	IBM

Fuente: Elaboración propia

Los programas académicos de pregrado que no tenían registros analizables, no aparecen en la tabla. Al discriminar los datos por periodo académico, se obtienen los registros requeridos de acuerdo a los objetivos del proyecto. La siguiente tabla presenta este resumen:

Tabla 19. Relación de registros obtenidos por periodo académico

Periodo Académico	Total Estudiantes	Desertores	No Desertores
2014-02	468	83	385
2015-01	1203	116	1087
2015-02	501	79	422
2016-01	1198	155	1043
TOTAL	**3370**	**246**	**2325**

Fuente: Elaboración propia

De acuerdo a la metodología CRISP-DM, el proceso de **comprensión de los datos** continúa con la identificación de problemas de calidad. Parte de este proceso fue realizado en la identificación de las variables. A continuación, se describen otros problemas identificados:

- **Calidad de los datos**
  Se identifica que los registros cuentan parcialmente con la información académica, psicológica y financiera de los estudiantes, algunos campos son vacíos, nulos o con valores inválidos.

  Los registros cuentan con un campo llamado **Caracterización,** el cual, identifica si un estudiante se encuentra o no caracterizado. Aquellos estudiantes que registran el valor "NO CARACTERIZADO", no cuentan con ningún tipo de información, por lo tanto, se descartan del proceso de minería de datos.

- **Dispersión de la información**
  La información recibida se encuentra en diferentes archivos en formato Excel, los cuales, presentan de forma separada los estudiantes matriculados, la información académica, la información financiera, la información psicológica y la información de deserción.

8.6 PREPARACIÓN DE LOS DATOS

De acuerdo al análisis realizado en la fase anterior, fue necesario consolidar la información en un único archivo de Excel, seleccionando solo aquellos estudiantes CARACTERIZADOS, relacionando la información a través del código del estudiante, el programa y el periodo académico. El resultado se resume en la siguiente tabla, donde se resume el total de estudiantes caracterizados por cada periodo académico y el número de estudiantes desertores y no desertores.

Tabla 20. Relación de registros caracterizados por periodo académico

Periodo Académico	Total Estudiantes	Desertores	No Desertores
2014-02	272	30	242
2015-01	944	71	873
2015-02	356	40	316
2016-01	999	105	894
TOTAL	**2571**	**246**	**2325**

Fuente: Elaboración propia

8.7 MODELADO DE LOS DATOS

La fase de modelado depende de cada uno de los análisis realizados. Dependiendo de la estrategia de análisis los datos pueden ser almacenados y procesados a través de un único archivo arff o, por el contrario, puede ser necesaria la generación varios archivos, algunos para entrenar el algoritmo y otros para la realización de las pruebas.

Partiendo de los resultados del proceso de preparación de los datos, en donde fue identificado un problema de dispersión de la información académica, psicológica y financiera, fue necesario consolidar todos los registros en un único archivo Excel, el cual, posteriormente fue exportado a formato csv. Este formato es aceptado por Weka, sin embargo, no es flexible para realizar el procesamiento. Desde Weka el archivo csv es exportado al formato arff. Este formato fue el utilizado en los diferentes análisis realizados.

En el primer análisis, solo fue necesaria la generación de un archivo arff con la totalidad de los registros caracterizados. En el segundo análisis, fue necesaria la generación de ocho archivos. Cuatro de ellos para el entrenamiento del algoritmo (uno por periodo académico) y los 4 restantes para las pruebas (uno por periodo académico).

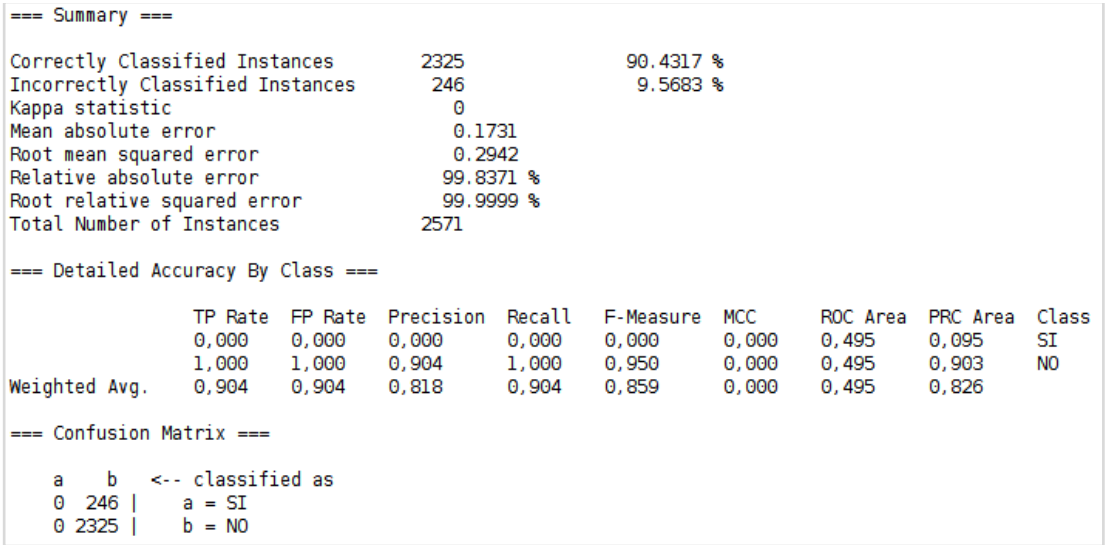
En el tercer análisis, al igual que en el segundo, se necesitaron nuevamente 8 archivos arff, sin embargo, los cuatro archivos de entrenamiento almacenan no solo la información de un único periodo académico, sino la información de los periodos anteriores al periodo de análisis.

Finalmente, el último análisis realizado, hace uso de los ocho archivos del análisis anterior, descartando en cada uno de ellos la variable rendimiento académico.

8.8 ANÁLISIS GENERAL DE LOS DATOS

Teniendo en cuenta la fase de modelado realizada, fue generado el archivo arff llamado ALL_DATA_CARACTERIZADOS.arff, el cual, fue obtenido a partir del archivo Excel con la totalidad de registros caracterizados. En general, fueron procesados los 2571 registros. En la siguiente imagen se pueden observar los resultados de predicción arrojados por Weka.

Figura 20. Resultados Weka al procesar la totalidad de datos caracterizados



```
=== Summary ===

Correctly Classified Instances        2325               90.4317 %
Incorrectly Classified Instances       246                9.5683 %
Kappa statistic                          0
Mean absolute error                      0.1731
Root mean squared error                  0.2942
Relative absolute error                 99.8371 %
Root relative squared error             99.9999 %
Total Number of Instances             2571

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000    0,495     0,095     SI
                 1,000    1,000    0,904      1,000    0,950      0,000    0,495     0,903     NO
Weighted Avg.    0,904    0,904    0,818      0,904    0,859      0,000    0,495     0,826

=== Confusion Matrix ===

    a    b   <-- classified as
    0  246 |    a = SI
    0 2325 |    b = NO
```

Fuente: Elaboración propia

Como se observa, Weka no fue capaz de reconocer ningún patrón de predicción 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 aquellos estudiantes no desertores, los estudiantes desertores tuvieron una clasificación nula.

La matriz de confusión que se observa al final de la imagen puede ser interpretada de acuerdo a los valores mostrados en la siguiente tabla, donde se identifican los datos clasificados correcta e incorrectamente, en los registros de estudiantes desertores y no desertores.

Tabla 21. Interpretación de la matriz de confusión del análisis general de los datos

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	2325 (90.43%)	2325	0
Desertores	246 (9.56%)	0	246
Total	2571 (100%)	2325	246

Fuente: Elaboración propia

Claramente los resultados obtenidos no fueron los esperados. Analizando en detalle los registros para identificar las causas de la pobre clasificación realizada por Weka se identifica el siguiente problema: Desproporción de los registros de los estudiantes no desertores en relación a los registros de estudiantes desertores. Mientras que los primeros corresponden a un 90.43%, los segundos solo representan el 9.56% del total de los datos.

Advirtiendo esta situación, los siguientes análisis realizados contemplan diferentes estrategias que tienen como objetivo obtener mejores resultados en el porcentaje de clasificación de estudiantes desertores.

8.9 ANÁLISIS POR PERIODO ACADÉMICO

De acuerdo al resultado anterior, fue necesario idear una estrategia que permitiera tener una proporción más equilibrada de los registros. Como describe la metodología CRISP-DM, el procesamiento de los datos puede requerir que algunas fases sean realizadas nuevamente para mejorar las predicciones finales. Se realizó una nuevamente la fase de comprensión de los datos para identificar una estrategia de análisis adecuada, en la cual, se definió:

- La cantidad de registro de los estudiantes desertores es muy inferior en cada uno de los periodos analizados.

- Es necesario utilizar la totalidad de los datos de los estudiantes desertores en el proceso de clasificación en Weka.

- De la población de estudiantes no desertores, se debe seleccionar el mismo número de estudiantes desertores, con el fin de lograr igual proporción de los datos. Los registros de los estudiantes no desertores se deben seleccionar aleatoriamente.

- Inicialmente, el procesamiento de los datos se debe realizar individualmente por periodo académico, después, se debe realizar acumulativamente con los datos seleccionados del periodo inmediatamente anterior.

  2014-02, 2015-01
  2014-02, 2015-01, 2015-02
  2014-02, 2015-01, 2015-02, 2016-01

- Se deben definir dos poblaciones de datos, una para entrenar el algoritmo (datos de entrenamiento) y otra para la realización de las pruebas (datos de prueba), lo que brinda un mayor control sobre la información.

De acuerdo a las observaciones anteriores, se definió la siguiente tabla, en la cual, se seleccionaron la cantidad de registros dispuestos para el análisis de acuerdo al total de registros de estudiantes desertores.

Tabla 22. Tamaño de la muestra seleccionada por periodo académico para equilibrar las poblaciones en el modelo de predicción

Periodo	Total Desertores	No Desertores Seleccionados	Total Periodo
2014-02	30	30	60
2015-01	72	72	144
2015-02	40	40	80
2016-01	105	105	210
Total	**247**	**247**	**494**

Fuente: Elaboración propia

A continuación, en la descripción del proceso del periodo académico 2014-02 se detalla la forma en la que se obtuvieron los nuevos datos para el análisis. Este proceso, fue igualmente replicado para los demás periodos analizados.

8.9.1 Análisis Periodo 2014-02

Para el modelado de los datos por periodo académico se tomó como base el archivo ALL_DATA_CARACTERIZADOS.xlsx. A partir de este archivo, se seleccionaron solo los registros correspondientes al periodo 2014-02. La siguiente tabla muestra el total de registros obtenidos, discriminados por desertores y no desertores.

Tabla 23. Total registros caracterizados del periodo 2014-02

Clasificación	Cantidad
Estudiantes Desertores	30
Estudiantes No Desertores	242
Total Estudiantes	**272**

Fuente: Elaboración propia

A partir de la tabla anterior, fueron creados los archivos **DESERTORES_201402.xlsx** y **NO_DESERTORES_201402.xlsx**. Teniendo en cuenta que solo se tienen 30 registros de estudiantes desertores, fue necesario seleccionar solo 30 de los 242 registros de estudiantes No Desertores. Este proceso fue realizado a través de selección aleatoria. Los 60 registros obtenidos fueron divididos, esta vez para disponer 30 registros para datos de entrenamiento y otros 30 registros como datos de prueba. Cada uno de estos 30 registros contaba con 15 datos de estudiantes no desertores y 15 datos de estudiantes desertores. La siguiente tabla muestra la relación de registros ajustados.

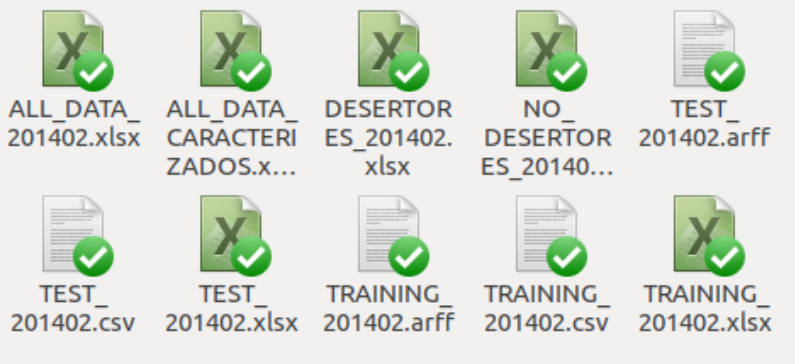
Tabla 24. Total registros del periodo 2014-02 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	15	15
No Desertores	15	15
Total Registros	**30**	**30**

Fuente: Elaboración propia

Estos registros fueron igualmente almacenados en dos archivos Excel, **TEST_201402.xlsx** y **TRAINING_201402.xlsx**, sin embargo, para poder cargar la información en Weka, es necesario exportar los archivos al formato csv. Desde la pestaña ***Preprocess*** estos archivos son guardados en formato arff. Este proceso es necesario ya que algunas de las variables en los archivos csv son interpretadas por Weka como variables de tipo String, las cuales, no pueden ser analizadas por los algoritmos de clasificación. La siguiente imagen muestra el directorio final de archivos obtenidos para realizar el proceso de clasificación a través de Weka.

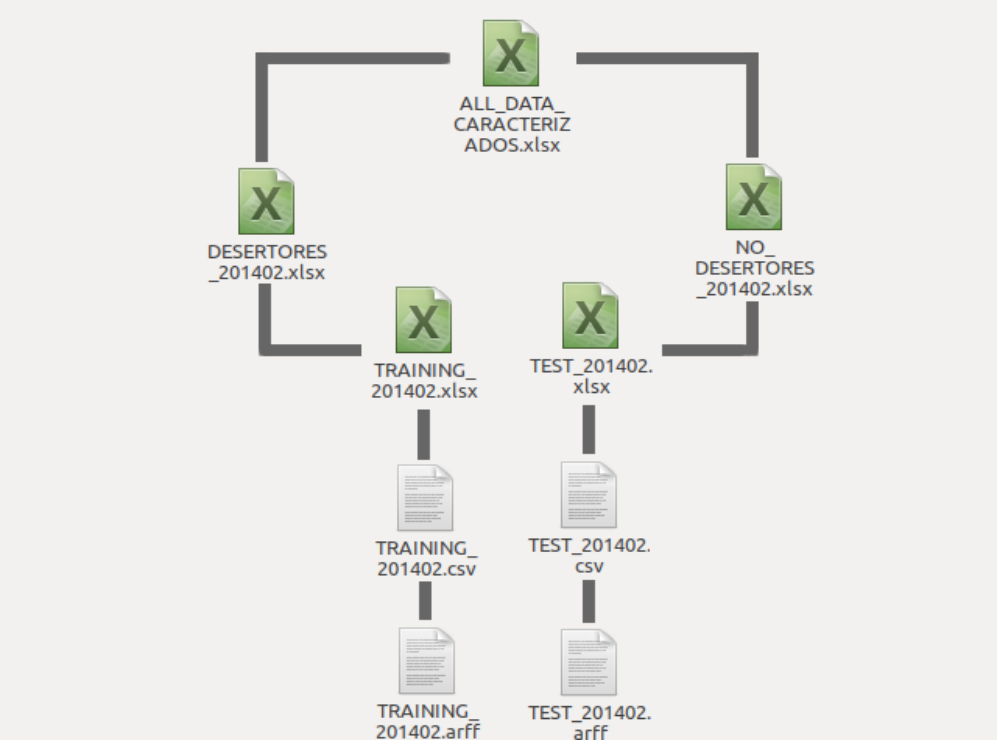
Figura 21. Archivos obtenidos durante la generación de los archivos arff



Fuente: Elaboración propia

Para entender el proceso de generación de los archivos arff con la información de los estudiantes caracterizados, la siguiente imagen muestra un mapa que describe la secuencia realizada:

Figura 22. Proceso de obtención de los archivos arff

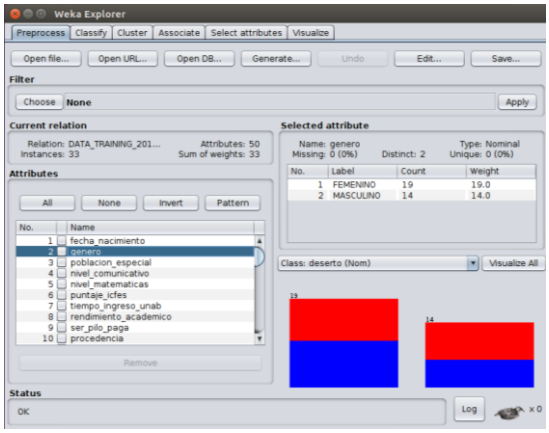


Fuente: Elaboración propia

El archivo TRAINING_201402.arff que contiene la información de los 30 registros de entrenamiento, fue cargado a través de la pestaña ***Preprocess*** de Weka. Es importante tener en cuenta que los datos de entrenamiento deben ser cargados desde esta pestaña. Los datos de prueba, se cargan desde la pestaña ***Classify***.

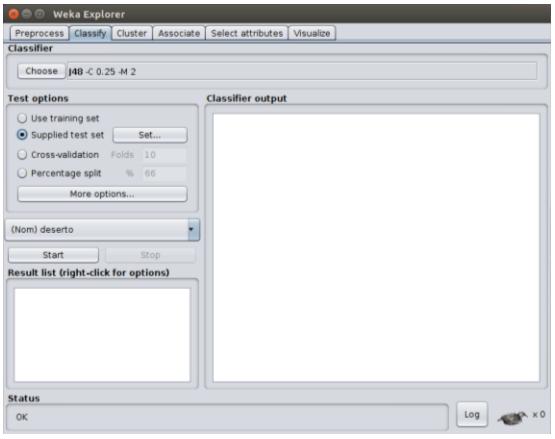
En la pestaña ***Classify*** inicialmente se debe seleccionar el algoritmo a utilizar. Para ello, el botón ***Choose*** muestra el conjunto de algoritmos disponibles. Se seleccionó la categoría **Trees** y el algoritmo **J48**. Adicionalmente, en la sección ***Test Option*** se seleccionó la opción ***Supplied Test Set***, la cual, habilita el botón ***set*** que permite seleccionar el archivo TEST_201402.arff con los datos de prueba. Las siguientes imágenes muestran el proceso hasta ahora realizado.

Figura 23. Cargue de los datos de entrenamiento desde la pestaña Preprocess.



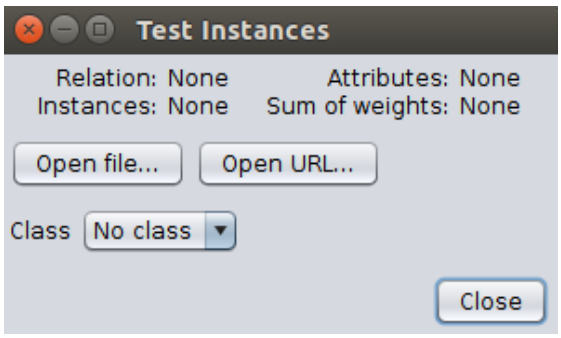
Fuente: Elaboración propia

Figura 24. Selección del algoritmo de clasificación de árboles de decisión J48



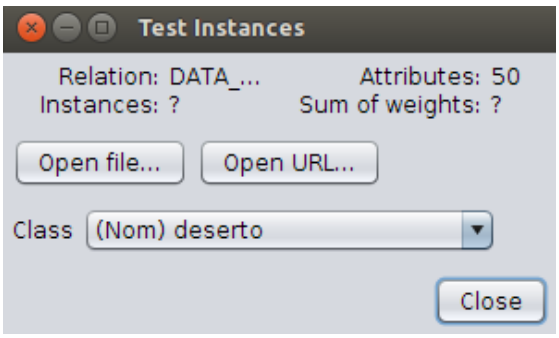
Fuente: Elaboración propia

Figura 25. Ventana para el cargue del archivo con los datos de prueba



Fuente: Elaboración propia

Figura 26. Ventana con el archivo seleccionado



Fuente: Elaboración propia

Una vez cargados los datos de entrenamiento y los datos de prueba, se procedió a realizar la evaluación de los datos. Haciendo clic en el botón **Start,** Weka inicia el proceso de evaluación. La siguiente imagen muestra los resultados obtenidos.

Figura 27. Resultados Weka con el algoritmo J48 en el periodo 2014-02

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

puntaje_icfes = 40-50 PUNTOS: SI (8.0)
puntaje_icfes = MAS DE 50 PUNTOS
|   rendimiento_academico = ENTRE 70% Y 90%: NO (3.0)
|   rendimiento_academico = ENTRE 50% Y 70%: NO (0.0)
|   rendimiento_academico = 100% APROBADOS: NO (16.0/2.0)
|   rendimiento_academico = MENOS DE 20%: SI (3.0)
|   rendimiento_academico = ENTRE 20% Y 50%: NO (0.0)
puntaje_icfes = 0-40 PUNTOS: SI (3.0)

Number of Leaves  :     7

Size of the tree :      9

=== Summary ===

Correctly Classified Instances          29               96.6667 %
Incorrectly Classified Instances         1                3.3333 %
Kappa statistic                          0.9333
Mean absolute error                      0.075
Root mean squared error                  0.1963
Relative absolute error                 15      %
Root relative squared error             39.248  %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0,933    0,000    1,000      0,933    0,966      0,935    0,944     0,967
                 1,000    0,067    0,938      1,000    0,968      0,935    0,944     0,903
Weighted Avg.    0,967    0,033    0,969      0,967    0,967      0,935    0,944     0,935

=== Confusion Matrix ===

  a  b   <-- classified as
 14  1 |  a = SI
  0 15 |  b = NO
```

Fuente: Elaboración propia

Del análisis textual se puede observar que el **Puntaje ICFES** y el **Rendimiento Académico** figuran como variables involucradas en la determinación de deserción de un estudiante. En general, se generó un análisis de 7 niveles con un árbol de tamaño 9. De los 30 datos analizados, 29 registros fueron clasificados correctamente, mientras que solo un registro se clasificó incorrectamente. La matriz de confusión obtenida que se observa en la imagen, se describe en detalle en la siguiente tabla:

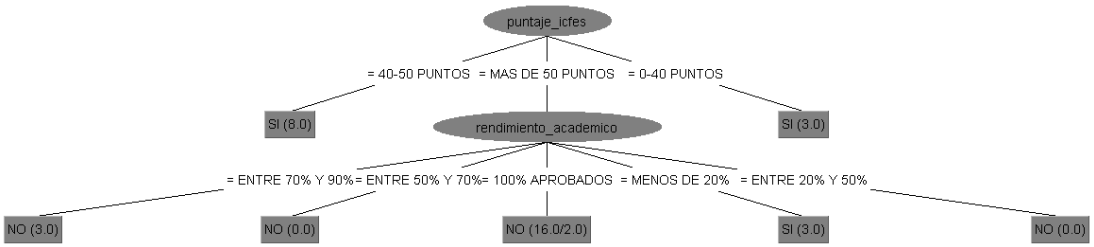
Tabla 25. Interpretación de la matriz de confusión de los resultados del periodo 2014-02

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
Desertores	15 (50%)	14	1
No Desertores	15 (50%)	15	0
Total	30 (100%)	29	1

Fuente: Elaboración propia

Para facilitar la comprensión general del análisis realizado por Weka con el algoritmo J48, se dispone de una serie de opciones por resultado. Una de estas, es la visualización gráfica del árbol generado. La siguiente imagen muestra el resultado obtenido.

Figura 28. Árbol de decisión Weka del análisis del periodo 2014-02 con el algoritmo J48



Fuente: Elaboración propia

De acuerdo a la imagen, se puede observar que la principal variable que determina la deserción de un estudiante es el Puntaje ICFES. Para puntajes entre los 40 y 50 puntos, 8 de los 15 estudiantes desertaron. Para puntajes entre 0 y 40 puntos, 3 de los 15 estudiantes desertaron.

Cuando los estudiantes obtuvieron un puntaje mayor a 50 puntos, se debe considerar el rendimiento académico. Si el rendimiento académico fue mayor al 50%, el estudiante no desertó, sin embargo, si el rendimiento académico se encuentra entre el 20% y 50%, los estudiantes registran un total de 3 deserciones del total de 15 registros analizados.

En general, por puntaje ICFES desertaron 11 estudiantes, mientras que por rendimiento académico y puntaje ICFES desertaron 3 estudiantes. La suma de

estudiantes desertores es de 14, que corresponde con el total de registros clasificados correctamente por Weka.

8.9.2 Análisis Periodo 2015-01

Al igual que el proceso realizado en el periodo 2014-02, se seleccionaron solo los registros del periodo 2015-01 y se obtuvo la siguiente tabla:

Tabla 26. Total registros caracterizados del periodo 2015-01

Clasificación	Cantidad
Estudiantes Desertores	72
Estudiantes No Desertores	872
Total Estudiantes	**944**

Fuente: Elaboración propia

De acuerdo a la tabla anterior, fueron creados los archivos **DESERTORES_201501.xlsx** y **NO_DESERTORES_201501.xlsx**. Teniendo en cuenta que solo se tienen 72 registros de estudiantes desertores, es necesario seleccionar 72 de los 872 registros de estudiantes No Desertores. Este proceso se realiza a través de selección aleatoria. El total de 144 registros fueron divididos, 72 registros para datos de entrenamiento y otros 72 registros como datos de prueba. Cada uno de estos 72 registros contaba con 36 datos de estudiantes no desertores y 36 datos de estudiantes desertores. La siguiente tabla muestra la relación de registros ajustados.

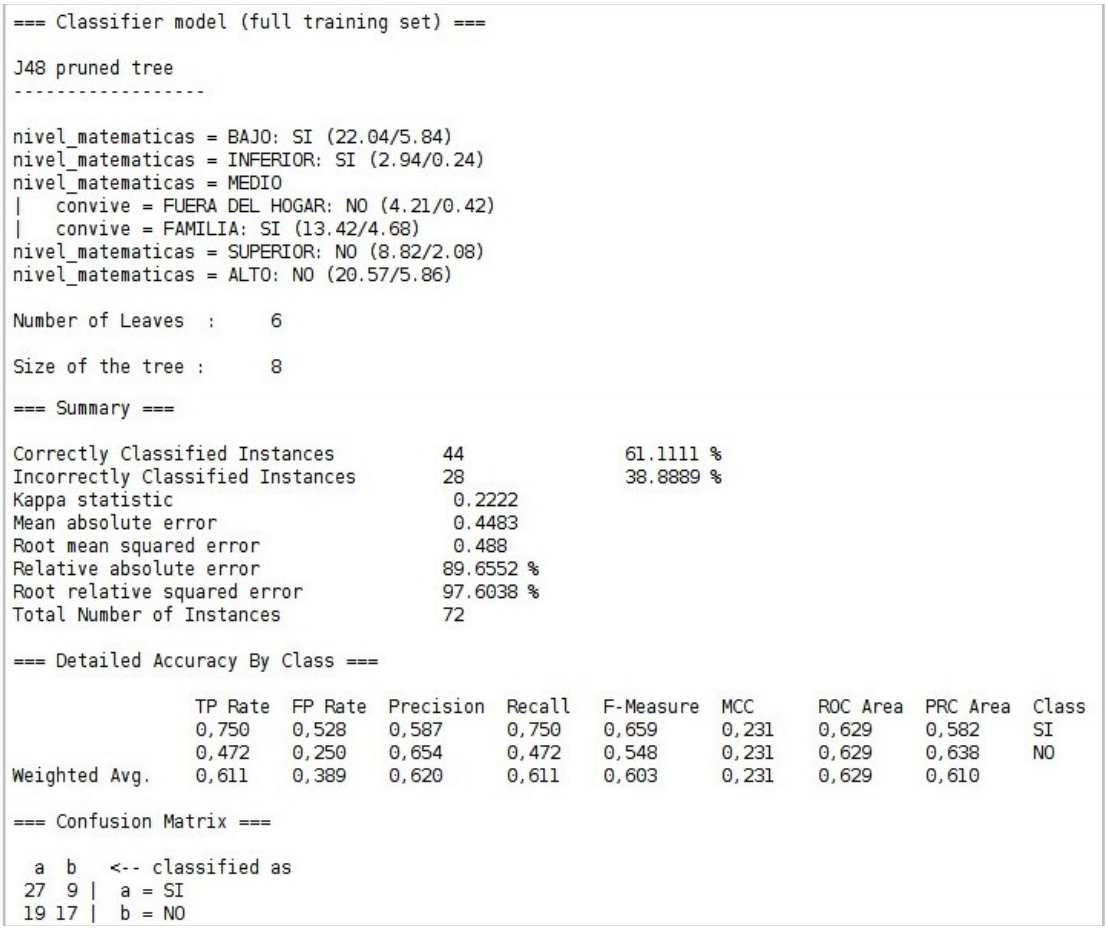
Tabla 27. Total registros del periodo 2015-01 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	36	36
No Desertores	36	36
Total Registros	**72**	**72**

Fuente: Elaboración propia

Procesando la cantidad de registros de la tabla, se obtienen los siguientes resultados:

Figura 29. Resultados Weka con el algoritmo J48 en el periodo 2015-01



```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

nivel_matematicas = BAJO: SI (22.04/5.84)
nivel_matematicas = INFERIOR: SI (2.94/0.24)
nivel_matematicas = MEDIO
|   convive = FUERA DEL HOGAR: NO (4.21/0.42)
|   convive = FAMILIA: SI (13.42/4.68)
nivel_matematicas = SUPERIOR: NO (8.82/2.08)
nivel_matematicas = ALTO: NO (20.57/5.86)

Number of Leaves  :     6

Size of the tree :      8

=== Summary ===

Correctly Classified Instances          44               61.1111 %
Incorrectly Classified Instances        28               38.8889 %
Kappa statistic                          0.2222
Mean absolute error                      0.4483
Root mean squared error                  0.488
Relative absolute error                 89.6552 %
Root relative squared error             97.6038 %
Total Number of Instances               72

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,750    0,528    0,587      0,750    0,659      0,231    0,629     0,582     SI
                 0,472    0,250    0,654      0,472    0,548      0,231    0,629     0,638     NO
Weighted Avg.    0,611    0,389    0,620      0,611    0,603      0,231    0,629     0,610

=== Confusion Matrix ===

  a  b   <-- classified as
 27  9 |  a = SI
 19 17 |  b = NO
```

Fuente: Elaboración propia

Del análisis textual se puede observar que el **Nivel de Matemáticas** y el **Núcleo Familiar** figuran como variables involucradas en la determinación de deserción de un estudiante. En general, se generó un análisis de 6 niveles con un árbol de tamaño 8. De los 72 datos analizados, 44 registros (61.11%) fueron analizados correctamente, mientras que 28 (38.88%) fueron clasificados incorrectamente. El análisis de la matriz de confusión se presenta en la siguiente tabla.

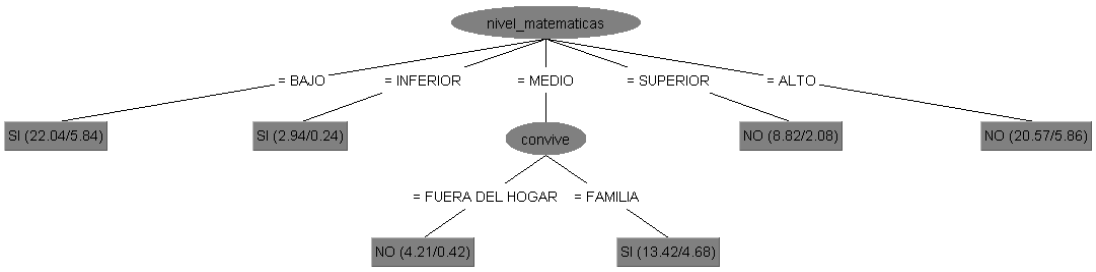
Tabla 28. Interpretación de la matriz de confusión de los resultados del periodo 2015-01

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	36 (50%)	17	19
Desertores	36 (50%)	27	9
Total	72 (100%)	44	28

Fuente: Elaboración propia

Se puede observar que la clasificación de los desertores en este caso, fue mucho mejor que la clasificación de los no desertores. Este tipo de situaciones se presenta por que los registros de los estudiantes No Desertores pueden ser inconsistentes dentro de la lógica del algoritmo, es decir, existen registros que en sus variables presentan valores que, de acuerdo a los datos de entrenamiento, indican que son estudiantes que desertaron, sin embargo, los datos de prueba registran lo contrario. Esta situación sucedió con 19 del total de 36 estudiantes No Desertores. La siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 30. Árbol de decisión Weka del análisis del periodo 2015-01 con el algoritmo J48



Fuente: Elaboración propia

Puntajes bajos e inferiores en el nivel de matemáticas, son determinantes en la deserción; para puntajes medios, es necesario verificar la variable ¿Con quién vive? Aquellas personas que viven con su familia, tienen un mayor riesgo de deserción en comparación con los que viven fuera del hogar. Los estudiantes con puntajes altos y superiores, presentan un bajo riesgo de deserción.

8.9.3 Análisis Periodo 2015-02

Del total de registros de estudiantes caracterizados, se seleccionaron solo los registros del periodo 2015-02 y se obtuvo la siguiente tabla:

Tabla 29. Total registros caracterizados del periodo 2015-02

Clasificación	Cantidad
Estudiantes Desertores	40
Estudiantes No Desertores	316
Total Estudiantes	**356**

Fuente: Elaboración propia

De acuerdo a la tabla, fueron creados los archivos **DESERTORES_201502.xlsx** y **NO_DESERTORES_201502.xlsx**. Teniendo en cuenta que solo se tienen 40 registros de estudiantes desertores, fue necesario seleccionar 40 de los 316 registros de estudiantes No Desertores. Este proceso se realizó a través de selección aleatoria. El total de 80 registros fueron divididos, 40 registros para datos de entrenamiento y otros 40 registros como datos de prueba. Cada uno de estos 40 registros contaba con 20 datos de estudiantes no desertores y 20 datos de estudiantes desertores. La siguiente tabla muestra la relación de registros ajustados.

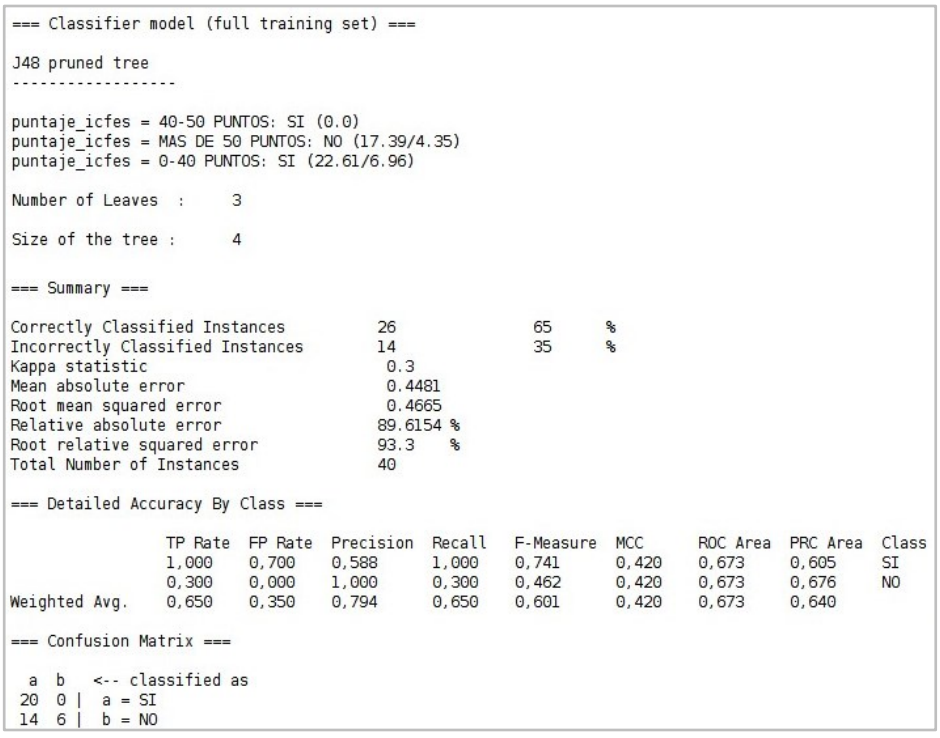
Tabla 30. Total de estudiantes del periodo 2015-02 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	20	20
No Desertores	20	20
Total Registros	**40**	**40**

Fuente: Elaboración propia

Procesando la cantidad de registros de la tabla, se obtuvieron los siguientes resultados:

Figura 31. Resultados Weka con el algoritmo J48 en el periodo 2015-02



```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

puntaje_icfes = 40-50 PUNTOS: SI (0.0)
puntaje_icfes = MAS DE 50 PUNTOS: NO (17.39/4.35)
puntaje_icfes = 0-40 PUNTOS: SI (22.61/6.96)

Number of Leaves  :     3

Size of the tree :      4


=== Summary ===

Correctly Classified Instances          26               65      %
Incorrectly Classified Instances        14               35      %
Kappa statistic                          0.3
Mean absolute error                      0.4481
Root mean squared error                  0.4665
Relative absolute error                 89.6154 %
Root relative squared error             93.3    %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,700    0,588      1,000    0,741      0,420    0,673     0,605     SI
                 0,300    0,000    1,000      0,300    0,462      0,420    0,673     0,676     NO
Weighted Avg.    0,650    0,350    0,794      0,650    0,601      0,420    0,673     0,640

=== Confusion Matrix ===

  a  b   <-- classified as
 20  0 |  a = SI
 14  6 |  b = NO
```

Fuente: Elaboración propia

Del análisis textual se puede observar que solo el **Puntaje ICFES** figura como variable involucrada en la determinación de deserción de un estudiante. En general, se generó un análisis de 3 niveles con un árbol de tamaño 4. De los 40 datos analizados, 26 registros (65%) fueron clasificados correctamente, mientras que 14 (35%) se clasificaron incorrectamente. El análisis de la matriz de confusión se presenta en la siguiente tabla:

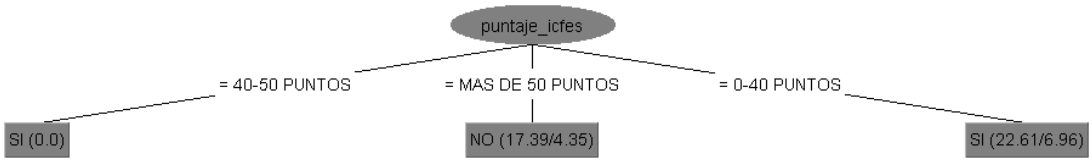
Tabla 31. Interpretación de la matriz de confusión de los resultados del periodo 2015-02

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	20 (50%)	6	14
Desertores	20 (50%)	20	0
Total	40 (100%)	26	14

Fuente: Elaboración propia

Se puede observar que la clasificación de los desertores fue correcta, mientras que la de los No Desertores fue muy pobre. Solo 6 de los 20 registros fueron clasificados correctamente. La siguiente imagen muestra el árbol de decisión construido:

Figura 32. Árbol de decisión Weka del análisis del periodo 2015-02 con el algoritmo J48



Fuente: Elaboración propia

Aquellos estudiantes que tienen un puntaje Icfes entre los 0 y 40 puntos o entre los 40 y 50 puntos, tiene una probabilidad considerable de deserción, cuando el puntaje supera los 50 punto, el riesgo disminuye.

8.9.4 Análisis Periodo 2016-01

Del total de registros de estudiantes caracterizados, se seleccionaron solo los registros del periodo 2016-01 y se obtuvo la siguiente tabla:

Tabla 32. Total registros caracterizados del periodo 2016-01

Clasificación	Cantidad
Estudiantes Desertores	105
Estudiantes No Desertores	894
Total Estudiantes	**999**

Fuente: Elaboración propia

De acuerdo a la tabla anterior, fueron creados los archivos **DESERTORES_201601.xlsx** y **NO_DESERTORES_201601.xlsx**. Teniendo en cuenta que solo se tienen 105 registros de estudiantes desertores, es necesario seleccionar 105 de los 999 registros de estudiantes No Desertores. Este proceso se realiza a través de selección aleatoria. El total de 210 registros fueron divididos, 105 registros para datos de entrenamiento y otros 105 registros como datos de prueba. Cada uno de estos 105 registros contaba con 52 datos de estudiantes no desertores y 53 datos de estudiantes desertores.

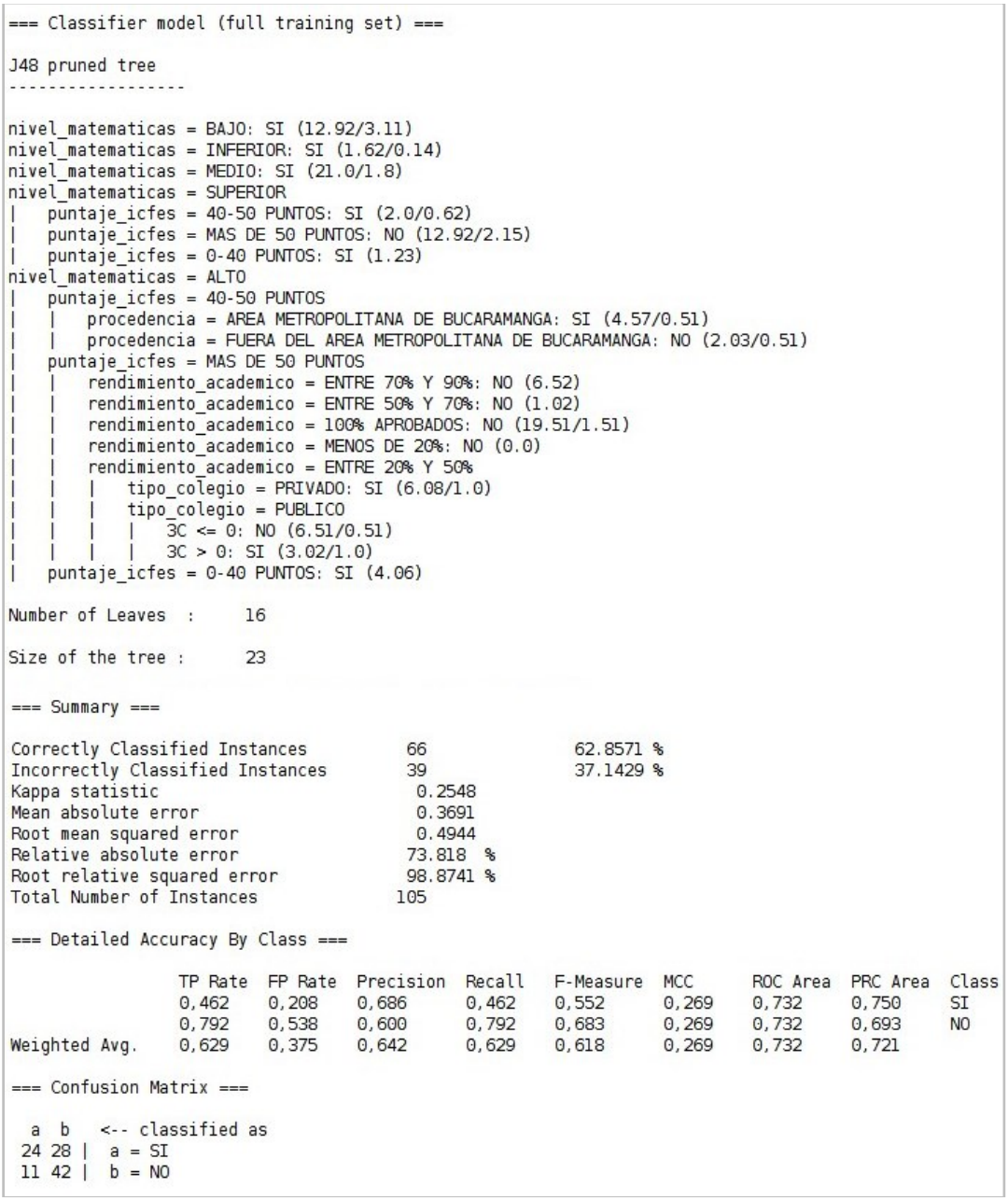
Tabla 33. Total de estudiantes del periodo 2016-01 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	53	52
No Desertores	52	53
Total Registros	**105**	**105**

Fuente: Elaboración propia

Procesando la cantidad de registros de la tabla anterior, se obtienen los siguientes resultados:

Imagen 33. Resultados Weka con el algoritmo J48 en el periodo 2016-01



```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

nivel_matematicas = BAJO: SI (12.92/3.11)
nivel_matematicas = INFERIOR: SI (1.62/0.14)
nivel_matematicas = MEDIO: SI (21.0/1.8)
nivel_matematicas = SUPERIOR
|   puntaje_icfes = 40-50 PUNTOS: SI (2.0/0.62)
|   puntaje_icfes = MAS DE 50 PUNTOS: NO (12.92/2.15)
|   puntaje_icfes = 0-40 PUNTOS: SI (1.23)
nivel_matematicas = ALTO
|   puntaje_icfes = 40-50 PUNTOS
|   |   procedencia = AREA METROPOLITANA DE BUCARAMANGA: SI (4.57/0.51)
|   |   procedencia = FUERA DEL AREA METROPOLITANA DE BUCARAMANGA: NO (2.03/0.51)
|   puntaje_icfes = MAS DE 50 PUNTOS
|   |   rendimiento_academico = ENTRE 70% Y 90%: NO (6.52)
|   |   rendimiento_academico = ENTRE 50% Y 70%: NO (1.02)
|   |   rendimiento_academico = 100% APROBADOS: NO (19.51/1.51)
|   |   rendimiento_academico = MENOS DE 20%: NO (0.0)
|   |   rendimiento_academico = ENTRE 20% Y 50%
|   |   |   tipo_colegio = PRIVADO: SI (6.08/1.0)
|   |   |   tipo_colegio = PUBLICO
|   |   |   |   3C <= 0: NO (6.51/0.51)
|   |   |   |   3C > 0: SI (3.02/1.0)
|   puntaje_icfes = 0-40 PUNTOS: SI (4.06)

Number of Leaves  :     16

Size of the tree :      23

=== Summary ===

Correctly Classified Instances          66               62.8571 %
Incorrectly Classified Instances        39               37.1429 %
Kappa statistic                          0.2548
Mean absolute error                      0.3691
Root mean squared error                  0.4944
Relative absolute error                 73.818  %
Root relative squared error             98.8741 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,462    0,208    0,686      0,462    0,552      0,269    0,732     0,750     SI
                 0,792    0,538    0,600      0,792    0,683      0,269    0,732     0,693     NO
Weighted Avg.    0,629    0,375    0,642      0,629    0,618      0,269    0,732     0,721

=== Confusion Matrix ===

  a  b   <-- classified as
 24 28 |  a = SI
 11 42 |  b = NO
```

Fuente: Elaboración propia

Del análisis textual se puede observar que el **Nivel de Matemáticas**, **Puntaje ICFES**, **Rendimiento Académico**, **Procedencia**, **Tipo de Colegio** y la variable psicológica **3C**, figuran como variables involucradas en la determinación de deserción de un estudiante. Hasta ahora, el resultado más nutrido en variables generado por Weka. En general, se generó un análisis de 16 niveles con un árbol de tamaño 23. De los 105 datos analizados, 66 registros (62.86%) fueron

clasificados correctamente, mientras que 39 (37.14%) se clasificaron incorrectamente. El análisis de la matriz de confusión se presenta a continuación:

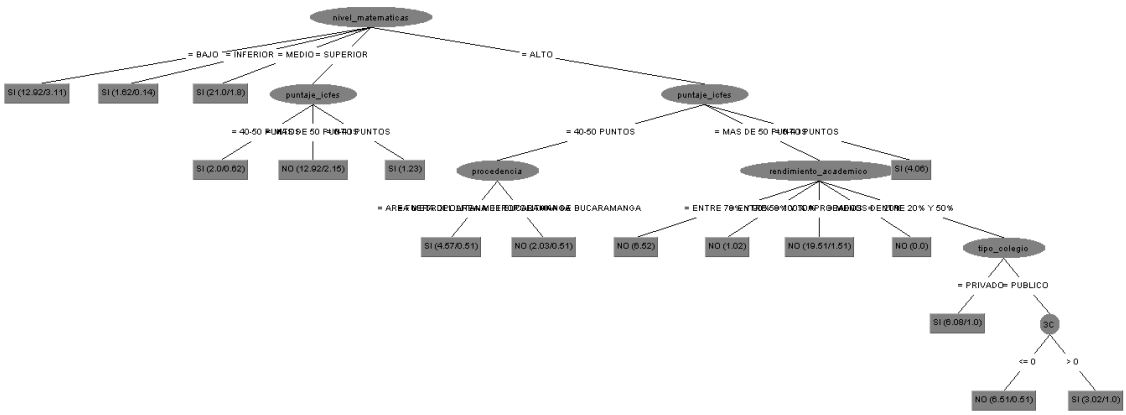
Tabla 34. Interpretación de la matriz de confusión de los resultados del periodo 2016-01

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	53 (51%)	42	11
Desertores	52 (49%)	24	28
Total	105 (100%)	66	39

Fuente: Elaboración propia

A diferencia de los resultados obtenidos en el análisis anterior, en 2016-01 fue más precisa la clasificación de los estudiantes No Desertores. En los estudiantes Desertores fue mayor la tasa de error (28) que la tasa de aciertos (24). La siguiente imagen muestra el árbol de decisión construido:

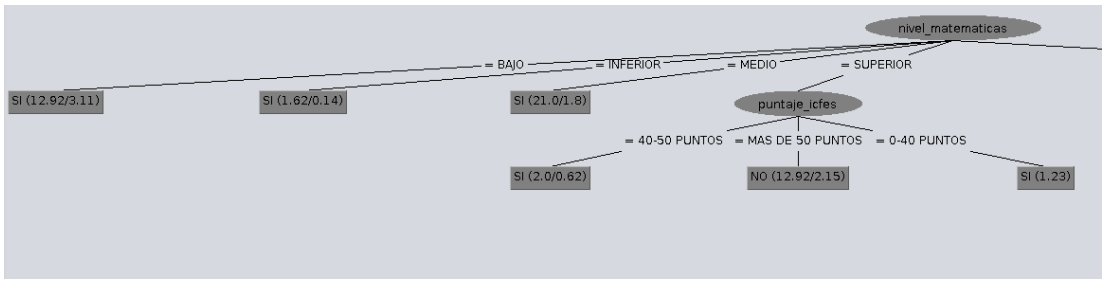
Figura 34. Árbol de decisión Weka del análisis del periodo 2016-01 con el algoritmo J48



Fuente: Elaboración propia

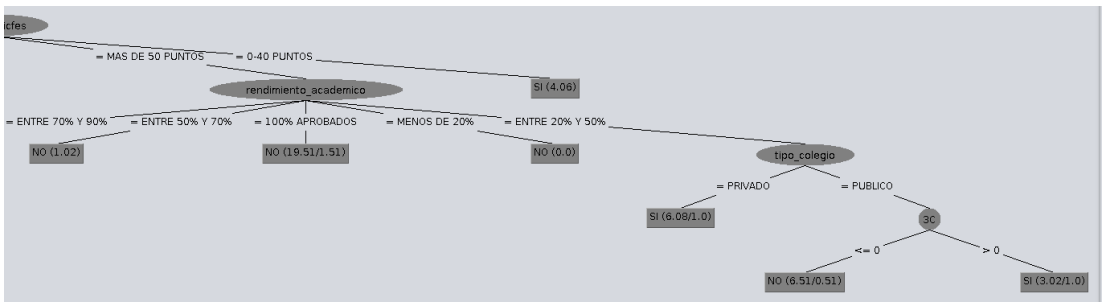
Teniendo en cuenta que el árbol obtenido no es visible claramente debido a su gran tamaño y a las limitaciones de Weka para exportar los resultados, las siguientes imágenes hacen un acercamiento de la zona izquierda y zona derecha del árbol respectivamente.

Figura 35. Árbol de decisión Weka del análisis del periodo 2016-01 con el algoritmo J48 (Izquierda)



Fuente: Elaboración propia

Figura 36. Árbol de decisión Weka del análisis del periodo 2016-01 con el algoritmo J48 (Derecha)



Fuente: Elaboración propia

El nivel de matemáticas es identificado como la principal variable de deserción, Incluso en sus valores superiores se presentaron deserciones a través de la variable puntaje Icfes. El tipo de colegio, dependiendo de un rendimiento académico de menos del 20% es de alto riesgo si este es privado. En caso de ser público, se analiza el valor de la variable psicológica 3C, donde el valor 1 determina un riesgo de deserción.

8.10 ANÁLISIS POR PERIODO ACADÉMICO ACUMULADO

Para continuar el proceso, se realizó el análisis acumulativo por semestre. A diferencia del proceso por periodo académico, en este caso, solo fueron realizados tres análisis, ya que para el periodo 2014-02 no se tienen datos anteriores. Para el primer análisis, los datos del periodo 2014-02 aplican como datos de entrenamiento y los datos del periodo 2015-01 como datos de prueba. Para el segundo análisis, los datos de los periodos 2014-02 y 2015-01 se utilizan como datos de entrenamiento, el periodo 2015-02 como datos de prueba. Finalmente, el tercer análisis toma los datos de los periodos 2014-02, 2015-01 y 2015-02 como datos de entrenamiento y el periodo 2016-01 para las pruebas.

Para intentar aumentar las predicciones previamente obtenidas en los análisis anteriores, se aumentan las muestras procesadas, teniendo en cuenta que se cuenta con mayor número de registros de estudiantes desertores. En el análisis realizado individualmente para el periodo 2014-01, fueron utilizados 60 datos en total, 30 para entrenamiento y 30 para pruebas, cada uno de estos bloques estaba conformado por 15 registros de estudiantes desertores y 15 registros de estudiantes no desertores, así, por ejemplo, el siguiente análisis del periodo 2015-01 fue realizado utilizando el total de estos 60 registros y no solo los 30 del análisis anterior. A continuación, se describen los diferentes análisis realizados.

8.10.1 Periodos 2014-02 y 2015-01

El primer análisis incluye el periodo 2014-02 para los datos de entrenamiento y 2015-01 para los datos de prueba. La siguiente tabla presenta el resumen de los registros analizados:

Tabla 35. Total registros de los periodos 2014-02 y 2015-01 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	72	30
No Desertores	72	30
Total Registros	**144**	**60**

Fuente: Elaboración propia

De acuerdo a la tabla anterior, fueron generados los archivos **TRAINING_201402.xlsx** y **TEST_201501.xslx**, los cuales, fueron procesados a

través del algoritmo J48. La siguiente imagen muestra el resultado del procesamiento realizado:

Figura 37. Resultados Weka con el algoritmo J48 en los periodos 2014-02 y 2015-01

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

puntaje_icfes = 40-50 PUNTOS: SI (16.0)
puntaje_icfes = MAS DE 50 PUNTOS
|   rendimiento_academico = ENTRE 70% Y 90%: NO (9.0/1.0)
|   rendimiento_academico = ENTRE 50% Y 70%: NO (0.0)
|   rendimiento_academico = 100% APROBADOS: NO (26.0/2.0)
|   rendimiento_academico = MENOS DE 20%: SI (3.0)
|   rendimiento_academico = ENTRE 20% Y 50%: NO (0.0)
puntaje_icfes = 0-40 PUNTOS: SI (9.0)

Number of Leaves  :     7

Size of the tree :      9

=== Summary ===

Correctly Classified Instances          72               50      %
Incorrectly Classified Instances        72               50      %
Kappa statistic                          0
Mean absolute error                      0.4679
Root mean squared error                  0.5572
Relative absolute error                 93.573  %
Root relative squared error            111.4362 %
Total Number of Instances              144

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,306    0,306    0,500      0,306    0,379      0,000    0,664     0,607     SI
                 0,694    0,694    0,500      0,694    0,581      0,000    0,663     0,704     NO
Weighted Avg.    0,500    0,500    0,500      0,500    0,480      0,000    0,664     0,656

=== Confusion Matrix ===

  a  b   <-- classified as
 22 50 |  a = SI
 22 50 |  b = NO
```

Fuente: Elaboración Propia

De los resultados anteriores se puede analizar que el **Puntaje ICFES** y el **Rendimiento Académico** figuran como variables involucradas en la determinación de la deserción de un estudiante. Se generó un análisis de 7 niveles con un árbol de tamaño 9. De los 144 datos de prueba, 72 registros (50%) fueron analizados correctamente y los restantes 72 (50%) fueron clasificados incorrectamente. A continuación, se analiza la matriz de confusión.

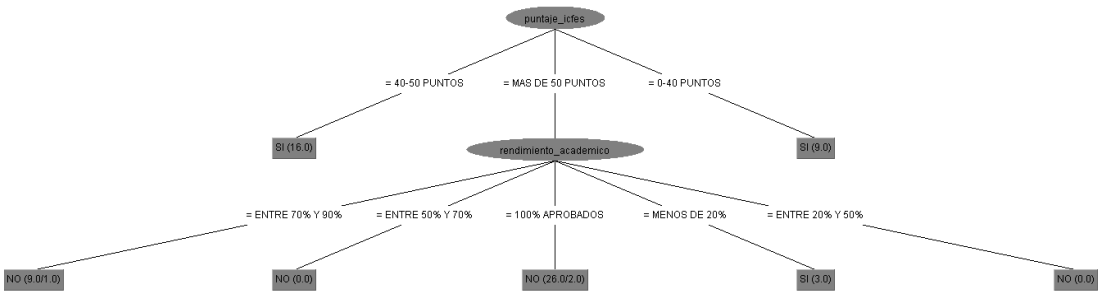
Tabla 36. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02 y 2015-01

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	72 (50%)	50	22
Desertores	72 (50%)	22	50
Total	144 (100%)	72	72

Fuente: Elaboración propia

Se observa en este caso que los datos de los estudiantes No Desertores fueron mejor clasificados que los estudiantes Desertores. La siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 38. Árbol de decisión Weka del análisis de los periodos 2014-02 y 2015-01 con el algoritmo J48



Fuente: Elaboración propia

La variable puntaje Icfes se hace ya habitual en los análisis realizados en conjunto con la variable rendimiento académico. Para puntajes Icfes por debajo de 50 puntos, existe riesgo de deserción, para puntajes mayores se analiza la variable rendimiento académico, si estos, son del menos del 20% existe una probabilidad de deserción considerable.

8.10.2 Periodos 2014-02, 2015-01 y 2015-02

El análisis incluye los periodos 2014-02 (60 registros) y 2015-01 (144 registros) para los datos de entrenamiento y 2015-02 para los datos de prueba. La siguiente tabla presenta el resumen de los registros analizados:

Tabla 37. Total registros de los periodos 2014-02, 2015-01 y 2015-02 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	40	102
No Desertores	40	102
Total Registros	**80**	**204**

Fuente: Elaboración propia

Fueron generados los archivos **TRAINING_201402_201501.xlsx** y **TEST_201502.xslx**, los cuales, fueron procesados a través del algoritmo J48. La siguiente imagen muestra el resultado del procesamiento realizado:

Figura 39. Resultados Weka con el algoritmo J48 en los periodos 2014-02, 2015-01 y 2015-02

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

rendimiento_academico = ENTRE 70% Y 90%
|   nivel_comunicativo = ALTO
|   |   nivel_matematicas = BAJO: SI (6.84/0.45)
|   |   nivel_matematicas = INFERIOR: SI (0.0)
|   |   nivel_matematicas = MEDIO: NO (1.37/0.28)
|   |   nivel_matematicas = SUPERIOR: NO (1.37/0.28)
|   |   nivel_matematicas = ALTO: NO (2.73/0.55)
|   nivel_comunicativo = BAJO: NO (5.96/1.35)
|   nivel_comunicativo = SUPERIOR: NO (18.03/0.84)
|   nivel_comunicativo = MEDIO
|   |   procedencia = AREA METROPOLITANA DE BUCARAMANGA
|   |   |   fuente_pago_estudios = RECURSOS PROPIOS: NO (2.7/0.2)
|   |   |   fuente_pago_estudios = OTRA ENTIDAD: SI (2.06/0.06)
|   |   |   fuente_pago_estudios = ICETEX: SI (2.06/0.06)
|   |   procedencia = FUERA DEL AREA METROPOLITANA DE BUCARAMANGA: NO (2.2)
rendimiento_academico = ENTRE 50% Y 70%: SI (35.01/8.34)
rendimiento_academico = 100% APROBADOS: NO (58.7/8.13)
rendimiento_academico = MENOS DE 20%: SI (27.81/7.27)
rendimiento_academico = ENTRE 20% Y 50%: SI (40.16/6.39)

Number of Leaves  :     15

Size of the tree :      20
=== Summary ===

Correctly Classified Instances          57               71.25   %
Incorrectly Classified Instances        23               28.75   %
Kappa statistic                          0.425
Mean absolute error                      0.364
Root mean squared error                  0.4777
Relative absolute error                 72.7911 %
Root relative squared error             95.543  %
Total Number of Instances               80

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,700    0,275    0,718      0,700    0,709      0,425    0,688     0,659     SI
                0,725    0,300    0,707      0,725    0,716      0,425    0,688     0,659     NO
Weighted Avg.   0,713    0,288    0,713      0,713    0,712      0,425    0,688     0,659

=== Confusion Matrix ===

  a  b   <-- classified as
 28 12 |  a = SI
 11 29 |  b = NO
```

Fuente: Elaboración propia

De los resultados anteriores se puede analizar que el **Rendimiento Académico** el **Nivel Comunicativo**, **Nivel de Matemáticas**, **Procedencia** y **Fuente de Pago de Estudios** figuran como variables involucradas en la determinación de la deserción de un estudiante. Se generó un análisis de 15 niveles con un árbol de tamaño 20. De los 80 datos de prueba, 57 registros (71.25%) fueron analizados correctamente y los restantes 23 (28.75%) fueron clasificados incorrectamente. En análisis de la matriz de confusión se presenta en la siguiente tabla.

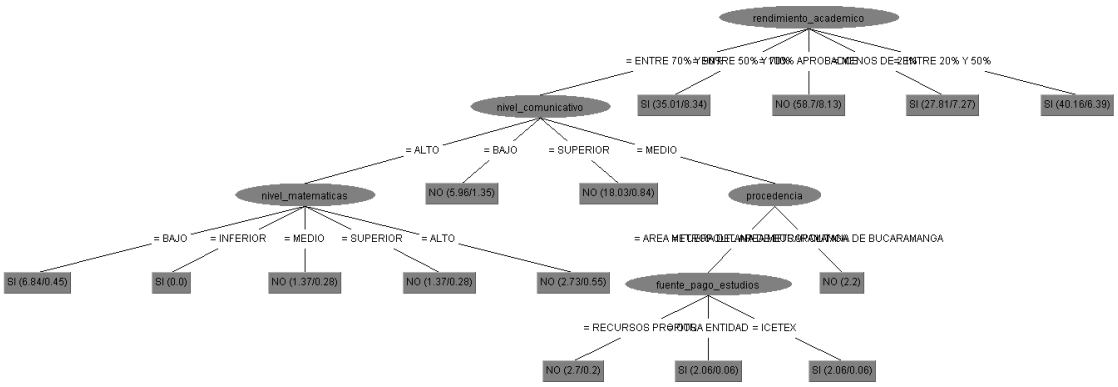
Tabla 38. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02, 2015-01 y 2015-02

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	40 (50%)	29	11
Desertores	40 (50%)	28	12
Total	80 (100%)	57	23

Fuente: Elaboración propia

Se observa que los resultados de los registros clasificados correctamente, tanto para estudiantes desertores (28) y no desertores (29) fue bastante equilibrada. A continuación, la siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 40. Árbol de decisión Weka del análisis de los periodos 2014-02, 2015-01 y 2015-02 con el algoritmo J48



Fuente: Elaboración propia

La variable rendimiento académico por sí sola, representa el mayor número de estudiantes desertores cuando se obtienen porcentajes entre el 20% y 50%, menos del 20% o entre 50% y 70%. Cuando esta variable tiene el valor de 100% aprobados, el estudiante no deserta. Para valores del entre el 70% y 90% entra

en juego la variable Nivel Comunicativo a través de las variables nivel de matemáticas y procedencia, esta última, en relación a la variable fuente de pago de estudios, donde la financiación determina la deserción.

8.10.3 Periodos 2014-02, 2015-01, 2015-02 y 2016-01

El análisis incluye los periodos 2014-02 (60 registros), 2015-01 (144 registros) y 2015-02 (80 registros) para los datos de entrenamiento y 2016-01 para los datos de prueba. La siguiente tabla presenta el resumen de los registros analizados:

Tabla 39. Total registros de los periodos 2014-02, 2015-01, 2015-02 y 2016-01 seleccionados para el análisis

Tipo de Registros	Datos de Prueba	Datos de Entrenamiento
Desertores	105	142
No Desertores	105	142
Total Registros	**210**	**284**

Fuente: Elaboración propia

De acuerdo a la tabla anterior, fueron generados los archivos **TRAINING_201402_201501_201502.xlsx** y **TEST_201601.xslx**, los cuales, fueron procesados a través del algoritmo J48. La figura 41 muestra el resultado del procesamiento realizado:

De los resultados, se puede analizar que el **Rendimiento Académico, Personas a Cargo** y **Trabaja Actualmente** figuran como variables involucradas en la determinación de la deserción de un estudiante. Se generó un análisis de 7 niveles con un árbol de tamaño 10. De los 210 datos de prueba, 145 registros (69.04%) fueron analizados correctamente y los restantes 65 (30.95%) fueron clasificados incorrectamente. El análisis de la matriz de confusión se presenta en la tabla 40.

Figura 41. Resultados Weka con el algoritmo J48 en los periodos 2014-02, 2015-01, 2015-02 y 2016-01

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

rendimiento_academico = ENTRE 70% Y 90%
|   personas_cargo = SI: SI (2.43)
|   personas_cargo = NO
|   |   trabaja = NO: NO (52.17/13.43)
|   |   trabaja = SI: SI (6.68/2.68)
rendimiento_academico = ENTRE 50% Y 70%: SI (40.85/12.28)
rendimiento_academico = 100% APROBADOS: NO (84.77/14.18)
rendimiento_academico = MENOS DE 20%: SI (36.77/7.26)
rendimiento_academico = ENTRE 20% Y 50%: SI (63.32/12.44)

Number of Leaves  :     7

Size of the tree :      10

=== Summary ===

Correctly Classified Instances         145               69.0476 %
Incorrectly Classified Instances        65               30.9524 %
Kappa statistic                          0.381
Mean absolute error                      0.3925
Root mean squared error                  0.4746
Relative absolute error                 78.4906 %
Root relative squared error             94.9195 %
Total Number of Instances              210

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,724    0,343    0,679      0,724    0,700      0,382    0,664     0,604     SI
                 0,657    0,276    0,704      0,657    0,680      0,382    0,664     0,643     NO
Weighted Avg.    0,690    0,310    0,691      0,690    0,690      0,382    0,664     0,624

=== Confusion Matrix ===

  a  b   <-- classified as
 76 29 |  a = SI
 36 69 |  b = NO
```

Fuente: Elaboración propia

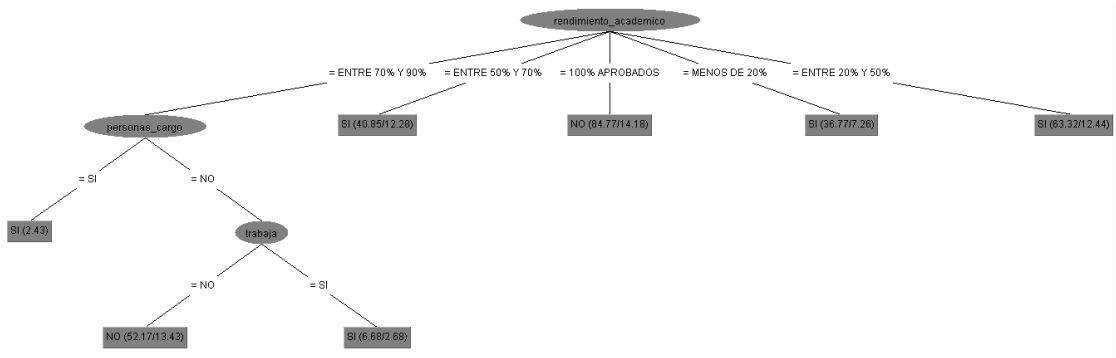
Tabla 40. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02, 2015-01, 2015-02 y 2016-01

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	105 (50%)	69	36
Desertores	105 (50%)	76	29
Total	210 (100%)	57	23

Fuente: Elaboración propia

Se observa que los resultados de los registros clasificados correctamente, al igual que en el análisis anterior, tanto para estudiantes desertores (76) y no desertores (69) fue bastante equilibrado. A continuación, la siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 42. Árbol de decisión Weka del análisis de los periodos 2014-02, 2015-01, 2015-02 y 2016-01con el algoritmo J48



Fuente: Elaboración propia

Como variable principal, se encuentra el rendimiento académico, que para porcentajes por debajo del 70% determina la deserción, en puntajes entre el 70% y 90% evalúa la variable ¿Personas a cargo? Si el estudiante tiene personas a cargo, deserta, sino, se evalúa la variable ¿Trabaja Actualmente? Si el estudiante traba, entonces deserta.

8.11 ANÁLISIS POR PERIODO ACADÉMICO ACUMULADO SIN LA VARIABLE RENDIMIENTO ACADÉMICO

Dentro de los análisis realizados anteriormente, la variable Rendimiento Académico destaca en la mayoría de los resultados como determinante de la deserción de un estudiante. Teniendo en cuenta que, solo hasta el final del semestre es posible por parte de la UNAB, conocer el porcentaje de rendimiento académico, la labor de identificación temprana de los estudiantes con riesgo de deserción se ve afectada. Atendiendo esta situación, se realizó el análisis acumulativo utilizando la misma cantidad de registros del análisis de la sección anterior, esta vez, descartando la variable rendimiento académico.

8.11.1 Periodos 2014-02 y 2015-01

De acuerdo al árbol de la figura 43 se puede observar que solo el **Puntaje ICFES** figura como variable involucrada en la determinación de la deserción de un estudiante, a diferencia del análisis equivalente realizado en la sección anterior, que incluyo la variable Rendimiento Académico, el porcentaje de predicción obtenido se redujo, pasando de un 50% a un 48.61%. Se generó un análisis de 3 niveles con un árbol de tamaño 4. De los 144 datos de prueba, 70 registros (48%) fueron analizados correctamente y los restantes 74 (51.38%) fueron clasificados incorrectamente. En análisis de la matriz de confusión se presenta en la tabla 41.

Figura 43. Resultados Weka con el algoritmo J48 en los periodos 2014-02 y 2015-01 sin la variable Rendimiento Académico.

```
=== Classifier model (full training set) ===

InputMappedClassifier:

J48 pruned tree
------------------

puntaje_icfes = 40-50 PUNTOS: SI (16.0)
puntaje_icfes = MAS DE 50 PUNTOS: NO (38.0/6.0)
puntaje_icfes = 0-40 PUNTOS: SI (9.0)

Number of Leaves  :     3

Size of the tree :      4

=== Summary ===

Correctly Classified Instances          70               48.6111 %
Incorrectly Classified Instances        74               51.3889 %
Kappa statistic                         -0.0278
Mean absolute error                      0.4838
Root mean squared error                  0.5436
Relative absolute error                 96.7697 %
Root relative squared error            108.7083 %
Total Number of Instances              144

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,181    0,208    0,464      0,181    0,260      -0,035   0,544     0,524     SI
                 0,792    0,819    0,491      0,792    0,606      -0,035   0,544     0,551     NO
Weighted Avg.    0,486    0,514    0,478      0,486    0,433      -0,035   0,544     0,538

=== Confusion Matrix ===

  a  b   <-- classified as
 13 59 |  a = SI
 15 57 |  b = NO
```

Fuente: Elaboración propia

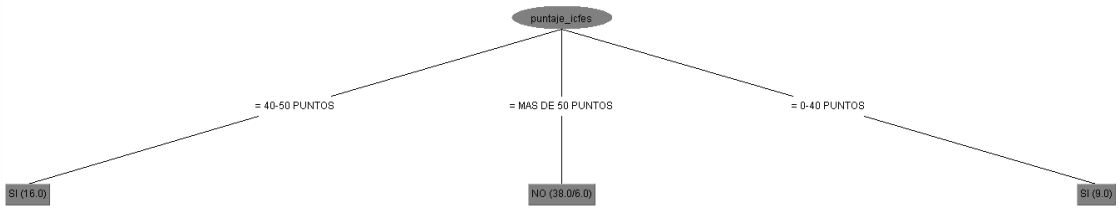
Tabla 41. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02 y 2015-01

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	72 (50%)	57	15
Desertores	72 (50%)	13	59
Total	144 (100%)	70	74

Fuente: Elaboración propia

Se observa en este caso que los datos de los estudiantes No Desertores fueron mejor clasificados que los estudiantes Desertores. A continuación, la siguiente imagen muestra el árbol de decisión construido por Weka, donde se observa que para puntajes Icfes por debajo del 50%, el estudiante desertó.

Figura 44. Árbol de decisión Weka del análisis de los periodos 2014-02 y 2015-01 con el algoritmo J48 sin la variable Rendimiento Académico

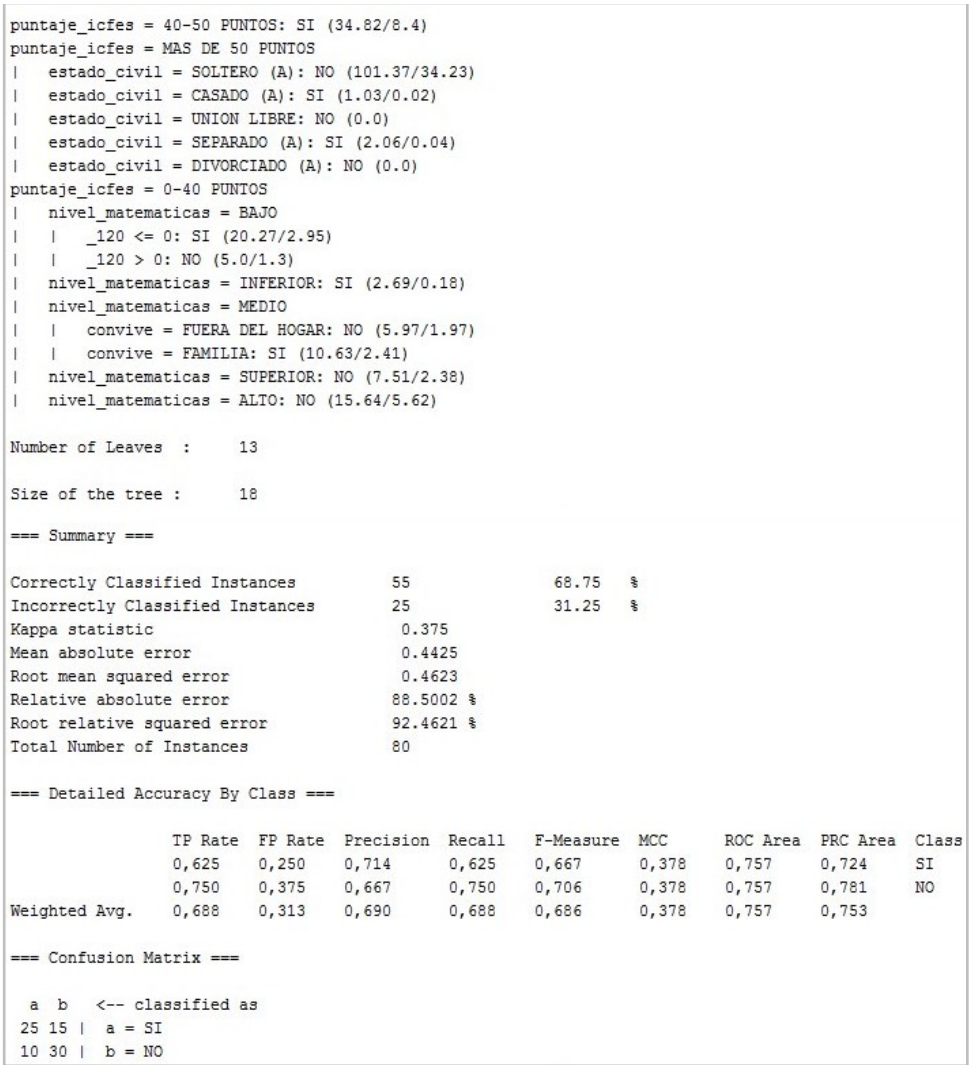


Fuente: Elaboración propia

8.11.2 Periodos 2014-02, 2015-01 y 2015-02

De los resultados de la figura 45 se puede analizar que el **Puntaje ICFES**, **Estado Civil**, **Nivel de Matemáticas** y **¿Con Quien Vive?**, figuran como variables involucradas en la determinación de la deserción de un estudiante. Se generó un análisis de 13 niveles con un árbol de tamaño 18. De los 80 datos de prueba, 55 registros (68.75%) fueron analizados correctamente y los restantes 25 (31.25%) fueron clasificados incorrectamente. El análisis de la matriz de confusión se presenta en la tabla 42.

Figura 45. Resultados Weka con el algoritmo J48 en los periodos 2014-02, 2015-01 y 2015-02 sin la variable Rendimiento Académico



```
puntaje_icfes = 40-50 PUNTOS: SI (34.82/8.4)
puntaje_icfes = MAS DE 50 PUNTOS
|   estado_civil = SOLTERO (A): NO (101.37/34.23)
|   estado_civil = CASADO (A): SI (1.03/0.02)
|   estado_civil = UNION LIBRE: NO (0.0)
|   estado_civil = SEPARADO (A): SI (2.06/0.04)
|   estado_civil = DIVORCIADO (A): NO (0.0)
puntaje_icfes = 0-40 PUNTOS
|   nivel_matematicas = BAJO
|   |   _120 <= 0: SI (20.27/2.95)
|   |   _120 > 0: NO (5.0/1.3)
|   nivel_matematicas = INFERIOR: SI (2.69/0.18)
|   nivel_matematicas = MEDIO
|   |   convive = FUERA DEL HOGAR: NO (5.97/1.97)
|   |   convive = FAMILIA: SI (10.63/2.41)
|   nivel_matematicas = SUPERIOR: NO (7.51/2.38)
|   nivel_matematicas = ALTO: NO (15.64/5.62)

Number of Leaves  :     13

Size of the tree :      18

=== Summary ===

Correctly Classified Instances          55               68.75   %
Incorrectly Classified Instances        25               31.25   %
Kappa statistic                          0.375
Mean absolute error                      0.4425
Root mean squared error                  0.4623
Relative absolute error                 88.5002 %
Root relative squared error             92.4621 %
Total Number of Instances               80

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,625    0,250    0,714      0,625    0,667      0,378    0,757     0,724     SI
                 0,750    0,375    0,667      0,750    0,706      0,378    0,757     0,781     NO
Weighted Avg.    0,688    0,313    0,690      0,688    0,686      0,378    0,757     0,753

=== Confusion Matrix ===

  a  b   <-- classified as
 25 15 |  a = SI
 10 30 |  b = NO
```

Fuente: Elaboración propia

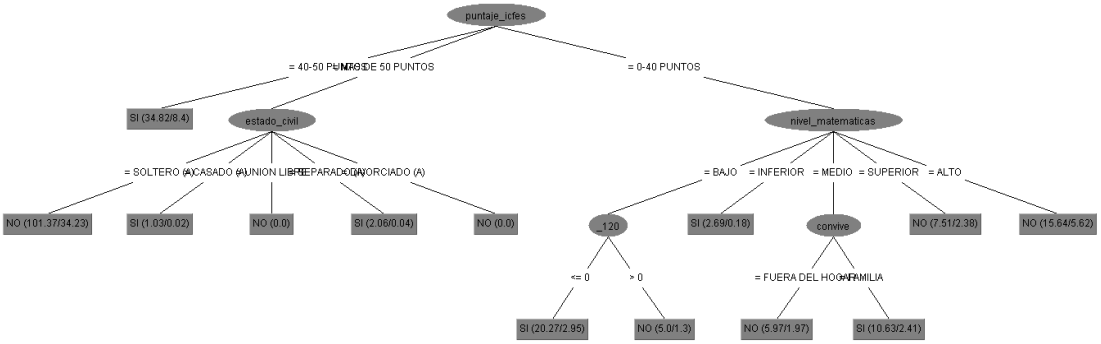
Tabla 42. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02, 2015-01 y 2015-02 sin la variable Rendimiento Académico

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	40 (50%)	30	10
Desertores	40 (50%)	25	15
Total	80 (100%)	55	25

Fuente: Elaboración propia

Se observa que los resultados de los registros clasificados correctamente, tanto para estudiantes desertores (30) y no desertores (25) en comparación con los desertores (29) y no desertores (28) del análisis anterior, resultan bastante similares. A continuación, la siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 46. Árbol de decisión Weka del análisis de los periodos 2014-02, 2015-01 y 2015-02 con el algoritmo J48 sin la variable Rendimiento Académico



Fuente: Elaboración propia

Nuevamente el puntaje Icfes aparece como el nodo raíz, sin embargo, por si solo es determinante de deserción cuando su valor se encuentra entre 40 y 50 puntos. Cuando es mayor a 50 puntos se evalúa el estado civil del estudiante, si este es casado o separado, deserta. Cuando el puntaje Icfes es de 0 a 40 puntos se evalúa el nivel de matemáticas, si este es inferior, el estudiante deserta, si es bajo o medio, se evalúan las variables 12O y ¿Con quién vive? Respectivamente.

8.11.3 Periodos 2014-02, 2015-01, 2015-02 y 2016-01

De los resultados de la figura 47 se puede analizar que el **Puntaje ICFES**, **Personas a Cargo, Nivel de Matemáticas**, **8I**, **16T** y **Con Quien Vive**, figuran como variables involucradas en la determinación de la deserción de un estudiante. Se generó un análisis de 11 niveles con un árbol de tamaño 17. De los 210 datos de prueba, 125 registros (59.52%) fueron analizados correctamente y los restantes 85 (40.47%) fueron clasificados incorrectamente. El análisis de la matriz de confusión se presenta en la tabla 43.

Figura 47. Resultados Weka con el algoritmo J48 en los periodos 2014-02, 2015-01, 2015-02 y 2016-01 sin la variable Rendimiento Académico.

```
J48 pruned tree
------------------

puntaje_icfes = 40-50 PUNTOS: SI (36.84/8.83)
puntaje_icfes = MAS DE 50 PUNTOS
|   personas_cargo = SI: SI (5.36/0.95)
|   personas_cargo = NO: NO (130.38/41.47)
puntaje_icfes = 0-40 PUNTOS
|   nivel_matematicas = BAJO
|   |   8I <= 0: SI (26.41/3.11)
|   |   8I > 0
|   |   |   16T <= 0: SI (3.61/0.53)
|   |   |   16T > 0: NO (4.13/0.52)
|   nivel_matematicas = INFERIOR: SI (4.93/0.24)
|   nivel_matematicas = MEDIO
|   |   convive = FUERA DEL HOGAR: NO (8.85/3.75)
|   |   convive = FAMILIA: SI (25.64/9.56)
|   nivel_matematicas = SUPERIOR: NO (13.02/5.4)
|   nivel_matematicas = ALTO: NO (27.82/12.28)

Number of Leaves  :     11

Size of the tree :      17

=== Summary ===

Correctly Classified Instances         125               59.5238 %
Incorrectly Classified Instances        85               40.4762 %
Kappa statistic                          0.1905
Mean absolute error                      0.4637
Root mean squared error                  0.5049
Relative absolute error                 92.735  %
Root relative squared error            100.9854 %
Total Number of Instances              210

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,343    0,152    0,692      0,343    0,459      0,221    0,589     0,556     SI
                 0,848    0,657    0,563      0,848    0,677      0,221    0,589     0,555     NO
Weighted Avg.    0,595    0,405    0,628      0,595    0,568      0,221    0,589     0,556

=== Confusion Matrix ===

  a  b   <-- classified as
 36 69 |  a = SI
 16 89 |  b = NO
```

Fuente: Elaboración propia

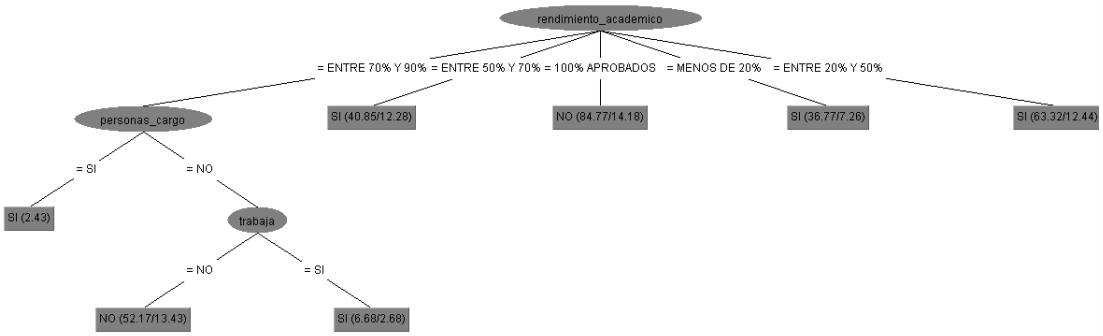
Tabla 43. Interpretación de la matriz de confusión de los resultados de los periodos 2014-02, 2015-01, 2015-02 y 2016-01 sin la variable Rendimiento Académico

Tipo de Registro	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente
No Desertores	105 (50%)	89	16
Desertores	105 (50%)	36	69
Total	210 (100%)	125	85

Fuente: Elaboración propia

Se observa que los resultados de los registros clasificados correctamente, de estudiantes desertores (36) y no desertores (89) fue desequilibrado, clasificando en una mayor proporción los registros de estudiantes No Desertores. A continuación, la siguiente imagen muestra el árbol de decisión construido por Weka:

Figura 48. Árbol de decisión Weka del análisis de los periodos 2014-02, 2015-01, 2015-02 y 2016-01con el algoritmo J48 sin la variable Rendimiento Académico



Fuente: Elaboración propia

El rendimiento académico en este caso, figura como la principal causa de deserción, donde solo los estudiantes que aprobaron el 100% de sus asignaturas no desertaron. En caso de aprobar entre el 70% y 90% se evalúa la variable ¿Tiene personas a cargo? Si la respuesta es afirmativa, el estudiante deserta, en caso contrario se evalúa la variable ¿Trabaja Actualmente? Si trabaja, el estudiante deserta en la mayoría de los casos.

8.12 ANÁLISIS DE RESULTADOS

En este apartado, se analizan comparativamente los resultados obtenidos en los diferentes análisis realizados, teniendo en cuenta los porcentajes de deserción calculados por el algoritmo en cada periodo académico, como los porcentajes de clasificación de registros de estudiantes desertores y no desertores. En general, se realizó el análisis de los datos por periodo académico individual, periodo académico acumulado y periodo académico acumulado sin la variable rendimiento académico. El resultado corresponde a la cantidad de registros clasificados correctamente por el algoritmo, el cual, se resume en la siguiente tabla.

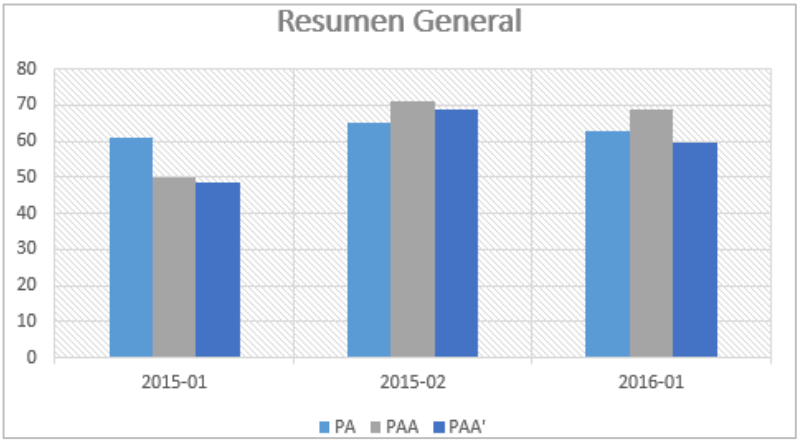
Tabla 44. Resumen general de los resultados obtenidos en los tres análisis realizados

Periodo Académico	Resultados por Periodo	Resultados por Periodo Acumulado	Resultados por Periodo Acumulado sin RA
2014-02	96,66	--	--
2015-01	61,11	50,00	48,61
2015-02	65,00	71,25	68,75
2016-01	62,85	69,04	59,52

Fuente: Elaboración propia

El periodo 2014-02 solo fue evaluado individualmente, teniendo en cuenta la ausencia de registros de periodos anteriores. por lo tanto, no es posible realizar un análisis profundo del resultado obtenido. En la siguiente gráfica, se observan la comparación de los resultados por periodo académico.

Figura 49. Resumen general de los resultados obtenidos por periodo académico en los tres análisis realizados. Fuente: Elaboración propia



Fuente: Elaboración propia

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 los registros clasificados correctamente en el análisis individual (PA) fue superior. En el periodo 2015-02 y 2016-01 los análisis por periodo académico acumulado (PAA), el porcentaje de clasificación fue mayor, superando hasta por 7 puntos porcentuales los resultados del análisis individual por periodo académico (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 cuando 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 análisis acumulados por periodo académico sin la variable rendimiento académico, los porcentajes de clasificación son inferiores en comparación con los resultados del análisis con la variable rendimiento académico, lo que indica la importancia que tiene esta variable para determinar la deserción de un estudiante.

A continuación, se realiza un segundo análisis de resultados que, a diferencia del análisis anterior, incluye de forma separada el porcentaje de clasificación de registros de estudiantes no desertores y desertores. Hay que tener en cuenta que la suma de los porcentajes individuales no debe ser 100%, puesto que, el porcentaje restante de las columnas 3 y 4 corresponde al porcentaje de los registros clasificados incorrectamente para estudiantes no desertores y desertores respectivamente. A continuación, las tablas 45, 46 y 47 presentan los resultados en cada uno de los periodos académicos analizados con su respectiva gráfica.

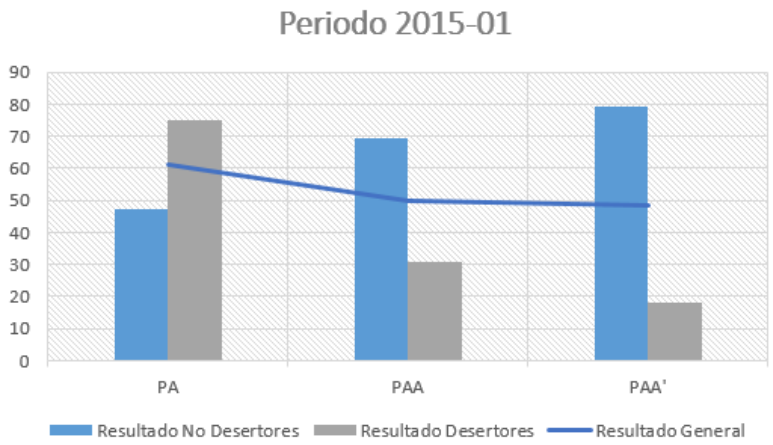
Tabla 45. Resultados de clasificación obtenidos en los 3 análisis realizados, discriminados por desertores y no desertores para el periodo 2015-01

Periodo Académico	Tipo de Análisis	Resultado No Desertores %	Resultado Desertores %	Resultado General %
2015-01	PA	47,22	75	61,11
	PAA	69,44	30,55	50
	PAA'	79,16	18,05	48,61

Fuente: Elaboración propia

El resultado general que se observa en la última columna de la tabla, corresponde al promedio del resultado de estudiantes desertores y no desertores. En la siguiente gráfica, se puede observar la tendencia de los resultados:

Figura 50. Porcentaje de registros de estudiantes desertores y no desertores clasificados correctamente en el periodo 2015-01



Fuente: Elaboración propia

Como se observa, 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.

En relación al resultado general (línea), se observa que el porcentaje de clasificación obtenido en el análisis por periodo académico (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 son marcadamente desiguales en cada uno de los análisis realizados.

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

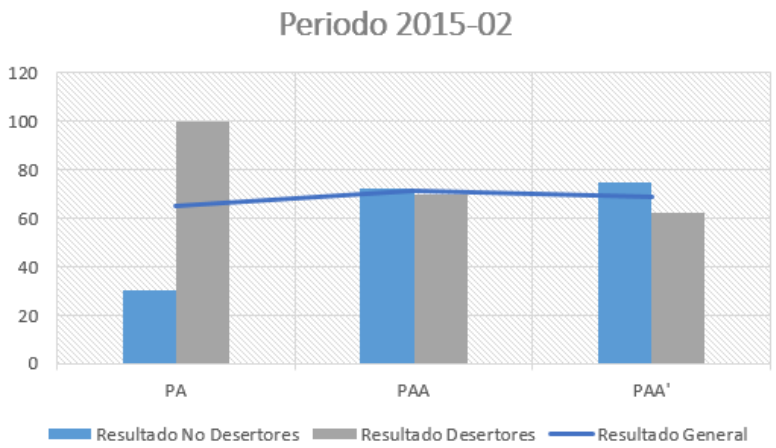
Tabla 46. Resultados de clasificación obtenidos en los 3 análisis realizados, discriminados por desertores y no desertores para el periodo 2015-02

Periodo Académico	Tipo de Análisis	Resultado No Desertores %	Resultado Desertores %	Resultado General %
2015-02	PA	30	100	65
	PAA	72,5	70	71,25
	PAA'	75	18,05	68,75

Fuente: Elaboración propia

Teniendo en cuenta que, para este análisis fueron tenidos en cuenta los registros del periodo 2014-02 y 2015-01 como datos de entrenamiento del algoritmo, se obtienen porcentajes de clasificación superiores a los resultados obtenidos en el análisis del periodo 2015-01. En la siguiente gráfica, se puede observar la tendencia de los resultados:

Figura 51. Porcentaje de registros de estudiantes desertores y no desertores clasificados correctamente en el periodo 2015-02



Fuente: Elaboración propia

Como se observa, al igual que en los resultados del periodo 2015-01, 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.

En relación al resultado general (línea), se observa que el porcentaje de clasificación obtenido en el análisis por periodo académico acumulado (PAA) es superior a los porcentajes obtenidos en los análisis restantes, siendo el resultado coherente a lo esperado, teniendo en cuenta que el registro histórico de PAA es superior al registro histórico del análisis del periodo 2015-01.

Igualmente, se observa que los porcentajes de clasificación de registros de estudiantes desertores en relación al porcentaje de clasificación de registros de estudiantes no desertores es bastante equilibrado en los análisis PAA y PAA' si se compara con los resultados de los análisis del periodo anterior (Figura 50). A continuación, en la siguiente tabla se presenta la clasificación de los resultados obtenidos en el análisis del periodo 2016-01.

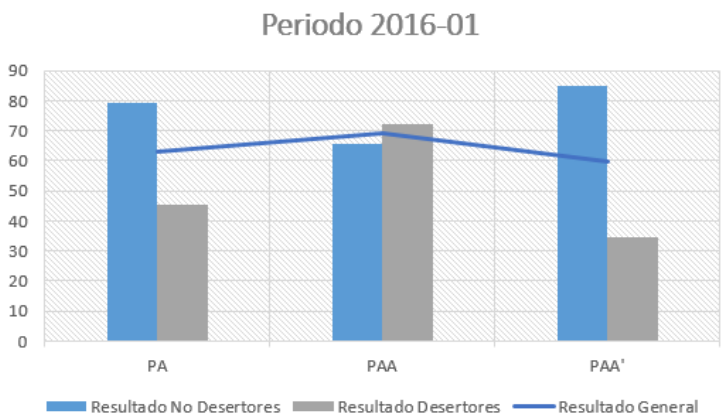
Tabla 47. Resultados de clasificación obtenidos en los 3 análisis realizados, discriminados por desertores y no desertores para el periodo 2016-01

Periodo Académico	Tipo de Análisis	Resultado No Desertores %	Resultado Desertores %	Resultado General %
2016-01	PA	79,24	45,28	62,85
	PAA	65,71	72,38	69,04
	PAA'	84,76	34,28	59,52

Fuente: Elaboración propia

Continuando con la tendencia esperada, 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, sin embargo, el porcentaje de clasificación estuvo dos puntos porcentuales si se compara con el mismo análisis del periodo 2015-02. En la siguiente gráfica, se puede observar la tendencia de los resultados:

Figura 52. Porcentaje de registros de estudiantes desertores y no desertores clasificados correctamente en el periodo 2016-01



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 registros de desertores y no desertores en el análisis PA teniendo en cuenta los resultados de los análisis de los periodos anteriores.

En relación al resultado general (línea), se observa que el porcentaje de clasificación obtenido en el análisis por periodo académico acumulado (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 tres gráficas, en los análisis por periodo acumulado teniendo en cuenta la variable rendimiento académico, 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, sin embargo, en el análisis del periodo 2015-02, donde se cuenta con un mayor número de datos de entrenamiento, los porcentajes de clasificación son más equilibrados, a pesar que la clasificación de los registros de estudiantes no desertores es mayor, la diferencia es solo de dos puntos porcentuales en comparación con los registros de estudiantes desertores. En el periodo 2016-01, donde se cuenta con más registros de entrenamiento, 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 los análisis por periodo acumulado sin tener en cuenta la variable rendimiento académico (PAA'), se observa que el porcentaje de estudiantes desertores clasificados correctamente, desciende drásticamente en comparación con la clasificación realizada en los análisis por periodo acumulado (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.

8.12.1 Análisis de 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, en cada uno de los análisis ejecutados, es necesario identificar cuáles de las variables analizadas determinan en mayor y menor medida la deserción de los estudiantes de primer ingreso a los programas de pregrado en la UNAB. Para ello, se tienen 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

De acuerdo a estos criterios y teniendo en cuenta los resultados Weka en los diferentes análisis los cuales se pueden consultar en las secciones 8.9, 8.10 y 8.11 se construyó la siguiente tabla:

Tabla 48. Relación de variables que determinan la deserción de los estudiantes

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

Fuente: Elaboración propia

Como se observa en tabla, 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, 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 las deserciones, 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, el cual, aparece en 9 del total de 10 análisis individuales realizados.

8.12.2 Comparación del Modelo BI Contra el Modelo Empírico de la UNAB

Teniendo en cuenta los resultados obtenidos en el Análisis por Periodo Académico Acumulado del periodo 2016-01 descrito en la unidad 8.10.3**,** y cumpliendo con el tercer objetivo específico del proyecto, 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 presenta 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 BI; estos mismos registros fueron consultados en el modelo matemático. A continuación, fue creado un archivo Excel con el nombre **Resultado_Final_Predicción_201601.xlsx** con el siguiente formato:

Figura 53. Formato de archivo generado para calcular los porcentajes de precisión del modelo BI y el modelo empírico de la UNAB.

DESERTO_REAL	DESERTO_BI	ALERTA_GENERAL	DESERTO_MM	ACERTO_BI	ACERTO_MM
NO	SI	34	SI	0	0
NO	NO	32	SI	1	0
SI	NO	28	SI	0	1
NO	SI	32	SI	0	0
SI	NO	8	NO	0	0
NO	SI	25	NO	0	1
NO	NO	23	NO	1	1
NO	SI	22	NO	0	1
SI	SI	14	NO	1	0
NO	NO	26	NO	1	1
SI	SI	30	SI	1	1

Fuente: Elaboración propia

Donde:

DESERTO_REAL: Corresponde al estado real de deserción del estudiante
DESERTO_BI: Valor de deserción calculado por el modelo BI
ALERTA_GENERAL: Valor de alerta calculado por el modelo matemático
DESERTO_MM: Valor de deserción calculado por el modelo matemático
ACERTO_BI: Indicador de acierto del modelo BI, donde 1 Acertó, 0 no Acertó
ACERTO_MM: Indicador de acierto del modelo matemático

Al final, la suma de los valores de la columna ACERTO_BI y ACERTO_MM corresponden al total de estudiantes clasificados correctamente por el modelo BI y por el modelo empírico respectivamente. La siguiente tabla resume los resultados obtenidos:

Tabla 49. Resultados finales de predicción del modelo BI y el modelo empírico de la UNAB para el periodo 2016-01

Modelo	Total Registros	Clasificados Correctamente	Clasificados Incorrectamente	Porcentaje de Acierto
BI	210	145	65	69,04
Empírico	210	109	101	51,90

Fuente: Elaboración propia

Como se observa en la tabla, el modelo BI basado en minería de datos supera las predicciones realizadas por el modelo matemático de la UNAB, 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 observan más en detalle los resultados, discriminando el porcentaje de clasificación tanto de desertores como de no desertores, se obtienen los siguientes resultados:

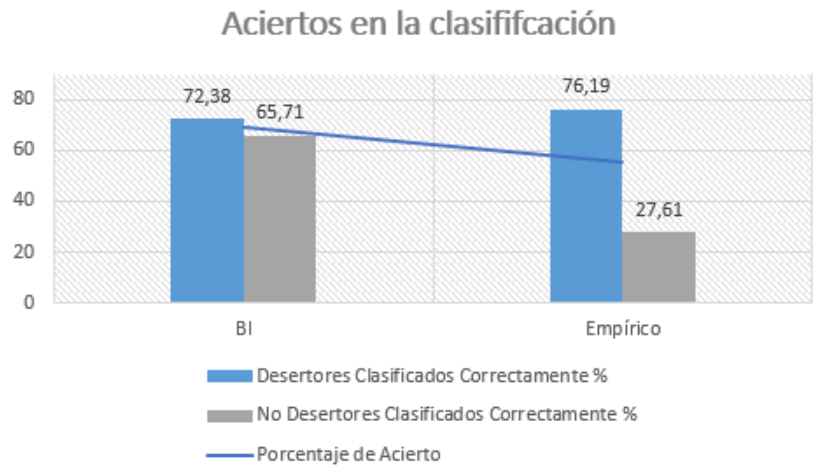
Tabla 50. Resultados finales de aciertos del modelo BI y el modelo empírico de la UNAB para el periodo 2016-01 discriminados por desertores y no desertores

Modelo	Desertores Clasificados Correctamente %	No Desertores Clasificados Correctamente %	Porcentaje de Acierto
BI	72,38	65,71	69,04
Empírico	76,19	27,61	55,71

Fuente: Elaboración propia

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

Figura 54. Gráfica de resultados finales de aciertos del modelo BI y el modelo empírico de la UNAB para el periodo 2016-01 discriminados por desertores y no desertores



Fuente: Elaboración propia

Si se observa el porcentaje de desertores clasificados correctamente del modelo empírico (76,19%), se tiene que este es mayor que el porcentaje del modelo BI (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, sobre la muestra dada, por lo tanto, su porcentaje de error debe ser mucho mayor, esto se comprueba observando los siguientes resultados:

Tabla 51. Resultados finales de errores del modelo BI y el modelo empírico de la UNAB para el periodo 2016-01 discriminados por desertores y no desertores.

Modelo	Desertores Clasificados Incorrectamente %	No Desertores Clasificados Incorrectamente %	Porcentaje de Errores
BI	34,28	27,61	30,95
Empírico	72,38	23,80	48,09

Fuente: Elaboración propia

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. La siguiente gráfica muestra la tendencia observada:

Figura 55. Gráfica de resultados finales de errores del modelo BI y el modelo empírico de la UNAB para el periodo 2016-01 discriminados por desertores y no desertores

Fuente: Elaboración propia

En los registros de estudiantes no desertores el porcentaje de error es ligeramente superior en el modelo BI, 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 BI 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.

9. 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, a diferencia de lo que ocurre cuando solo se procesa la información de un solo periodo académico,

donde los porcentajes de clasificación de estudiantes desertores, son muy bajos, mientras que los resultados de no desertores son bastante altos, por lo tanto, los resultados pierden calidad.

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. En el caso de la variable Rendimiento Académico, solo es posible conocer el porcentaje de asignaturas aprobadas una vez el periodo académico ha finalizado, es importante que se pueda idear un mecanismo que permita obtener este porcentaje de forma anticipada; teniendo en cuenta que el semestre académico de pregrado cuenta con tres cortes, es posible realizar el proceso de predicción haciendo uso de las calificaciones parciales de los estudiantes.

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 matemático, 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.

10. RECOMENDACIONES Y TRABAJOS FUTUROS

Es necesario por parte de la UNAB, mejorar el proceso de recolección de la información académica, psicología y financiera de los estudiantes de pregrado al momento de ingreso a la universidad y los datos que son recolectados en el transcurso del periodo académico. El mejoramiento en la calidad de los datos, influye directamente en los porcentajes de predicción obtenidos por parte de los algoritmos de clasificación.

El porcentaje de inasistencia se presume puede tener un impacto importante en la determinación de deserción de un estudiante, sin embargo, esta variable no fue tenida en cuenta por ausencia de datos. Es necesario que los valores de esta variable sean obtenidos y analizados durante el proceso de clasificación,

El proceso realizado, incluye el análisis de predicción con el algoritmo de Arboles de Decisión J48, sin embargo, existen muchas otras alternativas de clasificación, las cuales, pueden ser identificadas, modeladas y evaluadas para identificar mejores resultados de clasificación.

El modelo de clasificación utilizado en el proyecto, permitió la obtención de resultados aceptables de predicción, sin embargo, existen otras técnicas como el agrupamiento y las reglas asociación las cuales pueden ser evaluadas para determinar los mejores porcentajes de predicción.

El proceso en general, hizo uso de un total de cincuenta variables entre académicas, psicológicas y financieras. De acuerdo a los diferentes documentos analizados, tanto en el marco teórico como en el marco de antecedentes, se encuentra que muchos de ellos realizaron sus análisis haciendo uso de no más de diez variables. De acuerdo con esto, es importante analizar si el número de variables analizadas influye directamente en los porcentajes de predicción obtenidos, si es así, se recomienda realizar diferentes análisis que procesen un número reducido de variables, por ejemplo, las quince variables de deserción identificadas en la tabla 48.

Teniendo en cuenta las diferentes modalidades de estudio que ofrece la UNAB, como la educación virtual y los convenios con otras ciudades, es recomendable ajustar el modelo desarrollado y aplicarlos en estos escenarios para validar los resultados obtenidos.

BIBLIOGRAFÍA

Amaya, Y. y Barrientos, E. y Heredia, D. (2015). Student dropout predictive model using data mining techniques. IEEE Latin America Transactions, vol. (13).

Azoumana, K. (2013). Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos. Pensamiento Americano, 41-51

Bouckaert, R. (2010). WEKA---Experiences with a java open-source project. The Journal of Machine Learning Research, vol. (11), pp. 2533-2541. Recuperado de http://dl.acm.org/citation.cfm?id=1953016

Cabena, P. (1998). Discovering Data mining From Concept To Implementation. Estados Unidos: Prentice Hall

Castellanos Guarín, L. (2015). Incorporación de Elementos de Inteligencia de Negocios al Análisis de Deserción Estudiantil de la universidad Autónoma de Bucaramanga (Tesis de Maestría). Universidad Autónoma de Bucaramanga, Colombia.

Chapman, P. y Clinton, J. y Kerber, R. y Khabaza, T. y Reinartz, T. y Shearer, C. y Wirth, R. (2000). CRISP-DM 1.0. Recuperado de ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Demšar, J. (2016). Orange: Data mining toolbox in python. The Journal of Machine Learning Research, vol. (14), pp. 2349-2353. Recuperado de http://dl.acm.org/citation.cfm?id=2567736&CFID=789665709&CFTOKEN=93305719

Departamento Nacional de Planeación. (2016). Visión Colombia II Centenario. Recuperado de http://www.mineducacion.gov.co/cvn/1665/article-95980.html

Ekkachai, N. Jatsada, S. Nittaya, K. (2012). Classification Model Induction For Student Recruiting, Latest Advances In Educational Technologies. Recuperado de http://www.wseas.us/e-library/conferences/2012/Singapore/EDUC/EDUC-18.pdf

Fayyad, U. y Piatetsky-Shapiro, G. y Smyth, P. (1996). From data mining to knowledge discovery: an overview. AI Magazine, vol. (17), pp. 37-54. Recuperado de https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Han, J. y Kamber, M. y Pei, J. (2001). Data mining: Concepts and techniques. Amsterdam: Morgan Kaufmann Publishers.

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

Howson, C. (2007). Successful business intelligence: Secrets to making bi a killer App. Estados Unidos: McGraw-Hill Education.

International Educational Data Mining Society. (s.f.). Educational Data Mining. Recuperado el 10 de abril de 2016 de http://educationaldatamining.org

Kumar, S. y Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal, vol. (2), pp. 51-56.

The University Of Waikato (s.f.). Weka 3 - data mining with open source machine learning software in java. Recuperado de http://www.cs.waikato.ac.nz/ml/weka/

Ministerio de Educación Nacional. (2010). Deserción estudiantil en la educación superior colombiana. Recuperado el 15 abril de 2016 de http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

Ministerio de Educación Nacional. (2015a). Estrategias Para la Permanencia en Educación Superior: Experiencias Significativas. Recuperado de http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-350844_pdf.pdf.

Ministerio de Educación Nacional (2015b). Guía para la implementación del modelo de gestión de permanencia y graduación estudiantil en instituciones de educación superior. Recuperado de http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-355193_guia_.pdf

Ministerio de Educación Nacional (2016). Estadísticas de deserción y graduación 2015. Recuperado de http://www.colombiaaprende.edu.co/html/micrositios/1752/articles-350629_estadisticas_pdf2015.pdf

Moine, J. y Haedo, A. y Gordillo, S. (2001). Estudio comparativo de metodologías para minería de datos. XIII Workshop de Investigadores en Ciencias de la Computación. Recuperado de http://sedici.unlp.edu.ar/handle/10915/20034

Muenchen, B. (2017). The Popularity of Data Science Software. Recuperado de http://r4stats.com/articles/popularity

Oracle Help Center (2016). Data Warehousing and Business Intelligence. Recuperado de https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm

Parr Rud, O. (2000). Data mining cookbook: Modeling data for marketing, risk, and customer relationship management. United States: Wiley, John & Sons.

Rangra, K. (2014). Comparative study of data mining tools. International Journal of Advanced Research in Computer Science and Software Engineering, vol. (04), pp. 6

Remco, R. Eibe, F. (2016). Weka Manual for Version 3-8-1. Recuperado de http://www.cs.waikato.ac.nz/ml/weka/documentation.html

Rohanizadeh, S. y Moghadam, M. (2010). A proposed data mining methodology and its application to industrial procedures. Journal of Industrial Engineering, vol. (4), pp. 37-50.

Sauter, V. (2011). Decision support systems for business intelligence. New Jersey, Estados Unidos: United Kingdom: Wiley-Blackwell.

Society for Learning Analytics Research – SoLAR. (2014). About SOLAR. Recuperado el 10 de abril de 2016 de http://educationaldatamining.org/

Statistical Analysis System - SAS Institute Inc (2003). Data Mining Using SAS Enterprise Miner: A Case Study Approach. Recuperado de http://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf

Timaran Pereira, R. (2009). Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la Minería de Datos. Recuperado el 20 de mayo de 2016 de http://www.redalyc.org/html/1053/105317327011/

Turban, E. y Sharda, R. y Denle, D. y King, D. (2013). Business intelligence: A managerial perspective on Analytics. Boston, Estados Unidos: Prentice Hall

Universidad Autónoma de Bucaramanga (2016a). Acerca de la UNAB. Recuperado de http://unab.edu.co/nosotros/acerca-de

Universidad Autónoma de Bucaramanga (2016b). Bienestar Universitario. Recuperado de http://unab.edu.co/nosotros/bienestar-universitario

University of Ljubljana (2016). License Orange. Recuperado de http://orange.biolab.si/license/

Valenzuela, J. Flores, M. (2014). Fundamentos de Investigación Educativa. Ciudad de México, México: Editorial digital del tecnológico de Monterrey.

Vercellis, C. (2011). Business intelligence: Data mining and optimization for decision making. United States: Wiley, John & Sons.

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

ANEXO A: Formato de archivo arff

Encabezado de Archivo arff

```
@relation TRAINING_201601-weka.filters.unsupervised.attribute.Remove-R1-4

@attribute fecha_nacimiento numeric
@attribute genero {FEMENINO,MASCULINO}
@attribute poblacion_especial {'HABITANTE FRONTERA',AFRODESCENDIENTE,INDIGENA,'DISCAPACIDAD SENSORIAL',DESPLAZADO,'DISCAPACIDAD MOTRIZ','DISCAPACIDAD COGNITIVA'}
@attribute nivel_comunicativo {ALTO,BAJO,SUPERIOR,MEDIO}
@attribute nivel_matematicas {BAJO,INFERIOR,MEDIO,SUPERIOR,ALTO}
@attribute puntaje_icfes {'40-50 PUNTOS','MAS DE 50 PUNTOS','0-40 PUNTOS'}
@attribute tiempo_ingreso_unab {'0 - 2 AÑOS','2 - 4 AÑOS','MAS DE 6 AÑOS','4 - 6 AÑOS'}
@attribute rendimiento_academico {'ENTRE 70\% Y 90\%','ENTRE 50\% Y 70\%','100\% APROBADOS','MENOS DE 20\%','ENTRE 20\% Y 50\%'}
@attribute ser_pilo_paga {P}
@attribute procedencia {'AREA METROPOLITANA DE BUCARAMANGA','FUERA DEL AREA METROPOLITANA DE BUCARAMANGA'}
@attribute estado_civil {'SOLTERO (A)','CASADO (A)','UNION LIBRE','SEPARADO (A)','DIVORCIADO (A)'}
@attribute hijos {NINGUNO,UNO,DOS,'TRES O MAS'}
@attribute personas_cargo {SI,NO}
@attribute estrato {CUATRO,TRES,CINCO,SEIS,UNO,DOS}
@attribute convive {'FUERA DEL HOGAR',FAMILIA}
@attribute trabaja {NO,SI}
@attribute vinculacion_laboral {EMPLEADO,INDEPENDIENTE}
@attribute sector_productivo {SERVICIOS,INDUSTRIAL,MANUFACTURA}
@attribute cargo {OPERARIO,JEFATURA,DIRECTIVO,COORDINADOR,SUPERVISOR}
@attribute salario {'MENOS DE UN SALARIO MINIMO','ENTRE 1 Y 2 SALARIOS MINIMOS','ENTRE 3 SALARIOS MINIMOS O MAS'}
@attribute horario_laboral {'FINES DE SEMANA',TURNOS,NOCTURNO,'HORARIO DE OFICINA'}
@attribute fuente_pago_estudios {'RECURSOS PROPIOS','OTRA ENTIDAD',ICETEX}
@attribute tipo_colegio {PRIVADO,PUBLICO}
@attribute 10M numeric
@attribute 11N numeric
@attribute _120 numeric
@attribute 13Q numeric
@attribute 14R numeric
@attribute 15S numeric
@attribute 16T numeric
@attribute 1A numeric
@attribute 2B numeric
@attribute 3C numeric
@attribute 4E numeric
@attribute 5F numeric
@attribute 6G numeric
```

Data de Archivo arff

```
@attribute deserto {SI,NO}

@data
1997,MASCULINO,?,BAJO,?,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CINCO,FAMILIA,NO,?,?,?,?
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',P,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UNO,FAMILIA,NO,?,?,?
1999,MASCULINO,?,SUPERIOR,BAJO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','MENOS DE 20\%',P,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,'FUERA
1997,MASCULINO,?,BAJO,BAJO,'0-40 PUNTOS','2 - 4 AÑOS','MENOS DE 20\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,SEIS,FAMILIA,SI,INDEPENDIENTE,
1987,FEMENINO,DESPLAZADO,BAJO,BAJO,'40-50 PUNTOS','2 - 4 AÑOS','ENTRE 20\% Y 50\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',UNO,SI,UNO,'FUERA
1999,FEMENINO,?,BAJO,?,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',P,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UNO,'FUERA DEL
1995,FEMENINO,?,BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,FAMILIA,SI,INDEPENDIENTE
1998,MASCULINO,?,BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA,NO,?,?,?,?,?,'RE
1998,FEMENINO,?,BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UNO,FAMILIA,NO,?,?,?,?,?,ICET
1997,FEMENINO,'DISCAPACIDAD SENSORIAL',BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,F
1998,FEMENINO,?,BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','MENOS DE 20\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,'FUERA DEL HOGAR',
1996,MASCULINO,?,MEDIO,?,'0-40 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA,NO,?,?,?,?,?,ICE
1998,FEMENINO,?,BAJO,?,'0-40 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,SEIS,FAMILIA,NO,?,?,?,?,?,'REC
1998,MASCULINO,?,SUPERIOR,?,'0-40 PUNTOS','0 - 2 AÑOS','MENOS DE 20\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,SI,CUATRO,'FUERA DEL H
1998,MASCULINO,?,MEDIO,?,'0-40 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,FAMILIA,NO,?,?,?,?,?,'R
1999,FEMENINO,?,?,?,'0-40 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA,NO,?,?,?,?,?,
1997,FEMENINO,?,MEDIO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CUATRO,FAMILIA,NO,?,?,?,?,?,
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 70\% Y 90\%',?,'AREA METROPOLITANA DE BUCARAMANGA','CASADO (A)',NINGUNO,NO,CUATRO,FAMILIA,NO,?,?,?
1998,FEMENINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','MENOS DE 20\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,FAMILIA,NO
1995,MASCULINO,DESPLAZADO,BAJO,INFERIOR,'40-50 PUNTOS','4 - 6 AÑOS','ENTRE 20\% Y 50\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UN
1995,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','2 - 4 AÑOS','ENTRE 20\% Y 50\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CUATRO,FAM
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 70\% Y 90\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CINCO,FAMILIA,NO,?,?
1997,FEMENINO,?,SUPERIOR,BAJO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',P,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,FAMILIA,NO,?,
1995,MASCULINO,?,BAJO,BAJO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA
1998,MASCULINO,'HABITANTE FRONTERA',ALTO,?,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',P,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUN
1997,MASCULINO,?,BAJO,ALTO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',P,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UNO,'FUERA DEL HOGAR'
1997,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 70\% Y 90\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CINCO,'FUERA DEL HOG
1997,MASCULINO,'HABITANTE FRONTERA',BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',?,'FUERA DEL AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NI
1998,MASCULINO,?,BAJO,?,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,CUATRO,FAMILIA,?,?,?,?,?
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',P,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA,NO,?,?,?
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','MENOS DE 20\%',P,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,UNO,FAMILIA,NO,?,?,?,?,?
1999,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','100\% APROBADOS',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,TRES,FAMILIA,NO,?,?,?,
1997,MASCULINO,?,BAJO,MEDIO,'MAS DE 50 PUNTOS','0 - 2 AÑOS','ENTRE 50\% Y 70\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,DOS,FAMILIA,NO,?,?,?
1998,MASCULINO,?,BAJO,?,'40-50 PUNTOS','0 - 2 AÑOS','ENTRE 20\% Y 50\%',?,'AREA METROPOLITANA DE BUCARAMANGA','SOLTERO (A)',NINGUNO,NO,SEIS,FAMILIA,NO,?,?,?,?,?,'R
```