UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS Y COMPUTACIÓN

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

TRABAJO DE TITULACIÓN

"ANÁLISIS DE EXACTITUD DE LOS ALGORITMOS DE ÁRBOLES DE DECISIÓN APLICADOS EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH."

Autor:
Carlos Stiven Correa Salinas

Tutor:
MsC. Jorge Delgado

Riobamba - Ecuador

2019

PÁGINA DE ACEPTACIÓN

Los miembros del tribunal de graduación del proyecto de investigación de título: **"ANÁLISIS DE EXACTITUD DE LOS ALGORITMOS DE ÁRBOLES DE DECISIÓN EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH"**, presentado por el Sr. Carlos Stiven Correa Salinas y dirigida por: MsC. Jorge Delgado.

Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación con fines de graduación escrito en el cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la Universidad Nacional de Chimborazo.

Para constancia de lo expuesto firman:

MsC. Jorge Delgado

Director del proyecto

Firma

PhD. Lida Barba

Miembro del Tribunal

Firma

MsC. Lady Espinoza

Miembro del Tribunal

Firma

DERECHOS DE AUTORÍA

La responsabilidad del contenido de este proyecto de graduación corresponde exclusivamente al Sr. Carlos Stiven Correa Salinas bajo la dirección del MsC. Jorge Delgado y al patrimonio intelectual de la Universidad Nacional de Chimborazo.

Autor:

Carlos Stiven Correa Salinas
085023837-9

DEDICATORIA

Dedico este proyecto de investigación a mi familia, quienes siempre han creído en mi y me han apoyado incondicionalmente desde el inicio de esta etapa, a mi madre Olga Salinas que es mi amor más grande y ha sido mi apoyo y pilar fundamental en el logro de esta meta, a mis hermanos Byron y Odalys que siempre me han dado la fuerza necesaria para seguir en este largo camino.

Carlos Stiven Correa Salinas

AGRADECIMIENTO

Agradezco principalmente a Dios porque sin el nada de esto me sería posible, a los docentes de la Carrera de Ingeniería en Sistemas y Computación quienes apoyan a los estudiantes cada día para que logren su objetivo, a mi tutor por guiarme con sus conocimientos y apoyarme incondicionalmente durante todo el proceso de desarrollo y culminación de este proyecto de investigación.

Carlos Stiven Correa Salinas

ÍNDICE GENERAL

PÁGINA DE ACEPTACIÓN.....II
DERECHOS DE AUTORÍA.....III
DEDICATORIA.....IV
AGRADECIMIENTO.....V
LISTA DE TABLAS.....VIII
RESUMEN.....XII
ABSTRACT.....XIII
INTRODUCCIÓN.....1
CAPÍTULO I.....3
PLANTEAMIENTO DEL PROBLEMA.....3
1.1. Planteamiento del problema y justificación.....3
1.2. Objetivos.....4
1.2.1. Objetivo general.....4
1.2.2. Objetivos específicos.....4
CAPÍTULO II.....5
MARCO TEÓRICO.....5
2.1. Minería de datos.....5
2.2. Aplicaciones de la minería de datos.....5
2.3. Técnicas de minería de datos.....6
2.5. Exactitud en árboles de decisión.....7
2.6. Algoritmo ID3.....8
2.7. Algoritmo J48.....10
2.8. Metodología CRISP–DM.....12
CAPITULO III.....15
METODOLOGÍA.....15
3.1. Tipo de Investigación.....15
3.2. Unidad de análisis.....16
3.3. Técnicas de análisis e interpretación de la información.....16
3.4. Aplicación de la metodología CRISP-DM.....16

CAPÍTULO IV. 25
RESULTADOS Y DISCUSIÓN 25
4.1. Resultados 25
4.2. Discusión 46
5. CONCLUSIONES 47
6. RECOMENDACIONES 48
7. REFERENCIAS BIBLIOGRÁFICAS 49
8. ANEXOS 52
8.1. Anexo 1. Recursos software y hardware 52
8.2. Anexo 2. Descripción de los datos 53
8.3. Anexo 3. Exploración de los datos 60
8.4. Anexo 4. Análisis de calidad de los datos 67
8.6. Anexo 6. Atributos derivados y generados 71
8.7. Anexo 7. Modelos generados 73

LISTA DE TABLAS

Tabla 1 Matriz de confusión. 8
Tabla 2 Metodologías usadas en minería de datos.....12
Tabla 3 Datos del estudiante. 21
Tabla 4 Datos del docente. 21
Tabla 5 Datos de investigación. 22
Tabla 6 Exactitud de los algoritmos en datos de estudiantes de Ingeniería. 25
Tabla 7 Matriz de confusión algoritmo ID3 en datos de estudiantes de Ingeniería. 26
Tabla 8 Matriz de confusión algoritmo J48 en datos de estudiantes de Ingeniería. 26
Tabla 9 Mejor regla de clasificación de estudiantes de Ingeniería según su promedio. 26
Tabla 10 Porcentaje de probabilidad de un estudiante de Ingeniería de obtener un 26
Tabla 11 Exactitud de los algoritmos en datos de estudiantes de Ciencias de la Educación. 27
Tabla 12 Matriz de confusión algoritmo ID3 en datos de estudiantes de Educación. 27
Tabla 13 Matriz de confusión algoritmo J48 en datos de estudiantes de Educación. 27
Tabla 14 Mejor regla de clasificación de estudiantes de Educación según su promedio.....28
Tabla 15 Porcentaje de probabilidad de un estudiante de Educación de obtener un 28
Tabla 16 Exactitud de los algoritmos en datos de estudiantes de Ciencias Políticas. 29
Tabla 17 Matriz de confusión algoritmo ID3 en datos de estudiantes de Ciencias Políticas 29
Tabla 18 Matriz de confusión algoritmo J48 en datos de estudiantes de Ciencias Políticas 29
Tabla 19 Mejor reglas de clasificación de estudiantes de Ciencias Políticas 29
Tabla 20 Porcentaje de probabilidad de un estudiante d Políticas de obtener cierto promedio ..30
Tabla 21 Exactitud de los algoritmos en datos de estudiantes de Ciencias de la Salud. 30
Tabla 22 Matriz de confusión algoritmo ID3 en datos de estudiantes de Ciencias de la Salud. ...31
Tabla 23 Matriz de confusión algoritmo J48 en datos de estudiantes de Ciencias de la Salud. ...31
Tabla 24 Mejor regla de clasificación de estudiantes de Salud según su promedio. 31
Tabla 25 Porcentaje de probabilidad de un estudiante de Ciencias de la Salud de obtener un 31
Tabla 26 Exactitud de los algoritmos en datos de docentes de Ingeniería. 32
Tabla 27 Matriz de confusión algoritmo ID3 en datos de docentes de Ingeniería. 33
Tabla 28 Matriz de confusión algoritmo J48 en datos de docentes de Ingeniería. 33
Tabla 29 Mejor regla de clasificación de docentes de Ingeniería según su calificación.....33

Tabla 30 Porcentaje de probabilidad de un docente de Ingeniería de obtener una.....33
Tabla 31 Exactitud de los algoritmos en datos de docentes de Ciencias de la Educación.....34
Tabla 32 Matriz de confusión algoritmo ID3 en datos de docentes de Educación.....34
Tabla 33 Matriz de confusión algoritmo J48 en datos de docentes de Educación. 34
Tabla 34 Mejor regla de clasificación de docentes de Educación según su calificación. 35
Tabla 35 Porcentaje de probabilidad de un docente de Ciencias de la Educación 35
Tabla 36 Exactitud de los algoritmos en datos de docentes de Ciencias Políticass.....36
Tabla 37 Matriz de confusión algoritmo ID3 en datos de docentes de Ciencias Políticas.....36
Tabla 38 Matriz de confusión algoritmo J48 en datos de docentes de Ciencias Políticas. 36
Tabla 39 Mejor reglas de clasificación de docentes de Políticas según su calificación. 36
Tabla 40 Porcentaje de probabilidad de un docente de Ciencias Políticas. 37
Tabla 41 Exactitud de los algoritmos en datos de docentes de Ciencias de la Salud. 37
Tabla 42 Matriz de confusión algoritmo ID3 en datos de docentes de Ciencias de la Salud.....37
Tabla 43 Matriz de confusión algoritmo J48 en datos de docentes de Ciencias de la Salud. 38
Tabla 44 Mejor regla de clasificación de docentes de Ciencias de la Salud 38
Tabla 45 Porcentaje de probabilidad de un docente de Ciencias de la Salud de obtener una.....38
Tabla 46 Exactitud de los algoritmos en datos de publicaciones de docentes de Ingeniería. 39
Tabla 47 Matriz de confusión algoritmo ID3 en publicaciones de docentes de Ingeniería. 39
Tabla 48 Matriz de confusión algoritmo J48 en de publicaciones de docentes de Ingeniería. 39
Tabla 49 Mejores reglas de clasificación de publciaciones de docentes de Ingeniería 40
Tabla 50 Porcentaje de probabilidad de publciaciones de un docente de Ingeniería.....40
Tabla 51 Exactitud de los algoritmos en datos de publicaciones de docentes de Educación.....41
Tabla 52 Matriz de confusión algoritmo ID3 en publicaciones de docentes de Educación.....41
Tabla 53 Matriz de confusión algoritmo J48 en publicaciones de docentes de Educación. 41
Tabla 54 Mejores reglas de clasificación de publicaciones docentes de Educación. 41
Tabla 55 Porcentaje de probabilidad de publicaciones de un docente de Educación.....42
Tabla 56 Exactitud de los algoritmos en datos de publicaciones de docentes de Políticas.....42
Tabla 57 Matriz de confusión algoritmo ID3 en publicaciones de docentes de Políticas.....42
Tabla 58 Matriz de confusión algoritmo J48 en publicaciones de docentes de Políticas 43
Tabla 59 Mejores reglas de clasificación de publicaciones de docentes de Políticas.....43
Tabla 60 Porcentaje de probabilidad de publicaciones de un docente de Políticas.....43

Tabla 61 Exactitud de los algoritmos en datos de publicaciones de docentes de Salud. 44
Tabla 62 Matriz de confusión algoritmo ID3 en datos de publicaciones de docentes de Salud. ..44
Tabla 63 Matriz de confusión algoritmo J48 en datos de publicaciones de docentes de Salud. ..44
Tabla 64 Mejores reglas de clasificación de publicaciones de docentes de Salud 45
Tabla 65 Porcentaje de probabilidad de un docente de Salud de realizar o no publicaciones. 45
Tabla 66 Promedio general de exactitud de cada uno de los algoritmos. 45
Tabla 67 Recursos Hardware.....52
Tabla 68 Características de Recurso Hardware 52
Tabla 69 Descripción de la tabla Estudiante 53
Tabla 70 Descripción de la tabla Estudiante_Rendimiento.....54
Tabla 71 Descripción de la tabla Docente.....55
Tabla 72 Descripción de la tabla Docente_infAcadémica 56
Tabla 73 Descripción de la tabla Evaluación_Docente.....56
Tabla 74 Descripción de la tabla Investigación.....57
Tabla 75 Distribución de los estudiantes según su estado civil.....60
Tabla 76 Distribución de los estudiantes según su sexo 60
Tabla 77 Distribución de los estudiantes según su facultad.....61
Tabla 78 Distribución de los docentes según su estado civil. 62
Tabla 79 Distribución de los docentes según su sexo. 63
Tabla 80 Distribución de los docentes según su facultad. 63
Tabla 81 Distribución de los estudiantes según su estado. 64
Tabla 82 Distribución de las investigaciones según su tipo. 65
Tabla 83 Análisis de calidad de datos. 67
Tabla 84 Atributos eliminados. 70
Tabla 85 Atributos derivados. 71
Tabla 86 Atributos generados. 72
Tabla 87 Ponderación de la calificación del estudiante. 72
Tabla 88 Ponderación de la calificación del docente. 72

LISTA DE FIGURAS

Figura 1. Estudiantes según su estado civil. 60
Figura 2. Estudiantes según su sexo....61
Figura 3. Estudiantes según su facultad. 62
Figura 4. Docentes según su estado civil. 62
Figura 5. Docentes según su sexo. 63
Figura 6. Docentes según su facultad....64
Figura 7. Investigaciones según su estado....65
Figura 8. Investigaciones según su tipo. 66
Figura 9. Modelo generado para el algoritmo ID3 en tabla Estudiantes. 73
Figura 10. Validación cruzada del algoritmo ID3 en tabla Estudiantes. 73
Figura 11. Modelo generado para el algoritmo J48 en tabla Estudiantes. 74
Figura 12. Validación cruzada del algoritmo J48 en tabla Estudiantes. 74
Figura 13. Modelo generado para algoritmo ID3 en tabla Docentes. 74
Figura 14. Validación cruzada d el algoritmo ID3 en tabla Docentes. 74
Figura 15. Modelo generado para el algoritmo J48 en tabla Docentes. 74
Figura 16. Validación cruzada del algoritmo ID3 en tabla Docentes. 74
Figura 17. Modelo generado para el algoritmo ID3 en la tabla Investigación. 74
Figura 18. Validación cruzada del algoritmo ID3 en la tabla Investigación. 74
Figura 19. Modelo generado para el algoritmo J48 en la tabla Investigación. 74
Figura 20. Validación cruzada del algoritmo J48 en la tabla Investigación. 74

RESUMEN

En la Universidad Nacional de Chimborazo (Unach) existen grandes volúmenes de datos que han sido escasamente analizados por medio de técnicas de minería de datos, privándose así de información útil que pudiese apoyar a los directivos en la toma de decisiones y contribuir así al desarrollo de la institución.

Con la finalidad de encontrar el mejor modelo de árboles de decisión, se realizó el análisis de exactitud de los algoritmos árbol de decisión de inducción de arriba abajo (ID3) y árbol de inducción (J48), aplicados en la información personal, académica de estudiantes, docentes y producción científica, todo esto almacenado en la base de datos del Sistema de Control Académico (Sicoa) de la Unach. El proceso de análisis, preparación de los datos y aplicación de los algoritmos se realizó a través de la metodología CRISP-DM, misma que proporciona una guía detallada de cómo llevar a cabo este proceso.

Se realizaron un total de doce análisis con cada uno de los algoritmos que involucraron las variables independientes y dependiente de estudiantes, docentes y producción científica, se promedió los resultados conseguidos en cada uno de ellos obteniendo así un valor general, el cual demostró que el algoritmo J48 es más exacto que el algoritmo ID3, además se generaron las principales reglas de clasificación para cada una de las tablas. Los resultados obtenidos contribuirán al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

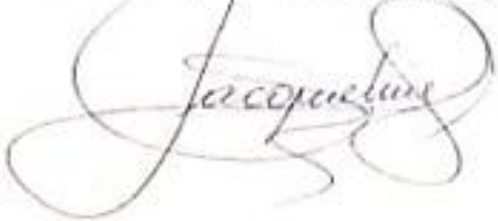
Palabras claves: Minería de datos, Árboles de decisión, ID3, J48, Exactitud, Clasificación.

ABSTRACT

At the National University of Chimborazo (UNACH), there is a great deal of data that have been poorly analyzed through data mining techniques. Therefore, the limited analysis deprives from useful information that may support UNACH's administrators to make decisions and contribute to the development of the institution. In the endeavor to find the best decision tree model, an accurate tree-algorithm induction decision analysis has been done from the top to the bottom (ID3), and induction tree (J48). Both of them were applied on personal and academic information of the students' performance, professors and scientific production. All the information is stored in a database named Academic Control System (SICOA) from the UNACH. The process of analysis, data preparation and application of the algorithms were carried out through CRISP-DM methodology, which provides a detailed guide on how to carry out the mentioned process. A total of twelve analyzes were performed with each of the algorithms that involved independent and dependent variables from students, professors and scientific production. The achieved results of each one of them were averaged. Thus, it allowed obtaining a general value, which showed that the J48 algorithm is more accurate than the ID3 algorithm. In addition, the main classification rules for each of the tables was generated. Finally, the obtained results will contribute to a project named "Designing strategies for continuous improvement in the academic and research management of UNACH by using data mining".

Keywords: Data mining, Decision trees, ID3, J48, Accuracy, Classification.

Reviewed and corrected by: Lic. Arnljos Monar Jacqueline, MsC.



INTRODUCCIÓN

La minería de datos en la educación no es un tópico nuevo y tanto su estudio como aplicación ha sido muy relevante en los últimos años en las instituciones de educación superior. Su uso permite, entre otras cosas, predecir cualquier acontecimiento dentro del ámbito educativo, de esta manera se puede generar nuevo conocimiento con un porcentaje muy alto de confiabilidad (Valero, Salvador y García, 2005).

Las técnicas de minería de datos permiten descubrir patrones en grandes volúmenes de información, las cuales son aplicadas mediante una serie de algoritmos que ayudan a extraer conocimiento de estos conjuntos (Robles y Sotolongo, 2013). Entre estas técnicas se encuentran los árboles de decisión, mismos que se encargan de brindar un modelo de predicción basado en construcciones lógicas a partir de un conjunto de variables proporcionadas (Escobar, Alcivar y Puris, 2016). Los principales algoritmos de árboles de decisión son ID3 y J48. El algoritmo ID3 se enfoca en la búsqueda de hipótesis o reglas dado un conjunto de ejemplos (Encarnación, 2014). El algoritmo J48 construye el árbol iterativamente al ir agregando nodos o ramas que minimicen la diferencia entre los datos (Barrientos et al., 2009) .

En la revisión de literatura sobre antecedentes de la investigación relacionados con la exactitud de los algoritmos se encontraron los siguientes proyectos relacionados, uno de los trabajos se titula *Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos* (Encarnación, 2014) aplicado a los estudiantes de la Universidad de Loja, como resultado del análisis comparativo se verifica que el algoritmo J48 posee un mayor porcentaje de valores clasificados correctamente, otro de los documentos es *Uso de minería de datos para determinar la disponibilidad de una red ip v.4 en una cadena de terminales distribuidos* realizado *por* (Pradenas y Parra, 2012), de igual forma este autor realiza una comparativa de resultados de los algoritmos

ID3 y J48, pero en esta ocasión se concluye que al algoritmo ID3 es mejor debido a que posee un mayor porcentaje de valores correctamente clasificados. Razón por la cual en este trabajo se realiza el análisis comparativo de la exactitud de clasificación de ID3 y J48, aplicados en la información personal y académica de estudiantes, docentes y producción científica, todo esto almacenado en la base de datos del Sistema de Control Académico (SICOA) de la Unach. Además, se busca generar las principales reglas de clasificación basadas en los valores de las variables independientes con respecto a la variable dependiente, en el caso del estudiante esta variable dependiente es su promedio general, del docente la calificación obtenida en la evaluación final que se les realiza y por último si un docente ha realizado o no publicaciones.

Para realizar un procedimiento de minería de datos eficiente es necesario utilizar un enfoque metodológico, CRISP-DM es la metodología para llevar a cabo el proceso de minería de datos en forma sistemática y no trivial más utilizada en la actualidad (Moine, Gordillo y Haedo, 2011).

El documento se encuentra organizado de la siguiente manera: en la sección I se presenta el planteamiento del problema y los objetivos tanto generales como específicos, la sección II está conformada por el estado del arte, en la sección III se detalla la metodología, la sección IV muestra el proceso de minería de datos, en la sección V se describen los resultados obtenidos y se realiza una discusión de los mismos, la sección VI y VII corresponde a conclusiones y recomendaciones, la sección VIII contiene las bibliografías utilizadas para realizar el trabajo y por último en la sección IX se adjuntan los anexos correspondientes al trabajo de investigación.

CAPÍTULO I.
PLANTEAMIENTO DEL PROBLEMA

1.1. Planteamiento del problema y justificación

En Ecuador el tema de minería de datos se encuentra en fase de evolución, sin embargo, existen algunas aplicaciones que están en continuo desarrollo, un ejemplo son los centros de investigación y universidades de Ecuador, quienes han utilizado datos históricos almacenados logrando así que estos aporten al conocimiento en la toma de decisiones (Camana, 2016). Un ejemplo de esto es el estudio realizado en la Universidad de Loja el cual consiste en determinar el rendimiento académico de los estudiantes aplicando algoritmos de predicción (Encarnación, 2014).

En la Universidad Nacional de Chimborazo existen grandes cantidades de información de docentes, estudiantes y producción científica almacenada en la plataforma académica Sicoa, de la cual se puede generar información nueva y útil mediante un proceso de minería de datos, y así brindar un apoyo al desarrollo de la institución.

En este contexto, se busca realizar un análisis del algoritmo de árboles de decisión con la mayor exactitud aplicados sobre los datos personales y académicos de docentes, estudiantes y producción científica, con la finalidad de establecer que valores de las variables independientes determinan cierto resultado de la variable dependiente en cada una de las tablas.

1.2. Objetivos

1.2.1. Objetivo general.

Identificar el mejor algoritmo de árboles de decisión aplicado en la base de datos del sistema académico de la Unach por medio del análisis comparativo de la exactitud, para apoyar al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigación, utilizando minería de datos".

1.2.2. Objetivos específicos.

- Utilizar la metodología CRISP-DM para el análisis, preparación y construcción de los datos académicos y personales de estudiantes, docentes y producción científica.
- Aplicar los algoritmos de árboles de decisión ID3 y J48 para establecer las principales reglas de clasificación sobre los datos de estudiantes, docentes y producción científica.
- Evaluar la exactitud de la clasificación de los algoritmos ID3 y J48 para establecer el algoritmo con mejor desempeño y así apoyar el proceso de toma de decisiones.

CAPÍTULO II.
MARCO TEÓRICO

2.1. Minería de datos

La minería de datos es definida como el proceso de descubrir conocimientos interesantes como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de información almacenada en bases de datos, datawarehouse, o cualquier otro medio de almacenamiento de información (Moreno, Baturone, Sanchéz y Barriga, 2008).

El objetivo principal de la Minería de Datos consiste en extraer información y transformarla en una estructura comprensible para su posterior uso, donde para ello se debe seguir un proceso de preparación y exploración de los datos para obtener la información oculta en ellos, esencialmente surge para ayudar a comprender el contenido de un repositorio de datos (Encarnación, 2014).

La minería de datos hace hincapié principalmente en lo siguiente, tal como indica (Riquelme, Ruiz y Gilbert, 2006):

- La estabilidad de atributos e instancias existentes.
- Algoritmos y arquitecturas los cuales proporcionan la estadística, los métodos y las formulaciones.
- La automatización para manejar grandes cantidades de datos.

2.2. Aplicaciones de la minería de datos

La minería de datos prácticamente se puede aplicar en todas las actividades que generen datos, así lo señala (Riquelme et al., 2006):

- Comercio y banca: segmentación de clientes, previsión de ventas, análisis de riesgos.
- Medicina y farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.

- Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas, accesos a redes no permitidos, etc.
- Recuperación de información no numérica: minería de texto, minería web, búsqueda e identificación de videos, imágenes y texto.
- Astronomía: identificación de nuevas estrellas y galaxias.
- Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o de pesca.

2.3. Técnicas de minería de datos

Las técnicas de minería de datos se pueden clasificar en dos grandes grupos:

Técnicas Predictivas o Supervisadas.

- Clasificación: Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto redefinido de clases) (Escobar et al., 2016).
- Regresión: Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística) (Escobar et al., 2016).
- Redes Neuronales: Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida (Corso, 2009).
- Arboles de Decisión: Es un modelo de predicción que dada una base de datos realiza un diagrama de construcciones lógicas (Corso, 2009).

Técnicas Descriptivas o No Supervisadas.

- Agrupamiento o clustering: Es el proceso de agrupar los datos en clases o clústeres de modo que los datos del mimo clúster tienen una gran similitud. (Encarnación, 2014)

- Reglas de asociación: Es la exploración de los datos con el objetivo de buscar relación entre ellos, dentro de una fuente o base de datos. (Encarnación, 2014)

2.4. Árboles de decisión

Los árboles de decisión son una técnica de minería de datos que explora los datos, los prepara y sondea para extraer la información oculta en ellos (Berlanga, Rubio y Vilá, 2013)**,** además, es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que permite determinar la decisión final que se debe tomar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. (Robles y Sotolongo, 2013)

Los árboles de decisión crean el modelo de clasificación basándose en diagramas de flujo, clasifican los diferentes casos en grupos o a su vez realizar predicciones de una variable dependiente basándose en los valores de las variables independientes (Berlanga et al, 2013).

Las ventajas de un árbol de decisión son (Pérez, 2004) :

- Facilita la interpretación de la decisión adoptada.
- Facilita la comprensión del conocimiento utilizado en la toma de decisiones.
- Explica el comportamiento respecto a una determinada decisión.
- Reduce el número de variables independientes.

2.5. Exactitud en árboles de decisión

La exactitud se refiere a cuán cerca del valor real se encuentran los valores de la predicción (Haro, Zúñiga, Meneses, Vera, & Escudero, 2018). La medida de exactitud de la clasificación se construye a través de una matriz de confusión que registra los siguientes valores (Marcano et al, 2011):

- **Verdaderos positivos (VP):** número de predicciones positivas que fueron clasificadas como positivas.
- **Falsos positivos (FP):** número de predicciones negativas que fueron clasificadas como positivas.
- **Falsos Negativos (FN):** número de predicciones negativas que fueron clasificadas como negativas.
- **Verdaderos Negativos (VN):** número de predicciones positivas que fueron clasificadas como negativas.

Partiendo de la matriz de confusión, se establece la fórmula para el análisis de la exactitud de nuestro modelo de clasificación:

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

Tabla 1
Matriz de confusión.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

2.6. Algoritmo ID3

Fue desarrollado por J. Ross Quinlan en 1983.Su uso se enfoca en la búsqueda de hipótesis o reglas dado un conjunto de ejemplos (Encarnación, 2014). Es un algoritmo que aprende a partir de la diferencia que existe entre los datos para analizar, esto es un procedimiento que maximiza la

información obtenida, la cual se utiliza como una métrica para seleccionar el mejor atributo que divida los datos en clases homogéneas. (Barrientos et al, 2009)

Permite construir un árbol de arriba abajo, de forma directa y sin ejecutar vuelta atrás en su búsqueda. Una vez que el algoritmo selección un atributo, nunca reconsidera esta elección. Los dominios de los atributos y de las clases deben ser discretos. (Encarnación, 2014)

Sus elementos son (Encarnación, 2014):

- Nodos: Los que contienen los atributos.
- Arcos: Contienen los valores posibles del nodo padre.
- Hojas: Son nodos que clasifican el ejemplo como positivo o negativo.

Sus principales características son las siguientes:

- Es recursivo.
- No realiza b*acktracking*, es decir evaluación hacia atrás.
- Encuentra el árbol más sencillo que pueda dividir mejor los atributos.
- Los resultados pueden expresarse como una regla Si-entonces.

Además, el algoritmo ID3 realiza una selección previa de variables denominada *preprunning*, la cual consiste en efectuar una prueba de independencia entra cada variable independiente, de tal manera que para la creación del árbol solo se consideran aquellas variables independientes para las que se rechaza el test de independencia (Barrientos et al, 2009).

Presenta los principales inconvenientes:

- Beneficia indirectamente a los atributos con mayor cantidad de valores, los cuales ni siempre son los más relevantes.
- Genera el árbol de decisión partiendo de ejemplos de partida.

- Presenta conflictos en soluciones diferentes que se alcanzan con variables con valores similares.
- Sus árboles son de gran tamaño, lo cual no es garantía de que sus reglas sean eficientes.
- Se pueden aplicar solo en problemas de clasificación y diagnóstico.

2.7. Algoritmo J48

También conocido como C4.5, este algoritmo fue creado por JR Quinlan en 1993, construye un árbol a partir de datos, se construye iterativamente al ir agregando nodos o ramas que minimicen la diferencia entre los datos. Este algoritmo es un descendiente del ID3 y se extiende en el sentido de su capacidad de utilizar atributos numéricos y vacíos para generar reglas del árbol (Barrientos et al., 2009). Como se dice es una mejora del ID3 se pueden describir estas mejoras (Vizcaíno, 2008):

- En vez de elegir los casos de entrenamiento de forma aleatoria este algoritmo sesga la selección para conseguir una distribución más uniforme de la clase inicial.
- En cuanto al límite de excepciones (casos clasificados incorrectamente) c4.5 incluye el 50% de las excepciones, el resultado es una convergencia más ágil hacia el árbol definitivo.
- C4.5 termina la construcción del árbol sin tener que clasificar los datos de todas las categorías posibles.

Sus principales características son las siguientes (Espino et al., 2015):

- Permite trabajar con atributos de valores continuos y separa los posibles resultados en dos ramas.
- Los árboles que genera con menos frondosos, ya que cada clase no cubre una clase en particular sino toda una distribución de clases.

- Genera el árbol inicial a partir de un grupo de datos designados para el entrenamiento.
- Genera el árbol en base al criterio de proporción de ganancia, de esta manera se evita beneficiar a la variable con mayor número de valores.
- Es recursivo.

La estructura del árbol de este algoritmo se compone de dos partes (Vizcaíno, 2008):

- Una hoja (nodo terminal) que indica una clase.
- Un nodo de decisión, que específica una comprobación a realizar sobre el valor de la variable.
- Rama: contiene los posibles valores del atributo asociado al nodo.

El algoritmo considera todas las posibles pruebas en las que se puede dividir los datos y selecciona la que presenta una mayor ganancia de información. Para un atributo discreto, se considera *n* resultados como prueba, siendo *n* el número de posibles valores que puede contener el atributo. Para los atributos continuos, se realiza una prueba binaria sobre cada valor que toma el atributo. En cada nodo, el sistema decide que prueba dividirá los datos (Espino et al., 2015).

Los tres tipos de pruebas propuestas son (Espino et al., 2015).

- La prueba estándar para las variables discretas, con una rama y un resultado para cada posible valor del atributo.
- Una prueba más compleja para variables discretas, donde se crean grupos, asignando a un determinado grupo cada posible valor de una variable.
- La prueba binaria la cual se realizar a los valores continuos.

2.8. Metodología CRISP–DM

CRISP-DM es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos tal como se puede observar en la Tabla 2. Esta gráfica, publicada por kdnuggets.com representa el resultado de múltiples encuestas aplicada durante los últimos años respecto a la utilización de las principales metodologías para realizar proyectos de minería de datos.

Los orígenes de CRISP-DM, se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases) el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining). (Chapman et al., 2000)

Tabla 2
Metodologías usadas en minería de datos.

Metodología	Porcentaje de uso
CRISP-DM	43 %
Propia	28 %
SEMMA	9 %
Otra, sin dominio específico	8 %
KDD	6 %
De la organización	4 %
Otra, de dominio específico	2 %
Ninguna	0 %

***Fuente**. (KDNuggets, 2014)*

Según (Arancibia, 2009), CRISP-DM está dividida en 4 niveles de abstracción, en tareas que van desde el nivel más general hasta los casos más específicos y organiza el desarrollo de un proyecto de minería de Datos en una serie de seis fases:

1. **Fase de comprensión del negocio o problema.** Es probablemente la fase más importante y reúne las tareas de comprensión de objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables.
2. **Fase de comprensión de los datos.** Comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos.
3. **Fase de preparación de los datos.** Se procede a la preparación de los datos para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.
4. **Fase de modelado.** En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios, dese ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos del problema, tiempo adecuado para obtener un modelo, conocimiento de la técnica.
5. **Fase de evaluación.** En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Además, debe considerarse que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el

proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error.

6. **Fase de implementación.** En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio. ya sea que se recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso.

CAPITULO III.
METODOLOGÍA

3.1. Tipo de Investigación.

Esta investigación tiene un enfoque mixto, debido a que implica la recolección y análisis de datos cualitativos y cuantitativos, así como su integración y discusión conjunta para lograr un mejor entendimiento del fenómeno en estudio (Hernández, Fernández y Baptista , 2010). Se aplicó la metodología CRISP-DM, la cual nos permitió llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. (Moine et al., 2011), esta investigación siguió el proceso que se describe a continuación:

1. Recolección y análisis de los datos: Los datos se obtuvieron de la base de datos del Sistema de Control Académico (SICOA) de la Universidad Nacional de Chimborazo en archivos .xlsx desde el año 2012 hasta la actualidad, en los cuales se almacenó información personal y académica del estudiante y docente, además los resultados de la evaluación aplicada al docente y las publicaciones realizadas por los mismos.

Se utilizó el método analítico y el método de síntesis, debido a que mediante el análisis de la información obtenida se reducirán las variables de estudio.

2. Comprensión de los datos: Se realizó a través de la investigación descriptiva y exploratoria, ya que se efectuó una descripción general de los datos mediante el uso de tablas y gráficos, además, se realizó un análisis de la calidad de los datos obteniendo el número de valores válidos, inválidos y nulos existentes.
3. Preparación de los datos: Se aplicó un método deductivo, ya que, tras una fase de observación repetida se seleccionaron las variables finales con las que se realizó el estudio.
4. Implementación y evaluación de los modelos: se crearon modelos con los cuales se realizó la aplicación de los algoritmos ID3 y J48 sobre los datos seleccionados, para luego proceder

al análisis y evaluación de los resultados obtenidos de los mismos mediante la métrica exactitud, todo este proceso se llevó a cabo en la herramienta Rapid Miner versión 9.1 con una licencia educacional que tiene una duración de un año.

5. Inducción científica: En este paso se aplicó también la investigación explicativa, debido a que, al observar los resultados se estableció que características debe cumplir un estudiante y un docente para obtener cierto tipo de promedio tanto en rendimiento académico, en la evaluación del docente y al momento de realizar publicaciones, colaborando así con el diseño de estrategias para el mejoramiento en la gestión académica e investigativa de la Unach.

3.2. Unidad de análisis

La información necesaria para este proyecto se recolectó del Sistema Informático de Control Académico (SICOA) de la Universidad Nacional de Chimborazo, y del Sistema de Publicaciones de la Dirección de Investigación, esta consiste en información académica, personal y socioeconómica de docentes y estudiantes, además se contó con información de las publicaciones realizadas por los docentes en la institución.

3.3. Técnicas de análisis e interpretación de la información.

El análisis de los datos se lo realizará mediante la técnica árboles de decisión, aplicando los algoritmos ID3 y J48 para obtener la exactitud y las principales reglas de clasificación, además se aplicó la metodología CRISP-DM para llevar a cabo el proceso de minería.

3.4. Aplicación de la metodología CRISP-DM.

En este apartado se detalla cada fase de la metodología CRISP-DM que se aplicó en el proceso de minería de datos. A continuación, se detalla cada uno de los pasos realizados:

3.4.1. Comprensión del Negocio

Determinar el objetivo del negocio

En objetivo de este proyecto es analizar la exactitud de los de algoritmos de árboles de decisión aplicados en la base de datos del Sistema de Control Académico (SICOA) de la Unach, además se generarán las principales reglas de clasificación para cada una de las tablas. Dicho estudio se realizará para apoyar al proyecto "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la institución, utilizando minería de datos".

Evaluación de la Situación

Para el desarrollo del proyecto se dispone de la base de datos del Sicoa proporcionada por las autoridades correspondientes en formato xmls desde el año 2012, la misma que almacena información académica y personal sobre estudiantes, docente y producción científica

Inventario de recursos. Los recursos tanto hardware como software utilizados en el proyecto se pueden revisar en el Anexo 1.

Fuente de datos. La fuente de datos es un archivo en formato xmls el cual contiene información desde el año 2012, este archivo almacena información académica y personal sobre estudiantes, docentes, y producción científica, un total de 16009 registros de estudiantes, 826 registros de docentes y 2051 registros de publicaciones.

Determinar los objetivos de la minería de datos

En este proyecto se han definido los siguientes objetivos:

- Determinar qué características presenta un estudiante para tener un promedio Excelente, Bueno o Insuficiente.
- Establecer qué características presenta un docente según su calificación en la Evaluación a docentes.

- Determinar qué características presenta un docente que realiza publicaciones.

Realizar el plan del proyecto

El plan de proyecto se ha organizado en seis etapas, las cuales se van a realizar en un tiempo total de trece semanas, a continuación, se detalla en que consiste cada una de ellas y el tiempo estimado que llevará realizarlas:

- Etapa 1: Comprensión del Negocio y planteamiento de los objetivos del proceso de minería de datos. Tiempo estimado: 1 semana.
- Etapa 2: Comprensión y Análisis de la información de la base de datos. Tiempo estimado: 2 semanas.
- Etapa 3: Preparación de los Datos (selección, conversión, limpieza, normalización) para facilitar el proceso de minería de datos sobre ellos. Tiempo estimado: 3 semanas.
- Etapa 4: Diseño y ejecución del modelado. Tiempo estimado: 2 semanas.
- Etapa 5: Evaluación de los resultados obtenidos de la etapa anterior. Tiempo estimado: 2 semanas.
- Etapa 6: Producción de informes y presentación de resultados finales. Tiempo estimado: 3 semanas

3.4.2. Comprensión de los datos

En esta segunda fase se realiza la recolección inicial de los datos para poder familiarizarnos con el problema, evaluar su calidad e identificar las relaciones entre ellos.

Recolectar Datos iniciales

Se realizó la recolección inicial de los datos que se encuentran relacionados con el problema de estudio, así mismo se efectúo un análisis de los mismo para identificar las relaciones que existen entre ellos. La información obtenida corresponde a estudiantes y docentes de la Universidad

Nacional de Chimborazo, además existen datos de las publicaciones de producción científica realizadas en la institución. Los datos se obtuvieron del Sistema Académico de la Unach, mismos que fueron almacenados en un archivo de formato xmls tal como se menciona anteriormente.

Descripción de los datos

Los datos recopilados se dividen en varias tablas, mismas que se detallan en el Anexo 2.

Exploración de los datos

Para la exploración de los datos se realizó una revisión de la información que se recolectó de estudiantes, docente y publicaciones de la Unach, misma que se puede observar en el Anexo 3.

Verificar la calidad de los datos

Luego de la exploración de los datos, se realizó un análisis de calidad de los datos, obteniendo valores como datos nulos, datos válidos y datos inválidos tal como se detalla en el Anexo 4.

3.4.3. Preparación de los datos

Selección de los datos

La selección de datos estuvo enfocada en los atributos que permitan cumplir con los objetivos planteados en la primera fase de la metodología (Comprensión del negocio) determinando así la inclusión o exclusión de algunos campos basado en un estudio de calidad de datos, los campos seleccionados se detallan en el apartado Datos finales para el estudio.

Limpieza de los datos

En esta fase se realizó una limpieza de todos los datos recolectados con el objetivo de tratar las inconsistencias encontradas en ellos y así poder realizar un modelo de alta calidad, esta limpieza consiste en eliminar atributos que contienen en su mayoría valores nulos, reducción del volumen de los datos analizando cuales son necesarios para la realización del análisis y completar campos

que se encuentran vacíos. Por esta razón se eliminaron varios de los atributos de las tablas, los cuales se describen en el Anexo 5.

Construir los datos

Atributos derivados y generados. En este apartado se describen los campos que han sido generados a partir de otros ya existentes, además, a partir de los atributos Promedio y ResultadoFinalEvaluación y respectivamente se generaron nuevos atributos, mismos que consisten en la ponderación de los valores contenidos en estas variables, estos atributos se describen en el Anexo 6.

Integración de los datos

En esta fase se realizó la integración entre tablas de la base de datos, las integraciones que se hicieron son las siguientes:

- Se relacionó la tabla Estudiante la cual contiene toda la información personal de los estudiantes con la tabla Estudiante_rendimiento la cual contiene toda la información académica de los mismos.
- También se realizó la relación entre las tablas Docente, Docente_infAcadémica y Evaluación_Docente, las cuales contienen la información personal del docente, información académica y el resultado obtenido en la evaluación que se les aplicó.
- Por último, se realizó la respectiva relación entre la tabla Docente, Docente_infAcadémica e Investigación, combinando así la información personal y académica del docente con la información obtenida sobre las publicaciones.

Formateo de los datos

No ha sido necesario el formateo de los datos, ya que se utilizaron en la estructura original en la que fueron obtenidos, ya que la técnica aplicada en este caso (árboles de decisión) no requerían de esto.

Datos finales para el estudio

Tabla 3
Datos del estudiante.

Estudiantes		
Atributos	**Tipo**	**Naturaleza**
EstudianteID	Código	Original
EstadoCivil	Independiente	Original
Género	Independiente	Original
ActividadDeportiva	Independiente	Derivado
ActividadCultural	Independiente	Derivado
EsForáneo	Independiente	Derivado
TieneHermanos	Independiente	Derivado
Trabaja	Independiente	Derivado
TieneHijos	Independiente	Derivado
Promedio	Independiente	Calculado
PonderaciónPromedio	Dependiente	Derivado

Los valores de la variable dependiente se asignaron en base a la equivalencia correspondiente a la nota del promedio que el estudiante tenga, tal como se detalla en la Tabla 88 del Anexo 6.

Tabla 4
Datos del docente.

Docentes		
Atributos	**Tipo**	**Naturaleza**
Cedula	Código	Original
EstadoCivil	Independiente	Original
Género	Independiente	Original
NivelInstrucción	Independiente	Original
EventosNacionales	Independiente	Derivado

EventosInternacionales	Independiente	Derivado
HorasActividadAcadémica	Independiente	Derivado
ExperienciaPrivada	Independiente	Original
ExperienciaPública	Independiente	Original
ResultadoFinalEvaluaciónDocente	Independiente	Calculado
EquivalenciaCalificación	Dependiente	Derivado

Los valores de la variable dependiente se asignaron en base a la equivalencia correspondiente a la nota final que el docente tenga en la evaluación que se les realiza, tal como se detalla en la Tabla 89 del Anexo 6.

Tabla 5
Datos de investigación.

Investigación		
Atributos	**Tipo**	**Naturaleza**
Cedula	Código	Original
EstadoCivil	Independiente	Original
Género	Independiente	Original
NivelInstrucción	Independiente	Original
EventosNacionales	Independiente	Derivado
HorasActividadAcadémica	Independiente	Derivado
HorasClase	Independiente	Derivado
EventosInternacionales	Independiente	Derivado
ExperinciaPrivada	Independiente	Original
ExperienciaPública	Independiente	Original
TienePublicaciones	Dependiente	Calculado

Los valores de la variable dependiente se calcularon en base a la cantidad de publicaciones que haya realizado el docente, asignando el valor de “No” en caso de que no tenga ninguna publicación y “Si” en caso de si tenerlas.

3.4.4. Modelado

En esta fase se escoge la técnica de minería de datos que se aplicará al conjunto de datos que han sido previamente seleccionados, la cuál debe ser la más apropiada para así lograr cumplir con los objetivos planteados.

Selección de técnica de modelado

En esta tarea se seleccionó la técnica de árboles de decisión específicamente, para lo cual se aplicarán los algoritmos ID3 y J48, ya que el proyecto se basa principalmente en el estudio y aplicación de estos.

Generar plan de prueba

La validez de los modelos se realizará mediante la validación cruzada, la cual divide el total de datos en dos grupos, el 70 % de los datos se utilizarán para el entrenamiento y el 30 % restante para las pruebas. Por otra parte, para medir la calidad de los mismos se utilizará la exactitud (accuracy) la cual expresa el porcentaje de valores que han sido correctamente clasificados, esta medida se obtiene automáticamente de la herramienta rapid miner al momento de ejecutar el modelo.

Construir el modelo

A continuación, se realizó la ejecución de los modelos realizados sobre los datos seleccionados para el estudio. En este apartado se muestran los modelos ejecutados en cada una de las tablas de datos para cada uno de los algoritmos tal como se muestra en el Anexo 7. Ya que se han definido tres objetivos para la minería de datos, esta sección se dividirá en tres partes, una por cada objetivo planteado.

Estudiante. En este caso el campo sobre el cual queremos hacer la predicción es "PonderaciónPromedio", el cual contiene la ponderación del promedio final del estudiante. En cuanto a los parámetros empleados se utilizaron los campos definidos en el apartado de la metodología (Datos finales para el estudio) para la tabla Estudiantes.

Docente. En este caso el campo sobre el cual queremos hacer la predicción es "EquivalenciaCalificación", el cual señala la ponderación de la calificación final de la evaluación

aplicada al docente. En cuanto a los parámetros empleados se utilizaron los campos definidos en el apartado de la metodología (Datos finales para el estudio) para la tabla Docentes.

Investigación. En este caso el campo sobre el cual queremos hacer la predicción es "TienePublicaciones", en el que se indica si un docente ha realizado publicaciones o no. En cuanto a los parámetros empleados se utilizaron los campos definidos en el apartado de la metodología (Datos finales para el estudio) para la tabla Investigación.

Evaluar el modelo

En esta sección se detalla la exactitud obtenida por cada uno de los modelos realizados, estos valores se describen en el Capítulo V. Resultados y discusión.

3.4.5. Evaluación

Evaluar Resultados

En este paso, luego de haber analizado el algoritmo que posee un valor de exactitud mayor, se procedió a calcular los pesos de cada una de las variables y se obtuvieron las principales reglas de clasificación de cada tabla, estos resultados se detallan en el Capítulo V. Resultados y Discusión.

Revisar el proceso

El proceso hasta este punto se ha ejecutado sin ningún inconveniente, ya que los tres modelos realizados para presenta resultados consistentes.

CAPÍTULO IV.
RESULTADOS Y DISCUSIÓN

4.1. Resultados

Para determinar los resultados de este proyecto, se utilizaron las tablas estudiantes, docentes e investigación mismas que se detallan a continuación:

Estudiantes: Al momento de la recolección de los datos, esta tabla contenía un total de 16009 registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019, una vez realizada la comprensión y preparación de los datos se disminuyó a 15793 registros que corresponden al 98.65% del total de los datos, los campos utilizados son EstudianteId, EstadoCivil, Género, EsForáneo, Trabaja, TieneHijos, TieneHermanos, ActividadDeportiva, ActividadCultural, Promedio, PonderaciónPromedio.

- **Estudiantes de Facultad de Ingeniería.**

En este análisis el algoritmo J48 obtuvo un total de 79,35 % de exactitud en su clasificación por encima del 78,39 % de exactitud obtenido por el algoritmo ID3 tal como se detalla en la Tabla 6, por lo tanto, el algoritmo J48 será el que se aplique para generar las principales reglas de clasificación de los datos de los estudiantes de la Facultad de Ingeniería.

Tabla 6
Exactitud de los algoritmos en datos de estudiantes de Ingeniería.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	78,39	22,61
J48	79,35	20,65

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 7 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 8.

Tabla 7
Matriz de confusión algoritmo ID3 en datos de estudiantes de Ingeniería.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	1320	305	0
pred. Bueno	527	1730	10
pred. Excelente	0	0	6

Tabla 8
Matriz de confusión algoritmo J48 en datos de estudiantes de Ingeniería.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	1317	389	0
pred. Bueno	400	1776	11
pred. Excelente	0	0	5

Las características de los estudiantes de la Facultad de Ingeniería según su promedio académico se describen a continuación:

Tabla 9
Mejor regla de clasificación de estudiantes de Ingeniería según su promedio.

	Promedio		
	Excelente	Bueno	Insuficiente
Género	Masculino	Masculino	Masculino
Estado Civil	Soltero	Soltero	Soltero
Es foráneo	Si	No	No
Trabaja	No	No	No
Tiene hijos	No	No	No
Tiene hermanos	Si	Si	No
Actividad Deportiva	No	Si	Si
Actividad Cultural	No	Si	Si

Tabla 10
Porcentaje de probabilidad de un estudiante de Ingeniería de obtener un promedio excelente, bueno o insuficiente.

Promedio	**% probabilidad**
Excelente	0,41
Bueno	55,54
Insuficiente	44,05

Análisis: Los estudiantes de ingeniería tienen una gran probabilidad de obtener un promedio insuficiente y una probabilidad mínima de obtener un promedio excelente, aquellos que obtienen un promedio excelente presentan como principal característica ser foráneos y no practicar actividades deportivas ni culturales.

- **Estudiantes de Facultad de Ciencias de la Educación.**

Tal como se puede observar en la Tabla 11, en este análisis el algoritmo J48 obtuvo un 90,06 % de exactitud en su clasificación y el algoritmo ID3 un 88,58 % de exactitud, por lo tanto, las principales reglas de clasificación de los datos de los estudiantes de la Facultad de Ciencias de la Educación se realizarán mediante la aplicación del algoritmo J48.

Tabla 11
Exactitud de los algoritmos en datos de estudiantes de Ciencias de la Educación.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	88,58	11,42
J48	90,06	9,94

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 12 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 13.

Tabla 12
Matriz de confusión algoritmo ID3 en datos de estudiantes de Ciencias de la Educación.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	250	9	0
pred. Bueno	113	2013	203
pred. Excelente	3	44	625

Tabla 13
Matriz de confusión algoritmo J48 en datos de estudiantes de Ciencias de la Educación.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	280	0	0
pred. Bueno	86	2066	238
pred. Excelente	0	0	590

Las características de los estudiantes de la Facultad de Ciencias de la Educación según su promedio académico se describen a continuación:

Tabla 14
Mejor regla de clasificación de estudiantes de Ciencias de la Educación según su promedio.

	Promedio		
	Excelente	Bueno	Insuficiente
Género	Femenino	Masculino	Femenino
Estado Civil	Soltero	Soltero	Soltero
Es foráneo	No	No	Si
Trabaja	No	No	Si
Tiene hijos	No	Si	No
Tiene hermanos	Si	No	No
Actividad Deportiva	No	Si	Si
Actividad Cultural	No	Si	Si

Tabla 15
Porcentaje de probabilidad de un estudiante de Ciencias de la Educación de obtener un promedio excelente, bueno o insuficiente.

Promedio	**% Probabilidad**
Excelente	25,40
Bueno	63,37
Insuficiente	11,23

Análisis: Los estudiantes de la Facultad de Ciencias de la Educación tienen una gran probabilidad de obtener un promedio bueno resaltando como principales características que sea de género masculino y que tenga hijos, además las más opcionadas a obtener un promedio excelente son las mujeres que no practiquen actividades deportivas ni culturales.

- **Estudiantes de Facultad de Ciencias Políticas y Administrativas.**

En este análisis luego de la aplicación de los algoritmos ID3 y J48 se pudo observar un mejor desempeño por parte del algoritmo J48, el cual obtuvo un valor de 89,89 % de exactitud en la clasificación por encima del 86,22 % obtenido por el algoritmo ID3 como se observa en la Tabla 16, basado en estos resultados el algoritmo aplicado para establecer las principales reglas de clasificación sobre los datos de los estudiantes de la Facultad de Ciencias Políticas y Administrativas es el J48.

Tabla 16
Exactitud de los algoritmos en datos de estudiantes de Ciencias Políticas y Administrativas.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	86,22	13,78
J48	89,89	10,11

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 17 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 18.

Tabla 17
Matriz de confusión algoritmo ID3 en datos de estudiantes de Ciencias Políticas y Administrativas.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	577	123	1
pred. Bueno	261	2571	120
pred. Excelente	0	2	25

Tabla 18
Matriz de confusión algoritmo J48 en datos de estudiantes de Ciencias Políticas y Administrativas.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	625	26	0
pred. Bueno	213	2680	106
pred. Excelente	0	0	30

Las características de los estudiantes de la Facultad de Ciencias Políticas y Administrativas según su promedio académico se describen a continuación:

Tabla 19
Mejor reglas de clasificación de estudiantes de Ciencias Políticas según su promedio.

	Promedio		
	Excelente	Bueno	Insuficiente
Género	Femenino	Femenino	Masculino
Estado Civil	Soltero	Soltero	Soltero
Es foráneo	Si	No	No
Trabaja	No	No	No
Tiene hijos	No	No	No
Tiene hermanos	Si	Si	No
Actividad Deportiva	No	No	Si
Actividad Cultural	No	No	No

Tabla 20
Porcentaje de probabilidad de un estudiante de Ciencias Políticas de obtener un promedio excelente, bueno o insuficiente.

Promedio	% Probabilidad
Excelente	3,97
Bueno	73,26
Insuficiente	22,77

Análisis: En la Facultad de Ciencias Políticas y Administrativas los estudiantes tienen una mayor probabilidad de obtener un promedio bueno y una mínima probabilidad de obtener un promedio excelente, aquellas estudiantes de género femenino son las más opcionadas a obtener los mejores promedios teniendo en común el hecho de no trabajar y no tener hijos, siendo irrelevante el hecho de ser o no foráneo

- **Estudiantes de Facultad de Ciencias de la Salud.**

En este análisis el algoritmo J48 obtuvo un total de 82,42 % de exactitud en su clasificación por encima del 81,60 % de exactitud obtenido por el algoritmo ID3 tal como se detalla en la Tabla 21, por lo tanto, el algoritmo J48 será el que se aplique para generar las principales reglas de clasificación de los datos de los estudiantes de la Facultad de Ciencias de la Salud.

Tabla 21
Exactitud de los algoritmos en datos de estudiantes de Ciencias de la Salud.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	81,60	18,40
J48	82,42	17,58

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 22 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 23.

Tabla 22
Matriz de confusión algoritmo ID3 en datos de estudiantes de Ciencias de la Salud.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	7	35	0
pred. Bueno	738	3972	111
pred. Excelente	1	12	0

Tabla 23
Matriz de confusión algoritmo J48 en datos de estudiantes de Ciencias de la Salud.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	0	0	0
pred. Bueno	746	4019	111
pred. Excelente	0	0	0

Las características de los estudiantes de la Facultad de Ciencias de la Salud según su promedio académico se describen a continuación:

Tabla 24
Mejor regla de clasificación de estudiantes de Ciencias de la Salud según su promedio.

	Promedio		
	Excelente	Bueno	Insuficiente
Género	Masculino	Femenino	Femenino
Estado Civil	Soltero	Soltero	Soltero
Es foráneo	Si	No	No
Trabaja	No	No	Si
Tiene hijos	No	No	No
Tiene hermanos	No	No	No
Actividad Deportiva	Si	No	No
Actividad Cultural	No	No	No

Tabla 25
Porcentaje de probabilidad de un estudiante de Ciencias de la Salud de obtener un promedio excelente, bueno o insuficiente.

Promedio	**% Probabilidad**
Excelente	2,23
Bueno	82,43
Insuficiente	15,30

Análisis: En la Facultad de Ciencias de la Salud los estudiantes que tienen un promedio excelente son en su mayoría de género masculinos, que sean foráneos y practiquen una actividad deportiva, pero a nivel general los estudiantes de esta facultad tienen una mayor probabilidad de obtener un

promedio bueno, además se demuestra que existen un índice muy bajo de estudiantes con promedios insuficientes.

Docentes: Al momento de la recolección de los datos , esta tabla contenía un total de 826 registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019, una vez realizada la comprensión y preparación de los datos se disminuyó a 448 registros que corresponden al 54,24% del total de los datos, los campos utilizados son Cedula, EstadoCivil, Género, NivelInstrucción, EventosNacionales, EventosInternacionales, HorasActividad Académica, ExperienciaPrivada, ExperienciaPública, ResultadoFinalEvaluación Docente, Equivalencia Calificación.

- **Docentes de Facultad de Ingeniería.**

Tal como se puede observar en la Tabla 26, en este análisis el algoritmo J48 obtuvo un 86,90 % de exactitud en su clasificación y el algoritmo ID3 un 82,76 % de exactitud, por lo tanto, las principales reglas de clasificación de los datos de los docentes de la Facultad de Ingeniería se realizarán mediante la aplicación del algoritmo J48.

Tabla 26
Exactitud de los algoritmos en datos de docentes de Ingeniería.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	82,76	17,24
J48	86,90	13,10

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 27 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 28.

Tabla 27
Matriz de confusión algoritmo ID3 en datos de docentes de Ingeniería.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	0	1	1
pred. Bueno	1	0	7
pred. Excelente	2	13	120

Tabla 28
Matriz de confusión algoritmo J48 en datos de docentes de Ingeniería.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	0	0	0
pred. Bueno	0	0	2
pred. Excelente	3	14	126

Las características de los docentes de la Facultad de Ingeniería según su calificación final en la evaluación que se les aplica se describen a continuación:

Tabla 29
Mejor regla de clasificación de docentes de Ingeniería según su calificación en la evaluación docente.

	Calificación		
	Excelente	Bueno	Insuficiente
EstadoCivil	Unión libre	Unión libre	Unión libre
Género	Femenino	Femenino	Masculino
NivelInstrucción	PhD	Maestría	Maestría
EventosNacionales	Si	Si	Si
HorasActividadAcadémica	Si	No	Si
EventosInternacionales	Si	Si	No
ExperienciaPrivada	Si	No	No
ExperienciaPública	Si	Si	No

Tabla 30
Porcentaje de probabilidad de un docente de Ingeniería de obtener una calificación excelente, buena o insuficiente en la evaluación docente.

Calificación	**% Probabilidad**
Excelente	88,28
Bueno	9,66
Insuficiente	2,06

Análisis: Los docentes de la Facultad de ingeniería tienen mayor probabilidad de obtener una calificación excelente en la evaluación docente, las principales características de los docentes para obtener esta calificación es que tengan título PhD y que cuenten con experiencia en instituciones

privadas, además se puede observar que las docentes de género femenino son las que obtienen mejores calificaciones.

- **Docentes de Facultad de Ciencias de la Educación.**

En este análisis luego de la aplicación de los algoritmos ID3 y J48 se pudo observar un mejor desempeño por parte del algoritmo J48, el cual obtuvo un valor de 84,82 % de exactitud en la clasificación por encima del 81,18 % obtenido por el algoritmo ID3 como se observa en la Tabla 31, basado en estos resultados el algoritmo aplicado para establecer las principales reglas de clasificación sobre los datos de los docentes de la Facultad de Ciencias de la Educación es el J48.

Tabla 31
Exactitud de los algoritmos en datos de docentes de Ciencias de la Educación.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	81,18	18,82
J48	84,82	15,18

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 32 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 33.

Tabla 32
Matriz de confusión algoritmo ID3 en datos de docentes de Ciencias de la Educación.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	0	0	2
pred. Bueno	0	1	3
pred. Excelente	4	11	85

Tabla 33
Matriz de confusión algoritmo J48 en datos de docentes de Ciencias de la Educación.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	0	0	0
pred. Bueno	0	0	0
pred. Excelente	4	12	90

Las características de los docentes de la Facultad de Ciencias de la Educación según su calificación final en la evaluación que se les aplica se describen a continuación:

Tabla 34
Mejor regla de clasificación de docentes de Ciencias de la Educación según su calificación en la evaluación docente.

	Calificación		
	Excelente	Bueno	Insuficiente
EstadoCivil	Casado	Casado	Casado
Género	Masculino	Masculino	Masculino
NivelInstrucción	Maestría	Maestría	Maestría
EventosNacionales	Si	Si	Si
HorasActividadAcadémica	Si	Si	Si
EventosInternacionales	Si	No	No
ExperienciaPrivada	Si	No	No
ExperienciaPública	Si	Si	No

Tabla 35
Porcentaje de probabilidad de un docente de Ciencias de la Educación de obtener una calificación excelente, buena o insuficiente en la evaluación docente.

Calificación	**% Probabilidad**
Excelente	84,91
Bueno	11,32
Insuficiente	3,77

Análisis: En el caso de los docentes de la facultad de Ciencias de la Educación, su calificación en la evaluación docente es independiente del título que tenga, lo que marca una diferencia es que cuente con experiencia en instituciones privadas y haya asistido a eventos internacionales, contar con estas cualidades les brinda una gran probabilidad de obtener una calificación excelente.

- **Docentes de Facultad de Ciencias Políticas y Administrativas.**

En este análisis el algoritmo J48 obtuvo un total de 92,56 % de exactitud en su clasificación por encima del 91,44 % de exactitud obtenido por el algoritmo ID3 tal como se detalla en la Tabla 36, por lo tanto, el algoritmo J48 será el que se aplique para generar las principales reglas de clasificación de los datos de los docentes de la Facultad de Ciencias Políticas y Administrativas.

Tabla 36
Exactitud de los algoritmos en datos de docentes de Ciencias Políticas y Administrativas.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	91,44	8,56
J48	92,56	7,44

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 37 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 38.

Tabla 37
Matriz de confusión algoritmo ID3 en datos de docentes de Ciencias Políticas y Administrativas.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	0	0	0
pred. Bueno	0	3	0
pred. Excelente	1	7	83

Tabla 38
Matriz de confusión algoritmo J48 en datos de docentes de Ciencias Políticas y Administrativas.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	0	0	0
pred. Bueno	0	5	1
pred. Excelente	1	5	82

Las características de los docentes de la Facultad de Ciencias Políticas y Administrativas según su calificación final en la evaluación que se les aplica se describen a continuación:

Tabla 39
Mejor reglas de clasificación de docentes de Ciencias Políticas según su calificación en la evaluación docente.

	Calificación		
	Excelente	Bueno	Insuficiente
EstadoCivil	Casada	Soltero	Divorciado
Género	Femenino	Masculino	Masculino
NivelInstrucción	PhD	Maestría	Maestría
EventosNacionales	Si	Si	Si
ActividadAcadémica	No	Si	Si
EventosInternacionales	Si	No	No
ExperienciaPrivada	Si	Si	No
ExperienciaPública	Si	No	Si

Tabla 40
Porcentaje de probabilidad de un docente de Ciencias Políticas de obtener una calificación excelente, bueno o insuficiente en la evaluación docente.

Calificación	% Probabilidad
Excelente	88,30
Bueno	10,64
Insuficiente	1,06

Análisis: Los docentes de la facultad de Ciencias Políticas y Administrativas que presentan mayor probabilidad de obtener una calificación excelente en la evaluación docente son aquellos que se encuentran casados, tengan un título de PhD, hayan asistido a eventos internacionales y sean de género femenino.

- **Docentes de Facultad de Ciencias de la Salud.**

Tal como se puede observar en la Tabla 41, en este análisis el algoritmo J48 obtuvo un 91,17 % de exactitud en su clasificación y el algoritmo ID3 un 90,19 % de exactitud, por lo tanto, las principales reglas de clasificación de los datos de los docentes de la Facultad de Ciencias de la Salud se realizarán mediante la aplicación del algoritmo J48.

Tabla 41
Exactitud de los algoritmos en datos de docentes de Ciencias de la Salud.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	90,19	9,81
J48	91,17	8,83

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 42 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 43.

Tabla 42
Matriz de confusión algoritmo ID3 en datos de docentes de Ciencias de la Salud.

	true Insuficiente	true Bueno	true Excelente
pred. Insuficiente	7	0	0
pred. Bueno	0	15	6
pred. Excelente	0	4	70

Tabla 43
Matriz de confusión algoritmo J48 en datos de docentes de Ciencias de la Salud.

	true Insuficiente	**true Bueno**	**true Excelente**
pred. Insuficiente	7	0	0
pred. Bueno	0	17	1
pred. Excelente	0	2	75

Las características de docentes de la Facultad de Ciencias de la Salud según su calificación final en la evaluación que se les aplica se describen a continuación:

Tabla 44
Mejor regla de clasificación de docentes de Ciencias de la Salud según su calificación en la evaluación docente.

	Calificación		
	Excelente	Bueno	Insuficiente
EstadoCivil	Casado	Unión libre	Soltera
Género	Masculino	Femenino	Femenino
NivelInstrucción	Maestría	Especialidad	Maestría
EventosNacionales	Si	Si	No
ActividadAcadémica	No	Si	Si
EventosInternacionales	Si	No	No
ExperienciaPrivada	No	No	No
ExperienciaPública	Si	No	Si

Tabla 45
Porcentaje de probabilidad de un docente de Ciencias de la Salud de obtener una calificación excelente, buena o insuficiente en la evaluación docente.

Promedio	**% Probabilidad**
Excelente	74,51
Bueno	18,63
Insuficiente	6,86

Análisis: Los docentes de la facultad de Ciencias de la Salud que presentan mayor probabilidad de obtener una calificación excelente en la evaluación docente son aquellos que se encuentran casados, hayan asistido a eventos internacionales, sean de género masculino y no realicen actividades académicas, su calificación es independiente del título que tengan.

Publicaciones de docentes: Al momento de la recolección de los datos, esta tabla contenía un total de 2051 registros desde el periodo académico septiembre 2012 - marzo 2013 hasta octubre 2018 - marzo 2019, una vez realizada la comprensión y preparación de los datos se disminuyó a

1321 registros que corresponden al 64,41% del total de los datos, los campos utilizados son: Cedula, EstadoCivil, Género, NivelInstrucción, EventosNacionales, EventosInternacionales, HorasActividad Académica, HorasClase ExperienciaPrivada, ExperienciaPública, Resultado FinalEvaluación Docente, EquivalenciaCalificación.

- **Publicaciones de docentes de Facultad de Ingeniería.**

En este análisis luego de la aplicación de los algoritmos ID3 y J48 se pudo observar un mejor desempeño por parte del algoritmo J48, el cual obtuvo un valor de 86,40 % de exactitud en la clasificación por encima del 85,49 % obtenido por el algoritmo ID3 como se observa en la Tabla 46, basado en estos resultados el algoritmo aplicado para establecer las principales reglas de clasificación sobre los datos de las publicaciones de los docentes de la Facultad de Ingeniería es el J48.

Tabla 46
Exactitud de los algoritmos en datos de publicaciones de docentes de Ingeniería.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	85,49	14,51
J48	86,40	13,60

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 47 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 48.

Tabla 47
Matriz de confusión algoritmo ID3 en datos de publicaciones de docentes de Ingeniería.

	true Si	true No
pred. Si	75	21
pred. No	27	208

Tabla 48
Matriz de confusión algoritmo J48 en datos de publicaciones de docentes de Ingeniería.

	true Si	true No
pred. Si	82	25
pred. No	20	204

Las características de los docentes de la Facultad de Ingeniería que realizan o no publicaciones se describen a continuación:

Tabla 49
Mejores reglas de clasificación de docentes de Ingeniería que realizan o no publicaciones.

	Publicaciones	
	Si	No
EstadoCivil	Casado	Casado
Género	Masculino	Masculino
NivelInstrucción	PhD	Maestría
EventosNacionales	Si	No
ActividadAcadémica	Si	Si
HorasClase	Si	Si
EventosInternacionales	Si	No
ExperinciaPrivada	No	No
ExperienciaPública	Si	Si

Tabla 50
Porcentaje de probabilidad de un docente de Ingeniería de realizar o no publicaciones.

Publicaciones	**% Probabilidad**
Si	30,82
No	69,18

Análisis: Los docentes de esta facultad tienen una mayor probabilidad de no realizar publicaciones, las características de un docente que si realiza publicaciones son: que tenga un título de PhD, haya asistido a eventos nacionales e internacionales, independientemente de su género o estado civil.

- **Publicaciones de docentes de Facultad de Ciencias de la Educación.**

En este análisis el algoritmo J48 obtuvo un total de 88,72 % de exactitud en su clasificación por encima del 84,58 % de exactitud obtenido por el algoritmo ID3 tal como se detalla en la Tabla 51, por lo tanto, el algoritmo J48 será el que se aplique para generar las principales reglas de clasificación de los datos de las publicaciones de los docentes de la Facultad de Ciencias de la Educación.

Tabla 51
Exactitud de los algoritmos en datos de publicaciones de docentes de Ciencias de la Educación.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	84,58	15,42
J48	88,72	11,28

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 52 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 53.

Tabla 52
Matriz de confusión algoritmo ID3 en datos de publicaciones de docentes de Ciencias de la Educación.

	true Si	true No
pred. Si	96	25
pred. No	16	129

Tabla 53
Matriz de confusión algoritmo J48 en datos de publicaciones de docentes de Ciencias de la Educación.

	true Si	true No
pred. Si	102	20
pred. No	10	134

Las características de los docentes de la Facultad de Ciencias de la Educación que realizan o no publicaciones se describen a continuación:

Tabla 54
Mejores reglas de clasificación docentes de Ciencias de la Educación que realizan o no publicaciones.

	Publicaciones	
	Si	No
EstadoCivil	Casado	Casado
Género	Masculino	Masculino
NivelInstrucción	Maestría	Maestría
EventosNacionales	Si	No
ActividadAcadémica	Si	Si
HorasClase	Si	No
EventosInternacionales	Si	No
ExperinciaPrivada	Si	Si
ExperienciaPública	Si	No

Tabla 55
Porcentaje de probabilidad de un docente de Ciencias de la Educación de realizar o no publicaciones.

Publicaciones	% Probabilidad
Si	42,11
No	57,89

Análisis: En este caso, si un docente ha asistido a eventos nacionales e internacionales y cuenta con experiencia en instituciones privadas tiene gran probabilidad de realizar publicaciones, pero a nivel general los docentes de esta facultad tienen una mayor probabilidad de no realizar publicaciones.

- **Publicaciones de docentes de Facultad de Ciencias Políticas y Administrativas.**

Tal como se puede observar en la Tabla 56, en este análisis el algoritmo J48 obtuvo un 83,90 % de exactitud en su clasificación y el algoritmo ID3 un 80,97 % de exactitud, por lo tanto, las principales reglas de clasificación de los datos de las publicaciones de los docentes de la Facultad de Ciencias de Ciencias Políticas y Administrativas se realizarán mediante la aplicación del algoritmo J48.

Tabla 56
Exactitud de los algoritmos en datos de publicaciones de docentes de Ciencias Políticas y Administrativas.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	80,97	19,03
J48	83,90	16,10

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 57 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 58.

Tabla 57
Matriz de confusión algoritmo ID3 en datos de publicaciones de docentes de Ciencias Políticas y Administrativas.

	true Si	true No
pred. Si	48	21
pred. No	18	118

Tabla 58
Matriz de confusión algoritmo J48 en datos de publicaciones de docentes de Ciencias Políticas y Administrativas.

	true Si	**true No**
pred. Si	51	18
pred. No	15	121

Las características de los docentes de la Facultad de Ciencias Políticas y Administrativas que realizan o no publicaciones se describen a continuación:

Tabla 59
Mejores reglas de clasificación de docentes de Ciencias Políticas que realizan o no publicaciones.

	Publicaciones	
	Si	No
EstadoCivil	Casado	Casado
Género	Masculino	Masculino
NivelInstrucción	PhD	Maestría
EventosNacionales	Si	Si
ActividadAcadémica	Si	Si
HorasClase	Si	No
EventosInternacionales	No	No
ExperinciaPrivada	No	Si
ExperienciaPública	Si	No

Tabla 60
Porcentaje de probabilidad de un docente de Ciencias Políticas de realizar o no publicaciones.

Publicaciones	**% Probabilidad**
Si	32,20
No	67,80

Análisis: En esta facultad los docentes que más realizan publicaciones son aquellas de título PhD y que cuentan con experiencia en instituciones públicas, su estado civil y su género es irrelevante, sin embargo, existe una mayor probabilidad de que un docente de esta facultad no realice ningún tipo de publicación.

- **Publicaciones de docentes de Facultad de Ciencias de la Salud.**

En este análisis luego de la aplicación de los algoritmos ID3 y J48 se pudo observar un mejor

desempeño por parte del algoritmo J48, el cual obtuvo un valor de 83,61 % de exactitud en la clasificación por encima del 81,90 % obtenido por el algoritmo ID3 como se observa en la Tabla 61, basado en estos resultados el algoritmo aplicado para establecer las principales reglas de clasificación sobre los datos de las publicaciones de los docentes de la Facultad de Ciencias de la Salud es el J48.

Tabla 61
Exactitud de los algoritmos en datos de publicaciones de docentes de Ciencias de la Salud.

Algoritmo	Exactitud (%)	Error de la clasificación (%)
ID3	81,90	18,10
J48	83,61	16,39

Además, se generó la matriz de confusión de la clasificación del algoritmo ID3 tal como se puede observar en la Tabla 62 y de la misma forma la matriz de confusión del algoritmo J48 tal como se detalla en la Tabla 63.

Tabla 62
Matriz de confusión algoritmo ID3 en datos de publicaciones de docentes de Ciencias de la Salud.

	true Si	true No
pred. Si	74	41
pred. No	54	356

Tabla 63
Matriz de confusión algoritmo J48 en datos de publicaciones de docentes de Ciencias de la Salud.

	true Si	true No
pred. Si	60	18
pred. No	68	379

Las características de docentes de la Facultad de Ciencias de la Salud que realizan o no publicaciones se describen a continuación:

Tabla 64
Mejores reglas de clasificación de docentes de Ciencias de la Salud que realizan o no publicaciones.

	Publicaciones	
	Si	No
EstadoCivil	Casada	Casado
Género	Femenino	Masculino
NivelInstrucción	Maestría	Maestría
EventosNacionales	Si	No
ActividadAcadémica	Si	Si
HorasClase	Si	Si
EventosInternacionales	Si	No
ExperinciaPrivada	No	No
ExperienciaPública	Si	Si

Tabla 65
Porcentaje de probabilidad de un docente de Ciencias de la Salud de realizar o no publicaciones.

Publicaciones	**% Probabilidad**
Si	22,91
No	77,09

Análisis: Los docentes de la facultad de Ciencias de la Salud que más realizan publicaciones son aquellas de género femenino y que hayan asistido a eventos nacionales e internacionales, el titulo que tenga es irrelevante en este caso, sin embargo, existe mayor probabilidad de que los docentes a nivel general no realicen publicaciones.

Evaluación de la exactitud

Luego de haber realizado cada uno de los análisis, para determinar cuál era el algoritmo más exacto se realizó un promedio general de cada uno de los valores obtenidos, con lo cual se obtuvo los resultados que se detallan en la Tabla 66:

Tabla 66
Promedio general de exactitud de cada uno de los algoritmos.

	Exactitud (%)					
	J48			**ID3**		
Datos	**Mínimo**	**Máximo**	**Media**	**Mínimo**	**Máximo**	**Media**
Estudiantes	79,35	90,06	85,43	78,39	88,58	83,70
Docente	84,82	92,56	88,86	81,18	91,44	86,39
Publicaciones	83,61	88,72	85,66	80,97	85,49	83,24
Promedio general			**86,65**			**84,44**

4.2. Discusión

El objetivo principal de la investigación es analizar la exactitud de los algoritmos J48 e ID3 para así determinar cuál de los dos es el más exacto, para esto se realizó 12 análisis, 4 sobre la tabla Estudiantes, 4 sobre la tabla Docentes y 4 sobre la tabla Publicaciones, obteniendo valores de exactitud en cada uno de estos análisis, con los cuales al final se realizó un promedio para calcular el rendimiento general de cada uno de estos algoritmos.

El resultado de este promedio entregó como resultado que el algoritmo J48 tuvo un total de 86,65 % de exactitud y el algoritmo ID3 un 84,44 % de exactitud, demostrando así que el algoritmo J48 es más exacto en sus predicciones tal y como lo menciona (Encarnación, 2014) en su trabajo *Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos*, el cual de la misma forma busca predecir el rendimiento académico de un estudiante de la Universidad de Loja, luego de realizar un análisis de los resultados de varios algoritmos de predicción determina que el algoritmo J48 es el más exacto.

5. CONCLUSIONES

La metodología CRISP-DM permitió llevar a cabo el proceso de minería de datos de forma ordenada, brindó la facilidad de dividir el proyecto en fases con lo cual se pudo realizar un trabajo más detallado y fácil de entender, estas fases consistieron en el análisis, depuración y construcción de los datos, para esto se realizó un análisis de la calidad de los datos de cada una de las tablas, encontrando así el porcentaje de valores nulos que contenía cada registro con lo cual se llevó a cabo la depuración, luego de esto se realizó la construcción de ciertos datos a partir de otros con lo cual se obtuvo como resultado una base de datos totalmente confiable.

La aplicación de los algoritmos se llevó a cabo sobre los datos de estudiantes, docentes y producción científica de cada facultad es decir 4 por cada tabla, en cada estudio se calculó la exactitud del algoritmo ID3 y J48, se detallaron las principales reglas de clasificación y por último se calculó el porcentaje de probabilidad de obtener cierto valor de la variable dependiente, , con lo cual se identificó que los estudiantes de Ciencias de la Educación poseen una mayor probabilidad de obtener un promedio Excelente, los Docentes de Ingeniería y Ciencias Políticas tienen una mayor posibilidad de obtener una calificación excelente en la evaluación docente y por último los docentes de Ciencias de la Educación e Ingeniería son los que más probabilidad presentan de realizar publicaciones.

Para determinar que algoritmo es más exacto se realizó un total de 12 análisis, 4 sobre la tabla Estudiantes, 4 sobre la tabla Docentes y 4 sobre la tabla Investigación, obteniendo valores de exactitud en cada uno de estos análisis, al final se realizó un promedio general de la exactitud de el cual dio como resultado un 86,65 % de exactitud para el algoritmo J48 y un 84,44 % para el algoritmo ID3, basado en estos resultados se concluye que el algoritmo más exacto es el J48.

6. RECOMENDACIONES

- Al momento de seleccionar los datos, tener muy en cuenta las variables que presenten menos irregularidades para así poder obtener los mejores resultados en el proceso de minería, eso se puede realizar mediante un análisis de calidad de datos.
- Realizar una amplia investigación sobre la interpretación de las gráficas de resultados, ya que suelen ser un poco difíciles de entender y sin la correcta preparación se tendrá algunos problemas al momento de analizarlas.
- Establecer de manera clara los objetivos que se desean alcanzar al final del proceso de minería de datos, para así obtener resultados concisos y que sean de gran aporte para la entidad o el lugar donde se esté aplicando el proyecto.
- Realizar un estudio exhaustivo sobre el manejo de cada una de las herramientas que se utilice, para generar un modelo eficiente y que permita obtener los mejores resultados posibles al momento de su aplicación.

7. REFERENCIAS BIBLIOGRÁFICAS

Arancibia, J. A. (2009). Metodología para la definición de requisitos en proyectos de Data Mining (ER-DM).

Barrientos , R., Cruz, N., Acosta, H., Suárez, I., Gogeascoechea , M., Pavón, P., & Blázquez, S. (18 de Septiembre de 2009). Árboles de decisión como herramienta en el diagnóstico médico. Xalapa, Veracruz, Mexico.

Berlanga, V., Rubio, M. J., & Vilá, R. (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE*, 15.

Camana, R. (2016). Potenciales aplicaciones de la Minería de Datos en Ecuador. *RTE*, 14.

Chapman , P., Clinton , J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.*

Corso, C. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. 11.

Encarnación, D. A. (2014). *Estudio del rendimiento académico aplicando técnicas de minería de datos.* Loja.

Escobar, H., Alcivar, M., & Puris, A. (2016). Aplicaciones de mineria de datos en marketing. *Publicando*, 10.

Espino, J., Tijerina , J., Cedano, M., Amaya, E., Pérez, J., & Chiñas, A. (2015). *Algoritmo C4.5.* Tamaulipas.

Franco, C., Pardo, F., Laborda, R., & Pérez , C. (2017). Aplicación de la técnica de árboles de clasificación y regresión en la valoración ecográfica de los nódulos tiroideos. *RAR*, 11.

García, F. J. (2013). *Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andalu de Medio Ambiente (CEAMA).* Granada.

Haro, S., Zúñiga, L., Meneses, A., Vera, L., & Escudero, A. (2018). Métodos de clasificación en minería de datos meteorológicos. *Perfiles*, 7.

Hernández, R., Fernández , C., & Baptista , P. (2010). *Metodología de la investigación.* México: Interamericana.

KDNuggets. (2014). *KDNuggets*. Obtenido de KDNuggets: www.kdnuggets.com

Marcano, A., Chausa , P., Cáceres , C., García , A., López, R., Tormos, J., & Gomez, E. (2011). Análisis comparativo de algoritmos de aprendizaje para predecir la evolución de pacientes con Daño Cerebral Adquirido. *Caseib*, 4.

Martínez, C. A. (2012). *Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión de ENTEL.* Santiago de Chile, Santiago de Chile, Chile.

Moine, J. M., Gordillo, S., & Haedo, A. S. (2011). *Análisis comparativo de metodologías para la gestión de proyectos de minería de datos* . Rosario.

Molina, L. C. (2012). Data mining: torturando a los datos hasta. 11.

Moreno, F., Baturone, I., Sanchéz, S., & Barriga, Á. (2008). Nuevos algoritmos de clasificación integrados en XFUZZY3. 8.

Pérez, C. (2004). *Técnicas de análisis multivariante de datos.* Madrid: Pearson Education.

Porcel, E., Dapozo, G., & López, M. (2009). Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios. 5.

Pradenas, L., & Parra, C. (2012). Uso de minería de datos para determinar la disponibilidad de una red ip v.4 en una cadena de terminales distribuidos . *PODes*, 14.

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *SciELO*, 8.

Robles, Y., & Sotolongo, A. (2013). Integración de los algoritmos de minería de datos 1R, PRISM e ID3 a PostgreSQL. *Jistem*, 18.

Salazar, J., & López, M. (2016). Aplicación de una metodología adaptada de minería de datos en información del sector público. *UGCiencia*, 14.

Suca, C., Cordova, A., Condori, A., Cayra, J., & Sulla, J. (2016). *Comparación de algoritmos de clasificación para la predicción de casos de obesidad infantil.*

Valero, S., Salvador, A., & García, M. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. 8.

Vizcaíno, P. A. (2008). *Aplicación de Técnicas de Inducción de Árboles de Decisión a problemas de clasificación mediante el uso de WEKA.* Bogotá.

8. ANEXOS

8.1. Anexo 1. Recursos software y hardware

Recursos software.

Tabla 67
Recursos Hardware

Software	Uso
Talend Data Quality Web	Proporciona herramientas para realizar el proceso de calidad de datos.
Rapid Miner Studio 9.1	Proporciona herramientas para realizar las tareas de minería de datos.

Recursos Hardware. El recurso de hardware del que se dispone es un ordenador portátil con las siguientes características:

Tabla 68
Características de Recurso Hardware

Característica	Descripción
Marca	Toshiba
Modelo	Satellite E55a
Procesador	Intel Core I5 4ta 2.50 Gz
Memoria RAM	6 GB
Disco duro	750 GB
Sistema Operativo	Windows 10

8.2.Anexo 2. Descripción de los datos

Tabla 69
Descripción de la tabla Estudiante

Atributos	Tipo	Descripción
EstudianteID	Numérico	Identificador único del estudiante
FechaNacimiento	Fecha	Fecha de nacimiento del estudiante
EstadoCivil	Texto	Estado Civil del estudiante (casado, soltero, divorciado, viudo)
OrientaciónSexual	Texto	Atracción afectiva del estudiante.
Genero	Texto	Tipo de Género del estudiante (Masculino, Femenino)
Etnia	Texto	Etnia de la cual se considera el estudiante.
NacionalidadIndígena	Texto	Indica si algún estudiante proviene de alguna nacionalidad indigna del país.
InstituciónEducativa	Texto	Institución secundaria de donde proviene el estudiante
EnfermedadCatastrófica Extraña	Texto	Indica si el estudiante posee alguna enfermedad.
TipoDiscapacidad	Texto	Indica si el estudiante tiene algún tipo de discapacidad.
ActividadCultural	Texto	Indica si el estudiante realiza alguna actividad cultural.
NumeroIntegrantesHogar	Numérico	Número de integrantes que existen en la familia del estudiante
PaísNacimiento	Texto	País de nacimiento del estudiante.
ProvinciaNacimiento	Texto	Provincia de nacimiento del estudiante
CantónNacimiento	Texto	Cantón de nacimiento del estudiante
PaísProcedencia	Texto	País de donde procede el estudiante
ProvinciaProcedencia	Texto	Provincia de donde procede el estudiante
CantónProcedencia	Texto	Cantón de procedencia del estudiante.
PaísDirección	Texto	Dirección del país donde reside el estudiante.
DirecciónProvincia	Texto	Dirección de la provincia donde reside el estudiante.
DirecciónCantón	Texto	Dirección del cantón donde reside el estudiante.

Parroquia	Texto	Parroquia donde reside el estudiante.
Tipo Parroquia	Texto	Tipo de parroquia de donde procede el estudiante (rural, urbana)
NumeroHermanos	Numérico	Numero de hermanos que tiene el estudiante
IngresosPadre	Numérico	Ingresos mensuales del padre del estudiante.
IngresosMadre	Numérico	Ingresos mensuales de la madre del estudiante.
OcupaciónMadre	Texto	Ocupación de la madre del estudiante
OcupaciónPadre	Texto	Ocupación del padre del estudiante
TotalIngresosPadres	Numérico	Total de ingresos mensuales de los padres del estudiante.
NúmeroDependenIngresos	Numérico	Número de personas que dependen de los ingresos de los padres del estudiante.
TipoVivienda	Texto	Indica si la vivienda es de propiedad del estudiante o es en alquiler.
TipoConstrucción	Texto	Indica si la vivienda del estudiante es de construcción mixta, ladrillo o bloque.
Ocupación	Texto	Indica si el estudiante hace actividades extra aparte de estudiar.
TotalIngresos	Numérico	Ingresos mensuales del estudiante.
NumeroHijos	Numérico	Número de hijos que posee el estudiante
OcupaciónCónyuge	Numérico	Ocupación del conyugue del estúdiate.
IngresosCónyuge	Numérico	Ingresos mensuales del cónyuge del estudiante.
TotalIngresosEstudiante	Numérico	Total de ingresos mensuales del estudiante.
PersonasDependenIngresos	Numérico	Número de personas que dependen de los ingresos del estudiante.

Tabla 70
Descripción de la tabla Estudiante_Rendimiento

Atributos	**Tipo**	**Descripción**
EstudianteID	Numérico	Identificador único del estudiante.
Facultad	Texto	Indica la facultad a la que pertenece el estudiante.
Carrera	Texto	Indica la carrera a la que pertenece el estudiante.

SituaciónActual	Texto	Muestra si el estudiante se encuentra graduado o no.
Nivel	Texto	Semestre del estudiante.
Periodo	Texto	Periodo en el que se matriculo de determinado semestre.
Promedio	Numérico	Promedio general que tuvo en todo ese semestre.

Tabla 71
Descripción de la tabla Docente

Atributos	Tipo	Descripción
Cedula	Texto	Número de cédula del docente.
País	Texto	País de procedencia.
Nacionalidad	Texto	Nacionalidad según el país de procedencia.
FechaNacimiento	Fecha	Fecha de nacimiento del docente.
EstadoCivil	Texto	Estado civil del docente.
Sexo	Texto	Sexo del docente.
Etnia	Texto	Etnia del docente.
TipoSangre	Texto	Tipo de sangre del docente.
GrupoGLBTI	Texto	Grupo LGBTI al que pertenece el docente.
NacionalidadIndígena	Texto	Nacionalidad indígena del docente.
Pais	Texto	País en el que radica actualmente.
Cantón	Texto	Cantón en el que radica actualmente.
Parroquia	Texto	Parroquia en la que radica actualmente.
NúmeroHijos	Numérico	Número de hijos que tiene el docente.
NivelIsntrucción	Texto	Tipo de título académico del docente.
País	Texto	País en el que se obtuvo el título.
TiempoEstudio	Texto	Tiempo que demoró en obtener el título.
Modalidad	Texto	Modalidad de estudio.
Área	Texto	Área en la que obtuvo el título.
Subárea	Texto	Subárea en la que obtuvo el título.
Campo	Texto	Campo en el que obtuvo el título.
EstáCursando	Texto	Si se encuentra estudiando actualmente,
InstituciónEducativa	Texto	Institución educativa en la que se obtuvo el título.
Título	Texto	Título que obtuvo.

NoEventosAprobados	Numérico	Numero de eventos aprobados del docente.
NoEventosAsistidos	Numérico	Numero de eventos asistidos por el docente.
HorasEventosAprobados	Numérico	Horas de eventos aprobados del docente.
HorasEventosAsistidos	Numérico	Horas de eventos asistidos por el docente.
NoEventosNacionales	Numérico	Numero de eventos nacionales del docente.
NoEventosInternacionales	Numérico	Numero de eventos internacionales docente.
ExperienciaPrivada	Texto	Si posee experiencia en entidades privadas.
ExperienciaPública	Texto	Si posee experiencia en entidades públicas.
FamiliarSustituto	Texto	Familiar sustituto en el trabajo.
EnfermedadCatastrófica	Texto	Si posee alguna enfermedad catastrófica.
TieneDiscapacidad	Texto	Si posee alguna discapacidad.
GestaciónLactancia	Texto	Si se encuentra en estado de gestación o lactancia.

Tabla 72
Descripción de la tabla Docente_infAcadémica

Atributos	Tipo	Descripción
NumeroDocumento	Texto	Contiene el número de cedula del docente.
Facultad	Texto	Indica la facultad a la que pertenece el docente.
Carrera	Texto	Indica la carrera a la que pertenece el docente.
Periodo	Texto	Muestra el periodo en que dio clases en esa facultad y carrera el docente.
ActividadAcadémica	Texto	Actividad académica que realiza el docente en la institución.
HorasActividadAcadémica	Numérico	Horas realizadas de actividad académica.
HorasClase	Numérico	Horas de clase impartidas por el docente.

Tabla 73
Descripción de la tabla Evaluación_Docente

Atributos	Tipo	Descripción
UsuarioEvaluado	Texto	Contiene el número de usuario del docente evaluado.
Cedula		Contiene el número de cedula del docente evaluado.

TipoEvaluación	Texto	Muestra el tipo de evaluación que se le realizó al docente (autoevaluación, heteroevaluación, etc.)
Componente	Texto	Indica el componente (docencia, gestión, investigación) en el que fue evaluado el docente.
ResultadoFinal	Texto	Contiene la calificación del docente en cada una de las evaluaciones.
Periodo	Texto	Indica el periodo en el que fue evaluado el docente.

Tabla 74
Descripción de la tabla Investigación

Atributos	Tipo	Descripción
ESTADO PUBLICACION	Texto	Describe el estado actual de la publicación.
TIPO PUBLICACION	Texto	Indica a que tipo pertenece la publicación realizada.
TITULO	Texto	Contiene el título del proyecto de investigación.
REVISTA	Texto	Nombre de la revista en la que fue publicada.
CEDULA	Texto	Número de cedula del docente que realiza la investigación.
ROL INSTITUCION	Texto	Rol que cumple en la institución el docente que realizó la investigación.
SEXO	Texto	Sexo del docente.
TIPO AUTOR	Texto	Si el autor es docente o no.
ORDEN AUTOR	Numérico	Orden de autor en la investigación.
NOMBRES	Texto	Nombres del docente.
APELLIDO MATERNO	Texto	Apellido materno del docente.
APELLIDO PATERNO	Texto	Apellido paterno del docente.
AREA DE INVESTIGACION	Texto	Área en la que se realizó la investigación.
LINEA DE INVESTIGACION	Texto	Línea en la que se realizó la investigación.
CAMPO AMPLIO	Texto	Campo amplio en la que se realizó la investigación.
CAMPO DETALLADO	Texto	Campo detallado en la que se realizó la investigación.
CAMPO ESPECIFICO	Texto	Campo específico en la que se realizó la investigación.
AÑO	Fecha	Año en el que se realizó la investigación.
AÑO-MES DE PUBLICACION	Fecha	Año y mes en el que se realizó la investigación.
AÑO-MES DE REGISTRO	Fecha	Año y mes de registro de la investigación.
AÑO-MES REGISTRO MOD	Fecha	Año y mes de registro de la investigación.
FECHA ACEPTACION	Fecha	Fecha en la que fue aceptada la investigación.
FECHA ACTUALIZACION	Fecha	Fecha en la que fue actualizada la investigación.

FECHA DE REGISTRO	Fecha	Fecha en la que fue registrada la investigación.
FECHA PUBLICACION	Fecha	Fecha en la que fue publicada la investigación.
FACULTAD	Texto	Facultad en la que se realizó la investigación.
CARRERA	Texto	Carrera en la que se realizó la investigación.
Codigo Carrera	Numérico	Código de la carrera en la que se realizó la investigación.
CIUDAD DE PUBLICACION	Texto	Ciudad en la que se realizó la publicación de la investigación.
COMITE CIENTIFICO U ORGANIZADOR	Texto	Comité científico u organizador.
COMITE EDITORIAL O EXPERTO	Texto	Comité editorial o experto.
CONGRESO O SEMINARIO	Texto	Congreso o seminario.
ES EDITORIAL DE PRESTIGIO	Texto	Indica si la editorial es de prestigio o no.
ES EDITORIAL DE PRESTIGIO	Numérico	Indica si la editorial es de prestigio o no.
EXISTE APROBACION DE COMISION	Numérico	Indica si fue aprobado o no por una comisión.
EXISTE COMITE CIENTIFICO U ORGANIZADOR	Numérico	Indica si existe un comité científico u organizador.
EXISTE COMITE EDITORIAL	Numérico	Indica si existe un comité editorial.
EXISTE PROCEDIMIENTO SELECTIVO	Numérico	Indica si existe un procedimiento selectivo
EXISTE REVISION POR PARES EXTERNOS	Numérico	Indica si existe una revisión por parte de pares externos.
FORMA PUBLICACION	Texto	Señala las características de la publicación.
FORMA PUBLICACION EN ARTICULO COMPLETO	Texto	Indica si forma o no publicación en artículo completo.
OBSERVACIONES AUTOR	Texto	Muestran las observaciones del autor.
OBSERVACIONES DE COMISION	Texto	Muestra las observaciones de la comisión.
OBSERVACIONES GENERALES	Texto	Muestra las observaciones generales.
LISTADO DE REVISTAS SENESCYT	Texto	Si se encuentra o no en las revistas del listado de la SENESCYT.
OBSERVACIONES PUBLICACION	Texto	Señalan las observaciones realizadas en la publicación.
DOAJ	Texto	Indica si la publicación se encuentra disponible en DOAJ.
DOI	Texto	Identificador digital de la publicación.
EBSCO	Texto	Indica si la publicación se encuentra disponible en EBSCO.
ESTADO PERSONAL ACADEMICO	Numérico	Estado personal académico del docente.
ISBN	Texto	Código ISBN que lo identifica.
ISI WEB KNOWLEDGE	Texto	Indica si la publicación se encuentra disponible en isi web Knowledge.
ISSN	Texto	Código ISNN que lo identifica.
JSTOR	Texto	Indica si la publicación se encuentra disponible en JSTOR.
LATINDEX	Texto	Indica si la publicación se encuentra disponible en LATINDEX.

LIBROS O CAPITULOS DE LIBROS REVISADOS POR PARES	Numérico	Indica el número de libros o capítulos revisados por pares.
LILACS	Numérico	Indica si la publicación se encuentra disponible en LILACS.
NACIONAL O INTERNACIONAL	Texto	Señala si la publicación es de tipo nacional o internacional.
OAJI	Texto	Indica si la publicación se encuentra disponible en OAJI.
PAIS	Texto	País en el que se realizó la investigación.
PROCEDIMIENTO SELECTIVO	Texto	Señala la resolución final sobre la investigación.
PROQUEST	Texto	Indica si la publicación se encuentra disponible en PROQUEST
REDALYC	Texto	Indica si la publicación se encuentra disponible en REDALYC.
REVISION POR PARES EXTERNOS	Texto	Resolución que tomaron los pares externos.
SCIELO	Texto	Indica si la publicación se encuentra disponible en Scielo.
SCIMAGO JOURNAL RANK	Texto	Indica si la publicación se encuentra disponible en Scielo.
SJR	Numérico	Indica el índice de impacto de la investigación.
PAGINAS	Texto	Número de páginas de la publicación.
VOLUMEN	Numérico	Volumen de la revista en el que fue publicado.
ORGANISMO DE AFILIACION	Texto	Organismo al que se encuentra afiliado el docente.

8.3. Anexo 3. Exploración de los datos

a) Distribución de los estudiantes según su estado civil.

Tabla 75
Distribución de los estudiantes según su estado civil

Estado Civil	Total
SOLTERO (A)	15131
CASADO (A)	589
UNION LIBRE	219
DIVORCIADO (A)	65
VIUDO (A)	4

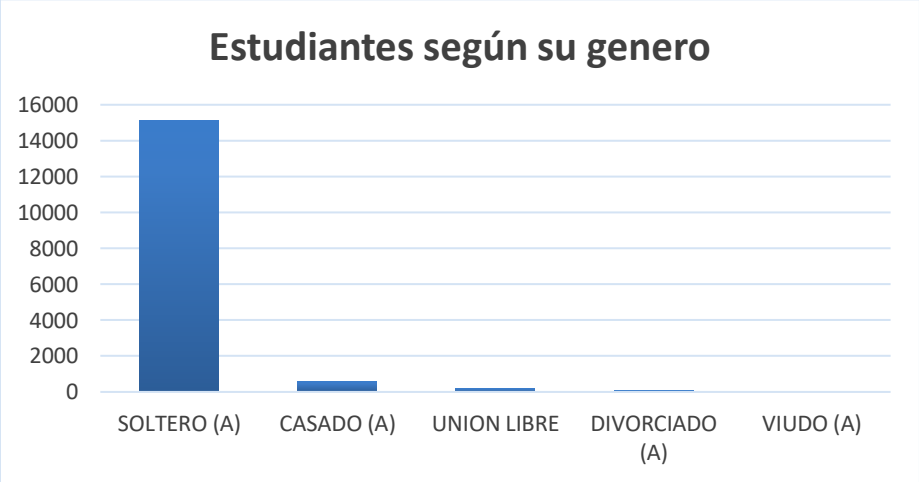


Figura 1. Estudiantes según su estado civil.

b) Distribución de los estudiantes según su sexo.

Tabla 76
Distribución de los estudiantes según su sexo

Género	Total
HOMBRES	8861
MUJERES	7147

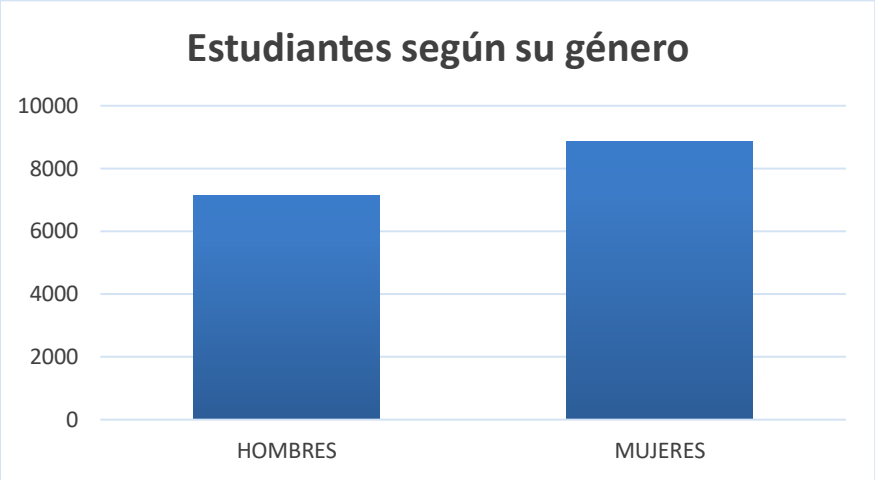


Figura 2. Estudiantes según su sexo.

c) Distribución de los estudiantes por facultad.

Tabla 77
Distribución de los estudiantes según su facultad

Facultad	Total
CIENCIAS DE LA SALUD	4984
INGENIERÍA	4054
CIENCIAS POLÍTICAS Y ADMINISTRATIVAS	3878
CIENCIAS DE LA EDUCACIÓN, HUMANAS Y TECNOLOGÍAS	3341

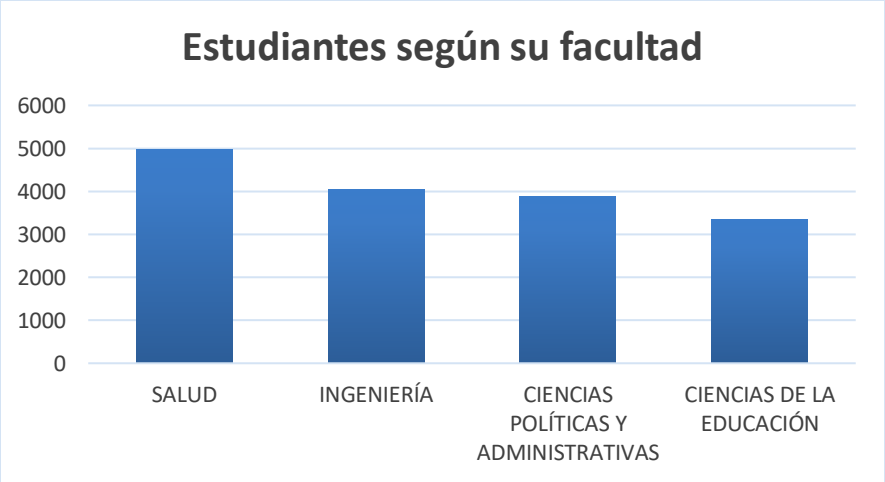


Figura 3. Estudiantes según su facultad.

d) Distribución de los docentes por estado civil.

Tabla 78
Distribución de los docentes según su estado civil.

Estado Civil	Total
SOLTERO (A)	996
CASADO (A)	2620
UNION LIBRE	22
DIVORCIADO (A)	430
VIUDO (A)	29

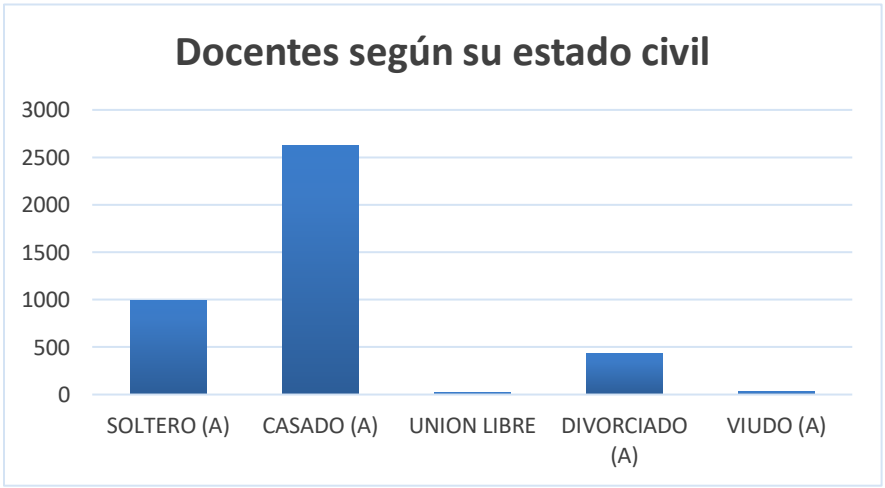


Figura 4. Docentes según su estado civil.

e) Distribución de los docentes por sexo.

Tabla 79
Distribución de los docentes según su sexo.

Género	Total
HOMBRES	2424
MUJERES	1673

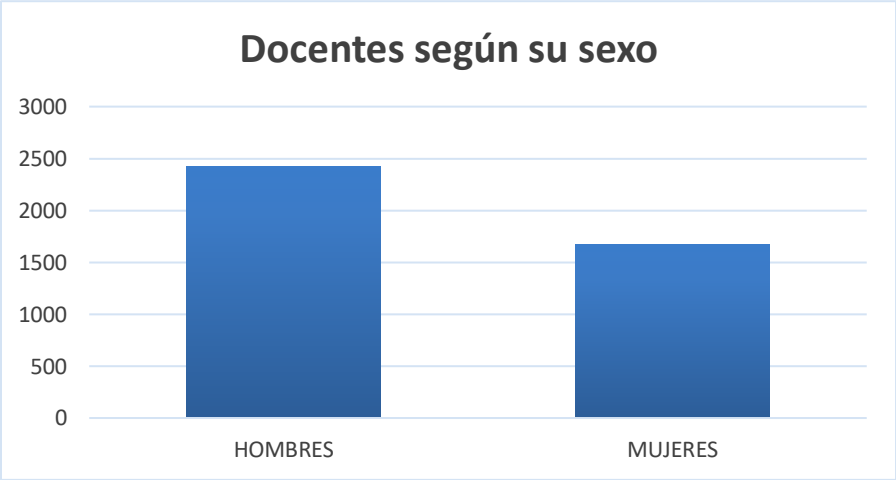


Figura 5. Docentes según su sexo.

f) Distribución de los docentes por facultad.

Tabla 80
Distribución de los docentes según su facultad.

Facultad	Total
CIENCIAS DE LA SALUD	529
INGENIERÍA	360
CIENCIAS POLÍTICAS Y ADMINISTRATIVAS	234
CIENCIAS DE LA EDUCACIÓN, HUMANAS Y TECNOLOGÍAS	297
UNIDAD DE FORMACIÓN ACADÉMICA Y PROFESIONALIZACIÓN	36

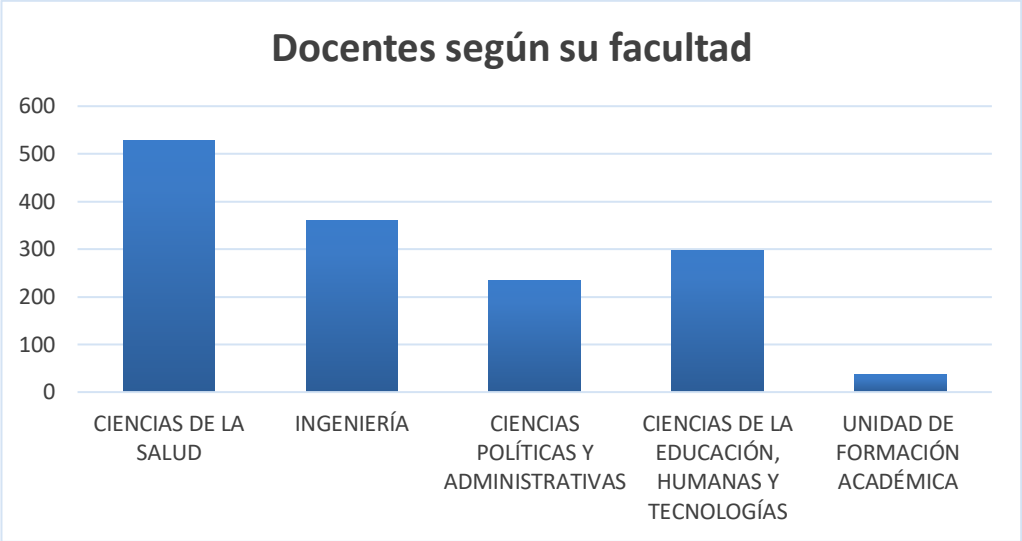


Figura 6. Docentes según su facultad.

g)Distribución de las investigaciones por su estado.

Tabla 81
Distribución de los estudiantes según su estado.

Estado	Total
PUBLICADO	10927
ACEPTADO	599
PROMOCION DOCENTE	92
PATENTE EN ETAPA DE PUBLICACION	12
IN PRESS	61
EN IMPRESION	16

Figura 7. Investigaciones según su estado.

h) Distribución de investigaciones según su tipo.

Tabla 82
Distribución de las investigaciones según su tipo.

Tipo	Total
PONENCIA	3807
INVESTIGACION REGIONAL	3463
PRODUCCION CIENTIFICA	3040
CAPITULO DE LIBRO	1077
LIBRO	341

Figura 8. Investigaciones según su tipo.

8.4.Anexo 4. Análisis de calidad de los datos

Tabla 83
Análisis de calidad de datos.

Tabla	Campo	Valores nulos		Valores válidos		Valores no válidos	
		#	%	#	%	#	%
Estudiante 16008	EstudianteID	0	0,00	15766	98,49	242	1,51
	Porcentaje Discapacidad	15901	99,33	107	0,67	0	0,00
	Numero Integrantes Hogar	1308	8,17	14700	91,83	0	0,00
	Numero Hermanos	66	0,41	15942	99,59	0	0,00
	Ingresos Padre	0	0,00	16008	100,00	0	0,00
	Ingresos Madre	0	0,00	16008	100,00	0	0,00
	Total Ingresos Padres	0	0,00	16008	100,00	0	0,00
	Numero Dependen Ingresos	0	0,00	16008	100,00	0	0,00
	Valor Mensual Servicios	0	0,00	16008	100,00	0	0,00
	Total Ingresos	0	0,00	16008	100,00	0	0,00
	Numero Hijos	0	0,00	16008	100,00	0	0,00
	Ingresos Cónyuge	0	0,00	16008	100,00	0	0,00
	Total Ingresos Estudiante	0	0,00	16008	100,00	0	0,00
	Personas Dependen Ingresos	0	0,00	16008	100,00	0	0,00
	Fecha Nacimiento	0	0,00	16008	100,00	0	0,00
	Estado Civil	0	0,00	16008	100,00	0	0,00
	Orientación Sexual	9	0,06	15999	99,94	0	0,00
	Sexo	0	0,00	16008	100,00	0	0,00
	Género	0	0,00	16008	100,00	0	0,00
	Etnia	9	0,06	15999	99,94	0	0,00
	Nacionalidad Indígena	9	0,06	15999	99,94	0	0,00
	Institución Educativa	0	0,00	16008	100,00	0	0,00
	Tipo	0	0,00	16008	100,00	0	0,00
	Enfermedad Catastrófica Extraña	0	0,00	16008	100,00	0	0,00
	Tipo Discapacidad	9	0,06	15999	99,94	0	0,00
	Actividad Deportiva	1560	9,75	14448	90,25	0	0,00
	Actividad Cultural	1694	10,58	14314	89,42	0	0,00
	País Nacimiento	0	0,00	15974	99,79	34	0,21
	Provincia Nacimiento	0	0,00	16008	100,00	0	0,00
	Cantón Nacimiento	5740	35,86	10268	64,14	0	0,00
	País Procedencia	5519	34,48	10406	65,00	83	0,52
	Provincia Procedencia	5629	35,16	10379	64,84	0	0,00
	Cantón Procedencia	5629	35,16	10379	64,84	0	0,00
	Tipo Parroquia	5416	33,83	10592	66,17	0	0,00
	Ocupación	14896	93,05	16008	100,00	0	0,00
Estudiante Rendimiento 87105	EstudianteID	0	0,00	85685	98,37	1420	1,63
	Facultad	0	0,00	87105	100,00	0	0,00
	Carrera	0	0,00	87105	100,00	0	0,00
	Situación Actual	0	0,00	87105	100,00	0	0,00
	Nivel	0	0,00	87105	100,00	0	0,00
	Período	0	0,00	87105	100,00	0	0,00
	Promedio	296	0,34	86809	99,66	0	0,00
Docente 4097	Cédula	0	0,00	3253	79,40	844	20,6
	País	13	0,32	4046	98,76	38	0,93
	Nacionalidad	3	0,07	4094	99,93	0	0,00

	Fecha Nacimiento	0	0,00	4097	100,00	0	0,00
	Número Hijos	0	0,00	4097	100,00	0	0,00
	Estado Civil	0	0,00	4097	100,00	0	0,00
	Sexo	0	0,00	4097	100,00	0	0,00
	Etnia	550	13,42	3547	86,58	0	0,00
	Tipo Sangre	556	13,57	3541	86,43	0	0,00
	GrupoGLBTI	0	0,00	4097	100,00	0	0,00
	Nacionalidad Indígena	560	16,67	3537	86,33	0	0,00
	Cantón	556	13,57	3541	86,43	0	0,00
	Parroquia	556	13,57	3541	86,43	0	0,00
	Nivel Instrucción	16	0,39	4081	99,61	0	0,00
	Modalidad	1058	25,82	2725	66,51	314	7,66
	Área	1847	45,08	2250	54,92	0	0,00
	Subárea	1847	45,08	2250	54,92	0	0,00
	Campo	1847	45,08	2250	54,92	0	0,00
	Está Cursando	1058	25,82	2725	66,51	0	0,00
	Institución Educativa	1	0,02	4096	99,98	0	0,00
	Titulo	1	0,02	4096	99,98	0	0,00
	Experiencia Privada	0	0,00	4097	100,00	0	0,00
	Experiencia Pública	0	0,00	4097	100,00	0	0,00
	Familiar Sustituto	0	0,00	4097	100,00	0	0,00
	Enfermedad Catastrófica	2116	51,65	1981	48,35	0	0,00
	Tiene Discapacidad	115	2,81	3982	97,19	0	0,00
	Gestión Lactancia	3248	79,28	849	20,72	0	0,00
	Tiempo Estudio	0	0,00	4097	100,00	0	0,00
	Nº Eventos Aprobados	0	0,00	4097	100,00	0	0,00
	Nº Eventos Asistidos	0	0,00	4097	100,00	0	0,00
	Horas Eventos Aprobados	930	22,70	3167	77,30	0	0,00
	Horas Eventos Asistidos	2191	53,48	1906	46,52	0	0,00
	Nº Eventos Nacionales	0	0,00	4097	100,00	0	0,00
	Nº Eventos Internacionales	0	0,00	4097	100,00	0	0,00
	Numero Documento	0	0,00	15479	80,00	3857	20,0
Docente Información Académica 19335	Facultad	0	0,00	19335	100,00	0	0,00
	Carrera	2717	14,00	16619	86,00	0	0,00
	Periodo	0	0,00	19335	100,00	0	0,00
	Actividad Académica	2717	14,00	16619	86,00	0	0,00
	Horas Actividad Académica	0	0,00	19335	100,00	0	0,00
Evaluación Docente 15276	Horas Clase	0	0,00	19335	100,00	0	0,00
	Usuario Evaluado	0	0,00	15276	100,00	0	0,00
	Tipo Evaluación	0	0,00	15276	100,00	0	0,00
	Componente	0	0,00	15276	100,00	0	0,00
	Periodo	0	0,00	15276	100,00	0	0,00
	Resultado Final	0	0,00	15276	100,00	0	0,00

Investigación 12050	Estado Publicación	0	0,00	12050	100,00	0	0,00
	Tipo Publicación	0	0,00	12050	100,00	0	0,00
	Revista	0	0,00	12050	100,00	0	0,00
	Cédula	0	0,00	8270	68,63	3780	31,37
	Rol Institución	4191	34,78	7859	65,22	0	0,00
	Sexo	0	0,00	12050	100,00	0	0,00
	Tipo Autor	0	0,00	12050	100,00	0	0,00
	Orden Autor	0	0,00	12050	100,00	0	0,00
	Nombres	0	0,00	12050	100,00	0	0,00
	Apellido Materno	1408	11,68	10642	88,32	0	0,00
	Apellido Paterno	0	0,00	12050	100,00	0	0,00
	Área de Investigación	22	0,18	12028	99,82	0	0,00
	Línea de Investigación	22	0,18	12028	99,82	0	0,00
	Año	0	0,00	12050	100,00	0	0,00
	Facultad	3949	32,77	8101	67,23	0	0,00
	Carrera	3949	32,77	8101	67,23	0	0,00
	Ciudad Publicación	183	1,52	11867	98,48	0	0,00
	Es Editorial de Prestigio	2050	17,01	10000	82,99	0	0,00
	Existe Aprobación de Comisión	9446	78,39	2604	21,61	0	0,00
	Existe Comité Científico u Organizador	9446	78,39	2604	21,61	0	0,00
	Existe Comité Editorial	9446	78,39	2604	21,61	0	0,00
	Existe Procedimiento Selectivo	9446	78,39	2644	21,61	0	0,00
	Existe Revisión por Pares Externos	9446	78,39	2604	21,61	0	0,00
	Listado de Revistas SENESCYT	2050	17,01	10000	82,99	0	0,00
	Estado Personal Académico	6077	50,43	5973	49,57	0	0,00
	ISBN	7957	66,03	4093	33,97	0	0,00
	ISSN	7066	58,64	4984	41,36	0	0,00
	Nacional o Internacional	2050	17,01	10000	82,99	0	0,00
	Organismo de afiliación	2050	17,01	10000	82,99	0	0,00
	SJR	10855	90,08	1195	9,92	0	0,00
	Volumen	6555	54,40	5495	82,99	0	0,00

8.5. Anexo 5. Atributos eliminados

Tabla 84
Atributos eliminados.

Tabla	Columna
Estudiante	InstituciónEducativa, EnfermedadCatastrófica Extraña, TipoDiscapacidad, Porcentaje Discapacidad, PaisNacimiento, Canton Nacimiento, Parroquia, TipoVivienda, Tipo construcción, ServicioAguaPotable, Servicio electricidad, ServicioTelefono, ServicioInternet, ServicioTV Pagada, ValorMensualServicios, Tiene Vehiculo, OcupaciónConyugue, IngresosConyugue, Personas DependenIngresos.
Estudiante_Rendimiento	EstudianteID, SituacionActual, Nivel.
Docente	TipoSangre, GrupoGLBTI, País, Cantón, Parroquia, País, TiempoEstudio, Modalidad, Área, SubáreaCampo EstáCursando, InstitucionEducativa, Titulo, ExperienciaPrivada, ExperienciaPública, FamiliarSustituto, EnfermedadCatastrófica, TieneDiscapacidad, Gestación Lactancia.
Docente_infAcadémica	NumeroDocumento, ActividadAcadémica.
Evaluación_Docente	UsuarioEvaluado, TipoEvaluación, Componente

8.6. Anexo 6. Atributos derivados y generados

Tabla 85
Atributos derivados.

Tabla	Campo	Descripción
Estudiante	TieneHermanos	Esta transformación consistió en asignar un valor de "Si" a todos los estudiantes que posean hermanos sin importar la cantidad, y "No" aquellos que tienen una cantidad de hermanos de cero.
	Promedio	Esta variable contiene el promedio general de los estudiantes de todo el tiempo de estudio realizado, ya que en la base original solo se tenía los promedios por semestres.
Docente	HorasActividadAcadémica	Esta variable contiene el valor general de las horas de actividad académica de los docentes ya que en la base original solo se tenía las horas por actividad académica, no de forma general.
	ResultadoFinalEvalaución	El contenido de esta variable consiste en el promedio general de la evaluación realizada al docente ya que en la base de datos original se contenía un valor por tipo de evaluación mas no un resultado general.
Investigación	TienePublicaciones	Este campo señala que docentes han realizado publicaciones y quienes no, conteniendo valores únicamente de si o no.
	HorasClase	En esta variable está contenido el valor general de las horas de clase de los docentes ya que en la base original no se tenía una variable que contenga dicho valor.

Tabla 86
Atributos generados.

Tabla	Atributo	Descripción
Estudiante	Ponderación Promedio	En este campo se realiza la ponderación del promedio general del estudiante, basándose en los valores contenidos en la Tabla 20.
Docente	Equivalencia Calificación	Este atributo contiene la ponderación del promedio general de la evaluación realizada al docente, basándose en los valores contenidos en la Tabla 21.

A continuación, se describen las respectivas ponderaciones de la calificación del estudiante en la Tabla 88:

Tabla 87
Ponderación de la calificación del estudiante.

Valor	Rango de la calificación
Excelente	9.00 – 10.00
Bueno	7.00 – 8.99
Insuficiente	Menos de 7.00

En esta sección, en la Tabla 89 se describen las respectivas ponderaciones del resultado a la evaluación del docente:

Tabla 88
Ponderación de la calificación del docente.

Valor	Rango de la calificación
Excelente	90.00 – 100.00
Bueno	70.00 – 89.99
Insuficiente	Menos de 70.00

8.7. Anexo 7. Modelos generados

Modelos de Estudiante

Algoritmo ID3. A continuación, en la Figura 11 se muestra el modelo aplicado para el algoritmo ID3 en la tabla Estudiante.

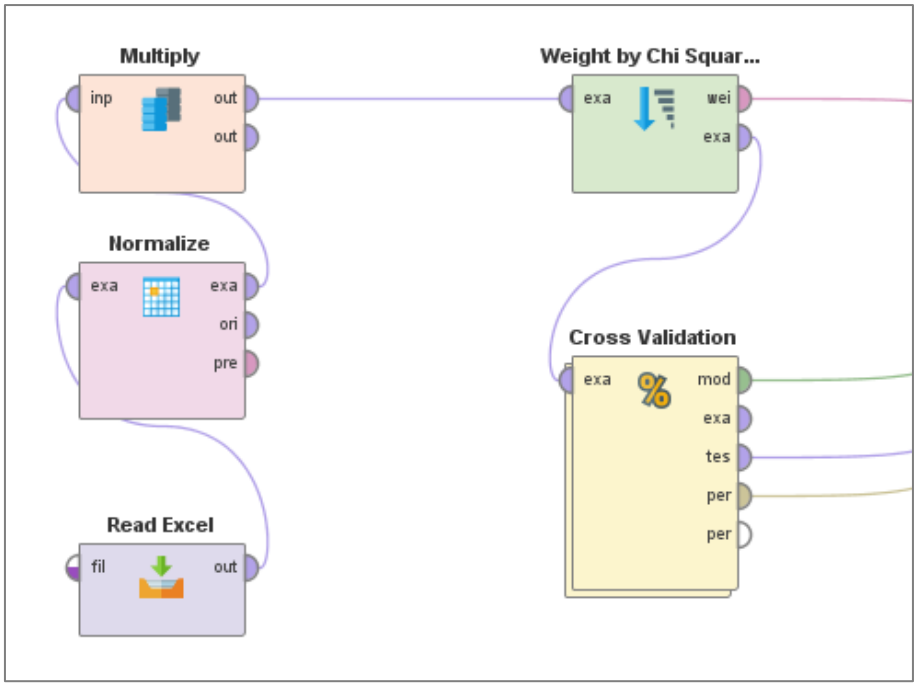


Figura 9. Modelo generado para el algoritmo ID3 en tabla Estudiantes.

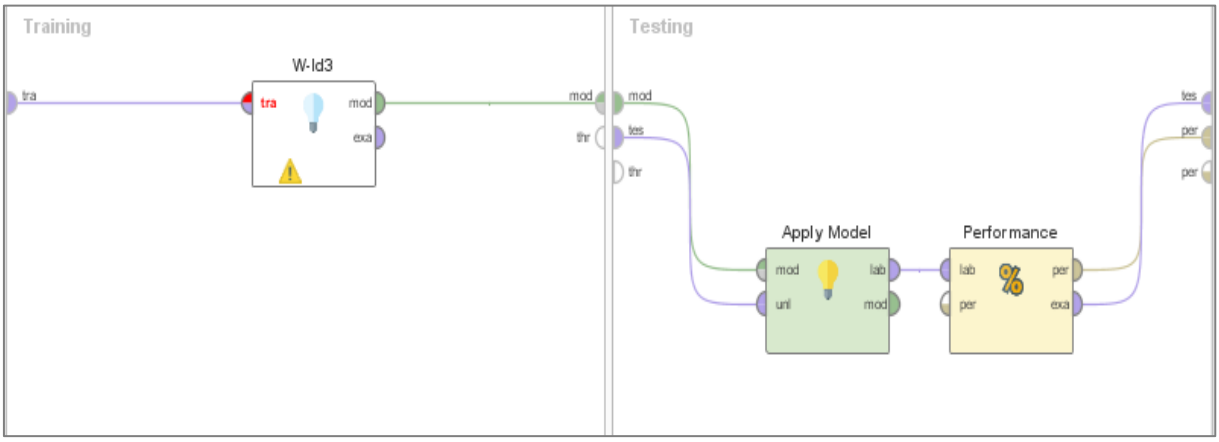


Figura 12. Validación cruzada del algoritmo ID3 en tabla Estudiantes.

Algoritmo J48. A continuación, en la Figura 13 se muestra el modelo aplicado para el algoritmo J48 en la tabla Estudiante.

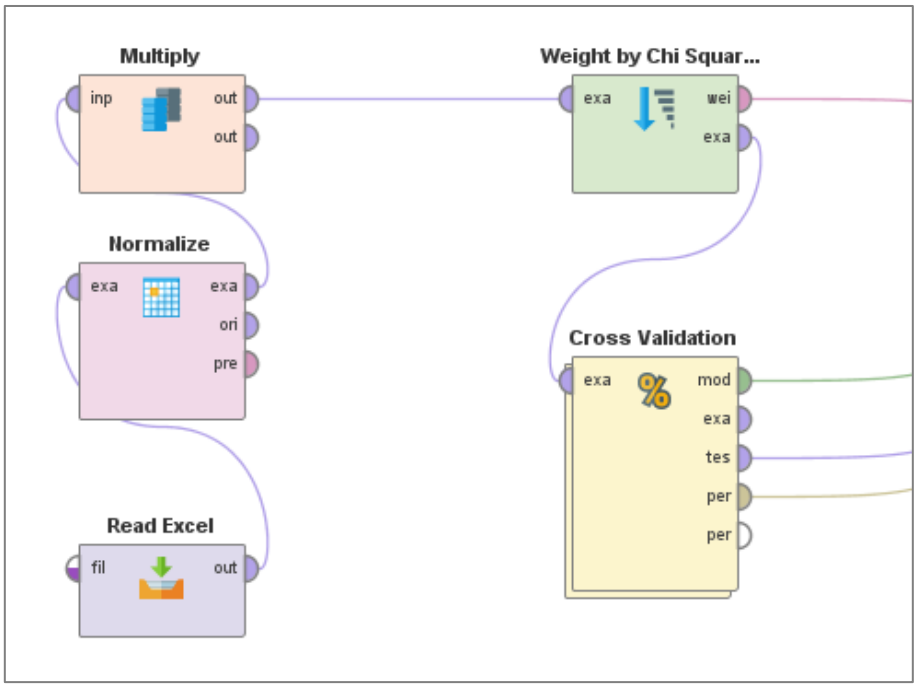


Figura 15. Modelo generado para el algoritmo J48 en tabla Estudiantes.

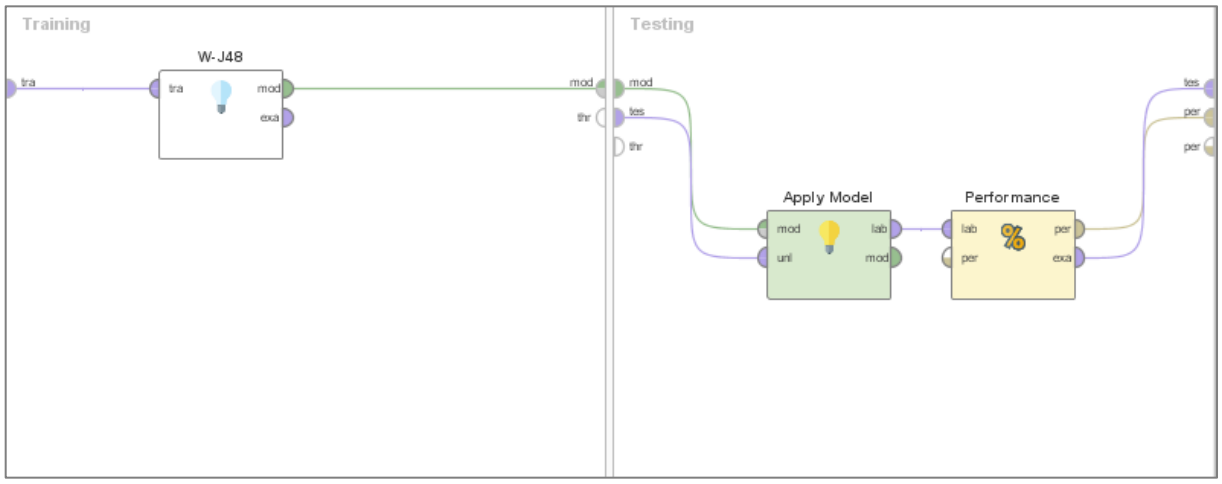


Figura 16. Validación cruzada del algoritmo J48 en tabla Estudiantes.

Modelos de tabla Docente

Algoritmo ID3. A continuación, en la Figura 15 se muestra el modelo aplicado para el algoritmo ID3 en la tabla Docente.

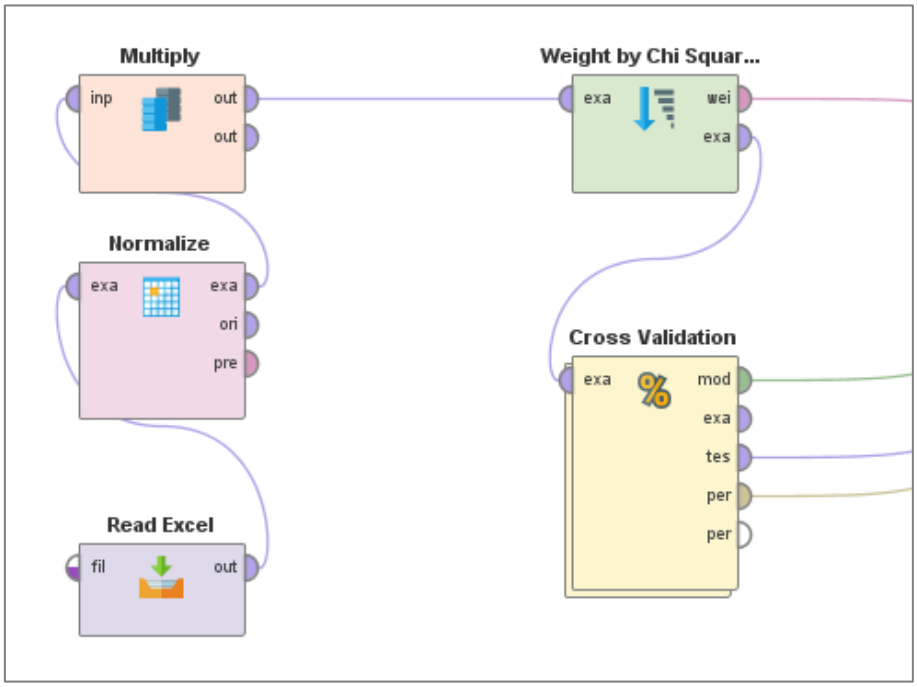


Figura 17. Modelo generado para algoritmo ID3 en tabla Docentes.

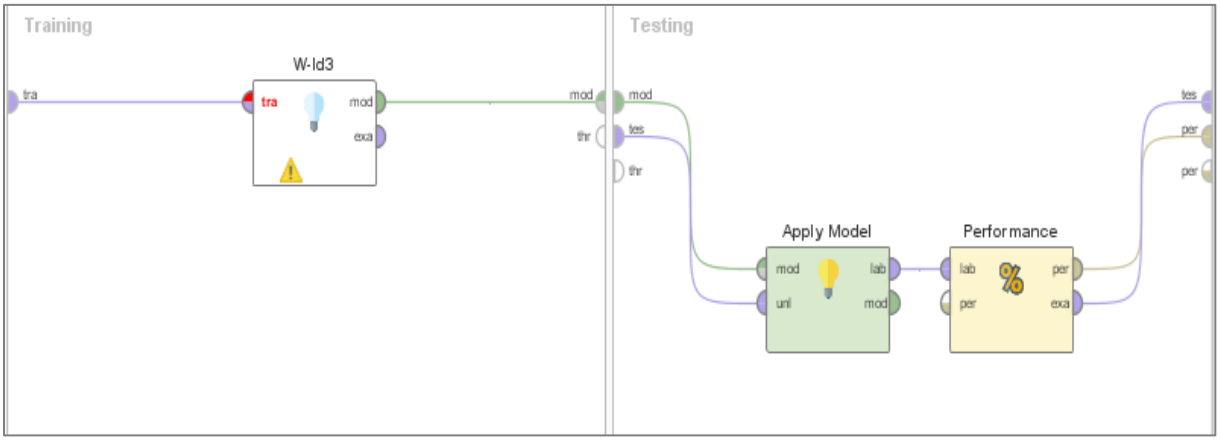


Figura 18. Validación cruzada d el algoritmo ID3 en tabla Docentes.

Algoritmo J48. A continuación, en la Figura 17 se muestra el modelo aplicado para el algoritmo J48 en la tabla Docente.

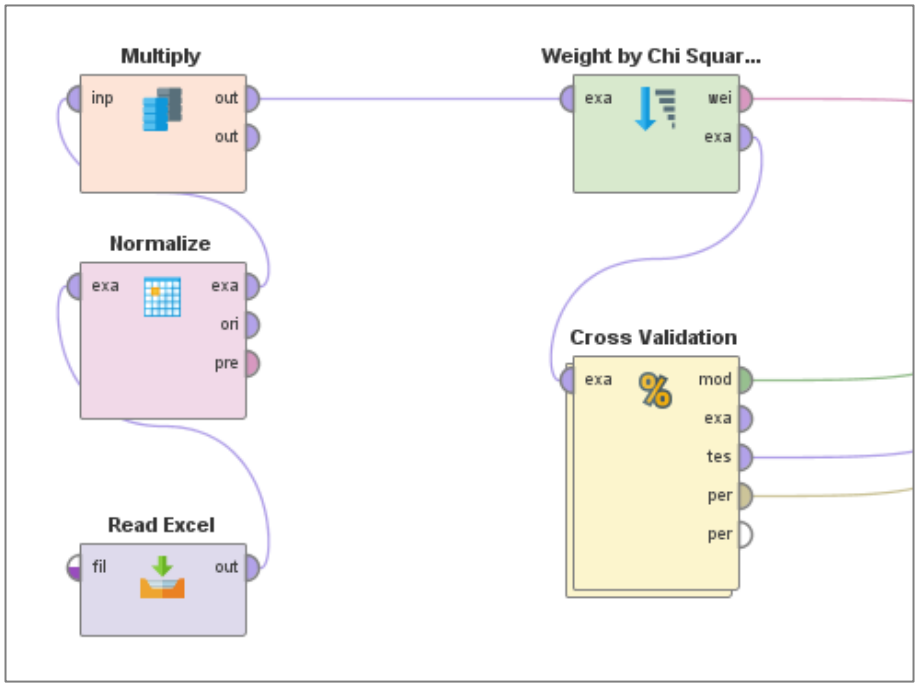


Figura 19. Modelo generado para el algoritmo J48 en tabla Docentes.

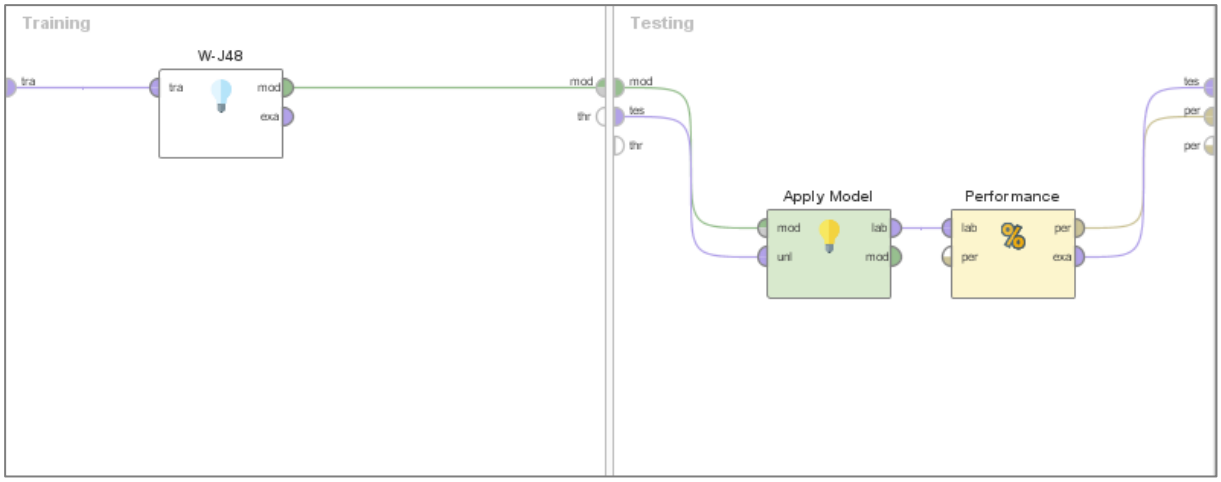


Figura 20. Validación cruzada del algoritmo ID3 en tabla Docentes.

Modelos de tabla Investigación

Algoritmo ID3. A continuación, en la Figura 19 se muestra el modelo aplicado para el algoritmo ID3 en la tabla Investigación.

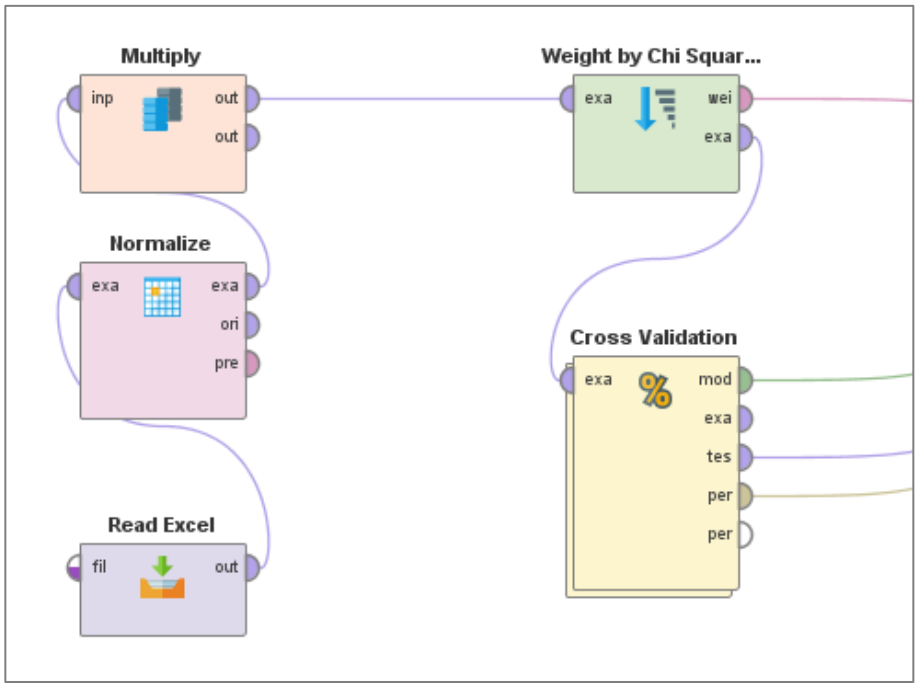


Figura 21. Modelo generado para el algoritmo ID3 en la tabla Investigación.

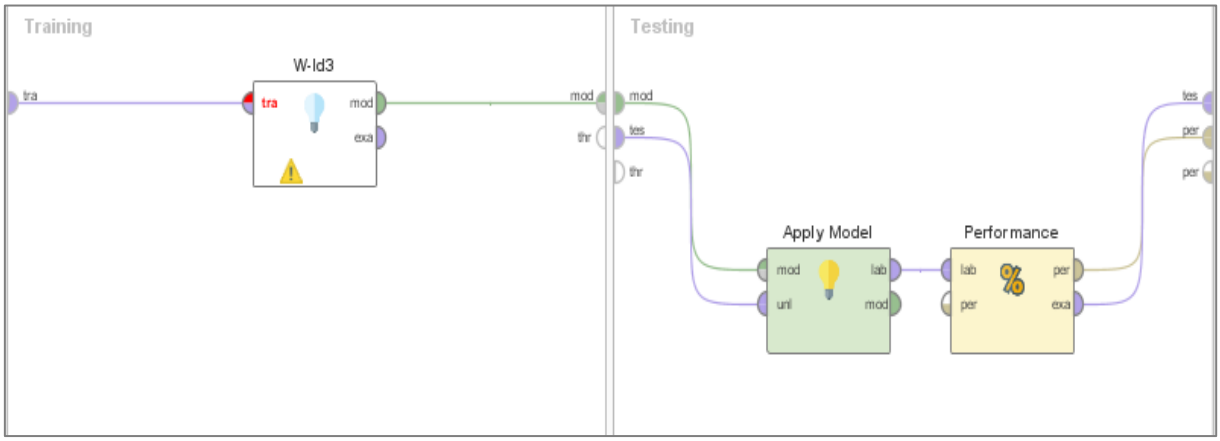


Figura 22. Validación cruzada del algoritmo ID3 en la tabla Investigación.

Algoritmo J48. A continuación, en la Figura 21 se muestra el modelo aplicado para el algoritmo J48 en la tabla Investigación.

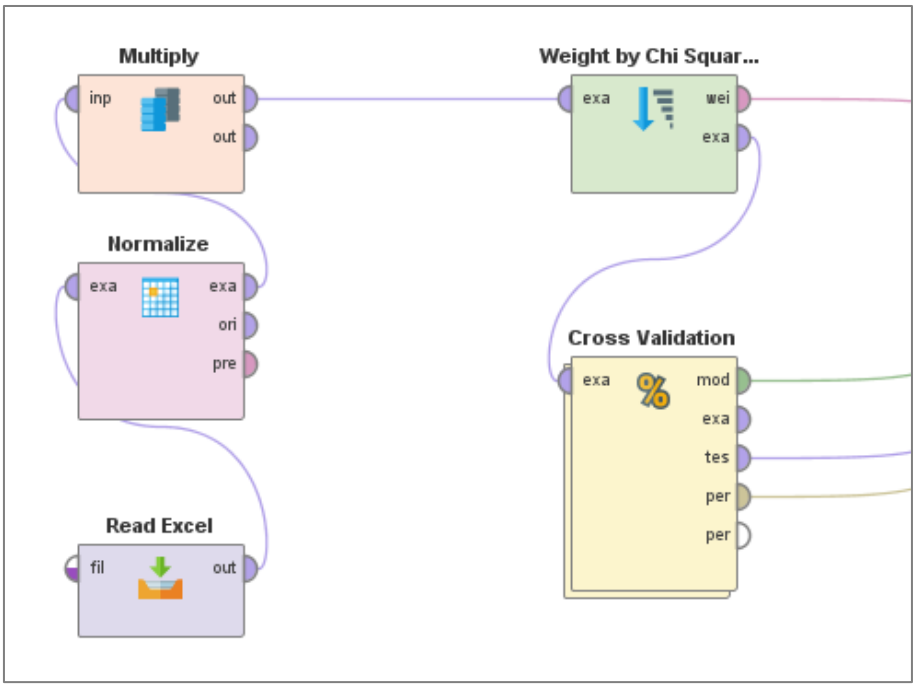


Figura 23. Modelo generado para el algoritmo J48 en la tabla Investigación.

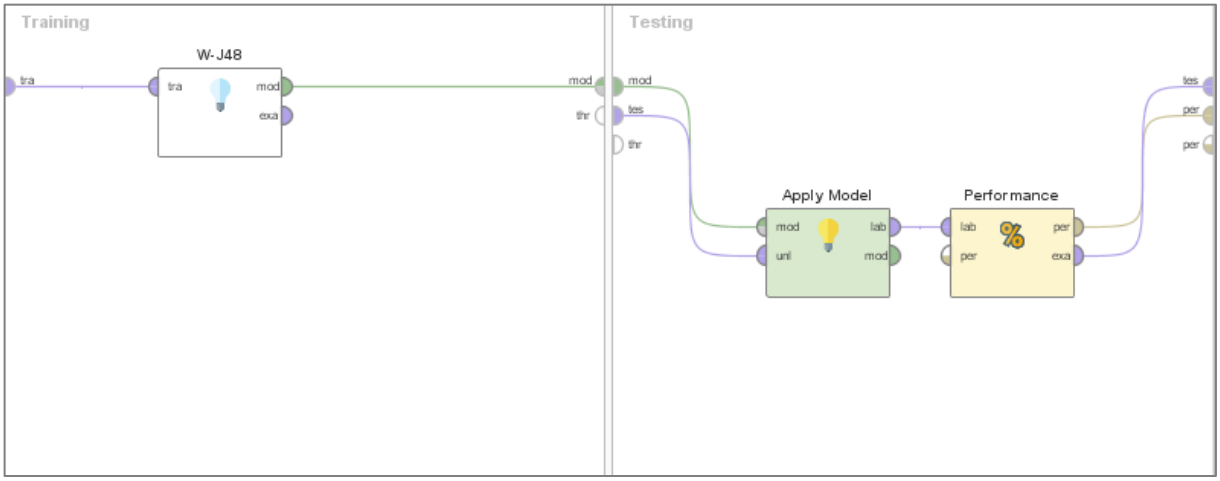


Figura 24. Validación cruzada del algoritmo J48 en la tabla Investigación.