UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

Carrera de Ingeniería en Sistemas

Proyecto de Investigación previo a la obtención del título de Ingeniero en Sistemas

TÍTULO DEL PROYECTO DE INVESTIGACIÓN

ANÁLISIS DE APROVECHAMIENTO ACADÉMICO EN LA UNIDAD EDUCATIVA 24 DE MAYO APLICANDO REGLAS DE ASOCIACIÓN

AUTOR:

Geisson David Coello Chabla

DIRECTOR DE PROYECTO DE INVESTIGACIÓN:

Ph.D. Amilkar Yudier Puris Cáceres

Quevedo – Los Ríos – Ecuador

2016

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

Carrera de Ingeniería en Sistemas

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, **Coello Chabla Geisson David**, declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Coello Chabla Geisson David

C.C. # 1206245787

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

Carrera de Ingeniería en Sistemas

CERTIFICACIÓN DE CULMINACIÓN DEL PROYECTO DE INVESTIGACIÓN

El suscrito **Ph.D. Amilkar Yudier Puris Cáceres** Docente de la Universidad Técnica Estatal de Quevedo, certifico que el estudiante **Coello Chabla Geisson David** realizó el Proyecto de Investigación de grado titulado **"ANÁLISIS DE APROVECHAMIENTO ACADÉMICO EN LA UNIDAD EDUCATIVA 24 DE MAYO APLICANDO REGLAS DE ASOCIACIÓN"** previo a la obtención del título de Ingeniero en Sistemas bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

Ph.D. Amilkar Puris Cáceres

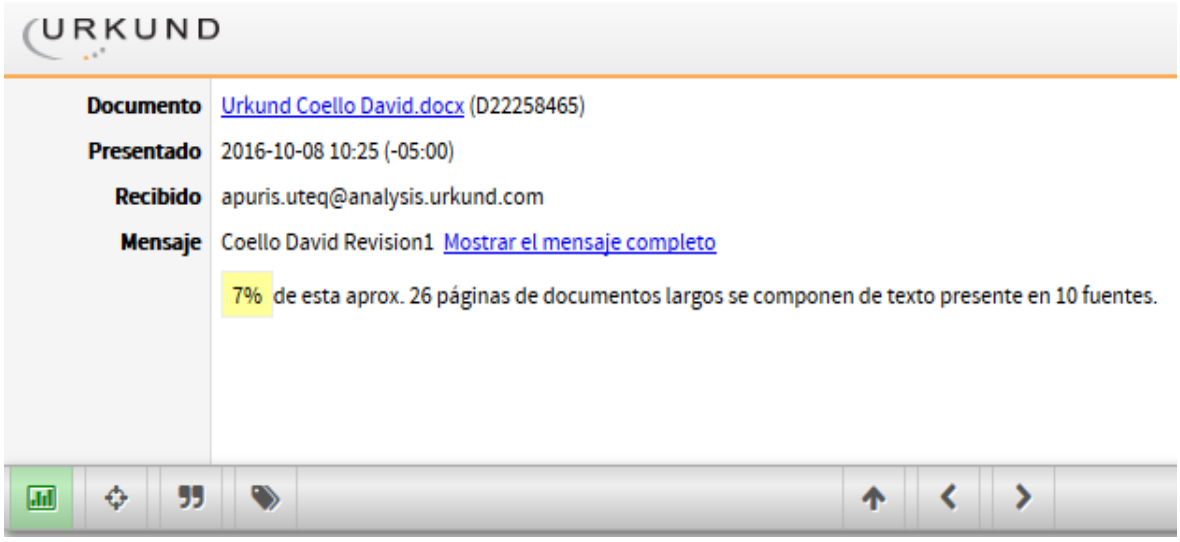
DIRECTOR DE PROYECTO DE INVESTIGACIÓN

CERTIFICADO DEL REPORTE DE LA HERRAMIENTA DE PREVENCIÓN DE COINCIDENCIA Y/O PLAGIO ACADÉMICO

INFORME DEL DIRECTOR DE TESIS SOBRE EL SISTEMA URKUND

Ph.D. Amilkar Yudier Puris Cáceres, en calidad de director del proyecto de investigación **"ANÁLISIS DE APROVECHAMIENTO ACADÉMICO EN LA UNIDAD EDUCATIVA 24 DE MAYO APLICANDO REGLAS DE ASOCIACIÓN"**, me permito manifestar a usted lo siguiente:

Que, el Sr. Coello Chabla Geisson David, egresado de la facultad de Ciencias de la Ingeniería, ha cumplido con las correcciones pertinentes, e ingresado su proyecto de investigación al sistema URKUND, tengo a bien certificar la siguiente información sobre el informe del sistema anti plagio con un porcentaje del **7%**



Ph.D. Amilkar Puris Cáceres
DIRECTOR DE PROYECTO DE INVESTIGACIÓN

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

Carrera de Ingeniería en Sistemas

CERTIFICACION

Los abajo firmantes que actuamos como **Presidente y Miembros del Tribunal** de la Unidad de Titulación, designada por el Consejo Académico de la Facultad de Ciencias de la Ingeniería, tenemos a bien **CERTIFICAR** que el estudiante **COELLO CHABLA GEISSON DAVID,** se le procedió a efectuar la respectiva revisión de las correcciones planteadas en la sustentación de su proyecto de investigación titulado **"ANÁLISIS DE APROVECHAMIENTO ACADÉMICO EN LA UNIDAD EDUCATIVA 24 DE MAYO APLICANDO REGLAS DE ASOCIACIÓN".**

Dando cumplimiento a lo estipulado por el reglamento firmado dando fe de lo actuado, los que conformamos dicho tribunal.

Quevedo, 28 de noviembre del 2016

Atentamente

Ing. Guerrero Ulloa Gleiston Cicerón, MBA.	Ph.D. Pavel Novoa Hernández
PRESIDENTE DEL TRIBUNAL	MIEMBRO DEL TRIBUNAL
Ing. Eduardo Samaniego Mena, MSc.	Ing. Moisés Arturo Menace Almea, MSc.
MIEMBRO DEL TRIBUNAL	REDACCIÓN TÉCNICA

AGRADECIMIENTO

En especial y sobre todas las cosas Dios por bendecirme en cada paso que doy y darme la fuerza necesaria para cumplir todas mis metas.

A mi familia por brindarme su apoyo incondicional, sus concejos y la manera en que cada día me inspiraron a la finalización de mi carrera.

A mi director de proyecto Ph.D. Amilkar Puris Cáceres, al Ph.D. Pavel Novoa, Ing. Iván Jaramillo e Ing. Nancy Rodríguez que con sus amplios conocimientos me han guiado para la consecución del mismo, y a cada una de las personas que de una u otra manera me brindaron su apoyo durante este tiempo de estudio, me es imposible mencionar a todos mis compañeros y mis amigos que siempre estuvieron impulsándome para el desarrollo de mi carrera, Gracias!.

DEDICATORIA

Dedico este trabajo a Dios quien siempre me acompaña y me ayuda a levantar durante mis continuos tropiezos.

A mi familia, que son las personas que siempre me apoyan de manera incondicional en cada paso que doy, apoyando cada una de mis decisiones.

RESUMEN EJECUTIVO Y PALABRAS CLAVES

El presente proyecto exhibe una nueva alternativa en la búsqueda de información para caracterizar el aprovechamiento académico de los estudiantes de la Unidad Educativa 24 de Mayo de la ciudad de Quevedo, tanto como lo es la promoción del año lectivo, es decir su aprobación o reprobación del mismo así como el rendimiento que hayan tenido a lo largo de sus estudios secundarios.

Este proyecto emplea la minería de datos mediante reglas de asociación para la extracción del conocimiento en cuanto a la asociación de variables que caractericen el aprovechamiento académico. Para obtener las variables de clase se adicionan dos atributos al conjunto de datos de los estudiantes, estas definen el aprovechamiento académico de los estudiantes, se recurre a formar dos bases de datos con las mismas variables, con la diferencia de que cada una de las variables de clase estarán en diferentes bases de datos.

Se realiza un análisis descriptivo para corroborar que exista un patrón y coherencia en los datos a través de los períodos lectivos desde el 2008 al 2014. Por medio de este análisis también se demuestra gráficamente que existen valores fuera de rango, por ende se emplean técnicas de pre-procesamiento para obtener bases de datos más concisas.

Se utiliza la técnica de selección de atributos, la misma permite identificar que variables son más significativas en dependencia de la variable clase. Posterior a esta etapa se procede a realizar el balanceo de las muestras para que no exista un favoritismo en los atributos dentro de las variables que más registros contengan. Finalmente se extraen las reglas de asociación mediante la ejecución de algoritmos, en cuanto a las reglas se establecen las mejores basadas en los paramentos de soporte y confianza para su posterior interpretación con la ayuda de dos expertos en educación que evaluaran las mismas.

Palabras clave: minería de datos, reglas de asociación, aprovechamiento académico.

ABSTRACT AND KEYWORDS

This project shows a new alternative in the search of information to characterize the academic achievement of students of the Unidad Educativa 24 de Mayo of the Quevedo, such as, the promotion of the academic year, it means it´s approbation or reprobation in the same way the performance they have had throughout their scholar year.

This project uses Data Mining association rules for the extraction of knowledge regarding the association of variables that characterize academic achievement. For obtaining class variables two attributes to the set of student data are added, these define the academic achievement of students, it is used to form two database with the same variables, with the difference that each one of the variables class will be on different databases.

A descriptive analysis was performed to confirm that there is a pattern and data consistency across the school periods from 2008 to 2014. Through this analysis graphically also it shows that there are values out of range, therefore techniques are used pre-processing for more concise databases.

The Feature Selection technique is used, it identifies which variables are most important in dependence of the variable class. After this stage proceeds to balancing the samples in order to there is no favoritism in attributes within the variables that contain records. Finally the association rules by running algorithms are extracted, as to the rules-based best facings support and confidence for later interpretation with the help of two education experts to evaluate them.

Keywords: data mining, association rules, academic achievement.

ÍNDICE

Índice de tablas

Índice de figuras

Índice de anexos

CÓDIGO DUBLIN

Título:	"Análisis de aprovechamiento académico en la Unidad Educativa 24 de Mayo aplicando Reglas de Asociación".		
Autor:	Coello Chabla Geisson David		
Palabras clave:	minería de datos	reglas de asociación	aprovechamiento académico.
Fecha de publicación:			
Editorial:			
Resumen:	El presente proyecto exhibe una nueva alternativa en la búsqueda de información para caracterizar el aprovechamiento académico de los estudiantes de la Unidad Educativa 24 de Mayo de la ciudad de Quevedo, tanto como lo es la promoción del año lectivo, es decir su aprobación o reprobación del mismo así como el rendimiento que hayan tenido a lo largo de sus estudios secundarios. Este proyecto emplea la minería de datos mediante reglas de asociación para la extracción del conocimiento en cuanto a la asociación de variables que caractericen el aprovechamiento académico. Para obtener las variables de clase se adicionan dos atributos al conjunto de datos de los estudiantes, estas definen el aprovechamiento académico de los estudiantes, se recurre a formar dos bases de datos con las mismas variables, con la diferencia de que cada una de las variables de clase estarán en diferentes bases de datos. Se realiza un análisis descriptivo para corroborar que exista un patrón y coherencia en los datos a través de los períodos lectivos desde el 2008 al 2014. Por medio de este análisis también se demuestra gráficamente que existen valores fuera de rango, por ende se emplean técnicas de pre-procesamiento para obtener bases de datos más concisas. Se utiliza la técnica de selección de atributos, la misma permite identificar que variables son más significativas en dependencia de la variable clase. Posterior a esta etapa se procede a realizar el balanceo de las muestras para que no exista		

	un favoritismo en los atributos dentro de las variables que más registros contengan. Finalmente se extraen las reglas de asociación mediante la ejecución de algoritmos, en cuanto a las reglas se establecen las mejores basadas en los paramentos de soporte y confianza para su posterior interpretación con la ayuda de dos expertos en educación que evaluaran las mismas. Palabras clave: minería de datos, reglas de asociación, aprovechamiento académico.
Descripción:	96 hojas : dimensiones, 29 x 21 cm + CD-ROM
URI:	

INTRODUCCIÓN

Como antecedente del presente proyecto investigativo se puede mencionar que la Unidad Educativa 24 de Mayo [1], surgió con la necesidad ya que en el sector Galo Plaza no existía un establecimiento de nivel medio. A partir del 8 de mayo de 1983 comienza una ardua labor docente y administrativa, con 60 alumnos y 5 profesores. Actualmente la institución cuenta con más de 800 alumnos en su sección diurna y vespertina.

Recientemente, en el año 2013 el Gobierno en curso del Presidente Rafael Correa Delgado implementó una nueva modalidad de gestión académica, siendo la misma llevada a cabo mediante calificaciones de dos notas quimestrales valoradas en un máximo de 10 puntos. Esto resultó un cambio con respecto a la modalidad anterior que era llevado por tres notas trimestrales valoradas en un máximo de 20 puntos.

La presente investigación aborda el problema del aprovechamiento académico de los estudiantes en la Unidad Educativa 24 de Mayo del Cantón Quevedo. Desde una perspectiva de análisis inteligente de los datos se emplea la aplicación de técnicas de minería de datos[1] para la identificación de las variables y asociaciones entre estas, identificando las más relevantes que favorezcan el aprovechamiento académico de los estudiantes.

Específicamente se trabajará con dos bases de datos entre los periodos lectivos desde 2008 hasta el 2014 de la referida Unidad Educativa, a los que se les aplicará técnicas de Selección de Rasgos[2] y de Minería de Reglas de Asociación (ARM)[3].

A partir de los resultados que se esperan de la presente investigación se pretende asociar variables que relacionen el aprovechamiento académico de la Unidad Educativa 24 de Mayo, implementando reglas de asociación de gran importancia para la toma de decisiones. Cabe mencionar que en el sistema educativo del cantón Quevedo no existen registros sobre el uso de técnicas de minería de datos, como lo es la minería de reglas de asociación para la extracción del conocimiento que se puede obtener mediante el empleo de la misma.

[1] En inglés Data Mining.
[2] En inglés: Feature selection, también conocido como selección de características o selección de variables.
[3] En inglés: Association rule Minning,

CAPÍTULO I

CONTEXTUALIZACIÓN DE LA INVESTIGACIÓN

1.1.Problematización

En la Unidad Educativa 24 de Mayo no existe un proceso en el cual desde el inicio de las actividades académicas en cada año lectivo, les permita conocer a las autoridades y docentes del plantel, cuáles son los alumnos candidatos a obtener un bajo rendimiento académico o pérdida del año lectivo, dado que desconocen los factores que inciden en el mismo.

No cuentan con herramientas informáticas que les permita conocer acerca del aprovechamiento académico de los estudiantes para encarar tempranamente acciones tendientes a revertir la situación de los alumnos que tengan problemas en cuanto a su rendimiento académico.

La institución no ha realizado estudios en años anteriores, que les pueda dar información acerca del aprovechamiento de los estudiantes, a más de esto se conoce que en la ciudad de Quevedo las instituciones secundarias no han realizado análisis alguno que les permita mejorar el rendimiento académico de los estudiantes y a la vez disminuir la deserción estudiantil o pérdida del año lectivo.

A continuación se detalla el planteamiento del problema con su respectivo diagnóstico y pronostico, para posteriormente dar a conocer la formulación y sistematización del mismo.

1.1.1. Diagnóstico

A partir de conversaciones mantenidas con los directivos de la Unidad Educativa 24 de Mayo se comprobó que no cuentan con instrumentos de análisis de datos, que les permita valorar de manera efectiva el aprovechamiento académico de los estudiantes.

Concretamente, los principales recursos con los que cuentan los directivos son las calificaciones tabuladas mediante la herramienta informática de Microsoft Excel por parte del personal docente de la institución y registros digitalizados con información acerca de los estudiantes en el departamento de secretaria de la institución.

A esto se le suma que hasta donde se conoce en el Cantón Quevedo, no existen estudios previos en el ámbito educativo acerca del análisis de los datos, desaprovechando el conocimiento que se podría obtener para favorecer así la toma de decisiones de los directivos de las instituciones secundarias.

1.1.2. Pronóstico

La falta de análisis en los datos de estudiantes de la Unidad Educativa 24 de Mayo sobre su aprovechamiento académico, determinará que los directivos de la institución no cuenten con información que respalde la toma de decisiones en el ámbito académico.

1.1.3. Formulación del problema

¿Cuáles son las relaciones que existen entre las variables que determinan el aprovechamiento académico en la Unidad Educativa 24 de Mayo?

1.1.3.1. Sistematización

- ¿Cómo comprobar si una base de datos presenta registros concisos?
- ¿Cuáles son los pasos a seguir para obtener una base de datos consistente, con el fin de realizar un correcto análisis de los datos?
- ¿Cuáles son los umbrales necesarios para la extracción de conocimiento?
- ¿De qué manera se puede dar a conocer el conocimiento obtenido al realizar un análisis datos basado en la ejecución de diferentes algoritmos?

1.2. Objetivos

1.2.1. Objetivo general

A partir del problema científico formulado anteriormente, la presente investigación se plantea como objetivo general:

- Obtener un modelo basado en reglas de asociación que caracterice el aprovechamiento académico en la Unidad Educativa 24 de Mayo.

1.2.2. Objetivos específicos

- Analizar de forma descriptiva la información proporcionada, buscando coherencia y patrones entre los datos.

- Aplicar técnicas de pre-procesamiento para obtener bases de datos consistentes y aplicar diferentes modelos de reglas de asociación basados en el criterio de soporte y confianza.

- Interpretar las reglas de asociación obtenidas, para transmitir el conocimiento aportado por las mismas.

1.3. Justificación

La recolección y almacenamiento de grandes volúmenes de datos ha sido una de las tareas más comunes en todo tipo de actividad realizada, ya sea esta comercial, educativa, investigativa, laboral, etc., puesto que se hace necesario contar con registros, los mismos pueden ser un histórico acerca de los diferentes tipos de datos o para almacenar información que en algún momento será de utilidad para la toma de decisiones.

Se desconoce de estudios desarrollados acerca de la efectividad del análisis inteligente de datos en el ámbito educativo del cantón Quevedo. Los factores que determinan el

aprovechamiento académico en la educación básica y de bachillerato pueden ser varios, sin embargo contar con unas pocas variables no ordenadas, con diferentes formatos e inclusive errores, no permiten realizar un análisis minucioso y detallado con herramientas tradicionales de análisis estadístico por no ser capaces de poder extraer conocimiento por medio de la ejecución de algoritmos.

Es allí donde se propone el uso de la minería de datos, conociendo que la misma se fundamenta en la búsqueda de patrones dentro de grandes bases de datos, empleando diversos métodos de inteligencia artificial como de estadística, haciendo uso de recursos informáticos y tecnológicos para la extracción de conocimiento con el fin de aportar información que no se obtiene al emplear análisis estadísticos tradicionales.

Se busca por tanto aprovechar los beneficios de este tipo de análisis con el fin de extraer conocimiento oculto en los datos, beneficiando a las autoridades educativas al contar con un información que favorezca la toma de decisiones en lo que respecta al aprovechamiento académico de los estudiantes mediante el modelo de reglas de asociación.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

2.1. Marco Conceptual

A continuación se caracteriza aquellos elementos que intervienen en el proceso de la investigación, siendo los mismos los más relevantes que forman parte del presente proyecto investigativo, con el fin de brindar información adicional.

2.1.1. Minería de datos

La minería de datos es considerada un conjunto de técnicas y herramientas como lo denominan varios autores, como se establece en [2] y [3], se define como un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos.

Es importante mencionar que otros autores dan una denominación semejante a la de un conjunto de técnicas y herramientas, solo que a la misma le suman la aplicación de algoritmos como lo da a entender [4] la minería de datos utiliza diversas técnicas y algoritmos para descubrir el conocimiento de grandes cantidades de datos e identificar patrones comprensibles a partir de los mismos. Es considerado como una de las tendencias más importantes de la tecnología de la información en la década anterior.

2.1.2. Fases del proceso de minería de datos

Para implementar la minería de datos se necesita de la aplicación de algoritmos, más una sucesión de actividades previas para la correcta aplicación de la misma, como por ejemplo el pre-procesamiento de los datos entre otras, en algunas ocasiones los datos a los que se les desea aplicar minería proceden de fuentes heterogéneas es decir están compuesto por características diferenciadoras, cuentan con un formato incorrecto o contienen datos erróneos incluso redundantes.

Cabe resaltar que una de las principales acciones al implementar la minería de datos, es darle una correcta interpretación y evaluación a los resultados obtenidos, por ende se han establecido fases para la ejecución de la minería como lo detallan [5] el proceso completo consta de las siguientes fases:

1. Definición de los objetivos.
2. Preparación de datos.
3. Análisis exploratorio de los datos.
4. Especificación de los métodos.
5. Análisis de los datos.
6. Evaluación de los métodos.
7. Implementación de los métodos.

2.1.3. Selección de rasgos

La selección de rasgos se centra en encontrar los atributos más importantes dentro de los que se han utilizado y eliminar aquellos que se consideran irrelevantes, para ello en [6] se define la selección de rasgos en un conjunto de datos como un problema en cuya solución se han utilizado diversas variantes dentro de la inteligencia artificial, de esta forma, grandes volúmenes de datos se pueden manipular rápidamente al extraer de ellos sólo la información necesaria que los describa, sin perder la calidad del sistema, y obteniendo conocimiento sobre ellos.

2.1.4. Reglas de asociación

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común como se establece en [7]. En este contexto, la denominada minería de reglas de asociación consiste en encontrar reglas de la forma (A1yA2y...yAm) => (B1yB2y...yBn), donde Ai y Bj son valores de atributos del conjunto de datos. Como por ejemplo, en una base de datos de compras en un supermercado, la regla de asociación correspondiente a que si un cliente compra leche, entonces compra pan.

2.1.5. Descripción de KEEL

Como se describe en [8] KEEL (Knowledge Extraction based on Evolutionary Learning), es una herramienta software no comercial escrita en Java, permite al usuario emplear algoritmos evolutivos en diferentes tipos de problemas de minería de datos: Regresión, clasificación, agrupamiento, asociación, etc.

Una de las ventajas del software KEEL, es que permite ejecutar sus algoritmos dentro del propio entorno conocido también como ejecución online, a su vez generarlos para una ejecución posterior, es decir poder ejecutarlos en distintas maquinas. A esto se le conoce como ejecución off-line. La versión que se emplea en el presenten proyecto investigativo está compuesta por los siguientes módulos que se aprecian en la Figura 1.

Figura 1: Pantalla principal de KEEL

A continuación se detallan los principales módulos de KEEL:

Tratamiento de datos[4]: Se encuentran una serie de herramientas de tratamiento de datos para la transformación de las bases de datos como lo es: Importación, exportación, edición y visualización de datos, aplicación de transformaciones, etc.

Experimentos[5]: Este módulo genera procedimientos de análisis y evaluación automática de algoritmos off-line, proporcionando numerosas opciones: Tipo de validación, tipo de

[4] En ingles: Data Management
[5] En ingles: Experiments

aprendizaje (clasificación, regresión, aprendizaje no-supervisado), etc, en esta sección se procederá a emplear los algoritmos utilizados en el presente proyecto investigativo.

Educacional[6]: Realiza la ejecución de algoritmos on-line. Tiene una estructura similar al módulo anterior, pero permite diseñar experimentos para ser ejecutados paso a paso.

Ayuda[7]: Manual acerca del funcionamiento del software KEEL.

2.1.6. Descripción de WEKA

WEKA se trata de un acrónimo derivado de Waikato Environment for Knowledge Analysis (Entorno para Análisis del Conocimiento de la Universidad de Waikato) como se denomina en [9] está compuesta por una serie de herramientas gráficas de visualización y diferentes algoritmos para el análisis de datos y modelado predictivo.

El software WEKA ofrece herramientas para implementar las reglas de asociación y el pre-procesamiento de datos, consta de 4 módulos principales que se puede observar en la Figura 2, son descritos a continuación:

Figura 2: Pantalla principal de WEKA

En Explorador[8], se encuentran varios paneles que dan acceso a las principales características del programa, como lo son el pre-procesamiento de los datos, la

[6] En ingles: Educational
[7] En ingles: Help
[8] En ingles: Explorer

implementación de algoritmos de clasificación, agrupamiento, asociación, selección de variables y por ultimo visualización.

- Experimentar[9], permite la comparación sistemática de una ejecución de los algoritmos predictivos de WEKA sobre una colección de conjuntos de datos.

- Flujo de conocimiento[10], soporta esencialmente las mismas opciones que la interfaz Explorer, pero esta permite "arrastrar y soltar". Ofrece soporte para el aprendizaje incremental.

- Interface de línea de comando simple[11], permite el acceso a través de consola de comandos a todas las opciones de WEKA.

2.1.7. Descripción de algoritmos para realizar la investigación

A continuación en la Tabla 1 se detalla los algoritmos con los que cuenta la herramienta WEKA y KEEL para la extracción de reglas de asociación.

Tabla 1: Algoritmos basados en Reglas de asociación en KEEL y WEKA.

Algoritmos	Basado en ítems frecuentes	Algoritmos Genéticos Mono Objetivo	Los algoritmos evolutivos Multi Objetivos (AEMOs)
Apriori (WEKA)	X		
Apriori (KEEL)	X		
Eclat (KEEL)	X		
GENAR (KEEL)		X	
GAR (KEEL)		X	
EARMGA (KEEL)		X	
Alatasetal (KEEL)		X	
MOEA Ghosh (KEEL)			X
MODENAR (KEEL)			X
ARMMGA (KEEL)			X
MOPNAR (KEEL)			X
QAR_CIP_NSGA (KEEL)			X

[9] En ingles: Experimenter
[10] En ingles: Knowledge Flow
[11] En ingles: Simple command line interface

Fuente: Web oficial de KEEL http://www.keel.es y web oficial de WEKA Machine Learning Group at the University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/
Elaborado por el autor

Apriori →Apriori-A [10] A priori, este algoritmo se basa en el conocimiento previo o "a priori" de los conjuntos frecuentes, esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia.

Equivalence Class Transformation → Eclat-A [10] Transformación clase de equivalencia, este algoritmo define subárboles de búsqueda, mediante las clases de equivalencia de los conjuntos de ítems.

Genetic Association Rules → GENAR-A [11] Reglas de asociación genéticas, emplea una longitud fija de manera que sólo el último atributo de la base de datos actúa como consiguiente, mientras que todas las anteriores representan la parte antecedente.

Genetic Association Rules → GAR-A [11] Reglas de asociación genéticas, descubre conjuntos de elementos frecuentes de un tamaño variable que abarca tanto como sea posible los intervalos de dominio de todos los atributos numéricos sin la necesidad de emplear la discretización.

Evolutionary Association Rules Mining with Genetic Algorithm → EARMGA-A [12] de Minería de reglas de asociación evolutivas con algoritmo genético, emplea una longitud fija en la que los atributos numéricos están representados por una unión finita de intervalos uniformes.

Association Rules Mining by means of a genetic algorithm proposed by Alatas et al. →Alatasetal-A [13], algoritmo genético propuesto por Alatas et al., diseña un algoritmo genético para simultáneamente buscar los intervalos de los atributos cuantitativos.

Multi-objective rule mining using genetic algorithms → [14] MOEA_Ghosh-A Minería de reglas multi-objetivo usando algoritmos genéticos, utiliza un algoritmo genético basado en

el frente de Pareto[12] para extraer algunas reglas útiles e interesantes de cualquier conjunto de datos.

Multi-objective differential evolution algorithm for mining association rules → MODENAR-A [13] algoritmo de evolución diferencial multi-objetivo para la minería de reglas de asociación, el aprendizaje genético se realiza de forma iterativa hasta que se alcanza un número máximo de las evaluaciones.

ARMMGA → ARMMGA-A: [15] Explora reglas de asociación sin tener en cuenta el apoyo mínimo, la confianza y el grado de interés, y extraer las mejores reglas con altos valores de estas medidas

Multi-Objective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules → MOPNAR-A [16], algoritmo evolutivo multi-objetivo propuesto por Martin et al., concebido para la minería de un conjunto reducido de reglas de asociación positivas y negativos que son interesantes, fáciles de entender y con un buen equilibrio entre el número de reglas, el apoyo y la cobertura del conjunto de datos.

QAR_CIP_NSGAII → QAR_CIP_NSGAII-A: [16] Minería de algoritmos evolutivos multi-objetivo emplea un conjunto reducido de reglas de asociación positivos y negativos, se realiza de forma iterativa hasta que se alcanza un número máximo de las evaluaciones.

Medidas de algoritmo en WEKA:

- Media de Confianza: la probabilidad de encontrar la parte derecha de una regla condicionada a que se encuentre también la parte izquierda.
- Media Soporte: El 'soporte' de un conjunto de ítems X en una base de datos D se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de ítems.

Medidas del algoritmo en KEEL:

[12] Realizar un cambio sin hacer que empeore la situación.

- #R: número medio de reglas obtenidas.
- MedSop: valores medios de soporte de cada regla.
- MedConf: valores medios de confianza.
- MedLift: valores medios de lift.

  Representa la confianza de la regla y la confianza esperada de la regla, toma valores en el intervalo [0,1) donde valores menores que 1 representan dependencia negativa, 1 representa independencia y mayores que 1 dependencia positiva.
- MedConv valores medios de conviction.

  Mide la dependencia entre X y Y , toma valores en el intervalo [0,1), donde valores menores que 1 representan dependencia negativa, 1 representa independencia y mayores que 1 representan dependencia positiva.
- MedFC: valores medios de factor certeza.

  Mide la variación de la probabilidad de que Y este en un registro cuando se consideran solo los registros donde está X. toma valores en el intervalo [-1,1] donde valores positivos y negativos representan dependencia positiva y negativa respectivamente y 0 representa independencia.
- MedNetconf: valores medios de netconf.

  Evalúa una regla basándose en el soporte de la regla, del antecedente y del consecuente. Netconf obtiene valores en el intervalo [-1,1] donde valores positivos y negativos representan dependencia positiva y negativa respectivamente y 0 representa independencia.
- MedYulesQ: valores medios de yule'sQ.

  Representa la correlación entre dos eventos posiblemente relacionados. Esta medida obtiene valores en el intervalo [-1,1], donde 1 implica una correlación positiva perfecta, -1 implica una correlación negativa perfecta y 0 implica que no hay correlación.
- MedAmp: número medio de atributos involucrados en las reglas .
- %Reg: tanto por ciento de registros de la BD cubiertos por las reglas generadas.
- Tiempo: tiempo en segundos que demoro la ejecución del algoritmo.

2.1.8. Definición de rendimiento académico

De acuerdo con [17] el rendimiento académico es un nivel de conocimientos demostrado en un área o materia comparado con la norma de edad y nivel académico, se puede

observar la similitud que tiene con la definición de [18] que lo define como el producto de la asimilación del contenido del programa de estudio expresado en calificaciones dentro de una escala convencional.

2.1.8.1 Factores del rendimiento académico

Los factores que influyen en el rendimiento escolar son muchos y variados, igualmente los criterios para clasificarlos, como lo menciona [19] dentro de las clasificaciones más importantes se encuentran los informes de PISA[13], TIMSS[14], PIRLS[15], el informe centrado en el análisis de los factores que condicionan la adquisición de conocimientos básicos, es manejado en cuatro niveles:

1. Nivel sistémico: Contempla las características del sistema educativo.
2. Nivel estructural: Formado por las características del entorno socioeconómico.
3. Nivel escolar: Relacionado básicamente con aspectos de la dirección del centro y el clima escolar.
4. Nivel individual: Concerniente con la trascendencia de las actitudes, la motivación y la conducta de cara al aprendizaje por parte de los alumnos.

El bajo rendimiento no tiene que ver solamente con los estudiantes, como lo menciona [20] sino con muchas otras personas y factores, por ende, las malas notas no solo son el resultado de las evaluaciones de los alumnos, sino también de los profesores, los textos, los métodos, el plantel y los padres de familia, Torres.

2.2.Marco Referencial

En lo que sigue se describirán los trabajos relacionados con el objeto de estudio de la presente investigación. Estos trabajos han sido ordenados de manera cronológica.

[13] PISA (Programme for International Student Assessment) en español: programa para la evaluación internacional de los estudiantes.
[14] TIMSS (Trends in International Mathematics and Science Study) en español: tendencias en matemáticas y ciencias
[15] PIRLS (Progress in International Reading Literacy Study)en español: estudio internacional de progreso de comprensión

2.2.1 Minería de datos educativa: una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo

Este trabajo implementa la minería de datos, para encontrar patrones que permitan mejorar el aprendizaje de los estudiantes, y por ende mejorar su rendimiento académico. Como se detalla por sus autores, [21] se presenta un estudio de los métodos y técnicas de Minería de Datos Educativa (Educational Data Mining – EDM, por sus siglas en inglés); aplicada a la educación, como una herramienta donde el investigador podrá tener un margen mayor de análisis de la información a través de los datos generados por la aplicación, apoyándose de igual manera en algoritmos de inteligencia artificial.

2.2.2 Predicción del fracaso escolar mediante técnicas de minería de datos

En este trabajo se propone la utilización de técnicas de minería de datos para detectar, cuáles son los factores que más influyen para que los estudiantes de enseñanza media o secundaria fracasen, es decir, suspendan o abandonen. Además como se menciona en [22] se propone utilizar diferentes técnicas de minería de datos debido a que es un problema complejo, los datos suelen presentar una alta dimensionalidad (hay muchos factores que pueden influir) y suelen estar muy desbalanceados (la mayoría de los alumnos suelen aprobar y sólo una minoría suele fracasar). El objetivo final es detectar lo antes posible a los estudiantes que presenten esos factores para poder ofrecerles algún tipo de atención o ayuda para tratar de evitar y/o disminuir el fracaso escolar.

2.2.3 Metodología de estudio del rendimiento académico mediante la minería de datos

Busca determinar los patrones de éxito y de fracaso académico de los alumnos mediante la aplicación de la minería de datos. Como se menciona en [23] la minería de datos orientada a la educación permite predecir determinado tipo de factor o característica de un caso, fenómeno o situación. Se consideran especialmente modelos de minería de agrupamiento, clasificación y asociación. En todos los casos se busca determinar los patrones de éxito y de fracaso académico de los alumnos para, de esta manera, predecir la probabilidad de los mismos de desertar o tener un bajo rendimiento académico, con la ventaja de poder hacerlo tempranamente, permitiendo así encarar acciones tendientes a revertir tal situación.

CAPÍTULO III
MÉTODOLOGÍA DE LA INVESTIGACIÓN

A continuación se presenta la localización del desarrollo de la investigación en el apartado 3.1, posteriormente se describe el tipo de investigación a emplear en la Sección 3.2. En la Sección 3.3, se formalizan los métodos y técnicas usados, luego en 3.4 se darán a conocer las fuentes de recopilación de información. En la Sección 3.5 se expone el diseño de la investigación, finalmente en el apartado 3.6 se dará a conocer los recursos humanos y materiales.

3.1.Localización

La presente investigación se basa en la información de los estudiantes de la Unidad educativa 24 de Mayo, ubicada en la calle Juan de Dios Avilés Zarate No. 100, en la ciudadela Santa Rosa, perteneciente a la parroquia urbana 24 de Mayo del Cantón Quevedo, provincia de Los Ríos, dentro del territorio Continental del Ecuador.

Cabe mencionar que parte del desarrollo de la investigación se la realizo en la Universidad Técnica Estatal de Quevedo (UTEQ), ubicada en la Av. Quito, Km 1.5 Vía Sto. Domingo en el Sector Nuevo Quevedo, perteneciente a la parroquia urbana Nicolás Infante Díaz del Cantón Quevedo, provincia de Los Ríos, dentro del territorio Continental del Ecuador.

La UTEQ cuenta con Acreditación Categoría B, dirigida por las siguientes autoridades:
Rector: Dr. Eduardo Díaz Ocampo, MSc.
Vicerrectora Académico: Ing. Guadalupe del Pilar Murillo Campuzano, MSc.
Vicerrector Administrativo: Ing. Bolívar Roberto Pico Saltos, MSc.

El cantón Quevedo se encuentra ubicada geográficamente en la región litoral-centro, 1° 20' 30" de Latitud Sur y los 79° 28' 30" de Longitud occidental, limitado con los siguientes cantones; al Norte con Buena Fe y Valencia, al Sur Mocache y Quinsaloma, al Este Quinsaloma y al Oeste El Empalme. La ciudad es considerada un punto de enlace entre la sierra y la costa ecuatoriana, dado que está fuertemente comunicada por vías terrestres que la conectan a varias provincias.

3.2. Tipos de Investigación

Para el desarrollo de este proyecto se utilizaron los siguientes tipos de investigación: de campo e investigación aplicada. La investigación asumió un nivel exploratorio dado que se buscó la realidad que enfrenta la Unidad Educativa 24 de Mayo con respecto al aprovechamiento académico de los estudiantes, mientras que también se menciona el desconocimiento de aplicar minería de reglas de asociación en este campo.

3.2.1. Investigación de campo

Uno de los tipos de investigación empleados en el presente proyecto es la de campo, por las siguientes razones:

- Se basó en el testimonio del secretario de la Unidad Educativa 24 de Mayo, el cual mencionó la dificultad de predecir el aprovechamiento académico de los estudiantes.

- Otra razón es el desconocimiento de la aplicación de la minería de datos en este campo de estudio, se plantearon reglas de asociación en cuanto al aprovechamiento académico.

3.2.2. Investigación aplicada

Este proyecto se considera también como una investigación aplicada, lo anterior se sustenta en las siguientes razones:

- Se obtuvo toda la información necesaria para el desarrollo, como lo fue la base de datos de los alumnos proporcionada por el secretario de la Unidad Educativa 24 de Mayo.

- Se planteó caracterizar la relación entre un conjunto de variables que influyen en el aprovechamiento académico de los estudiantes, mediante la ejecución de reglas de asociación.

3.3.Métodos de Investigación

Los métodos empleado en el presente proyecto investigativo son: el método de observación, deductivo y el de extracción de conocimiento de bases de datos, a continuación se da una definición de método y metodología en general como lo manifiesta [24] se define método como una forma planificada y estructurada de llevar a cabo una acción, la metodología, sin embargo explica, analiza y describe los métodos, sus limitaciones, sus fuentes así como las presuposiciones y consecuencias de los resultados de aplicar cada uno de los distintos métodos.

3.3.1.Método observación

La investigación siguió el método observacional con el que los resultados estuvieron limitados a establecer reglas que caractericen la relación entre un conjunto de variables que contiene la base de datos de los estudiantes, permitiendo caracterizar el aprovechamiento académico tanto como lo es la promoción y el rendimiento, dando a sugerir causalidad (ya que ésta sólo puede demostrarse mediante estudios experimentales).

Este método fue elegido dado que, por un lado, existen ciertas variables fundamentales que no pueden ser objeto de manipulación como sexo, edad, etc. y por otro lado los diseños experimentales exigen una alta complejidad y ciertos impedimentos éticos, derivados de la propia naturaleza del objeto de estudio, como lo sería el conocer por menores de la vida privada de los estudiantes.

3.3.2. Método deductivo

Es parte de un principio general ya conocido para inferir en él, consecuencias particulares, expresados de una forma más sencilla, la deducción consistió en partir de una teoría general para expresar hechos o fenómenos particulares. En el caso del presente proyecto investigo, por medio de la aplicación de este método se pudo deducir la manera de caracterizar las variables que definan el aprovechamiento académico de los estudiantes.

3.3.3. Método de extracción de conocimiento de bases de datos

El proceso de extracción de conocimiento de base de datos para el presente proyecto investigativo sigue una metodología que consta de una serie de fases que llevaron un orden respectivo empleando cinco fases según [25] la *Integración y recopilación* es la primera, se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas, se convertirán los datos en bruto a los formatos adecuados para trabajar según los requerimientos de software utilizados.

Se procedió a realizar una de las fases más importante dentro de la minería de datos como lo es el *Pre-procesado de los datos* en dicha fase se emplea la limpieza, transformación y completamiento de datos incompletos.

Posteriormente se obtuvo la fase de *Selección de Variables* en la misma se ejecutan algoritmos que presentan las variables más significativas dependiendo de la clase empleada, la misma que permitirán alcanzar los objetivos propuestos, es decir se realiza una reducción de variables, se emplean técnicas de selección de rasgos conocidas como *Feature Selection*[16].

Cuando se han tenido los datos pre-procesados se puede proceder a realizar la siguiente fase que es *Extracción de conocimiento* en la cual se identificarán las reglas de asociación importantes que permitan tomar decisiones - una vez establecido el umbral de soporte y confianza, se emplea el algoritmo adecuado para la extracción de reglas.

Como fase final se tiene la *Interpretación y evaluación* la misma que consiste en una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

[16] Selección de variables

3.4. Fuente de Recopilación de la Información

En esta sección se revisó las principales características de la base de datos concedida para el desarrollo de la presente investigación, a más de esto, se describen los conjuntos de datos que serán utilizados. Para el desarrollo del presente proyecto investigativo y en dirección a lograr los objetivos expuestos se utilizó la siguiente fuente de investigación:

3.4.1. Fuente primaria

La base de datos en el formato de Excel, la misma que cuenta con información acerca de los estudiantes de los diferentes periodos lectivos, desde el año 2008 hasta el año 2014.

3.4.2. Fuente secundaria

Responsable: Secretario general de la Unidad Educativa 24 de Mayo, quien fue el facilitador de la base de datos de los alumnos de dicha institución, entrega de la información realizada el 7 de diciembre del 2015, otra fuente que se obtuvo es la obtenida por medio de la ejecución de los algoritmos para la minería de datos, dado que los mismos permitieron la extracción del conocimiento al emplear la caracterización de las variables por medio de reglas para definir el aprovechamiento académico de los estudiantes.

3.5. Diseño de la Investigación

El presente proyecto investigativo empleo como diseño de la investigación, el diseño experimental, como lo define [26] la experimentación es el método que permite descubrir con mayor grado de confianza, relaciones de tipo causal entre hechos o fenómenos de la realidad, por ello es el tipo y nivel más alto de investigación científica. Este diseño es un tipo de investigación experimental en la que el investigador tiene control sobre las variables que se van a utilizar.

Esta investigación se basa fundamentalmente en la observación de fenómenos tal y como se dan en su contexto natural para analizarlos con posterioridad. Como lo detalla [26] en síntesis, la experimentación se caracteriza por la provocación del fenómeno que se estudia,

la manipulación de las variables, el control de la situación experimental y la utilización de la comparación.

3.5.1. Diagrama de flujo acerca de la investigación

A continuación se detalla el diagrama de flujo de las actividades realizadas en la presente investigación con el fin de dar a conocer los procesos que se han realizadas en la misma. Para esto se detalla en la Figura 3 el significado de cada grafico empleado en el diagrama.

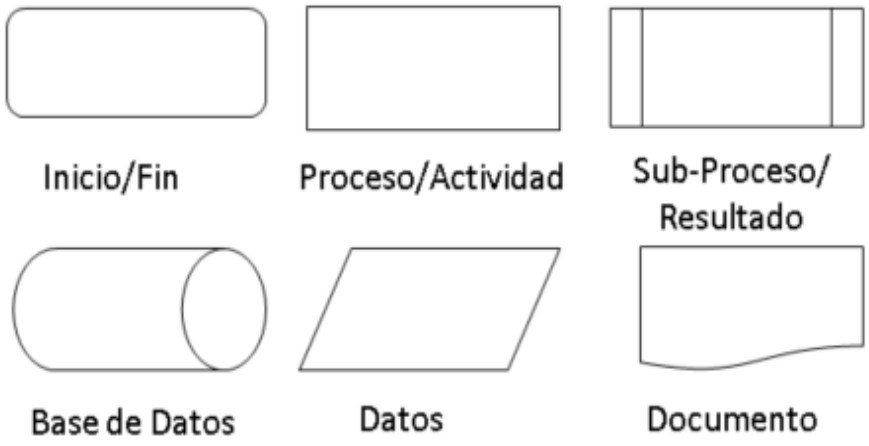
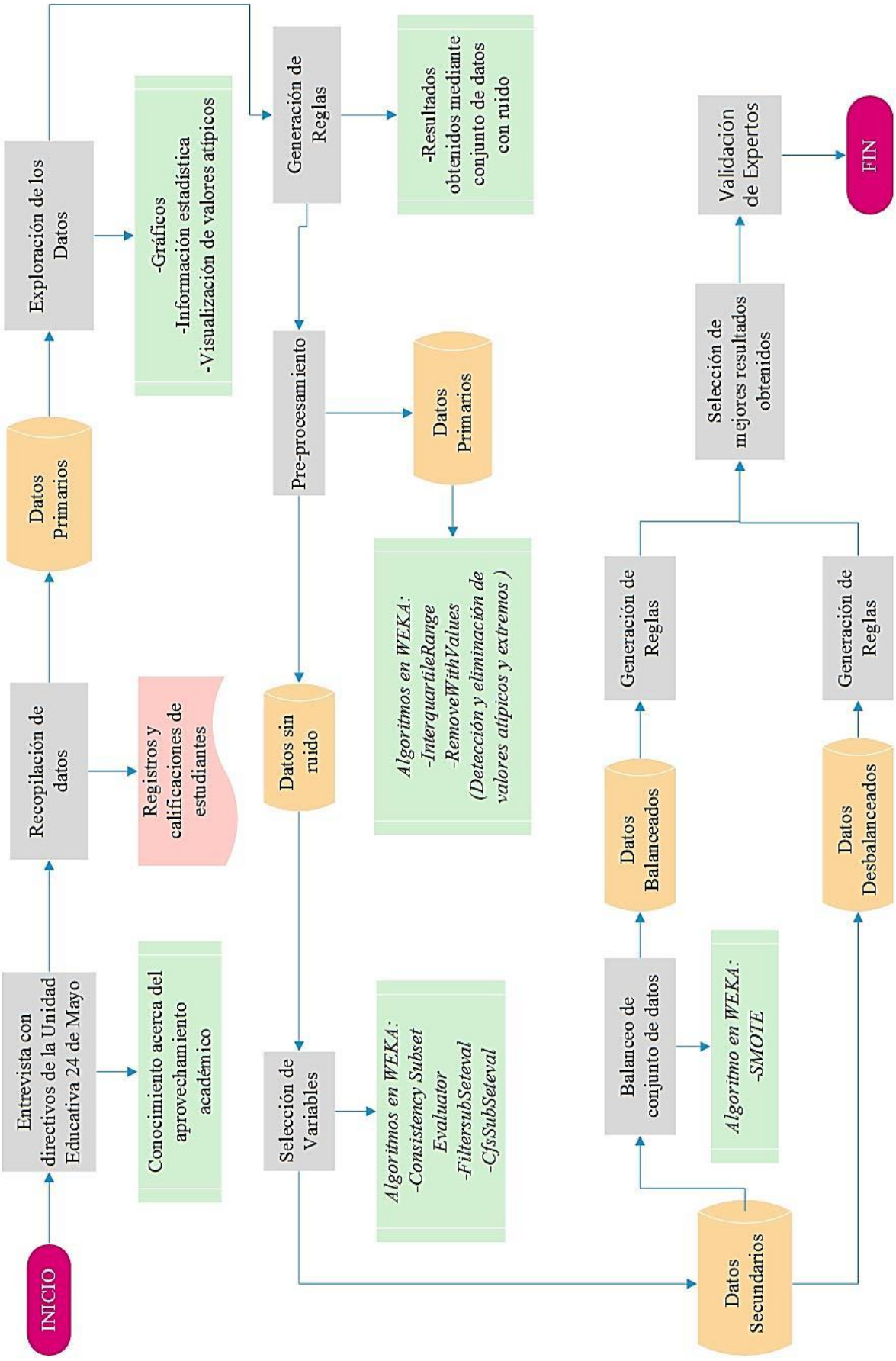


Figura 3: Descripción de gráficos empleados para el diagrama de flujo.

3.5.1.1 Diagrama de flujo de actividades



Elaborado por el autor.

3.6. Recursos Humanos y Materiales

Es importante mencionar que como en todo proyecto investigativo es necesario el equipo humano y materiales empleados para el desarrollo del mismo, a continuación se detallan el equipo humano que estuvo a cargo de la realización del proyecto y los materiales necesarios para la ejecución del mismo.

3.6.1. Equipo humano

El presente proyecto investigativo está elaborado por el estudiante David Coello Chabla, el mismo que se encargó de la recopilación de información, desarrollo, documentación, guiado por el Ph.D. Amilkar Puris Cáceres como director del proyecto de investigación.

3.6.2. Equipos y materiales

El presente proyecto fue realizado con varios materiales así mismo gracias a equipos de importancia, dentro de estos se puede encontrar equipos informáticos, de oficina, que se detallarán a continuación.

3.6.2.1 Hardware

- Computadora (Laptop)
  - Procesador: Intel(R) Core(TM) i5-4210U CPU @1.70GHz 2.40 GHz
  - Memoria instalada (RAM): 8,00 GB
  - Tipo de sistema: Sistema operativo de 64 bits, procesador x64
  - Disco Duro: 500 GB
- Otros
  - Impresora EPSON L220
  - Pendrive HP 8 GB

3.6.2.2 Software

- Sistema Operativo
  - Windows 10 Home Single Lenguaje

- Software para minería de datos
  - WEKA
  - KEEL
- Utilitarios
  - Microsoft Office Professional Plus 2013

3.7. Presupuesto

3.7.1. Presupuestos del proyecto en suministros de oficina

Tabla 2: Suministro de oficina

Recurso	Cant.	Valor unitario	Valor total
Servicio de Internet(mensual)	6	$22,40	$134,40
Resma de Papel A4	5	$3,00	$15,00
Tinta para impresora continua	4	$12,00	$48,00
Carpetas	3,00	$0,50	$1,50
Pendrive 8GB	1,00	$10,00	$10,00
Anillado	3,00	$1,00	$3,00
Total			$211,90

Fuente: Gastos realizados por el autor, 2016.
Elaborado por el autor

3.7.2. Presupuesto general del proyecto

Tabla 3: Gastos generales

Recurso	**Valor**	**Valor Total**
Suministros de oficina	$ 211,90	$211,90
Movilización	$ 50,00	$50,00
Gastos varios	$ 100,00	$100,00
TOTAL		$361,90

Fuente: Gastos realizados por el autor, 2016.
Elaborado por el autor

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

4.1. Resultados

En el presente capítulo se da a conocer la experimentación que se desarrolló para la extracción de reglas que caractericen la relación entre un conjunto de variables. Se parte de la descripción de las mismas, para posteriormente realizar un análisis estadístico simple con el fin de corroborar la relación entre las mismas.

Se emplearon técnicas de minería de datos, pasando por varias fases como lo son la preparación de los datos, ejecución de algoritmos para reglas de asociación, extracción de conocimiento e interpretación y evaluación de los resultados. Para la aplicación de las técnicas de minería se emplearon las herramientas informáticas WEKA y KEEL.

4.1.1. Descripción de información a utilizar

Para la implementación del presente proyecto investigativo se conformó, una única base de datos de estudiantes de la Unidad Educativa 24 de Mayo desde el año 2008 hasta el 2014 dado que la información otorgada por el secretario de la institución, estaba dividida en periodos lectivos.

Principalmente se emplearon dos tipos de archivos en EXCEL, siendo los mismos, la base de matrículas de cada periodo lectivo, y los archivos que contenían las calificaciones y nombres de los estudiantes, desde octavo curso hasta sexto curso.

Las variables con las cuales se trabajaron están detalladas en la Tabla 4 donde se da a conocer el nombre de la misma, el significado, el tipo y cuál será la descripción que toma dependiendo de los valores que tome.

Tabla 4: Variables empleadas en la Base de datos.

VARIABLE	SIGNIFICADO	TIPO	DESCRIPCIÓN	
periodclases	Periodo de clases comprendido entre 2008 y 2014	Numérico	1 --> 2013-2014	4 --> 2010-2011
			2 --> 2012-2013	5 --> 2009-2010
			3 --> 2011-2012	6 --> 2008-2009
curso	Identificación del curso	Numérico	1 --> Octavo	4 --> Primero de Bach
			2 --> Noveno	5 --> Segundo de Bach

			3 --> Décimo	6 --> Tercero de Bach
edadanoscurso	Edad del estudiante.	Numérico	Edad comprendida entre 10 y 26 años	
sexo	Sexo del estudiante	Numérico	1 Masculino	2 Femenino
nacionalidad	Nacionalidad del estudiante	Numérico	1 Ecuatoriana 2 Colombiana 3 Otro	
lugnacprov	Provincia de Nacimiento	Numérico	Ver anexo1 para codificación	
lugnaccant	Cantón de Nacimiento	Numérico	Ver Anexo 1 para codificación	
lugnacparroq	Parroquia de Nacimiento	Numérico	Ver Anexo 1 para codificación	
lugresiprov	Provincia de Residencia	Numérico	Ver Anexo 1 para codificación	
lugresicant	Cantón de Residencia	Numérico	Ver Anexo 1 para codificación	
lugresiparroq	Parroquia de Residencia	Numérico	Ver Anexo 1 para codificación	
asisteeducacinicial	Asistencia a Educación Inicial	Numérico	1-->No Asiste 0--> Asiste	
jornacadm	Jornada Académica	Numérico	1 Diurna 3 Nocturna	
vivecon	Con quien convive el estudiante	Numérico	1 --> Padre y Madre	4--> Abuelos
			2--> Solo Madre	5 --> Otro Familiar
			3 --> Solo Padre	6 --> Solo/a
diasnoasistidos	Días no asistidos	Numérico	Días no asistidos en el año lectivo	
numsuplet	Cantidad de supletorios	Numérico	Cantidad de supletorios en el año lectivo	
notaanual	Nota anual	Numérico	Calificación anual en el año lectivo	
COMPORTAM	Comportamiento anual	Numérico	Calificación del comportamiento en el año lectivo.	

Fuente: Base de estudiantes Unidad Educativa 24 de Mayo 2008-2014
Elaborado por el autor

4.1.1.1. Análisis estadístico de los datos

Es importante mencionar que la calidad de los resultados conseguidos por medio del método de extracción de conocimiento empleados en el presente proyecto investigativo, no solo depende del método en sí, sino también de cómo se haya conformado la base de datos en la cual se va a emplear las reglas de asociación.

Al aplicar las técnicas respectivas para obtener una base de datos lo más sólida posible, se asegura un resultado más confiable en cuanto al conocimiento extraído, dado que las mencionadas técnicas permiten obtener información veraz acerca del objeto de estudio.

Luego de obtener una sola base de datos de todos los registros de los estudiantes comprendidos entre los periodos 2008 y 2014 dando un total de 6245 registros, se realizó el primer análisis, detallado a continuación.

En el Gráfico 1 se puede observar la cantidad de estudiantes por cada año lectivo, seguidamente de la jornada académica en la que estudiaron, el sexo de los mismos y la cantidad de alumnos por cada año lectivo, a continuación en la Tabla 5 se detallan dichos registro.

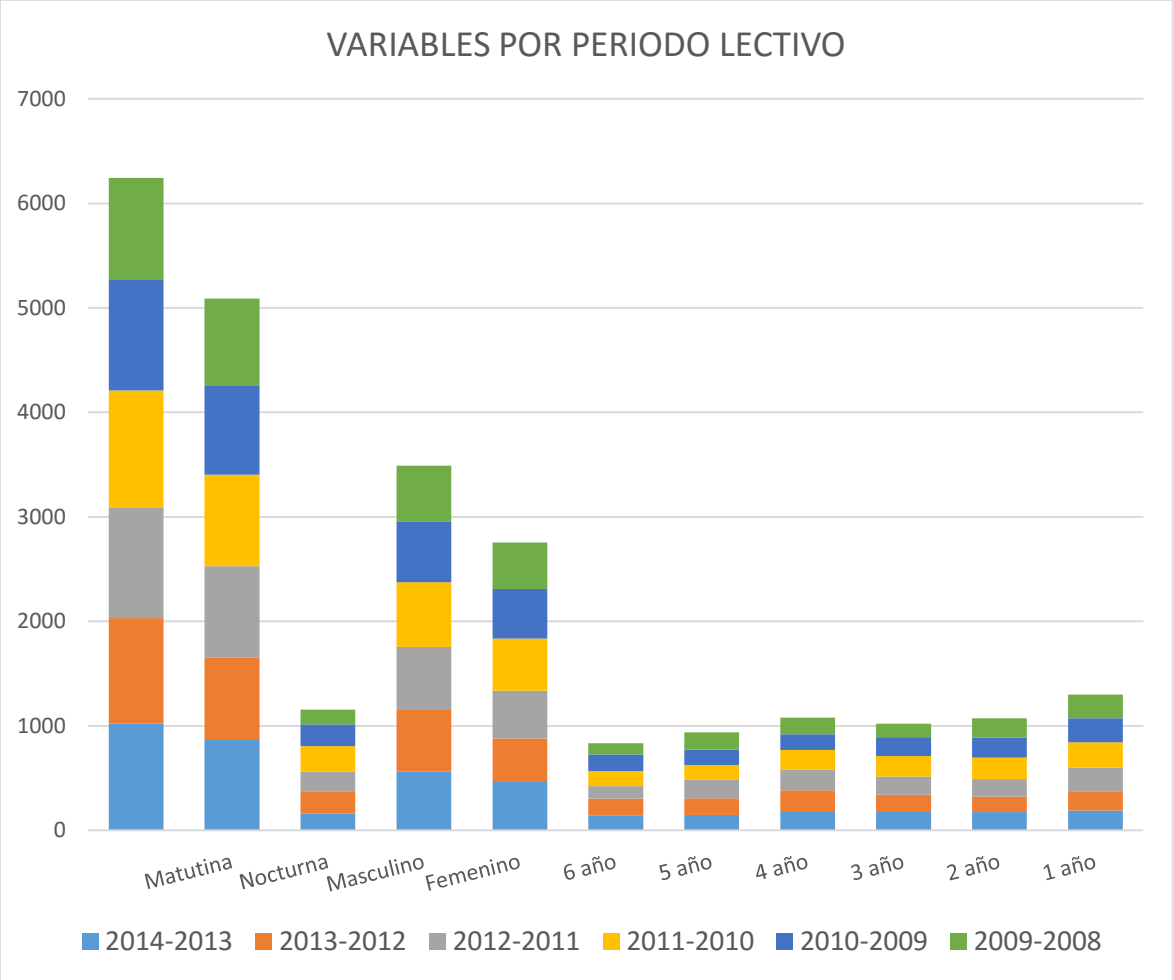


Gráfico 1: Análisis descriptivo de la base de datos.

Tabla 5: Variables por periodo lectivo de la base de datos.

Periodo Lectivo		2014-2013	2013-2012	2012-2011	2011-2010	2010-2009	2009-2008
Registros		1023	1008	1060	1120	1056	978
Jornada	Matutina	865	791	873	875	849	837
	Nocturna	158	217	187	245	207	141
Sexo	Masculino	560	592	603	621	582	532
	Femenino	463	416	457	499	474	446
Cursos	6 año	143	162	118	145	158	109
	5 año	148	158	178	142	145	166
	4 año	183	200	195	190	150	161
	3 año	184	156	175	196	181	128
	2 año	174	149	168	205	192	186
	1 año	191	183	226	242	230	228

Fuente: Base de estudiantes Unidad Educativa 24 de Mayo 2008-2014
Elaborado por el autor

Un segundo análisis estadístico representado en el Gráfico 2, es el obtenido por medio de las variables edad y sexo de los registros con los que cuenta la base de datos, el mismo gráfico hace referencias a edades mayores a los mil años, o en la variable sexo el valor de 10 o 0, datos que demuestran que existe valores atípicos, por ende se emplearon técnicas de pre-procesamiento de datos.

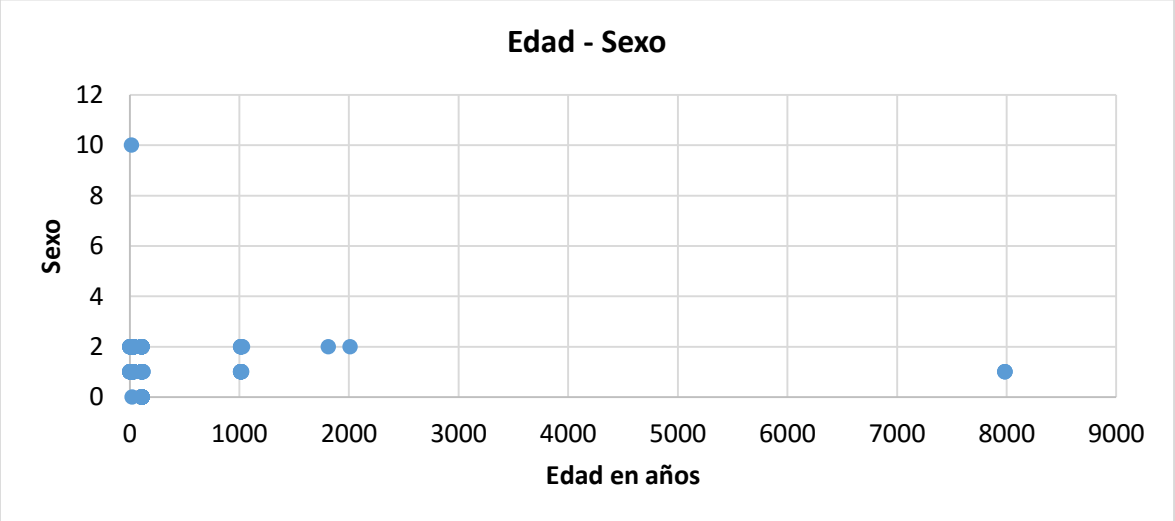


Gráfico 2: Edad y sexo de los estudiantes en la base de datos.

Una vez conformada la base de datos se procedió a realizar una copia de la misma, con el objetivo de adicionar tres variables, dos de las cuales son las *variables de clase*[17] que

[17] Es uno de los atributos simbólicos disponibles, que se convierte en la variable objetivo a predecir

fueron las reemplazantes de la variable “NotaAnual”, esto con el fin de contar con dos bases de datos para sus respectivos estudios.

Las variables adicionadas antes mencionadas se pueden apreciar en la Tabla 6, posterior a este paso se cuenta con dos bases de datos las mismas que se denominaran *promocion*[18] y *rendimiento*[19].

Tabla 6: Variables adicionadas en la Base de datos

VARIABLE	Significado	Tipo	Descripción
distancolegio	Distancia desde la residencia del alumno a la institución	Numérico	Ver Tabla 7
*pierdeanio	Promoción del estudiante	Numérico	1 Pierde año 2 No pierde año
*rendimiento	Rendimiento académico	Numérico	1) >=9 <=10 2) >=7 <=8.99 3) >=4.01 <=6.99 4) >=0 <=4
***Variable de clase para las reglas de asociación**			

Fuente: Base de estudiantes Unidad Educativa 24 de Mayo 2008-2014
Elaborado por el autor

Para la obtención de información acerca de la distancia desde la residencia del estudiante hasta la institución educativa, como lo es la variable “DistanColegio” empleada en la Tabla 6, se recurrió a una distribución basada en las parroquias con las que cuenta la ciudad de Quevedo.

La variable “DistanColegio” está definida en varias categorías, como lo son “corta”, la misma que hace referencia a que el domicilio del estudiante se encuentra a menos de 5 cuadras de la Unidad Educativa 24 de Mayo, se puede observar está medida y las demás empleadas para la definición de la distancia a la institución en la Tabla 7.

Tabla 7: Distancia de residencia a la institución

SÍMBOLO	DISTANCIA	DESCRIPCIÓN
1	Corta (<5 cuadras)	Parroquia 8 sector 24 Mayo

[18] Base de datos sobre la promoción académica del estudiante.
[19] Base de datos sobre el rendimiento académico del estudiante.

2	Media (>5 cuadras y < 10 cuadras)	Parroquia 5 sector Nicolás Infante Díaz
		Parroquia 9 sector VENUS DEL RIO QUEVEDO"
3	Grande (>10 cuadras y < 20 cuadras)	Parroquia 7 sector SIETE DE OCTUBRE
		Parroquia 10 sector VIVA ALFARO
		Parroquia 1 sector CENTRO Quevedo
4	Dentro del perímetro urbano	Parroquia 4 sector GUAYACAN
		Parroquia 3 sector SAN JOSE
		Parroquia 2 sector SAN CAMILO
5	Fuera de la ciudad	Parroquia 55 LA ESPERANZA
		Cantón 54 VALENCIA
		Parroquia 53 SAN CARLOS
		Cantón 52 MOCACHE
		Demás códigos fuera de la ciudad

Fuente: Ministerio de educación del Ecuador**,** anexo de codificación de la provincia de Los Ríos

Un tercer análisis estadístico representado en el Gráfico 3, es el obtenido por medio de las variables agregadas a las bases de datos en estudio, como se detalla a continuación, clasificado por periodo lectivo.

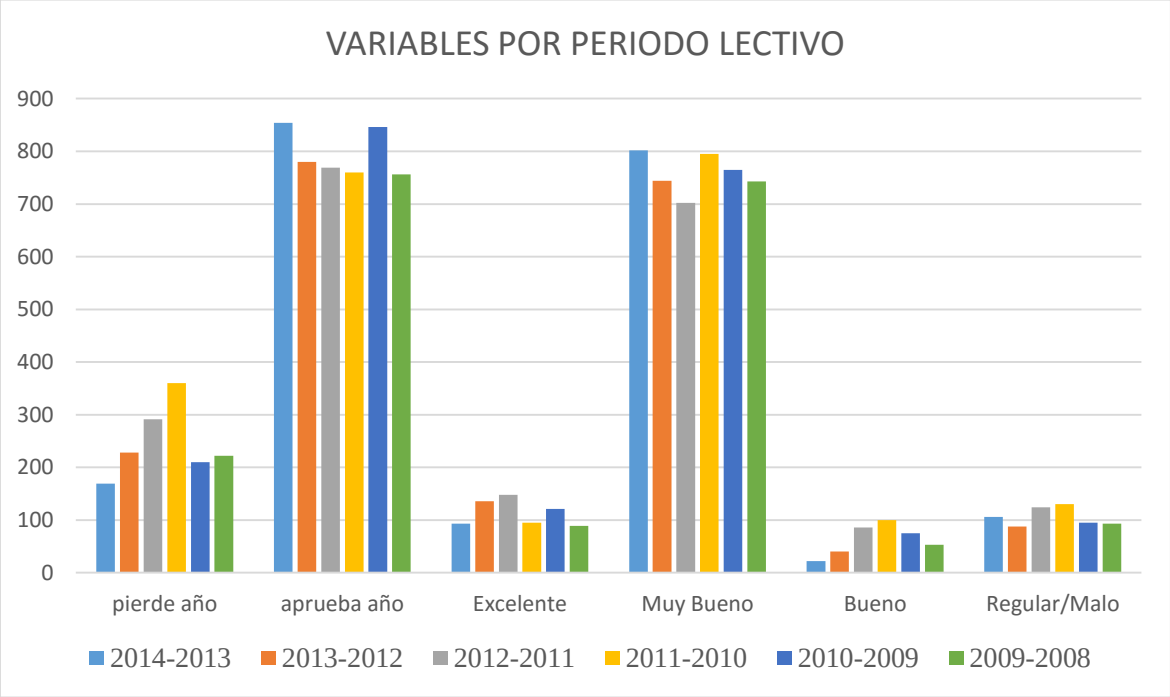


Gráfico 3: Análisis descriptivo de las variables de clase, visualización de patrones.

En el grafico 3 se observa la clasificación de los estudiantes por año lectivo, siendo las mismas establecidas en la cantidad de alumnos que reprueban y aprueban el año lectivo, es decir la promoción del estudiante, mientras que por otro lado muestra el rendimiento de los estudiantes a lo largo de los diferentes años que comprende esta investigación.

El rendimiento académico está catalogado en 4 secciones, como lo son "Excelente", "Muy Bueno", "Bueno" y "Regular/Malo" las cuales están en el rango de calificaciones anuales >=9 hasta 10, >=7 hasta <8.99, >=4,01 hasta 6.99 y por último las calificaciones <=4 respectivamente. A continuación se detalla la Tabla 8 de las mismas.

Tabla 8: Cantidad de registros por periodo lectivo acerca de la promoción y rendimiento.

PERIODO LECTIVO		**2014-2013**	**2013-2012**	**2012-2011**	**2011-2010**	**2010-2009**	**2009-2008**
PROMOCIÓN	Pierde	169	228	291	360	210	222
	Aprueba	854	780	769	760	846	756
RENDIMIENTO ACADÉMICO	Excelente	93	136	148	95	121	89
	MuyBueno	802	744	702	795	765	743
	Bueno	22	40	86	100	75	53
	Regular	106	88	124	130	95	93

Fuente: Base de estudiantes Unidad Educativa 24 de Mayo 2008-2014
Elaborado por el autor

Luego de haber realizado los primeros análisis estadísticos, se muestra como los datos tienen relación entre los distintos periodos, es decir no existen diferencias abismales entre los mismos, también recordando la visualización de *valores atípicos*[20] en los mismos. A continuación se muestra el proceso de minería de datos para la extracción de reglas que caractericen la relación entre un conjunto de variables.

En la Sección 4.1 y 4.1.1 se ha realizado el objetivo específico referente al análisis de forma descriptiva de los datos que conforman la base de datos para la presente investigación, obteniendo la descripción de las variables empleadas en la misma, corroborando que existe un patrón y coherencia de los datos.

[20] En ingles: Outliers

4.2.1. Análisis de partida base de datos con ruido

En esta sección se describen los resultados obtenidos a partir de las herramientas informáticas KELL y WEKA para la obtención de reglas de asociación.

4.2.1.1 Reglas de asociación en KEEL

Una vez conformada las bases de datos *promocion* y *rendimiento,* con las respectivas variables y registros de los alumnos de los diferentes periodos lectivos, se procede a experimentar la misma en la herramienta informática KEEL, con los diferentes algoritmos que cuentan el mismo.

Este análisis permite obtener un punto de partida en la presente investigación, para posteriormente proceder a realizar los pasos del método de extracción de conocimiento de bases de datos y comparar los resultados obtenidos en ambos casos, determinando cual es la mejor forma de obtener variables asociadas mediante reglas.

En la Tabla 9 se puede observar los valores de los parámetros obtenidos en los diferentes algoritmos ejecutados en KEEL, con la base de datos *promoción*.

Tabla 9: Resultados de la ejecución de algoritmos en KEEL mediante los registros de la base de datos *promocion* con valores atípicos.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	#R	MedSop	MedConf	MedLift	MedConv	MedFC	MedNetconf	MedYulesQ	MedAmp	%Reg	Tiempo
Apriori	no genera										
Eclat	no genera										
GENAR	30	0,10	0,87	1,14	1,86	0,45	0,12	0,39	19,00	59,44	0:00:23
GAR	205	0,87	0,95	1,03	10,28	0,40	0,36	0,74	2,11	100	0:10:59
EARMGA	67	0,49	1,00	1,01	Infinity	0,02	0,01	0,00	2,00	100	0:00:51
Alatasetal	genera 0 reglas										0:00:26
MOEA Ghosh	19	0,64	0,89	1,00	Infinity	0,73	0,53	0,62	10,74	100	0:00:48
MODENAR	76	0,13	0,97	1,53	Infinity	0,89	0,34	0,87	12,00	30,08	0:01:01
ARMMGA	1	1,00	1,00	1,00	1,00	-0,01	0,00	0,00	2,00	99,96	0:00:56
MOPNAR	55	0,36	0,92	9,75	Infinity	0,89	0,67	0,99	2,91	100	0:00:45
QAR_CIP_NSGA	76	0,17	0,93	349,45	Infinity	0,92	0,73	0,84	3,14	100	0:00:39

Fuente: Algoritmos ejecutados en KEEL
Elaborado por el autor

En la Tabla 10 se puede observar los valores de los parámetros obtenidos en los diferentes algoritmos ejecutados en KEEL, con la base de datos *rendimiento*.

Tabla 10: Resultados de la ejecución de algoritmos en KEEL mediante los registros de la base de datos *rendimiento* con valores atípicos.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	#R	MedSop	MedConf	MedLift	MedConv	MedFC	MedNetconf	MedYulesQ	MedAmp	%Reg	Tiempo
Apriori	no genera										
Eclat	no genera										
GENAR	30	0,1	0,93	1,15	3,18	0,62	0,13	0,54	19	58,26	0:00:31
GAR	200	0,87	0,95	1,03	10,78	0,42	0,38	0,73	2,09	100	20:47:00
EARMGA	91	0,49	1	1,01	Infinity	0,15	0,01	0	2	100	0:01:02
Alatasetal	1	0,01	1	9,82	Infinity	1	0,91	1	5	0,25	0:00:15
MOEA Ghosh	2	0,6	0,62	0,94	0,98	-0,07	-0,1	-0,2	14,5	100	0:01:09
MODENAR	66	0,62	0,97	1,1	Infinity	0,87	0,28	0,93	11	87,64	0:01:46
ARMMGA	1	1	1	1	1	-0,01	0	0	2	99,92	0:01:01
MOPNAR	47	0,38	0,94	11,7	Infinity	0,91	0,74	1	3,09	100	0:01:05
QAR_CIP_NSGA	123	0,25	0,91	391,77	Infinity	0,89	0,68	0,87	3,39	100	0:00:55

Fuente: Algoritmos ejecutados en KEEL
Elaborado por el autor

En las tablas, los algoritmos como MOEA Ghost y QAR_CIP_NSGA representan un valor cercano a 1 en la medida yule'sQ pero la variable MedAmp posee una cifra mínima, dado que la misma representa el número de atributos involucrados en las reglas.

Mientras otros dan como resultado 0 en la medida yule'sQ, es decir que los mismos no encuentran correlación entre los datos, una de las causas es que la base está en su estado puro, es decir, se encuentra con ruido, desbalanceada, sin pre-procesamiento alguno.

4.2.1.2 Reglas de asociación en WEKA

En WEKA el principal algoritmo para la ejecución de reglas de asociación es el de APRIORI, siendo el primer criterio para establecer las reglas de asociación el parámetro de precisión o confianza, dada por el porcentaje de confianza que se da en cada regla. Se detalla en el Anexo 2, los parámetros por defecto de las medidas que pertenecen al algoritmo WEKA, esto con el fin de conocer la descripción y empleo de cada una.

Una de las ventajas de emplear WEKA en la minería de datos, es que la menciona herramienta cuenta con una interfaz amable con el usuario, inclusive permite obtener gráficos en representación de nuestro atributo clase.

En la Figura 4 se representa los datos para la promoción de los estudiantes y la Figura 5 se representa el rendimiento académico de los estudiantes, se muestra claramente que dichas bases de datos contienen ruido, se encuentran desbalanceadas en cuanto a las variables de clase.

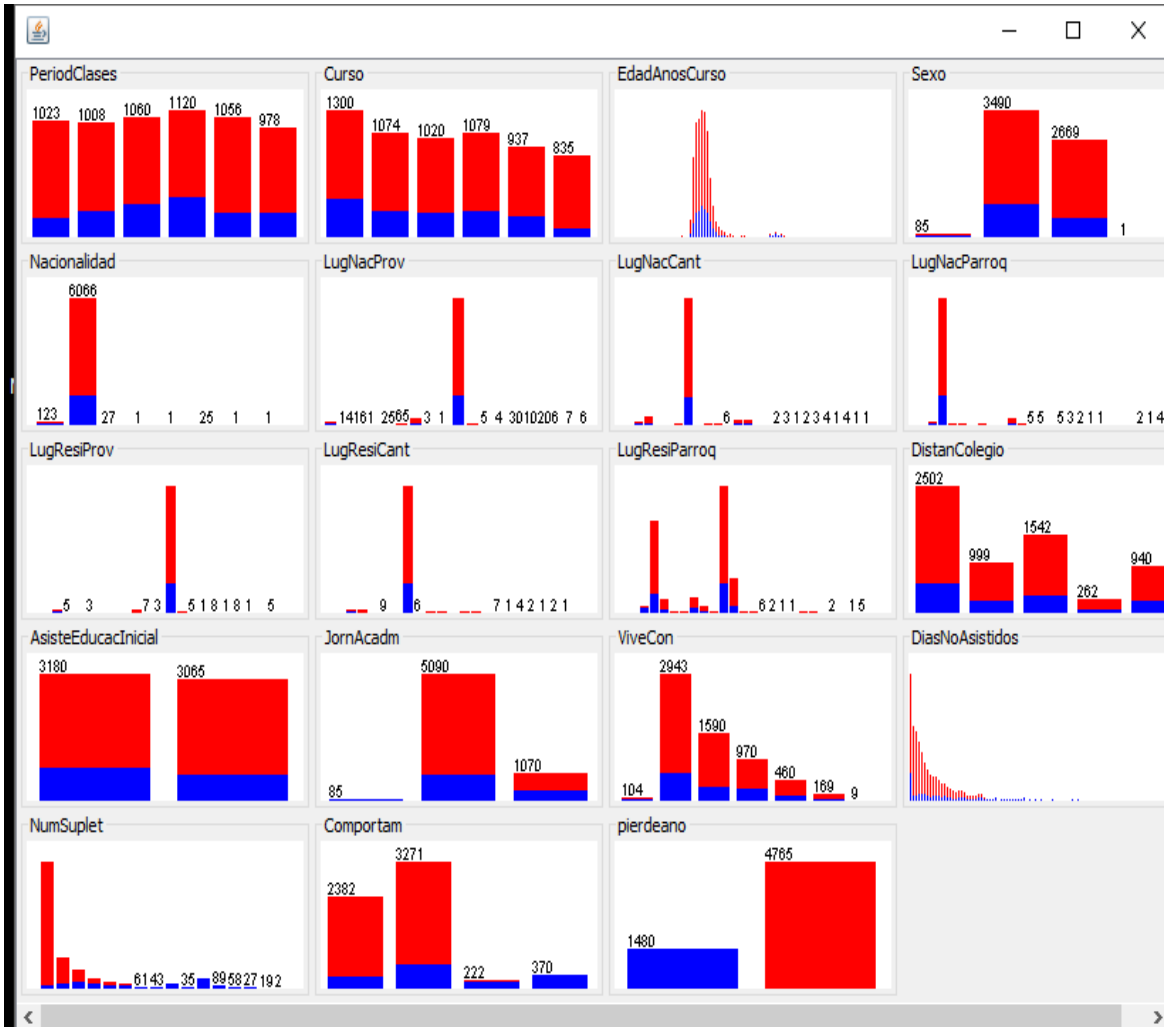


Figura 4: Base de datos *promocion* en relación a la variable de clase *pierdeanio*.

En las figuras mencionadas en el párrafo anterior, se observa una mayoría de estudiantes que no han perdido el año lectivo, mientras que el rendimiento académico no es igual a lo largo de los periodos lectivos, se visualizan valores fuera de rango, recordando que este será el punto de partida para el análisis respectivo de la presente investigación, por ende se

justifica el estado de la base. Una vez concluidos los análisis preliminares en la herramienta informática de KEEL, se procede a realizar estudios similares con WEKA.

Figura 5: Base de datos *rendimiento* en relación a la variable de clase *rendimiento*.

Como se ilustra en la Sección 4.2.1 empleando la herramienta KEEL con la ejecución de algoritmos para la extracción de reglas, se efectúa el mismo proceso en WEKA, obteniendo como resultado un conjunto de reglas por cada base de datos al ejecutar el algoritmo APRIORI. En la Anexo 3 se encontrara la descripción de las medidas que tomaran los parámetros para la generación de las reglas, tanto dependientes de la clase, como las reglas generales de cada base de datos.

La Tabla 11, muestra la cantidad de reglas generadas, la media de confianza y soporte mediante el algoritmo APRIORI en dependencia de las variables de clase *pierdeanio* y *rendimiento*, en tanto que también presenta las asociaciones sin clase de las bases de datos

empleadas en el presente proyecto, mientras que en los Anexos 4 y 5 se da a conocer las primeras 12 reglas generadas en todos los casos descritos.

Tabla 11: Resultados de la ejecución del algoritmo APRIORI en WEKA mediante los registros de la base de datos *promocion* y *rendimiento* con valores atípicos.

DESCRIPCIÓN	Base de datos *promocion*		Base de datos *rendimiento*	
	Asociación con clase	Asociación sin clase	Asociación con clase	Asociación sin clase
Numero de reglas generadas	1000	1000	101	1000
Media de confianza	0,98	0,95	0,96	0,95
Media soporte	1346	3642	712	3646

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Cabe resaltar que la diferencia entre las dos herramientas informáticas, es que en WEKA los resultados obtenidos son una media calculada del número de reglas y los indicadores de confianza y soporte, mientras que en KEEL además de lo mencionado se obtienen parámetros adicionales en las medidas de cada algoritmo ejecutado.

Aunque se observe una confianza muy alta en las reglas obtenidas mostradas en las Tabla 11, siendo las mismas con dependencia de la variable clase o reglas de asociación sin clase, al interpretar las mismas, no se puede asegurar que dichas reglas sean las mejores que caractericen la relación entre un conjunto de variables acerca del aprovechamiento académico de los alumnos.

Para comprobar lo antes mencionado, se procedió aplicar técnicas de pre-procesamiento, para posteriormente implementar algoritmos que nos permitan realizar la selección de variables, cabe mencionar que se realizaron pruebas con las bases de datos con sus registros balanceados y desbalanceados.

4.1.3. Pre-procesamiento de los datos

El pre-procesamiento de datos se emplea en la restructuración de la información que tiene algún problema de valor ausente o inconsistencia como valores fuera de rango. En especial

estos dos problemas antes mencionados afectan en gran medida a la extracción del conocimiento en cualquier base de datos de datos.

Como se menciona en [10] la preparación de datos genera datos de calidad, los cuales pueden conducir a su vez a patrones o reglas de calidad. Los principales motivos que causan esta problemática se originan por error humano a la hora de incorporar la información, ya sea esta por el ingreso de registros, copia de los mismos o trasformación.

En el apartado de descripción de las variables 4.2.2, se realiza una visualización de dichos valores fuera de rangos en las Figuras 4 y 5, conocido en la minería de datos como valores atípicos o ruido, a continuación se empleara la herramienta WEKA para la detección de los mismos en ambas bases de datos para su posterior detección y eliminación, con el fin de poder realizar un nuevo análisis en la investigación.

En la sección de *Preprocess*[21] de WEKA se escoge la opción *Filter*[22] para luego dirigirse a *Unsupervise*[23], escoger *Attribute*[24] y posteriormente *InterquartileRange* la misma herramienta lo detalla cómo, un filtro para la detección de valores atípicos y extremos basados en rangos intercuartil. En el Anexo 6 se detallan las medidas que emplea *InterquartileRange* con su respectiva descripción.

Una vez que se han identificado los valores extremos y valores atípicos, se procede a eliminar dichos registros, para esto se implementa el filtro *RemoveWithValues* el mismo que se encuentra en la sección *Preprocess de WEKA*, donde luego se dirige a *Filter* para seleccionar *Unsupervise*, por último se elige *Instances*[25] y se selecciona el filtro mencionado que como su nombre en inglés lo indica procederá a eliminar los atributos escogidos.

En la Anexo 7 se procederá a describir las medidas por defecto que contiene el filtro *RemoveWithValues*. En la Tabla 12 se detallan las medidas designados para realizar la detección y posterior eliminación de valores extremos y valores atípicos, tanto para la base de *promocion* como de *rendimiento*, mediante los filtros mencionados.

[21] En español: Preprocesado
[22] En español: Filtro
[23] En español: No supervisado
[24] En español: Atributo
[25] En español: Instanceas

Tabla 12: Medidas para la detección y eliminación de valores extremos y valores atípicos.

Medidas del Filtro InterquartileRange		**Medidas del Filtro RemoveWithValues**	
Parámetro	**Medida**	**Parámetro**	**Medida**
AttributeIndices	1-5,13,15	AttributeIndices	19-20
Debug	False	DontFilterAfterFirstBatch	False
DetectionperAttribute	False	InvertSelection	False
Extremevaluesasoutlier	False	MatchMissingValues	False
Extremevaluesfactor	10.0	ModifyHeader	False
Outlierfactor	1.0	NominalIndices	last
Outputoffsetmultiplier	False	SplitPoint	0.0

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Una vez aplicado los filtros correspondientes, se procede a la detección de valores atípicos y valores extremos en ambas bases de datos, representando una cantidad de 100 registros valores atípicos y 218 Valores Extremos.

Como se lo aprecia en la Figura 6 una vez eliminado dichos registros, se ha dado por terminado el apartado de pre-procesamiento de los datos, dado que al obtener variables numéricas y en rangos antes definidos, no es necesario aplicar otras técnicas de pre-procesado, como lo son la Discretización o Normalización.

Base de datos *promoción* **Base de datos *rendimiento***

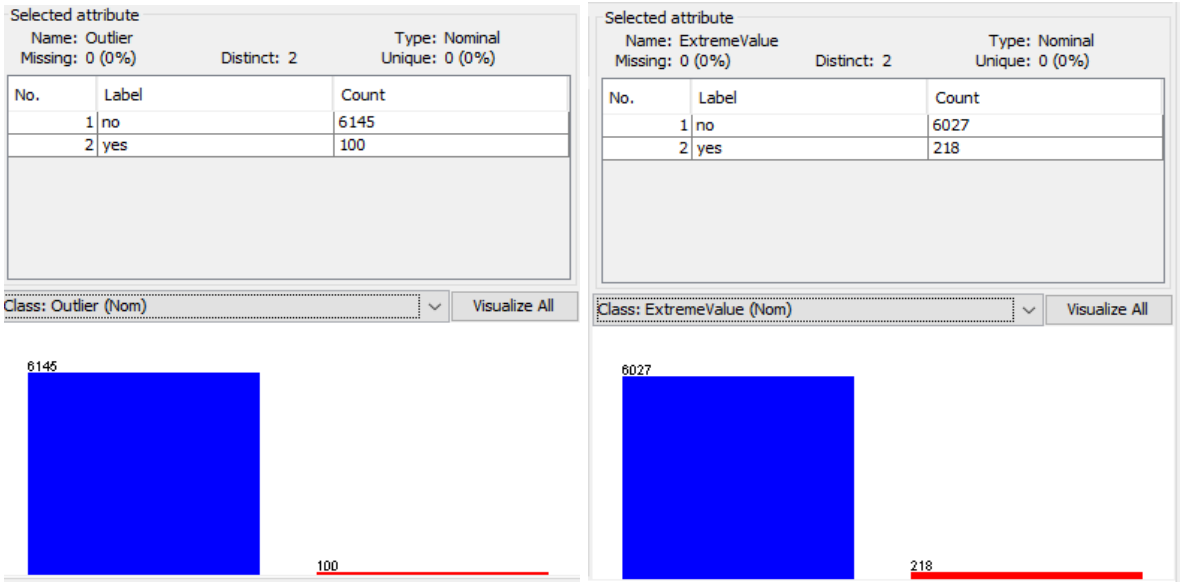


Figura 6: Detección de valores atípicos y fuera de rango.

4.1.4. Selección de variables

La selección de atributos o de variables permite descartar información redundante que no conlleva a representar un mayor conocimiento en el proceso de extracción de conocimiento, permitiendo que los algoritmos se ejecuten de forma más rápida y precisa.

De acuerdo a [12] el objetivo de los métodos de selección de variables es buscar un modelo que se ajuste bien a los datos y que a la vez sea posible buscar un equilibrio entre bondad de ajuste y sencillez. Al implementar esta técnica en la obtención de reglas, se reduce gradualmente el número obtenidas de las mismas, dado que se emplea menos atributos para su asociación.

El presente proyecto investigativo cuenta con un número de 19 variables en cada base de datos como lo son *promocion* y *rendimiento*, se aplicaran una serie de algoritmos que cuenta la herramienta WEKA para determinar cuáles de los atributos no está introduciendo ningún tipo de información relevante para el proceso de extracción de conocimiento con el objetivo de obtener reglas más concisas al momento de aplicar los algoritmos correspondientes.

Con el fin de aplicar la técnica de selección de atributos, en WEKA se emplearon algoritmos de tipo filtro, específicamente el algoritmo *Consistency Subset Evaluator* como método de evaluación.

Este algoritmo se basa en evaluar el valor de un subconjunto de atributos por el nivel de consistencia en los valores de la clase, eliminando atributos que tienen una correlación muy alta como atributos redundantes. En las Figuras 7 y 8 se muestra la configuración del mencionado algoritmo en la herramienta de WEKA, con las variables obtenidas en cada base de datos.

Al disminuir el número de variables como se indica en [27], existe un cierto riesgo de pérdida de información, por ese motivo las variables deben seleccionarse cuidadosamente, por ello la aplicación de algoritmos mediante la herramienta informática WEKA.

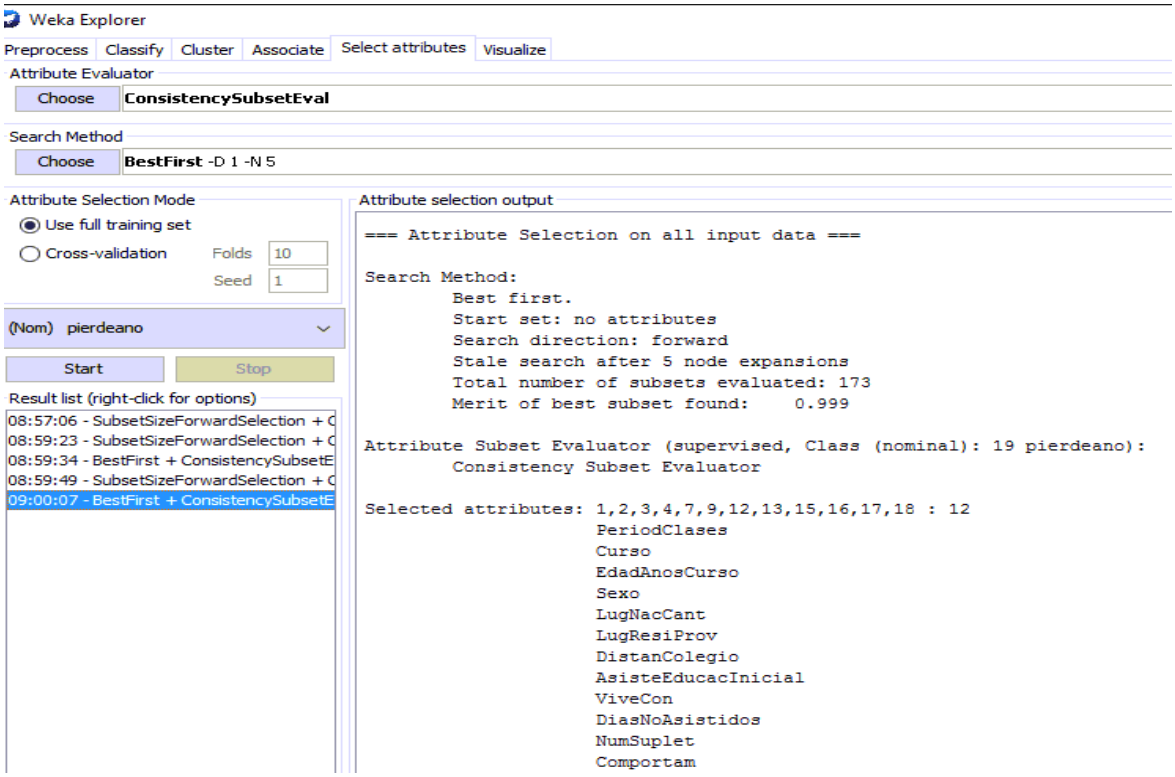


Figura 7: Selección de variables mediante WEKA en la base de datos *promocion* con la variable clase *pierdeanio*.

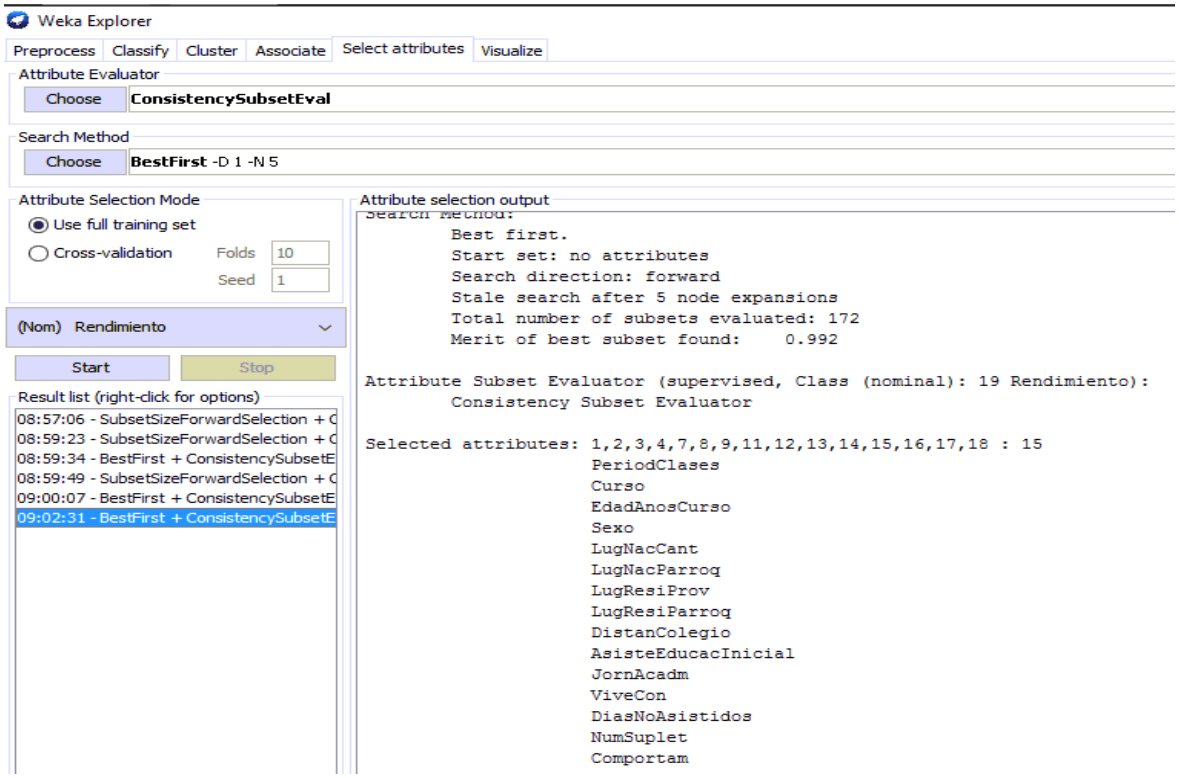


Figura 8: Selección de variables mediante WEKA en la base de datos *rendimiento* con la variable clase *rendimiento*.

A continuación se detalla en la Tabla 13 dos algoritmos más con sus respectivas descripciones, numero de variables y el nombre de las mismas.

Tabla 13: Selección de variables en las distintas bases de datos

Base de datos *promocion*			**Base de datos *rendimiento***		
Algoritmo	**#**	**Variables**	**Algoritmo**	**#**	**Variables**
Cfssubseteval	2	NumSuplet, Comportam	CfsSubsetEval	2	NumSuplet, Comportam
Filteredsubseteval	2	NumSuplet, Comportam	FilteredSubsetEval	2	NumSuplet, Comportam
Consistency subset evaluator	11	Curso, EdadAnosCurso, Sexo, LugNacCant, LugResiProv, DistanColegio, AsisteEducacInicial, ViveCon, diasNoAsistidos, NumSuplet, Comportam	Consistency Subset Evaluator	14	Curso, EdadAnosCurso, Sexo, LugNacCant, LugNacParroq, LugResiProv, LugResiParroq, DistanColegio, AsisteEducacInicial, JornAcadm, ViveCon, DiasNoAsistidos, NumSuplet, Comportam

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

4.1.5. Análisis de desbalance de clases

Uno de los problemas frecuentes que se presenta en muchas aplicaciones de aprendizaje automático ocurre al momento de contar con la variable de clase con registros desbalanceados, cuyos efectos sobre el desempeño de los clasificadores son notables, o como en nuestro caso la creación de reglas mediante asociaciones.

El problema ocurre cuando el número de instancias de la variable clase poseen diferentes cantidades en sus registros, provocando en los clasificadores un sobre aprendizaje acerca de la instancia predominante y por otro lado el poco aprendizaje de las demás instancias.

En este sentido las bases de datos *promocion* y *rendimiento* cuentan cada una con una clase diferente, representando la aprobación o reprobación del año lectivo, y el rendimiento académico del estudiante respectivamente.

Como lo muestra la Figura 9, se puede apreciar que aparece una dominancia de la instancia aprobación con un total de 4563 casos mientras que reprobación cuenta con 1366, también se da una representación alta en rendimiento medio, con un total de 4373, en alto 637, bueno 361 y regular 556 casos.

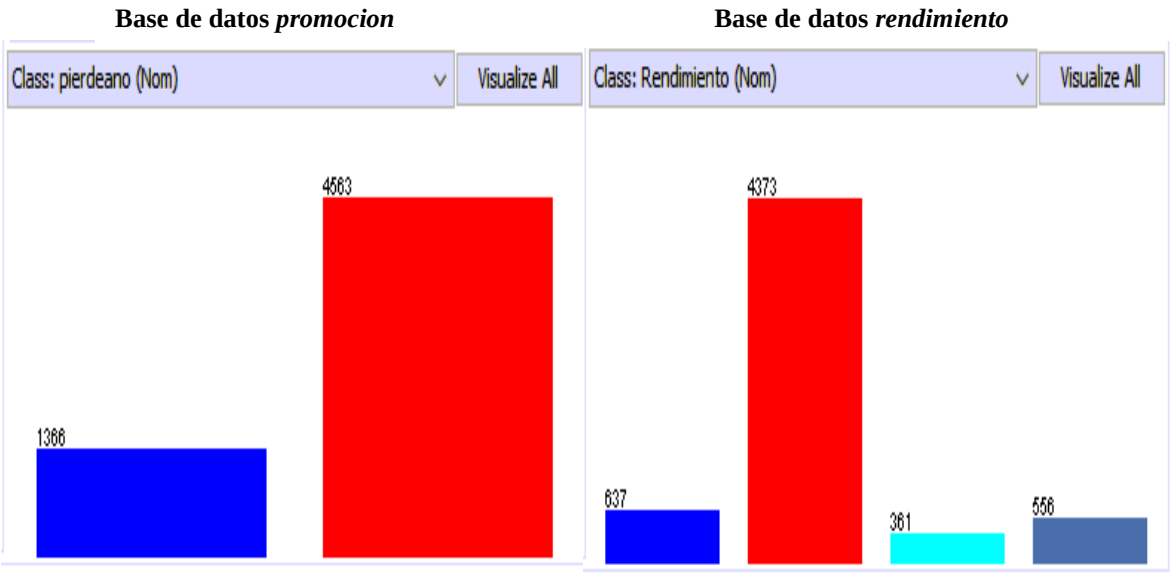


Figura 9: Desbalanceo de las variables de clase, promoción y rendimiento

Con el fin de poder balancear las bases de datos empleadas en la presente investigación, se puede incurrir en la estrategia de sobre muestreo, la misma consiste en incorporar instancias de la clase minorista para emparejar los casos, es decir implementar registros en la variable de clase que se encuentre desbalanceada, existen dos métodos fundamentales:

- Sobre Muestreo Aleatorio: se generan instancias aleatorias hasta completar la clase minoritaria.

- Sobre Muestreo Aleatorio Centrado: al igual que el anterior se generan casos de forma aleatoria pero siguiendo un patrón de los ejemplos existentes.

En esta investigación se utiliza el algoritmo SMOTE (Synthetic Minority Oversampling Technique), [28] sigue una estrategia de sobre muestreo donde las nuevas instancias se crean a través de la interpolación de los casos más cercanos a la clase minoritaria. En la Figura 10 se representa las clases *pierdeanio* y *rendimiento* ya balanceadas para su posterior estudio.

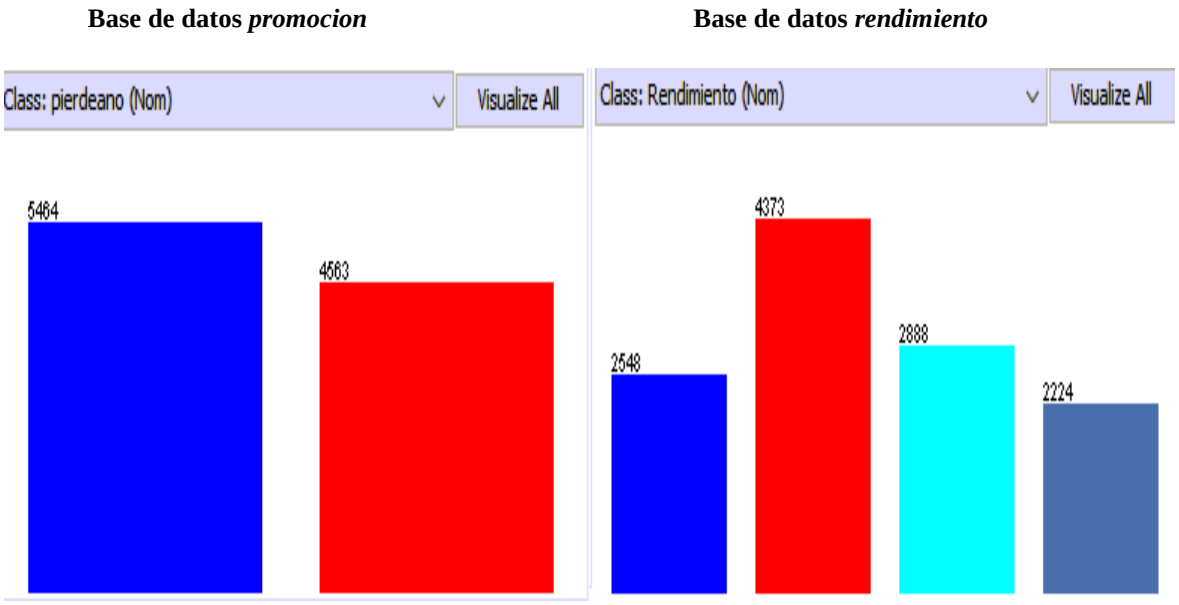


Figura 10: Desbalanceo de las variables de clase, promoción y rendimiento

4.1.6. Extracción de conocimiento

Una vez que se ha realizado el pre-procesamiento de los datos, la selección de variables y el balanceo de las bases de datos, se procede a realizar la ejecución de los algoritmos para reglas de asociación implementados en la herramienta de WEKA y KEEL.

Se implementaron dos estudios, en ambos se tomó en cuenta el pre-procesamiento de los datos y la selección de variables, pero cambiará el empleo de la base de datos, dado que en una se lo realizo con los registros balanceados y en otra con los registros desbalanceados. Esto se hace con el fin de poder observar las diferencias en ambos casos, dado que al balancear las bases de datos se están introduciendo registros aleatorios.

En las Tablas 14 y 15 se puede observar los valores de los parámetros obtenidos en los diferentes algoritmos ejecutados en KEEL, con la base de datos *promocion* con sus registros balanceados y desbalanceados.

Tabla 14: Resultados de la ejecución de algoritmos en KEEL en la base de datos *promocion* con los registros de la variable clase desbalanceados.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	**#R**	**MedSop**	**MedConf**	**MedLift**	**MedConv**	**MedFC**	**MedNetconf**	**MedYulesQ**	**MedAmp**	**%Reg**	**Tiempo**
Apriori	11	0,62	0,92	1,26	4,85	0,39	0,39	0,41	2,37	97,14	0:00:01
Eclat	11	0,62	0,92	1,26	4,85	0,39	0,39	0,41	2,37	97,14	0:00:01
GENAR	3	0,01	1,00	1,30	Infinity	1,00	0,24	0,00	13,00	0,14	0:00:03
GAR	86	0,70	0,91	1,03	1,71	0,10	0,03	-0,05	2,03	99,92	0:08:24
EARMGA	92	0,20	1,00	1,00	1,00	0,00	0,00	0,00	2,00	100,00	0:00:24
Alatasetal	89	0,07	1,00	1,08	Infinity	0,92	0,08	0,66	6,86	70,56	0:00:17
MOEA Ghosh	19	0,55	0,75	1,04	Infinity	0,33	0,03	0,21	5,00	100,00	0:00:14
MODENAR	12	0,02	0,44	3,15	Infinity	0,36	0,19	0,41	6,17	19,25	0:00:11
ARMMGA	1	0,16	1,00	1,22	Infinity	1,00	0,21	1,00	2,00	15,79	0:00:16
MOPNAR	38	0,32	0,89	4,47	Infinity	0,85	0,57	0,97	3,16	99,67	0:00:18
QAR_CIP_NSGA	44	0,07	0,79	149,87	Infinity	0,76	0,45	0,62	3,03	93,63	0:00:15

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

Tabla 15: Resultados de la ejecución de algoritmos en KEEL en la base de datos *rendimiento* con los registros de la variable clase desbalanceados.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	**#R**	**MedSop**	**MedConf**	**MedLift**	**MedConv**	**MedFC**	**MedNetconf**	**MedYulesQ**	**MedAmp**	**%Reg**	**Tiempo**
Apriori	10	0,61	0,95	1,29	4,44	0,55	0,48	0,61	2,21	97,82	0:00:01
Eclat	10	0,61	0,95	1,29	4,44	0,55	0,48	0,61	2,21	97,82	0:00:02
GENAR	30	0,01	1,00	1,82	Infinity	0,98	0,45	0,01	13,00	4,70	0:00:11
GAR	80	0,72	0,94	1,01	1,08	0,07	-0,01	-0,06	2,11	99,92	0:35:30
EARMGA	100	0,54	1,00	1,00	1,00	0,00	0,00	0,00	2,00	100,00	0:02:01
Alatasetal	94	0,01	1,00	1,05	Infinity	0,74	0,05	0,32	6,99	93,01	0:00:30
MOEA Ghosh	23	0,57	0,81	1,04	Infinity	0,30	0,01	-0,02	4,83	100,00	0:01:09
MODENAR	11	0,05	0,46	2,10	Infinity	0,31	0,26	0,46	5,46	32,63	0:00:50
ARMMGA	1	0,60	0,97	1,01	1,11	0,10	0,01	0,13	2,00	59,66	0:01:10
MOPNAR	40	0,37	0,89	3,00	Infinity	0,83	0,61	0,95	3,05	100,00	0:01:06
QAR_CIP_NSGA	43	0,06	0,81	574,01	Infinity	0,80	0,62	0,78	3,52	85,80	0:00:54

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

En las Tablas 16 y 17 se puede observar los valores de los parámetros obtenidos en los diferentes algoritmos ejecutados en KEEL, con la base de datos *rendimiento* en su estado balanceado y desbalanceado.

Tabla 16: Resultados de la ejecución de algoritmos en KEEL en la base de datos *rendimiento* con los registros de la variable clase desbalanceados.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	**#R**	**MedSop**	**MedConf**	**MedLift**	**MedConv**	**MedFC**	**MedNetconf**	**MedYulesQ**	**MedAmp**	**%Reg**	**Tiempo**
Apriori	56	0,64	0,89	1,15	2,17	0,38	0,37	0,54	2,90	99,82	0:00:01
Eclat	56	0,64	0,89	1,15	2,17	0,38	0,37	0,54	2,90	99,82	0:00:15
GENAR	2	0,01	0,88	1,19	2,10	0,53	0,14	0,00	16,00	0,14	0:00:04
GAR	172	0,76	0,88	1,01	1,05	0,05	0,03	0,06	2,02	100,00	0:09:15
EARMGA	96	0,26	1,00	1,01	Infinity	0,11	0,01	0,00	2,00	100,00	0:00:32
Alatasetal	NO GENERA										
MOEA Ghosh	26	0,61	0,89	1,07	Infinity	0,36	0,03	0,15	5,50	99,97	0:00:18
MODENAR	11	0,06	0,65	5,17	Infinity	0,46	0,13	0,25	8,00	35,77	0:00:11
ARMMGA	1	0,84	0,94	1,00	0,93	-0,01	-0,05	-0,57	2,00	83,27	0:00:26
MOPNAR	29	0,34	0,90	4,53	Infinity	0,88	0,67	0,99	2,87	100,00	0:00:14
QAR_CIP_NSGA	41	0,08	0,87	143,43	Infinity	0,85	0,68	0,94	3,49	91,18	0:00:11

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

Tabla 17: Resultados de la ejecución de algoritmos en KEEL en la base de datos *rendimiento* con los registros de la variable clase balanceados.

	número medio de reglas	valores medios de soporte	valores medios de confianza	valores medios de lift	valores medios de conviction	valores medios de factor certeza	valores medios de netconf	valores medios de yule'sQ,	# medio de atributos involucrados en las reglas	% de registros de la BD cubiertos por las reglas generadas.	tiempo de ejecucion del algoritmo
Algoritmos	**#R**	**MedSop**	**MedConf**	**MedLift**	**MedConv**	**MedFC**	**MedNetconf**	**MedYulesQ**	**MedAmp**	**%Reg**	**Tiempo**
Apriori	53	0,66	0,93	1,22	3,94	0,48	0,47	0,65	2,76	99,86	0:00:01
Eclat	53	0,66	0,93	1,22	3,94	0,48	0,47	0,65	2,76	99,86	0:00:01
GENAR	30	0,01	1,00	4,40	Infinity	1,00	0,77	0,94	16,00	5,56	0:00:11
GAR	151	0,80	0,92	1,01	1,04	0,04	0,02	-0,01	2,14	100,00	0:23:50
EARMGA	75	0,19	1,00	1,01	Infinity	0,09	0,01	0,00	2,00	100,00	0:02:18
Alatasetal	80										
MOEA Ghosh	25	0,54	0,85	1,06	Infinity	0,44	0,10	0,17	6,56	100,00	0:01:15
MODENAR	16	0,03	0,62	8,55	Infinity	0,51	0,32	0,54	9,57	26,22	0:01:00
ARMMGA	1	0,25	0,98	1,01	1,26	0,21	0,01	0,16	2,00	25,00	0:02:17
MOPNAR	55	0,25	0,80	4,95	Infinity	0,76	0,63	0,98	3,62	97,65	0:01:12
QAR_CIP_NSGA	60	0,09	0,89	445,84	Infinity	0,88	0,69	0,83	3,57	91,45	0:01:01

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

Posterior a los análisis realizados en KEEL, se utiliza el algoritmo APRIORI en WEKA, con los parámetros que se han implementado a lo largo de la investigación, la Tabla 18 describe las reglas obtenidas luego de emplear técnicas de pre-procesamiento de datos y

selección de variables en la base de datos *promocion* con sus registros desbalanceados y balanceados.

Tabla 18: Resultados de la ejecución del algoritmo APRIORI en WEKA de la base *promocion* con los registros de su variable clase balanceados y desbalanceados.

	DESBALANCEADOS		BALANCEADOS	
DESCRIPCIÓN	**Asociación con clase**	**Asociación sin clase**	**Asociación con clase**	**Asociación sin clase**
NUMERO DE REGLAS GENERADAS	123	760	27	700
MEDIA DE CONFIANZA	0,97	0,97	0,95	0,96
MEDIA SOPORTE	990	966	1305	1608

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Por otra parte en la Tabla 19 se muestran los resultados de la base de datos *rendimiento* con sus registros desbalanceados y balanceados.

Tabla 19: Resultados de la ejecución del algoritmo APRIORI en WEKA de la base *rendimiento* con los registros de su variable clase balanceados y desbalanceados.

	DESBALANCEADOS		BALANCEADOS	
DESCRIPCIÓN	**Asociación con clase**	**Asociación sin clase**	**Asociación con clase**	**Asociación sin clase**
NUMERO DE REGLAS GENERADAS	16	1000	10,00	1000
MEDIA DE CONFIANZA	0,98	0,99	0,71	0,96
MEDIA SOPORTE	697	1187	1274	3459

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Una vez terminada la sección de pre-procesamiento de la información, la selección de variables, del balanceo de las bases de datos y el emplear varios algoritmos para la obtención de reglas de asociación, se cumple con el segundo objetivo específico del presente proyecto investigativo, el mismo que se enfocaba en la pre-procesamiento de datos, para obtener bases de datos consistente y emplear diferentes modelos de reglas de asociación basados en el criterio de soporte y confianza.

4.1.7. Interpretación y Evaluación

A continuación en las Tablas 20 y 21 se detallan los parámetros de soporte y confianza obtenida en la base de datos *promocion* y *rendimiento* en la herramienta informática KEEL, a lo largo del presente proyecto investigativo, para su posterior análisis.

Tabla 20: Comparación de los resultados de la ejecución de algoritmos sobre la base de datos *promocion* en KEEL.

ALGORITMO	**Con Valores Atípicos**		**Registros desbalanceados**		**Registros balanceados**	
	Confianza	**Soporte**	**Confianza**	**Soporte**	**Confianza**	**Soporte**
APRIORI	no genera		0,62	0,92	0,61	0,95
ECLAT	no genera		0,62	0,92	0,61	0,95
GENAR	0,10	0,87	0,01	1,00	0,01	1,00
GAR	0,87	0,95	0,70	0,91	0,72	0,94
EARMGA	0,49	1,00	0,20	1,00	0,54	1,00
ALATASETAL	no genera		0,07	1,00	0,01	1,00
MOEA GHOSH	0,64	0,89	0,55	0,75	0,57	0,81
MODENAR	0,13	0,97	0,02	0,44	0,05	0,46
ARMMGA	1,00	1,00	0,16	1,00	0,60	0,97
MOPNAR	0,36	0,92	0,32	0,89	0,37	0,89
QAR_CIP_NSGA	0,17	0,93	0,07	0,79	0,06	0,81

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

Tabla 21: Comparación de los resultados de la ejecución de algoritmos sobre la base de datos *rendimiento* en KEEL

ALGORITMO	**Con Valores Atípicos**		**Registros desbalanceados**		**Registros balanceados**	
	Confianza	Soporte	Confianza	Soporte	Confianza	Soporte
APRIORI	no genera		0,64	0,89	0,66	0,93
ECLAT	no genera		0,64	0,89	0,66	0,93
GENAR	0,10	0,93	0,01	0,88	0,01	1,00
GAR	0,87	0,95	0,80	0,92	0,72	0,88
EARMGA	0,49	1,00	0,26	1,00	0,19	1,00
ALATASETAL	0,01	1,00	no genera		0,02	1,00
MOEA GHOSH	0,60	0,62	0,61	0,89	0,54	0,85
MODENAR	0,62	0,97	0,06	0,65	0,03	0,62
ARMMGA	1,00	1,00	0,84	0,94	0,25	0,98

MOPNAR	0,38	0,94	0,34	0,90	0,25	0,80
QAR_CIP_NSGA	0,25	0,91	0,08	0,87	0,09	0,89

Fuente: Algoritmos ejecutado en KEEL
Elaborado por el autor

En las Tablas 22 y 23 se detallan los parámetros de soporte y confianza obtenida en las bases de datos *promocion* y *rendimiento* en la herramienta informática WEKA, a lo largo del presente proyecto investigativo, para su posterior análisis.

Tabla 22: Comparación sobre la base de datos *promocion* con la ejecución del algoritmo APRIORI en WEKA

Descripción	**CON VALORES ATÍPICOS**		**REGISTROS DESBALANCEADOS**		**REGISTROS BALANCEADOS**	
	Asociación con clase	Asociación sin clase	Asociación con clase	Asociación sin clase	Asociación con clase	Asociación sin clase
Media de Confianza	0,78	0,75	0,97	0,97	0,98	0,99
Media Soporte	1346	3642	990	966	697	1187

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Tabla 23: Comparación sobre la base de datos *rendimiento* con la ejecución del algoritmo APRIORI en WEKA

Descripción	**CON VALORES ATÍPICOS**		**REGISTROS DESBALANCEADOS**		**REGISTROS BALANCEADOS**	
	Asociación con clase	Asociación sin clase	Asociación con clase	Asociación sin clase	Asociación con clase	Asociación sin clase
Media de Confianza	0,76	0,75	0,95	0,96	0,71	0,96
Media Soporte	712	3646	1305	1608	1274	3459

Fuente: Algoritmos ejecutado en WEKA
Elaborado por el autor

Para la evaluación de las reglas se tomó en cuenta el criterio de los expertos en educación como lo son la Lcda. Guisella Chabla Galarza, MSc.[26] sumado al Psic. Darwin Coello Goyburo[27] quienes juntos suman más de 30 años en el ámbito educativo secundario.

A continuación se realiza el análisis de las reglas obtenidas en la herramienta informática KEEL, mediante el algoritmo que obtuvo el resultado más óptimo en cuanto a los índices de medición de soporte y confianza.

El análisis referente a la base de datos *promocion* mediante la herramienta KEEL, como se observa en la Tabla 20, el algoritmo GAR obtuvo mejores resultados de confianza y soporte con la base de datos con sus registros balanceados, por ende en el Anexo 8 se encuentran las reglas generadas por dicho algoritmo, el mismo conformo reglas de asociación en su mayoría con 2 variables.

Las reglas generadas al estar compuestas por dos variables no aportan mayor información según la opinión de los expertos, como por ejemplo las reglas 2 y 5 es información que ellos no consideran como un nuevo conocimiento.

A continuación se realiza el análisis de las reglas obtenidas en KEEL mediante la base de datos *rendimiento* con sus registros desbalanceados por medio del algoritmo GAR dado que el mismo obtuvo los mejores resultados incluso sobre la base de datos con registros balanceados.

La regla 0 no aporta mayor información, sin embargo la regla 17 que obtiene como resultado que los estudiantes que tengan un rendimiento medio estarán en la jornada matutina con una confianza del 88% les parece interesante a los expertos en educación.

Una vez concluido el estudio de las reglas obtenidas en KEEL por medio de la base *promocion* y *rendimiento* se procede a realizar el mismo análisis con las reglas obtenidas en la herramienta informática de WEKA, complementado dicho estudio una vez más con la evaluación de las reglas obtenidas por medio del criterio de los expertos en educación.

[26] Licenciada en Ciencias de la Educación en la Especialidad de Ingles, Magister en Educación a Distancia y Abierta
[27] Psicólogo Educativo y Orientador Vocacional

El análisis de la base de datos *promocion* como se observa en la Tabla 22, obtuvo un mejor resultado en sus índices de medición como lo son el soporte y confianza en la base de datos con los registros balanceados, por ende se exponen las reglas obtenidas en el Anexo 10.

Los expertos en educación hacen referencia a la regla 7 la misma que al interpretarla da como resultado que un comportamiento "malo" trae como consecuencia la pérdida del año lectivo, así mismo una cantidad de 10 supletorios en los estudiantes obtiene la misma consecuencia que la primera regla menciona.

Otra de las reglas que les interesaron fue la 13, la misma que menciona que los estudiantes de sexo femenino y con un comportamiento excelente, tienden aprobar el año lectivo igual a la regla 25 que hace referencia a los estudiantes de sexo masculino que tengan un comportamiento muy bueno.

A continuación se detalla el análisis de la base de datos *rendimiento* como se observa en la Tabla 23, obtuvo un mejor resultado en sus índices de medición como lo son el soporte y confianza en la base de datos con sus registros desbalanceados, por ende se exponen las reglas obtenidas en el Anexo 11.

En cuanto al rendimiento académico la regla número 4 y la 10 les interesó a los expertos, la misma detalla que los estudiantes de jornada matutina y que tengan un supletorio, contarán con un rendimiento académico entre >=7 y <=8.99, es decir bueno, dado que el tener un supletorio no descarta que el estudiante tenga buenas notas.

Al final de la Sección 4.7 se ha desarrollado el último objetivo específico del presente proyecto investigativo, como lo es el interpretar las reglas de asociación obtenidas, para poder transmitir el conocimiento aportado por las mismas, habiendo cumplido con los tres objetivos específicos planteados en la presente investigación.

4.2. Discusión

Esta investigación ha obtenido como resultado la extracción 6 reglas mediante la evaluación de los expertos en educación en cuanto a la promoción del año lectivo y 4 sobre el rendimiento académico.

Las variables que se impusieron para caracterizar la promoción del año lectivo fueron las de: Curso, EdadAnosCurso, Sexo, LugNacCant, LugResiProv, DistanColegio, AsisteEducacInicial, ViveCon, diasNoAsistidos, NumSuplet, Comportam, mientras que las que variables se relacionan con el rendimiento académico de los estudiantes son: Curso, EdadAnosCurso, Sexo, LugNacCant, LugNacParroq, LugResiProv, LugResiParroq, DistanColegio, AsisteEducacInicial, JornAcadm, ViveCon, DiasNoAsistidos, NumSuplet, Comportam.

Se detallan las reglas elegidas por los expertos en educación para formar un modelo basado en reglas de asociación que caracterice el aprovechamiento académico de los estudiantes de la Unidad Educativa 24 de Mayo:

Sobre la promoción de los estudiantes un total de 6 reglas:

- LugNacCant=5 Comportam=4 1161 ==> pierdeanio=1 1161 conf:(1)
  Los estudiantes que nacieron en la ciudad de Quevedo, y cuentan con un comportamiento muy malo, obtendrán como resultado la pérdida del año lectivo con un total de 1161 casos registrados en la base de datos, obteniendo una confianza del 100% en esta regla.

- NumSuplet=10 1155 ==> pierdeanio=1 1155 conf:(1)
  Los estudiantes que cuenten con una cantidad de diez supletorios en el año lectivo, obtendrán como resultado la pérdida del año lectivo con un total de 1155 casos registrados en la base de datos, obteniendo una confianza del 100% en esta regla.

- Comportam=4 1250 ==> pierdeanio=1 1249 conf:(1)
  Los estudiantes que tengan un comportamiento muy malo, obtendrán como resultado la pérdida del año lectivo con un total de 1249 casos registrados en la base de datos, obteniendo una confianza del 100% en esta regla.

- Sexo=2 LugResiProv=12 NumSuplet=0 Comportam=1 1175 ==> pierdeanio=2 1108 conf:(0.94)
  Los estudiantes de sexo femenino, que residan en la provincia de Los Ríos, que no tengan ningún supletorio y cuenten con un comportamiento excelente, obtendrán

como resultado la aprobación del año lectivo con un total de 1108 casos registrados en la base de datos, obteniendo una confianza del 94% en esta regla.

- NumSuplet=0 Comportam=2 1699 ==> pierdeanio=2 1566 conf:(0.92)
  Los estudiantes que no tengan ningún supletorio, y cuenten con un comportamiento muy bueno obtendrán como resultado la aprobación del año lectivo con un total de 1566 casos registrados en la base de datos, obteniendo una confianza del 92% en esta regla.

- Sexo=1 NumSuplet=0 Comportam=2 1197 ==> pierdeanio=2 1085 conf:(0.91)
  Los estudiantes de sexo masculino que no tengan ningún supletorio en el año lectivo y cuenten con un comportamiento muy bueno, obtendrán como resultado la aprobación del año lectivo con un total de 1085 casos registrados en la base de datos, obteniendo una confianza del 91% en esta regla.

Sobre el rendimiento académico un total de 4 reglas:

- JornAcadm=1 NumSuplet=1 718 ==> Rendimiento=2 714 conf:(0.99)
  Los estudiantes que asistan a clases en la sección matutina y aunque tengan un supletorio, contaran con un rendimiento académico de mayor o igual a 7 y menor o igual a 8.99 con un total de 714 casos registrados y una confianza del 99%.

- NumSuplet=1 832 ==> Rendimiento=2 824 conf:(0.99)
  Los estudiantes que cuenten con un supletorio, obtendrán un rendimiento académico de mayor o igual a 7 y menor o igual a 8.99, con un total de 824 casos registrados y una confianza del 99%.

- Sexo=1 LugResiProv=12 JornAcadm=1 NumSuplet=0 Comportam=2 938 ==> Rendimiento=2 847 conf:(0.9)
  Los estudiantes de sexo masculino que residan en la provincia de Los Ríos, estén asistiendo a clases en la sección matutina, no cuenten con ningún supletorio y tengan un comportamiento de muy bueno, obtendrán un rendimiento académico de mayor o igual a 7 y menor o igual a 8.99, con un total de 847 casos registrados y una confianza del 90%.

- Sexo=1 JornAcadm=1 NumSuplet=0 Comportam=2 736 ==> Rendimiento=1 663 conf:(0.9)

  Los estudiantes de sexo masculino que estén asistiendo a clases en la sección matutina, no cuenten con ningún supletorio y tengan un comportamiento de muy bueno, obtendrán un rendimiento académico de mayor o igual a 9 y menor o igual a 10, con un total de 663 casos registrados y una confianza del 90%.

En la literatura se dan a conocer investigaciones realizadas mediante la minería de datos en el ámbito educativo, siendo relevante que las mismos se dan en otros países, en el presente proyecto investigativo se dan a conocer cuáles son la variables que influyen tanto en la promoción del año lectivo como en el rendimiento académico.

En la investigación de [22] la predicción del fracaso escolar mediante técnicas de minería de datos, las variables de mayor influencia empleando el mismo clasificador que el usado en la presente investigación, son las de: humanidades, nivel de motivación, fumas, edad, sexo y discapacidad física, comportamiento.

Por otra parte en la investigación de [23] metodología de estudio del rendimiento académico mediante la minería de datos, las variables de mayor relevancia son: sexo, nivel de educación de los padres, situación laboral del estudiante, edad, lugar de residencia, actitud general hacia el estudio y vive con.

El presente proyecto guarda relación en cuanto a las variables seleccionadas por las dos investigaciones antes mencionadas, dado que existen variables como edad, sexo, comportamiento, vive con y lugar de residencia del estudiante, que si bien no todas se encuentran en una misma investigación, permite dar a conocer la importancia de dichas variables seleccionadas en otras investigaciones realizadas.

Las reglas de asociación obtenidas presentan información que puede ser discutida en el ámbito académico por los especialistas en la materia, pero cabe resaltar que la confianza obtenida en las mismas reglas provee un soporte para demostrar la calidad de las mismas, es decir permite tener fundamentos en la elaboración del modelo que caracterice el aprovechamiento académico en los estudiantes.

CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

El presente proyecto investigativo se lo realizó para encontrar un modelo basado en reglas que caractericen el aprovechamiento académico en la Unidad Educativa 24 de Mayo en dependencia de 20 variables que contienen información de los estudiantes, en base a la experiencia obtenida al finalizar esta investigación se puede concluir lo siguiente**:**

- Se logró comprobar que existe relación y patrones entre los datos mediante análisis descriptivos realizados en la Sección 4.1 y 4.1.1 por medio de los Gráficos 1 y 3, donde se observa que se cumple con la lógica de contar con más registros de estudiantes aprobados de año que reprobados, así mismo se observa que la mayoría de ellos posee un rendimiento académico medio.

- Las variables que obtuvieron mayor cantidad de registros atípicos y fuera de rango fueron las de: edad y sexo, se emplearon los algoritmos *InterquartileRange* y *RemoveWithValues* en la herramienta informática WEKA para la detección y posterior eliminación de estos registros.

- El algoritmo que mejores resultados obtuvo en cuando a la selección de atributos, fue el de *ConsistencySubsetEvaluator,* el mismo que obtuvo como resultado que de las 19 variables empleadas en cada base de datos, solo 11 variables se relacionan directamente con la aprobación o reprobación del año lectivo las mismas que son: Curso, EdadAnosCurso, Sexo, LugNacCant, LugResiProv, DistanColegio, AsisteEducacInicial, ViveCon, diasNoAsistidos, NumSuplet, Comportam, mientras que 14 variables se relacionan con el rendimiento académico de los estudiantes las mismas que son: Curso, EdadAnosCurso, Sexo, LugNacCant, LugNacParroq, LugResiProv, LugResiParroq, DistanColegio, AsisteEducacInicial, JornAcadm, ViveCon, DiasNoAsistidos, NumSuplet, Comportam.

- Mediante la Figura 9 se observa como las bases de datos cuentan con sus variables de clase desbalanceadas, registrando la aprobación con 4563 casos mientras que la reprobación del año lectivo cuenta con 1366, también se da una representación mayor en rendimiento académico medio con 4373 casos, versus el rendimiento alto 637,

rendimiento bueno 361 y rendimiento regular 556, por ende se emplea el algoritmo SMOTE para introducir datos ficticios con el fin de obtener bases de datos consistentes al emparejar los registros para que no exista un favoritismo al buscar reglas de asociación que caractericen el aprovechamiento académico.

- Se aplicaron un conjunto de 11 algoritmos basados en reglas asociación en la herramienta informática KEEL y 1 en WEKA, la mayoría responden a modelos evolutivos de búsqueda, para encontrar reglas que obtengan los mejores resultados basados en los parámetros de confianza y soporte. Identificando al clasificador GAR para KEEL, superando por más de 11% en la medida de confianza a su predecesor en la base de datos *promocion* y con 16% en la base de datos *rendimiento*.

- El modelo obtenido por medio de la evaluación de los expertos en educación para la promoción del año lectivo de los estudiantes responde a 6 reglas, la interpretación se ajusta en parte a los conocimientos previos de los expertos como cuando los estudiantes que cuenten con una cantidad de diez supletorios o un comportamiento malo en el año lectivo, obtendrán como resultado la pérdida del año lectivo.

  No obstante otras reglas no son tan fáciles de predecir por los expertos, por ejemplo los estudiantes de sexo femenino, que residan en la provincia de Los Ríos, que no tengan ningún supletorio y cuenten con un comportamiento excelente, obtendrán como resultado la aprobación del año lectivo, igual resultado se dará con los estudiantes de sexo masculino que no tengan ningún supletorio en el año lectivo y cuenten con un comportamiento muy bueno.

- En el caso del rendimiento académico el modelo obtenido con la ayuda de los expertos en educación, cuenta con 4 reglas las mismas no son fáciles de predecir por los expertos, dado que la información obtenida al interpretar estas reglas trasmiten un nuevo conocimiento; como por ejemplo los estudiantes que asistan a clases en la sección matutina y aunque tengan un supletorio, contaran con un rendimiento académico mayor o igual a 7 y menor o igual a 8.99, también los estudiantes de sexo masculino que estén asistiendo a clases en la sección matutina, no cuenten con ningún supletorio y tengan un comportamiento de muy bueno, obtendrán un rendimiento académico mayor o igual a 9 y menor o igual 10.

Con esto se puede concluir la siguiente investigación como satisfactoria, debido a que se obtuvo un modelo mediante 10 reglas de asociación escogidas por los expertos en educación, el cual permite caracterizar el aprovechamiento académico de los estudiantes de la Unidad Educativa 24 de Mayo.

5.2 Recomendaciones

La presente investigación es de mayor interés en el área de la educación secundaria, debido a que constantemente se plantea descubrir cuáles son los factores que influyen en el aprovechamiento académico de los estudiantes, de tal manera al finalizar el presente proyecto de investigación se recomienda lo siguiente:

- Implementar nuevas herramientas de minería para la ejecución de reglas de asociación que permita establecer relaciones entre las variables para extraer información acerca del aprovechamiento académico de los estudiantes.

- Sugerir a los directivos de la institución educativa que tengan en cuenta el modelo conformado por las 10 reglas obtenidas mediante la evaluación de los expertos en educación, para tomar decisiones al respecto, especialmente con el objetivo de favorecer el aprovechamiento académico de los estudiantes propensos a perder el año lectivo o contar con un rendimiento académico regular.

- Implementar nuevas variables al modelo, en la búsqueda de mejorar los índices de soporte y confianza, con el fin de obtener reglas que aporten más información sobre el aprovechamiento académico.

CAPÍTULO VI
BIBLIOGRAFÍA

[1] «Unidad Educativa 24 de Mayo,» 30 Enero 2011. [En línea]. Available: http://www.colegio24demayo.edu.ec/historia.aspx.

[2] Servente M. R. García-Martínez, «Tesis Doctoral Algoritmos TDIDT aplicados a la minería de datos inteligente.,» Universidad de Buenos Aires, Buenos Aires, Argentina, 2002.

[3] Perichinsky García Martínez, «A Data Mining Approach to Computational,» *Proceedings del Workshop de Investigadores en Ciencias de la Computación.,* pp. 107-110, 2000.

[4] Adeyemi Adejuwon Amir Mosavi, «Domain Driven Data Mining- Application to Business,» *IJCSI International Journal of Computer Science Issues,* vol. VII, nº 2, pp. 41-44, Julio 2010.

[5] Cabena Peter & Hadjinian Pablo & Stadler Rolf & Verhees Jaap & Zanasi Alessandro, *Discovering Data Mining. From Concept To Implementation*, I. U. S. River, Ed., New Jersey: Prentice-Hall, 1998.

[6] Y. Caballero, D. Álvarez, A. Baltá, R. Bello y M. García, *Un nuevo algoritmo de selección de rasgos basado en la teoría de los conjuntos aproximados,* Antioquia.: Rev. Fac. Ing. Univ. Antioquia, 2007, pp. 132-144.

[7] M.-S. Chen, J. Han y P. S. Yu, *Data mining: An overview from database perspective,* vol. 8, IEEE Transactions on Knowledge and Data Engineering , 2002, pp. 866 - 883.

[8] J. Derrac, A. Fernández, J. Luengo, S. García, L. Sánchez , J. Alcalá y F. Herrera, *KEEL: Una herramienta software para el análisis de sistemas difusos evolutivos,* Huelva: ESTYLF 2010, 2010, pp. 417-422.

[9] A. J. Calleja Gómez, «Minería De Datos Con Weka Para La Predicción Del Precio De Automóviles De Segunda Mano,» Valencia, 2010.

[10] C. Borgelt, *Efficient implementations of Apriori and Eclat.,* Melborne, Florida, 2003, pp. 280-296.

[11] J. Mata, J. Alvarez y J. Riquelme, *Mining numeric association rules with genetic algorithms.,* 2001, pp. 264-267.

[12] X. Yan, C. Zhang y S. Zhang, *Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support.,* 2009, p. 3066–3076.

[13] B. Alatas y E. Akin, *An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules.,* 2006, p. 230–237.

[14] A. Ghosh y B. Nath, *Multi-objective rule mining using genetic algorithms,* 2004, pp. 123-133.

[15] H. Qodmanan, M. Nasiri y B. Minaei-Bidgoli, *Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence.,* 2011, p. 288–298.

[16] D. Martín, A. Rosete, J. Alcalá-Fdez y F. Herrera, 2014, pp. 54-69.

[17] M. Jiménez Hernández, *Competencia social: intervención preventiva en la escuela,* 2000, pp. 21-48.

[18] C. Figueroa , *Sistemas de evaluación académica,* 1 ed., Editorial Universitaria, 2004.

[19] I. Sánchez López, *Apoyo parental y rendimiento académico,* Cd. Victoria, Tamaulipas, 2013.

[20] L. E. Torres Velázquez y N. Y. Rodríguez Soriano, *El rendimiento académico y contexto familiar en estudiantes universitarios,* vol. 11, Xalapa,: Enseñanza e Investigación en Psicología, 2006, pp. 255-270.

[21] A. Ballesteros Román, D. Sánchez Guzmán y R. García Salcedo, *Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo,* vol. 7, México D. F., 2013, pp. 662-668.

[22] C. Márquez Vera, C. Romero Morales y S. Ventura Soto, *Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos,* vol. 7, IEEE-RITA, 2012, pp. 109-117.

[23] D. L. De la Red Martínez y C. E. Podestá Gómez, *Metodología de Estudio del Rendimiento Académico Mediante la Minería de Datos,* vol. III, Revista Científica de Tecnología Educativa|, 2014, pp. 56-73.

[24] M. D. Saiz González, *Minería de datos para el análisis de los antecedentes familiares de la conducta suicida: hacia una definición del endofenotipo,* Alcalá , Madrid.

[25] J. Hernandez Orallo, M. Ramírez Quintana y C. Ferri Ramírez, *Introducción a la Minería de Datos,* 1 ed., Madrid: PEARSON EDUCACION S.A., 2005, pp. 19-39.

[26] N. Rodríguez, *Diseños Experimentales en Educación,* 91 ed., vol. XXXII, Caracas: Revista de Pedagogía, 2011, pp. 147-158.

[27] O. E. Gualdrón Guerrero, *Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales*, Tarragona, 2006, pp. 10-19.

[28] N. V. Chawla, K. W. Bowyer, L. . O. Hall y P. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*, AI Access Foundation and Morgan Kaufmann Publishers. , 2002, p. 321–357.

CAPÍTULO VII

ANEXOS

Anexo 1: códigos provinciales, cantonales y parroquiales para definir la distancia desde la residencia hasta la Unidad Educativa 24 de Mayo.

CÓD 12 LOS RIOS

CANTONES Y PARROQUIAS

Código	Nombre	Código	Nombre	Código	Nombre
CÓD 01	**CANTON BABAHOYO**				
50	BABAHOYO	04	EL SALTO	53	FEBRES CORDERO
01	CLEMENTE BAQUERIZO	51	*BARREIRO	54	PIMOCHA
02	DR. CAMILO PONCE	52	CARACOL	55	LA UNION
03	BARREIRO				
COD 02	**CANTON BABA**				
50	BABA	51	GUARE	52	ISLA DE BEJUCAL
COD 03	**CANTON MONTALVO**				
50	MONTALVO				
COD 04	**CANTON PUEBLOVIEJO**				
50	PUEBLOVIEJO	51	PUERTO PECHICHE	52	SAN JUAN
COD 05	**CANTON QUEVEDO**				
50	QUEVEDO	06	SAN CRISTOBAL	51	*BUENA FE
01	QUEVEDO	07	SIETE DE OCTUBRE	52	*MOCACHE
02	SAN CAMILO	08	24 DE MAYO	53	SAN CARLOS
03	*SAN JOSE	09	VENUS DEL RIO QUEVEDO	54	*VALENCIA
04	GUAYACAN	10	VIVA ALFARO	55	LA ESPERANZA
05	NICOLAS INFANTE DIAZ				
COD 06	**CANTON URDANETA**				
50	CATARAMA	51	RICAURTE		
COD 07	**CANTON VENTANAS**				
50	VENTANAS	51	QUINSALOMA	52	ZAPOTAL
COD 08	**CANTON VINCES**				
50	VINCES	51	ANTONIO SOTOMAYOR	52	*PALENQUE
COD 09	**CANTON PALENQUE**				
50	PALENQUE				
COD 10	**CANTON BUENA FE**				
50	SAN JACINTO DE BUENA FE	02	7 DE AGOSTO	51	PATRICIA PILAR
01	SAN JACINTO DE BUENA FE	03	11 DE OCTUBRE		
COD 11	**CANTON VALENCIA**				
50	VALENCIA				
COD 12	**CANTON MOCACHE**				
50	MOCACHE				

CÓD 09 GUAYAS

CANTONES Y PARROQUIAS

Código	Nombre	Código	Nombre	Código	Nombre
CÓD 01	**CANTON GUAYAQUIL**				
50	GUAYAQUIL	09	ROCA	51	*CHONGON
01	AYACUCHO	10	ROCAFUERTE	52	JUAN GOMEZ RENDON
02	BOLIVAR	11	SUCRE	53	MORRO
03	CARBO (CONCEPCION)	12	TARQUI	54	*PASCUALES
04	FEBRES CORDERO	13	URDANETA	55	*PLAYAS
05	GARCIA MORENO	14	XIMENA	56	POSORJA
06	LETAMENDI	15	CHONGON	57	PUNA
07	NUEVE DE OCTUBRE	16	PASCUALES	58	TENGUEL
08	OLMEDO				
COD 02	**CANTON ALFREDO BAQUERIZO MORENO (JUJAN)**				
50	ALFREDO BAQUERIZO M				
COD 03	**CANTON BALAO**				
50	BALAO				
COD 04	**CANTON BALZAR**				
50	BALZAR				
COD 05	**CANTON COLIMES**				
50	COLIMES	51	SAN JACINTO		
COD 06	**CANTON DAULE**				
50	DAULE	52	JUAN BAUTISTA AGUIRRE	55	*LOMAS DE SARGENTILLO
01	DAULE	53	LAUREL	56	LOS LOJAS
02	LA AURORA	54	LIMONAL	57	*PIEDRAHITA
51	*ISIDRO AYORA				
COD 07	**CANTON DURAN**				
50	ELOY ALFARO				

Código	Nombre	Código	Nombre	Código	Nombre
COD 08	**CANTON EMPALME**				
50	VELASCO IBARRA	51	GUAYAS	52	EL ROSARIO
COD 09	**CANTON EL TRIUNFO**				
50	EL TRIUNFO				
COD 10	**CANTON MILAGRO**				
50	MILAGRO	52	*GENERAL ELIZALDE	54	ROBERTO ASTUDILLO
51	CHOBO	53	MARISCAL SUCRE		
COD 11	**CANTON NARANJAL**				
50	NARANJAL	52	SAN CARLOS	54	TAURA
51	JESUS MARIA	53	SANTA ROSA DE FLANDES		
COD 12	**CANTON NARANJITO**				
50	NARANJITO				
COD 13	**CANTON PALESTINA**				
50	PALESTINA				
CDO 14	**CANTON PEDRO CARBO**				
50	PEDRO CARBO,	51	VALLE DE LA VIRGEN	52	SABANILLA
COD 15	**CANTON SALINAS**				
50	SALINAS	03	VICENTE ROCAFUERTE	52	JOSE LUIS TAMAYO
01	CARLOS ESPINOZA L	04	SANTA ROSA	53	*LA LIBERTAD
02	GRAL. ALBERTO ENRIQUEZ	51	ANCONCITO		
COD 16	**CANTON SAMBORONDON**				
50	SAMBORONDON	51	TARIFA		
COD 17	**CANTON SANTA ELENA**				
50	SANTA ELENA	51	ATAHUALPA	54	MANGLARALTO
01	BALLENITA	52	COLONCHE	55	SIMON BOLIVAR
02	SANTA ELENA	53	CHANDUY	56	SAN JOSE DE ANCON
COD 18	**CANTON SANTA LUCIA**				
50	SANTA LUCIA				
COD 19	**CANTON SALITRE**				
50	EL SALITRE	03	CENTRAL	51	GENERAL VERNAZA
01	BOCANA	04	PARAISO	52	LA VICTORIA
02	CANDILEJOS	05	SAN MATEO	53	JUNQUILLAL
COD 20	**CANTON SAN JACINTO DE YAGUACHI**				
50	S JACINTO DE YAGUACHI	53	GRAL. PEDRO J. MONTERO	55	YAGUACHI VIEJO
51	*CRL. LORENZO DE GARAICOA	54	*SIMON BOLIVAR	56	VIRGEN DE FATIMA
52	*CRL. MARCELINO MARIDUEÑA				
COD 21	**CANTON PLAYAS**				
50	GENERAL VILLAMIL				
COD 22	**CANTON SIMON BOLIVAR**				
50	SIMON BOLIVAR	51	CRL.LORENZO DE GARAICOA		
COD 23	**CANTON CORONEL MARCELINO MARIDUEÑA**				
50	CRL MARCELINO MARIDUEÑA				
COD 24	**CANTON LOMAS DE SARGENTILLO**				
50	LOMAS DE SARGENTILLO	51	*ISIDRO AYORA		
COD 25	**CANTON NOBOL**				
50	NARCISA DE JESUS				
COD 26	**CANTON LA LIBERTAD**				
50	LA LIBERTAD				
COD 27	**CANTON GENERAL ANTONIO ELIZALDE**				
50	GRAL ANTONIO ELIZALDE				
COD 28	**CANTON ISIDRO AYORA**				
50	ISIDRO AYORA				

Fuente: http://web.educacion.gob.ec/CNIE/pdf/Anexo%20con%20Codificacion.pdf

Anexo 2: Medidas por defecto del algoritmo APRIORI en WEKA

Medida	***Valor por defectos***	***Descripción***
car	False	Si las reglas de asociación de clase habilitado se extraen en lugar de reglas de asociación (general).
classIndex	-1	Índice del atributo de clase. Si se establece en -1, el último atributo se toma como atributo de clase.
delta	0,05	Reduce el soporte hasta que se alcance el soporte requerido mínimo o número de reglas que se ha generado.
lowerBoundMinSupport	0,1	Límite inferior para el soporte mínimo.
metricType	Confidence	La confianza es la proporción de los ejemplos incluidos en la premisa de que también están cubiertos por la consecuencia (reglas de asociación de clase sólo pueden ser extraídos utilizando confianza).
minMetric	0,9	Puntuación mínima métrica, para la medida por la cual se va a clasificar.
numRules	10	Número de reglas a encontrar.
outItemSets	False	Permitir a los conjuntos de elementos
removeAllMissingCols	False	Eliminar columnas con todos los valores que faltan.
significanceLevel	-1,0	El nivel de significación. Prueba de significación (sólo la confianza métrica).
upperBoundMinSupport	1,0	Límite superior de soporte mínimo. Iniciar de forma iterativa la disminución del soporte mínimo de este valor.
verbose	False	Si está activado el algoritmo se ejecuta en modo detallado.

Anexo 3: Parámetros de medidas para generación de Reglas de asociación en WEKA.

Asociación con clase		**Asociación sin clase**	
Medida	**Valor para indicadores**	**Medida**	**Valor para indicadores**
car	True	car	false
classIndex	19	classIndex	19
delta	0,05	delta	0,05
lowerBoundMinSupport	0,1	lowerBoundMinSupport	0,1
metricType	Confidence	metricType	Confidence
minMetric	0,85	minMetric	0,85
numRules	1000	numRules	1000
outItemSets	False	outItemSets	False
removeAllMissingCols	False	removeAllMissingCols	False

significanceLevel	-1,0	significanceLevel	-1,0
upperBoundMinSupport	1,0	upperBoundMinSupport	1,0
verbose	False	verbose	False

Anexo 4: Primeras 12 reglas obtenidas en la base de datos *promocion* con valores atípicos.

Reglas de asociación con clase		**% conf**	**Reglas de asociación sin clase**		**% conf**
1. JornAcadm=1 NumSuplet=0 2921	==> pierdeanio=2 2862	conf:(0.98)	1. LugResiCant=5 5499	´==> LugResiProv=12 5489	conf:(1)
2. Nacionalidad=1 JornAcadm=1 NumSuplet=0 2886	==> pierdeanio=2 2827	conf:(0.98)	2. Nacionalidad=1 LugResiCant=5 5465	LugResiProv=12 5455	conf:(1)
3. LugNacProv=12 NumSuplet=0 3109	==> pierdeanio=2 3027	conf:(0.97)	3. LugResiCant=5 5499	Nacionalidad=1 5465	conf:(0.99)
4. Nacionalidad=1 LugNacProv=12 NumSuplet=0 3075	==> pierdeanio=2 2993	conf:(0.97)	4. LugResiProv=12 LugResiCant=5 5489	Nacionalidad=1 5455	conf:(0.99)
5. LugNacProv=12 LugResiProv=12 NumSuplet=0 2917	==> pierdeanio=2 2836	conf:(0.97)	5. LugResiProv=12 5756	Nacionalidad=1 5719	conf:(0.99)
6. Nacionalidad=1 LugNacProv=12 LugResiProv=12 NumSuplet=0 2897	==> pierdeanio=2 2816	conf:(0.97)	6. LugResiCant=5 5499	Nacionalidad=1 LugResiProv=12 5455	conf:(0.99)
7. Nacionalidad=1 NumSuplet=0 3434	==> pierdeanio=2 3335	conf:(0.97)	7. LugNacParroq=1 5420	Nacionalidad=1 5375	conf:(0.99)
8. LugNacParroq=1 NumSuplet=0 3045	==> pierdeanio=2 2957	conf:(0.97)	8. LugNacProv=12 5526	Nacionalidad=1 5471	conf:(0.99)
9. Nacionalidad=1 LugNacParroq=1 NumSuplet=0 3019	==> pierdeanio=2 2931	conf:(0.97)	9. Nacionalidad=1 LugResiProv=12 5719	LugResiCant=5 5455	conf:(0.95)
10. NumSuplet=0 3501	==> pierdeanio=2 3398	conf:(0.97)	10. LugResiProv=12 5756	LugResiCant=5 5489	conf:(0.95)
11. LugResiProv=12 NumSuplet=0 3263	==> pierdeanio=2 3165	conf:(0.97)	11. LugResiProv=12 5756	Nacionalidad=1 LugResiCant=5 5455	conf:(0.95)
12. Nacionalidad=1 LugResiProv=12 NumSuplet=0 3238	==> pierdeanio=2 3140	conf:(0.97)	12. Nacionalidad=1 6066	LugResiProv=12 5719	conf:(0.94)

Anexo 5: Primeras 12 reglas obtenidas en la base de datos *rendimiento* con valores atípicos

BASE CON RUIDO *rendimiento*					
Reglas de asociación de clase		**% conf**	**Reglas de asociación generales**		**% conf**
1. LugNacCant=5 LugResiProv=12 JornAcadm=1 Comportam=2 2221	==> Rendimiento=2 1909	conf:(0.86)	1. LugResiCant=5 5499	==> LugResiProv=12 5489	conf:(1)
2. LugNacProv=12 LugNacCant=5 LugResiProv=12 JornAcadm=1 Comportam=2 2213	==> Rendimiento=2 1901	conf:(0.86)	2. Nacionalidad=1 LugResiCant=5 5465	==> LugResiProv=12 5455	conf:(1)
3. Nacionalidad=1 LugNacCant=5 LugResiProv=12 JornAcadm=1 Comportam=2 2206	==> Rendimiento=2 1894	conf:(0.86)	3. LugResiCant=5 5499	==> Nacionalidad=1 5465	conf:(0.99)
4. Nacionalidad=1 LugNacProv=12 LugNacCant=5 LugResiProv=12 JornAcadm=1 Comportam=2 2198	==> Rendimiento=2 1886	conf:(0.86)	4. LugResiProv=12 LugResiCant=5 5489	==> Nacionalidad=1 5455	conf:(0.99)
5. LugNacCant=5 JornAcadm=1 Comportam=2 2397	==> Rendimiento=2 2056	conf:(0.86)	5. LugResiProv=12 5756	==> Nacionalidad=1 5719	conf:(0.99)
6. LugNacProv=12 LugNacCant=5 JornAcadm=1 Comportam=2 2389	==> Rendimiento=2 2048	conf:(0.86)	6. LugResiCant=5 5499	==> Nacionalidad=1 LugResiProv=12 5455	conf:(0.99)
7. Nacionalidad=1 LugNacCant=5 JornAcadm=1 Comportam=2 2369	==> Rendimiento=2 2030	conf:(0.86)	7. LugNacParroq=1 5420	==> Nacionalidad=1 5375	conf:(0.99)
8. Nacionalidad=1 LugNacProv=12 LugNacCant=5 JornAcadm=1 Comportam=2 2361	==> Rendimiento=2 2022	conf:(0.86)	8. LugNacProv=12 5526	==> Nacionalidad=1 5471	conf:(0.99)
9. LugNacCant=5 LugNacParroq=1 JornAcadm=1 Comportam=2 2220	==> Rendimiento=2 1901	conf:(0.86)	9. Nacionalidad=1 LugResiProv=12 5719	==> LugResiCant=5 5455	conf:(0.95)
10. LugNacProv=12 LugNacCant=5 LugNacParroq=1 JornAcadm=1 Comportam=2 2213	==> Rendimiento=2 1894	conf:(0.86)	10. LugResiProv=12 5756	==> LugResiCant=5 5489	conf:(0.95)

11. Nacionalidad=1 LugNacCant=5 LugNacParroq=1 JornAcadm=1 Comportam=2 2198	==> Rendimiento=2 1881	conf:(0.86)	11. LugResiProv=12 5756	==> Nacionalidad=1 LugResiCant=5 5455	conf:(0.95)
12. Nacionalidad=1 LugNacProv=12 LugNacCant=5 LugNacParroq=1 JornAcadm=1 Comportam=2 2191	==> Rendimiento=2 1874	conf:(0.86)	12. Nacionalidad=1 6066	==> LugResiProv=12 5719	conf:(0.94)

Anexo 6: Medidas del Filtro *InterquartileRange* en WEKA para la detección de valores extremos y atípicos.

Parámetro	VALOR POR DEFECTOS	DESCRIPCIÓN
AttributeIndices	fist-last	Especifica la gama de atributos donde se ejecutara
Debug	False	Activa la salida de información depurada
DetectionperAttribute	False	Genera outliers / ExtremeValue atributo par para cada atributo numérico
ExtremEvaluesasOutliers	False	Etiqueta valores extremos u outliers
ExtremeValuesFactor	6,0	Factor para la determinación de los umbrales de los valores extremos
OutlierFactor	3,0	Factor para la determinación de los umbrales de outliers
OutputOffSetMultiplier	False	Genera un atributo adicional "offset"

Anexo 7: Medidas del filtro *RemoveWithValues* en WEKA para la eliminación de valores extremos y atípicos.

Parámetro	VALOR POR DEFECTOS	DESCRIPCIÓN
Attributeinidces	last	Seleccione el atributo que se utilizará para la selección (por defecto última).
Dontfilterafterfirstbatch	False	si desea aplicar el proceso de filtrado a los casos que se reciben después del primer lote
Invertselection	False	Invertir sentido de atributos
Matchmissingvalues	False	Los valores perdidos se cuentan como un valor
Modifyheader	False	Al seleccionar los atributos nominales, elimina las referencias a las cabeceras de valores excluidos.

Nominalindices	first-last	Rango de los índices de etiqueta que se utilizará para la selección en el atributo nominal
Splitpoint	0.0	Se seleccionarán los casos con valores menores que el valor dado.

Anexo 8: Resultados obtenidos en la base de datos *promocion* con sus registros balanceados mediante el algoritmo GAR

rule id=0 antecedents LugResiProv =12 consequents LugNacCant =5
rule id=0 rule_support=0.86 confidence=0.89
rule id=1 antecedents LugNacCant =5 consequents LugResiProv =12
rule id=1 rule_support=0.86 confidence=0.97
rule id=2 antecedents pierdeanio =1 consequents LugNacCant =5
rule id=2 rule_support=0.51 confidence=0.94
rule id=3 antecedents ViveCon =1 consequents LugResiProv =12
rule id=3 rule_support=0.54 confidence=0.98
rule id=4 antecedents AsisteEducacInicial =0 consequents LugResiProv =12
rule id=4 rule_support=0.53 confidence=0.96
rule id=5 antecedents pierdeanio =1 consequents LugResiProv =12
rule id=5 rule_support=0.54 confidence=0.98
rule id=6 antecedents Sexo =1 consequents LugNacCant =5
rule id=6 rule_support=0.56 confidence=0.9
rule id=7 antecedents Sexo =1 LugResiProv =12 consequents LugNacCant =5
rule id=7 rule_support=0.54 confidence=0.9
rule id=8 antecedents Sexo =1 LugNacCant =5 consequents LugResiProv =12
rule id=8 rule_support=0.54 confidence=0.97
rule id=9 antecedents Sexo =1 consequents LugResiProv =12
rule id=9 rule_support=0.6 confidence=0.97

Anexo 9: Resultados obtenidos en la base de datos *rendimiento* con sus registros desbalanceados mediante el algoritmo GAR

rule id=0 antecedents Rendimiento=2 consequents LugResiProv=12
rule id=0 rule_support=0.7 confidence=0.95

rule id=1 antecedents LugNacParroq=1 consequents LugNacCant=5

rule id=1 rule_support=0.77 confidence=0.86

rule id=2 antecedents LugNacCant=5 consequents LugNacParroq=1

rule id=2 rule_support=0.77 confidence=0.93

rule id=3 antecedents LugNacParroq=1 consequents LugNacCant=5

rule id=3 rule_support=0.77 confidence=0.86

rule id=4 antecedents LugNacCant=5 consequents LugNacParroq=1

rule id=4 rule_support=0.77 confidence=0.93

rule id=5 antecedents JornAcadm=1 consequents LugNacCant=5

rule id=5 rule_support=0.7 confidence=0.83

rule id=6 antecedents LugNacCant=5 consequents JornAcadm=1

rule id=6 rule_support=0.7 confidence=0.84

rule id=7 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=7 rule_support=0.8 confidence=0.95

rule id=8 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=8 rule_support=0.8 confidence=0.85

rule id=9 antecedents LugResiProv=12 consequents LugNacParroq=1

rule id=9 rule_support=0.84 confidence=0.89

rule id=10 antecedents LugNacParroq=1 consequents LugResiProv=12

rule id=10 rule_support=0.84 confidence=0.94

rule id=11 antecedents JornAcadm=1 consequents LugNacParroq=1

rule id=11 rule_support=0.75 confidence=0.9

rule id=12 antecedents LugNacParroq=1 consequents JornAcadm=1

rule id=12 rule_support=0.75 confidence=0.85

rule id=13 antecedents LugNacParroq=1 consequents LugNacCant=5

rule id=13 rule_support=0.77 confidence=0.86

rule id=14 antecedents LugNacCant=5 consequents LugNacParroq=1

rule id=14 rule_support=0.77 confidence=0.93

rule id=15 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=15 rule_support=0.8 confidence=0.95

rule id=16 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=16 rule_support=0.8 confidence=0.85

rule id=17 antecedents Rendimiento=2 consequents JornAcadm=1

rule id=17 rule_support=0.65 confidence=0.88

rule id=18 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=18 rule_support=0.8 confidence=0.95

rule id=19 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=19 rule_support=0.8 confidence=0.85

rule id=20 antecedents JornAcadm=1 consequents LugNacParroq=1

rule id=20 rule_support=0.75 confidence=0.9

rule id=21 antecedents LugNacParroq=1 consequents JornAcadm=1

rule id=21 rule_support=0.75 confidence=0.85

rule id=22 antecedents JornAcadm=1 consequents LugNacCant=5

rule id=22 rule_support=0.7 confidence=0.83

rule id=23 antecedents LugNacCant=5 consequents JornAcadm=1

rule id=23 rule_support=0.7 confidence=0.84

rule id=24 antecedents JornAcadm=1 consequents LugNacParroq=1

rule id=24 rule_support=0.75 confidence=0.9

rule id=25 antecedents LugNacParroq=1 consequents JornAcadm=1

rule id=25 rule_support=0.75 confidence=0.85

rule id=26 antecedents LugResiProv=12 consequents LugNacParroq=1

rule id=26 rule_support=0.84 confidence=0.89

rule id=27 antecedents LugNacParroq=1 consequents LugResiProv=12

rule id=27 rule_support=0.84 confidence=0.94

rule id=28 antecedents LugResiProv=12 consequents LugNacCant=5

rule id=28 rule_support=0.78 confidence=0.83

rule id=29 antecedents LugNacCant=5 consequents LugResiProv=12

rule id=29 rule_support=0.78 confidence=0.94

rule id=30 antecedents Rendimiento=2 consequents JornAcadm=1

rule id=30 rule_support=0.65 confidence=0.88

rule id=31 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=31 rule_support=0.8 confidence=0.95

rule id=32 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=32 rule_support=0.8 confidence=0.85

rule id=33 antecedents Rendimiento=2 consequents JornAcadm=1

rule id=33 rule_support=0.65 confidence=0.88

rule id=34 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=34 rule_support=0.8 confidence=0.95

rule id=35 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=35 rule_support=0.8 confidence=0.85

rule id=36 antecedents Rendimiento=2 consequents JornAcadm=1

rule id=36 rule_support=0.65 confidence=0.88

rule id=37 antecedents LugResiProv=12 consequents LugNacParroq=1

rule id=37 rule_support=0.84 confidence=0.89

rule id=38 antecedents LugNacParroq=1 consequents LugResiProv=12

rule id=38 rule_support=0.84 confidence=0.94

rule id=39 antecedents LugResiProv=12 consequents LugNacParroq=1

rule id=39 rule_support=0.84 confidence=0.89

rule id=40 antecedents LugNacParroq=1 consequents LugResiProv=12

rule id=40 rule_support=0.84 confidence=0.94

rule id=41 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=41 rule_support=0.8 confidence=0.95

rule id=42 antecedents JornAcadm=1 consequents LugNacProv=12

rule id=42 rule_support=0.8 confidence=0.85

rule id=43 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=43 rule_support=0.8 confidence=0.95

rule id=44 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=44 rule_support=0.8 confidence=0.85

rule id=45 antecedents JornAcadm=1 consequents LugResiProv=12

rule id=45 rule_support=0.8 confidence=0.95

rule id=46 antecedents LugResiProv=12 consequents JornAcadm=1

rule id=46 rule_support=0.8 confidence=0.85

rule id=47 antecedents Rendimiento=2 consequents JornAcadm=1

rule id=47 rule_support=0.65 confidence=0.88

rule id=48 antecedents JornAcadm=1 consequents LugNacCant=5

rule id=48 rule_support=0.7 confidence=0.83

rule id=49 antecedents LugNacCant=5 consequents JornAcadm=1

rule id=49 rule_support=0.7 confidence=0.84

rule id=50 antecedents JornAcadm=1 consequents LugNacParroq=1

rule id=50 rule_support=0.75 confidence=0.9

rule id=51 antecedents LugNacParroq=1 consequents JornAcadm=1

rule id=51 rule_support=0.75 confidence=0.85

rule id=52 antecedents Rendimiento=2 consequents JornAcadm=1
rule id=52 rule_support=0.65 confidence=0.88
rule id=53 antecedents JornAcadm=1 consequents LugResiProv=12
rule id=53 rule_support=0.8 confidence=0.95

Anexo 10: Resultados obtenidos en la base de datos *promocion* con sus registros balanceados mediante el algoritmo APRIORI en WEKA

1. LugNacCant=5 Comportam=4 1161 ==> pierdeanio=1 1161 conf:(1)
2. NumSuplet=10 1155 ==> pierdeanio=1 1155 conf:(1)
3. LugResiProv=12 NumSuplet=10 1132 ==> pierdeanio=1 1132 conf:(1)
4. LugNacCant=5 LugResiProv=12 Comportam=4 1122 ==> pierdeanio=1 1122 conf:(1)
5. LugNacCant=5 NumSuplet=10 1063 ==> pierdeanio=1 1063 conf:(1)
6. LugNacCant=5 LugResiProv=12 NumSuplet=10 1040 ==> pierdeanio=1 1040 conf:(1)
7. Comportam=4 1250 ==> pierdeanio=1 1249 conf:(1)
8. LugResiProv=12 Comportam=4 1211 ==> pierdeanio=1 1210 conf:(1)
9. AsisteEducacInicial=0 NumSuplet=0 1638 ==> pierdeanio=2 1597 conf:(0.97)
10. LugResiProv=12 AsisteEducacInicial=0 NumSuplet=0 1523 ==> pierdeanio=2 1483 conf:(0.97)
11. LugNacCant=5 AsisteEducacInicial=0 NumSuplet=0 1389 ==> pierdeanio=2 1350 conf:(0.97)
12. LugNacCant=5 LugResiProv=12 AsisteEducacInicial=0 NumSuplet=0 1284 ==> pierdeanio=2 1246 conf:(0.97)
13. Sexo=2 NumSuplet=0 Comportam=1 1249 ==> pierdeanio=2 1182 conf:(0.95)
14. Sexo=2 LugResiProv=12 NumSuplet=0 Comportam=1 1175 ==> pierdeanio=2 1108 conf:(0.94)
15. NumSuplet=0 Comportam=1 1763 ==> pierdeanio=2 1660 conf:(0.94)
16. LugResiProv=12 NumSuplet=0 Comportam=1 1679 ==> pierdeanio=2 1576 conf:(0.94)
17. LugNacCant=5 NumSuplet=0 Comportam=1 1444 ==> pierdeanio=2 1349 conf:(0.93)
18. LugNacCant=5 LugResiProv=12 NumSuplet=0 Comportam=1 1374 ==> pierdeanio=2 1279 conf:(0.93)

19. NumSuplet=0 Comportam=2 1699 ==> pierdeanio=2 1566 conf:(0.92)
20. LugNacCant=5 NumSuplet=0 Comportam=2 1441 ==> pierdeanio=2 1324 conf:(0.92)
21. LugResiProv=12 NumSuplet=0 Comportam=2 1596 ==> pierdeanio=2 1464 conf:(0.92)
22. LugNacCant=5 LugResiProv=12 NumSuplet=0 Comportam=2 1344 ==> pierdeanio=2 1228 conf:(0.91)
23. Sexo=2 NumSuplet=0 1827 ==> pierdeanio=2 1667 conf:(0.91)
24. Sexo=2 LugResiProv=12 NumSuplet=0 1721 ==> pierdeanio=2 1561 conf:(0.91)
25. Sexo=1 NumSuplet=0 Comportam=2 1197 ==> pierdeanio=2 1085 conf:(0.91)
26. Sexo=2 LugNacCant=5 NumSuplet=0 1522 ==> pierdeanio=2 1376 conf:(0.9)
27. Sexo=1 LugResiProv=12 NumSuplet=0 Comportam=2 1126 ==> pierdeanio=2 1015 conf:(0.9)

Anexo 11: Resultados obtenidos en la base de datos *rendimiento* con sus registros desbalanceados mediante el algoritmo APRIORI en WEKA

1. LugNacParroq=1 JornAcadm=1 NumSuplet=1 654 ==> Rendimiento=2 651 conf:(1)
2. LugNacParroq=1 LugResiProv=12 JornAcadm=1 NumSuplet=1 621 ==> Rendimiento=2 618 conf:(1)
3. LugNacCant=5 JornAcadm=1 NumSuplet=1 603 ==> Rendimiento=2 600 conf:(1)
4. JornAcadm=1 NumSuplet=1 718 ==> Rendimiento=2 714 conf:(0.99)
5. LugResiProv=12 JornAcadm=1 NumSuplet=1 684 ==> Rendimiento=2 680 conf:(0.99)
6. LugNacCant=5 NumSuplet=1 700 ==> Rendimiento=2 694 conf:(0.99)
7. LugNacCant=5 LugNacParroq=1 NumSuplet=1 653 ==> Rendimiento=2 647 conf:(0.99)
8. LugNacCant=5 LugResiProv=12 NumSuplet=1 652 ==> Rendimiento=2 646 conf:(0.99)
9. LugNacParroq=1 NumSuplet=1 752 ==> Rendimiento=2 745 conf:(0.99)
10. NumSuplet=1 832 ==> Rendimiento=2 824 conf:(0.99)
11. LugNacCant=5 LugNacParroq=1 LugResiProv=12 NumSuplet=1 605 ==> Rendimiento=2 599 conf:(0.99)

12. LugNacParroq=1 LugResiProv=12 NumSuplet=1 704 ==> Rendimiento=2 697 conf:(0.99)

13. LugResiProv=12 NumSuplet=1 783 ==> Rendimiento=2 775 conf:(0.99)

14. Sexo=1 LugNacCant=5 LugResiProv=12 JornAcadm=1 NumSuplet=0 Comportam=2 788 ==> Rendimiento=2 712 conf:(0.9)

15. Sexo=1 LugResiProv=12 JornAcadm=1 NumSuplet=0 Comportam=2 938 ==> Rendimiento=2 847 conf:(0.9)

16. Sexo=1 JornAcadm=1 NumSuplet=0 Comportam=2 736 ==> Rendimiento=1 663 conf:(0.9)