
**ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ**

## CARRERA DE INFORMÁTICA

**TRABAJO DE TITULACIÓN PREVIA LA OBTENCIÓN DEL TÍTULO DE INGENIERA EN INFORMÁTICA**

## MODALIDAD: SISTEMATIZACIÓN DE EXPERIENCIAS

## TEMA:

# MINERÍA DE DATOS APLICADA A LA CLASIFICACIÓN DEL RENDIMIENTO ACADÉMICO

**AUTORAS:**

**SULAY KATERINE CEVALLOS MOLINA**

**VIVIANA KATHERINE TRUJILLO UTRERAS**

**TUTORA:**

**ING. JÉSSICA JOHANNA MORALES CARRILLO, MG.**

**CALCETA, NOVIEMBRE 2018**

# DERECHOS DE AUTORÍA

Sulay Katerine Cevallos Molina y Viviana Katherine Trujillo Utreras, declaran bajo juramento que el trabajo aquí descrito es de nuestra autoría, que no ha sido previamente presentado para ningún grado o calificación profesional, y que hemos consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedemos los derechos de propiedad intelectual a la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, según lo establecido por la Ley de Propiedad Intelectual y su reglamento.

| | |
|---|---|
| ..... | ..... |
| **SULAY K. CEVALLOS MOLINA** | **VIVIANA K. TRUJILLO UTRERAS** |

# CERTIFICACIÓN DEL TUTOR

Jessica Johanna Morales Carrillo certifica haber tutelado el trabajo de titulación **MINERÍA DE DATOS APLICADA A LA CLASIFICACIÓN DEL RENDIMIENTO ACADÉMICO,** que ha sido desarrollada por Sulay Katerine Cevallos Molina y Viviana Katherine Trujillo Utreras, previa la obtención del título de Ingeniero en Informática, de acuerdo al **REGLAMENTO DE LA UNIDAD DE TITULACIÓN ESPECIAL DE PROGRAMAS DE GRADO** de la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López.

……

**ING. JÉSSICA J. MORALES CARRILLO, MG**

# APROBACIÓN DEL TRIBUNAL

Los suscritos integrantes del tribunal correspondiente, declaran que han **APROBADO** el Trabajo de Titulación **MINERÍA DE DATOS APLICADA A LA CLASIFICACIÓN DEL RENDIMIENTO ACADÉMICO,** que ha sido propuesto, desarrollado y sustentado por Sulay Katerine Cevallos Molina y Viviana Katherine Trujillo Utreras, previa la obtención del título de Ingeniero en Informática, de acuerdo al **REGLAMENTO DE LA UNIDAD DE TITULACIÓN ESPECIAL DE PROGRAMAS DE GRADO** de la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López.

……...  ……...

ING. RICARDO A. VÉLEZ VALAREZO  DR. INF. JORGE A. PÁRRAGA ÁLAVA

**MIEMBRO**  **MIEMBRO**

……...

ING. DANIEL A. MERA MARTÍNEZ, MGTR

**PRESIDENTE**

# AGRADECIMIENTO

A la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López que nos dio la oportunidad de estudiar y ser profesionales.

Al Ing. Luis Ortega director encargado de la Carrera de Computación por su esfuerzo y dedicación,

A nuestra tutora, Ing. Jessica Morales Carrillo por la dedicación y apoyo que ha brindado a este trabajo, por las sugerencias e ideas y

A los docentes que durante el transcurso de la carrera nos han impartido conocimientos significativos, gracias por la ayuda, consejos, amistad y tiempo para atender nuestras inquietudes.

# DEDICATORIA

Quiero dedicar este trabajo a Dios quien me ha brindado la vida y al mismo tiempo me da la fortaleza y la sabiduría para seguir adelante y así poder superar obstáculos.

A mis padres que me formaron como mujer de bien para poder ser una gran profesional, a mi esposo que siempre está presente en los momentos en donde más lo he necesitado brindándome su apoyo incondicional y a mi abuelo que siempre me ha aconsejado día a día para ser una persona llena de humildad y respeto.

**SULAY K. CEVALLOS MOLINA**

# DEDICATORIA

Dedico este trabajo de titulación a Dios por ser mi fortaleza en todos los momentos de mi vida, siendo mi aliento y esperanza en los momentos más difíciles para no rendirme y poder culminar mis estudios.

Para mis padres quienes son las personas que siempre me han apoyado emocional y económicamente en cada paso para lograr mis metas. Me han formado como una mujer de principios y valores que con altivez y orgullo puedo decir que gracias a ello he logrado obtener mi profesión. A mi novio que siempre ha estado en cada momento que lo he necesitado dándome ánimos para seguir, y demás familiares por sus consejos dados en cada etapa universitaria.

**VIVIANA K. TRUJILLO UTRERAS**

# CONTENIDO GENERAL

# CONTENIDO DE CUADROS Y FIGURAS

## CUADROS



## FIGURAS

**GRÁFICOS**

# RESUMEN

El presente trabajo de titulación fue desarrollado, con el objetivo de aplicar los principales algoritmos de minería de datos utilizados en la educación para realizar inferencias en la clasificación del rendimiento académico de los estudiantes de la carrera de computación de la ESPAM MFL. La herramienta utilizada para el proceso de análisis fue el software WEKA (Waikato Environment for Knowledge Analysis), además se usó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) la cual está estructurada en fases de desarrollo: inicialmente se buscó información de investigaciones ya existentes lo cual permitió realizar un enfoque de los algoritmos y variables a aplicar, también se realizó la adaptación de los datos de acuerdo a la herramienta seleccionada, una vez preparados los datos se aplicaron las técnicas más apropiadas y se realizaron las pruebas pertinentes para determinar la utilidad de los modelos obtenidos a partir de las variables que se evaluaron. Se estableció que los principales algoritmos de la técnica de clasificación son el J48, Naïve Bayes, Random Forest y OneR ya que estos son los más utilizados en la minería de datos educacional debido a su precisión en la clasificación de datos, con los modelos se determinó que las variables que más inciden en la clasificación del rendimiento académico de los estudiantes son: estado académico, semestre y sub total, se aplicaron estos algoritmos con el fin de obtener un modelo que genere conocimiento que apoye la toma de decisiones en el proceso de educación superior.

# PALABRAS CLAVE

Minería de datos, rendimiento académico, árbol de decisión, reglas de asociación, naïve bayes.

# ABSTRACT

The present work of degree was developed with the objective of applying the main data mining algorithms used in education to make inferences in the classification of the academic performance of the students of the computing career of the ESPAM MFL. The tool used for the analysis process was the WEKA (Waikato Environment for Knowledge Analysis) software. In addition, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was used, which is structured in phases of development. Initially, the information was sought from existing research which allowed us to perform an approach to the algorithms and variables to apply, also the adaptation of the data was carried out according to the selected tool, once the data were prepared, the most appropriate techniques were applied and the pertinent tests were carried out to determine the usefulness of the models obtained from the variables that were evaluated. It was established that the main algorithms of the classification technique are J48, Naïve Bayes, Random Forest and OneR since these are the most used in educational data mining due to their accuracy in the classification of data, with the models determined that the variables that most affect the classification of academic performance of students are: academic state, semester and sub-total, these algorithms were applied in order to obtain a model that generates knowledge that supports decision-making in the education process higher.

# KEY WORDS

Data mining, academic performance, decision tree, association rules, naïve bayes.

# CAPÍTULO I. ANTECEDENTES

## 1.1 DESCRIPCIÓN DE LA INSTITUCIÓN

Las gestiones de la ESPAM MFL (2015a) empezaron en el Congreso Nacional y luego en otras instancias desde 1995. El 29 de abril de 1996 se crea el INSTITUTO TECNOLÓGICO SUPERIOR AGROPECUARIO DE MANABÍ, ITSAM, mediante Ley Nº. 116, publicada en el R.O. Nº. 935. Y es así que, en el 1999, el Congreso Nacional expidió la Ley Reformatoria que transformaba el Instituto Tecnológico Superior Agropecuario de Manabí, ITSAM, en ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ, ESPAM, cuya Ley 99-25 fue publicada en el R.O. el 30 de abril de 1999.

De acuerdo con la reseña histórica la ESPAM MFL (2015a) nace como persona jurídica de derecho público, autónoma, que se rige por la Constitución Política del Estado, Ley de Educación Superior, su Estatuto Orgánico y Reglamentos para preparar a la juventud ecuatoriana y convertirla en profesionales, conforme lo exigen los recursos naturales de su entorno. Iniciando sus labores con las Carreras de Agroindustrias, Medio Ambiente, Agrícola y Pecuaria. Luego mediante un estudio de mercado, se crea la Carrera de Informática emprendiendo así, un riguroso programa de fortalecimiento académico, con el fin de formar profesionales que ejecuten proyectos sustentables, que generen fuentes de trabajo.

La carrera de Informática en la Escuela Superior Politécnica Agropecuaria de Manabí se creó con el propósito de satisfacer la demanda de los estudiantes ávidos de conocimientos, que querían ingresar a esta carrera para obtener una profesión que les permitiera estar a la par con los avances de última tecnología. (ESPAM MFL, 2015a)

Atendiendo lo dispuesto en la ley de creación de la Escuela Superior Politécnica Agropecuaria de Manabí – ESPAM, su Estatuto y Ley de Educación Superior,

Consejo Politécnico de la ESPAM, el Ing. Leonardo Félix López rector de la ESPAM, informó al presidente del CONESUP Ing. Vinicio Baquero Ordoñez que con fecha 16 de diciembre del 2002 se creó la carrera de Informática, con la modalidad presencial en los predios de la institución, localizada en el Cantón Bolívar - Provincia Manabí (ESPAM MFL, 2015a).

Con el mismo oficio se le acompañó el estudio correspondiente, característica de la carrera modalidad presencial, títulos de Tecnólogos en Informática con seis semestres, y de Ingeniero en Informática con diez semestres, su diseño curricular centrado en el perfil profesional teniendo como norte la Misión y Visión Institucional. La misión de la carrera es: Formar profesionales con compromiso ético y social, que aporten soluciones computacionales, garantizándolo desde la docencia, investigación y vinculación. Y como visión: Ser un centro de referencia en la formación de profesionales que aporten innovaciones computacionales en el sector agroproductivo o de servicios (ESPAM MFL, 2015a).

La Carrera de Computación de la ESPAM MFL (2015b) responde a la necesidad de generar conocimientos que contribuyan a la trasformación de la matriz productiva haciendo uso de la industria tecnológica de servicio; software, hardware y servicios informáticos. Dentro de esta carrera se presentan varias propuestas de investigación que aportan a las líneas de investigación Institucional y de la carrera. Ante ello, presenta periódicamente propuestas de investigación de acuerdo a las convocatorias anuales que se dan desde la Coordinación General de Investigación; es así que en la VII CONVOCATORIA DE PROYECTOS I+D+i, participa con el proyecto titulado “Minería de datos aplicada al rendimiento académico de los estudiantes en la ESPAM MFL” del cual toma parte el presente trabajo de titulación.

## 1.2. DESCRIPCIÓN DE LA INTERVENCIÓN

Eckert y Suénaga (2013) expresan que en la actualidad la sociedad se encuentra en la época de la información, donde los datos se incrementan de forma considerable. El exceso de datos no siempre aumenta el conocimiento porque al procesarlos resultan difíciles, por lo que es conveniente poseer información necesaria y así incrementar su efectividad. Así Troche (2014) resalta la necesidad de almacenar información y menciona que la experiencia nos ha enseñado que esta información con el tiempo llega a ser un insumo importante para mejora o no cometer los mismos errores de tiempos pasados, es de acuerdo a este aspecto que existe la necesidad de analizar información histórica por lo que indica que ya teniendo este repositorio de información pre procesado y con un grado de consistencia muy alto, se puede aplicar distintas técnicas de análisis de información, es entonces que se pone en manifiesto la minería de datos (MD).

Según Pérez *et al.,* (2014) el objetivo principal de la minería de datos consiste en extraer información oculta de un conjunto de datos. En esta misma línea Jaramillo y Paz (2015) señalan que la minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, para así encontrar modelos a partir de los datos. La MD es la etapa de descubrimiento de conocimiento en bases de datos o también conocida como KDD por sus siglas en inglés Knowledge Discovery from databases (La Red *et al.,* 2015). Para mejor comprensión del proceso de la MD Natek y Zwilling (2014) explican los pasos para el análisis: primero, crear los datos conjuntos; segundo, definir la herramienta de minería de datos a utilizar; tercero, evaluar las técnicas de minería de datos; y cuarto, analizar los datos por cada modelo y elegir el mejor.

Riquelme *et al.,* (2016) complementa lo anterior y explica que la minería de datos es un campo interdisciplinario ya que existen numerosas áreas donde se puede aplicar, prácticamente en todas las actividades humanas que generen datos como, por ejemplo: Comercio y banca, Medicina y Farmacia, Seguridad y

detección de fraude, Recuperación de información no numérica, y Astronomía. García (2016) considera que algunas de las más importantes son: el análisis y fragmentación del mercado (marketing), la predicción bursátil y gestión de finanzas, la seguridad y auditoría informática, la optimización de procesos industriales, las diferentes ramas de la ingeniería, el análisis de tráfico y saturación de redes, las telecomunicaciones, las ciencias de la salud, las distintas aplicaciones en ciencias físicas y astrofísica, sistemas de aviso preventivo de fraudes y la clasificación de riesgos en seguros.

Rosado y Verjel (2016) mencionan que en el campo de la educación recientemente ha cobrado importancia la llamada la minería de datos que definen como el proceso de extraer conocimiento útil y comprensible previamente desconocido desde grandes cantidades de datos almacenados en distintos formatos, para así encontrar modelos a partir de los datos. Mientras Pereira *et al.,* (2013) señala que la minería de datos en la educación ha generado gran importancia, a su vez asevera que su estudio y aplicación ha sido muy relevante en los últimos años y que el uso de estas técnicas permite, entre otras cosas, predecir cualquier fenómeno dentro del ámbito educativo.

Menacho (2017) menciona que uno de los retos que tienen que enfrentar las instituciones de educación superior para ofrecer una mayor calidad educativa, es mejorar el rendimiento académico de los estudiantes. Así que García *et al.,* (2017) señala que el abandono de los estudios superiores es un fenómeno cuya magnitud se ha visto incrementada a la par que el número de estudiantes en la universidad lo cual a su vez supone un importante desaprovechamiento de recursos.

Existen varias investigaciones donde se han aplicado diferentes técnicas de minería de datos en la educación, dos proyectos similares son: el de Saleem *et al.,* (2015) que señala que una de las aplicaciones más comunes de la minería de datos es el uso de diferentes algoritmos y herramientas para estimar eventos futuros basados en experiencias previas, y que usando diferentes técnicas de

clasificación para construir un modelo de predicción de rendimiento basado en los registros académicos de los estudiantes anteriores se puede ayudar a los estudiantes en su selección de cursos. Al calcular la precisión promedio para cada algoritmo, se encontró que el algoritmo J48 logró un mejor rendimiento de precisión para permitir a los estudiantes obtener una predicción para uno o más cursos. Por otra parte, Sanvitha *et al.,* (2018) presenta un estudio de investigación que se llevó a cabo para determinar los posibles factores que afectan el rendimiento de los estudiantes en el sistema de educación superior, los métodos de clasificación utilizados en este análisis proporcionaron mejores resultados de la predicción que varía entre 92 y 98%. De una cantidad de 12 atributos: la edad del estudiante, la cantidad de módulos de falla y el rendimiento de semestres pasados fueron identificados como los factores más correlacionados que predicen la calificación final de los estudiantes.

Con base en los resultados de estas investigaciones y al beneficio para todas las partes interesadas en el sector de la educación las autoras se motivaron a contribuir al proyecto de investigación de la carrera el cual tiene como tema “Minería de datos aplicada al rendimiento académico de los estudiantes en la ESPAM MFL”, donde la presente investigación contribuye con dos de sus objetivos: Seleccionar la técnica adecuada de análisis de datos requerida y emplear técnicas de minería de datos; el propósito es utilizar las técnicas de minería de datos de clasificación ya que están siendo ampliamente aplicadas para predecir el rendimiento académico de los estudiantes, con la finalidad de detectar los factores que más influyen en su proceso de aprendizaje.

## 1.3. OBJETIVOS

### 1.3.1. OBJETIVO GENERAL

Aplicar los principales algoritmos de minería de datos utilizados en educación para realizar inferencias en la clasificación del rendimiento académico de los estudiantes de la carrera de computación de la ESPAM MFL.

### 1.3.2. OBJETIVOS ESPECÍFICOS

- Establecer los principales algoritmos de minería de datos aplicados en la educación.
- Determinar las variables y el método a utilizar para la experimentación.
- Preparar el conjunto de datos.
- Emplear las técnicas adecuadas de minería de datos en la educación.
- Evaluar los resultados de los modelos.

# CAPÍTULO II. DESARROLLO METODOLÓGICO DE LA INTERVENCIÓN

El presente trabajo de titulación se realizó con el propósito de determinar los principales algoritmos de minería de datos aplicados en la educación para realizar inferencias sobre el rendimiento académico de los estudiantes de la carrera de computación de la ESPAM MFL se utilizó la metodología CRISP-DM que es una de las metodologías más usadas en la actualidad para la generación de proyectos de minería de datos. Esta metodología estándar sirve para la construcción de proyectos de minería de datos con sus fases no necesariamente rígidas. Organiza el desarrollo de un proyecto de minería de datos en una serie de fases o etapas que permitan cumplir con los objetivos del proyecto (Chapman *et al.,* 2000).

Además, se utilizó el software WEKA que como menciona Moreno (2016), es un programa que está basado en un conjunto de librerías Java bajo licencia GPL, y ha sido desarrollado en la Universidad de Waikato, de ahí el nombre de WEKA (Waikato Environment for Knowledge Analysis). Está orientado a la extracción de conocimiento a través de bases de datos con gran cantidad de información, contiene una gran colección de algoritmos y herramientas para analizar los datos junto con una interfaz sencilla, que hace que el usuario pueda usar este software de manera muy simple.

Para el primer objetivo se empleó la fase de comprensión del negocio, que es la primera etapa del proceso investigativo que proporcionó el conocimiento de las diferentes técnicas de minería de datos descriptivas ya existentes, a través de una amplia búsqueda de información además gracias a esta búsqueda se determinaron los principales algoritmos de minería de datos aplicados a la educación realizando un cuadro resumen de estas técnicas.

Para el segundo objetivo se aplicó la fase de compresión de los datos para definir las variables a evaluar en los algoritmos y observar su naturaleza y efectos mediante una ficha técnica, a continuación, se diseñó un diagrama del método a utilizar en la experimentación.

Para el tercer objetivo se aplicó la fase preparación de los datos, que permitió la discretización del conjunto de datos, para adquirir una vista minable además presenta los algoritmos seleccionados mediante una ficha técnica.

El cuarto objetivo comprende la fase de modelado se aplicaron los algoritmos de la técnica de minería de datos seleccionada y después se aplicaron fórmulas estadísticas a los resultados de cada algoritmo.

Para el cuarto objetivo se aplicó la fase de evaluación, que permitió el análisis de los resultados de cada algoritmo, teniendo en cuenta el cumplimiento de cada una de las fases.

# CAPÍTULO III. DESCRIPCIÓN DE LA EXPERIENCIA

## 3.1. METODOLOGÍA CRISP-DM

Se utilizó la metodología CRISP-DM porque proporcionó una descripción normalizada del ciclo de vida de minería de datos en el análisis de datos, además sus fases permiten ir hacia adelante y hacia atrás siempre que fue necesario para tener mejores resultados en la investigación.

### 3.1.1. COMPRENSIÓN DEL NEGOCIO

Esta primera fase permitió cumplir con el primer objetivo planteado que radicó en determinar los principales algoritmos de minería de datos aplicados en educación, se integró algunas de las investigaciones científicas que proporcionaron información de las diferentes técnicas de minería de datos ya existentes, y de acuerdo con esto se pudo elaborar el cuadro resumen de estas técnicas que se muestra a continuación:

**Cuadro 3.1.** Cuadro Resumen de las Técnicas y Métodos de Minería de Datos aplicadas a la Educación

| AÑO | PROYECTO | MÉTODOS/TÉCNICAS DE MINERÍA DE DATOS | | ALGORITMO | HERRAMIENTAS | AUTORES |
|---|---|---|---|---|---|---|
| 2018 | A Data Mining Approach to Identify the Factors Affecting the Academic Success of Tertiary Students in Sri Lanka | Clasificación | Redes Bayesianas | Naïve Bayes | Software R | Sanvitha, K; Liyanage, S; Bhatt C. |
| | | | Árbol de decisión | C4.5 | | |
| | | | | Random Forest | | |
| 2017 | Predicción del rendimiento académico aplicando técnicas de minería de datos | Clasificación | Árbol de decisión | C4.5 | Weka | Menacho, C. |
| | | | | ID3 | | |
| | | | | CART | | |
| | | | | M5P | | |
| | | | | J48 | | |
| | | | Redes Bayesianas | Naïve Bayes | | |
| | | | | Bayes Net | | |
| | | | | TAN | | |
| | | | Redes Neuronales | Perceptrón multicapa | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | Back Propagación | | |
| | Educational Data Mining: Discovery Standards of Academic Performance by Students in Public High Schools in the Federal District of Brazil | Clasificación | Árbol de decisión | Máquina de Impulso de Gradiente (GBM) | H2O | Fernandes, E; Carvalho, R; Holanda, M; Van, G. |
| | Minería de datos aplicada para la identificación de factores de riesgo en alumnos | Asociación | Reglas de asociación | A priori | Weka | Reyes, A; Flores, A; Alejo, R; Rendon, E. |
| | | Agrupamiento | Métodos basados en casos y en vecindad | Kmeans | | |
| 2016 | Minería de datos en egresados de la Universidad de Caldas | Clasificación | Árbol de decisión | OneR | SQL | Bedoya, O; López, M; Marulanda, C. |
| | | | | J48 | | |
| | | | Redes Bayesianas | Stacking | | |
| | | | | Naïve Bayes | | |
| | Prediction of Student Dropout Using Personal Profile and Data Mining Approach | Clasificación | Árbol de decisión | J48 | Weka | Meedech, P; Iam-On, N; Boongoen, T. |
| | | | | SimpleCart | | |
| | | | | ADTree | | |
| | | | | RandomTree | | |
| | | | | REPTree | | |
| | | | Basados en reglas | JRip | | |
| | | | | OneR | | |
| | | | | Ridor | | |
| | Aplicación de la minería de datos de la educación en línea | Clasificación | Árbol de decisión | M5P | Cross-Platform, Platform-Specific, Ad Hoc Tools, Learning Analytics Tools, Learning Analytic Frameworks and Tools | Rosado, A y Verjel, A. |
| | | | | J48 | | |
| | | | | RepTree | | |
| 2015 | Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje | Clasificación | Árbol de decisión | CHAID | RapidMiner | Jaramillo, Paz. |
| | | | | ID3 | | |
| | | | | J48 | | |
| | Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos | Clasificación | Árbol de decisión | CART | SPSS Clementine v.9.0 | Alcover, R; Benlloch, P; Blesa, M; Calduch, M; Celma, C; Ferri, J; Hernández, L; Iniesta, J; Ramírez, A; Robles, J. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Mining Educational Data to Predict Students' Academic Performance | Clasificación | Árbol de decisión | J48 | NetBeans IDE, Weka | Saleem, M; Kathiry, N; Osimi, S; Badr, G. |
| | | | | ID3 | | |
| | Caracterización de la deserción estudiantil en educación superior con minería de datos. | Agrupamiento | Métodos basados en casos y en vecindad | Kmeans | Weka | Jiménez, J y Toledo, S. |
| | | Clasificación | Árbol de decisión | J48 | | |
| 2014 | Using Data Mining Techniques to Detect the Personality of Players in an Educational Game | Clasificación | Máquinas de vectores de soporte | | Weka | Keshtkar, F; Burkett, C; Li, H; Graesser, A. |
| | | | Redes Bayesianas | Naïve Bayes | | |
| | | | Árbol de decisión | J48 | | |
| 2013 | Data Mining and Social Network Analysis in the Educational Field: An Application for Non-Expert Users | Clasificación | Árbol de decisión | J48 | E-Learning Web Miner (ElWM) | García, D; Palazuelos, C; Zorrilla, M. |
| | Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD | Clasificación | Árbol de decisión | J48 | Weka | Eckert y Suénaga. |
| | | Agrupamiento | Métodos basados en casos y en vecindad | Kmeans | | |
| | | Asociación | Reglas de asociación | A priori | | |
| | Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil | Clasificación | Árbol de decisión | J48 | Weka | Pereira, R; Calderón, A; Jiménez, A. |
| | Mining for Marks: A Comparison of Classification Algorithms when Predicting Academic Performance to Identify "Students at Risk" | Clasificación | Árbol de decisión | J48 | Weka | Mashiloane, L; Mchunu, M. |
| | | | Redes Bayesianas | Naïve Bayes | | |
| | | | Tablas de decisión | Algoritmo de hallazgo | | |
| 2012 | Use of machine learning techniques for educational proposes: a decision support | Clasificación | Árbol de decisión | M5P | SPSS AHA | Kotsiantis, S. |
| | | | Redes Neuronales | Naïve Bayes | | |
| | | | Regresión Lineal | Regresión Lineal | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | system for forecasting students' grades | | Regresión Lineal localmente ponderado | Regresión Lineal localmente ponderado | | |
| | | | Máquinas de vectores de soporte | Optimización de mínimos secuencial (SMO) | | |
| 2011 | Aplicando minería de datos al marketing educativo | Agrupamiento | Métodos basados en casos y en vecindad | Kmeans | RapidMiner | Pinzón, L. |
| | A data mining approach to guide students through the enrollment process based on academic performance | Clasificación | Árbol de decisión | C4.5 | CRISP-DM methodology | Vialardi, C; Chue, J; Peche, JP y col |
| | | | Máquinas de vectores de soporte | KNN (k-vecinos más cercano) | | |
| | | | Redes Bayesianas | Naïve Bayes | | |
| | Accuracy Comparison of Predictive Algorithms of Data Mining: Application in Education Sector | Clasificación | Árbol de decisión | SOTA (algoritmo de árbol autoorganiza-dor) | Software de código abierto Knime Moodle - Logs del Sistema de Gestión de Aprendizaje de Código Abierto (LMS) | Sharma, M; Mavani, M. |
| | | | Redes Bayesianas | Naïve Bayes | | |
| 2010 | Classification and Prediction of Academic Talent Using Data Mining Techniques | Clasificación | Árbol de decisión | C4.5 | Weka y ROSETTA toolkit | Jantan, H; Hamdan, A; Othman, Z. |
| | | | | Random Forest | | |
| | | | Redes Neuronales | Perceptrón multicapa | | |
| | | | | Radial función de red básica | | |
| | | | Máquinas de vectores de soporte | KNN (k-vecinos más cercano) | | |
| | Minería de Datos en la Educación | Clasificación | Árbol de decisión | ID3 | Answer Tree, SPSS | Jiménez, G; Álvarez, H. |
| | | | | J48 | | |
| | Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales. Un caso de estudio en el norte de Coahuila | Agrupamiento | VARK | FarthestFirst | Moodle y Weka | Olague, J; Torres, S; Morales, F; Valdez, A; Silva, A. |
| 2009 | Minería de Datos aplicada al análisis de la deserción en la Carrera de | Clasificación | Árbol de decisión | Árbol de decisión binario | IBM DB2 Warehouse 9.5 | Pautsch, J. |
| | | | | ID3 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Analista en Sistemas de Computación. | | Redes Bayesianas | Naïve Bayes | | |
| | | Agrupamiento | Generación de clústeres | Algoritmo Kohonen | | |
| | Herramientas de Minería de Datos | Clasificación | Árbol de decisión | J48 | SPSS Clementine, Oracle Data Miner y Weka | Rodríguez, Y; Díaz, A. |
| 2008 | Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil | Clasificación | Árbol de decisión | FT<br>J48 | MS SQL Server<br>SPSS<br>Weka | Spositto, O; Etcheverry. M; Ryckeboer, H; Bossero, J. |
| | Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado | Clasificación | Reglas Difusas | Mamdani | Matlab | Domínguez, M. |
| 2007 | Minería de datos para descubrir estilos de aprendizaje | Agrupamiento | Generación de clústeres | FarthestFirs | Weka | Durán, E. y Costaguta, R. |
| 2005 | Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web | Agrupamiento | Redes de Kohonen | Algoritmo Kohonen | Sistemas Tutores Inteligentes (STI), Sistemas Hipermedia Adaptativos (SHA). | Romero, C; Ventura, S; Hervás, C. |
| | | Clasificación | Redes Bayesianas | Naïve Bayes | | |
| | Algoritmos evolutivos para descubrimiento de reglas de predicción en la mejora de sistemas educativos | Clasificación | Árbol de decisión | ID3 | EPRules (Education Prection Rules) | Romero, C; Ventura, S; Castro, C; García, E. |
| | | Asociación | Reglas de asociación | A priori | | |
| | | | Inducción de reglas | Prism | | |

El Cuadro 3.1, muestra varias investigaciones de minería de datos aplicada a la educación realizadas entre los años 2005-2018 y de acuerdo con estas investigaciones se pudo observar que la aplicación de estas técnicas ha cobrado mucha importancia en los últimos años. Se puede confirmar que la información bibliográfica recopilada predomina la técnica de clasificación como la más utilizada en los datos que genera la educación respecto a información académica, entre los cuales se encuentran los métodos de Árbol de decisión, Naïve Bayes y Redes neuronales.

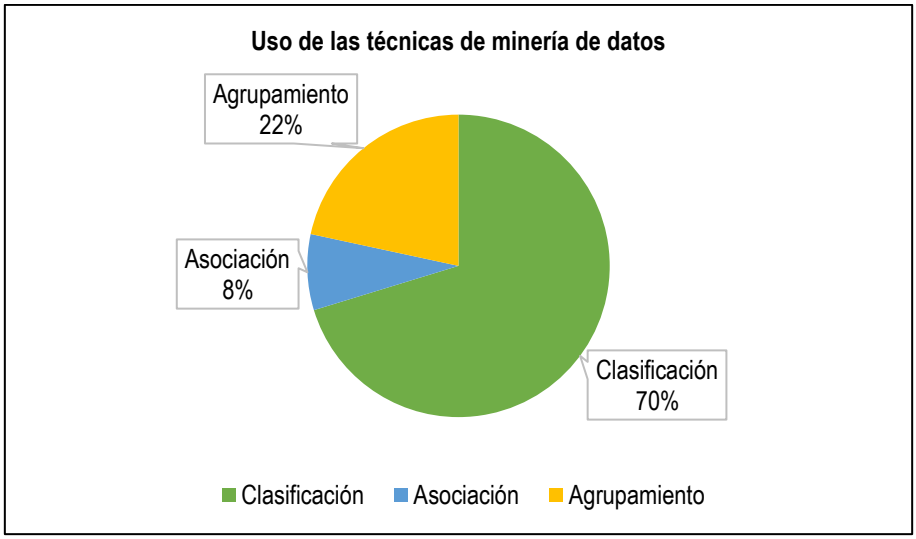



**Gráfico 3.1.** Resultados del uso de las Técnicas de Minerías de Datos aplicados en la educación.

El Gráfico 3.1, muestra los porcentajes de las técnicas que han sido utilizadas en las treinta revisiones bibliográficas de minería de datos educacional tomadas de referencia, de acuerdo con el gráfico se observa que la técnica de clasificación tiene el porcentaje más alto (70%) debido a su capacidad predictiva a partir de datos que se le proporcione, luego se encuentra la técnica de agrupamiento como la segunda más utilizada (22%) y por último la técnica de asociación (8%).

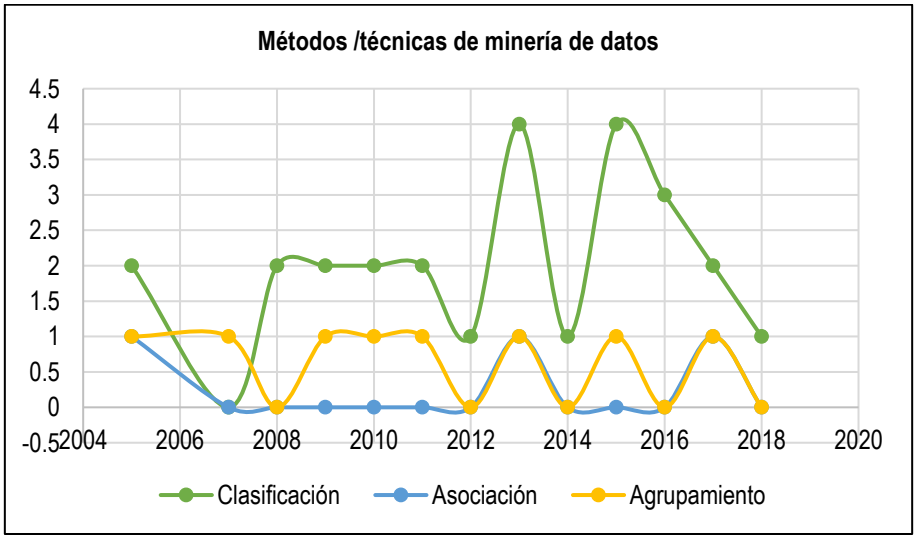



**Gráfico 3.2.** Uso de las Técnicas de Minería de Datos aplicadas en la Educación (2005-2018).

El Gráfico de dispersión 3.2, revela que entre los años 2005-2018 se ha experimentado más con la técnica de clasificación, dentro de la información bibliográfica se pudo notar que no existe una diversidad de algoritmos que

puedan ser usados para este tipo de investigaciones, es por eso que las técnicas de agrupación y asociación son las menos utilizadas debido a que muy pocos de sus algoritmos tienen prioridad para ser usados en datos académicos.

De acuerdo con el análisis planteado se realiza una categorización de los proyectos donde se aplican las técnicas de clasificación:

**Cuadro 3. 2.** Técnicas de Minería de Datos de Clasificación utilizadas en periodos del 2005-2017.

| PROYECTO | MÉTODOS /TÉCNICAS DE MINERÍA DE DATOS | | ALGORITMO | HERRAMIENTAS |
|---|---|---|---|---|
| Un enfoque de minería de datos para identificar los factores que afectan el éxito académico de los estudiantes terciarios en Sri Lanka | Clasificación | Redes Bayesianas | Naïve Bayes | Software R |
| | | Árbol de decisión | C4.5 | |
| | | | Random Forest | |
| Predicción del rendimiento académico aplicando técnicas de minería de datos | Clasificación | Árbol de decisión | C4.5 | Weka |
| | | | ID3 | |
| | | | CART | |
| | | | M5P | |
| | | | J48 | |
| | | Redes Bayesianas | Naïve Bayes | |
| | | | Bayes Net TAN | |
| | | Redes Neuronales | Perceptrón multicapa | |
| | | | Back Propagación | |
| Minería de Datos Educativos: Estándares de Descubrimiento del Desempeño Académico de Estudiantes en Escuelas Secundarias Públicas en el Distrito Federal de Brasil | Clasificación | Árbol de decisión | Máquina de Impulso de Gradiente (GBM) | H2O |
| Minería de datos en egresados de la Universidad de Caldas | Clasificación | Árbol de decisión | OneR | SQL |
| | | | J48 | |
| | | Redes Bayesianas | Stacking | |
| | | | Naïve Bayes | |
| Predicción de la deserción estudiantil mediante el perfil personal y el enfoque de minería de datos | Clasificación | Árbol de decisión | J48 | Weka |
| | | | SimpleCart | |
| | | | ADTree | |
| | | | RandomTree | |
| | | | REPTree | |
| | | Basados en reglas | JRip | |
| | | | OneR | |
| | | | Ridor | |
| Aplicación de la minería de datos de la educación en línea | Clasificación | Árbol de decisión | M5P | Cross-Platform, Platform-Specific, Ad Hoc Tools, Learning Analytics Tools, Learning Analytic Frameworks and Tools |
| | | | J48 | |
| | | | RepTree | |

| | | | | |
|---|---|---|---|---|
| Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje | Clasificación | Árbol de decisión | CHAID<br>ID3<br>J48 | RapidMiner |
| Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos | Clasificación | Árbol de decisión | CART | SPSS Clementine v.9.0 |
| Datos educativos mineros para predecir el rendimiento académico de los estudiantes | Clasificación | Árbol de decisión | J48<br>ID3 | NetBeans IDE, Weka |
| Caracterización de la deserción estudiantil en educación superior con minería de datos. | Clasificación | Árbol de decisión | J48 | Weka |
| Usar técnicas de minería de datos para detectar la personalidad de los jugadores en un juego educativo | Clasificación | Máquinas de vectores de soporte<br>Redes Bayesianas<br>Árbol de decisión | <br>Naïve Bayes<br>J48 | Weka |
| Minería de datos y análisis de redes sociales en el campo educativo: una aplicación para usuarios no expertos | Clasificación | Árbol de decisión | J48 | E-Learning Web Miner (EIWM) |
| Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD | Clasificación | Árbol de decisión | J48 | Weka |
| Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil | Clasificación | Árbol de decisión | J48 | Weka |
| Minería para marcas: una comparación de algoritmos de clasificación al predecir el rendimiento académico para identificar a "estudiantes en riesgo" | Clasificación | Árbol de decisión<br>Redes Bayesianas<br>Tablas de decisión | J48<br>Naïve Bayes<br>Algoritmo de hallazgo | Weka |
| Uso de técnicas de aprendizaje automático para propuestas educativas: un sistema de apoyo a la decisión para pronosticar las calificaciones de los estudiantes | Clasificación | Árbol de decisión<br>Redes Neuronales<br>Regresión Lineal<br>Regresión Lineal localmente ponderado<br>Máquinas de vectores de soporte | M5P<br>Naïve Bayes<br>Regresión Lineal<br>Regresión Lineal localmente ponderado<br>Optimización de mínimos secuencial (SMO) | SPSS<br>AHA |
| Un enfoque de minería de datos para guiar a los estudiantes a través del proceso de inscripción basado en el rendimiento académico | Clasificación | Árbol de decisión<br>Máquinas de vectores de soporte<br>Redes Bayesianas | C4.5<br>KNN (k-vecinos más cercano)<br>Naïve Bayes | CRISP-DM methodology |

| | | | | |
|---|---|---|---|---|
| Comparación de precisión de los algoritmos predictivos de la minería de datos: aplicación en el sector de la educación | Clasificación | Árbol de decisión | SOTA (algoritmo de árbol autoorganizador) | Software de código abierto Knime Moodle - Logs del Sistema de Gestión de Aprendizaje de Código Abierto (LMS) |
| | | Redes Bayesianas | Naïve Bayes | |
| Clasificación y predicción del talento académico mediante el uso de técnicas de minería de datos | Clasificación | Árbol de decisión | C4.5 | Weka y ROSETTA toolkit |
| | | | Random Forest | |
| | | Redes Neuronales | Perceptrón multicapa | |
| | | | Radial función de red básica | |
| | | Máquinas de vectores de soporte | KNN (k-vecinos más cercano) | |
| Minería de Datos en la Educación | Clasificación | Árbol de decisión | ID3 | Answer Tree, SPSS |
| | | | J48 | |
| Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. | Clasificación | Árbol de decisión | Árbol de decisión binario | IBM DB2 Warehouse 9.5 |
| | | | ID3 | |
| | | Redes Bayesianas | Naïve Bayes | |
| Herramientas de Minería de Datos | Clasificación | Árbol de decisión | J48 | SPSS Clementine, Oracle Data Miner y Weka |
| Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil | Clasificación | Árbol de decisión | FT | MS SQL Server SPSS Weka |
| | | | J48 | |
| Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado | Clasificación | Reglas Difusas | Mamdani | Matlab |
| Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web | Clasificación | Redes Bayesianas | Naïve Bayes | Sistemas Tutores Inteligentes (STI), Sistemas Hipermedia Adaptativos (SHA). |
| Algoritmos evolutivos para descubrimiento de reglas de predicción en la mejora de sistemas educativos | Clasificación | Árbol de decisión | ID3 | EPRules (Education Prection Rules) |

Como se muestra en el Cuadro 3.2, la mayoría de los proyectos de esta investigación hacen uso de las técnicas de clasificación y esto se debe a que estas técnicas brindan la función de pronosticar el rendimiento académico y detectar patrones respecto a grupos. Por tal motivo y la fuerte relación de cada

una de estas investigaciones se escoge de manera general la técnica de clasificación.

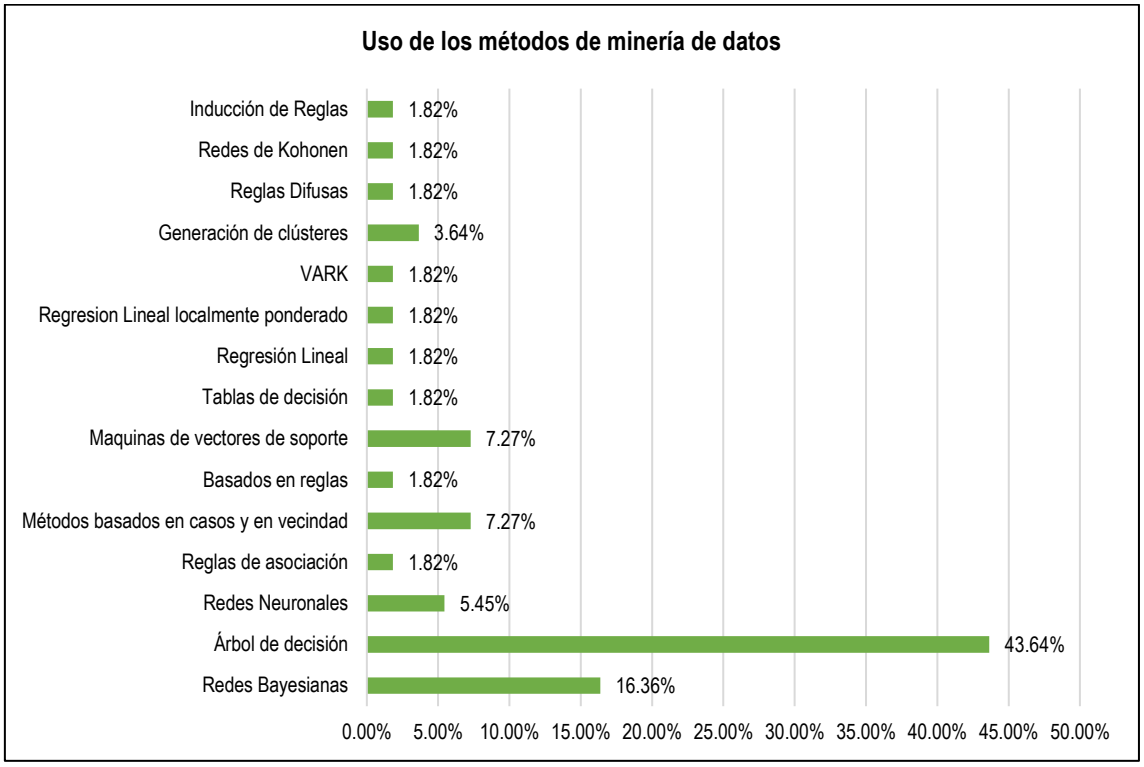



**Gráfico 3.3**. Uso de los métodos de minería de datos aplicados a la educación

El Gráfico 3.3, presenta los métodos que son más utilizados en la minería de datos referente a la educación, se puede observar que el método de árboles de decisión es el más usado (43.64%) para estos casos debido a que tiene varias soluciones para un mismo problema además permite analizar las posibles consecuencias de tomar una decisión; después de este método el más usado son las redes bayesianas.

Una vez analizado el uso de los métodos a continuación se muestra un gráfico donde se presenta el uso general de los algoritmos que constan en las investigaciones.

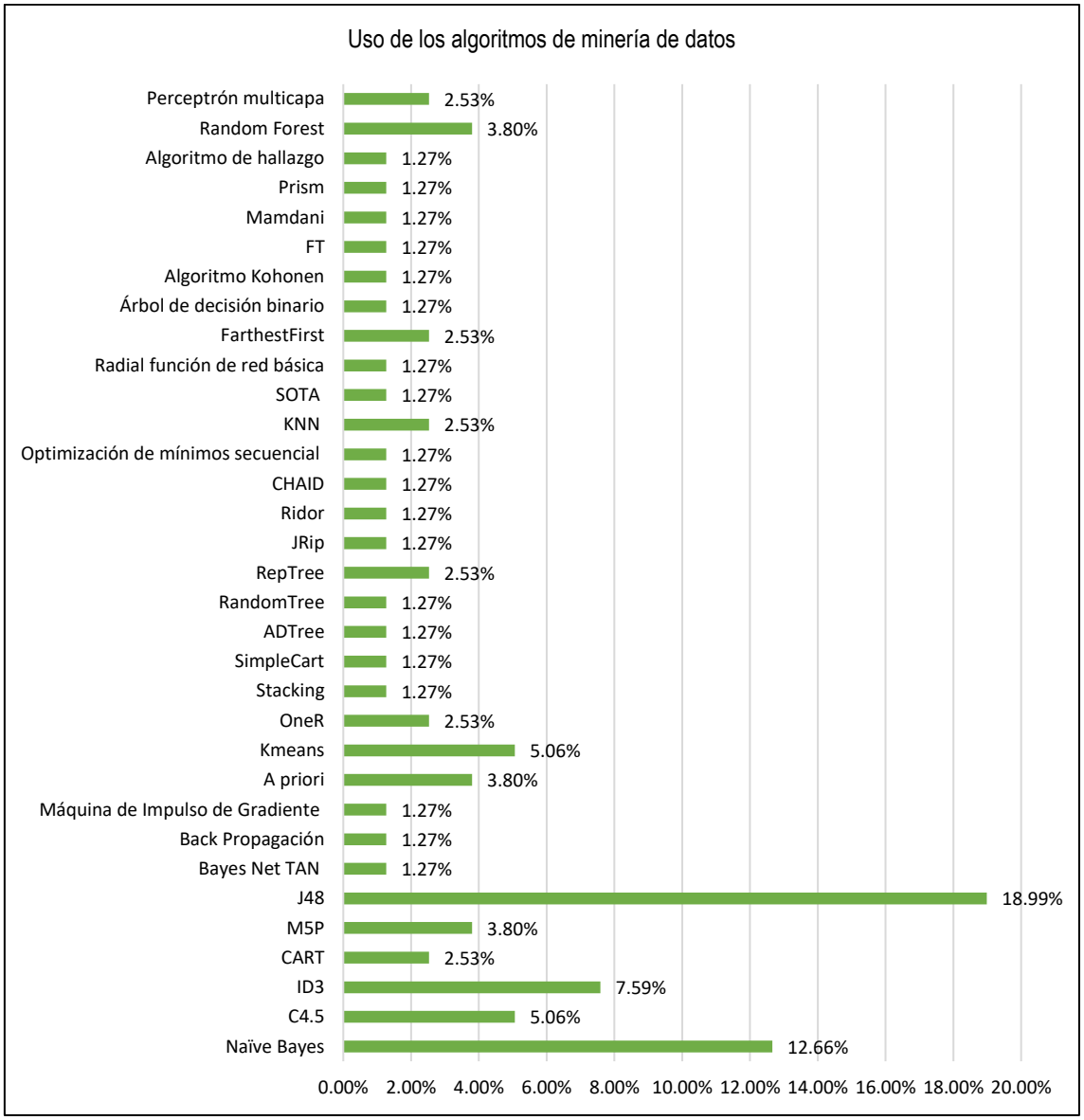



**Gráfico 3.4.** Uso de los algoritmos de minería de datos aplicados a la educación

En el Gráfico 3.4, se observa que los algoritmos con mayor uso son el J48 (18.99 %) como primera instancia perteneciente al método de árbol de decisión, seguido se encuentra Naïve Bayes perteneciente a redes bayesianas y en tercera instancia se encuentra el algoritmo de árbol de decisión ID3.

### 3.1.2. COMPRENSIÓN DE LOS DATOS

Para el cumplimiento de la segunda fase fue necesario realizar dos actividades. La primera consistió en realizar una ficha técnica de las variables a utilizar en la

aplicación de minería de datos, la cual se estructuró por un identificador, nombre de la variable, contenido, tipo y fuente de dato.

**Cuadro 3. 3.** Ficha técnica de la variable fecha de nacimiento

| | |
|---|---|
| **IDENTIFICADOR** | X1 |
| **NOMBRE** | Fecha de Nacimiento |
| **CONTENIDO** | El año en que nació la persona; puede incluir también el mes y el día de nacimiento de la persona. |
| **TIPO** | Date |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.4.** Ficha técnica de la variable genero

| | |
|---|---|
| **IDENTIFICADOR** | X2 |
| **NOMBRE** | Género |
| **CONTENIDO** | Termino especifico que distingue entre masculino y femenino. |
| **TIPO** | Varchar |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.5.** Ficha técnica de la variable ciudad

| | |
|---|---|
| **IDENTIFICADOR** | X3 |
| **NOMBRE** | Ciudad |
| **CONTENIDO** | Localización geográfica, donde radica una persona. |
| **TIPO** | Varchar |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.6.** Ficha técnica de la variable escuela secundaria

| | |
|---|---|
| **IDENTIFICADOR** | X4 |
| **NOMBRE** | Escuela Secundaria |
| **CONTENIDO** | Institución donde se efectúan los estudios secundarios de acuerdo a una especialidad. |
| **TIPO** | Varchar |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.7.** Ficha técnica de la variable tipo

| | |
|---|---|
| **IDENTIFICADOR** | X5 |
| **NOMBRE** | Tipo |
| **CONTENIDO** | Tipo de escuela donde realizó sus estudios secundarios el/la estudiante: fiscal, particular o fiscomisional. |
| **TIPO** | Varchar |

| | |
|---|---|
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.8.** Ficha técnica de la variable estado

| | |
|---|---|
| **IDENTIFICADOR** | X6 |
| **NOMBRE** | Estado |
| **CONTENIDO** | Situación en la que se encuentra un estudiante: aprobado o en recuperación. |
| **TIPO** | Varchar |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.9.** Ficha técnica de la variable nivel

| | |
|---|---|
| **IDENTIFICADOR** | X7 |
| **NOMBRE** | Nivel |
| **CONTENIDO** | Semestre en la que se encuentra legalmente matriculado un estudiante (primero a décimo). |
| **TIPO** | Varchar |
| **FUENTE DE DATO** | Base de datos Administración Académica |

**Cuadro 3.10.** Ficha técnica de la variable sub total

| | |
|---|---|
| **IDENTIFICADOR** | X8 |
| **NOMBRE** | Sub total |
| **CONTENIDO** | Nota de un estudiante antes del examen final. |
| **TIPO** | Decimal |
| **FUENTE DE DATO** | Base de datos Administración Académica |

La ficha técnica presenta las variables iniciales que fueron escogidas y consideradas características importantes para el rendimiento académico de los estudiantes, estas variables deberán ser exploradas en los datos para continuar con la siguiente fase.

La segunda actividad permitió establecer la secuencia del proceso de la metodología planteada mediante un diagrama de experimentación que se muestra a continuación:

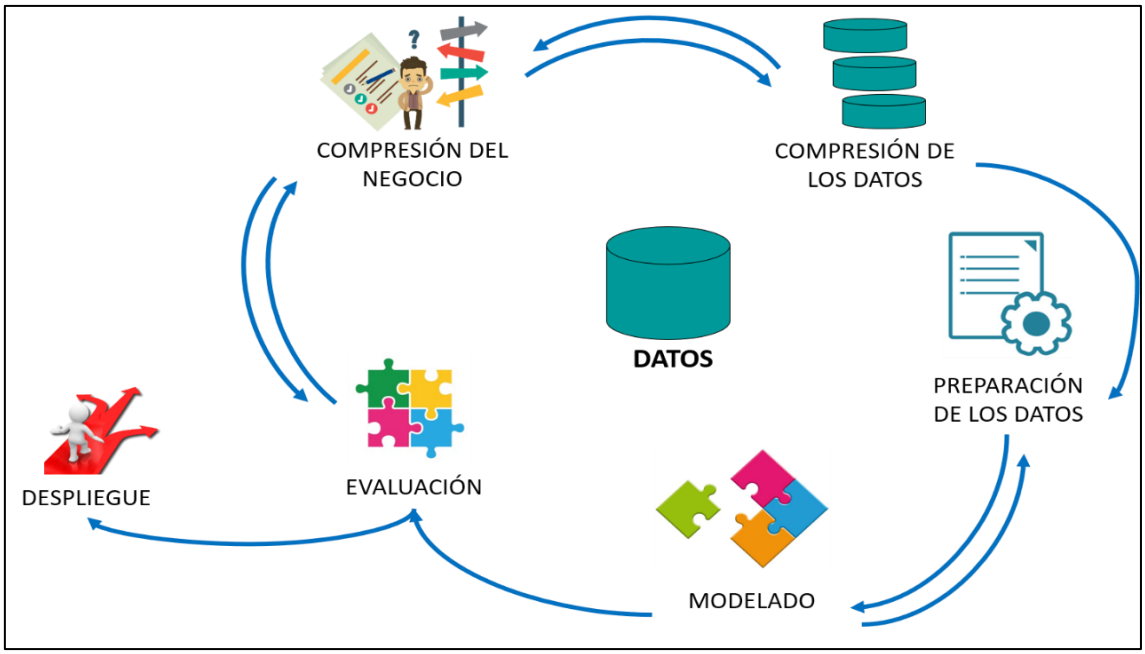



**Figura 3.1.** Diagrama de la metodología CRISP-DM
**Fuente:** Chapman *et al.,* 2000

Además, se elaboró el siguiente diagrama para especificar la aplicación de las técnicas de minería de datos desde la entrada de los datos hasta los resultados.

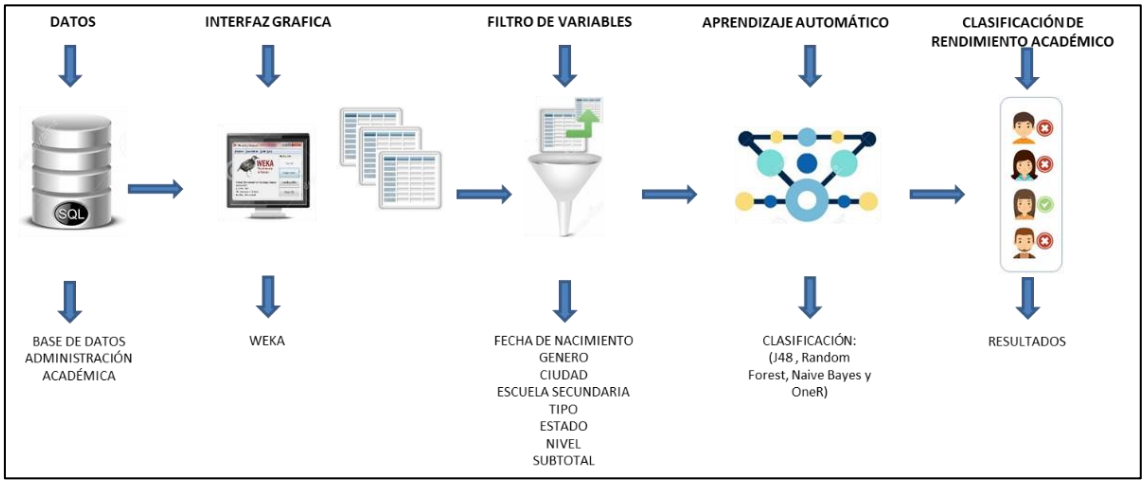



**Figura 3.2.** Diagrama de aplicación de las técnicas de minería de datos en la educación
**Fuente:** Las autoras

### 3.1.3. PREPARACIÓN DE LOS DATOS

Para empezar con la aplicación de la técnica de minería de datos se procedió a la preparación del conjunto de datos, actividad que es parte del tercer objetivo. En el cual se encuentran los 1164 registros que existen en la carrera de

computación de los semestres profesionalizantes de los últimos cinco años, almacenados en la base de datos Administración Académica.

Con el conjunto de registro, se procedió a la generación del conjunto de datos con la información correspondiente a la carrera de computación además, se realizó una tipificación de datos a las variables independientes y se discretizó el rendimiento académico (variable dependiente o clase) en 4 categorías: Insuficiente (calificados con nota menor a 7), Aceptable (calificados con nota menor a 8), Bueno (calificados con nota menor a 9) y excelente (calificados con nota mayor o igual a 9) para hacer esta discretización y tener resultados con mejor precisión se utilizó la nota global la cual es dependiente del Subtotal y el Examen final, se creó esta variable dependiente debido a que es necesario una selección de segmentos con características diferentes que correspondan a una clase que los diferencie (Cuadro 3.11).

**Cuadro 3.11**. Cuadro de variables y reglas aplicadas para la tipificación de los datos

| VARIABLE | DESCRIPCIÓN | VALORES |
|---|---|---|
| **X1** | Fecha de Nacimiento | Si Fecha >=1970 Y <= 1975 = 1;<br>Si Fecha >=1976 Y <=1980 = 2;<br>Si Fecha >=1981 Y <=1985 = 3;<br>Si Fecha >=1986 Y <=1990=4;<br>Si Fecha >1991 =5 |
| **X2** | Género | Si Genero = Masculino = 1;<br>Si Genero = Femenino =2 |
| **X3** | Ciudad | Si Ciudad = Bolívar = 1;<br>Si Ciudad = Chone = 2;<br>Si Ciudad = Junín = 3;<br>Si Ciudad = Manta = 4;<br>Si Ciudad = Pedernales;<br>Si Ciudad = Portoviejo;<br>Si Ciudad = Sucre;<br>Si Ciudad = Tosagua;<br>Si Ciudad = False = 9 |
| **X4** | Escuela Secundaria | Si Secundaria = Agropecuaria = 1;<br>Si Secundaria = Ciencias = 2;<br>Si Secundaria = Administración = 3;<br>Si Secundaria = Informática = 4;<br>Si Secundaria = Secretariado = 5;<br>Si Secundaria = Falso = 6 |
| **X5** | Tipo de Escuela Secundaria | Si Tipo = Fiscal = 1;<br>Si Tipo = Particular = 2;<br>Si Tipo = Fiscomisional = 3 |

| | | |
|---|---|---|
| **X6** | Estado | Si Estado = Aprobado = 1;<br>Si Estado = Reprobado =2;<br>Si Estado = Recuperación= 3 |
| **X7** | Semestre | Si Semestre = Quinto = 5;<br>Si Semestre = Sexto = 6;<br>Si Semestre = Séptimo = 7;<br>Si Semestre = Octavo= 8;<br>Si Semestre = Noveno = 9 |
| **X8** | Nota Subtotal (Nota parcial) | Si Subtotal < 5= 1;<br>Si Subtotal >= 5 Y <= 6 = 2;<br>Si Subtotal >= 6 Y <= 7 = 3;<br>Si Subtotal >= 7 = 4 |
| **Clase** | Rendimiento Académico (Atributos Clasificadores) | Si Nota Global < 7 = Insufficient;<br>Si Nota Global < 8 = Acceptable;<br>Si Nota Global < 9 = Good;<br>Si Nota Global >9 = Excellent |

El software que será utilizado para la aplicación de las técnicas de minería de datos propuestas es WEKA, debido a que esta herramienta trabaja con archivos en formato .csv se realizó una tipificación del conjunto de datos de origen .xlsx y se guardó en formato .csv como se muestra en el Anexo 2.

Para un mejor análisis de los registros que contiene el conjunto de datos a continuación se presenta un resumen de la cantidad de datos por cada variable:

**Cuadro 3.12.** Cantidad de registros por variables de acuerdo con su categorización

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **X1** | 17 | 38 | 134 | 617 | 358 | | | | | | **1164** |
| **X2** | 607 | 557 | | | | | | | | | **1164** |
| **X3** | 659 | 293 | 0 | 22 | 12 | 48 | 23 | 37 | 70 | | **1164** |
| **X4** | 17 | 330 | 75 | 692 | 14 | 36 | | | | | **1164** |
| **X5** | 966 | 137 | 61 | | | | | | | | **1164** |
| **X6** | 1141 | 20 | 3 | | | | | | | | **1164** |
| **X7** | | | | | 74 | 103 | 113 | 373 | 242 | 259 | **1164** |
| **X8** | 13 | 125 | 487 | 539 | | | | | | | **1164** |
| **Clase** | 377 | 411 | 353 | 23 | | | | | | | **1164** |

En el cuadro 3.12, el encabezado muestra el valor que se asignó en la tipificación a cada una de las variables para diferenciar la cantidad de datos que corresponden a cada valor, además al final se observa cuantos registros pertenecen a cada segmento clase, por ejemplo, existen 377 registros para la clase Excelente, 411 pertenecientes a Bueno, 353 Aceptable y 23 Insuficiente.

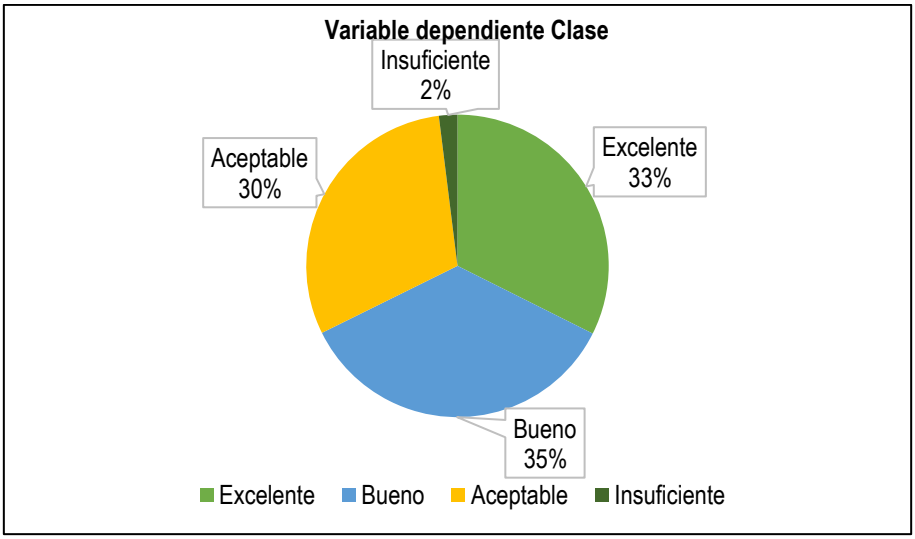



**Gráfico 3.5.** Total, de registros por categoría o clase

La variable clase representa el conjunto de datos dividido en 4 categorías, dando como resultado de un total de 1164 instancias (100%), un 35% para la categoría Bueno, un 33% para Excelente, un 30% para Aceptable y un 2% para insuficiente (Gráfico 3.5).

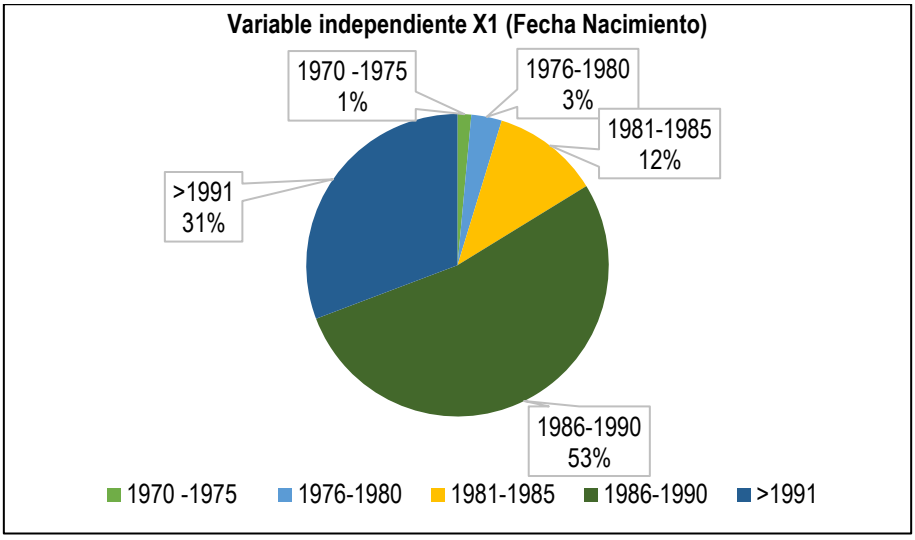



**Gráfico 3.6.** Porcentaje de estudiantes nacidos desde 1970 hasta 1991.

La variable que corresponde a la fecha de nacimiento se dividió en grupos de rangos dando como resultado que el mayor índice de estudiantes (53%) de los periodos académicos establecidos nacieron entre 1986 y 1990 (Gráfico 3.6).

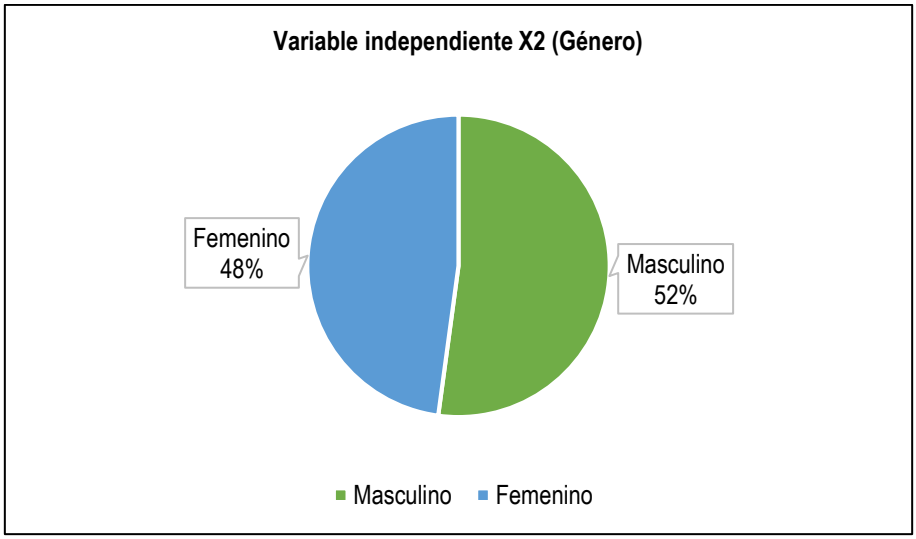



**Gráfico 3. 7.** Porcentaje de género masculino y femenino existentes en el conjunto de datos

Según la estadística existen más estudiantes del género masculino (52%), mientras que el género femenino tiene muy poca diferencia en porcentaje (48%) (Gráfico 3.7).

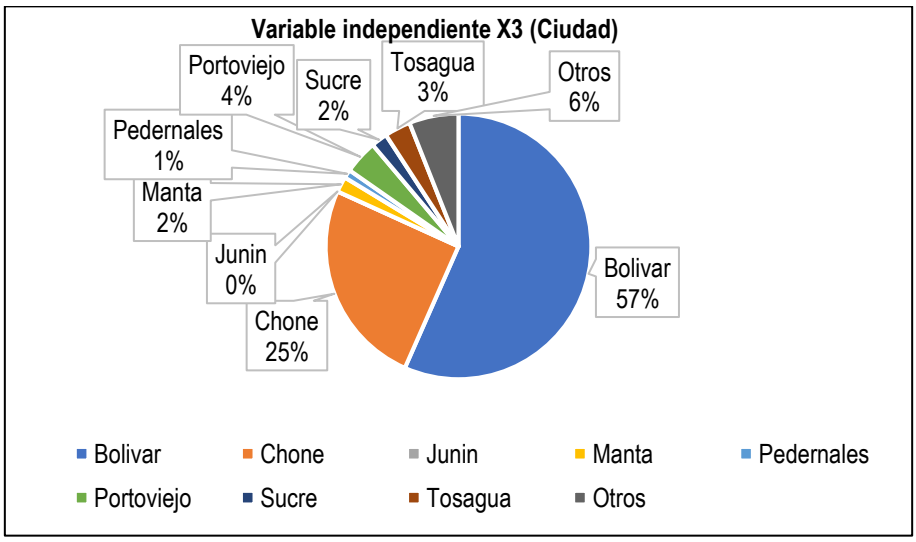



**Gráfico 3.8.** Porcentaje de lugar de origen de los estudiantes

Según los registros del conjunto de datos el 57% de los estudiantes pertenecen al cantón Bolívar, y otro gran porcentaje pertenece al cantón Chone (25%) así lo demuestra el Gráfico 3.8.

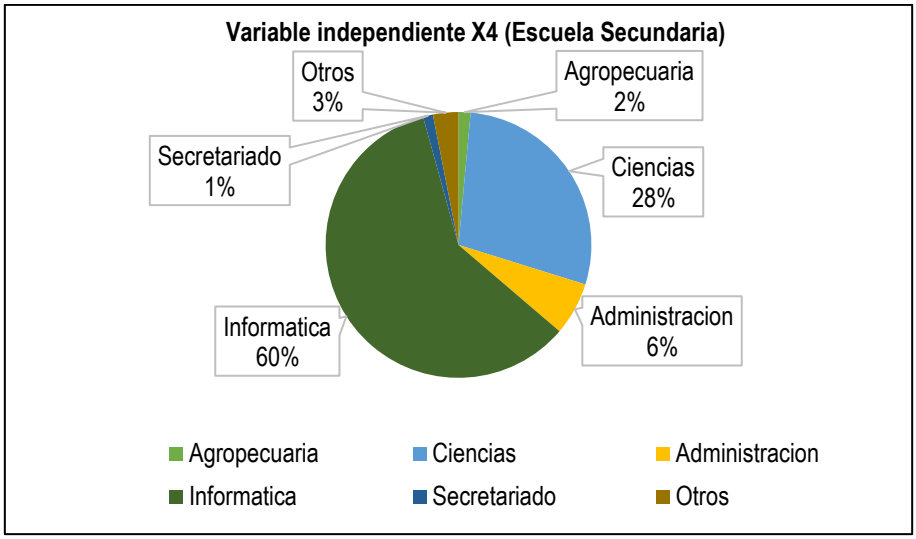



**Gráfico 3.9.** Porcentaje de especialidad de escuelas secundaria

De acuerdo al Gráfico 3.9, los registros del conjunto de datos el 60% de los estudiantes terminaron sus estudios secundarios en la especialidad de informática, y otro porcentaje considerable pertenece a la especialidad de ciencias (28%).

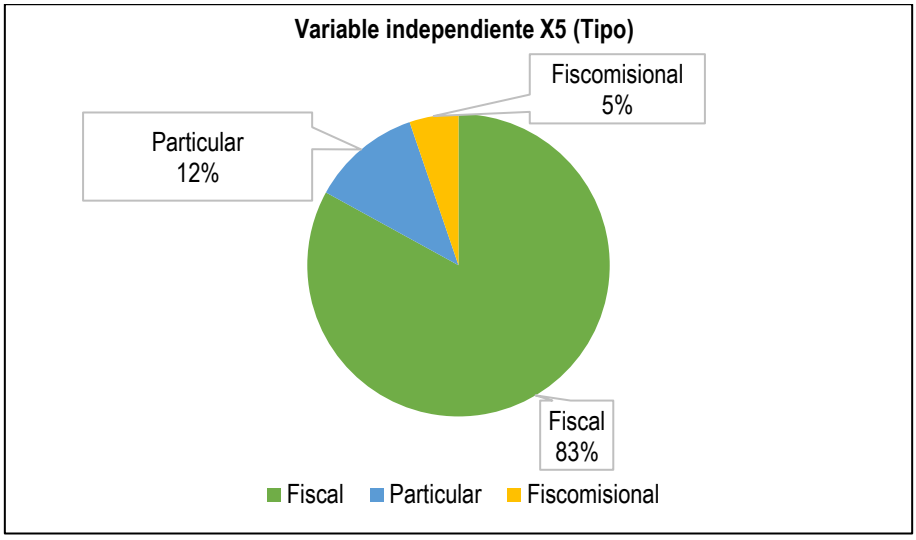



**Gráfico 3.10.** Porcentaje de tipo de escuela secundaria

Otra variable analizada es el tipo de escuela secundaria, el análisis estadístico indica que el 83% de los estudiantes finalizaron sus estudios secundarios en una institución fiscal, mientras que otro pequeño porcentaje finalizaron en instituciones Particulares (12%) y Fiscomisionales (5%) así lo demuestra el Gráfico 3.10.

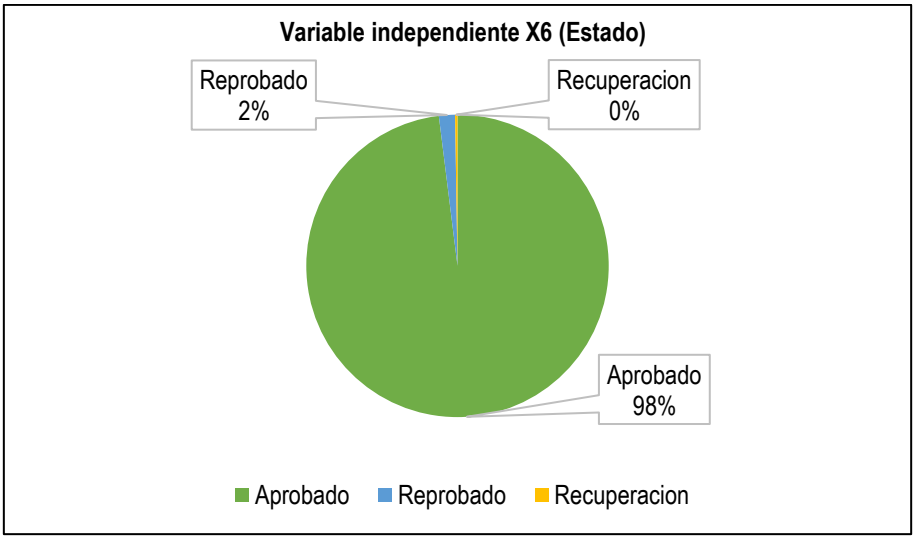



**Gráfico 11.** Porcentaje de estado académico de los estudiantes

De acuerdo con el conjunto de datos el 98% de los estudiantes han aprobado los semestres profesionalizantes y un 2% ha reprobado alguna vez, así lo demuestra los resultados estadísticos presentes en el Gráfico 3.11.

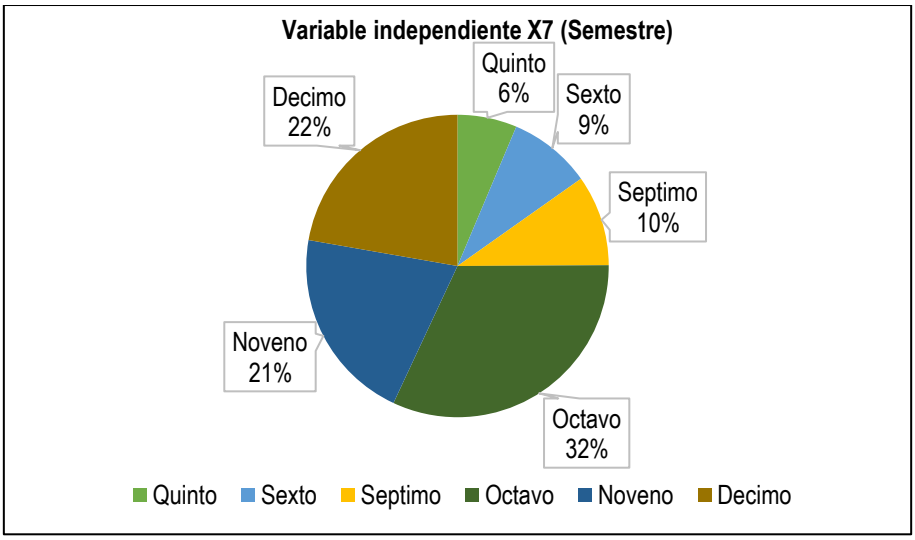



**Gráfico 3.12.** Porcentaje de estudiantes por semestre

El Gráfico 3.12, muestra el porcentaje de registros que existen por cada semestre en el conjunto de datos, al ser semestres con materias profesionalizantes el porcentaje varía debido que hay mayor riesgo de reprobar.

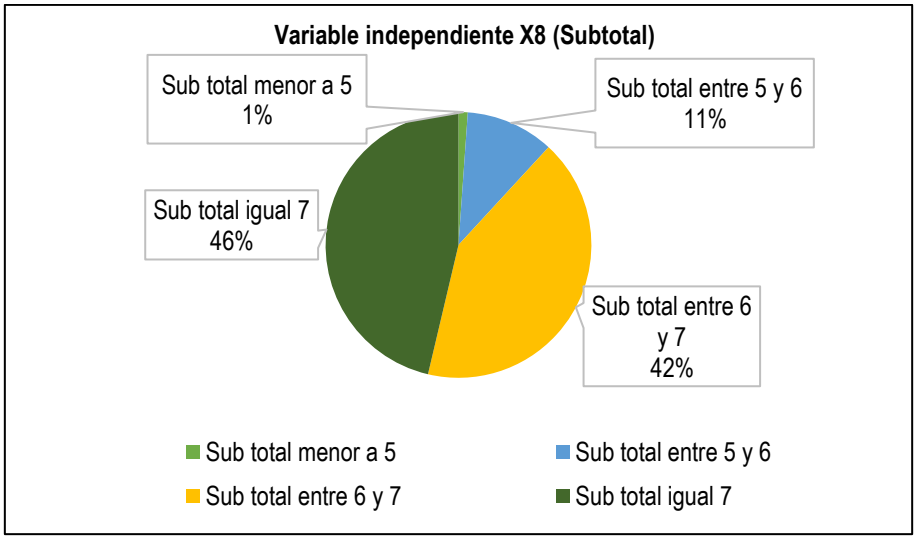



**Gráfico 3.13.** Porcentaje de promedio subtotal

El subtotal que equivale a la nota promedio antes del examen final estaba valorado sobre 8 puntos como nota máxima y 5 como nota mínima, el análisis estadístico del Gráfico 3.13, muestra que el 46% de estudiantes obtuvieron una nota mayor a 7, y el 42% obtuvo una nota entre 6 y 7 puntos siendo estas consideradas aceptables para aprobar el semestre.

Una vez que se analizaron las variables se realizó una ficha técnica de los algoritmos seleccionados la cual comprende: el método al que pertenece el algoritmo, el nombre del algoritmo, la descripción y los autores de la descripción.

**Cuadro 3.13.** Ficha Técnica de los algoritmos aplicados

| MÉTODO DE MINERÍA DE DATOS | ALGORITMO | DESCRIPCIÓN | AUTOR(ES) |
|---|---|---|---|
| Árboles | J48 | Este algoritmo genera un árbol de decisión estadístico. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. Cuando todas las muestras en la lista pertenecen a la misma clase, se crea un nodo de hoja para el árbol de decisión. | Díaz, H; Alemán, Y; Cabrera, L; Morales, A; Chávez, M; Casas, G. |
| Bayes | Naïve Bayes | Este es un algoritmo muy usado en procesos de clasificación; por su efectividad en el aprendizaje inductivo se lo considera como uno de los más eficientes dentro de la minería de datos. Este algoritmo trabaja basado en la hipótesis de que todos los atributos son independientes entre sí, claro está, si el valor de la variable clase es conocido. | Bedoya, O; López, M; Marulanda, C. |

| | | | |
|---|---|---|---|
| Reglas | OneR | Este algoritmo tiene como particularidad la selección del atributo que mejor revela la clase de salida. Las características propias de este método de clasificación de 117 resumen en su rapidez y buenos resultados, en contraste con otros algoritmos más complejos | Bedoya, O; López, M; Marulanda, C. |
| Árboles | Ramdom Forest | Está entre los métodos de clasificación supervisada más utilizados. Se trata de un algoritmo robusto y fácil de interpretar. Funciona haciendo particiones sucesivas en el espacio de variables buscando siempre la variable y el valor umbral de la misma que maximizan la homogeneidad de las particiones resultantes. | García, F; Sarria, F; Castillo, G. |

## 3.1.4. MODELADO

### 3.1.4.1. PROCESAMIENTO DE LOS DATOS

El procesamiento de los datos se realiza en la pantalla inicial que se muestra al abrir el archivo con WEKA (Figura 3.3 y 3.4) donde el conjunto de datos tiene 1164 instancias y 9 atributos anteriormente comentados.

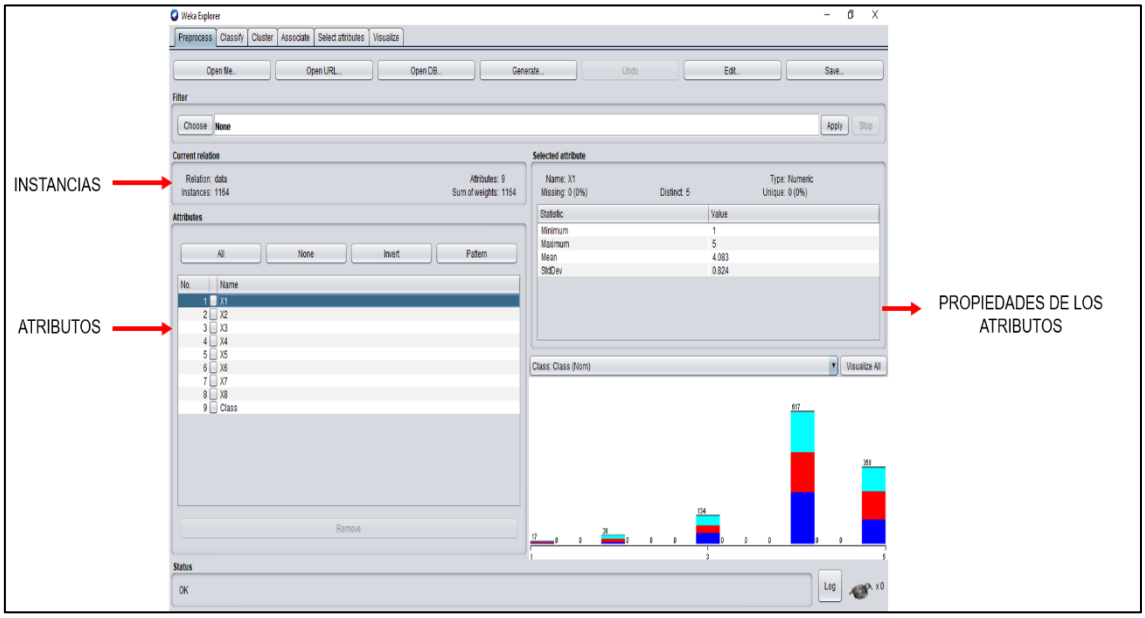



**Figura 3.3.** Base de datos en WEKA

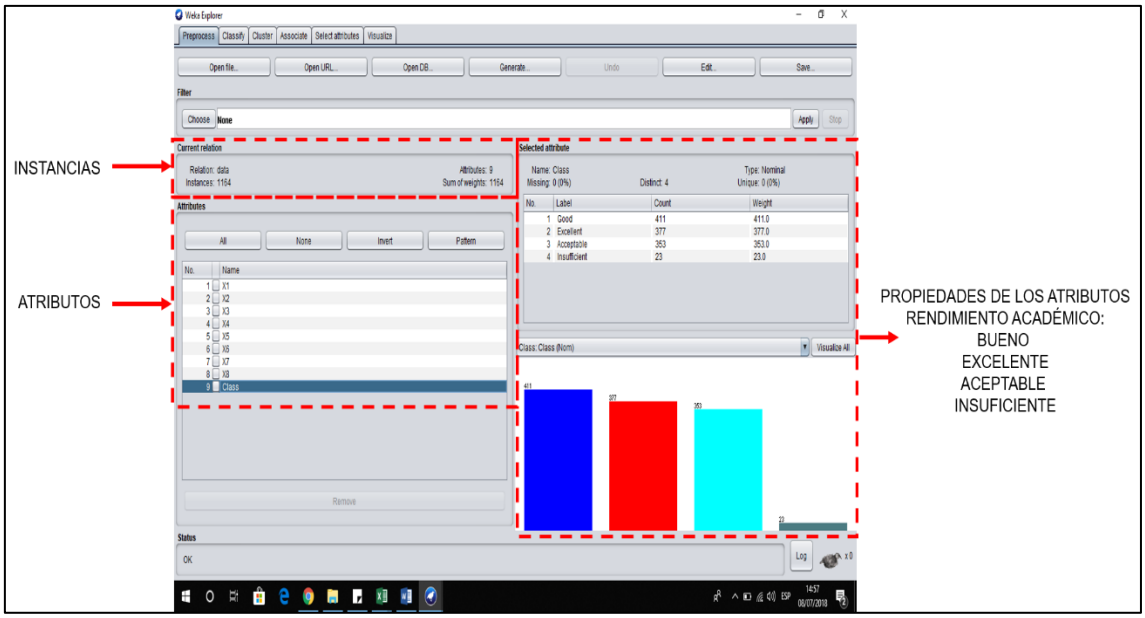



**Figura 3.4.** Base de datos en WEKA atributo Class

En esta pantalla al seleccionar uno de los atributos podemos observar en la parte derecha las propiedades de los atributos de forma estadística y de forma gráfica, en los datos estadísticos se presentan los valores máximos y mínimos que alcanzan cada uno de los atributos, además también se presenta su media y desviación típica, en la parte grafica están representados en colores para este caso un color para bueno, excelente, aceptable e insuficiente como se muestra en la Figura 3.4.

### 3.1.4.2. ANÁLISIS

Los algoritmos seleccionados para la aplicación de la técnica de clasificación son: J48, Naïve Bayes, Random Forest y OneR, para la ejecución de los algoritmos es necesario dividir el conjunto de datos en datos de prueba y entrenamiento, para lo cual se emplea la validación cruzada de K iteraciones o K-fold cross-validation, para este caso en cada algoritmo los datos se dividen en 10 iteraciones es decir un subconjunto se utiliza como datos de prueba y los restantes como datos de entrenamiento lo que al final genera una media aritmética de los resultados de las 10 iteraciones y presenta un único resultado por cada modelo clasificador.

Para una mejor compresión de los resultados se utiliza una matriz de evaluación que comprende la tasa de aciertos que indica el nivel de precisión de la clasificación de cada modelo, el estadístico de kappa que muestra el coeficiente de concordancia de las variables, la F-medida la cual mide la bondad de confiabilidad del modelo y por último el área ROC (Receiver Operating Characteristic, o Característica operativa del receptor) que indica la exactitud global de la prueba.

Para la determinación de aceptabilidad de los criterios de evaluación se entiende que la valoración máxima será de 1 y la mínima de 0.50, además Landis y Koch (1977) citado por Menacho (2017), proponen la escala de la valoración de k que se presenta a continuación:

**Cuadro 3.14.** Valoración del coeficiente Kappa.

| Kappa | Grado De Concordancia |
|---|---|
| < 0,00 | Sin Acuerdo |
| >0,00 - 0,20 | Insignificante |
| 0,21 - 0,40 | Discreto |
| >0,41 - 0,60 | Moderado |
| 0,61 - 0,80 | Sustancial |
| 0,81 - 1,00 | Casi Perfecto |

A continuación, se presenta la ejecución y resultados obtenidos por cada uno de los algoritmos:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     data
Instances:    1164
Attributes:   9
              X1
              X2
              X3
              X4
              X5
              X6
              X7
              X8
              Class
Test mode:    10-fold cross-validation
```

**Figura 3.5.** Información general del algoritmo clasificador J48

Como se muestra en la Figura 3.5, el algoritmo empieza mostrando el nombre del mismo, la relación, el número de instancias y atributos que tienen los datos

así mismo su identificador. A continuación, se presenta el algoritmo funcionando con los atributos de entrada:

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

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

X8 <= 3
|   X6 <= 1
|   |   X8 <= 2: Aceptable (115.0/3.0)
|   |   X8 > 2
|   |   |   X5 <= 1
|   |   |   |   X7 <= 8
|   |   |   |   |   X7 <= 6
|   |   |   |   |   |   X7 <= 5
|   |   |   |   |   |   |   X3 <= 1: Aceptable (13.0/5.0)
|   |   |   |   |   |   |   X3 > 1: Bueno (5.0/1.0)
|   |   |   |   |   |   X7 > 5: Bueno (49.0/14.0)
|   |   |   |   |   X7 > 6
|   |   |   |   |   |   X1 <= 2: Aceptable (8.0)
|   |   |   |   |   |   X1 > 2
|   |   |   |   |   |   |   X7 <= 7: Aceptable (61.0/16.0)
|   |   |   |   |   |   |   X7 > 7
|   |   |   |   |   |   |   |   X3 <= 7: Bueno (109.0/46.0)
|   |   |   |   |   |   |   |   X3 > 7: Aceptable (14.0/4.0)
|   |   |   |   |   X7 > 8
|   |   |   |   |   |   X4 <= 4
|   |   |   |   |   |   |   X1 <= 4: Bueno (90.0/27.0)
|   |   |   |   |   |   |   X1 > 4
|   |   |   |   |   |   |   |   X7 <= 9
|   |   |   |   |   |   |   |   |   X2 <= 1: Aceptable (10.0/3.0)
|   |   |   |   |   |   |   |   |   X2 > 1
|   |   |   |   |   |   |   |   |   |   X4 <= 3: Aceptable (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   X4 > 3: Bueno (15.0/6.0)
|   |   |   |   |   |   |   |   X7 > 9
|   |   |   |   |   |   |   |   |   X2 <= 1: Bueno (2.0)
|   |   |   |   |   |   |   |   |   X2 > 1: Aceptable (5.0/2.0)
|   |   |   |   |   X4 > 4: Aceptable (7.0/2.0)
|   |   |   |   X5 > 1
|   |   |   |   |   X4 <= 3
|   |   |   |   |   |   X7 <= 9: Aceptable (43.0/18.0)
|   |   |   |   |   |   X7 > 9: Bueno (7.0/1.0)
|   |   |   |   |   X4 > 3
|   |   |   |   |   |   X3 <= 5
|   |   |   |   |   |   |   X1 <= 4: Aceptable (31.0/12.0)
|   |   |   |   |   |   |   X1 > 4: Bueno (4.0/1.0)
|   |   |   |   |   |   X3 > 5: Bueno (10.0/3.0)
|   X6 > 1: Insuficiente (23.0)
X8 > 3
|   X7 <= 7
|   |   X7 <= 5
|   |   |   X3 <= 5: Excelente (36.0/10.0)
|   |   |   X3 > 5: Bueno (3.0)
|   |   X7 > 5
|   |   |   X3 <= 6
|   |   |   |   X3 <= 2
|   |   |   |   |   X1 <= 3: Excelente (2.0)
|   |   |   |   |   X1 > 3
|   |   |   |   |   |   X4 <= 3
|   |   |   |   |   |   |   X2 <= 1: Excelente (5.0/2.0)
|   |   |   |   |   |   |   X2 > 1: Aceptable (3.0/1.0)
|   |   |   |   |   |   X4 > 3
|   |   |   |   |   |   |   X4 <= 5: Bueno (41.0/14.0)
|   |   |   |   |   |   |   X4 > 5: Excelente (5.0/1.0)
|   |   |   |   X3 > 2: Bueno (3.0/1.0)
|   |   |   X3 > 6: Excelente (3.0)
|   X7 > 7: Excelente (438.0/121.0)

Number of Leaves  :     31

Size of the tree :      61
```

**Figura 3.6.** Modelo clasificador J48

Como se observa en la Figura 3.6, el árbol tiene como nodo raíz el atributo X8 (Subtotal) es decir de este atributo depende la salida que tenga a cada raíz haciendo uso de los demás atributos, el árbol de decisión ha demostrado que existe una fuerte relación significativa entre las variables X8 (Sub Total), X7 (Semestre) y X6 (Estado).

También se puede observar que el número de hojas para este árbol fue de 31 teniendo un tamaño de 61.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         787               67.6117 %
Incorrectly Classified Instances       377               32.3883 %
Kappa statistic                          0.5226
Mean absolute error                      0.2035
Root mean squared error                  0.3282
Relative absolute error                 59.9994 %
Root relative squared error             79.7055 %
Total Number of Instances             1164
```

**Figura 3.7.** Resultado general del algoritmo clasificador J48

Como resultado del algoritmo la Figura 3.7, muestra que se han clasificado 787 instancias correctas (67,61 %) y 377 instancias incorrectas (32,38%), de un total de 1164, lo que indica que el rendimiento académico de alrededor de 7 de cada 10 estudiantes se encuentra clasificados correctamente, además presenta que el coeficiente del estadístico de Kappa tiene un grado de concordancia moderado de 0,52, y si analizamos por cada una de las clases tenemos:

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
                 0,496     0,203     0,571       0,496    0,531       0,682     Bueno
                 0,923     0,175     0,716       0,923    0,806       0,877     Excelente
                 0,601     0,106     0,711       0,601    0,651       0,877     Aceptable
                 1,000     0,000     1,000       1,000    1,000       1,000     Insuficiente
Weighted Avg.    0,676     0,161     0,669       0,676    0,666       0,811
```

**Figura 3.8.** Resultado individual por clase del algoritmo clasificador J48

- La clase Bueno, presenta un TP Rate (Tasa de verdaderos positivos) de 0,49 aciertos, una FP Rate (Tasa de falsos positivos) de 0,20 lo que se

designa como un nivel de error elevado, una precisión de 0,57 lo que indica un nivel instancias correctamente clasificadas ligeramente bajo, una Recall (Cobertura) de 0,49, la bondad del modelo presentada en la F-medida se interpreta como baja debido a que el 0,53 se encuentra muy lejos del valor máximo de 1 y mientras el valor se aproxime más al 1 mayor será la confiabilidad, y el nivel ROC (característica operativa del receptor) Area tiene un nivel exactitud aceptable debido a que presenta un valor próximo al nivel máximo de 1 este es de 0,68 (Figura 3.8).

- La clase Excelente, presenta una tasa de verdaderos positivos con un nivel de aciertos elevado de 0,92, una tasa de falsos positivos de 0,17 lo que se designa como un nivel de error bajo, una precisión de 0,71 lo que indica un nivel instancias correctamente clasificadas muy bueno, una cobertura de 0,92, una medida de 0,80 y un nivel de área ROC muy bueno de 0,87 (Figura 3.8).
- La clase Aceptable, presenta una tasa de verdaderos positivos con un nivel de aciertos aceptable de 0,60, una tasa de falsos positivos de 0,10 lo que se indica que fueron pocas las instancias clasificadas como erróneas, una precisión de 0,71 lo que indica un nivel instancias correctamente clasificadas muy bueno, una cobertura de 0,60, una medida de 0,65 y un nivel de área ROC muy bueno de 0,87 (Figura 3.9).
- La clase Insuficiente, presenta una tasa de verdaderos positivos con un nivel de ciertos perfecto de 1,00, una tasa de falsos positivos de 0,0 lo que indica que no existió error en la clasificación, una precisión de 1,00 lo que indica un nivel instancias correctamente clasificadas perfecto, una cobertura de 1,00, una medida de 1,00 y un nivel de área ROC excelente de 1,00 (Figura 3.8).

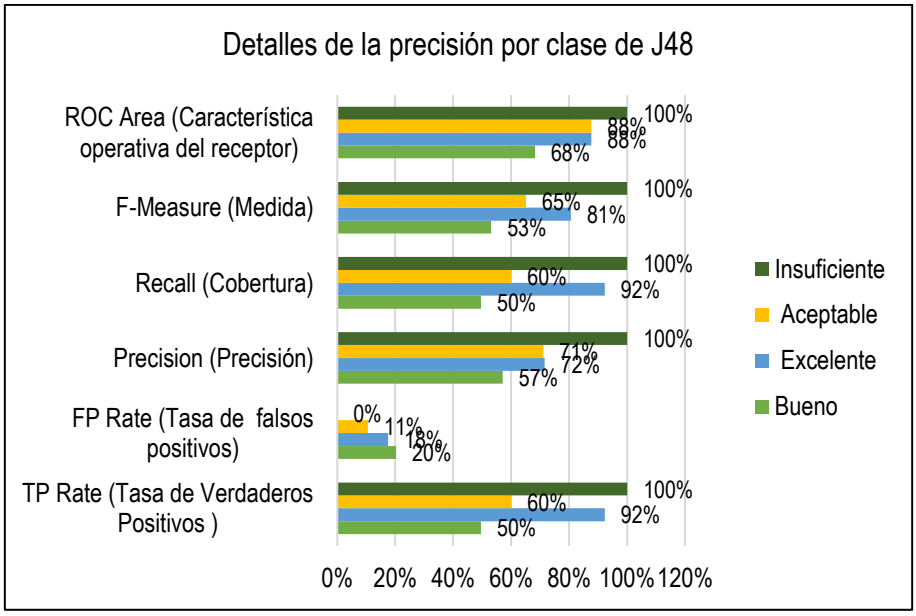



**Gráfico 3. 14.** Detalles de la precisión por clase del algoritmo J48

El Gráfico 3.14, presenta en escala los datos de los resultados porcentuales obtenidos, como se aprecia la longitud de las barras indican que la clase mejor clasificada es el rendimiento académico considerado Insuficiente con un 100% en total de aciertos, la segunda con mejor tasa de aciertos es la Excelente con un 92%, la siguiente es Aceptable con una tasa de aciertos de 60% y la peor clasificada es el rendimiento académico considerado Bueno con una tasa de aciertos del 50%.

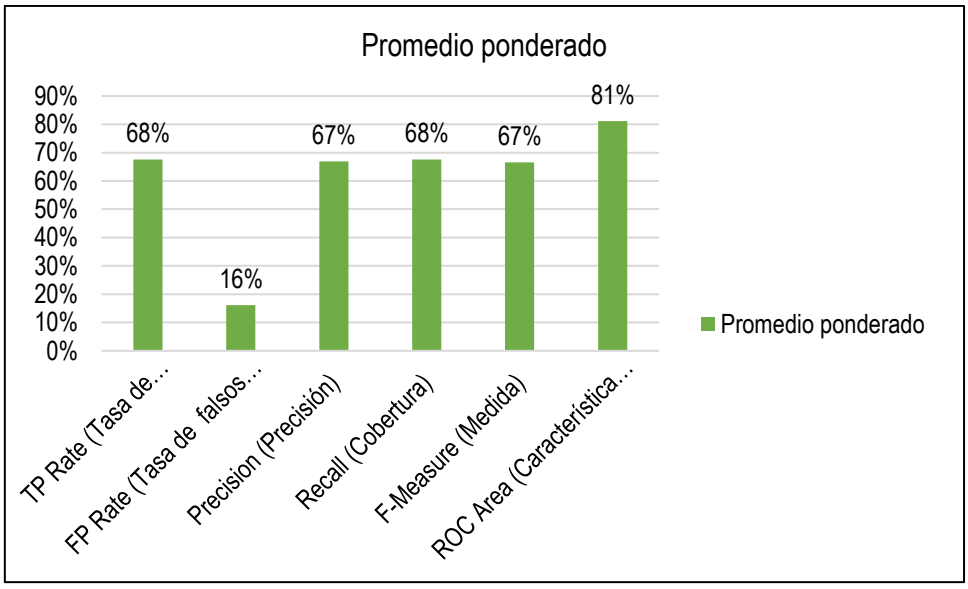



**Gráfico 3.15.** Resultado del promedio ponderado del algoritmo J48

Ante los resultados de la precisión detallada por clase también se obtiene una media ponderada en cada medida estadística lo cual indica que los valores obtenidos son aceptables debido a que la media ponderada de precisión para este clasificador es de 67% (Gráfico 3.15).

Para aclarar estos resultados también se presenta la matriz de confusión:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 204 126  81   0 |   a = Bueno
  24 348   5   0 |   b = Excelente
 129  12 212   0 |   c = Aceptable
   0   0   0  23 |   d = Insuficiente
```

**Figura 3.9.** Matriz de confusión del algoritmo clasificador J48

Según la matriz de confusión Figura 3.9, el algoritmo ha considerado lo siguiente:

- Los valores que se encuentran de manera diagonal en la matriz de confusión son el número de instancias que el clasificador predice como correctas, la suma de la fila de la clase representa el valor real de instancias clasificadas.
- De un total de 411 instancias consideradas como rendimiento académico Bueno el J48 clasificó 204 instancias como correctas, de las 207 instancias predichas por el clasificador como incorrectas 126 fueron clasificadas como Excelente y 81 como Aceptable.
- De las 377 instancias que se consideraban como rendimiento académico Excelente se clasificaron 348 como instancias correctas, de las 29 instancias predichas por el clasificador como incorrectas 24 fueron clasificadas como Bueno y 5 como Aceptable.
- De las 353 instancias que se consideraban como rendimiento académico aceptable se clasificaron 212 como instancias correctas, 12 como Excelente y 129 como Bueno.
- Por último, de los 23 registros considerados como Insuficiente todos fueron clasificados como insuficiente.

Otro algoritmo de clasificación seleccionado es el Naïve Bayes del método de Redes Bayesianas, la ejecución de este algoritmo presenta los siguientes resultados:

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     data
Instances:    1164
Attributes:   9
              X1
              X2
              X3
              X4
              X5
              X6
              X7
              X8
              Class
Test mode:    10-fold cross-validation
```

**Figura 3.10.** Información general del algoritmo clasificador Naïve Bayes

En la Figura 3.10, se muestra la información principal del algoritmo, el número de instancias y atributos del conjunto de datos y a continuación de muestra el modelo clasificador utilizado.

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

Naive Bayes Classifier

                        Class
Attribute               Bueno     Excelente     Aceptable Insuficiente
                       (0.35)        (0.32)         (0.3)        (0.02)
============================================================================
X1
  mean                 4.0754        4.1273        4.0652        3.7826
  std. dev.            0.7556        0.8647        0.8407        0.9305
  weight sum              411           377           353            23
  precision                 1             1             1             1

X2
  mean                 1.4623        1.5225        1.4533        1.4348
  std. dev.            0.4986        0.4995        0.4978        0.4957
  weight sum              411           377           353            23
  precision                 1             1             1             1

X3
  mean                 2.4637        2.3312        2.8264        2.4845
  std. dev.            2.3037        2.1437        2.6121        2.4914
  weight sum              411           377           353            23
  precision            1.1429        1.1429        1.1429        1.1429

X4
  mean                 3.4282        3.4271        3.3626        2.9565
  std. dev.            1.0448        1.0533        1.0559        0.9079
  weight sum              411           377           353            23
  precision                 1             1             1             1

X5
  mean                 1.2117        1.1724        1.2663        1.5652
  std. dev.             0.514         0.488        0.5505        0.7704
  weight sum              411           377           353            23
  precision                 1             1             1             1
X6
  mean                      1             1             1        2.1304
  std. dev.            0.1667        0.1667        0.1667        0.3368
  weight sum              411           377           353            23
  precision                 1             1             1             1

X7
  mean                 8.1314        8.5305        7.9093        7.8696
  std. dev.            1.4887        1.4122        1.3114        1.8952
  weight sum              411           377           353            23
  precision                 1             1             1             1

X8
  mean                 3.3747        3.9761        2.7224        1.4348
  std. dev.            0.4989        0.1667         0.529        0.4957
  weight sum              411           377           353            23
  precision                 1             1             1             1
```

**Figura 3.11.** Modelo clasificador Naïve Bayes

En la Figura 3.11, se observa un nodo raíz que pertenece a la variable dependiente denominada Clase y un conjunto de atributos independientes que contienen: media, desviación estándar, suma de valor y precisión.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         738               63.4021 %
Incorrectly Classified Instances       426               36.5979 %
Kappa statistic                          0.4591
Mean absolute error                      0.219
Root mean squared error                  0.3353
Relative absolute error                 64.5806 %
Root relative squared error             81.4326 %
Total Number of Instances             1164
```

**Figura 3.12.** Resultado general del algoritmo clasificador Naïve Bayes

Se muestra en la Figura 3.12, que, de un total de 1164 instancias, 738 se han considerado como instancias correctamente clasificadas con una equivalencia del 63,40% a diferencia de las 426 instancias no clasificadas correctamente que contienen un porcentaje del 36,60% lo que indica que el rendimiento académico de alrededor de 7 de cada 10 estudiantes se encuentra clasificados correctamente. Al clasificar el total de instancias se obtuvo un promedio de error absoluto del 0,22, también presenta que el coeficiente del estadístico de Kappa tiene un grado de concordancia moderado de 0.45. Al clasificar el total de instancias se obtuvo un promedio de error absoluto del 0.21.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
               0,477     0,260     0,500       0,477    0,488       0,672     Bueno
               0,912     0,201     0,685       0,912    0,783       0,888     Excelente
               0,496     0,089     0,709       0,496    0,583       0,868     Aceptable
               1,000     0,000     1,000       1,000    1,000       1,000     Insuficiente
Weighted Avg.  0,634     0,184     0,633       0,634    0,623       0,808
```

**Figura 3.13.** Resultado individual por clase del algoritmo clasificador Naïve Bayes

El algoritmo Naïve Bayes muestra en la Figura 3.13, los detalles de la precisión por clase, donde se observa TP Rate (Tasa de verdaderos positivos), FP Rate (Tasa de falsos positivos), Precision, Recall (Cobertura), F-Measure (Medida) y ROC Area (área característica de funcionamiento del receptor).

- Para la clase Bueno la TP Rate (Tasa de verdaderos positivos) ha clasificada las instancias con un valor del 0,47, la FP Rate (Tasa de falsos positivos) tiene un valor bajo del 0,26, la precisión tiene un acierto del 0.50, Recall (Cobertura) posee un 0,47, mientras que F-Measure (Medida) tiene 0,48 que especifica la bondad del modelo, misma que determina que entre más cerca este de 1, mayor es la seguridad y por ultimo se encuentra ROC Area que contiene un valor del 0,67, es decir que tiene una exactitud normal ya que lo máximo es 1 y lo mínimo 0,50.
- Para la clase Excelente la tasa de verdaderos positivos tiene un valor elevado del 0,91, la tasa de falsos positivos posee un valor bajo del 0,20, la precisión tiene un acierto del 0,68, la cobertura tiene un resultado del 0,91, significa que es favorable ya que identifica las instancias que se han considerado correctas, mientras que la medida tiene 0,78 y por ultimo se encuentra el ROC Area del 0,88.
- La clase Aceptable tiene la tasa de verdaderos positivos mas baja a diferencia de las clases restantes con un 0,49, la tasa de falsos positivos es del 0,089, mientras que en la precisión es del 0,70, la cobertura del 0,49, medida del 0,58 y ROC Area del 0,86.
- La clase Insuficiente es la clase que se ha clasificado de la mejor manera al ejecutar el algoritmo Naïve Bayes, ya que posee la tasa de verdaderos positivos, la precisión, la cobertura la medida y area bajo la curva del 1,00, obteniendo una tasa de falsos positivos al clasificar las instancias del 0,00.

```
  a   b   c   d   <-- classified as
196 147  68   0 |   a = Bueno
 29 344   4   0 |   b = Excelente
167  11 175   0 |   c = Aceptable
  0   0   0  23 |   d = Insuficiente
```

**Figura 3.14.** Matriz de confusión del algoritmo clasificador Naïve Bayes

La matriz de confusión es de gran importancia ya que permite comparar y evaluar la precisión que el algoritmo de minería ha realizado para cada clase. En la

Figura 3.12, se observa que la tasa de aciertos es 63,40% que corresponde a 738 instancias correctamente clasificadas. Al observar la matriz de confusión Figura 3.14, se deduce que los valores de la diagonal son los clasificados correctamente y los demás son los que se han considerados como errores.

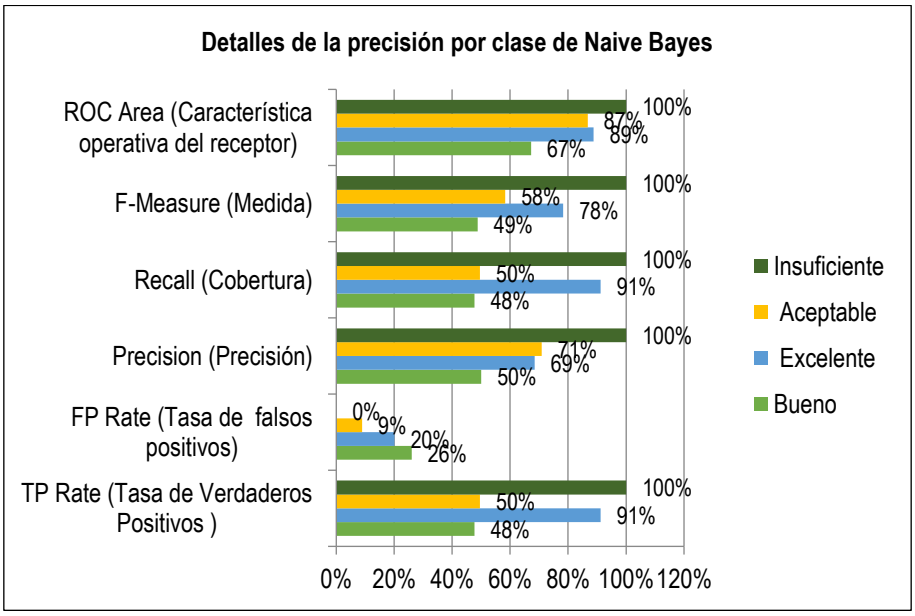



**Gráfico 3.16.** Detalles de la precisión por clase del algoritmo Naïve Bayes

En el Gráfico 3.16, se observa que la clase Insuficiente es la que ha clasificado correctamente las instancias ya que la TP Rate (Tasa de verdaderos positivos) tienen un acierto de 100%, la clase Excelente posee un valor elevado alcanzando un 91% en total de aciertos, la clase Aceptable alcanza un 50% y finalmente la clase Bueno un 47%, siendo esta la clase con un total de verdaderos positivos mas bajo al clasificar cada instancia.

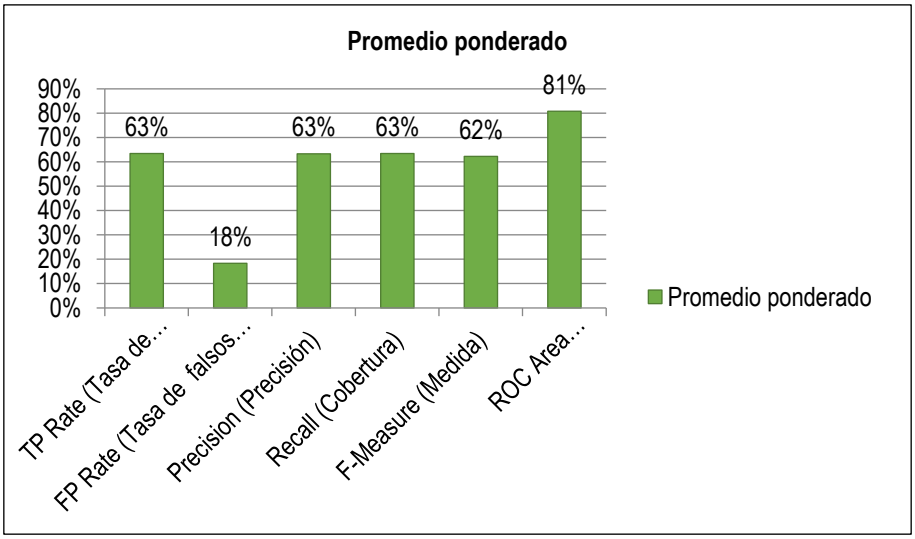



**Gráfico 3.17**. Resultado del promedio ponderado del algoritmo Naïve Bayes

El Gráfico 3.17, contiene un resumen de los detalles de la precisión por clase ya que muestra un promedio ponderado de TP Rate, FP Rate, Precision, Recall, F-Measure y ROC Area, lo cual indica que superan el 50% al clasificar el conjunto de datos.

A continuación, se observa el algoritmo OneR, que pertenece al método de Reglas.

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     data
Instances:    1164
Attributes:   9
              X1
              X2
              X3
              X4
              X5
              X6
              X7
              X8
              Class
Test mode:    10-fold cross-validation
```

**Figura 3.15.** Información general del algoritmo clasificador OneR

En la Figura 3.15, se observa el algoritmo OneR ejecutado en Weka, muestra algunos parámetros del conjunto de datos.

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

X8:
        < 1.5    -> Insuficiente
        < 2.5    -> Aceptable
        < 3.5    -> Bueno
        >= 3.5   -> Excelente
(744/1164 instances correct)
```

**Figura 3.16.** Modelo clasificador del algoritmo OneR

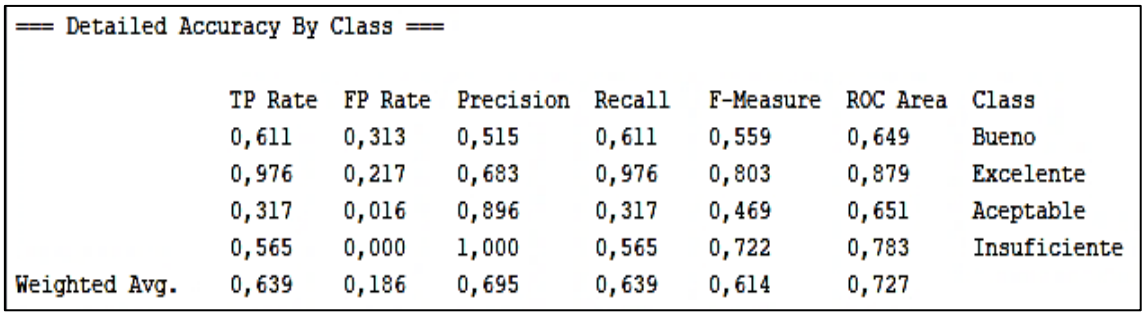
En la Figura 3.16, Se especifica que este algoritmo posee la característica de escoger el atributo que cree conveniente para mostrar resultados sobre la clase, en este proceso utilizo la nota global misma que corresponde a la variable X8.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         744               63.9175 %
Incorrectly Classified Instances       420               36.0825 %
Kappa statistic                          0.4611
Mean absolute error                      0.1804
Root mean squared error                  0.4247
Relative absolute error                 53.192  %
Root relative squared error            103.1628 %
Total Number of Instances             1164
```

**Figura 3.17.** Resultado general del algoritmo clasificador OneR

En la Figura 3.17, se especifica que, de un total de 1164 instancias, 744 fueron clasificadas correctamente con un porcenaje de 63,92% y 420 instancias clasificadas de forma incorrecta con un 36,08% lo que indica que el rendimiento académico de 6 de cada 10 estudiantes se encuentra clasificados correctamente. Ademas, al tratar de clasificar el total de instancias, se logro un promedio de error absoluto del 0,18 y también presenta un estadístico de Kappa con un grado de concordancia moderado 0,46.



=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0,611 | 0,313 | 0,515 | 0,611 | 0,559 | 0,649 | Bueno |
| | 0,976 | 0,217 | 0,683 | 0,976 | 0,803 | 0,879 | Excelente |
| | 0,317 | 0,016 | 0,896 | 0,317 | 0,469 | 0,651 | Aceptable |
| | 0,565 | 0,000 | 1,000 | 0,565 | 0,722 | 0,783 | Insuficiente |
| Weighted Avg. | 0,639 | 0,186 | 0,695 | 0,639 | 0,614 | 0,727 | |

**Figura 3.18.** Resultado individual por clase del algoritmo clasificador OneR

El algoritmo OneR muestra en la Figura 3.18, los detalles de la precisión por clase, donde se observa TP Rate (Tasa de verdaderos positivos), FP Rate (Tasa de falsos positivos), precisión, Recall (Cobertura), F-Measure (Medida) y ROC Área (Área bajo la curva).

- Para la clase Bueno la TP Rate (Tasa de verdaderos positivos) ha clasificado las instancias con un 0,61, la FP Rate (Tasa de falsos positivos) tiene un valor bajo del 0,31, la precisión tiene un acierto del 0,51, Recall (Cobertura) posee un 0,61, mientras que F-Measure (Medida) tiene 0.56 que especifica la bondad del modelo, misma que determina que entre más cerca este de 1, mayor es la seguridad y por ultimo se encuentra ROC Area que contiene un valor del 0,64 que se encuentra dentro de un rango moderado, ya que lo mínimo es 0.50 y el limite 1.
- Para la clase Excelente la tasa de verdaderos positivos tiene un valor del 0,97 que indica que las instancias han sido clasificadas de forma correcta, la tasa de falsos positivos posee un valor bajo del 0,21, la precisión tiene un acierto del 0,68, la cobertura tiene un resultado alto del 0,97, significa que es favorable ya que identifica las instancias que se han considerado correctas, mientras que la medida tiene 0,80 y por ultimo se encuentra el ROC Area que contiene un valor del 0,87.
- La clase Aceptable tiene la tasa de verdaderos positivos mas baja a diferencia de las clases restantes con un 0,31, la tasa de falsos positivos es del 0,016, mientras que en la precisión es del 0,89, cobertura del 0,32, medida del 0,46 y area bajo la curva del 0.65.
- La clase Insuficiente posee la tasa de verdaderos positivos del 0,56 lo cual significa que no han sido clasificadas de la mejor manera, obteniendo la tasa

de falsos positivos del 0,00, con respecto a la precisión el valor es elevado del 1,00, luego de hacer las clasificasiones en cada clase, lo que señala que las instancias se han evaluado correctamente, la cobertura es del 0,57 lo que señala que las instancias no han sido de un todo reconocidas, la medida indica una confiabilidad del 0,72 y un ROC Area del 0,78.

A continuación se muestra un gráfico estadistico de la figura 3.18.

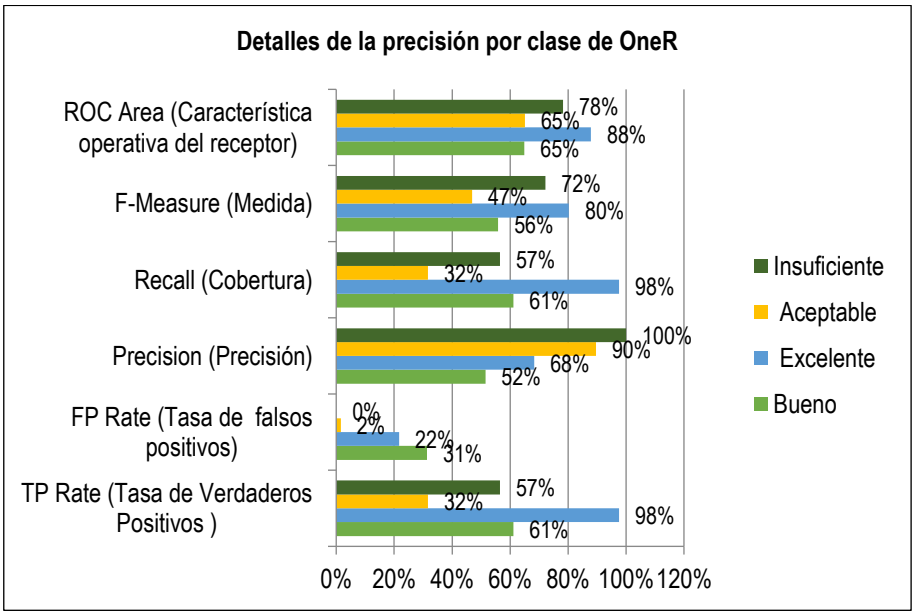



**Gráfico 3.18.** Detalles de la precisión por clase del algoritmo OneR

En el Gráfico 3.18. se observa que la tasa de verdaderos positivos para la clase Excelente tiene un valor de 98% por lo que se considera como la que ha clasificado las instancias de una mejor manera, seguida por la clase Bueno que contiene un valor del 61%, y a continuacion se encuentra la clase Insuficente con 57% y por ultimo la clase Aceptable con un valor porcentual de 32% en total de aciertos lo que indica que es la clase con niveles mas bajo.

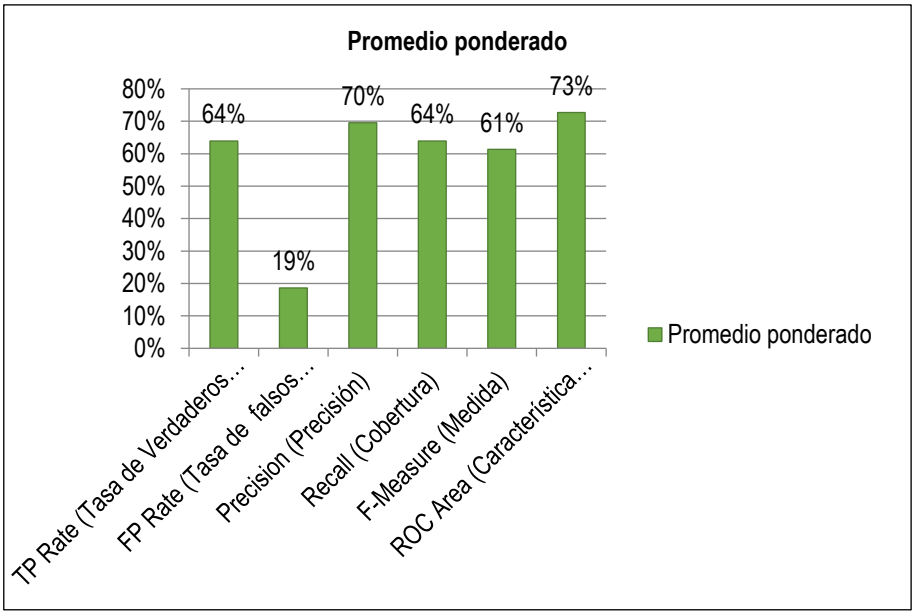



**Gráfico 3.19.** Resultado del promedio ponderado del algoritmo OneR

En el Gráfico 3.19, se observa un analisis detallado del promedio ponderado de TP Rate (Tasa de verdaderos positivos), FP Rate (Tasa de falsos positivos), precisión, Recall (Cobertura), F-Measure (Medida) y ROC Area ya que se trata de un cálculo más puntual de cada uno de los resultados de las clases.

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 251 157   3   0 |   a = Bueno
   9 368   0   0 |   b = Excelente
 227  14 112   0 |   c = Aceptable
   0   0  10  13 |   d = Insuficiente
```

**Figura 3.19.** Matriz de confusión del algoritmo clasificador OneR

En la matriz de confusión del algoritmo OneR de la Figura 3.19. se identifica que cada variable posee el siguiente valor a=251, b=368, c=112 y d=13, es decir, que el número asignados para cada variable en forma diagonal son las que se han considerados como clasificados de forma correcta, los valores restantes se han tomado como valores clasificados incorrectamente.

El Algoritmo del método de árbol de decisión Random Forest (Bosque al azar) presenta a continuación su información general:

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -
Relation:     data
Instances:    1164
Attributes:   9
              X1
              X2
              X3
              X4
              X5
              X6
              X7
              X8
              Class
Test mode:    10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.42 seconds
```

**Figura 3.20.** Información general y modelo clasificador del algoritmo Random Forest

Como se muestra en la Figura 3.20, el Random Forest empieza mostrando el nombre del mismo, la relación, el número de instancias y atributos que tienen los datos así mismo su identificador, en la misma figura también se presenta el modelo clasificador de este algoritmo.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         718               61.6838 %
Incorrectly Classified Instances       446               38.3162 %
Kappa statistic                          0.4356
Mean absolute error                      0.2094
Root mean squared error                  0.3608
Relative absolute error                 61.7257 %
Root relative squared error             87.6244 %
Total Number of Instances             1164
```

**Figura 3.21.** Resultado general del algoritmo clasificador Random Forest

Como resultado del algoritmo la Figura 3.21, muestra que se han clasificado 718 instancias correctas (61,68 %) y 446 instancias incorrectas (38,31%), de un total de 1164 lo que indica que el rendimiento académico de alrededor de 6 de cada 10 estudiantes se encuentra clasificados correctamente, lo que demuestra un

nivel de aciertos muy bueno y un estadístico de Kappa con un grado de concordancia moderado de 0,43; y si analizamos por cada una de las clases tenemos:

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0,433 | 0,255 | 0,481 | 0,433 | 0,456 | 0,643 | Bueno |
| | 0,780 | 0,183 | 0,671 | 0,780 | 0,721 | 0,881 | Excelente |
| | 0,635 | 0,134 | 0,673 | 0,635 | 0,653 | 0,868 | Aceptable |
| | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | Insuficiente |
| Weighted Avg. | 0,618 | 0,190 | 0,611 | 0,618 | 0,612 | 0,795 | |

**Figura 3.22.** Resultado individual por clase del algoritmo clasificador Random Forest

- La clase Bueno, presenta un TP Rate (Tasa de verdaderos positivos) de 0,43 aciertos, una FP Rate (Tasa de falsos positivos) de 0,25 lo que se designa como un nivel de error elevado, una precisión de 0,48 lo que indica un nivel instancias correctamente clasificadas ligeramente bajo, una Recall (Cobertura) de 0,43, una medida de 0,45 y un nivel ROC Area aceptable de 0.64 (Figura 3.22).
- La clase Excelente, presenta una tasa de verdaderos positivos con un nivel de aciertos aceptable de 0,78, una tasa de falsos positivas de 0,18 lo que se designa como un nivel de error bajo, una precisión de 0,67 lo que indica un nivel instancias correctamente clasificadas bueno, una cobertura de 0,78, una medida de 0,72 y un nivel de ROC Area muy bueno de 0,88 (Figura 3.22).
- La clase Aceptable, presenta una tasa de verdaderos positivos con un nivel de aciertos aceptable de 0,63, una tasa de falsos positivos de 0,13 lo que indica que fueron pocas las instancias clasificadas como erróneas, una precisión de 0,66 lo que indica un nivel instancias correctamente clasificadas bueno, una cobertura de 0,63, una medida de 0,65 y un nivel de ROC Area muy bueno de 0,86 (Figura 3.22).
- La clase Insuficiente, presenta una tasa de verdaderos positivos con un nivel de ciertos perfecto de 1,00, una tasa de falsos positivos de 0,0 lo que indica que no existió error en la clasificación, una precisión de 1,00 lo que indica un nivel instancias correctamente clasificadas perfecto, una

cobertura de 1,00, una medida de 1,00 y un nivel de ROC Area excelente de 1,00 (Figura 3.22).

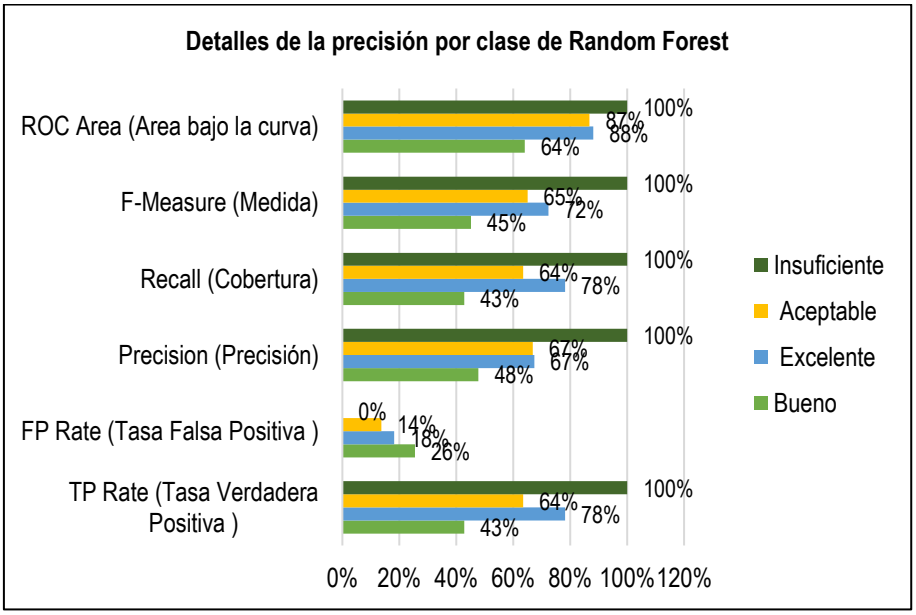



**Gráfico 3.20.** Resultado de la precisión por clase del algoritmo Random Forest

El Gráfico 3.20, presenta en escala los datos de los resultados obtenidos, como se aprecia la longitud de las barras indican que la clase mejor clasificada es el rendimiento académico considerado Insuficiente con un 100% en total de aciertos, la segunda con mejor tasa de aciertos es la Excelente con un 78%, la siguiente es Aceptable con una tasa de aciertos de 64% y el clasificador con más instancias erróneas encontradas es el rendimiento académico considerado Bueno con una tasa de aciertos del 43%.

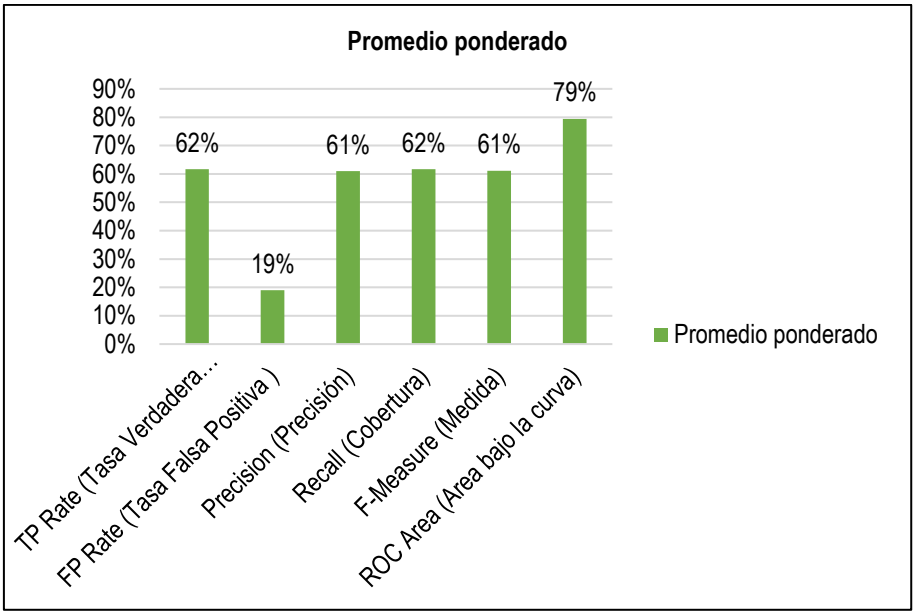



**Gráfico 3.21.** Promedio ponderado del algoritmo Random Forest

Ante los resultados de la precisión detallada por clase también se obtiene una media ponderada en cada medida estadística lo cual indica que los valores obtenidos son aceptables debido a que la media ponderada de precisión para este clasificador es de 61% (Gráfico 3.21).

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 178 131 102   0 |   a = Bueno
  76 294   7   0 |   b = Excelente
 116  13 224   0 |   c = Aceptable
   0   0   0  23 |   d = Insuficiente
```

**Figura 3.23.** Matriz de confusión del algoritmo clasificador Random Forest

Según la matriz de confusión el algoritmo ha considerado lo siguiente:

- Los valores que se encuentran de manera diagonal en la matriz de confusión son el número de instancias que el clasificador predice como correctas, la suma de la fila de la clase representa el valor real de instancias clasificadas.
- De un total de 411 instancias consideradas como rendimiento académico Bueno el Random Forest clasificó 176 instancias como correctas, de las

235 instancias predichas por el clasificador como incorrectas 131 fueron clasificadas como Excelente y 104 como Aceptable.

- De las 377 instancias que se consideraban como rendimiento académico Excelente se clasificaron 295 como instancias correctas, de las 82 instancias predichas por el clasificador como incorrectas 75 fueron clasificadas como Bueno y 7 como Aceptables.
- De las 353 instancias que se consideraban como rendimiento académico aceptable se clasificaron 224 como instancias correctas, 12 como Excelente y 117 como Bueno.
- Por último, de los 23 registros considerados como Insuficiente todos fueron clasificados como insuficiente.

### 3.1.5. EVALUACIÓN

Como resultado final una vez aplicadas cada una de las técnicas de minería de datos se emplearon métricas de evaluación para determinar que técnica o algoritmo tiene mejores resultados respecto a precisión, en el cuadro 3.15 se muestra el rendimiento de los cuatro modelos empleados evaluados a través de las métricas obtenidas a partir de los resultados de la matriz de confusión (Precisión de predicción), estadística de kappa y el área bajo la curva ROC.

**Cuadro 3.15**. Métricas de evaluación

| Métricas de Evaluación | J48 | Naïve Bayes | OneR | Random Forest |
|---|---|---|---|---|
| Instancias clasificadas correctas | 787 | 738 | 744 | 718 |
| Instancias clasificadas incorrectas | 377 | 426 | 420 | 446 |
| Precisión de la predicción (%) | 67,61 | 63,4 | 63,91 | 61,68 |
| Estadística de Kappa | 0,52 | 0,45 | 0,46 | 0,43 |
| Precisión | 0,66 | 0,63 | 0,69 | 0,61 |
| Cobertura | 0,67 | 0,63 | 0,63 | 0,61 |
| F-Medida | 0,66 | 0,62 | 0,61 | 0,61 |
| Área Bajo la curva ROC | 0,81 | 0,80 | 0,72 | 0,79 |

De acuerdo con el Cuadro 3.15, se puede observar que los cuatro algoritmos de clasificación producen relativamente buenos resultados mismos que son similares entre sí, el resultado más alto se obtiene por la clasificación del J48 el

cual presenta un porcentaje de precisión de la clasificación de 67,61%, lo que indica que de las 1164 instancias 787 fueron clasificadas como correctas, es decir que el rendimiento académico de alrededor de 7 de cada 10 estudiantes se encuentra clasificado correctamente.

El coeficiente del estadístico de Kappa, obtenido por el J48 tiene el valor más alto de 0,52 lo que indica que el nivel de concordancia de las variables moderado según lo presenta la escala de valores de kappa del Cuadro 3.2. pero hay que mencionar que los cuatro clasificadores presentan el mismo nivel de concordancia, pero cuanto más cercano al 1 la coincidencia de los datos es casi perfecta.

La precisión y la cobertura se encuentran relacionadas, pero si la precisión aumenta la cobertura disminuye, en los resultados se puede notar que para el J48 aumenta la cobertura y para el OneR disminuye mientras que en los demás se mantiene.

La bondad de los modelos presentada en la F-medida demuestra ser confiable debido a que se encuentran cerca del 1 y mientras el valor se aproxime más al 1 mayor será la confiabilidad.

Por último, de acuerdo con el resultado del área ROC todos los clasificadores tienen un nivel exactitud aceptable debido a que superan el nivel mínimo de 0,50 y presentan valores próximos al nivel máximo de 1.

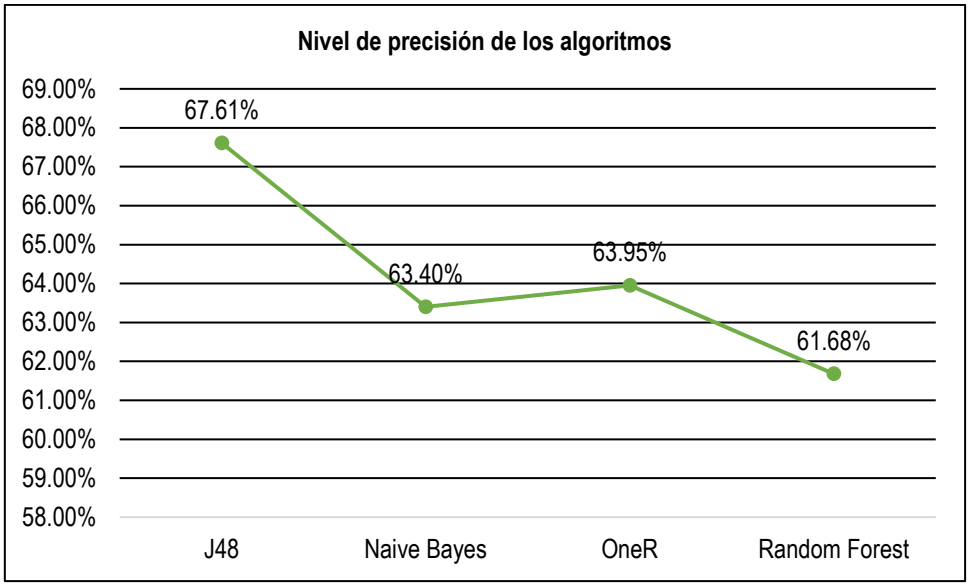



**Gráfico 3.22.** Nivel de precisión de los algoritmos aplicados

El nivel de precisión que se obtiene por los algoritmos aplicados al conjunto de datos con un tamaño de 1164 registros se muestra en el Gráfico 3.22, lo que presenta un nivel mayor para el algoritmo clasificador J48, el nivel de precisión para estos algoritmos como lo demuestra el gráfico de línea se encuentra en intervalos de 58% a 69%. El clasificador J48 es el algoritmo que demostró mejores niveles de precisión.

**Cuadro 3.16.** Cuadro comparativo de resultados de proyectos similares

| PROYECTO | J48 | Random Forest | OneR | Naïve Bayes |
|---|---|---|---|---|
| Minería de datos en egresados de la Universidad de Caldas (Bedoya *et al.,* 2016) | 92,04% | | 84,83% | 81,98% |
| Predicción del rendimiento académico aplicando técnicas de minería de datos de (Menacho 2017) | 68,3% | | | 71,0% |
| Un enfoque de minería de datos para identificar los factores que afectan el éxito académico de los estudiantes terciarios en Sri Lanka (Sanvitha *et al*., 2018) | 96,52% | 98,89% | | 91,22% |
| Minería de datos aplicada a la clasificación de rendimiento académico (Las Autoras) | 67,61 | 61,68 | 63.95 | 63,40 |

- El tema denominado minería de datos en egresados de la Universidad de Caldas, utiliza un registro de 15494. Usa algoritmos como el J48, que muestra un porcentaje más elevado, con un valor de 92,04%, además emplea el

algoritmo OneR que obtiene un 84,83% de fiabilidad al clasificar las instancias y por último el algoritmo Naïve Bayes con un porcentaje más bajo de 81,98%.

- El tema, predicción del rendimiento académico aplicando técnicas de minería de datos, utilizó 1547 registros, donde el algoritmo J48 obtuvo una tasa de buena clasificación del 68,3% y en el Naïve Bayes un 71,0%.
- Un enfoque de minería de datos para identificar los factores que afectan el éxito académico de los estudiantes terciarios en Sri Lanka, es un tema que utiliza algoritmos con tasa de aciertos elevada; el J48 con un valor de 96,52%, el algoritmo Random Forest con 98,89%, además el algoritmo Naïve Bayes con un valor del 91,22%.

Una vez culminado el presente trabajo, se elaboró un cuadro comparativo con tres temas que presentan mayor relacionados a esta área investigativa, donde se deduce que el algoritmo J48 muestra mayor similitud al clasificar la tasa de aciertos con el segundo tema del Cuadro 3.16.

# CAPÍTULO IV. CONCLUSIONES Y RECOMENDACIONES

## 4.1. CONCLUSIÓN

Al establecer los diferentes algoritmos de técnicas de minería de datos aplicada a la educación, se pudo determinar mediante la elaboración de una matriz resumen que los algoritmos más utilizados en este ámbito son algoritmos dependientes de la técnica de clasificación ya que éstos permiten obtener varias soluciones a partir de datos académicos para la toma de decisión.

Al definir las variables que influyen en el rendimiento académico de los estudiantes de la carrera de computación se tomaron datos como: fecha nacimiento, género, ciudad, escuela secundaria, tipo, estado, semestre y subtotal (Notas parciales), que fueron analizados por los algoritmos, además el diagrama de la metodología CRISP-DM permitió tener una estructura organizada y precisa del proceso de aplicación de la minería de datos, lo que determinó precisar el alcance de la experimentación.

Para el manejo de la herramienta WEKA fue necesario preparar los datos mediante una discretización que muestra la valoración de los datos cualitativos en datos cuantitativos, este proceso se realiza para que la clasificación del conjunto de datos se divida en clases que permitan tener una vista minable para una mejor interpretación.

Al analizar la información de los estudiantes de la carrera de Computación de la ESPAM MFL se aplicó minería de datos utilizando la técnica de clasificación, donde se procedió a utilizar los siguientes algoritmos: OneR, J48, Random Forest y Naïve Bayes con el propósito de ejecutar el conjunto de datos, ante los resultados se concluye que el algoritmo J48 además de ser uno de los clasificadores más utilizado en minería de datos educativa, es el modelo que mayor precisión presenta en la prueba de entrenamiento aprendizaje aplicada al

conjunto de datos, el modelo clasificador del J48 reveló que las variables que más inciden en la precisión de la clasificación del rendimiento académico son el estado, semestre y subtotal debido a la fuerte relación que estas presentan.

Graficar los resultados con criterio estadístico arrojados por los clasificadores brinda una mejor interpretación del rendimiento académico de los estudiantes de la carrera de computación, es así como se evidencia que el algoritmo J48 tiene un 67,61% de precisión es decir que alrededor de 7 de cada 10 estudiantes se encuentran categorizado adecuadamente, lo que deduce que en base a este modelo se pueden realizar pruebas para conocer el rendimiento académico de nuevos estudiantes y así mejorar la toma decisiones.

## 4.2. RECOMENDACIÓN

Para conseguir una estructura organizada de las revisiones bibliográficas sobre minería de datos educacional se recomienda la elaboración de una matriz resumen que contenga las investigaciones pertinentes, de tal manera que permita verificar el estado del arte en que actualmente se encuentra la aplicación de las técnicas de minería de datos en el área educativa.

Se recomienda definir bajo una ficha técnica los atributos puntuales del conjunto de datos de la información académica de los estudiantes, para una mejor compresión de los datos y ejecución de los algoritmos lo que permitirá revelar resultados precisos.

Es preciso realizar una preparación de los datos para el proceso de entrenamiento aprendizaje en la aplicación de los algoritmos, ya que al agrupar los datos el proceso se desarrolla con más facilidad.

En investigaciones relacionadas con minería de datos se recomienda el uso de la metodología CRISP-DM ya que está permite optimizar el desarrollo del ciclo de vida de la minería de datos.

Para aplicar minería de datos en investigaciones educativas es recomendable revisar bibliografía con resultados aprobados científicamente y en base a ello elaborar una ficha con información que describa el funcionamiento de los algoritmos a utilizar, además al elegir un algoritmo se debe tener en cuenta que en los resultados de los algoritmos y técnicas de minería influyen diversos factores como la cantidad o tipo de variables independientes.

Para presentar e interpretar los resultados de los modelos desarrollados por los algoritmos se recomienda realizar análisis mediante métricas de evaluación que determinen qué modelo es mejor en la clasificación de datos, específicamente relacionados al rendimiento académico.

# BIBLIOGRAFÍA


Alcover, R; Benlloch, P; Blesa, M; Calduch, M; Celma, C; Ferri, J; Hernández, L; Iniesta, J; Ramírez, A; Robles, J. 2015. Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. Revista en Ciencias de Computacion. Valencia, ES. Vol. XII. p 163-170

Bedoya, O; López, M; Marulanda, C. 2016. Minería de datos en egresados de la Universidad de Caldas. Medellín, CO. Revista Virtual Universidad Católica del Norte. Vol. 6. p.110-124

Chapman, P; Clinton, J; Kerber, R; Khabaza, T; Reinartz; T. 2000. Metodología CRISP-DM. (En línea). Consultado, 30 de may. 2018. Formato PDF. Disponible en: https://www.the-modeling-agency.com/crisp-dm.pdf

Díaz, H; Alemán, Y; Cabrera, L; Morales, A; Chávez, M; Casas, G. 2015. Algoritmos de aprendizaje automático para clasificación de Splice Sites en secuencias genómicas. Placetas, CU. Revista Cubana de Ciencias Informáticas. Vol. 9. p 155-170

Domínguez, M. 2008. Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado. Tesis. Maestría en Ciencias de la Computación. Universidad Valle Del Grijalva. MX. p 12-20

Durán, E. y Costaguta, R. 2007. Minería de datos para descubrir estilos de aprendizaje. ARG. Revista Iberoamericana de Educación. Vol. 21. p. 2-10

Eckert, K y Suénaga, R. 2013. Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD. Posadas-Misiones, AR. XV Workshop de Investigadores en Ciencias de la Computación. Vol. 1. p 94

ESPAM MFL (Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López). 2015a. Historia de ESPAM MFL. (En línea). EC. Consultado, 23 de dic. 2017. Formato HTML. Disponible en: http://web1.espam.edu.ec/?item=15

ESPAM MFL (Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López). 2015b. Misión y visión de la Carrera de Computación. (En línea). EC. Consultado, 21 de may. 2018. Formato HTML. Disponible en: http://computacion.espam.edu.ec/

Fernándes, E; Carvalho, R; Holanda, M; Van, G. 2017. Educational Data Mining: Discovery Standards of Academic Performance by Students in Public High Schools in the Federal District of Brazil. Advances in Intelligent Systems and Computing. Vol. 569. p 287-296

García, D; Palazuelos, C; Zorrilla, M. 2014. Data Mining and Social Network Analysis in the Educational Field: An Application for Non-Expert Users. Educational Data Mining. Studies in Computational Intelligence. Vol. 524. p 411-439

García, J. 2016. Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. Madrid, ES. Arxiv, Artificial Intelligence (cs.AI). Vol. 3. p 1-6

García, F; Sarria, F; Castillo, G. 2016. Modificación del algoritmo Random Forest para su empleo en clasificación de imágenes de teledetección. Murcia, ES. Revista Tecnológica de la Información Geográfica. Vol. 1. p. 359-368

García, M; Bernardo, A; Rodríguez. L. 2016. Permanencia en la universidad: la importancia de un buen comienzo. Oviedo-Asturias, ES Open Access funded by Instituto de Ciencias de la Educación. Vol. 44. p 1-6.

Jantan, H; Hamdan, A; Othman, Z. 2010. Clasificación y predicción del talento académico utilizando técnicas de minería de datos. Knowledge-Based and Intelligent Information and Engineering Systems. Notas de la conferencia en Ciencias de la computación. Vol. 6276. p 491-500

Jaramillo, A. y Paz, H. 2015. Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. Loja, EC. Revista Tecnológica ESPOL – RTE. Vol. 28. p 64-90

Jiang, L; Yang, Z; Liu; Q; Wei, H. 2008. Summarization on the Data Mining Application Research in Chinese Education. Leung E.W.C. LNCS 5328. Conference notes in Computer Science. Vol. 5328. p 110–120

Jiménez, G. y Álvarez, H. 2010. Minería de Datos en la Educación. ES. Revista de Inteligencia en Redes de Comunicación. p 2-9

Jiménez, J. y Toledo, S. 2015. Caracterización de la deserción estudiantil en educación superior con minería de datos. Pasto-Nariño, CO. Revista Tecnológica ESPOL – RTE. Vol. 28. p 447-463

Keshtkar, F; Burkett, C; Li, H; Graesser, A. 2014. Using Data Mining Techniques to Detect the Personality of Players in an Educational Game. Studies in Computational Intelligence. Vol. 524. p 125-150

Kotsiantis, S. 2012. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artif Intell Rev. Vol. 37. p 331-344

La Red, D; Karanik, M; Giovannini, M; Pinto, N. 2015. Perfiles de Rendimiento Académico: Un Modelo basado en Minería de datos. Campus Virtuales. ARG. Vol. 4. p 12-30

Mashiloane, L; Mchunu, M. 2013. Mining for Marks: A Comparison of Classification Algorithms when Predicting Academic Performance to Identify "Students at Risk". Inteligencia minera y exploración del conocimiento. Conference notes in Computer Science. Vol. 8284. p 403-414

Menacho, C. 2017. Predicción del rendimiento académico aplicando técnicas de minería de datos. La Molina-Lima, PE. Revista de Facultad de Economía y Planificación. Vol. 78. p 26-33

Meedech, P; Iam-On, N; Boongoen, T. 2016. Prediction of Student Dropout Using Personal Profile and Data Mining Approach. Intelligent and Evolutionary Systems, Proceedings in Adaptation, Learning and Optimization. Vol. 5.p 143-155

Natek, S. y Zwilling, M. 2014. Student data mining solution–knowledge management system related to higher education institutions. Celje, SI. Expert Systems with Applications. Vol. 41. p 6400–6407

Navas, F. 2016. Introducción a la minería de datos con Weka: Aplicación a un problema económico. Tesis. Trabajo Fin de Grado. Universidad de Jaén Facultad de Ciencias Sociales y Jurídicas. Jaén-Andalucía, ES. p 19-25

Olague, J; Torres, S; Morales, F; Valdez, A; Silva, A. 2010. Sistemas de gestión de contenidos de aprendizaje y técnicas de minería de datos para la enseñanza de ciencias computacionales. Un caso de estudio en el norte de Coahuila. Distrito Federal, MEX. Revista Mexicana de Investigación Educativa. Vol. 15. p. 391-421

Pautsch, J. 2009. Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. Tesis. Lic. Sistemas de Información. Universidad Nacional de Misiones. ARG. p 112 – 120

Pereira, R; Calderón, A; Jiménez, A. 2013. Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. Manizales, CO. Revista de Facultad de Ciencias e Ingeniería. Vol. 28. p. 31-47

Pérez, T; Caballero, D; Caro, A; Rodríguez, P y Antequera, T. 2014. Applying data mining and Computer Vision Techniques to MRI to estimate quality traits in Iberian hams. Cáceres, ES. Journal of Food Engineering. Vol. 131. p 82–88

Pinzón, L. 2011. Aplicando minería de datos al marketing educativo. CO. Notas D Marketing. Vol. 1. p 45-61

Reyes, A; Flores, A; Alejo, R; Rendón, E. 2017. Minería de datos aplicada para la identificación de factores de riesgo en alumnos. MX. Revistas de Ciencias de la Computación Vol. 139. p. 177–189

Riquelme, J; Ruiz, R. y Gilbert, K. 2016. Minería de Datos: Conceptos y Tendencias. Valencia, ES. Revista Iberoamericana de Inteligencia Artificial. Vol. 10. p 11-18

Rodríguez, Y. y Díaz A. 2009. Herramientas de Minería de Datos. CU. Revista Cubana de Ciencias Informáticas, Vol. 3. p. 73-80

Rosado, A. y Verjel, A. 2016. Aplicación de la minería de datos de la educación en línea. CO. Revista Colombiana de Tecnologías de Avanzada. Vol. 1. p 92-96

Romero, C; Ventura, S; Hervás, C. 2005. Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. (En línea). ES. Consultado, 04 de mar. 2018. Formato PDF. Disponible en: http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CICA_2009_Aplicacion_Mineria_Datos_basada_ense%F1anza_web.pdf

Romero, C; Ventura, S; Castro, C; García, E. 2005. Algoritmos evolutivos para descubrimiento de reglas de predicción en la mejora de sistemas educativos adaptativos basados en web. ES. Revista Iberoamericana de Informática Educativa. Vol. 2. p 47 – 60

Sharma, M. y Mavani, M. 2011. Accuracy Comparison of Predictive Algorithms of Data Mining: Application in Education Sector. DE. Journal of Computer Science. Vol. 11. p. 189–194

Spositto, O; Etcheverry, M; Ryckeboer, H; Bossero, J. 2008. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. San Justo, ARG. Revista de Ingeniería e Investigaciones Tecnológicas. Vol. 1. p 2-7

Saleem, M; Kathiry, N; Osimi, S; Badr, G. 2015. Mining Educational Data to Predict Students' Academic Performance. Machine Learning and Data Mining in Pattern Recognition. MLDM 2015. Lecture Notes in Computer Science. Vol. 9166. p 403-414

Sanvitha, K; Liyanage, S; Bhatt, C. 2018. A Data Mining Approach to Identify the Factors Affecting the Academic Success of Tertiary Students in Sri Lanka. Software Data Engineering for Networke Learning Environments Notes on Data Engineering and Communications Technologies. Vol. 11. p 179-197

Troche, A. 2014. Aplicación de la minería de datos sobre bases de datos transaccionales. Salle, BO. Fides et Ratio - Revista de Difusión cultural y científica de la Universidad La Salle en Bolivia. Vol. 7. p 58-66

Vaira, S; Avila, O; Ricardi, P; Bergesio, A. 2010. Deserción universitaria. Un caso de estudio: variables que influyen y tiempo que demanda la toma de decisión. El Pozo-Santa Fe, ARG. Revista FABICIB. Vol. 14. p 107-115

Vialardi, C; Chue, J; Peche, J. 2011. Un enfoque de minería de datos para guiar a los estudiantes a través del proceso de inscripción basado en el rendimiento académico. User Modeling and User-Adapted Interaction. Vol. 21. p 217–248

# ANEXOS

# ANEXO 1. NECESIDAD DE LA INTERVENCIÓN SISTEMATIZACIÓN DE EXPERIENCIA

Calceta, 16 de enero de 2018

Ingeniero
Luis Alberto Ortega Arcia, Mg.
**DIRECTOR ENCARGADO DE LA CARRERA DE COMPUTACIÓN**
En su despacho.-

De mi consideración:

Como directora del proyecto de investigación titulado "MINERÍA DE DATOS APLICADA AL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES EN LA ESPAM MFL", propuesto por la Carrera de Computación con CUP 381224, le solicito muy respetuosamente, se considere como propuesta para trabajo de titulación. De acuerdo con lo expuesto se pretende proponer el desarrollo de la Predicción del rendimiento académico aplicada a los estudiantes de la Carrera de Computación, de tal manera que permita contribuir de manera directa a la consecución de uno de los objetivos del proyecto.

Por la atención brindada a la presente, le agradezco de antemano.

Atentamente,

Jéssica Morales Carrillo
**DIRECTORA DEL PROYECTO DE INVESTIGACIÓN**

ESPAM MFL
ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ
RECIBIDO
FECHA: 16-01-2018 HORA 15:00

## ANEXO 2. CONJUNTO DE DATOS DISCRETIZADO EN FORMATO .CSV

| |
|---|
| X1,X2,X3,X4,X5,X6,X7,X8,Class |
| 2,1,2,2,1,1,10,3,Bueno |
| 2,1,2,2,1,1,10,4,Excelente |
| 3,1,1,2,1,1,10,4,Excelente |
| 3,1,1,2,1,1,10,4,Excelente |
| 4,1,1,2,1,1,10,4,Bueno |
| 4,1,1,2,1,1,10,4,Excelente |
| 5,2,7,3,1,1,9,2,Aceptable |
| 5,2,7,3,1,1,9,2,Aceptable |
| 5,2,7,3,1,1,9,2,Aceptable |
| 5,2,7,3,1,1,9,2,Aceptable |
| 5,2,7,3,1,1,9,2,Aceptable |
| 5,2,7,3,1,1,9,2,Aceptable |