UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Presentado por
Br. NOELIA HERMELINDA LUNA REYES

IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES

Asesor:
MSc. HERWIN ALAYN HUILLCEN BACA

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

ANDAHUAYLAS – APURÍMAC – PERÚ
2020

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Presentado por:
Br. NOELIA HERMELINDA LUNA REYES

Asesor:
MSc. HERWIN ALAYN HUILLCEN BACA

ANDAHUAYLAS – APURÍMAC – PERÚ
2020

APROBACIÓN DEL ASESOR

Quién suscribe:
MSc. Herwin Alayn Huillcen Baca, por la presente:

CERTIFICA,
Que, el Bachiller en Ingeniería de Sistemas, NOELIA HERMELINDA LUNA REYES ha culminado satisfactoriamente el informe final de tesis intitulado: "IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DE RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES" para optar el Título Profesional de Ingeniero de Sistemas

San Jerónimo, 28 de julio de 2020

MSc. Herwin Alayn Huilcen Baca
Asesor

Br. Noelia Hermelinda Luna Reyes
Tesista

ACTA DE SUSTENTACIÓN DE TESIS

En la Universidad Nacional José María Arguedas ubicado en el distrito de San Jerónimo de la Provincia de Andahuaylas, siendo las 11:00 horas del día 05 de agosto del año 2020, se reunieron los docentes: Dr. Julio César Huanca Marín, MSc. Iván Soria Solís, Mg. Humberto Silvera Reynaga en condición de integrantes del Jurado Evaluador del Informe Final de Tesis intitulado: **"IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES"**, cuyo autor es la Bachiller en Ingeniería de Sistemas **NOELIA HERMELINDA LUNA REYES**, el Asesor MSc. Herwin Alayn Huillcen Baca, con el propósito de proceder a la sustentación y defensa de dicha tesis.

Luego de la sustentación y defensa de la tesis, el Jurado Evaluador ACORDÓ: **APROBAR** por **UNANIMIDAD** a la Bachiller en Ingeniería de Sistemas **NOELIA HERMELINDA LUNA REYES**, obteniendo la siguiente calificación y mención:

Nota escala vigesimal		Mención
Números	Letras	
15	Quince	Bueno

En señal de conformidad, se procedió a la firma de la presente acta.

.....
Dr. Julio César Huanca Marín
Presidente del Jurado Evaluador

.....
MSc. Iván Soria Solís
Primer Miembro del Jurado Evaluador

.....
Mg. Humberto Silvera Reynaga
Segundo Miembro del Jurado Evaluador

APROBACIÓN DEL JURADO DICTAMINADOR

LA TESIS: **IMPLEMENTACIÓN DE UN SISTEMA DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS PARA LA ADECUADA TOMA DE DECISIONES**; para optar el Título Profesional de INGENIERO DE SISTEMAS, ha sido evaluado por el Jurado Dictaminador conformado por:

PRESIDENTE: Dr. JULIO CÉSAR HUANCA MARÍN
PRIMER MIEMBRO: MSc. IVÁN SORIA SOLÍS
SEGUNDO MIEMBRO: Mg. HUMBERTO SILVERA REYNAGA

Habiendo sido aprobado por **UNANIMIDAD**, en la ciudad de Andahuaylas el día 05 del mes de agosto 2020.

Andahuaylas, 05 de agosto de 2020.

DR. JULIO CÉSAR HUANCA MARÍN
PRESIDENTE DEL JURADO DICTAMINADOR

MSc. IVÁN SORIA SOLÍS
PRIMER MIEMBRO DEL JURADO DICTAMINADOR

Mg. HUMBERTO SILVERA REYNAGA
SEGUNDO MIEMBRO DEL JURADO DICTAMINADOR

DEDICATORIAS

Dedico este Trabajo de investigación a Dios por guiarme en mi camino, darme fuerzas para seguir adelante y no desmayar en los problemas que se me Presentan.

A mis padres por haberme dado la vida, por su gran apoyo, consejos, comprensión y amor.

A mis hermanos(as) Walter, Noé, Alex, Olga, Haydeé, María, Reyna que siempre me apoyaron incondicionalmente en la realización de mi trabajo de investigación, para cumplir con mis objetivos.

AGRADECIMIENTO

Deseo expresar mis sinceros agradecimientos a las personas que hicieron posible la realización de este trabajo de investigación:

A la Universidad Nacional José María Arguedas y la Escuela Profesional de Ingeniería de Sistemas por permitirme alcanzar este importante logro en mi formación profesional.

Un agradecimiento especial al MSc. Herwin Alayn Huillcen Baca, por el asesoramiento de tesis, por otorgarme la motivación necesaria de no rendirme, brindarme su apoyo, la orientación y consejo para la realización de esta tesis.

A mis jurados al Dr. Julio César Huanca Marín, MSc. Iván Soria Solís, Mg. Humberto Silvera Reynaga por su interés y sugerencias para la realización de este trabajo de investigación.

Gracias a todas las personas que me ayudaron directa e indirectamente en la realización de este proyecto les agradezco y siempre los tendré en mi corazón.

“Nadie puede llegar a la cima armado solo de su talento. Dios da el talento; el trabajo transforma el talento en genio”. Anna Pavlova.

ÍNDICE

Pág.

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE GRÁFICOS

Pág.

ÍNDICE DE IMÁGENES

Pág.

RESUMEN

EL rendimiento académico es un problema que se presenta en muchas instituciones educativas de todas partes del mundo, especialmente en niveles superiores, donde hay una cantidad de estudiantes que no aprueban sus materias o que abandonan sus estudios.

Este Trabajo de investigación ayudó a evaluar el problema de los estudiantes de Ingeniería de Sistemas, el sistema de predicción del rendimiento académico predice las notas de cada estudiante, lo cual permitirá tomar una mejor decisión para el siguiente semestre.

La metodología que se utilizó fue KDD (Knowledge Discovery in Databases), que permitió determinar la influencia del corpus de conocimiento (database) en la predicción de rendimiento académico de los estudiantes de Ingeniería de Sistemas.

Las etapas desarrolladas fueron: 1. Selección de datos, 2. Procesamiento, 3. Transformación, 4. Data Mining, 5. Interpretación y evaluación. En la primera etapa, se recolectó los datos socioeconómicos y el promedio ponderado de los estudiantes de Ingeniería de Sistemas. En la segunda etapa, se realizó la limpieza de los datos. En la tercera etapa, se convirtió los datos en el formato ARFF. En la cuarta etapa, se hizo uso de algoritmos de clasificación utilizando la herramienta Weka, en la quinta etapa, se evaluó e interpreto los resultados obtenidos por los algoritmos SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree.

De todos los algoritmos mencionados el Error Absoluto Medio más óptimo fue 0.85 adquirido por el algoritmo KStar con una data de 1674 datos de los Estudiantes de Ingeniería de Sistemas. Utilizando el modelo de clasificación del algoritmo KStar se implementó el sistema de predicción del rendimiento académico.

Se comprobó el error de la predicción del rendimiento académico de los estudiantes de Ingeniería de Sistemas en el semestre 2018-II es 0,11.

Palabras Claves: *Rendimiento académico, Técnicas de minería de datos, Weka, modelo de clasificación, KDD, base de datos, sistema de predicción.*

ABSTRACT

Academic performance is a problem that occurs in many educational institutions around the world, especially at higher levels, where there are a number of students who do not pass their subjects or who drop out.

This research work helped to evaluate the problem of the Systems Engineering students, the academic performance prediction system predicts the grades of each student, which will allow making a better decision for the following semester.

The methodology used was KDD (Knowledge Discovery in Databases), which made it possible to determine the influence of the corpus of knowledge (database) on the prediction of academic performance of Systems Engineering students.

The stages developed were: 1. Data selection, 2. Processing, 3. Transformation, 4. Data Mining, 5. Interpretation and evaluation. In the first stage, the socioeconomic data and the weighted average of the Systems Engineering students were collected. In the second stage, the data was cleaned. In the third stage, the data was converted into the ARFF format. In the fourth stage, the classification algorithms were used using the Weka tool, in the fifth stage, the results obtained by the algorithms SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, were evaluated and interpreted. RandomTree, REPTree.

Of all the aforementioned algorithms, the most optimal Mean Absolute Error was 0.85, acquired by the KStar algorithm with a data of 1674 data from Systems Engineering Students. Using the KStar algorithm classification model, the academic performance prediction system was implemented.

The error of the prediction of the academic performance of the Systems Engineering students in the 2018-II semester was 0.11.

Keywords: *Academic performance, Data mining techniques, Weka, classification model, KDD, database, prediction system.*

CHUMASQA

Rendimiento académico kan uk llaki iwka rikurin achkakuna yachay wasikunapi tukuykunapak piti pachapak, asuwanqa hatun yachachik wasikunapi, maypi, mayman tiyan uk achka yackakukpak iwka mana atipanchu uki kaytaku rini iwka jinchunkichi uk rikurichinakuna.

Kay llankay yanaparun kamaykuy llaki yachakukpak Ingeniería de Sistemas nisqanpata,llika pasananta musyaspa rendimiento académico killkaskapak sapanpak yachakuk, maykanyaskara atinqa atipayta allinta uknin parti wata qamukpak.

Taripanapaqmi ruwakurqa uk KDD (Knowledge Discovery in Databases) nisqanwan, iwka tukun rimanakuy chayachi corpus de conocimiento (database) nisqanwan predicción de rendimiento académico yachakukpak Ingeniería de Sistemas nisqanpi.

Kay etapaspi mayturay casqa: 1. Chanikunata akllan, 2. ruwanakuna, 3. Tukuykuna, 4. Maykayuq Mining nisqan, 5. Yachachiykuna chaymanta kamaykuykuna. puntay etapapi, huñukurqam wakachikukuna qullqikunata chaymantataq promedio chakchachiska yackakukpak Ingeniería de Sistemas nisqanpi, chini etapapi, kan ruwakurun llunpa wakachikukunapak. Kimsaniki etapapi, kan tarikun wakachikukuna sinaku ARFFpi. Uku etapapi, kan ruraspa mawkanaku algoritmospa rakispa mawkasapaktilla Weka, pichqa etapapi, kan kamaykuy rin asirtachina rayku qispiqninkunañataq chaninchasqa algoritmos SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree.

Tukuykunapak algoritmos nisqanwan pantay Absoluto chawpi allin karun 0.85 kuwan tarisqa rayku algoritmo KStar wan uk wakachiku 1674 wakachikunapak yachakukpak Ingeniería de Sistemas nisqanpi. Wawkasha wanki rakiywan algoritmo KStar kanpak mawkasha awari chay pasananta musyaspa rendimiento académico nisqanpi.

Kay chiqakchaspa pantay pasananta musyaspa rendimiento académico nisqanpi yachakukpak Ingeniería de Sistemas kay parti wata 2018-II kan 0,11.

Chanin simikuna: *Rendimiento académico, alliruray minería wakachikukunapak, Weka, wanki akllay, KDD, sinchiku wakachikukunapak, awari musyaspa.*

INTRODUCCIÓN

El Rendimiento Académico es un concepto que se utiliza de manera excluyente en el ámbito educativo para referirse a la evaluación que, en las diversas instituciones educativas de los niveles primarios, secundarios, universitarios, es llevada a cabo por los profesionales idóneos para justamente evaluar el conocimiento aprendido por parte de los estudiantes. Se considerará que un estudiante tendrá un buen rendimiento cuando tras las evaluaciones a las que es sometido a lo largo de la cursada en gestión de notas son buenas y satisfactoria. Por el contrario, estaremos hablando de un bajo rendimiento académico de un estudiante cuando las calificaciones que obtiene tras los exámenes no alcanza el nivel mínimo de aprobación (Ucha, 2007). Factores que influyen en el Rendimiento Académico pueden ser endógenos o exógenos. Endógeno es relacionado directamente con aspecto Psicológico del estudiante dentro de esto contemplan sexo, edad, la frecuencia de estudio y hábitos como leer prensa, ver noticieros y trayectoria de la vida académica. Exógenos son los factores que influyen desde el exterior, donde el estudiante contempla aspectos como la comunidad se relaciona con el entorno inmediato. Familia el nivel de ingresos, el nivel educativo de los padres, el clima de afecto y seguridad, la infractora física de hogar, los recursos disponibles para el aprendizaje, la relación de la familia con la escuela (Salinas et al., 2010).

En las últimas épocas con regularidad se observa el bajo rendimiento académico de los estudiantes universitarios. Pues una serie de factores intervienen en el rendimiento académico como endógeno o exógeno que impide aprobar los cursos satisfactoriamente. Combinando el corpus de conocimiento con las técnicas de minería de datos, se obtuvo varios modelos de clasificación. Con el modelo de clasificación que dio mejores resultados, se implementó un Sistema de Predicción de Rendimiento Académico, que predice la Promedio Ponderado del estudiante para tomar mejores decisiones para el siguiente Semestre.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del problema

El rendimiento académico es la suma de diferentes factores que interactúan en las personas que aprende y se define con un valor atribuido al aprendizaje de los estudiantes. Se mide mediante las calificaciones obtenidas en las tareas académicas, con una valoración cuantitativa, cuyos resultados muestran las materias y el grado del fracaso o éxito académico (Hernández et al., 2013). El rendimiento académico es una medida de las capacidades de estudiante, que lo expresa lo que este ha aprendido a lo largo del proceso formativo. Existen distintos factores que inciden en el rendimiento académico. Desde la dificultad propia de algunas asignaturas, hasta la gran cantidad de exámenes que pueden coincidir en una fecha, pasado por la amplia extensión de ciertos programas educativos, son muchos los motivos que pueden llevar a un estudiante a mostrar bajo rendimiento y hace referencia a la evaluación del conocimiento adquirido en el ámbito universitario, un estudiante con buen rendimiento académico es aquel que obtiene calificaciones positivas en los exámenes que debe rendir a lo largo de una cursada (Perez y Gardey, 2008)

Ahora la Universidad Nacional José María Arguedas (UNAJMA), es una universidad pública ubicada en la ciudad de Andahuaylas, Perú. Fue creada por la Ley N° 28372, en Andahuaylas el 01 de noviembre de 2004 es la primera universidad de la provincia que brinda a la juventud andahuaylina, la oportunidad de prepararse profesionalmente. Es una de las quince universidades de reciente creación que se encuentran bajo la supervisión del Consejo Nacional para el Funcionamiento de Universidades (CONAFU), Se creó gracias al apoyo y esfuerzo del pueblo andahuaylina, con el permanente empeño de quienes integramos esta casa superior de estudios: profesores, estudiantes, trabajadores y egresados. Actualmente cuenta con 2 facultades: Facultad de Ingeniería y Facultad de Ciencias de la Empresa, así mismo cuenta con tres carreras profesionales: Administración de Empresas, Ingeniería de Agroindustrial, Ingeniería de sistemas (universia, 2011).

Mientras tanto la Universidad Nacional José María Arguedas cuenta con más de 1200 estudiantes (SUNEDU, 2017). En esta tesis se implementó un sistema de predicción de rendimiento académico, que permita predecir su rendimiento académico de los estudiantes mediante la información de base de datos históricos académicas, nivel socioeconómico.

1.2. Formulación de problema

1.2.1. Problema general

- ¿Cuál es la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de Datos?

1.2.2. Problemas específicos

- ¿Cuál es la eficiencia en la elección de algoritmo más óptimo para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos?

- ¿Cuál es la eficiencia en el resultado del error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos?

1.3. Justificación

La Universidad Nacional José María Arguedas está en funcionamiento más de 12 años y como se observa el rendimiento académico de los estudiantes es muy bajo a la comparación de los estudiantes de administración de empresas por esa razón al culminar el semestre académico los estudiantes salen desaprobados en mayoría de los cursos por eso existen muchos fracasos y deserciones. Tanto los administrativos y docentes no

pueden hacer mucho al respecto. Puesto que generalmente no se le da el seguimiento adecuado a cada estudiante y tampoco se hace un diagnóstico al tiempo.

Por lo cual se plantea en este proyecto de investigación implementar un sistema utilizando información histórica académicas, demográficos y nivel socioeconómico para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional José María Arguedas utilizando técnicas Minería

Para poder hacer posible la clasificación de los datos se generó varios modelos de clasificación, combinando con las técnicas de minería de datos de los cuales se escogió el menor Error Absoluto medio. Con el modelo más óptimos y se implementó el sistema de predicción de rendimiento académico que servirá como herramienta para predecir las notas de los estudiantes.

1.4. Objetivos

1.4.1. Objetivos generales

- Evaluar la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.

1.4.2. Objetivos específicos

- Determinar la eficiencia en la elección del algoritmo más óptimo y que genere el mejor modelo para predecir el rendimiento de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.

- Medir el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.

CAPÍTULO II: ANTECEDENTES

2.1. Antecedentes a nivel nacional

Menacho, (2017) hicieron un estudio de la predicción del rendimiento académico aplicando técnicas de minería de datos, con los estudiantes de la Universidad Nacional Agraria La Molina, usando los datos académicos de los estudiantes matriculados en el curso de estadista general de la UNALM en los semestres 2013 II Y 2014 I en el curso de Estadística General, con la finalidad de predecir la clasificación final (aprobado o desaprobado), de los futuros estudiantes matriculados en el curso.

Para la demostración del estudio utilizó varias técnicas de minería de datos aplicadas a la tarea de la predicción como: las TMD de regresión logística, árboles de decisión, redes bayesianas y redes neuronales. Asimismo, utilizo técnicas para evaluar clasificadores lo que es la matriz de confusión, área bajo la curva ROC y el coeficiente Kappa.

Para la aplicación de las TMD propuestas, usaron el programa WEKA (Waikato Environment for Knowledge Analysis). Para evaluar los modelos predictivos sobre el resultado (Desaprobado, Aprobado).

Llegaron a un resultado de que Las TMD demuestran ser herramientas eficaces para obtener modelos que permitan predecir el resultado de los estudiantes matriculados en el curso de Estadística General. La técnica de la red Naive de Bayes resultó con una la mayor precisión, al obtener un 71,0% de correcta clasificación.

2.2. Antecedentes a nivel internacional

Los investigadores (Visitación et al., 2000) realizaron un estudio sobre La predicción del rendimiento académico regresión lineal versus regresión logística en la Universidad Complutense de Madrid con el objetivo de evaluar la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito o fracaso académico.

La muestra constituyeron con 175 estudiantes (140 mujeres y 35 hombres) de primer curso de Psicología, tomándose los datos fueron tomados durante el curso académico 1997/1998 en la asignatura de “Métodos y Diseños de Investigación en Psicología I”, del área de Metodología, de la UCM.

Las técnicas de análisis de datos empleadas fueron la regresión lineal múltiple y la regresión logística. Los Resultados que obtuvieron en:

a). Regresión lineal múltiple:

Obtuvieron un valor de Regresión Múltiple igual a 0,41906.

b) Regresión logística

El modelo permite hacer una estimación correcta (c2 4= 28.424, p<0.01) del 68,57% de los casos y llegaron a una conclusión que el procedimiento de regresión múltiple no nos permite hacer un buen pronóstico del rendimiento académico, mientras que la regresión logística sí parece ser un instrumento idóneo para hacer una buena predicción del éxito o fracaso académico.

Según (Cedano, 2015). En su tesis modelo de minería de datos para la identificación de patrones que influyen en el aprovechamiento académico, en la presente investigación realiza análisis de la aplicación de técnicas de minería de datos para identificar patrones de comportamiento con el fin de predecir fracaso escolar y el abandono.

Las principales fuentes de datos con la que trabajo fueron, los registros históricos de las bases de datos académica, socioeconómicos, psicopedagógica, registrados en la preparatoria Juan Pablo II, correspondientes al periodo 2008-2013. Con registro de 1719 alumnos.

Para preparación de datos utilizo ETIL y Se aplicó técnicas de minería de datos utilizando la metodología CRIP-DM que estructura el proceso de minería de datos en 6 fases: que interactúan entre ellas de forma reiterativa. REITERATIVA Se aplicaron los modelos de redes neuronales, árboles de decisión y Cluster K-medianas para analizar el comportamiento de los estudiantes.

Porcel et al. (2010). En su investigación desarrollaron un artículo científico sobre la predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE), en función de su caracterización socioeducativa en Argentina. La población que utilizaron para analizar estuvo integrada por los estudiantes ingresantes a la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste (FACENA-UNNE), durante en los años 2004 y 2005. La información sobre las características socioeducativas y el desempeño académico de los estudiantes fue obtenida del sistema informático de gestión de estudiantes de la unidad académica. El rendimiento fue medido por la aprobación de los exámenes parciales o finales de la primera materia de Matemática que los alumnos cursan. Para el modelamiento emplearon la técnica de regresión logística. Lo cual se ajustó un modelo de regresión logística binaria, el cual clasificó adecuadamente el 75% de los datos.

CAPÍTULO III: MARCO TEÓRICO

3.1. Bases teórico científicas

3.1.1. Base teórica (variable independiente)

X: Sistema de predicción del rendimiento académico

3.1.2. Base teórica (variable dependiente)

Y: Rendimiento académico de los estudiantes.

3.2. Definición conceptual

3.2.1. Minería de datos

Minería de datos es como el proceso de seleccionar, explorar, modificar, modelizar y valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. El proceso de Data Mining es aplicable a lo largo de una amplia variedad de industrias y proporciona distintas metodologías de análisis según el tipo de problema que queremos analizar (Perez y Santín, 2007).

Minería de datos es un proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. La tarea fundamental de minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debe ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a toma de decisiones más seguras que reporten, por tanto, algún beneficio a la organización (Witten y Frank, 2000).

Minería de datos son los datos que pasan ser productos (resultado histórico de los sistemas de información) a ser una “materia prima” que hay que explotar para obtener un verdadero” producto elaborado”, el conocimiento; un conocimiento que sea de ser especialmente valioso para la

ayuda en toma de decisiones sobre el ámbito en el que sean recopilado o extraído los datos es bien cierto que la estadística es la primera ciencia que considera los datos como su materia prima, pero las nuevas necesidades y, en particular las nuevas características de los datos(en volumen y tipología) hacen que las disciplinas sean numerosas y heterogéneas (Orallo et al.,2004)..

3.2.2. Técnicas de minería de datos

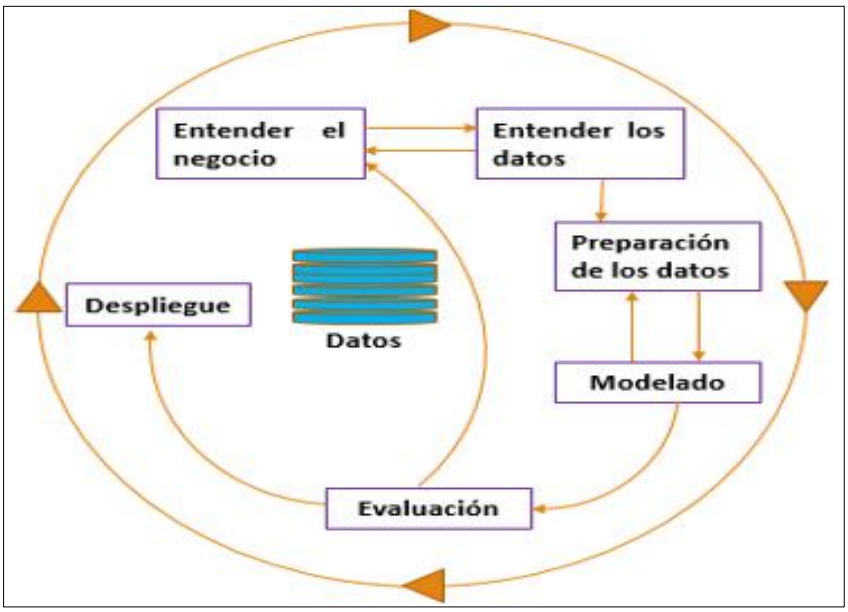


Figura 1: Metodología de minería de datos (CRISP-DM)

Las técnicas de la minería de datos provienen de Inteligencia Artificial y de la estadística, que son algoritmos que se aplica a un conjunto de datos con el fin de obtener resultados requeridos. Las técnicas más representativas que se utilizan en la minería de datos son:

- **Redes neuronales.-** Es un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales que se trata de un sistema de

interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

- ❖ Perceptrón
- ❖ perceptrón multicapa
- ❖ mapas autoorganizados o redes de Kohonen

- **Regresión lineal.-** Es la más utilizada para formar relaciones entre datos además es rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan.

- **Árboles de decisión.**- Es un modelo de predicción utilizando el ámbito de la inteligencia artificial, dada a una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirve para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema, los algoritmos que utilizan son:
  - ❖ Algoritmo ID3.
  - ❖ Algoritmo C4.5.

- **Modelos Estadísticos.-** Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modificar la variable de respuesta.

- **Agrupamiento o Clustering.-** Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente la distancia; se tratará de disponer los vectores de entrada de forma que estén cercanos aquellos que tengan características comunes como:
  - ❖ Algoritmo K-means.
  - ❖ Algoritmo K-medoids.

- **Reglas de asociación.**- se utilizan para descubrir hechos que ocurre en común dentro de un determinado conjunto de todos. Algoritmos supervisados (o predictivos): predicen a un dato o un conjunto de datos desconocidos a priori, a partir de otros conocidos. Algoritmos no supervisados (o descubrimiento del conocimiento): se descubren patrones y tendencias en los datos. Las técnicas son implementadas específicas de los algoritmos que se utilizan para llevar a cabo las operaciones de construcción de modelos. No todos los algoritmos para resolver un determinado problema de minería de datos son iguales y cada uno de ellos tendrá una serie de ventajas e inconvenientes (Perez y Santín, 2007). Técnicas de minería de datos que se puede aplicar para resolver en las diferentes operaciones como:

- **Modelos predictivos:** clasificación: En estos modelos se utiliza aprendizaje supervisado. Se suelen utilizar árboles de decisión, regresión logística, y redes neuronales. Estos modelos utilizan un conjunto de datos de entrenamiento para crear el modelo, que posteriormente se utiliza para clasificar individuos desconocidos.

- **Modelos predictivos**: predicción de valores: Para la predicción de valores se utilizan, junto a los métodos anteriores, la regresión lineal y la regresión no lineal.

- **Segmentación de base de datos: clustering no jerárquico:** Se compara cada registro de base de datos con todos los segmentos o semillas creados por la función de creación de las semillas. Se mide la distancia del registro de entrada con los segmentos ya creados y se asigna el registro de entrada al segmento correspondiente. El número de clusters se adjunta automáticamente. Método de las k-medidas.

- **Segmentación de base de datos: clustering jerárquico:** Este tipo de técnica es apropiado cuando no conocemos ni tenemos información acerca de los grupos en los que se clasifican los clusters. Se suelen utilizar algoritmos de tipo jerárquico como los aglomerativos o divisivos. Junto a ellos se utilizan redes neuronales basadas en el aprendizaje no supervisado, como por ejemplo los mapas de Kohonen.

- **Análisis de relaciones: asociaciones:** El objetivo de esta técnica es encontrar elementos que implica la presencia de otros elementos dentro de una misma transacción. El resultado de esta técnica son reglas del tipo “if X then Y”. En las reglas, X se denomina cabeza de la regla y se denomina cuerpo.

- **Análisis de relaciones: Patrones secuenciales:** trata de descubrir patrones entre transacciones en las que un conjunto de elementos distanciados un periodo de tiempo determinado.

- **Análisis de relaciones: Patrones en serie temporales:** con esta técnica se pretenden descubrir ocurrencias similares a una dada en un base de datos que almacene información que presente una serie temporal, como puede ser la evolución de los precios del mercado o datos de telemetría provenientes de algún sensor, es decir, una serie temporal en un conjunto de valores de una variable en un periodo de tiempo.

3.2.3. Árbol de decisión

Es una de las técnicas de inteligencia artificial es más útil y utilizadas para la toma de decisiones adecuadas teniendo varias alternativas posibles de acción, que es una buena herramienta que ayuda para elegir entre varios cursos de acción, semejante a un árbol. Está formado por múltiples nodos

cuadraos, que representan los puntos de decisión y de los cuales surgen ramas que representa las distintas alternativas (Marquez, 2012).

3.2.3.1. Características de árboles de decisión

- Es una forma gráfica y analítica de representar todos los sucesos posibles a partir de una decisión asumida.
- Ayuda a tomar la decisión más acertada, desde el punto de vista probabilístico, ante todas las posibles soluciones.
- Permite visualizar todas las posibles soluciones a un problema.
- Plantea el problema desde distintas perspectivas de acción.
- Permite analizar de manera completa todas las posibles soluciones.
- Su estructura permite analizar las alternativas, los eventos, las probabilidades y los resultados.

3.2.3.2. Algoritmos ID3

Es un algoritmo matemático para construir un árbol de decisión hace uso de la teoría de información, construye un árbol de arriba debajo de forma directa sin hacer uso de backtracking también se usa de ganancia de información para seleccionar el atributo más útil en cada paso.

3.2.3.3. Algoritmos C4.5

Es un algoritmo usado para generar un árbol de decisión que es una extensión del algoritmo ID3, que pueden ser usados para clasificación y por esa razón C4.5 está casi siempre referido como un clasificador.

3.2.4. Redes bayesianas

Son una representación gráfica de dependencias para razonamiento probabilístico en sistemas expertos. Una red bayesiana es un grafo acíclico dirigido y anotado que describe la distribución de probabilidad conjunta que gobierna un conjunto de varias aleatorias, los nodos pueden representar

cualquier tipo de variable, ya sea un parámetro medible (o medido), una variable latente o una hipótesis. Existe algoritmos que realizan inferencias y aprendizaje basados en redes bayesianas. Si existe un arco que une un nodo A con otro nodo B, A es denominado un padre de B, y B es llamado un hijo de A. el conjunto de nodos padre de un nodo Xi se denota como padres (Xi). Un gráfico acíclico es una red bayesiana relativa a un conjunto de variable si la distribución conjunta de los valores del nodo puede ser escrita como el producto de las distribuciones local de cada nodo y sus padres. Si el nodo Xi no tiene padres, su distribución local de probabilidad se toma como incondicional, en otro caso es condicional. Si el valor de nodo es observable y por tanto etiquetado como observado, dicho nodo es un nodo de evidencia (Quintan, 2018) .

Las redes bayesianas son elaboradas basándose en un conocimiento experto desarrollando diversas técnicas para aprender a partir de ciertos datos que estructuran parámetros asociados al modelo, siendo posibles compartir conocimientos obtenidos de los datos del caso de estudio (Rivera, 2011). Los Algoritmos que se utilizaron son:

- SMOreg
- IBK
- Kstar
- LWL
- Bagging
- Vote
- ZeroR
- M5P
- RandomForest
- RandomTree, REPTree.

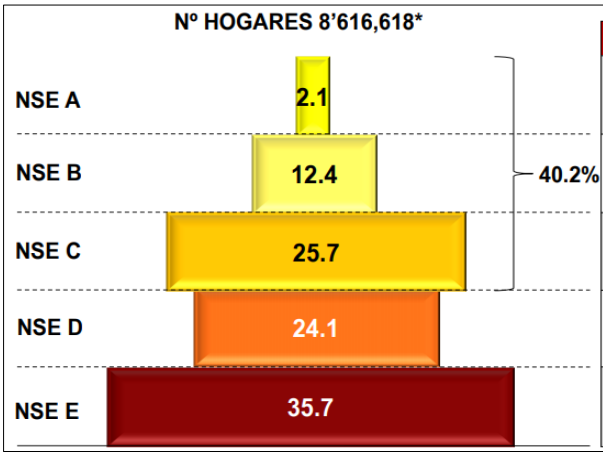
3.2.5. Nivel socioeconómico (NSE)

Es un atributo del hogar que caracteriza su inserción social y económica. Está basado en el nivel de educación, el nivel de ocupación y el patrimonio. Este nivel está dividido en varios segmentos: alto, medio y bajo.

Como ya se viene haciendo años anteriores, la distribución de niveles socioeconómicos de APEIM es calculada en base a la Encuesta Nacional de Hogares (ENAHO) que realiza el INEI para elaborar los indicadores de pobreza. Es importante seguir trabajando con esta base de datos, pues nos permite tener información de niveles socioeconómicos para todo el país y para cada uno de los departamentos que los conforman.

Debido a que la encuesta es publicada en su totalidad a mediados de año, APEIM ha decidido dar a conocer la distribución de niveles socioeconómicos también a mediados de años. Tanto la distribución, como la fórmula de NSE son utilizadas por todos los miembros de APEIM para sus estudios lo que permite la comparabilidad de sus datos. El nivel socioeconómico de una persona u hogar no se define a partir de sus ingresos sino en función a un grupo de variables definidas a partir de estudios realizados por (APEIM, 2017) .

Tabla 1: NSE 2017 - Departamento (Urbano + Rural)



NSE	Estrato	Porcentaje	
A	A1	0.3	2.1
	A2	1.8	
B	B1	4.4	12.4
	B2	8.0	
C	C1	15.6	25.7
	C2	10.1	
D	D	24.1	24.1
E	E	35.7	35.7

Fuente: Ingreso estimado en base al gasto – INEI el ingreso de las zonas rural.

Tabla 2: Ingreso socioeconómico

PROMEDIOS	Total Perú urbano y rural						
	TOTAL	NSE AB	NSE C	NSE C1	NSE C2	NSE D	NSE E
Grupo 1 : Alimentos – **gasto promedio**	S/. 940	S/. 1,414	S/. 1,194	S/. 1,235	S/. 1,129	S/. 925	S/. 576
Grupo 2 : Vestido y Calzado – **gasto promedio**	S/. 155	S/. 284	S/. 181	S/. 195	S/. 159	S/. 130	S/. 100
Grupo 3 : Alquiler de vivienda, Combustible, Electricidad y Conservación de la Vivienda – **gasto promedio**	S/. 270	S/. 623	S/. 359	S/. 385	S/. 318	S/. 221	S/. 97
Grupo 4 : Muebles, Enseres y Mantenimiento de la vivienda – **gasto promedio**	S/. 158	S/. 415	S/. 151	S/. 160	S/. 137	S/. 111	S/. 90
Grupo 5 : Cuidado, Conservación de la Salud y Servicios Médicos – **gasto promedio**	S/. 174	S/. 391	S/. 229	S/. 254	S/. 190	S/. 144	S/. 68
Grupo 6 : Transportes y Comunicaciones – **gasto promedio**	S/. 281	S/. 765	S/. 340	S/. 378	S/. 281	S/. 182	S/. 109
Grupo 7 : Esparcimiento, Diversión, Servicios Culturales y de Enseñanza – **gasto promedio**	S/. 280	S/. 813	S/. 351	S/. 393	S/. 286	S/. 181	S/. 79
Grupo 8 : Otros bienes y servicios – **gasto promedio**	S/. 163	S/. 339	S/. 203	S/. 216	S/. 183	S/. 140	S/. 79
PROMEDIO GENERAL DE GASTO FAMILIAR MENSUAL	**S/. 2,422**	**S/. 5,044**	**S/. 3,007**	**S/. 3,216**	**S/. 2,683**	**S/. 2,034**	**S/. 1,198**
PROMEDIO GENERAL DE INGRESO FAMILIAR MENSUAL*	**S/. 3,282**	**S/. 7,826**	**S/. 4,059**	**S/. 4,436**	**S/. 3,472**	**S/. 2,594**	**S/. 1,341**

FUENTE: Ingreso estimado en base al gasto – INEI El ingreso de las zonas rural APEIM 2017: Data ENAHO 2016**.**

Tabla 3: Ingresos y gastos según NSE 2017 - Perú (Urbano + Rural)

DEPARTAMENTO	HOGARES - NIVEL SOCIOECONÓMICO - URBANO+RURAL (%)						
	TOTAL	AB	C	D	E	MUESTRA	ERROR (%)*
Amazonas	100%	3.4	10.9	18.2	67.5	1,250	2.8
Ancash	100%	7.2	23.3	24.2	45.3	1,464	2.6
Apurimac	100%	1.9	6.5	16.0	75.6	939	3.2
Arequipa	100%	15.9	32.1	31.8	20.2	1,755	2.3
Ayacucho	100%	2.9	9.8	14.9	72.4	1,171	2.9
Cajamarca	100%	4.0	8.8	15.7	71.5	1,729	2.4
Cusco	100%	6.5	14.3	21.1	58.1	1,248	2.8
Huancavelica	100%	1.3	5.3	11.6	81.8	1,045	3.0
Huanuco	100%	5.2	11.6	15.3	67.9	1,284	2.7
Ica	100%	13.4	36.2	34.3	16.1	1,678	2.4
Junin	100%	7.7	20.0	25.2	47.1	1,551	2.7
La Libertad	100%	10.1	18.9	25.4	45.6	1,568	2.8

FUENTE: Ingreso estimado en base al gasto – INEI El ingreso de las zonas rural APEIM 2017: Data ENAHO 2016**.**

3.2.6. Rendimiento académico

Según (Pérez y Gardey, 2008) menciona que el rendimiento académico es una medida de las capacidades del alumno, que expresa lo que éste ha aprendido a lo largo del proceso formativo. También supone la capacidad del alumno para responder a los estímulos educativos. En este sentido, el rendimiento académico está vinculado a la aptitud.

También Requena (1998) afirma que el rendimiento académico es fruto del esfuerzo y la capacidad de trabajo del estudiante, de las horas de estudio, de la competencia y el entrenamiento para la concentración.

Según (Jaspe, 2010). El rendimiento académico escolar es una de las variables fundamental de la actividad docente, que actúa como halo de la calidad de un Sistema Educativo.

Chadwick (1979) define el rendimiento académico como la expresión de capacidades y de características psicológicas del estudiante desarrolladas y actualizadas a través del proceso de enseñanza-aprendizaje que le posibilita obtener un nivel de funcionamiento y logros académicos a lo largo de un período o semestre.

3.2.6.1. Factores que influyen en el rendimiento académico

Según los estudios de los investigadores (Salinas et al., 2010) realizados en el estudio sobre los factores que influyen en el rendimiento escolar, los factores o variables que inciden en el rendimiento académico, pueden ser exógenos o endógenos. Exógenos son los factores que influyen desde el exterior y endógenos relacionados directamente con aspectos personales psicológicos o somáticos del alumno.

a) ENDÓGENO

- **Estudiantes:** Dentro de esta se contemplan aspectos como la frecuencia de estudio y hábitos como el leer libros, ver noticieros, y trayectoria de la vida académica.

b) EXÓGENOS

- **Comunidad:** Se relaciona con el entorno inmediato o vecindario donde vive la familia y el involucramiento de los alumnos en las actividades tanto positivas o negativas.

- **Familia:** tiene que ver no sólo con el nivel de ingresos, sino con la composición de la familia, la ocupación y el nivel educativo de los padres, la vida familiar, el clima de afecto y seguridad, la infraestructura física del hogar, los recursos disponibles para el aprendizaje, el uso del tiempo, las prácticas de crianza, la relación de la familia con la escuela, etc.

- **Escuela:** Opera a nivel del sistema escolar en su conjunto, a nivel de cada institución (y/o redes de instituciones) y a nivel de aula. Tiene que ver no únicamente con la enseñanza, sino con todas las dimensiones del quehacer y la cultura escolares, incluyendo la infraestructura y los materiales de enseñanza, enseñanza, el uso del espacio y del tiempo, la organización, las rutinas y las normas, la relación entre directivos y docentes y entre estos, los alumnos, los padres de familia y la comunidad.

3.2.7. Weka

Según (Fallas, 2011) Weka es una herramienta de tipo software para el aprendizaje automático y minería de datos diseñado a base de Java y desarrollado en la universidad de Waikato en Nueva Zelanda en el año 1993, esta herramienta por su nombre en inglés (Waikato Environment for Knowledge Analysis) además es una herramienta de distribución de licencia GNU-GLP o software libre.

Weka contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, también tiene herramientas para la visualización

de estos datos, además provee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición.

3.2.7.1. Historia de Weka

Inicialmente en el año 1193 la universidad de Waikato ubicada en Nueva Zelanda se dio a la tarea de desarrollar la primera implementación para analizar datos procedentes de la agricultura y con esto se dio lo que fue la primera versión de WEKA, esta fue desarrollada en TCL/TK y lenguaje C, cuatro años más tarde en 1997 se decide escribir todo el código original en java, además se le incluyeron implementaciones de algoritmos de modelado.

En el año 2005 esta herramienta muy flexible y fácil de utilizar, recibe el galardón “Data Mining and Knowledge Discovery Service”, por parte de la ACM (Asociación for Computing Machinery) que es la Sociedad Científica Para el Desarrollo de la Computación educacional.

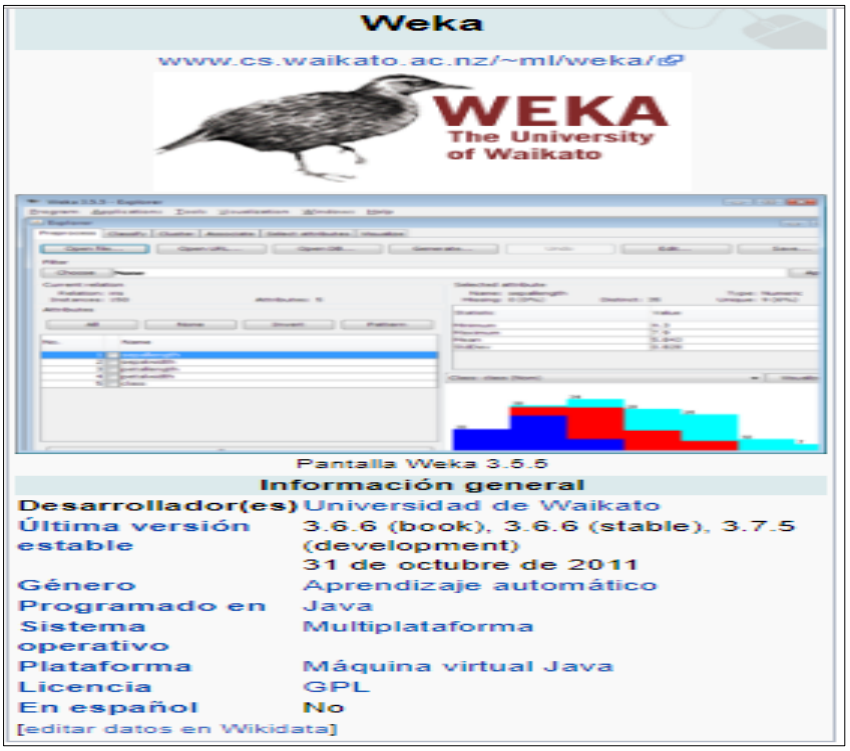


Imagen 1: Imagen de Weka

3.2.7.2. Características

Esta es una herramienta muy versátil que soporta muchas tareas estándar de la minería de datos en especial tareas de procesamiento de datos, regresión, clasificación, clusterin entre otras, así mismo permite la visualización y la selección de los datos.

Todas las técnicas en WEKA están basadas en la sunción de datos que están disponibles en un fichero plano o una relación, en donde cada registro de datos esta descrito por un número fijo de atributos nominales o numéricos.

Permite el acceso a otras instancias de bases de datos por medio de SQL, gracias al JDBC, además puede procesar un resultado generado a base de una consulta hecha a una base de datos. También (María y García, 2017) menciona que Weka se denomina a sí mismo un conjunto de Librerías para tareas de minería de datos.

Las librerías pueden ser llamadas desde la interficie de Weka o desde tus propias clases Java. Weka contiene herramientas para diferentes tareas básicas:

a). Preprocess: Multitud de herramientas para el preprocesamiento de los datos (como por ejemplo discretización de variables).

b). Classify: Algoritmos de clasificación, distribuidos por paquetes, como por ejemplo ID3 o C4.5

c). Clúster: Diferentes algoritmos de segmentación como el simple k-medias.

d). Associate: Algoritmos para encontrar relaciones de asociación entre variables (Apriori entre otros).

e). Select atributtes: Aquí, una vez cargados los datos, Weka es capaz de buscar por nosotros las mejores variables del modelo.

f). Visualize: Herramienta de visualización de datos en los ejes cartesianos, con muchas posibilidades.

3.2.7.3. Ventajas

- Es un software de fácil acceso por los usuarios ya que está disponible libremente bajo licencia pública GNU.
- Contiene una gran gama de técnicas para modelado y procesamiento de datos.
- Posee una interfaz de usuario muy sencilla que permite su uso hasta por usuarios sin experiencia
- Funciona en la mayoría de las plataformas actuales ya que es una implementación en java.

3.2.7.4. Desventajas

- No incluye algoritmos para el modelado de secuencias.
- Al utilizar métodos de combinación de modelos, los resultados tienden a complicarse, perdiendo comprensibilidad.

3.2.7.5. Campos de aplicación

Weka posee un grupo de técnicas que se pueden aplicar con éxito a múltiples capos, tales como el Marketing, manufactura, salud, energía, finanzas, medicina, entre otros, para su aplicación se debe de tener en cuenta el tipo de tarea que se desea realizar. Con base a la tarea se puede implementar en para dar la solución necesaria en cada uno de estos campos, a continuación, se muestran una serie de tareas incluidas en Weka.

- Métodos de clasificación basados en redes neuronales.
- Manipulación de métodos numéricos sobre datos (Funciones estadísticas).
- Métodos de clasificación basados en máquinas de soporte vectorial.
- Meta clasificadores.
- Métodos de implementación de árboles de decisión.

- ❖ Implementación de clusterin.
- ❖ Métodos de estimación de probabilidades.
- ❖ Clases para implementar clasificadores bayes.

3.2.7.6. Cuadros comparativos de herramientas de minería de datos

Actualmente además de Weka existen en el mercado un gran número de herramientas para la explotación de la minería de datos, algunas con mejores características en algunos aspectos, pero casi todas son diseñadas con la misma finalidad, a continuación, se muestra un cuadro comparativo.

Tabla 4 : Cuadro comparativo – herramientas de minería de datos

Característica	Clementine	SAS Enterprise Miner	Tariykdd	Weka
Licencia libre	No	No	Si	Si
Requiere conocimientos avanzados	No	No	No	No
Acceso a SQL	Si	No	Si	Si
Multiplataforma	NO	Si	Si	Si
Requiere bases de datos especializadas	NO	---	No	NO
Métodos de máquinas de soporte vectorial	Si	Si	No	Si
Métodos bayesianos	Si	---	No	Si
Puede combinar modelos	Si	Si	No	Si (no resulta muy eficiente)
Modelos de clasificación	Si	Si	Si	Si
Implementa arboles de decisión	Si	Si	Si	Si
Modelos de regresión	Si	Si	No	Si
Clusterin y agrupamiento	Si	Si	No	Si
Interfaz amigable	Si	Si	Si	Si
Permite visualización de datos	Si	Si	Si	Si

Fuente: Elaboración propia. Utilizando Recortes de Windows 8.1

3.2.8. Metodología cascada

Según (Soloriio, 2013) el modelo cascada es conocido como modelo clásico o lineal secuencial. Este método, implica un desarrollo rígido que está dividido en una secuencia de actividades.

3.2.9. Medidas estadísticas

3.2.9.1. Coeficiente de correlación (Correlation coefficient)

Según (Alejandro, 2017) El coeficiente correlación es una medida que permite conocer el grado que se asocian lineal entre dos variables cuantitativas (x, y) lo cual cuantifica algún tipo de relación y/o dependencia entre dos variables aleatorias o valores de datos observados.

Coeficiente de correlación está dado por la siguiente fórmula.

$$r = \frac{S_{XY}}{S_X S_Y}$$

¿De dónde sacamos estos valores?

S_{XY} = Covarianza

$S_X S_Y$ = Desviacion Estandar de X multiplicada por la Desviacion Estandar de Y

Luego de haber aplicado la formula, según el resultado se puede clasificar en este rango.

Tabla 5: Rango de correlación

Rango		**Relación lineal**
+- 0,96	+-1,0	Perfecta
+- 0,85	+- 0,95	Fuerte
+- 0,70	+- 0,84	Significativa
+- 0,50	+- 0,69	Moderada
+- 0,20	+- 0,49	Débil
+- 0,10	+- 0,19	Muy Débil
+- 0,09	+- 0,0	Nula

$$r_{\text{(coeficiente de correlación)}} = \begin{cases} -1 \\ 0 \\ 1 \end{cases}$$

3.2.9.2. Error absoluto medio MAE (Mean absolute error).

El error absoluto medio es una medida de la diferencia entre dos variables continuas. Considerando dos series de datos (unos calculados y otros observaos) relativos a un mismo, fenómeno, que sirve para cuantificar la precisión de una técnica de la predicción comparando por ejemplo los valores predichos frente a los observados, el tiempo real frente al tiempo provisto, o una técnica de medición a otra técnica alternativa de medición.

El error absoluto nos indica el grado de aproximación y da un indicio de la calidad de la medida. El conocimiento de la calidad se complementa con el error relativo. También puede ser un valor positivo o negativo, según si la medida es superior al valor real o inferior y además tiene las mismas unidades que las de medida.

La fórmula de error absoluto medio es:

$$\boldsymbol{MAE} = \frac{SAE}{N} = \frac{\sum_{i=1}^{N} |\mathrm{x_i} - \hat{x}_i|}{N}$$

También el Error Relativo se puede definir como el coeficiente entre en error absoluto y el valor verdadero esto es,

$$\mathrm{E_r} = \frac{\text{valor verdadero} - \text{valor aproximado}}{valor\ verdadero}$$

Y también se define el error relativo porcentual, como sigue:

$$\in_r = \mathrm{E_{y}} \times 100\%$$

Es decir:

$$\in_r = \frac{\text{valor verdadero} - \text{valor aproximado}}{valor\ verdadero} x100$$

3.2.9.3. Error cuadrático medio (Root mean squared error).

El error cuadrático medio es una forma de evaluar la diferencia entre un estimador y el valor real de cantidad que se requiere calcular, también mide el promedio del cuadrado del “error’’, siendo el error el valor en la que el estimador difiere de la cantidad a ser estimada.

El MSE En términos generales equivale a la suma de varianza y la desviación estándar al cuadro del estimador, usado para determinar la medida en la que el modelo no se ajusta a la información, o si el quitar ciertos términos puede simplificar el modelo de manera benéficas, que proporciona una forma para elegir el mejor estimador.

La fórmula para calcular el error cuadrático medio:

$$\boldsymbol{MSE} = \frac{1}{M} \sum_{i=1}^{M} (reali - estimsdoi)^2$$

3.2.9.4. Error absoluto relativo (Relative absolute error).

a) **Error absoluto.-** Es la diferencia entre el valor de la medida y el valor tomado como exacto. Puede ser positivo o negativo, según la medida si la medida es superior al valor real o inferior. Tiene unidades, las mismas que las de medida.

b) **Error relativo.-** Es el coeficiente (la división) entre el error absoluto y el valor exacto. Si se multiplica por 100 se obtiene el tanto por ciento (%) de error. Al igual que el error absoluto puede ser positivo o negativo (según sea el error absoluto) porque puede ser por exceso o por defecto. No tiene unidades.

c) **Error porcentual.-** El error porcentual es simplemente el error relativo expresado en un (%).

Es la diferencia entre el valor real y el valor aproximado.

$e = |Valor\ real - Valor\ Aproximado|$

Donde:

- El valor real es el valor que en teoría mide la magnitud a medir
- El valor aproximado es la medida de las diferentes medidas.

3.2.9.5. Error cuadrado relativo de raíz (Root relative squared error).

Error cuadrado relativo de raíz (RMSE) mide la cantidad de error que hay entre dos conjuntos de datos. Que compara un valor predicho y un valor observado o conocido.

La fórmula de error cuadrado relativo de raíz está dada de la siguiente manera:

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Pi - Oi)^2}{n}}$$

CAPÍTULO IV: DISEÑO METODOLÓGICO

4.1. Hipótesis de investigación

4.1.1. Hipótesis general

- Es posible evaluar la eficiencia de un sistema de rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando Minería de datos.

4.1.2. Hipótesis especifica

- Es posible determinar el algoritmo más óptimo y que genere el mejor modelo predictivo para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.

- Es posible medir el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando Minería de datos, con un grado de error mínimo 15%.

4.2. Operacionalización de variables

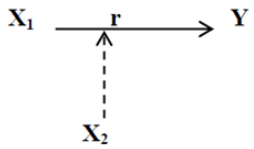
Tabla 6: Operacionalización de variables

VARIABLES	DEFINICION CONCEPTUAL	DEFINICION OPERACIONAL	DIMENCIONES	INDICADORES	HERRAMIENTA
VARIABLE INDEPENDIENTE Sistema de Predicción del Rendimiento Académico	Es una herramienta de inteligencia artificial que brinda información para la adecuada toma de decisiones.	La herramienta sistema de predicción es para evaluar los indicadores del rendimiento académico de los estudiantes.	Técnicas de Minería de datos	Tiempo de Procesamiento de Modelo	WEKA
VARIABLE DEPENDIENTE Rendimiento Académico de los Estudiantes	Es el fruto del esfuerzo y la capacidad de trabajo del estudiante, de las horas del estudio, de la competencia y el entrenamiento para la concentración.	Determinar los resultados del rendimiento de los estudiantes como aprobados y desaprobados.	Demográficos	Edad Genero Estado civil Situación Laboral	Matriz de Datos
			Nivel Socioeconómico	Ingreso Económico Familiar	
				Nivel educativo de los padres	
			Rendimiento Académicos	Calificación final (Promedio Ponderado) al terminar el semestre.	

Fuente: Elaboración propia. Utilizando Windows 8

4.3. Diseño de investigación

Transeccional Correlacional causal: estos diseños describen relaciones entre dos o más categorías en un momento dado.



Donde:

X= Variable independiente (Sistema de predicción del rendimiento académico).
r= posibles correlaciones.
Y= Variable dependiente (Rendimiento Académico de los estudiantes).

Para realizar una adecuada contrastación de los resultados, se llevó acabo varios experimentos con diferentes cantidades de muestra que se mencionaran a continuación:

Experimento Número 01.

El primer experimento se realizó con una muestra socioeconómico de 1380 estudiantes de Ingeniera de Sistemas, con 13 variables y una clase.

Experimento Número 02.

El primer experimento se realizó con una muestra socioeconómico de 1674 estudiantes de Ingeniera de Sistemas, con 13 variables y una clase.

Experimento Número 03.

El primer experimento se realizó con una muestra socioeconómico de 1975 estudiantes de Ingeniera de Sistemas, con 13 variables y una clase.

4.4. Población y muestra

4.4.1. Población

La población, está conformada por los datos históricos de los Estudiantes de Ingeniería de Sistemas de la Universidad Nacional José María Arguedas como el nivel socioeconómico familiar y promedio ponderado de los años (2013-I a 2018-II).

4.4.2. Muestra

La muestra fue tomada de manera no probabilística, de tal manera que aporte al estudio, los datos tomados fueron el nivel socioeconómico familiar y el promedio ponderado de los Estudiantes de Ingeniería de Sistemas.

De todos los datos, se obtuvo un total de 1975 datos, los cuales fueron separados en tres muestras en 1380,1674 y finalmente en un total 1975.

Los datos de los estudiantes de Ingeniería de Sistemas, fueron divididos proporcionalmente para formar cada una de las muestras.

Ultimo se tomó los datos del semestre 2018_I para calcular el error de la predicción realizado por el sistema de predicción de rendimiento académico.

4.5. Método de investigación

En la presente tesis, se desarrolló con el método experimental es un tipo de investigación que bien se utiliza experimentos y los principios encontrados en el medido científico. Método experimental, consiste en implicar la observación, manipulación, registro de las variables (dependiente, independiente, interviniente, etc.).

Según (Martinez, 2015) Método experimental es un procedimiento que permite a la verdad objetiva de los fenómenos que consiste en 4 fases:

1. **Pregunta.-** falta de conocimiento pleno sobre un fenómeno natural que está siendo observado, se expresa en forma de preguntas.

2. **Hipótesis.-** Es una suposición razonable acerca de la respuesta a una pregunta para que una hipótesis sea científica, debe referirse la comprensión de la naturaleza y de ser susceptible de probarse experimentalmente.

3. **Predicción.-** La hipótesis científica se utiliza para predecir las consecuencias que tendría ser cierta y experimentación la predicción.

4. **Experimentación.-** La predicción permite planear entonces un experimento bajo condiciones controladas, cuya realización proporciona medidas y datos que son registrados. La descripción exacta de cada experimento hace posible reproducirlo y obtener los mismos resultados.

5. **Análisis de resultados.-** Consiste en procesar los datos para obtener información, cuya interpretación proporciona conocimiento sobre la predicción, por lo general incluye:
    - ❖ conversión de unidades de medida, tabulación, gráfico y procesamiento matemático de los registrados.
    - ❖ estudio de la relación que tiene los datos entre sí.
    - ❖ correlación de la información procesada y la predicción.

4.6. Técnicas de instrumento de acopio de datos

El base de datos construido, consta de 1975 datos extraídos de base de datos de registros académicos y nivel socioeconómico familiar, se determinó la predicción de los estudiantes de Ingeniería de Sistemas, se determinó la predicción en las siguientes categorías:

Tabla 7: Categorías de clasificación de los estudiantes

Simbología	Categorías	Rango de valores
Sem	Semestre	2013-I a 2018-I
Sex	Sexo	Masculino, Femenino
Ed	Edad	18 – 29
Est_civil	Estado Civil	Soltero , Conviviente
Trab	Trabaja	sí, no
dep_eco_padr	Depende económicamente de sus padres	sí, no
n_herm	Numero de hermanos	1 – 9
t_hijo	Tiene hijo	sí, no
Proce	Procedencia	Andahuaylas, Talavera, San Jerónimo, otros distritos
i_e_familiar	Ingreso económico familiar	100 – 3000
gr_inst_padre	Nivel académico del padre	sin nivel, primaria, secundaria, técnico, universitario
gr_inst_madre	Nivel académico de la madre	sin nivel, primaria, secundaria, técnico, universitario
p_pond	Promedio ponderado	9.86 – 16.38

Fuente: Elaboración propia.

Los datos, fueron transformados al formato ARFF (formato de weka), para un adecuado aprendizaje de las técnicas de minería de datos. El archivo ARFF consta de 3 partes: @relation, @attribute y @data.

a) **@relation.-** Es una cadena de caracteres que indica lo que está tratando en este formato.

b) **@attribute.-** En esta sección se declaró los atributos (palabras) y su respectivo tipo (en este caso son de tipo numérico y nominal), Así también de declaro los atributos de clase como real.

c) **@data.-** Aquí, se encuentra los datos de los Estudiantes de Ingeniería de Sistemas transformados en formato ARFF de weka.

```
@relation prediccion
@attribute sem {2013_I, 2013_II, 2014_I, 2014_II, 2015_I, 2015_II, 2016_I, 2016_II, 2017_I, 2017_II,
@attribute sex {masculino, femenino}
@attribute ed numeric
@attribute Est_civil {soltero, conviviente}
@attribute trab {si, no}
@attribute dep_eco_padr {si, no}
@attribute n_herm numeric
@attribute t_hijo {si, no}
@attribute proce {Andahuaylas, Talavera, San_Jeronimo, Pacucha, Ocobamba, Andarapa, Cusco, Ayacucho,
@attribute i_e_familiar numeric
@attribute gr_inst_padre {sin_estudio, primaria_incompleta, primaria_completa, secundaria_incompleta
@attribute g_inst_madre {sin_estudio, primaria_incompleta, primaria_completa, secundaria_incompleta,
@attribute p_pond real
@data
2013_I,femenino,21,soltero,si,si,2,no,Andahuaylas,600,secundaria_completa,secundaria_completa,10.67
2013_I,masculino,23,soltero,no,si,5,no,Andahuaylas,800,secundaria_completa,primaria_incompleta,10.9
2013_I,femenino,22,soltero,si,no,5,no,Andahuaylas,480,secundaria_completa,primaria_incompleta,11.3
2013_I,masculino,21,soltero,no,si,4,no,Andahuaylas,583,primaria_incompleta,sin_estudio,11.45
2013_I,femenino,20,soltero,si,si,6,no,Andahuaylas,400,primaria_completa,primaria_completa,11.57
2013_I,masculino,21,soltero,no,si,2,no,Andahuaylas,1700,secundaria_incompleta,primaria_incompleta,1
2013_I,masculino,21,soltero,no,si,3,no,Andahuaylas,2500,superior_completa,superior_completa,11.68
```

Primer Parte

Secundo Parte

Imagen 2: Fichero con formato ARFF

4.7. Técnicas de análisis de datos

Para el análisis de datos (proceso KDD) se utilizará la herramienta Weka que permitirá dirigir la búsqueda y hacer refinamientos, con una interpretación adecuada de los resultados generados.

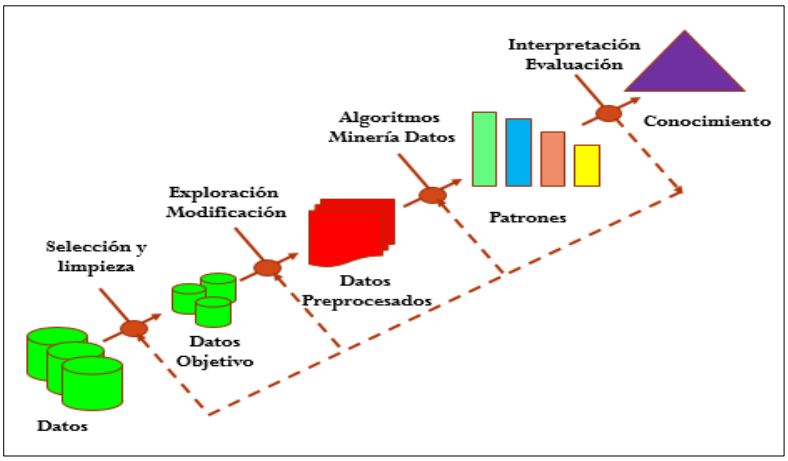


Figura 2: Proceso de extracción de datos (KDD)

Una vez que las técnicas de minería de datos (algoritmos), hayan aprendido del corpus de conocimiento, se hizo un experimento con 3 conjuntos de datos de prueba. Es allí donde se vio la precisión y los errores con lo que se clasifica para la predicción. Para el análisis de datos se utilizó la estadística implementada en weka como: Correlation coefficient, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error.

La prueba se realizó con diferentes algoritmos como: KStar, LWL, IBK, ZeroR, M5Rules, M5P, RandomForest, RandomTree, REPTree para ver cuál de los algoritmos es más óptimo para la predicción. De todos los algoritmos el menor error que arroja es el KStar.

CAPÍTULO V RESULTADOS

5.1. Elección de algoritmo de clasificación

5.1.1. Descripción de la solución desarrollada

Para poder saber cómo funciona el sistema de predicción basado en minería de datos, influye en clasificación de nivel socioeconómico familiar y promedio ponderado, me apoye de la metodología KDD. Esta metodología me permitió organizar en 4 fases para llegar a un objetivo. Las etapas que se siguieron fueron:

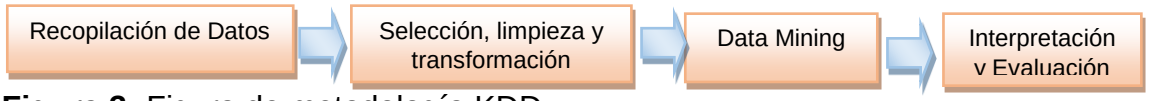


Figura 3: Figura de metodología KDD

5.1.2. Recolección de datos

En esta primera etapa, se realizó la recolección de datos de los Estudiantes de Ingeniería de Sistemas como: su edad, sexo, estado civil, número de hermanos, dependencia económicamente de sus padres, nivel socioeconómico familiar, grado de instrucción de su padre, grado de instrucción de su madre, procedencia y su promedio ponderado de cada semestre desde el año 2013-I hasta el año 2018-I.

5.1.3. Selección, limpieza, y transformación

En esta segunda etapa, es el procesamiento de datos. En donde se realizó la limpieza de forma manual.

Para el uso de técnicas de minería de datos, los datos fueron convertidos en el formato ARFF de Weka. Weka es una herramienta de minería de datos que posee un conjunto de técnicas para el análisis de datos. Los datos utilizados son nominales, numéricos y una determinada clase.

La transformación de los datos al formato weka, nos servirá para entrenar y evaluar con los algoritmos de clasificación. A continuación, se mostrará una pequeña parte en el formato ARFF utilizado.

```
@relation prediccion
@attribute sem {2013_I, 2013_II, 2014_I, 2014_II, 2015_I, 2015_II, 2016_I, 2016_II, 2017_I, 2017_II,
@attribute sex {masculino, femenino}
@attribute ed numeric
@attribute Est_civil {soltero, conviviente}
@attribute trab {si, no}
@attribute dep_eco_padr {si, no}
@attribute n_herm numeric
@attribute t_hijo {si, no}
@attribute proce {Andahuaylas, Talavera, San_Jeronimo, Pacucha, Ocobamba, Andarapa, Cusco, Ayacucho,
@attribute i_e_familiar numeric
@attribute gr_inst_padre {sin_estudio, primaria_incompleta, primaria_completa, secundaria_incompleta
@attribute g_inst_madre {sin_estudio, primaria_incompleta, primaria_completa, secundaria_incompleta,
@attribute p_pond real

@data
2013_I,femenino,21,soltero,si,si,2,no,Andahuaylas,600,secundaria_completa,secundaria_completa,10.67
2013_I,masculino,23,soltero,no,si,5,no,Andahuaylas,800,secundaria_completa,primaria_incompleta,10.91
2013_I,femenino,22,soltero,si,no,5,no,Andahuaylas,480,secundaria_completa,primaria_incompleta,11.3
2013_I,masculino,21,soltero,no,si,4,no,Andahuaylas,583,primaria_incompleta,sin_estudio,11.45
2013_I,femenino,20,soltero,si,si,6,no,Andahuaylas,400,primaria_completa,primaria_completa,11.57
2013_I,masculino,21,soltero,no,si,2,no,Andahuaylas,1700,secundaria_incompleta,primaria_incompleta,11
2013_I,masculino,21,soltero,no,si,3,no,Andahuaylas,2500,superior_completa,superior_completa,11.68
2013_I,masculino,18,soltero,si,no,3,no,Andahuaylas,800,secundaria_completa,primaria_completa,11.73
```

Imagen 3: Formato ARFF construido

5.1.4. Data mining

En esta tercera etapa, se utilizó el conjunto de librerías de java que posee la herramienta weka para la extracción de conocimientos de base de datos.

La herramienta weka realiza varias tareas como: clasificación, regresión, clustering, asociación y visualización, por lo tanto, para esta tesis se utilizó los algoritmos de clasificación.

A continuación, se detallará los pasos como se aplicó las diferentes técnicas de algoritmos de weka al corpus de conocimiento.

- **Primer paso:** se cargó la base de datos ya transformados en el formato ARFF a la herramienta Weka.

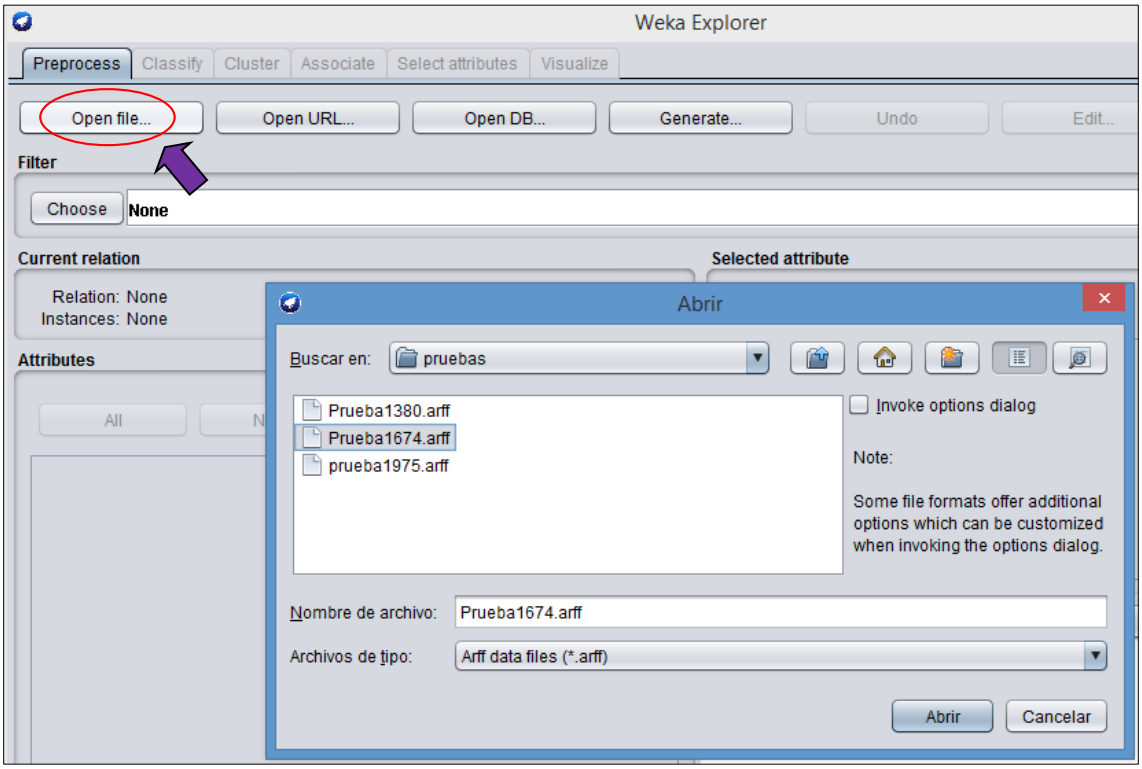


Imagen 4: Cargado del archivo ARFF en la herramienta Weka.

Fuente: Elaboración propia.

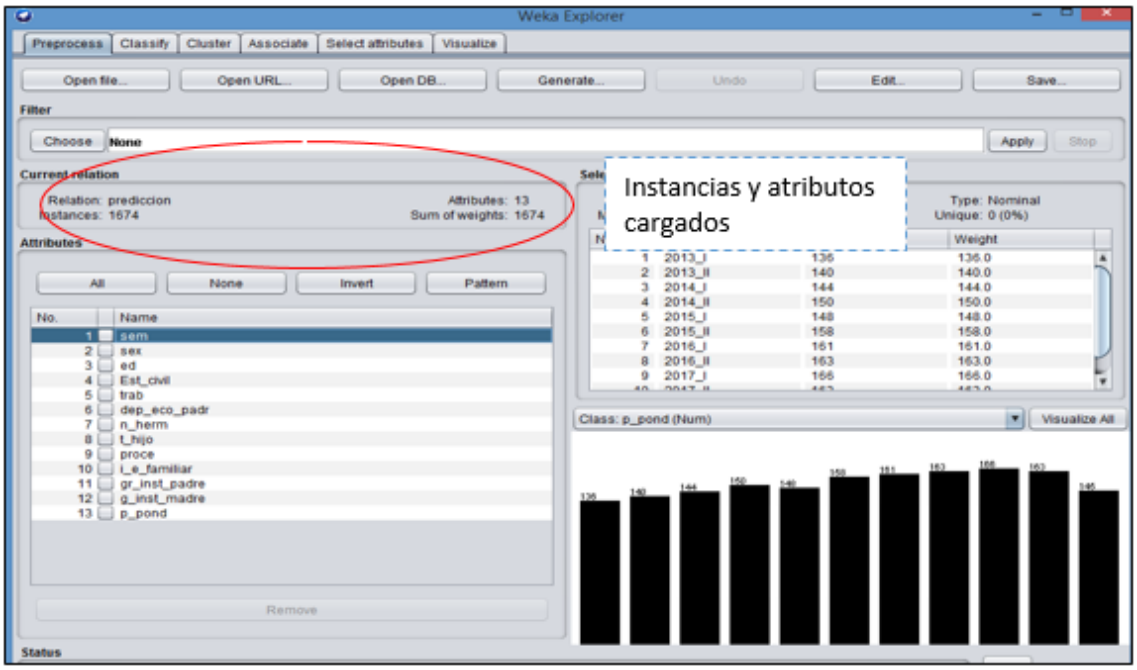


Imagen 5: Verificación de la cantidad de instancias y atributos cargados

Fuente: Elaboración propia.

- **Segundo paso:** se eligió la tarea a realizar, en este caso fue elegido el de clasificación. Luego la técnica y algoritmo a evaluar.

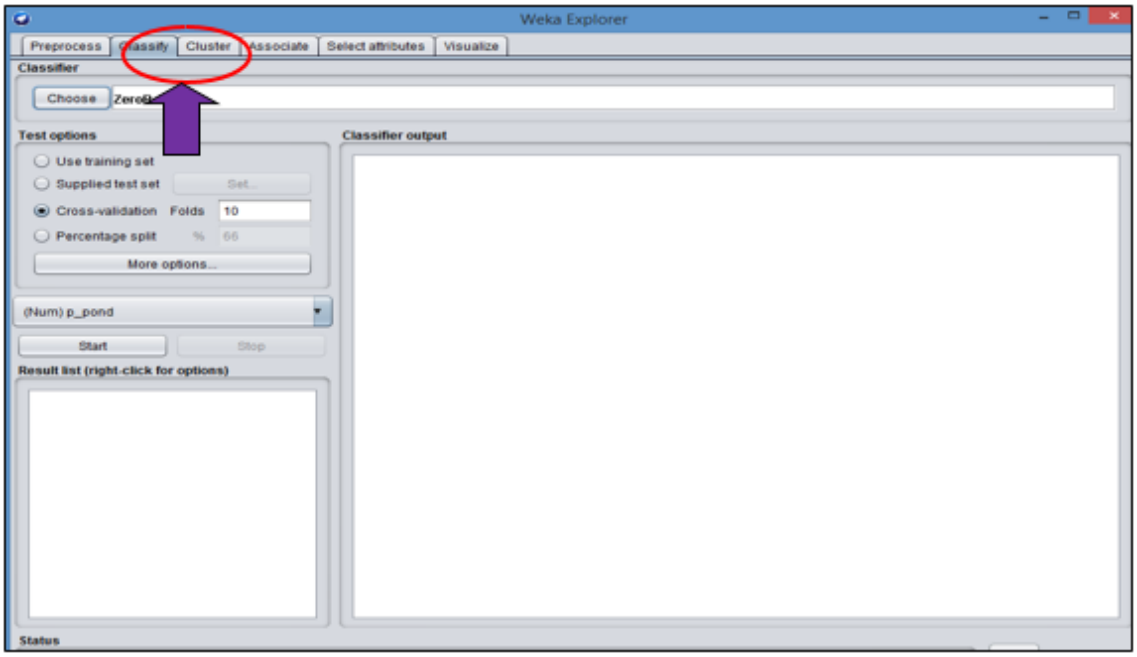


Imagen 6: Elección de la tarea que se realizó (clasificación)

Fuente: Elaboración propia.

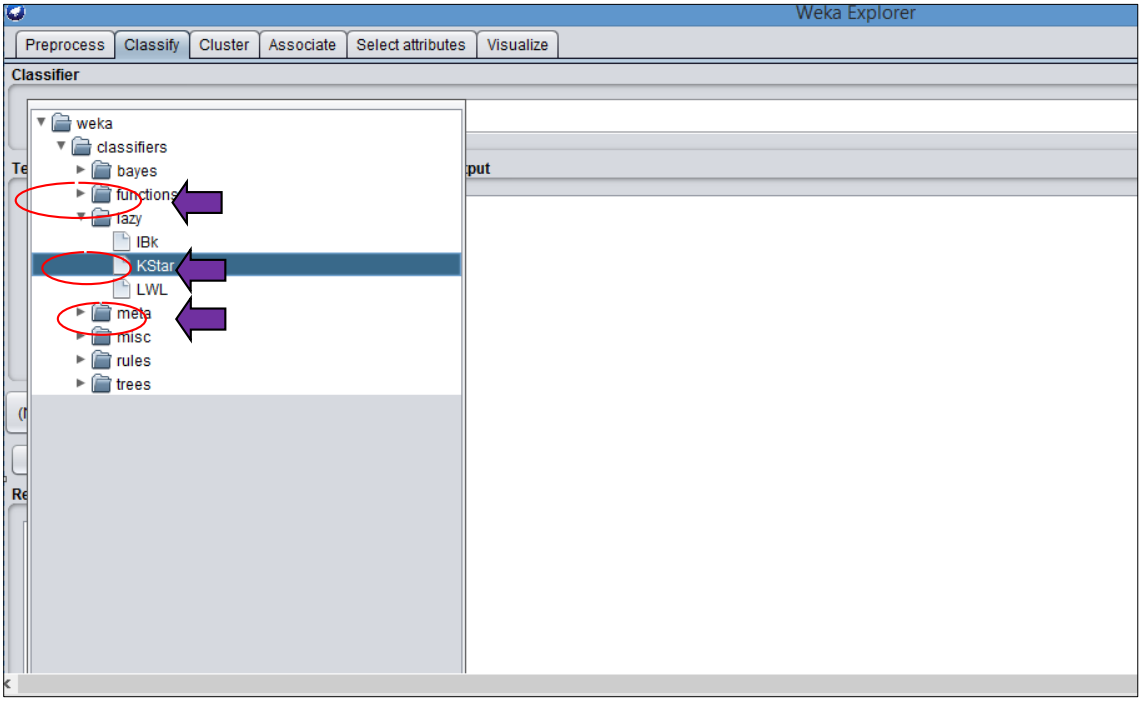


Imagen 7: Técnicas y algoritmos que se evaluaron

Fuente: Elaboración propia.

- **Tercer paso:** en este caso para la prueba fue elegido cross-validation de tamaño 10.

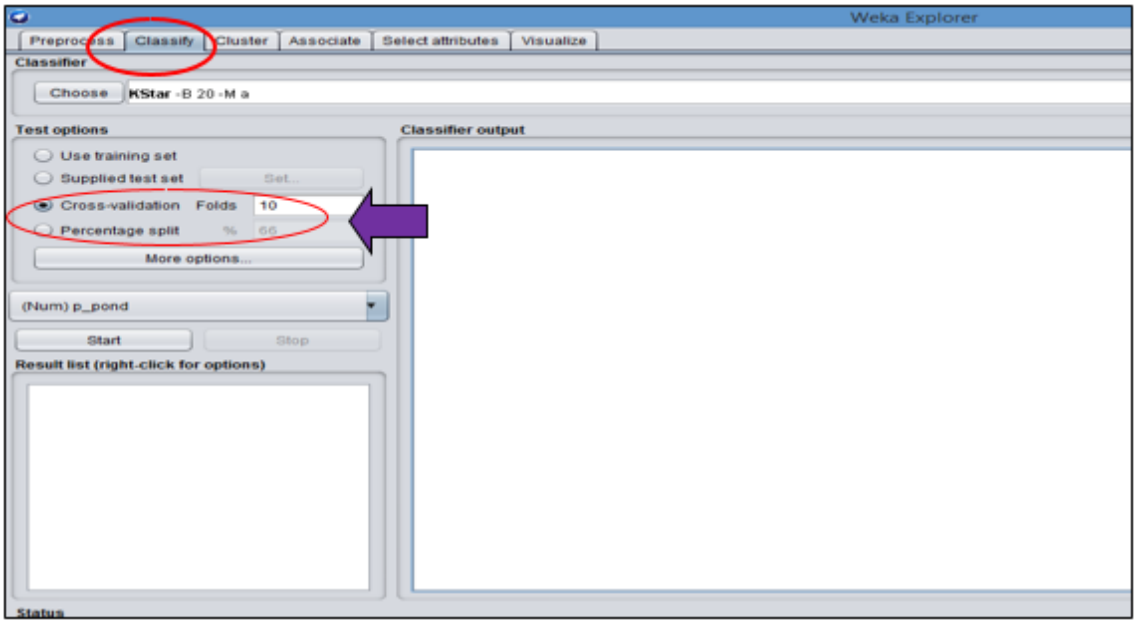


Imagen 8: Prueba Crossvalidation de tamaño 10

Fuente: Elaboración propia.

- **Cuarto paso:** Inicio de construcción y validación del modelo de A continuación, se detallará cada uno de las fases:

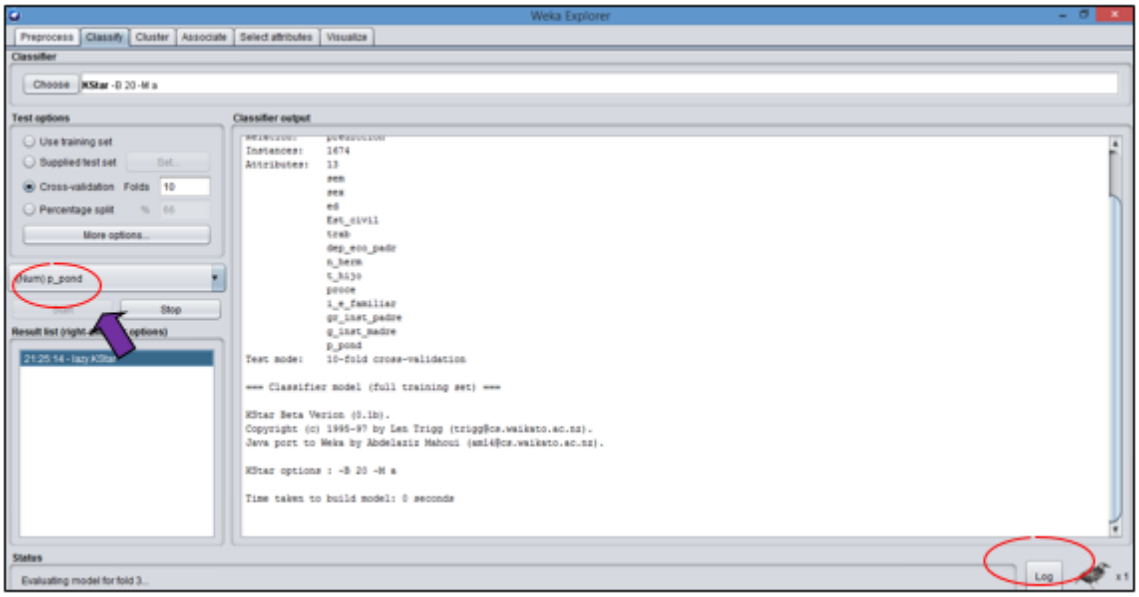


Imagen 9: Construcción del modelo de clasificación

Fuente: Elaboración propia

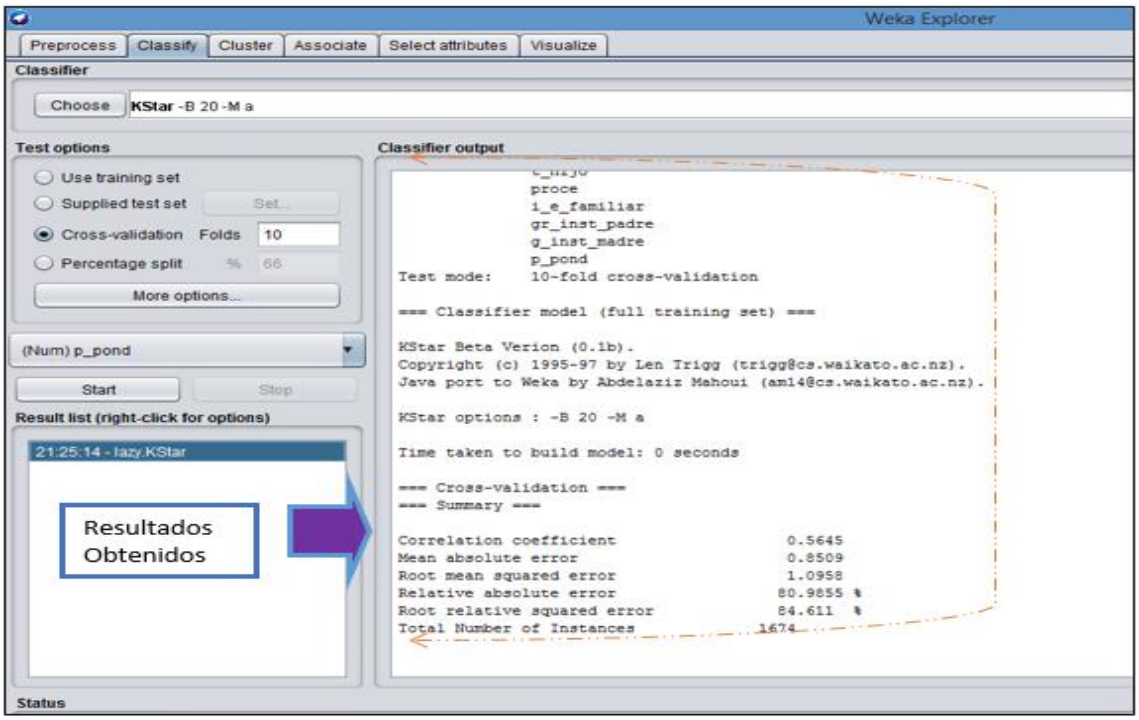


Imagen 10: Validación del modelo de clasificación

Fuente: Elaboración propia.

- **Quinto paso:** En este paso se guardó el modelo de clasificación para la implementación de un sistema de predicción del rendimiento académico de los estudiantes de Ingeniería de Sistemas.

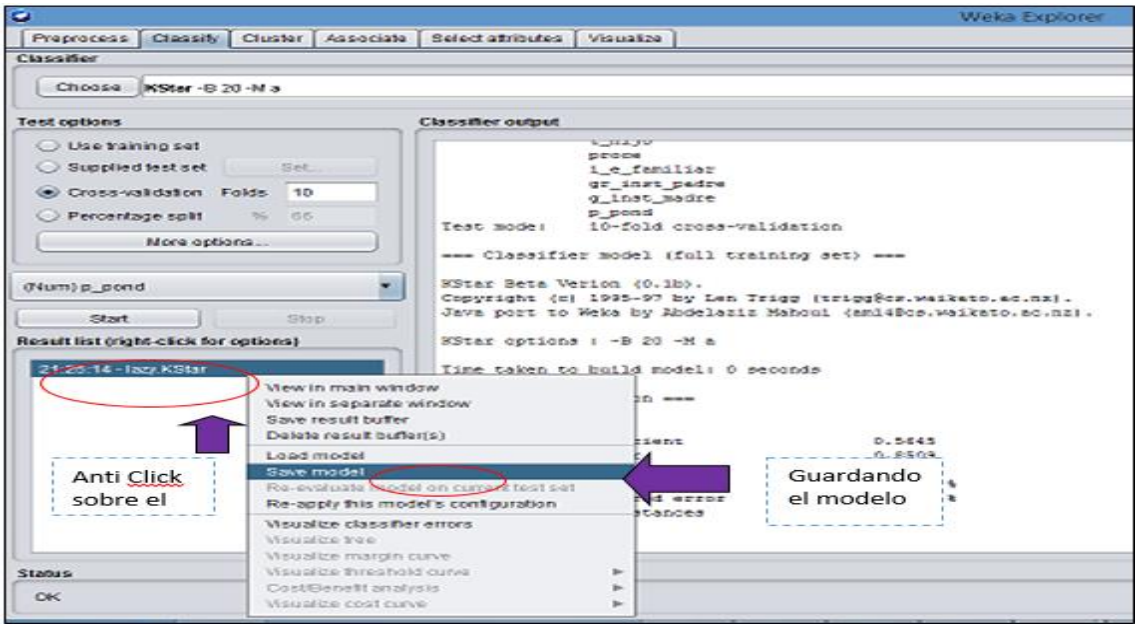


Imagen 11: Guardado del modelo de clasificación

Fuente: Elaboración propia.

5.1.5. Interpretación y evaluación

Como se mostró en las etapas anteriores (Data Mining), los algoritmos que fueron evaluados para el corpus de conocimiento son: SMOreg, IBK, Kstar, LWL, Bagging, Vote, M5Rules, ZeroR, M5P, RandomForest, RandomTree, REPTree. Para la evaluación de estos se usó la prueba estadística del tipo Cross – Validation 10 que nos ofrece Weka, este tipo de prueba hace que la muestra original se divida aleatoriamente en 10 submuestras. De la cual una de ellas se retiene como dato de validación para probar el modelo y el resto como datos de formación de modelo.

De todos los algoritmos que se mencionó el KStart fue el algoritmo que dio mejor resultado, la clasificación que realiza es más preciso a la comparación de los demás algoritmos, por lo cual se escogió ese modelo para la implementación del sistema de predicción de rendimiento académico de los estudiantes de Ingeniería de Sistemas.

5.2. Presentación de resultados

Resultados del experimento Número 01.

A continuación, se presenta los resultados obtenidos de los algoritmos como: SMOreg, IBK, Kstar, LWL, Bagging, Vote, M5Rules, ZeroR, M5P, RandomForest, RandomTree, REPTree con una muestra de 1380 datos de los estudiantes de Ingeniería de Sistemas.

Tabla 8: Resultados del experimento Número 01

Algoritmos	**SMOreg**	**IBK**	**Kstar**	**LWL**	**Bagging**	**Vote**	**M5Rules**	**ZeroR**	**M5P**	**RandomForest**	**RandomTree**	**REPTree**
Error Absoluto Medio	1.34	1.52	1.18	1.40	1.34	1.45	1.35	1.45	1.35	1.25	1.74	1.41
Tiempo de generación del modelo	3.32 S	0.01 S	99 ms	27 ms	0.23 S	49 ms	0.7 S	59 ms	0.81 S	0.62 S	0.02 S	0.03 S

Fuente: Elaboración propia.

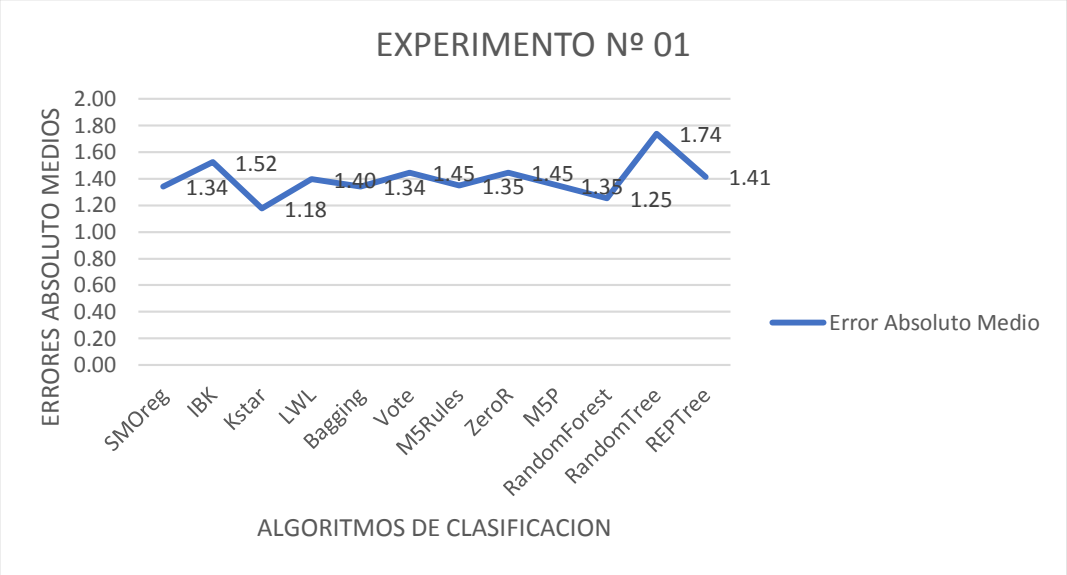


Gráfico 1: Gráfico de estadística del error absoluto medio (experimento 01)

Fuente: Elaboración propia.

Resultados del experimento Número 02.

A continuación, se presenta los resultados obtenidos de los algoritmos como: SMOreg, IBK, Kstar, LWL, Bagging, Vote, M5Rules, ZeroR, M5P, RandomForest, RandomTree, REPTree con una muestra de 1674 datos de los estudiantes de Ingeniería de Sistemas.

Tabla 9: Resultados del experimento Número 02

Algoritmos	SMOreg	IBK	Kstar	LWL	Bagging	Vote	M5Rules	ZeroR	M5P	RandomForest	RandomTree	REPTree
Error Absoluto Medio	0.98	1.09	0.85	1.02	1.03	1.05	0.97	1.05	0.97	0.94	1.20	1.05
Tiempo de generación del modelo	4.39 S	0.01 S	99 ms	98 ms	0.26 S	50 ms	0.57 S	0.01 S	1 S	0.56 S	0.02 S	0.06 S

Fuente: Elaboración propia.

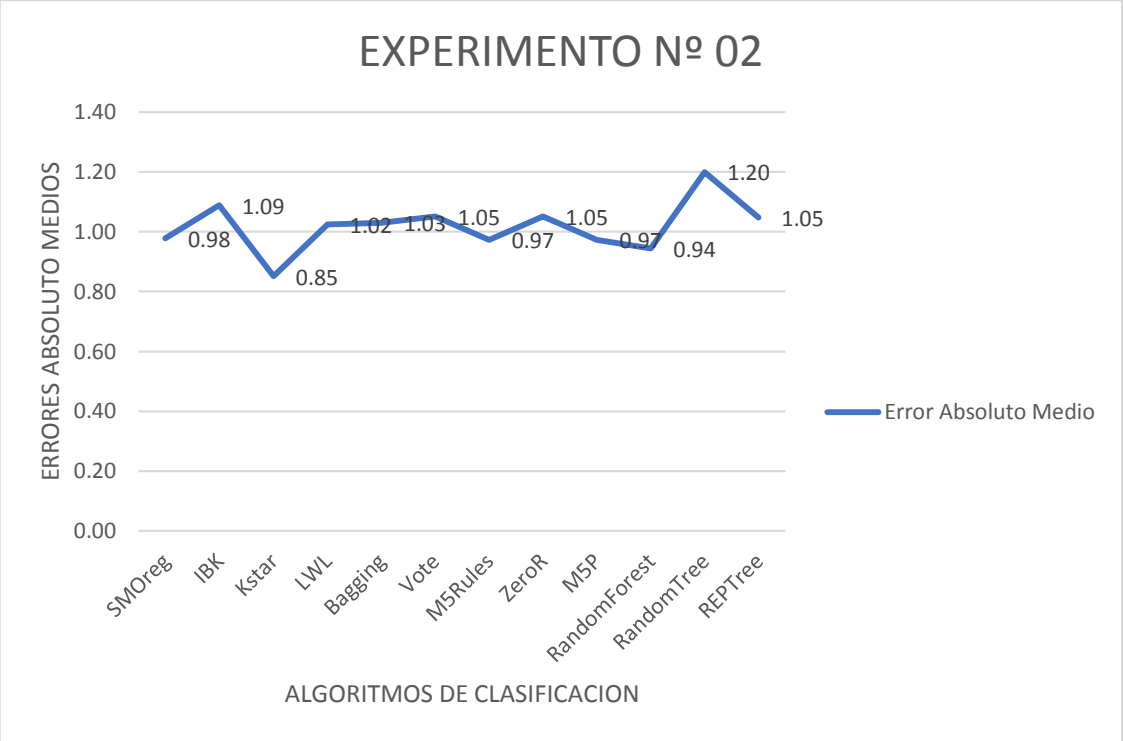


Gráfico 2: Gráfico de estadística del error absoluto medio (experimento 02)

Resultados del experimento Número 03.

A continuación, se presenta los resultados obtenidos de los algoritmos como: SMOreg, IBK, Kstar, LWL, Bagging, Vote, M5Rules, ZeroR, M5P, RandomForest, RandomTree, REPTree con una muestra de 1975 datos de los estudiantes de Ingeniería de Sistemas.

Tabla 10: Resultados del experimento Número 03

Algoritmos	SMOreg	IBK	Kstar	LWL	Bagging	Vote	M5Rules	ZeroR	M5P	RandomForest	RandomTree	REPTree
Error Absoluto Medio	1.35	1.58	1.17	1.44	1.38	1.47	1.37	1.47	1.35	1.26	1.66	1.40
Tiempo de generación del modelo	13.71 S	38 ms	99 ms	59 ms	0.18 S	0.79 S	1.71 S	29 ms	1.71 S	0.93 S	0.02 S	0.09 S

Fuente: Elaboración propia.

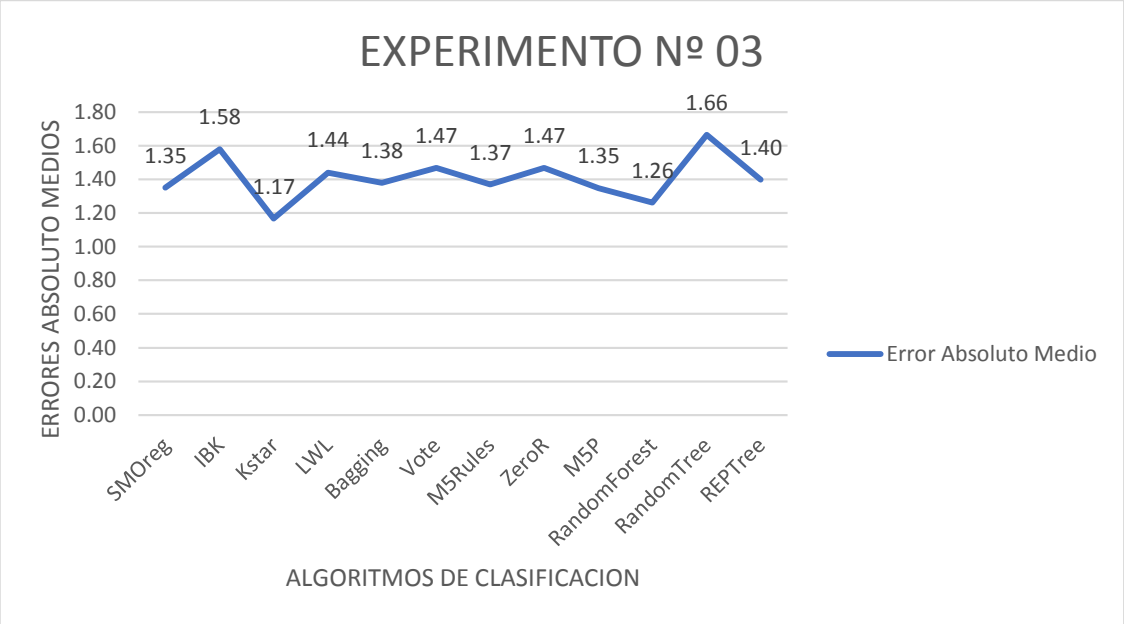


Gráfico 3: Gráfico de estadística del error absoluto medio (experimento 03)

Resumen de los resultados del experimento 01,02 y 03.

A continuación, se presenta el resumen de los resultados obtenidosn en el experimento 01, 02 y 03 de los algoritmos SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree.

Tabla 11: Tabla con el resumen de resultados de los experimentos 01, 02 y 03

Algoritmos	SMOreg	IBK	Kstar	LWL	Bagging	Vote	M5Rules	ZeroR	M5P	RandomForest	RandomTree	REPTree
Error Absoluto Medio	1.34	1.52	1.18	1.40	1.34	1.45	1.35	1.45	1.35	1.25	1.74	1.41
Tiempo de generación del modelo	3.32 s	0.01 s	99 ms	27 ms	0.23 s	49 ms	0.7 s	59 ms	0.81 s	0.62 s	0.02 s	0.03 s
Algoritmos	SMOreg	IBK	Kstar	LWL	Bagging	Vote	M5Rules	ZeroR	M5P	RandomForest	RandomTree	REPTree
Error Absoluto Medio	0.98	1.09	0.85	1.02	1.03	1.05	0.97	1.05	0.97	0.94	1.20	1.05
Tiempo de generación del modelo	4.39 S	0.01 S	99 ms	98 ms	0.26 S	50 ms	0.57 S	0.01 S	1 S	0.56 S	0.02 S	0.06 S

Algoritmos	SMOreg	IBK	Kstar	LWL	Bagging	Vote	M5Rules	ZeroR	M5P	RandomForest	RandomTree	REPTree
Error Absoluto Medio	1.35	1.58	1.17	1.44	1.38	1.47	1.37	1.47	1.35	1.26	1.66	1.40
Tiempo de generación del modelo	13.71 S	38 ms	99 ms	59 ms	0.18 S	0.79 S	1.71 S	29 ms	1.71 S	0.93 S	0.02 S	0.09 S

Fuente: Elaboración propia.

También se muestra el grafico estadístico con el resumen de los Errores Absolutos Medios en los experimentos ya mencionados.

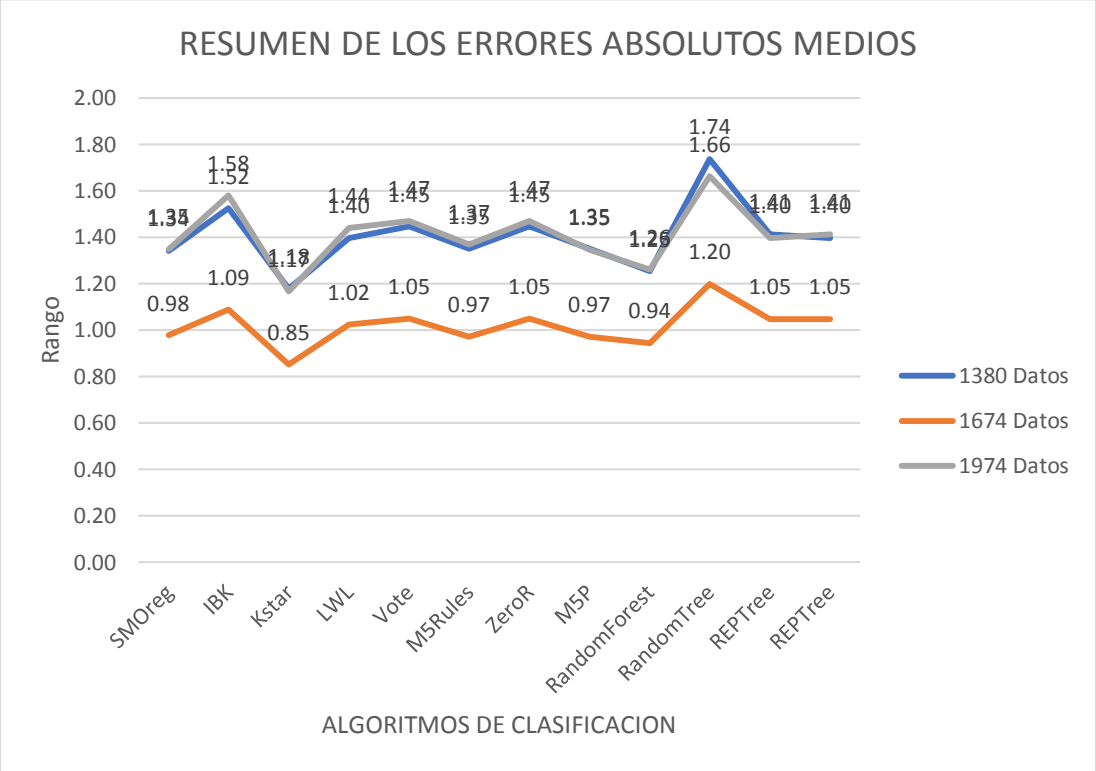


Gráfico 4: Resumen de los errores absolutos medios de los tres experimentos

5.3. Análisis y discusión de resultados

5.3.1. Resultados del experimento Número 01.

El experimento se realizó con una muestra de 1380 datos de los Estudiantes de Ingeniería de Sistemas. Los algoritmos que se utilizaron fueron: SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree. Se realizó una prueba estadística de tipo de Cross-validation 10 que nos ofrece weka, este tipo de prueba hace que la muestra original se divida aleatoriamente en 10 submuestras. De las 10 muestras, una de ellas se retiene como datos de validación para probar el modelo, esto quiere decir que se realizó 10 validaciones cruzadas en 10 hojas cada una, lo cual implica que el algoritmo de aprendizaje se ejecutó 10 veces. En el experimento se comparó los Errores Medios Absolutos de cada algoritmo. En el cuadro de estadística se muestra que el algoritmo KStar es el que tiene menor error absoluto medio siendo 1.18, mientras el resto de los algoritmos el Error Medio Absoluto es mayor a ello como se muestra en tabla 8. Por lo tanto, el algoritmo KStar es más eficiente.

5.3.2. Resultados del experimento Número 02.

El experimento se realizó con una muestra de 1674 datos de los Estudiantes de Ingeniería de Sistemas. Los algoritmos que se utilizaron en el experimento fueron: SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree. Se realizó una prueba estadística de tipo de Cross-validation 10 que nos ofrece weka, este tipo de prueba hace que la muestra original se divida aleatoriamente en 10 submuestras. De las 10 muestras, una de ellas se retiene como datos de validación para probar el modelo, esto quiere decir que se realizó 10 validaciones cruzadas en 10 hojas cada una, lo cual implica que el algoritmo de aprendizaje se ejecutó 10 veces. En el experimento se comparó los errores medios absolutos de cada algoritmo. En el cuadro de estadística se muestra que el algoritmo KStar es el que tiene menor error Absoluto Medio siendo 0.85, mientras el resto de los algoritmos el Error Medio Absoluto es mayor a ello como se muestra en tabla 9. Por lo tanto, el algoritmo KStar es más eficiente.

5.3.3. Resultados del experimento Número 03.

El experimento se realizó con una muestra de 1975 datos de los estudiantes de Ingeniería de Sistemas. Los algoritmos que se utilizaron en el experimento fueron: SMOreg, IBK, Kstar, LWL, Bagging, Vote, ZeroR, M5P, RandomForest, RandomTree, REPTree. Se realizó una prueba estadística de tipo de Cross-validation 10 que nos ofrece weka, este tipo de prueba hace que la muestra original se divida aleatoriamente en 10 submuestras. De las 10 muestras, una de ellas se retiene como datos de validación para probar el modelo, esto quiere decir que se realizó 10 validaciones cruzadas en 10 hojas cada una, lo cual implica que el algoritmo de aprendizaje se ejecutó 10 veces. En el experimento se comparó los errores medios absolutos de cada algoritmo. En el cuadro de estadística se muestra que el algoritmo KStar es el que tiene menor error Absoluto Medio siendo 1.17, mientras el resto de los algoritmos el Error Medio Absoluto es mayor a ello como se muestra en tabla 10. Por lo tanto, el algoritmo KStar es más eficiente.

5.4. Comprobación de las hipótesis

5.4.1. Comprobación de hipótesis especificas

Hipótesis especifica I

H1: Es posible determinar el algoritmo más óptimo y que genere el mejor modelo predictivo para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.

De acuerdo a los resultados obtenidos en el experimento 01, 02 y 03, el algoritmo más óptimo para el modelo de clasificación construido es el algoritmo KStar, que pertenece a los clasificadores de Lazy. Por lo tanto, la hipótesis H1 es verdadera.

Hipótesis especifica II

H2: Es posible medir el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas en el semestre 2018-I de UNAJMA utilizando Minería de datos, con un grado de error mínimo 15%.

De acuerdo a los resultados obtenidos con los datos reales de los estudiantes de Ingeniería de Sistemas del semestre 2018-I y los datos predichos de dichos datos el error es 11%. Por lo tanto, la hipótesis H2 es verdadera.

5.4.2. Comprobación de hipótesis general

Ho: Es posible evaluar la eficiencia de un sistema de rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando Minería de datos.

Con el resultado obtenido con el experimento 02 se creó el modelo de clasificación con el cual se implementó el sistema de predicción de rendimiento académico, demuestra que el sistema es eficiente. Por lo tanto, la hipótesis Ho es verdadera.

CAPÍTULO VI: IMPLEMENTACIÓN DEL SISTEMA

Después de haber obtenido el modelo de clasificación más óptimo con los 1674 datos de los Estudiantes de Ingeniería de Sistemas. Se prosiguió a desarrollar el sistema de predicción de rendimiento académico. Para ello se utilizó la metodología en cascada lo cual consiste en 4 fases como: Análisis de requisitos, Diseño del sistema, implementación del sistema y pruebas del sistema. A continuación, se mencionará cada una de las fases:

6.1. Análisis de requisitos.

En esta etapa, se definió la función que debe cumplir el sistema de predicción de rendimiento académico de los estudiantes de Ingeniería de Sistemas. El objetivo del sistema de predicción es la precisión de las notas que se sacara en el siguiente semestre.

6.2. Diseño del sistema.

Para el desarrollo del sistema de predicción del rendimiento académico, se eligió NetBeans IDE 8.2 como entorno de desarrollo, star uml para el diagrama de componentes. También se realizó un diagrama de componentes para poder identificar los archivos de clase y ver la dependencia que existe entre ellas. A continuación, se muestra el diagrama de componentes.

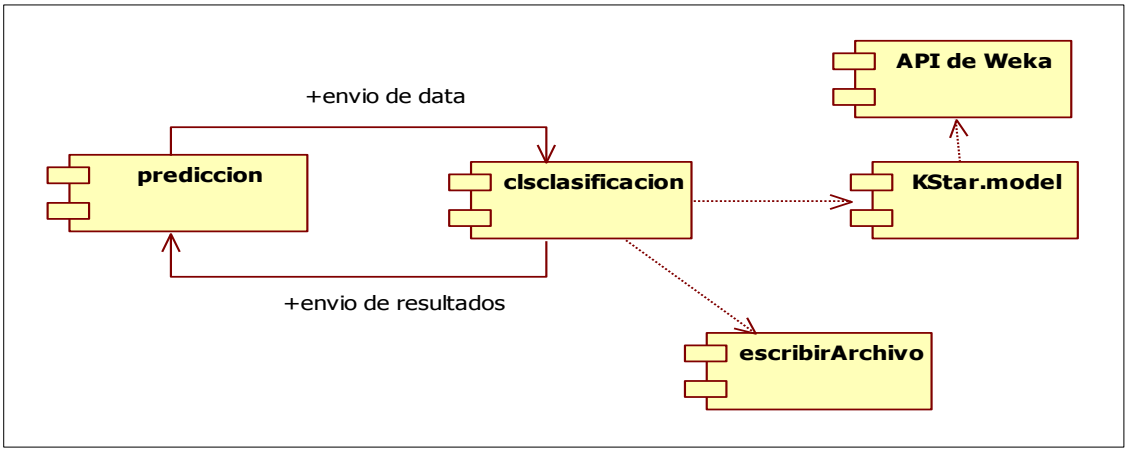


Imagen 12: Diagrama de componentes

- **API de weka:** Este API de Weka se usó la colección de algoritmos de Weka versión 3.9 para el análisis de datos y construcción del modelo de clasificación cuya extensión es “. model”.

- **Kstar.model:** Es un archivo generado por la API de Weka. El modelo el modelo de clasificación. El algoritmo Kstar fue utilizado por haber obtenido mejores resultados.

- **Clsclasificacion.java:** Es la clase que incorpora al modelo KStar y el producto generado por escribirAchivo.java, para darnos la clasificación con los datos de Los Estudiantes de Ingenieria de Sistemas.

- **EscribirArchivo.java:** Esta clase fue implementada en NetBeans, realiza escribir archivos ingresado por el usuario en un formato ARFF para ser conocido por el modelo de clasificación.

- **Prediccion.java:** aquí, se muestra al usuario el resultado de su predicción de rendimiento académico para el siguiente semestre.

6.3. Implementación del sistema

En esta fase, se programó el sistema de predicción del rendimiento Académico en el programa NetBeans IDE 8.2.

- **Predicción.java:** En este predicción.java está desarrollado el interfaz, donde los estudiantes ingresan los datos socioeconómicos para la clasificación o predicción del rendimiento académico.

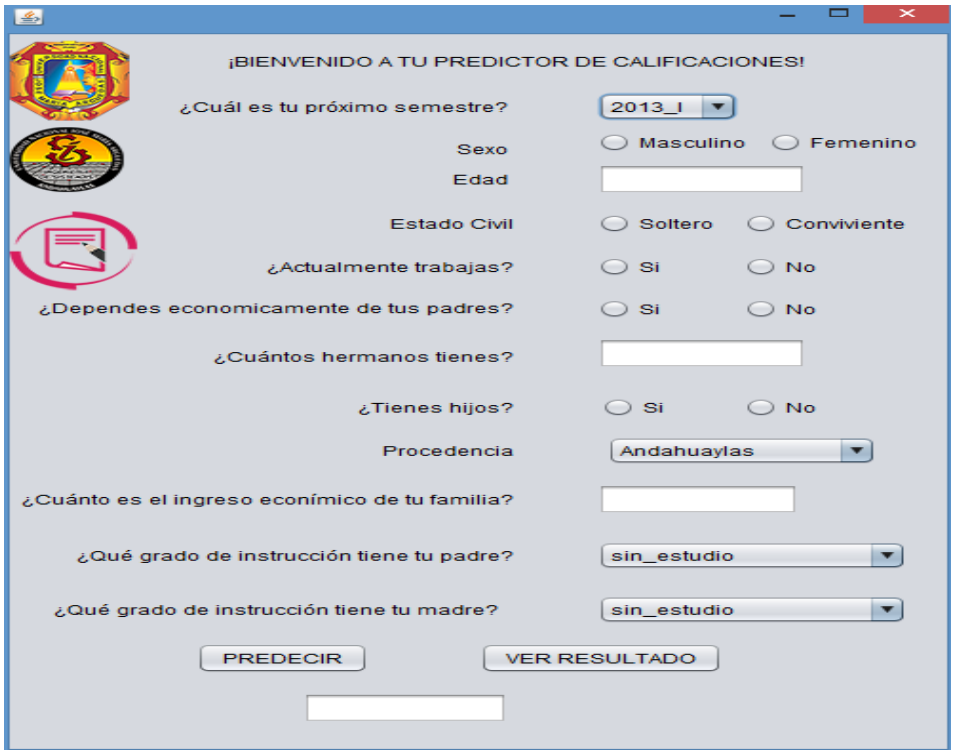


Imagen 13: Interfaz para el usuario

- **Modelo KStar.model:** será utilizado por la clase de clsclasificacion.java para clasificar nuevos datos ingresados por el usuario. A continuación, se muestra la información.

```
=== Run information ===

Scheme:       weka.classifiers.lazy.KStar -B 20 -M a
Relation:     prediccion
Instances:    1674
Attributes:   13
              sem
              sex
              ed
              Est_civil
              trab
              dep_eco_padr
              n_herm
              t_hijo
              proce
              i_e_familiar
              gr_inst_padre
              g_inst_madre
              p_pond
Test mode:    10-fold cross-validation

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

KStar Beta Version (0.1b).
Copyright (c) 1995-97 by Len Trigg (trigg@cs.waikato.ac.nz).
Java port to Weka by Abdelaziz Mahoui (am14@cs.waikato.ac.nz).

KStar options : -B 20 -M a

Time taken to build model: 0.03 seconds
```

Imagen 14: Información del modelo de clasificación.

Fuente: Elaboración propia.

- **EscribirArchivo.java:** escribe los datos ingresados por el usuario al formato ARFF, para poder ser reconocido y clasificado por el modelo de clasificación.

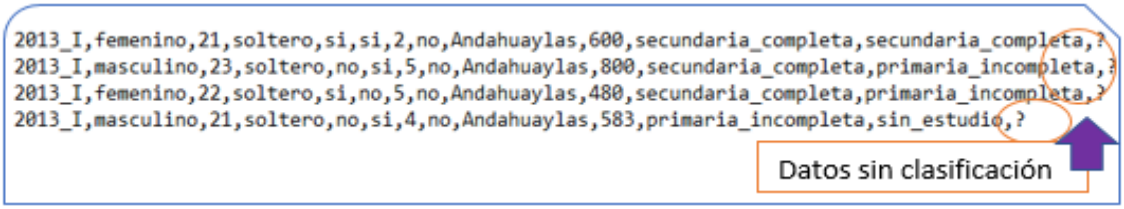


Imagen 15: Resultado obteniendo de la clase transformacion.java

Fuente: Elaboración propia.

- **Clsclasificacion.java:** se encarga de cargar el modelo de clasificación (model. KStar) para darle una clase al escribirarcivo.java, para devolver un resultado deseado.

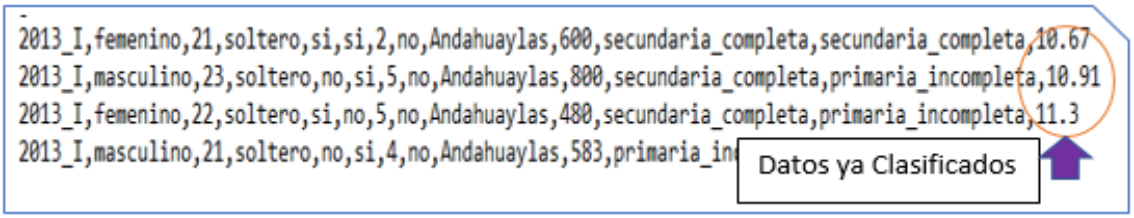


Imagen 16: Resultado obteniendo de la clase clasificacion.java

- **Clsclasificar:** Muestra el resultado final de la clasificación al usuario.

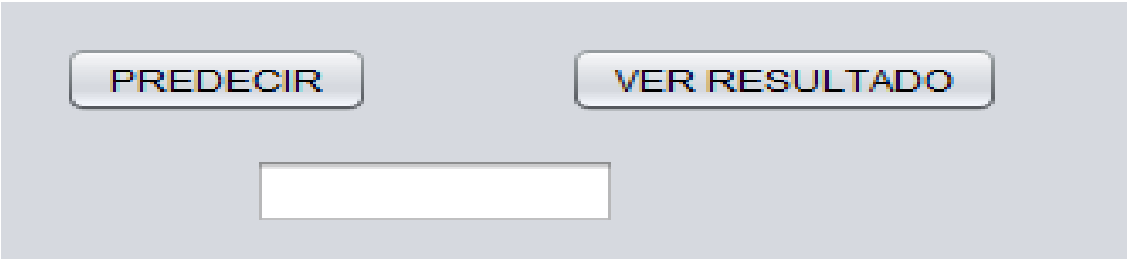


Imagen 17: Resultado final de predicción

- **Pruebas del Sistema:** una vez implementado el sistema de predicción del rendimiento académico, se verifico que todas sus clases realicen las tareas de manera correcta, de esta manera se pudo obtener el resultado final de la clasificación.

CONCLUSIONES

- El corpus de conocimiento es eficiente de manera proporcional en el tiempo de generación de los modelos de clasificación. El algoritmo Kstar con 1380 datos, genero el modelo de clasificación en 99 milisegundo; con 1674 datos, lo realizo en 99 milisegundo y con 1975 datos se realizó 99 milisegundo.

- El algoritmo que tiene menor Error Absoluto Medio (0.85), que los de más para el corpus de conocimiento construido con 1674 datos es Kstar. Por lo tanto, su modelo es más eficiente para la predicción de rendimiento académico.

- Se determino el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos. Los resultados del cálculo del Error Absoluto Medio es 0.11 es menor a 15%.

RECOMENDACIONES

- Se recomienda clasificar los datos de los estudiantes no solo para predecir el rendimiento académico, sino para otros indicadores, siendo la minería de datos una herramienta con un potencial increíble y aplicable en un sinfín de circunstancias, finalidades y proyectos.

- Para encontrar una mayor exactitud se recomienda tener base de datos socioeconómicos veraces de todos los estudiantes de Ingeniería de Sistemas en digital.

- Es importante que la universidad recolecte suficiente información referente a las características de los estudiantes al momento de ingresar a estudiar, ya que facilitan establecer diferentes técnicas para detectar los factores que afecten al estudiante, como las vimos en este estudio de tesis.

REFERENCIAS BIBLIOGRÁFICAS

Alejandro, B. P. (31 de 05 de 2017). *Enciclopedia Financiera*. Obtenido de Enciclopedia Financiera: https://www. /definicion-coeficiente-de-correlacion.

APEIM. (08 de 2017). *APEIM.* Obtenido de APPEIM: http://www.apeim.com.pe /wp-content/themes/apeim/docs/nse/APEIM-NSE-2017.pdf

Cedano, J. A. (2015). *Modelo de minería de datos para identificación de patrones que influyen en el aprovechamiento academico.* La Paz.B.C.S.

Cesar Perez y Daniel Santín. (2007). *Data Mining soluciones con Enterprise Miner.* Mexico: Madrid,España.

César Perez y Daniel Santín. (2007). *Data Mining Soluciones Con Enterprise Miner.* México: MADRID, España.

Diccionario, A. (2017). *Definicion de ABC.* Obtenido de https://www.definicionabc com/general/rendimiento-academico.php

Eduardo Adolfo Porcel, Gladys Noemí Dapozo y María Victoria López. (2010). Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. *Revista Electrónica de Investigación Educativa*, 21.

Fallas, L. C. (16 de Junio de 2011). *Minería de Datos.* Obtenido de Minería de Datos: http://cor-mineriadedatos.blogspot.pe/2011/06/weka.html

Francisca Salinas Esteban, Andres Bodensie, Roberto Lucio Alvarez & Alicia Molina Lizcano. (2010). *Estudio sobre los factores que influyrn en el rendimiento academico.*

Hernández Jácquez, Luis Fernando y Barraza Macías Arturo. (2013). *Rendimiento academico y autoeficiencia percibida.* Mexico: Instituto Universitario Anglo Español.

Jaspe, C. (7 de Julio de 2010). *El Rendimiento Estudiantil y las Estrategias de Enseñanza y Aprendizaje.* Obtenido de El Rendimiento Estudiantil y las Estrategias de Enseñanza y Aprendizaje: http://wwwestrategias264.

blogspot.pe/2010/07/rendimiento-academico-escolar.html

José Hernández Orallo, José Ramírez Quintana y César Ferri Ramírez. (2004). *Introducción a la Minería de Datos.* Madrid: PERSON EDUCACIÓN .S.A.

Julian Perez y Ana Gardey. (2008). *Definiciones.* Obtenido de Definiciones: https://definicion.de/rendimiento-academico/

Julián Pérez Porto y Ana Gardey. (2008). *REDMINE.* Obtenido de REDMINE: https://definicion.de/rendimiento-academico/

María y García . (2017). Análisis de Datos en WEKA – Pruebas de Selectividad. *EcuRed*, 2.

Marquez, A. (06 de 06 de 2012). *Arbol de Decision*. Obtenido de Arbol de Decision: http://descuadrando.com/%C3%81rbol_de_decisi%C3%B3n

Martinez, S. (29 de octubre de 2015). *Metodo Experimental*. Obtenido de Metodo Experimental: https://prezi.com/y6izpav6-ht4/metodo-experimental/

Menacho, C. (2017). Prediction of academic performace applying data mining techniques. *Artículo cientifico*, 32.

Quintan. (10 de 06 de 2018). ECURED. *ECURED*, 01. Obtenido de ECURED: https://www.ecured.cu/Red_Bayesiana

Rivera, M. (02 de 02 de 2011). *El papel de las redes Bayesianas en la toma de deciciones.* Obtenido de El papel de las redes Bayesianas en la toma de deciciones.

Soloriio. (16 de abril de 2013). *metodologia cascada*. Obtenido de metodologia cascada: http://metodologiaencascada.blogspot.com/

stitute, I. (15 de 11 de 2017). *RANKING BY COUNTRY*. Obtenido de RANKING BY COUNTRY: https://es.wikipedia.org/wiki/Clasificaci%C3%B3n_acad%

C3%A9mica_de_universidades_del_Per%C3%BA

SUDO. (11 de Junio de 2014). *SUDO.UTERO.PE*. Obtenido de SUDO.UTERO.PE: http://sudo.utero.pe/2014/06/11/universidades-en-el-peru-parte-3-la-cruda-realidad/

SUNEDU. (02 de octubre de 2017). *portal de traspariencia*. Obtenido de portal de traspariencia: https://www.sunedu.gob.pe/sunedu-otorga-licenciamiento-institucional-a-universidad-nacional-jose-maria-arguedas/

Ucha, F. (15 de Julio de 2007). *Rendimiento Academico*. Obtenido de Rendimiento Academico: https://www.definicionabc.com/general/rendimiento-academico.php

Ucha, F. (Julio de Julio). *Rendimiento Academico*. Obtenido de Rendimiento Academico.

universia. (27 de octubre de 2011). *Universidad Nacional José María Arguedas celebra aniversario*. Obtenido de Universidad Nacional José María Arguedas celebra aniversario: http://noticias.universia.edu.pe/vida-universitaria/noticia/2011/10/27/883059/universidad-nacional-jose-maria-arguedas-celebra-aniversario.html

Universidades, R. W. (2 de Julio de 2017). *Newsletter*. Obtenido de Newsletter: http://www.webometrics.info/es/Latin_America_es/Per%C3%BA

Visitación García Jiménez, Jesús Alvarado Izquierdo y Amelia Jiménez Blanco. (2000). *La predicción del rendimiento académico: regresión lineal versus regresión logística.*

Witten y Frank. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with java implementations.* Madrid: Morgan Kaufmann publishers.

ANEXOS

1. Resultados de modelos de clasificación de algoritmo SMOreg

Modelo generado con 1380 estudiantes:

```
Number of kernel evaluations: 3459563 (95.031% cached)

Time taken to build model: 3.32 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3755
Mean absolute error                      1.3419
Root mean squared error                  1.7738
Relative absolute error                 92.7797 %
Root relative squared error             92.8402 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Number of kernel evaluations: 8515679 (87.219% cached)

Time taken to build model: 4.39 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.378
Mean absolute error                      0.9776
Root mean squared error                  1.2103
Relative absolute error                 93.0496 %
Root relative squared error             93.4566 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 13.71 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3458
Mean absolute error                      1.3495
Root mean squared error                  1.9031
Relative absolute error                 91.8639 %
Root relative squared error             94.2494 %
Total Number of Instances             1975
```

2. Resultados de modelos de clasificación de algoritmo IBK

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3779
Mean absolute error                      1.5243
Root mean squared error                  2.051
Relative absolute error                105.3981 %
Root relative squared error            107.347  %
Total Number of Instances              1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.4042
Mean absolute error                      1.0882
Root mean squared error                  1.3887
Relative absolute error                103.5794 %
Root relative squared error            107.2271 %
Total Number of Instances              1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.34
Mean absolute error                      1.5798
Root mean squared error                  2.2352
Relative absolute error                107.5375 %
Root relative squared error            110.6943 %
Total Number of Instances              1975
```

3. Resultados de modelos de clasificación de algoritmo Kstar

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5929
Mean absolute error                      1.1786
Root mean squared error                  1.5783
Relative absolute error                 81.4897 %
Root relative squared error             82.6058 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5645
Mean absolute error                      0.8509
Root mean squared error                  1.0958
Relative absolute error                 80.9855 %
Root relative squared error             84.611  %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5517
Mean absolute error                      1.1659
Root mean squared error                  1.7317
Relative absolute error                 79.3646 %
Root relative squared error             85.7593 %
Total Number of Instances             1975
```

4. Resultados de modelos de clasificación de algoritmo LWL

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2856
Mean absolute error                      1.3969
Root mean squared error                  1.8327
Relative absolute error                 96.5882 %
Root relative squared error             95.9233 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2158
Mean absolute error                      1.024
Root mean squared error                  1.2639
Relative absolute error                 97.4658 %
Root relative squared error             97.592  %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.1953
Mean absolute error                      1.4404
Root mean squared error                  1.9795
Relative absolute error                 98.0506 %
Root relative squared error             98.0328 %
Total Number of Instances             1975
```

5. Resultados de modelos de clasificación de algoritmo Bagging

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.23 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3859
Mean absolute error                      1.3445
Root mean squared error                  1.7701
Relative absolute error                 92.9606 %
Root relative squared error             92.6472 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.26 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2352
Mean absolute error                      1.0298
Root mean squared error                  1.2839
Relative absolute error                 98.0163 %
Root relative squared error             99.1417 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0.18 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3678
Mean absolute error                      1.3767
Root mean squared error                  1.8855
Relative absolute error                 93.7161 %
Root relative squared error             93.3747 %
Total Number of Instances             1975
```

6. Resultados de modelos de clasificación de algoritmo Vote

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  -0.0943
Mean absolute error                       1.4463
Root mean squared error                   1.9106
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances              1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  -0.0736
Mean absolute error                       1.0506
Root mean squared error                   1.2951
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances              1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0.79 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                   0.3678
Mean absolute error                       1.3767
Root mean squared error                   1.8855
Relative absolute error                  93.7161 %
Root relative squared error              93.3747 %
Total Number of Instances              1975
```

7. Resultados de modelos de clasificación de algoritmo M5Rules

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.7 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3816
Mean absolute error                      1.3496
Root mean squared error                  1.7738
Relative absolute error                 93.3161 %
Root relative squared error             92.8422 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.57 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3899
Mean absolute error                      0.9719
Root mean squared error                  1.1939
Relative absolute error                 92.5082 %
Root relative squared error             92.1916 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 1.71 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3696
Mean absolute error                      1.368
Root mean squared error                  1.8843
Relative absolute error                 93.1208 %
Root relative squared error             93.3176 %
Total Number of Instances             1975
```

8. Resultados de modelos de clasificación de algoritmo ZeroR

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  -0.0943
Mean absolute error                       1.4463
Root mean squared error                   1.9106
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances              1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  -0.0736
Mean absolute error                       1.0506
Root mean squared error                   1.2951
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances              1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  -0.0667
Mean absolute error                       1.4691
Root mean squared error                   2.0193
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances              1975
```

9. Resultados de modelos de clasificación de algoritmo M5P

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.81 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3878
Mean absolute error                      1.3496
Root mean squared error                  1.7713
Relative absolute error                 93.3147 %
Root relative squared error             92.712  %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 1 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3899
Mean absolute error                      0.9719
Root mean squared error                  1.1939
Relative absolute error                 92.5082 %
Root relative squared error             92.1916 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 1.71 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.4017
Mean absolute error                      1.3467
Root mean squared error                  1.8634
Relative absolute error                 91.6703 %
Root relative squared error             92.2825 %
Total Number of Instances             1975
```

10. Resultados de modelos de clasificación de algoritmo RandomForest

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.62 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5015
Mean absolute error                      1.2533
Root mean squared error                  1.6597
Relative absolute error                 86.6576 %
Root relative squared error             86.8707 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.56 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.424
Mean absolute error                      0.9435
Root mean squared error                  1.1722
Relative absolute error                 89.8051 %
Root relative squared error             90.5125 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0.93 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.4823
Mean absolute error                      1.2603
Root mean squared error                  1.7724
Relative absolute error                 85.7929 %
Root relative squared error             87.7744 %
Total Number of Instances             1975
```

11. Resultados de modelos de clasificación de algoritmo RandomTree

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2056
Mean absolute error                      1.7375
Root mean squared error                  2.3009
Relative absolute error                120.1371 %
Root relative squared error            120.4285 %
Total Number of Instances              1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2532
Mean absolute error                      1.1989
Root mean squared error                  1.5171
Relative absolute error                114.1143 %
Root relative squared error            117.1485 %
Total Number of Instances              1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2622
Mean absolute error                      1.6626
Root mean squared error                  2.2873
Relative absolute error                113.1716 %
Root relative squared error            113.274  %
Total Number of Instances              1975
```

12. Resultados de modelos de clasificación de algoritmo REPTree

Modelo generado con 1380 estudiantes:

```
Time taken to build model: 0.03 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3272
Mean absolute error                      1.4135
Root mean squared error                  1.8493
Relative absolute error                 97.7312 %
Root relative squared error             96.7913 %
Total Number of Instances             1380
```

Modelo generado con 1674 estudiantes:

```
Time taken to build model: 0.06 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2048
Mean absolute error                      1.0473
Root mean squared error                  1.2879
Relative absolute error                 99.6866 %
Root relative squared error             99.4448 %
Total Number of Instances             1674
```

Modelo generado con 1975 estudiantes:

```
Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3328
Mean absolute error                      1.3965
Root mean squared error                  1.9322
Relative absolute error                 95.0599 %
Root relative squared error             95.6893 %
Total Number of Instances             1975
```

13. ClsClasificar.java

```
public class clsClasificar {

    public String clsClasificar(Instances label)
        {
            // escribiendo el resultado en un archivo
            FileWriter fichero = null;
            PrintWriter pw = null;
            String clase = null;

            try
                {
                    fichero = new FileWriter("resultado.txt");
                    pw = new PrintWriter(fichero);
                    pw.println(label);
                } catch (Exception e) {
                    e.printStackTrace();
                }
            finally {
                try {
                    if (null != fichero)
                        fichero.close();
                    } catch (Exception e2) {
                     e2.printStackTrace();
                    }
                }

            File archivo = null;
            FileReader fr = null;
            BufferedReader br = null;
            try {
                // Lectura del archivo
                archivo = new File ("resultado.txt");
                fr = new FileReader (archivo);
                br = new BufferedReader(fr);

                String linea;
                String lineaClase;
                ArrayList<String> encontrarClase = new ArrayList<String>();
                int c=0;
                while((linea=br.readLine())!=null)
                    {
                        encontrarClase.add(linea);
                        c=c+1;
                    }
                lineaClase=encontrarClase.get(17);
                String palabras[] = lineaClase.split(",");//extrayendo la clase
                for(String cls : palabras)
                    clase=cls;
                }
            catch(Exception e){
                e.printStackTrace();
                }
            finally{
                 try{
                    if( null != fr ){
                        fr.close();}
                     }catch (Exception e2){
                    e2.printStackTrace();
                    }
                }
        return(clase);
        }
}
```

14. Clsclacificacion.java

```
public class clsclasificacion {
    public String clsclasificacion() throws Exception{

        BufferedReader breader = null;
        breader = new BufferedReader(new FileReader("C:/Users/MANAGER/Desktop/JavaApplication10/prueba1.arff"));
        Instances train = new Instances(breader);
        train.setClassIndex(train.numAttributes() -1);// Seleccionando la columna de los datos a estimar

        breader = new  BufferedReader(new FileReader("C:/Users/MANAGER/Desktop/JavaApplication10/prueba2.arff"));
        Instances test = new Instances(breader);
        test.setClassIndex(test.numAttributes() -1);

        breader.close();

        KStar NB= new KStar();
        NB.buildClassifier(train);//Entrenando el modelo
        Instances labeled = new Instances(test);

        //* Testeando el modelo
        for (int i=0; i< test.numInstances(); i++){
            //double clsLabel = tree.classifyInstance(test.instance(i));
            double clsLabel = NB.classifyInstance(test.instance(i));
            labeled.instance(i).setClassValue(clsLabel);
        }
        //System.out.println(labeled);
        clsClasificar clasificar = new clsClasificar();
        String resultado = clasificar.clsClasificar(labeled);
        return (resultado);
    }

}
```

15. EscribirArchivo.java

```
public class escribirArchivos {
    FileWriter fichero = null;
        PrintWriter pw = null;
        ArrayList<String> procesar;
        public escribirArchivos(ArrayList<String> aprocesar)
        {
            procesar=aprocesar;

            try
                {
                    String[] palabra = {"semestre {2013_I, 2013_II, 2014_I, 2014_II, 2015_I, 2015_II, 2016_I, 2016_II, 2017_I,
                    fichero = new FileWriter("C:/Users/MANAGER/Desktop/JavaApplication10/prueba2.arff");
                    pw = new PrintWriter(fichero);
                    pw.println("@relation prediccionunajma");
                    pw.println("");

                     for (int i=0; i<palabra.length; i++)
                    {
                    {
                       pw.println("@ATTRIBUTE  " + palabra[i]);
                    }
                    pw.println("@DATA");

                    for(int j=0; j<procesar.size(); j++)
                    {
                      pw.print(procesar.get(j)+",");
                    }
                    pw.print("?");

                } catch (Exception e) {
                    e.printStackTrace();
                } finally {
                    try {
                    // asegurar que se cierra el fichero.
                        if (null != fichero)
                        fichero.close();
                    } catch (Exception e2) {
                     e2.printStackTrace();
                    }
                }

        }
}
```

16. Matriz de consistencia

<table>
<tr><th>Problemas</th><th>Objetivos</th><th>Hipótesis</th><th>Variables</th><th colspan="2">Dimensión/indicador</th><th>Metodología</th></tr>
<tr><td><u>Problema Principal</u>
¿Cuál es la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos?</td><td><u>Objetivo General</u>
Evaluar la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.</td><td><u>Hipótesis General</u>
Es posible evaluar la eficiencia de un sistema de rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando Minería de datos.</td><td><u>Independiente</u>
Sistema de Predicción del Rendimiento Académico</td><td colspan="2"></td><td rowspan="5">Método Experimental</td></tr>
<tr><td><u>Problemas Secundarios</u></td><td><u>Objetivos Específicos</u></td><td><u>Hipótesis Especificas</u></td><td><u>Dependiente</u></td><td><u>Dimensión:</u></td><td><u>Indicador</u></td></tr>
<tr><td>¿Cuál es la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos?</td><td>Determinar la eficiencia en la elección del algoritmo más óptimo y que genere el mejor modelo para predecir el rendimiento de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.</td><td>Es posible determinar el algoritmo más óptimo y que genere el mejor modelo predictivo para predecir el rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.</td><td rowspan="3">Rendimiento Académico de los estudiantes</td><td>Demográficos</td><td>Edad
Genero
Estado civil
Situación Laboral</td></tr>
<tr><td>¿Cuál es la eficiencia de un sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos?</td><td>Medir el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando minería de datos.</td><td>Es posible medir el error predictivo generado por el sistema de predicción del rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas de UNAJMA utilizando Minería de datos, con un grado de error mínimo 15%.</td><td>Nivel Socioeconómico</td><td>Ingreso Económico Familiar
Nivel educativo de los padres</td></tr>
<tr><td></td><td></td><td></td><td>Rendimiento Académicos</td><td>Calificación final (Promedio Ponderado) al terminar el semestre.</td></tr>
</table>

Fuente: Elaboración propia. Utilizando Windows 8.1.