Benemérita Universidad Autónoma de Puebla

Facultad De Ciencias de la Computación

Tesis

“Predicción del rendimiento académico de los estudiantes de nuevo ingreso de computación usando minería de datos.”

Para obtener el grado de:
Licenciatura en Ingeniería en Tecnologías de la Información

Presenta

Mireya Luna Jiménez.

ASESOR: Dr. Roberto Contreras Juárez

Octubre de 2024. Puebla, México

Dedicatoria

Dedico este trabajo de tesis a mi familia. A mis hermanos Maricruz y Osvaldo, mis padres Raymundo Luna Miranda y Carolina Jiménez Solís motor y motivo para alcanzar mis metas, todo lo que hago es por ustedes.

Contenido

RESUMEN

Las predicciones del rendimiento académico analizan varios factores como hábitos de estudio, circunstancias personales entre otros que pueden influir en el éxito académico de un estudiante. Usando datos históricos y patrones de comportamiento se pueden identificar áreas de mejora.

El presente trabajo muestra la Predicción del rendimiento académico de los estudiantes de nuevo ingreso de computación usando minería de datos como algoritmos de árboles de decisión J48, BAYES NET, JRip entre otros con el software WEKA.

PALABRAS CLAVES

Minería de datos, algoritmos de clasificación, WEKA, aprendizaje automático, Rendimiento académico.

CAPÍTULO 1

INTRODUCCIÓN

Una de las cosas más importantes y valiosas a la que un individuo puede aspirar es la adquisición de conocimiento. Para adquirir el conocimiento, un sujeto debe incorporarse desde temprana edad a instituciones educativas donde logrará desarrollar habilidades y destrezas que resultan fundamentales para la adquisición de nuevos conocimientos.

De acuerdo con Kabakchieva, D. (2013), las universidades modernas operan en un entorno muy complejo y altamente competitivo y su principal desafío es analizar profundamente su desempeño, para identificar su naturaleza y construir estrategias para su desarrollo y acciones futuras. Además, asegura que la gestión de las universidades debería centrarse más en el perfil de los estudiantes admitidos, tomando conciencia de los diferentes tipos y de las características específicas de los estudiantes en función de los datos recibidos. También deberían considerar si disponen de todos los datos necesarios para analizar a los estudiantes en el punto de entrada de la universidad o si necesitan otros datos para ayudar a los gestores a apoyar sus decisiones sobre cómo organizar la campaña de marketing y acercarse a los estudiantes potenciales prometedores.

Como es sabido, el ingreso de nuevos estudiantes a la Benemérita Universidad Autónoma de Puebla está limitado por la infraestructura física y académica, lo que lleva a la institución a tener que aplicar un instrumento de selección de aspirantes.

Hasta el día de hoy, este instrumento de selección solo ha sido utilizado para determinar que estudiante ingresa y quien no. Estamos convencidos que la información que se obtiene en el examen de admisión podría permitir predecir el rendimiento académico de los estudiantes y por ende la probabilidad de que un estudiante finalice la carrera. Predecir el rendimiento académico de los estudiantes de nuevo ingreso dará la oportunidad de mejorar y ayudar a los estudiantes en la toma de decisiones correctas.
Obtener estas predicciones con los estudiantes de nuevo ingreso de la Facultad de Ciencias de la Computación, podrían ser utilizadas ello mismos como una razón o advertencia para

empezar a mejorar, además de prever el resultado de su sesión actual y dar un enfoque extra a los que tienen más probabilidades de fracasar y rediseñar o mejorar los cursos en los que hay un número significativo de fracasos previstos como se afirma en (Ravinder A. and Yash K., 2017).

Para obtener tales predicciones, las nuevas tecnologías educativas ofrecen a los investigadores oportunidades únicas para analizar cómo los estudiantes aprenden y qué enfoques de aprendizaje conducen al éxito. Se pueden utilizar los datos académicos de los estudiantes obtenidos en el examen de admisión junto con los datos de sus antecedentes socioeconómicos, familia, residencia, división del tiempo de varios hábitos/tareas diarias como estudio/viaje/tiempo libre, etc. Los algoritmos de minería de datos serán muy útiles para estas predicciones, ya que pueden extraer, clasificar, agrupar y agrupar información de los registros de los estudiantes (Xindong Wu, et al., 2015).

Uno de los Objetivos de esta tesis es describir la metodología para la ejecución del proyecto de extracción de datos de los estudiantes de nuevo ingreso de la Facultad de Ciencias de la Computación y presentar los resultados de un estudio destinado a analizar el rendimiento de diferentes algoritmos de clasificación de la minería de datos sobre el conjunto de datos proporcionado a fin de evaluar su posible utilidad para el cumplimiento de la meta y los objetivos del proyecto. Para analizar los datos, utilizaremos algoritmos conocidos de minería de datos contenidos en el software WEKA. La idea de utilizar WEKA es porque está a disposición del público de forma gratuita y es utilizado ampliamente para la investigación en el campo de la minería de datos.

1.1 Antecedentes del Proyecto.

Las instituciones de educación superior consideran crucial el rendimiento académico de los estudiantes, ya que esto les permite egresar excelentes profesionistas al mundo laboral. Por esta razón, las instituciones educativas se enfocan en implementar mejores programas educativos, así como excelentes actividades extracurriculares. Es por esta razón que el objetivo principal de cualquier institución educativa es potenciar la capacidad de aprendizaje del estudiante y fortalecer sus habilidades y destrezas.

Un buen aprendizaje y rendimiento académico dependen, en gran medida, de la dedicación que el estudiante preste a las tareas escolares durante el tiempo que está en su hogar, aunque, al contrario de lo que pudiera pensarse, no es tan importante el tiempo que se invierte en el estudio como la calidad de este (Torres y Rodríguez, 2006).

Los padres son el agente socializador fundamental y, desde el comienzo de la vida, se comunican con los alumnos, transmitiéndoles su nivel cultural por medio del lenguaje y la relación afectiva. En la mayor parte de las investigaciones efectuadas al respecto, el nivel de educación formal de las madres tiene una enorme potencialidad explicativa en el desarrollo del estudiante (Mella y Ortiz, 1999).

El rendimiento escolar, incluyendo aspectos tales como el nivel de logro alcanzado en materias específicas, tasas de repetición y de retención escolar, han sido analizado tomando en cuenta dos conjuntos de causas: aquellos aspectos relacionados con la escuela como sistema educativo, y las características que los alumnos exhiben a partir de su contexto social, de sus capacidades personales, de sus motivaciones (Mella y Ortiz, 1999).

En la explicación del rendimiento escolar, lo más importante son las características de los propios estudiantes, sus capacidades, vocación, experiencias previas, esfuerzo y disposición a aprender, sin embargo, las instituciones deben

ofrecer oportunidades y ambientes formativos, en términos de su calidad y pertinencia para propiciar el desempeño de los estudiantes (Aldana, et al., 2010).

En su búsqueda por mejorar la calidad de la educación y con la finalidad de responder a su compromiso social, las escuelas han prestado atención al quehacer académico de sus estudiantes; dentro de algunas instituciones de educación superior se están llevando a cabo diversas acciones para apoyar la mejora del rendimiento académico para evitar la deserción de los estudiantes (Aldana, et al., 2010).

La implementación de métodos y herramientas de la minería de datos para el análisis de datos obtenidos en las instituciones educativas conocido como Minería de Datos Educativos (EDM, por sus siglas en inglés), es un campo relativamente nuevo en las investigaciones de la minería de datos (Romero and Ventura, 2007).

Como es bien sabido, el rendimiento de los estudiantes y el desarrollo de sus habilidades depende mucho de su capacidad de aprendizaje, la cual está estrechamente relacionada con los métodos de enseñanza de los docentes y muchos otros factores, mediante los resultados que se obtengan a través del uso de técnicas EDM, las diferentes academias de la Facultad de Ciencias de la Computación podrían desarrollar estrategias que ayuden a los docentes a tratar eficazmente los problemas que enfrentan los estudiantes en su vida académica, mejorando sus destrezas para enfrentar con éxito su vida profesional.

1.2 Objetivos Generales Y Específicos Del Proyecto.

Objetivo General.

Aplicar técnicas y métodos de minería de datos para predecir el rendimiento académico de los estudiantes de nuevo ingreso de la Facultad de Ciencias de la Computación a partir de los puntajes obtenidos en cada área que conforman el examen de admisión.

Objetivos Específicos

Describir la metodología para la extracción de datos de los estudiantes de nuevo ingreso de la Facultad de Ciencias de la Computación.
Analizar que algoritmos de minería de datos contenidos en el software WEKA se prestan mejor a nuestros intereses.
Revelar el gran potencial de las aplicaciones de la minería de datos para la gestión universitaria.
Presentar los resultados del estudio destinado a analizar el rendimiento de diferentes algoritmos de clasificación de la minería de datos.

1.3 Metodología.

En esta tesis se utiliza técnicas y métodos de Minería de Datos, por lo que se implementará siguiendo el modelo Cross-Industry Standard Process for Data Mining (CRISP-DM) descrito en Chapman, P., et al. (2000).

El CRISP-DM se elige como método de investigación porque está disponible libremente y es de aplicación estándar-neutral para proyectos de minería de datos que ha sido utilizado ampliamente por investigadores en el campo durante los últimos diez años. Además, es un método cíclico, que incluye seis fases principales - *comprensión de los negocios, comprensión de los datos, preparación de los datos, modelización, evaluación y despliegue*. Existe una serie de bucles de retroalimentación interna entre las fases, derivado de la muy compleja naturaleza no lineal del proceso de minería de

datos y de asegurar el logro de resultados consistentes y confiables.

El software que será utilizado para la implementación de este proyecto de tesis es el software de código abierto WEKA, el cual ofrece un amplio rango de métodos de clasificación para minería de datos (Witten, I. and E. Frank., 2005).

Durante la fase de *comprensión de los negocios,* se realizará una extensa revisión de la literatura con la finalidad de estudiar problemas semejantes que hayan sido resueltos mediante la aplicación de la minería de datos. El principal objetivo de este proyecto es predecir el rendimiento académico de los estudiantes de la facultad de ciencias de la computación a partir de los resultados obtenidos en cada una de las áreas que componen el examen de admisión que aplicaron al ingresar, en la minería de datos se considera un problema de clasificación que debe resolverse utilizando los datos disponibles de los estudiantes. Esta es una tarea para el aprendizaje supervisado, ya que los modelos de clasificación se construyen a partir de datos en los que se conoce la variable objetivo (o respuesta).

En la fase de *comprensión de datos,* se analizará todo el proceso de inscripción a la Universidad, incluyendo los procedimientos formales y documentos de solicitud, con el fin de identificar los tipos de datos recogidos de cada uno de los solicitantes. Todos los datos relacionados con el proceso de admisión a la universidad se almacenan en la base de datos de la Dirección de Administración Escolar de la Universidad, incluidos los datos personales de los aspirantes (número de solicitud, nombres, direcciones, promedio del nivel medio superior, promedio académico de la carrera, el desempeño en las áreas que componen el examen de admisión (puntajes), etc.

En la fase de *preprocesamiento de datos*, se extraerán y organizarán los datos de los estudiantes. Los datos proporcionados están sujetos a muchas transformaciones. Algunos de los parámetros que contienen datos que no son de interés en la investigación serán removidos, por ejemplo, el campo del país contiene sólo un valor – México, o el tipo de educación previa que también tiene un solo valor, porque sólo concierne a los estudiantes con la educación secundaria. Otras de las variables que

contienen datos importantes se procesarán y convertirán en variables nominales con un número limitado de valores distintos.

Finalmente, durante la fase de *modelación*, se seleccionarán los métodos para construir un modelo para clasificar a los estudiantes que permita predecir el rendimiento académico. Los clasificadores populares de WEKA (con su configuración por defecto a menos que se especifique lo contrario) que utilizaremos en el estudio experimental incluyen un algoritmo de árbol de decisión común C4.5 (J48), dos clasificadores bayesianos (NaiveBayes y BayesNet), un algoritmo del vecino (IBk) y dos aprendices de reglas (OneR y JRip).

1.4 Rendimiento Académico

El desempeño del estudiante depende de su capacidad de aprendizaje y está influenciado por muchos factores.
La predicción del desempeño del estudiante es una tarea desafiante ya que depende de muchos factores como las calificaciones, el desempeño en clase, los datos demográficos y las características emocionales.
El desempeño del estudiante no solo depende de lo académico, sino que también depende de otras actividades personales, sociales y extracurriculares. (Amjad Abu Saa, 2016).

El trabajo de investigación realizado por Ali Daud y Farhat Abbas introduce el proceso de predicción escolar del estudiante que utiliza cuatro tipos de atributos: gasto familiar, ingreso familiar, información personal del estudiante y patrimonio familiar. Concluyeron que el gasto familiar y los atributos de información personal tienen un efecto crucial en el desempeño del estudiante debido a razones instintivas.
Blanco, L (1989), el tema de rendimiento académico o el fracaso escolar universitario hay que estudiarlo dentro de un contexto sociocultural-económicopolítico a la vez que familiar, personal y académico, ya que los factores que influyen en el mismo son numerosos y se encuentran muy interrelacionados. Exigirá, por tanto, un tratamiento interdisciplinar con una metodología común y análisis multivariados.

Según Solano, L. (2015), El estudio del rendimiento escolar constituye, hoy en día, un tema destacado en la investigación educativa. Y es que, para nuestra sociedad actual, caracterizada por el bombardeo continuado de información desde distintos medios, el gran desafío de la educación es transformar esa gran cantidad de información en conocimiento personal valido para poder desenvolverse con eficacia en la vida. De ahí que lograr el éxito o el fracaso en los estudios es de vital importancia para el futuro profesional individual.

El principal desafío para las universidades modernas es analizar profundamente su desempeño, identificar su singularidad y construir una estrategia para un mayor desarrollo y acciones futuras. La gestión universitaria debería centrarse más en el perfil de los estudiantes admitidos, conociendo los diferentes tipos y las características específicas de los estudiantes en función de los datos recibidos. También deben considerar si tienen todos los datos necesarios para analizar a los estudiantes en el punto de entrada de la universidad o si necesitan otros datos para ayudar a los gerentes a respaldar sus decisiones sobre cómo organizar la campaña de marketing y acercarse a los prometedores estudiantes potenciales.

La implementación de modelos predictivos para maximizar el reclutamiento y la retención de estudiantes se presenta en el estudio. Estos problemas también son discutidos por De Long et al. El desarrollo de modelos de predicción de matrícula basados en datos de admisión de estudiantes mediante la aplicación de diferentes métodos de minería de datos es el foco de investigación de N andeshwar y C haudhari. D ekker.

El modelado del desempeño de los estudiantes en varios niveles y la comparación de diferentes algoritmos de minería de datos se analizan en muchos trabajos de investigación publicados recientemente.utiliza técnicas de minería de datos para explorar las variables sociodemográficas (edad, género, etnia, educación, estado laboral y discapacidad) y el entorno de estudio (programa del curso y curso bloque) que pueden influir en la persistencia o deserción de los estudiantes, identificando los factores más importantes para el éxito de los estudiantes y desarrollando un perfil de los típicos estudiantes exitosos y no exitosos.

Como resultado y conclusiones se revela que las tasas de predicción no son notables (varían entre 52-67%). Además, los clasificadores funcionan de manera

diferente para las diversas clases. Los atributos de datos relacionados con el puntaje de admisión a la universidad de los estudiantes y el número de fallas en los exámenes universitarios de primer año se encuentran entre los factores que más influyen en el proceso de clasificación.

La investigación de Amjad Abu Saa,2016, concluyó que el desempeño del estudiante no solo depende de lo académico, sino que también depende de otras actividades personales, sociales y extracurriculares. Junto con el algoritmo Naïve Bayes utilizó tres algoritmos de árbol de decisión para la clasificación de datos. En primer lugar, hizo una encuesta y recopiló datos de los estudiantes y luego preprocesó y exploró los datos para tareas de minería de datos. En segundo lugar, los algoritmos de minería de datos se implementaron en el conjunto de datos para generar modelos de clasificación para predecir el desempeño de los estudiantes.

CAPÍTULO 2
MINERÍA DE DATOS Y LA PLATAFORMA WEKA

2.1 Minería de Datos

La minería de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.

La minería de datos surgió con la intención o el objetivo de ayudar a comprender una enorme cantidad de datos, y que estos, pudieran ser utilizados para extraer conclusiones para contribuir en la mejora y crecimiento.

El proceso de hurgar en los datos para descubrir conexiones ocultas y predecir tendencias futuras tiene una larga historia. Conocido algunas veces como "descubrimiento de conocimientos en bases de datos", el término "minería de datos" no se acuño sino hasta la década de 1990. Pero su base comprende tres disciplinas científicas entrelazadas: estadística (el estudio numérico de relaciones de datos), inteligencia artificial (inteligencia similar a la humana exhibida por software y/o máquinas) machine learning (algoritmos que pueden aprender de datos para hacer predicciones). Lo que era antiguo es nuevo otra vez, ya que la minería de datos continúa evolucionando para igualar el ritmo del potencial sin límites del big data y poder de cómputo asequible.

Principales características y objetivos

- Explorar los datos que se encuentran en las profundidades de las bases de datos (por ejemplo los Almacenes de Datos), que algunas veces contienen información almacenada durante varios años.

- En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet.

- El entorno de la minería de datos suele tener una Arquitectura Cliente Servidor

- Las herramientas de la minería de datos ayudan a extraer el mineral de la información registrado en archivos corporativos o en registros públicos, archivados.

- El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias, para efectuar preguntas ad-hoc y obtener rápidamente respuestas.

- Hurgar y sacudir a menudo implica el descubrimiento de resultados valiosos e inesperados.

- Las herramientas de la minería de datos se combinan fácilmente y pueden analizarse y procesarse rápidamente.

- Debido a la gran cantidad de datos, algunas veces resulta necesario usar procesamiento en paralelo para la minería de datos.

- La minería de datos produce cinco tipos de información:

    1. Asociaciones.
    2. Secuencias.
    3. Clasificaciones.
    4. Agrupamientos.
    5. Pronósticos.

- Los mineros de datos usan varias herramientas y técnicas.

La minería de datos es un proceso que invierte la dinámica del método científico en el siguiente sentido:

- En el método científico, primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten la hipótesis. Si

esto se hace con la formalidad adecuada (cuidando cuáles son las variables controladas y cuáles experimentales), se obtiene un nuevo conocimiento.

- En la minería de datos, se coleccionan los datos y se espera que de ellos emerjan hipótesis. Luego entonces, se valida esa hipótesis inspirada por los datos en los datos mismos, será numéricamente significativa, pero experimentalmente inválida. De ahí que la minería de datos debe presentar un enfoque exploratorio, y no confirmador. Usar la minería de datos para confirmar las hipótesis formuladas puede ser peligroso, pues se está haciendo una inferencia poco válida.

2.2 Machine Learning

Kelleher, J. & Mac, B. & D'Arcy, A. (2017), El aprendizaje automático es una ciencia que le permite a las maquinas desarrollar técnicas para poder aprender. Este aprendizaje es un proceso automatizado que extrae patrones de los datos para la construcción de modelos que permitan realizar la predicción utilizando algoritmos supervisados, relacionando características descriptivas actuales con características de destino basadas en un conjunto de ejemplos históricos o instancias.

Ali Reza Samanpour, André Ruegenberg & Robin Ahlers (2017), El aprendizaje automático ofrece predicciones automatizadas y precisas a partir de información densa y desordenada y la convierte en un formato útil para los humanos. Dependiendo del propósito de un algoritmo o sistema de aprendizaje automático, estos modelos recomiendan acciones futuras basadas en los llamados valores empíricos o probabilidades.

Moreno, et al (1994), el aprendizaje se refiere a un amplio espectro de situaciones en las cuales el aprendiz incrementa su conocimiento o sus habilidades para cumplir una tarea. El aprendizaje aplica inferencias a determinada información para construir una representación apropiada de algún aspecto relevante de la realidad o de algún proceso

Ali Reza Samanpour, André Ruegenberg & Robin Ahlers (2017), *La minería de datos se* refiere a un proceso de reconocimiento de patrones en datos estructurados existentes. Este proceso puede realizarse automáticamente con la ayuda del aprendizaje automático o semiautomáticamente aplicando métodos estadísticos.

2.3 Tipos de aprendizaje en machine learning

Zambrano, J. (2018) el machine learning desarrolla algoritmos que hacen que las máquinas puedan aprender por su cuenta y responder a determinadas preguntas con bastante certeza. Para desarrollar estos algoritmos, existen dos modalidades: aprendizaje supervisado y no supervisado.

Aprendizaje supervisado: La primera modalidad de aprendizaje que tiene el machine learning es la de aprendizaje supervisado. Usándola, se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas.

Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones. Existen, a su vez, dos tipos de aprendizaje supervisado:

Regresión: Es uno de los algoritmos del aprendizaje supervisado, el cual es utilizado en aprendizaje automático y en la estadística, el cual consiste en dibujar una recta la cual indicará la tendencia de un grupo de datos continuos (números), y si fuesen discretos (cadenas de texto), se utilizaría regresión logística.

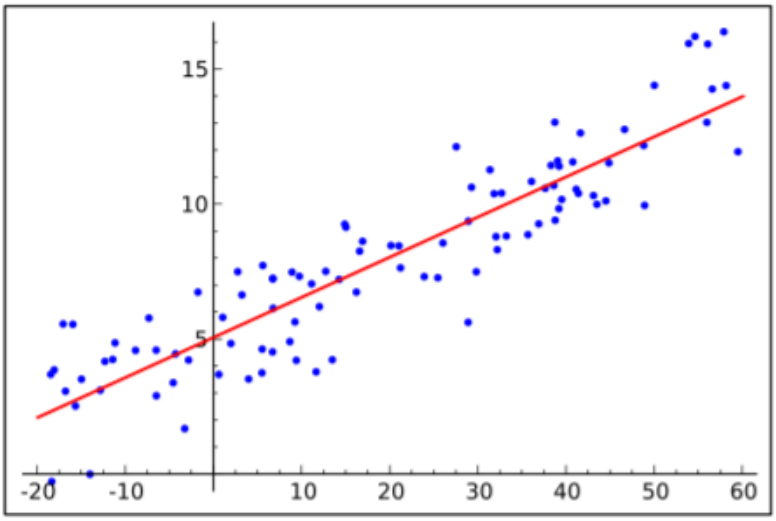


Imagen 1 – Aprendizaje automático Regresión

Fuente: Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado?

Clasificación: Los algoritmos de clasificación como su nombre lo indica busca o encuentra patrones que luego le permitirá clasificar a los elementos y determinar a qué grupos o clases pertenecen, se debe mencionar que los valores para estos algoritmos deben ser valores discretos.

Aprendizaje no supervisado: Los algoritmos no supervisados funcionan sin ningún entrenamiento adecuado. Funciona tan pronto como recibe los datos. El algoritmo toma sus propias decisiones y encuentra maneras de clasificar las variables y comprobar si encajan. Calvo, D. (2017) describe al Aprendizaje No Supervisado, como el conjunto de técnicas que permiten inferir modelos para extraer conocimiento de conjuntos de datos donde a priori se desconoce.

Para Gonzalo, A (2018) Se llama no supervisado porque, contrariamente al supervisado, tiende a ser más subjetivo ya que no tiene respuestas correctas. Los algoritmos sirven para descubrir y presentar estructuras interesantes en los datos. El objetivo del aprendizaje no supervisado es modelizar la estructura o distribución de los datos para aprender más sobre ellos. Sirve tanto para entender como para resumir un conjunto de datos.

Aprendizaje reforzado: El algoritmo ejecuta las acciones que serán las más recompensadas. Con frecuencia se usa en IA para juegos, o en robots de navegación.

Clasificador de árbol de decisión

Los árboles de decisión son herramientas poderosas y populares para la clasificación. Un árbol de decisión es una estructura similar a un árbol, que comienza con los atributos raíz y termina con nodos hoja. Generalmente, un árbol de decisión tiene varias ramas que constan de diferentes atributos, el nodo hoja en cada rama representa una clase o un tipo de distribución de clases. Los algoritmos del árbol de decisiones describen la relación entre los atributos y la importancia relativa de los atributos. Las ventajas de los árboles de decisión son que representan reglas que los usuarios pueden entender e interpretar fácilmente, no requieren una preparación compleja de datos y funcionan bien para variables numéricas y categóricas.

Clasificadores bayesianos

Los clasificadores bayesianos son clasificadores estadísticos que predicen la pertenencia a una clase mediante probabilidades, como la probabilidad de que una muestra determinada pertenezca a una clase en particular. Se han desarrollado varios algoritmos de Bayes, entre los cuales las redes bayesianas y los ingenuos Bayes son los dos métodos fundamentales. Los algoritmos Naive Bayes asumen que el efecto que juega un atributo en una clase dada es independiente de los valores de otros atributos.

El clasificador k-vecino más cercano

El algoritmo k-Vecino más cercano (k-NN) es un método para clasificar objetos según los ejemplos de entrenamiento más cercanos en el espacio de características. k-NN es un tipo de aprendizaje basado en instancias, o aprendizaje diferido, donde la función solo se aproxima localmente y todos los cálculos se aplazan hasta la clasificación. El algoritmo k-NN se encuentra entre los más simples de todos los algoritmos de aprendizaje automático: un objeto se clasifica por el voto mayoritario de sus vecinos, y el objeto se asigna a la clase más común entre sus k

vecinos más cercanos (k es un número entero positivo, normalmente pequeña). La mejor opción de k depende de los datos; en general, los valores más altos de k reducen el efecto del ruido en la clasificación, pero hacen que los límites entre clases sean menos distintos. La precisión del algoritmo k-NN puede verse seriamente degradada por la presencia de características ruidosas o irrelevantes.

2.4 ¿QUÉ ES WEKA?

Es una plataforma de software para el aprendizaje automático y la minería de datos escrito en java y desarrollado en la Universidad de Waikato.
Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelo predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.
Sus características más importantes son:

Weka soporta varias tareas estándar de minería de datos especialmente, preprocesamiento de datos, clustering, clasificación, regresión, visualización, y selección.
Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un fichero plano (*flat file*) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (normalmente numéricos o nominales, aunque también se soportan otros tipos).

Weka también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (*Java Database Connectivity*) y puede procesar el resultado devuelto por una consulta hecha a la base de datos.

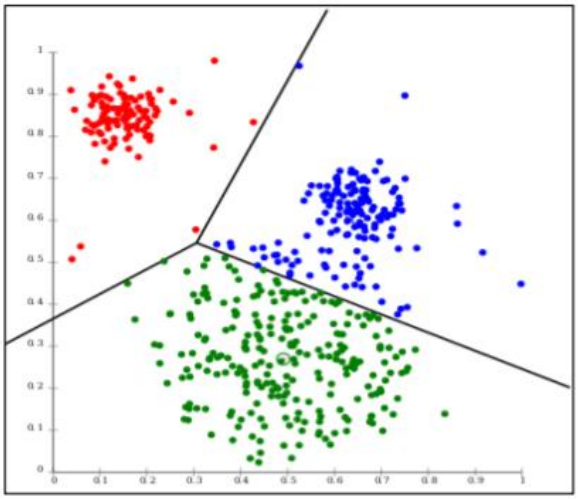


Imagen 2 – Aprendizaje automático Clasificación

Fuente: Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado?

El algoritmo no está en capacidad de determinar a qué grupo pertenece un valor o cuál es el resultado de una operación. Solamente logra relacionar características con etiquetas y así obtener un resultado.

2.4.1 ¿Cómo funciona WEKA?

Al ejecutar la aplicación de WEKA 3.8.6 muestra la primera pantalla de **selector de interfaz de Weka** que da la opción de seleccionar entre cinco posibles interfaces de usuario para acceder a las funcionalidades del programa, y son las siguientes: EXPLORER, EXPERIMENTER, KNOWLEDGEFLOW, WORKBENCH, SIMPLE CLI.

Imagen 3 – Pantalla de Inicio de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

El Explorer permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos. Cada una de las tareas de minería de datos viene representada por una pestaña en la parte superior. Estas son:

• Preprocess: visualización y preprocesado de los datos (aplicación de filtros)
• Classify: Aplicación de algoritmos de clasificación y regresión.
• Cluster: Agrupación
• Associate: Asociación
• Select Attributes: Selección de atributos
• Visualize: Visualización de los datos por parejas de atributos.

Y muestra la siguiente pantalla.

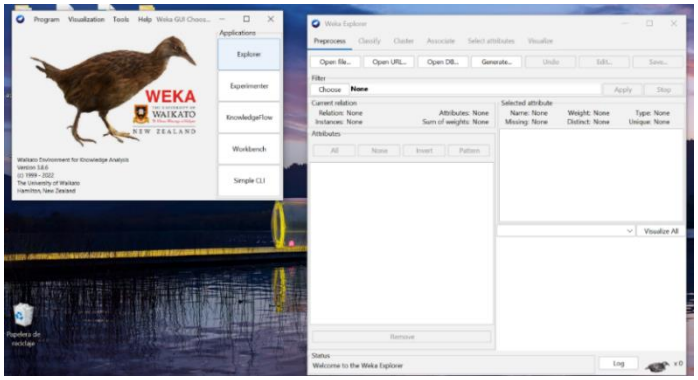
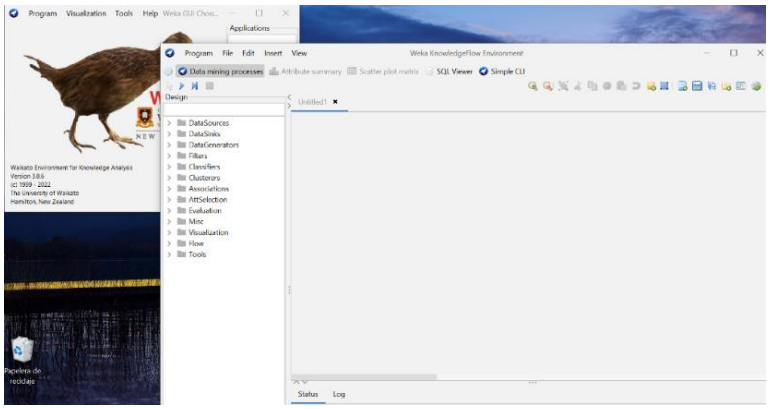


Imagen 4 – Pantalla de Interfaz Explorer de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

La interfaz **Experimenter** («experimentador») permite la comparación sistemática de una ejecución de los algoritmos predictivos de Weka sobre una colección de conjuntos de datos.

Imagen 5 – Pantalla de Interfaz KnowledgeFlow de Weka 3.8.6



Fuente: Propia, Imagen capturada de Weka 3.8.6

Knowledge Flow («flujo de conocimiento») es una interfaz que en esencia implementa las mismas funciones que Explorer, y además permite "*arrastrar y soltar*".

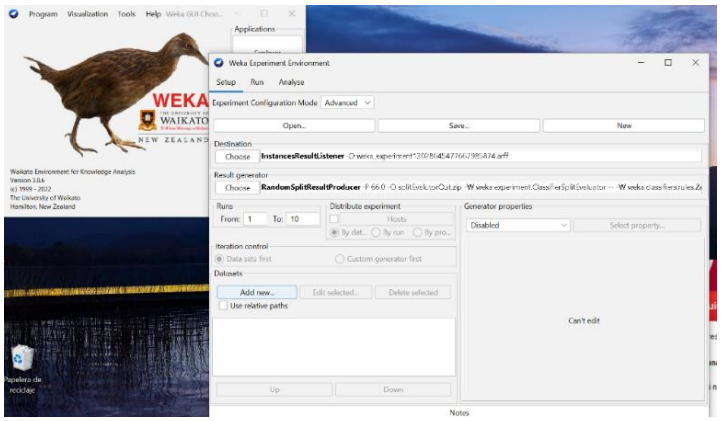


Imagen 6 – Pantalla de Interfaz Experimenter de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

También puede ofrecer aprendizaje incremental Workbench

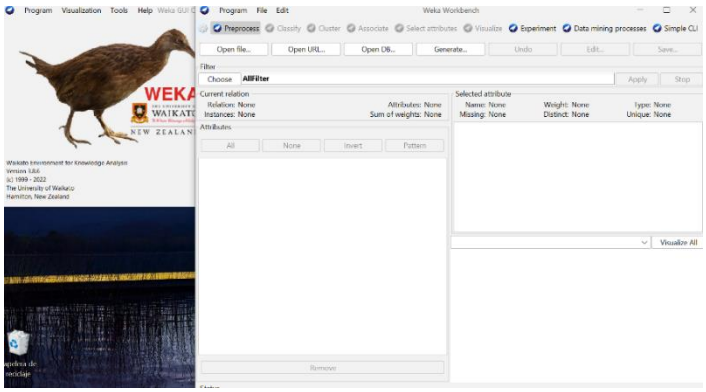


Imagen 7 – Pantalla de Interfaz Workbench de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Simple CLI, Se trata de una consola que permite acceder a todas las opciones de Weka desde Línea de comandos.

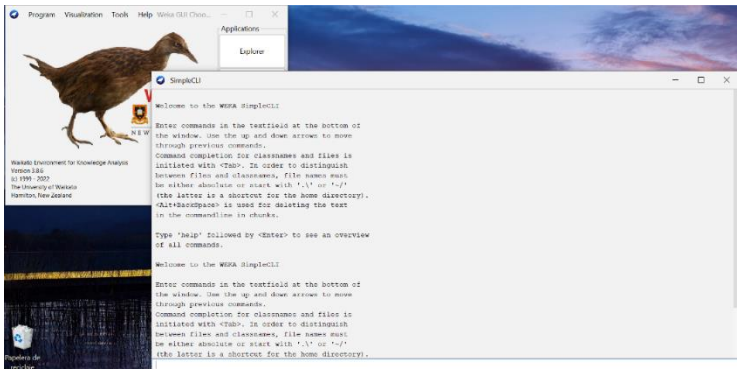


Imagen 8 – Pantalla de Interfaz Simple CLI de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Alternativas y competidores de Weka

Machine-learning in Python

Scikit-learn

XGBoost

	Gratuito	Algoritmos para análisis de datos	Fácil de modificar en código
WEKA	X	X	X
Machine-learning in Python		X	
Scikit-learn		X	
XGBoost		X	

Tabla 1 – Comparación de Weka con software semejantes

Fuente: Propia.

CAPÍTULO 3

METODOLOGÍA Y ANÁLISIS

3.1 Introducción

En esta tesis se utiliza técnicas y métodos de Minería de Datos, por lo que se implementará siguiendo el modelo Cross-Industry Standard Process for Data Mining (CRISP-DM) descrito en Chapman, P., et al. (2000), descrito en la metodología, en sus diferentes fases que se describen a continuación.

FASE I: **Comprensión del negocio**

En esta fase se realizará una extensa revisión de la literatura con la finalidad de estudiar problemas semejantes que hayan sido resueltos mediante la aplicación de la minería de datos.

Contexto de la Universidad

Las universidades como instituciones educativas tienen como propósito fundamental garantizar la calidad de la educación de sus estudiantes en sus diferentes programas, los cuales generalmente están expresados en las notas que estos obtienen, pero muchas veces en el intento de lograr lo anterior no se logra, por falta de herramientas que permitan monitorear o apoyar al rendimiento académico de los principales actores de las Universidades (los estudiantes).

Objetivos del Proyecto

El principal objetivo de este proyecto es predecir el rendimiento académico de los estudiantes de la facultad de ciencias de la computación a partir de los resultados obtenidos en cada una de las áreas que componen el examen de admisión que aplicaron al ingresar, en la minería de datos se considera un problema de clasificación que debe resolverse utilizando los datos disponibles de los estudiantes. Esta es una tarea para el aprendizaje supervisado, ya que los modelos de clasificación se construyen a partir de datos en los que se conoce la variable objetivo (o respuesta).

Con el presente trabajo buscaremos predecir el rendimiento académico de los estudiantes del primer semestre a partir de sus datos de ingreso o admisión a la BUAP, para que las autoridades correspondientes puedan tomar las acciones correspondientes para mejorar el éxito de los estudiantes en su rendimiento académico y evitar o eliminar el fracaso de estos.

Para determinar un modelo que permita predecir el rendimiento académico de los estudiantes del primer semestre, se obtuvo el puntaje de ingreso de los estudiantes del periodo 2017 así como la información del promedio del primer semestre, proporcionado por la Dirección de Administración Escolar de la Benemérita Universidad Autónoma de Puebla.

Evaluación de la situación

Desde una perspectiva institucional

Considerando que la Facultad de Ciencias de la Computación, BUAP, es una entidad de formación profesional de alto nivel académico y los procesos de licenciamiento y acreditación de las Escuelas profesionales, sería importante contar con una herramienta que les permita a las autoridades competentes, tomar decisiones respecto de los estudiantes que podrían fracasar en su rendimiento académico en el primer semestre de estancia en la institución.

Desde una perspectiva del aprendizaje automático y minería de datos

El aprendizaje automático se clasificaron las técnicas de algoritmos supervisados y no supervisados, y dentro de los supervisados a su vez las técnicas de regresión y de clasificación, es precisamente en las técnicas de clasificación donde utilizamos los algoritmos para determinar el de mejor desempeño, la estadística descriptiva también nos ayuda a la valoración de los resultados encontrados en la etapa inicial. Y respecto de la metodología utilizada para los resultados del presente trabajo se utilizó CRISP-DM, por la flexibilidad que ofrece nos ayudara a trabajar de manera ordenada.

Determinación de los Objetivos de Minería de Datos aplicados al proyecto

Los objetivos de la minería de datos aplicados al proyecto de aprendizaje automático de la presente tesis, se resumen en los siguientes pasos:

- Realizar la limpieza de los datos proporcionada por el centro de computo
- Determinar los factores más importantes que afectan al rendimiento académico
- Analizar y determinar el mejor algoritmo de aprendizaje automático para desarrollar la predicción
- Predecir el rendimiento académico con el algoritmo más óptimo de predicción. El criterio de la predicción se basa primero en el análisis estadístico de los datos para determinar los factores que influyen en el rendimiento académico y segundo la utilización de la herramienta WEKA para determinar la performance de los algoritmos y la selección del mejor predictor para realizar la predicción.

El criterio de la predicción se basa primero en el análisis estadístico de los datos para determinar los factores que influyen en el rendimiento académico y segundo la utilización de la herramienta WEKA para determinar la performance de los algoritmos y la selección del mejor predictor para realizar la predicción.

FASE II: **Comprensión de los datos**

En esta fase se enfoca en la comprensión del contexto para eso se analizará todo el proceso de inscripción a la Universidad.

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta. En la labor de investigación de antecedentes al presente, se pudo mostrar que la mayoría de autores coinciden que existen 3 factores que afectan al rendimiento académico de los estudiantes universitarios y en concreto son los siguientes:

- El estudiante
- La institución

- Los docentes

Respecto al factor estudiante es aún más complejo por que intervienen aspectos psicológicos, económicos y emocionales así como también la familia y otros aspectos inherentes a la persona; estos no serán estudiados en este trabajo de investigación, solo se considerarán factores de sus datos de ingreso a la Facultad de Ciencias de la Computación, y sobre las que se realizaron actividades de conocimiento preliminar de la data, para mejorar la consistencia de los mismos (inserción, modificación y eliminación) de datos, utilizando herramienta de Microsoft Excel 2019.

Recopilación de datos iniciales

La información procesada sobre los datos iniciales en bruto, contiene la siguiente información:

1. Matricula
2. FISICA Score
3. GEOMET Score
4. TRIGONO Score
5. PROBA Score
6. FISICA SE
7. GEOMET SE
8. TRIGONO SE
9. PROBA SE
10. Scaled Full Test
11. FISICA Scaled
12. GEOMET Scaled
13. TRIGONO Scaled
14. PROBA Scaled
15. Full Percentile
16. FISICA Percentile
17. GEOMET Percentile
18. TRIGONO Percentile
19. PROBA Percentile

FASE III.- Preparación de los datos

En esta fase de Preparación, se desarrollará todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales.

Las tareas incluirán la selección de atributos y registros, así como la transformación y la limpieza de datos para las herramientas que permitan generar los modelos predictivos.

Selección de los datos más relevantes para determinar los datos más o factores más relevantes que influyen en el rendimiento académico de los estudiantes desarrollaremos el análisis estadístico del dataset, desarrollando una limpieza preliminar de los mismos, para lo cual utilizaremos herramientas estadísticas.

Limpieza de Datos

Para la limpieza de los datos realizaremos un análisis de cada uno de los atributos de nuestra información y lo clasificaremos considerando 3 apreciaciones de importancia para desarrollar el modelo (Alta, Media y Baja),

1. Matricula
   Importancia: Alta
   Justificación: Se requiere para realizar la comparación con los resultados de la predicción y los resultados reales.

2. FISICA Score
   Importancia: Baja
   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.
3. GEOMET Score
   Importancia: Baja
   Justificación: el valor de de este atributo se incluye en el puntaje promedio obtenido.

4. TRIGONO Score

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

5. PROBA Score

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

6. FISICA SE

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

7. GEOMET SE

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

8. TRIGONO SE

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

9. PROBA SE

   Importancia: Baja

   Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido.

10. Scaled Full Test

    Importancia: Alta

    Justificación: el valor de este atributo es el promedio obtenido de todos los puntaje de cada materia.

11. FISICA Scaled

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

12. GEOMET Scaled

    Importancia: Baja

    Justificación: el valor de de este atributo se incluye en el puntaje promedio obtenido

13. TRIGONO Scaled

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

14. PROBA Scaled

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

15. Full Percentile

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

16. FISICA Percentile

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

17. GEOMET Percentile

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

18. TRIGONO Percentile

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

19. PROBA Percentile

    Importancia: Baja

    Justificación: el valor de este atributo se incluye en el puntaje promedio obtenido

Construcción de nuevos datos

Para la construcción de los nuevos datos se agruparon en intervalos quedando de la siguiente manera:

Alto

730 ≤ Scaled Full Test ≤ 100

Medio

587 ≤ Scaled Full Test ≤ 729

Bajo

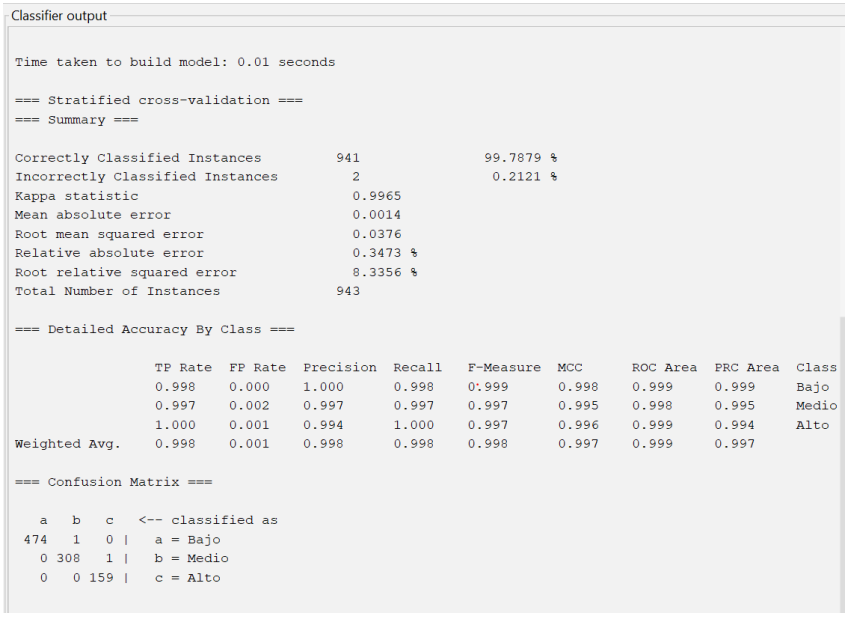
200 ≤ Scaled Full Test ≤ 586

CAPÍTULO 4

RESULTADOS

4.1. Algoritmo Árboles de Decisión J48

Los resultados generados por la herramienta WEKA 3.8.6 para el algoritmo “Arboles de decisión J-48”, se muestran a continuación.



Classifier output

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         941               99.7879 %
Incorrectly Classified Instances         2                0.2121 %
Kappa statistic                          0.9965
Mean absolute error                      0.0014
Root mean squared error                  0.0376
Relative absolute error                  0.3473 %
Root relative squared error              8.3356 %
Total Number of Instances              943

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.000    1.000      0.998    0.999      0.998    0.999     0.999     Bajo
                 0.997    0.002    0.997      0.997    0.997      0.995    0.998     0.995     Medio
                 1.000    0.001    0.994      1.000    0.997      0.996    0.999     0.994     Alto
Weighted Avg.    0.998    0.001    0.998      0.998    0.998      0.997    0.999     0.997

=== Confusion Matrix ===

   a   b   c   <-- classified as
 474   1   0 |   a = Bajo
   0 308   1 |   b = Medio
   0   0 159 |   c = Alto
```

Imagen 9 – Pantalla de Salida algoritmo J48 de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

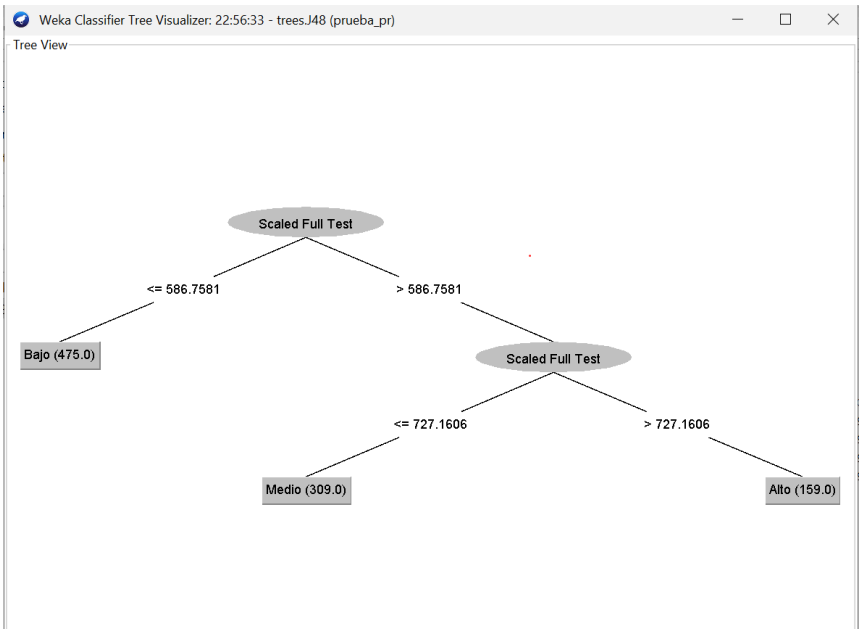


Imagen 10 – Pantalla árbol con algoritmo J48 de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: ARBOLES DE DECISIÓN J-48

Instancias correctamente clasificadas	**941**	**99.7879%**
Instancias incorrectamente clasificadas	**2**	**0.2121%**
Total	**943**	**100%**

Tabla 2 – Resultados de clasificación con algoritmo J48

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	474	1	0	**475**
Medio	0	308	1	**309**
Alto	0	0	159	**159**

Tabla 3 – Matriz de Confusión, algoritmo J48

Fuente: Propia.

Interpretación de resultados con árboles de decisión J-48

Como se puede apreciar clasifico de manera correcta hasta un 99.7879%, y en la matriz de confusión se aprecia que:

De 475 registros 474 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 308 fueron clasificados correctamente como “Medio” y finalmente de 159 registros 159 fueron clasificados correctamente como “Alto”.

4.2. Algoritmo de Vecinos más cercanos

Los resultados generados por la herramienta WEKA 3.8.6 para el algoritmo “Vecinos más

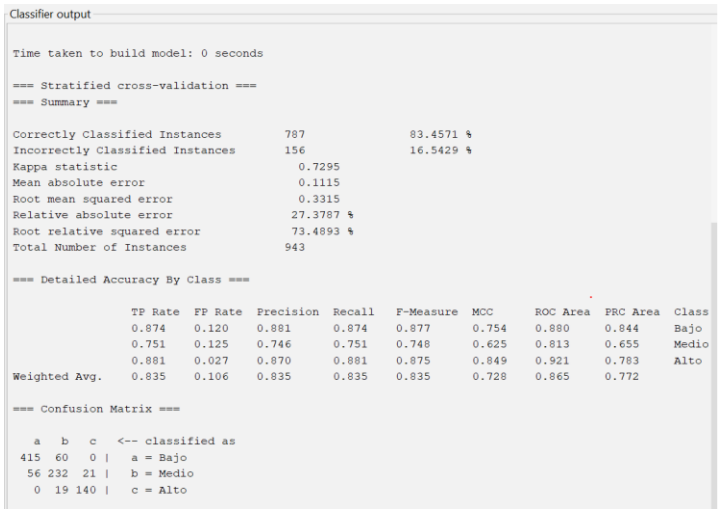


Imagen 11 – Pantalla de salida con algoritmo KNN 1 Vecino de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

cercanos”, se muestra a continuación:

Algoritmo: Vecinos más cercanos

Instancias correctamente clasificadas	**787**	**83.4571%**
Instancias incorrectamente clasificadas	**156**	**16.5429 %**
Total	**943**	**100%**

Tabla 4 – Resultados de clasificación con algoritmo KNN

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	415	60	0	**475**
Medio	56	232	21	**309**

Alto	0	19	140	**159**

Tabla 5 –Matriz de confusión algoritmo KNN

Fuente: Propia.

Interpretación de resultados con Vecinos más cercanos

Como se puede apreciar clasifico de manera correcta hasta un 83.4571%, y en la matriz de confusión se aprecia que:

De 475 registros 415 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 232 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 140 fueron clasificados correctamente como “Alto”

Algoritmo: Vecinos más cercanos con 4 vecinos

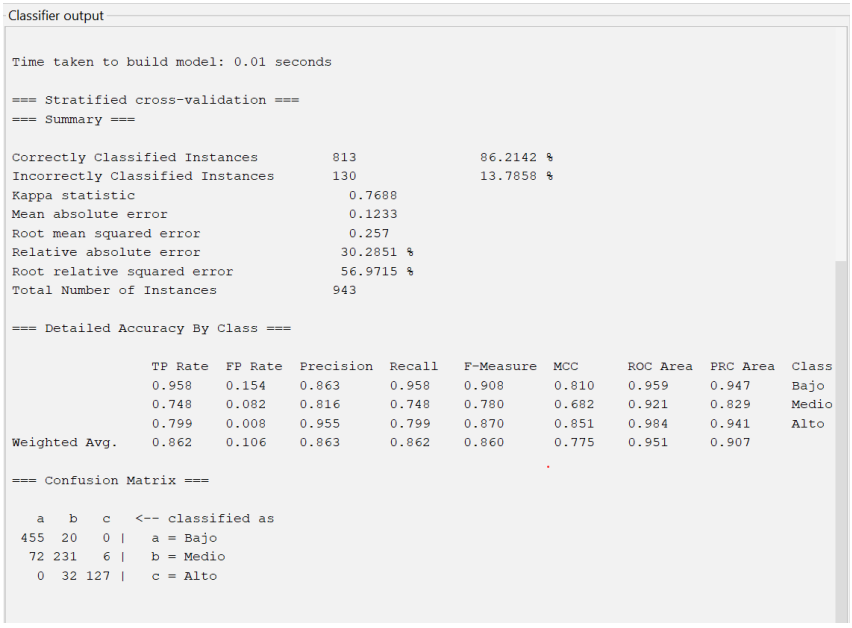


Imagen 12 – Pantalla de salida algoritmo KNN 4 vecinos de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: Vecinos más cercanos

Instancias correctamente clasificadas	**813**	**86.2142%**
Instancias incorrectamente clasificadas	**130**	**13.7858 %**
Total	**943**	**100%**

Tabla 6 – Resultados de clasificación con algoritmo KNN 4 vecinos

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	455	20	0	**475**
Medio	72	231	6	**309**
Alto	0	32	127	**159**

Tabla 7 – Matriz de confusión algoritmo KNN 4 vecinos

Fuente: Propia.

Interpretación de resultados con Vecinos más cercanos (4 vecinos)

Como se puede apreciar clasifico de manera correcta hasta un 86.2142%, y en la matriz de confusión se aprecia que:

De 475 registros 455 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 231 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 127 fueron clasificados correctamente como “Alto”

4.3. Algoritmo de Bayes NET

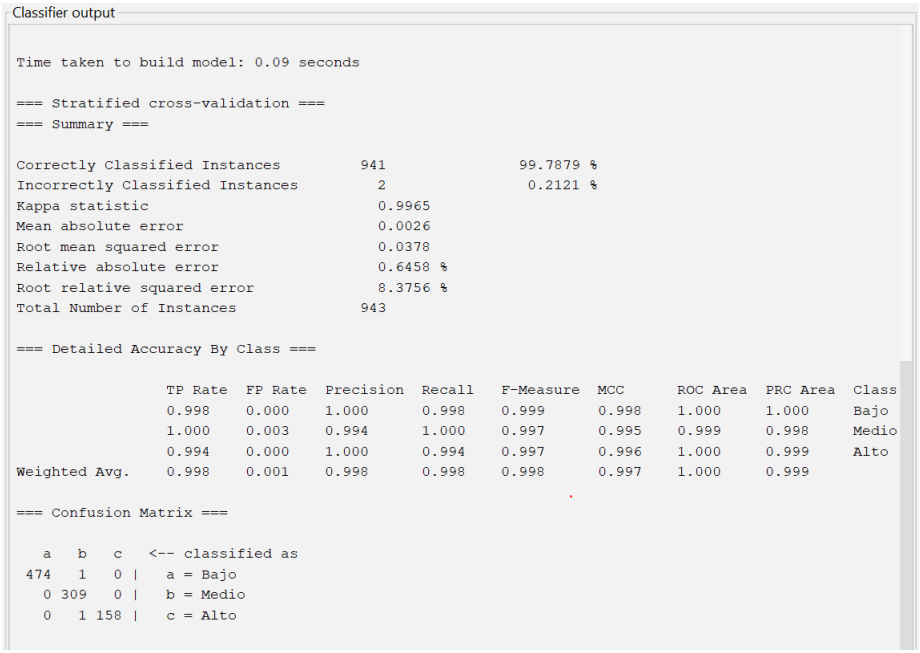


Imagen 13 – Pantalla de salida algoritmo BayesNet de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: BAYESNET

Instancias correctamente clasificadas	**941**	**99.7879%**
Instancias incorrectamente clasificadas	**2**	**0.2121%**
Total	**943**	**100%**

Tabla 8 –Resultados de clasificación algoritmo BayesNET

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	474	1	0	**475**
Medio	0	309	0	**309**
Alto	0	1	158	**159**

Tabla 9 –Matriz de Confusión algoritmo BayesNET

Fuente: Propia.

Interpretación de resultados con BAYESNET

Como se puede apreciar clasifico de manera correcta hasta un 99.7879%, y en la matriz de confusión se aprecia que:

De 475 registros 474 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 309 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 158 fueron clasificados correctamente como “Alto”

4.4. Algoritmo de NaiveBayes

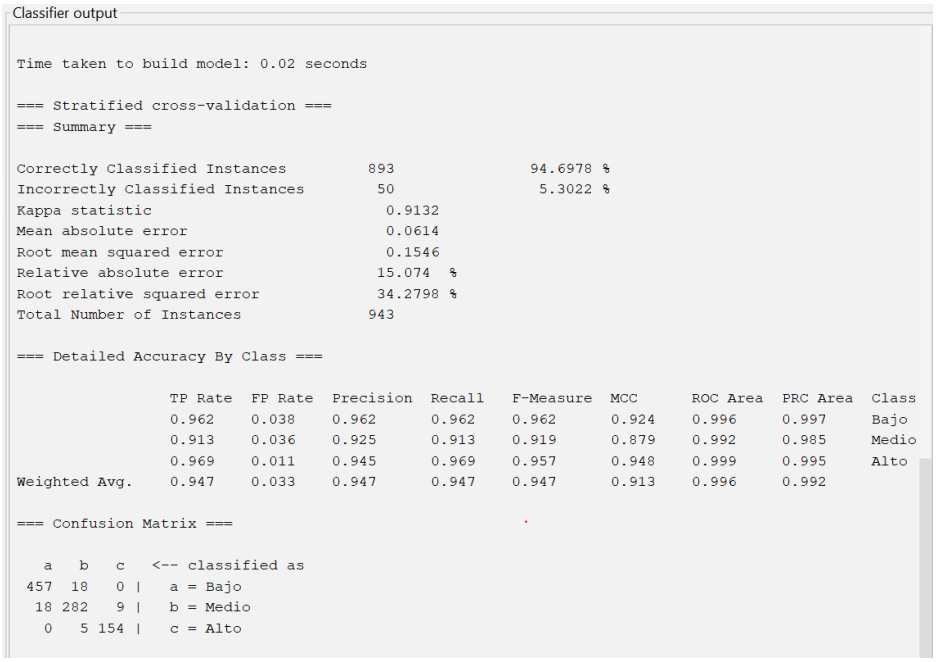


Imagen 14 – Pantalla de salida algoritmo NaiveBayes de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: NAIVEBAYES

Instancias correctamente clasificadas	**893**	**94.6978%**
Instancias incorrectamente clasificadas	**50**	**5.3022 %**
Total	**943**	**100%**

Tabla 10 –Resultados de clasificación algoritmo NaiveBayes

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	457	18	0	**475**
Medio	18	282	9	**309**
Alto	0	5	154	**159**

Tabla 11 –Matriz de confusión algoritmo NaiveBayes

Fuente: Propia.

Interpretación de resultados con NAIVEBAYES

Como se puede apreciar clasifico de manera correcta hasta un 94.6978%, y en la matriz de confusión se aprecia que:

De 475 registros 457 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 282 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 154 fueron clasificados correctamente como “Alto”

5. ALGORITMO ONER

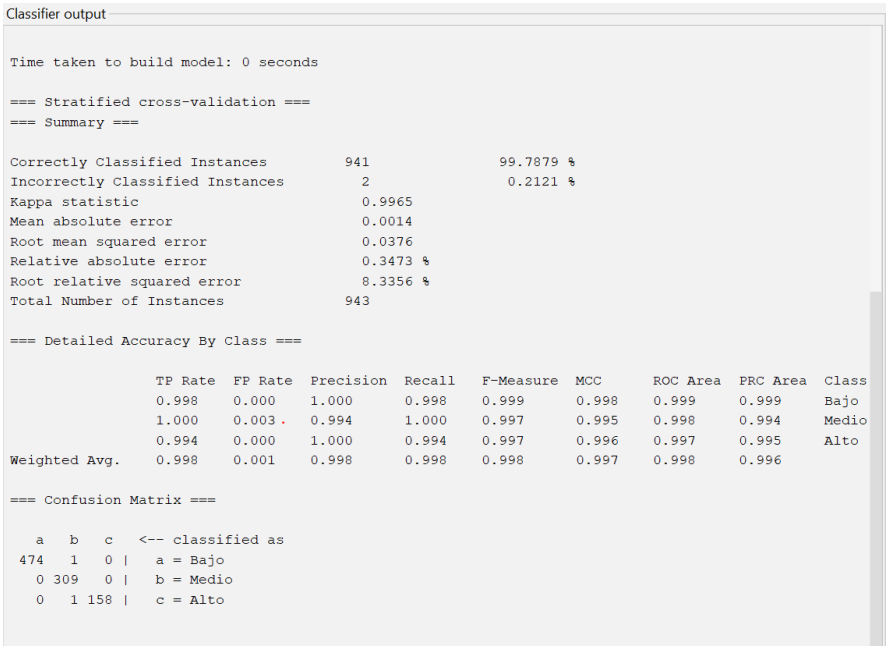


Imagen 15 – Pantalla de salida algoritmo OneR de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: ONER

Instancias correctamente clasificadas	**941**	**99.7879 %**
Instancias incorrectamente clasificadas	**2**	**0.2121 %**
Total	**943**	**100%**

Tabla 12 –Resultados de clasificación algoritmo ONER

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	474	1	0	**475**
Medio	0	309	0	**309**
Alto	0	1	158	**159**

Tabla 13 –Matriz de confusión algoritmo ONER

Fuente: Propia.

Interpretación de resultados con ONER

Como se puede apreciar clasifico de manera correcta hasta un 99.7879%, y en la matriz de confusión se aprecia que:

De 475 registros 474 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 309 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 158 fueron clasificados correctamente como “Alto”

4.5. Algoritmo JRip

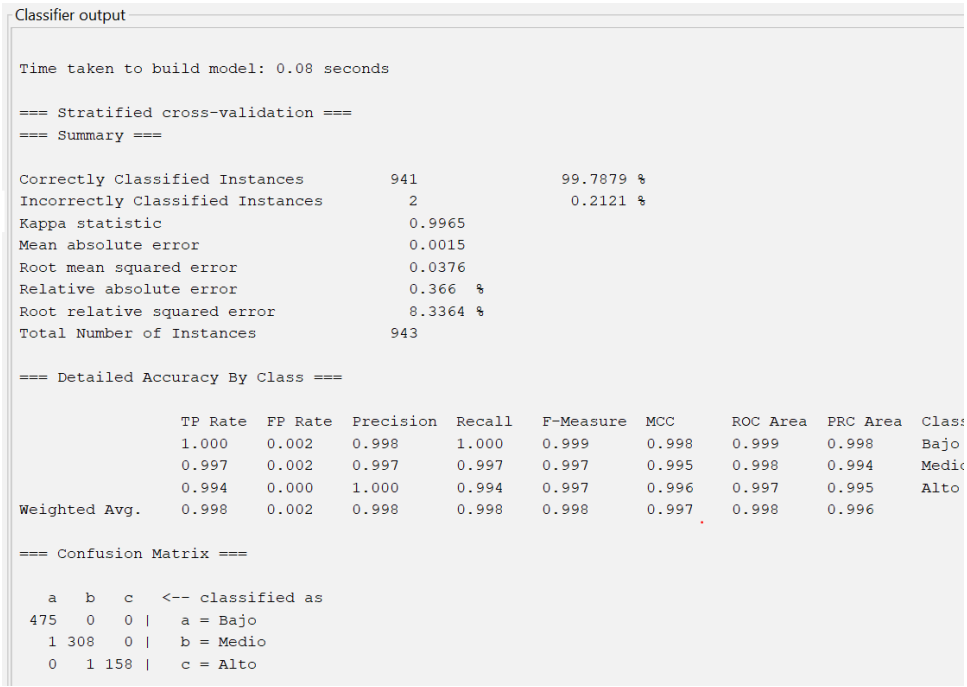


Imagen 16 – Pantalla de salida algoritmo JRip de Weka 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Algoritmo: JRip

Instancias correctamente clasificadas	**941**	**99.7879**
Instancias incorrectamente clasificadas	**2**	**0.2121 %**
Total	**943**	**100%**

Tabla 14 –Resultados de clasificación algoritmo Jrip

Fuente: Propia.

Matriz de confusión

	Bajo	Medio	Alto	Total
Bajo	475	0	0	**475**
Medio	1	308	0	**309**
Alto	0	1	158	**159**

Tabla 15 –Matriz de confusión algoritmo Jrip

Fuente: Propia.

Interpretación de resultados con JRip

Como se puede apreciar clasifico de manera correcta hasta un 99.7879 %, y en la matriz de confusión se aprecia que:

De 475 registros 475 fueron clasificados correctamente, para la condición de “Bajo”,

De 309 registros 308 fueron clasificados correctamente como “Medio”

y finalmente de 159 registros 158 fueron clasificados correctamente como “Alto”

Tabla comparativa de algoritmos

Algoritmo	***Porcentaje Clasificado correctamente***
J48	99.7879%
VECINOS MÁS CERCANOS	86.2142%

BAYESNET	99.7879%
NAIVEBAYES	94.6978%
ONER	99.7879 %
JRip	99.7879 %

Tabla 16 –Comparación de resultados de todos los algoritmos usados

Fuente: Propia.

Como se puede apreciar 4 de los algoritmos utilizados tuvieron un buen desempeño, sin embargo el algoritmo que se utilizará para la predicción del rendimiento académico de los estudiantes de nuevo ingreso de computación usando minería de datos.

Generación del modelo

Para la generación de los modelos utilizaremos el algoritmo J48 descrito anteriormente, y para esto dividiremos nuestra información en 2 grupos:

Primer grupo (Entrenamiento):

Formado por todas las instancias con un total de 943 Registros.

Segundo Grupo (Test):

Formado por cerca del 33% del total de las instancias o registros del grupo de entrenamiento. Para este segundo grupo (Test), el atributo Observación (1,0): “Alto”, “Medio” o “Bajo”, será desconocido y lo representaremos con un signo de interrogación “?”, el algoritmo será el que determine en la interrogación si el rendimiento del estudiante será “Alto”, “Medio” o “Bajo”,

Evaluación y comprobación del modelo.

Para la evaluación y comprobación del modelo se implementó el algoritmo J48 en WEKA 3.8.6, generándose los siguientes resultados:

El software generó 2 resultados en función a los datos utilizados para el Entrenamiento y para el Test, similares a las ventanas siguientes:

```
Instances:    943
Attributes:   3
              Matricula
              Scaled Full Test
              Valor
Test mode:    10-fold cross-validation

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

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

Scaled Full Test <= 586.7581: Bajo (475.0)
Scaled Full Test > 586.7581
|   Scaled Full Test <= 727.1606: Medio (309.0)
|   Scaled Full Test > 727.1606: Alto (159.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0 seconds

=== Predictions on test data ===

    inst#     actual  predicted error prediction
        1    2:Medio    2:Medio       1
        2    2:Medio    2:Medio       1
        3    2:Medio    2:Medio       1
        4    2:Medio    2:Medio       1
```

Imagen 17 – Pantalla con los datos de entrenamiento

Fuente: Propia, Imagen capturada de Weka 3.8.6

Classifier output

```
Model attributes                    Incoming attributes
----------------------------        ----------------
(numeric) Matricula             --> 1 (numeric) Matricula
(numeric) Scaled Full Test      --> 2 (numeric) Scaled Full Test
(nominal) Valor                 --> 3 missing (type mis-match)


Time taken to build model: 0.01 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?    2:Medio       1
        2        1:?    2:Medio       1
        3        1:?     3:Alto       1
        4        1:?    2:Medio       1
        5        1:?     1:Bajo       1
        6        1:?    2:Medio       1
        7        1:?     1:Bajo       1
        8        1:?    2:Medio       1
        9        1:?     1:Bajo       1
       10        1:?    2:Medio       1
       11        1:?     1:Bajo       1
       12        1:?     1:Bajo       1
       13        1:?     3:Alto       1
       14        1:?     3:Alto       1
       15        1:?     3:Alto       1
       16        1:?    2:Medio       1
       17        1:?    2:Medio       1
       18        1:?     1:Bajo       1
       19        1:?     1:Bajo       1
```

Imagen 18 – Pantalla con los datos del test

Fuente: Propia, Imagen capturada de Weka 3.8.6

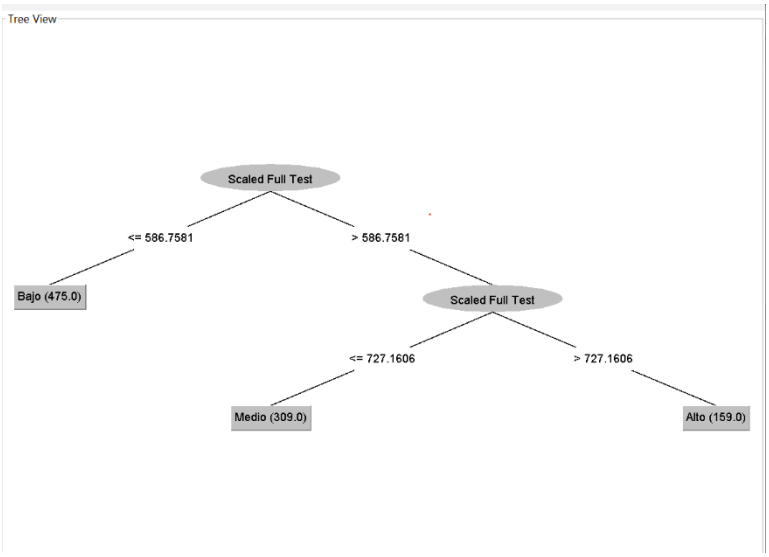


Imagen 19 – Pantalla de árbol generado con algoritmo J48 3.8.6

Fuente: Propia, Imagen capturada de Weka 3.8.6

Evaluación de Resultados

Los resultados son obtenidos a partir del test generado por WEKA 3.8.6, como se muestra a continuación en un fragmento de la siguiente pantalla.

Classifier output

```
Model attributes                    Incoming attributes
----------------------------        ----------------
(numeric) Matricula             --> 1 (numeric) Matricula
(numeric) Scaled Full Test      --> 2 (numeric) Scaled Full Test
(nominal) Valor                 --> 3 missing (type mis-match)


Time taken to build model: 0.01 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?    2:Medio       1
        2        1:?    2:Medio       1
        3        1:?     3:Alto       1
        4        1:?    2:Medio       1
        5        1:?     1:Bajo       1
        6        1:?    2:Medio       1
        7        1:?     1:Bajo       1
        8        1:?    2:Medio       1
        9        1:?     1:Bajo       1
       10        1:?    2:Medio       1
       11        1:?     1:Bajo       1
       12        1:?     1:Bajo       1
       13        1:?     3:Alto       1
       14        1:?     3:Alto       1
       15        1:?     3:Alto       1
       16        1:?    2:Medio       1
       17        1:?    2:Medio       1
       18        1:?     1:Bajo       1
       19        1:?     1:Bajo       1
```

Imagen 20 – Pantalla con datos predictivos

Fuente: Propia, Imagen capturada de Weka 3.8.6

Para la interpretación de la pantalla anterior se utilizó una hoja de Excel, en donde primero se agruparon en intervalos del más bajo al más alto de los promedios reales del primer semestre de los alumnos de computación, quedando de la siguiente manera.

Alto

9 ≤ Promedio Real ≤ 10

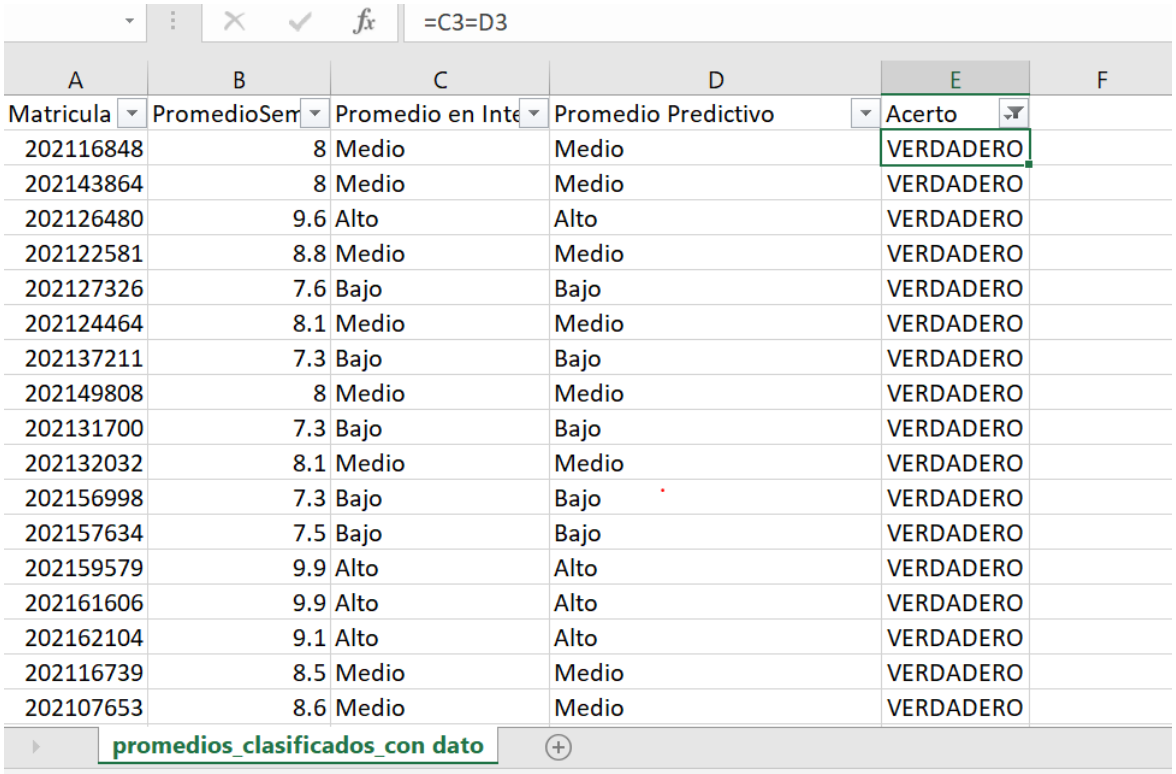
Medio

8 ≤ Promedio Real ≤ 9

Bajo

7 ≤ Promedio Real ≤ 8

Posterior a lo anterior, se aplicó una función sencilla pero muy útil con "=", para determinar si acertó o no, se mostrará con un "VERDADERO" o un "FALSO" respectivamente, a continuación, se muestran un fragmento de los resultados.



=C3=D3

	A	B	C	D	E	F
	Matricula	PromedioSem	Promedio en Inte	Promedio Predictivo	Acerto	
	202116848	8	Medio	Medio	VERDADERO	
	202143864	8	Medio	Medio	VERDADERO	
	202126480	9.6	Alto	Alto	VERDADERO	
	202122581	8.8	Medio	Medio	VERDADERO	
	202127326	7.6	Bajo	Bajo	VERDADERO	
	202124464	8.1	Medio	Medio	VERDADERO	
	202137211	7.3	Bajo	Bajo	VERDADERO	
	202149808	8	Medio	Medio	VERDADERO	
	202131700	7.3	Bajo	Bajo	VERDADERO	
	202132032	8.1	Medio	Medio	VERDADERO	
	202156998	7.3	Bajo	Bajo	VERDADERO	
	202157634	7.5	Bajo	Bajo	VERDADERO	
	202159579	9.9	Alto	Alto	VERDADERO	
	202161606	9.9	Alto	Alto	VERDADERO	
	202162104	9.1	Alto	Alto	VERDADERO	
	202116739	8.5	Medio	Medio	VERDADERO	
	202107653	8.6	Medio	Medio	VERDADERO	

promedios_clasificados_con dato

Imagen 21 – Pantalla de Aciertos y desaciertos en Excel

Fuente: Propia, Imagen capturada de Excel

En este proceso, las 340 instancias que se utilizaron datos del Test fueron acertados el 100%.

CAPÍTULO 5

CONCLUSIONES

Las aportaciones de la minería de datos educativos han influido en las teorías de la pedagogía y el aprendizaje y han impulsado la mejora del software educativo, sobre todo respecto a su capacidad para personalizar la experiencia del estudiante.

Los métodos de minería de datos a menudo se implementan en universidades avanzadas hoy en día para analizar los datos disponibles y extraer información y conocimiento para apoyar la toma de decisiones.

Según los resultados encontrados, el algoritmo de árboles de decisión J48, BAYESNET y JRIP, fueron los algoritmos que tuvieron el mejor resultado para la predicción del rendimiento académico de los ingresados del primer semestre a la facultad de ciencias de la computación con un 99.7879% de predicción. Sin embargó el algoritmo que se utilizó para este trabajo fue el J48 los resultados que se obtuvieron con los datos reales y el dataset fue de 100% acertado.

BIBLIOGRAFÍA.

Aldana, Kelsy, Reyna Pérez de Roberti y Ayolaida Rodríguez Miranda. (2010). Visión del desempeño académico estudiantil en la Universidad Centroccidental Lisandro Alvarado. *Revista Compendium,* 13(24)*, Universidad Centroccidental Lisandro Alvarado. Venezuela,* 5 – 21.

Astha Soni, Vivek Kumar, Rajwant Kaur and D. Hemavathi. (2018). Predicting student performance using data mining techniques. *International Journal of Pure and Applied Mathematics,* 119(12), 221 – 227.

B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and W. F. Punch. (2003). Predicting student performance: an application of data mining methods with an educational Web-based system. *33rd Annual Frontiers in Education, 2003. FIE 2003., Westminster, CO*, T2A – 13.

Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-Step Data Mining Guide 2000. SPSS Inc. CRISPWP-0800, *recuperado de* http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf

Edel, Rubén. (2003). El rendimiento académico: concepto, investigación y desarrollo. *Revista electrónica de iberoamericana sobre calidad, eficacia y cambio en educación, REICE*, 1(2). Madrid, España, 1 – 15.

Mella, Orlando e Iván Ortiz. (1999). Rendimiento escolar. Influencias diferenciales de factores externos e internos. *Revista Latinoamericana de Estudios Educativos*, 29(1), *Centro de Estudios Educativos, A. C. México*, 69 – 92.

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetic and Information Technologies, Bulgarian Academic of Sciences,* 13(1), 61 – 72.

K. Sangeeta, T. PanduRanga Vital, Kalyana Kiran Kumar. (2020). Student Classification Based on Cognitive Abilities and Predicting Learning Performances using Machine Learning Models. *International Journal of Recent Technology and Engineering,* 8(6), 3554 – 3569.

Panth, M. K.; Sahu, V. and Gupta, M. (2015). A comparative study of emotional intelligence and intelligence quotient between introvert and extrovert personality. *International Journal of Research in Humanities, Arts and Literature,* 3(5), 41 – 54.

Ravinder A. and Yash K. (2017). Predicting the Probability of Student's Degree Completion by Using Different Data Mining Techniques. *Fourth International Conference on Image Information Processing* (ICIIP), 474 – 477.

R. Lakshmi, K. S. Narayanan, R. Swathikrishna, K. M. Sameera. (2017). Predicting Student Performance Using Data Mining. *Journal of Network and Information Security*, 8 – 11.

Romero, C., S. Ventura. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications,* 33, 135 – 146.

Witten, I. and E. Frank. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. USA, Morgan. Kaufmann Publishers, Elsevier Inc.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. (2007). *Top 10 algorithms in data mining*. London: Springer-Verlag

Kelleher, J. & Mac, B. & D'Arcy, A. (s.f.) Fundamnetals of Machine learning for predictive data analyTIC's algorithms, worked, and case of studies, London, Inglaterra.

Samanpour AR, Ruegenberg A., Ahlers R. (2018) El futuro del aprendizaje automático y el análisis predictivo. En: Linnhoff-Popien C., Schneider R., Zaddach M. (eds) Digital Marketplaces Unleashed. Springer, Berlín, Heidelberg. https://doi-org.proxydgb.buap.mx/10.1007/978-3-662-49275-8_30

Moreno, A. & Armengol, E. & Béjar, J. & Belanche, L. & Cortés, U. & Gavalda, R. & Gimeno, J. & López, B. & Martin, M. & Sánchez, M. (1994), Aprendizaje automático, Barcelona, España. Edicions de la Universidad Politécnica de Cataluña.

Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado? Recuperado 26 de enero de 2019, de https://medium.com/@juanzambrano/aprendizaje-supervisado-ono-supervisado-39ccf1fd6e7b