BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Facultad de Ciencias de la Computación

Ingeniería en Tecnologías de la Información

“Técnicas de Clasificación para la Predicción del Desempeño de los Estudiantes usando Minería de Datos”

TESIS PRESENTADA PARA OBTENER EL GRADO DE

LICENCIATURA EN

Ingeniera en Tecnologías de la Información

PRESENTA

KAREN JOSEFINA RIVERA TORRES

DIRECTORES DE TESIS

ROBERTO CONTRERAS JUÁREZ
HÉCTOR DAVID RAMÍREZ HERNÁNDEZ

DICIEMBRE 2022

Dedicatoria

Este trabajo y esfuerzo está dedicado a las personas que me han apoyado en todo momento, en la buena y en las malas, principalmente a mis padres.

A mi madre Maria del Pilar Torres Zapata y a mi padre Jose Arturo Rivera Jaramillo.

Gracias por enseñarme a afrontar las dificultades, por enseñarme a ser la persona que soy hoy, mis principios y valores, mi perseverancia y mi empeño. Todo esto con una enorme dosis de amor y sin pedir nada a cambio. También a mis hermanas

Jaqueline Trinidad Rivera Torres y H. Gabriela Barragán Rivera

Gracias por tenerme paciencia, gracias por las lecciones de vida.

Y por último pero no menos importante a mis amigos y compañeros de la universidad, que me apoyaron, me cuidaron y siempre me enseñaron y me explicaban temas que no comprendía, gracias.

Contenido

Índice de tablas y Figuras

Capítulo 1.

Fundamentos Básicos

1.1 Introducción a la Minería de Datos

La minería de datos es el descubrimiento del conocimiento en las bases de datos y básicamente consiste en la extracción de información que se oculta de manera implícita en los datos [1]. La minería de datos o DM por sus siglas en ingles tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta [2, 4]. El objetivo fundamental es aprovechar el valor de la información localizada y usar los patrones preestablecidos para que los directivos tengan un mejor conocimiento de su negocio y puedan tomar decisiones más confiables [3, 4].

Las tareas de minería de datos generalmente se dividen en 2 categorías principales, tareas predictivas y descriptivas. El modelado predictivo se refiere a la tarea de construir un modelo para la variable objetivo en función de la variable explicativa. Los dos tipos de tareas de modelado predictivo son la clasificación, que se utiliza para predecir atributos de destino continuos. El objetivo de ambas tareas es crear un modelo que minimice el error entre los valores predichos y verdaderos de la variable de destino. [5]

Generar un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los siguientes pasos básicos (Microsoft, 2016): a) definir el problema, b) preparar los datos, c) explorar los datos, d) generar modelos, e) explorar y validar los modelos, f) implementar y actualizar los modelos.

En la figura 1 se puede visualizar a cada uno de los pasos del proceso, el cual es cíclico, lo que significa que la creación de un modelo de DM es un proceso dinámico e iterativo (Microsoft, 2016).

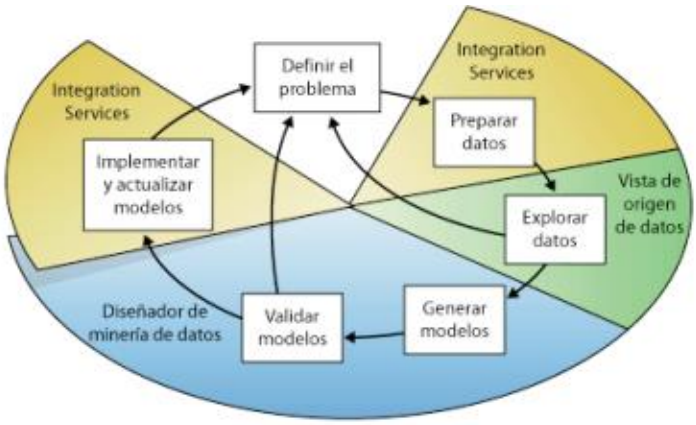

Figura 1. Pasos del proceso de la creación de un modelo de DM [4]

La minería de datos conjuga la estadística y las ciencias de la computación para intentar descubrir patrones en grandes volúmenes de datos. La técnica más aplicada de minería de datos es la clasificación, que consiste en utilizar un conjunto preclasificado de ejemplos para desarrollar un modelo que permita clasificar un gran volumen de información.

El proceso de clasificación de datos implica dos fases: el aprendizaje y la clasificación. En la fase de aprendizaje, mediante el algoritmo de clasificación, se analizan los datos de entrenamiento. El algoritmo de aprendizaje de clasificación usa los ejemplos preclasificados para determinar el conjunto de parámetros que se necesitan para realizar una discriminación adecuada. El algoritmo codifica estos parámetros en un modelo llamado clasificador. La fase de clasificación de los datos de prueba sirve para estimar la precisión de las reglas de clasificación. Si la precisión es aceptable, entonces las reglas se pueden aplicar a los datos. [6]

La minería de datos se ha utilizado en muchas áreas, especialmente en el área comercial, pero en los últimos años ha cobrado un gran interés en el ámbito educativo. La minería de datos aplicada a la educación se conoce como minería de datos educativos (EDM, por sus siglas en ingles) y es la encargada de extraer información útil de las bases de datos de las instituciones educativas aplicando herramientas y técnicas de minería de datos para una mejor comprensión o con el fin de obtener mayores conocimientos [7].

1.2 Minería de datos Educativa

La minería de datos en la educación no es un tópico nuevo, pero su estudio y aplicación ha sido muy relevante en los últimos años. El uso de estas técnicas permite, entre otras cosas, predecir cualquier fenómeno dentro del ámbito educativo. De esta forma, utilizando las técnicas que nos ofrece la minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de desertar de cualquier estudiante. [8]

Como es sabido, la educación es uno de los factores más importantes para el desarrollo de un país, esto obliga a que las instituciones educativas tengan como objetivo primordial ofrecer programas educativos de calidad y una forma de cumplir con esta calidad es a través de la evaluación del rendimiento o desempeño académico de los estudiantes.

Este proceso de evaluación es bastante complejo ya que no puede ser limitado a una simple calificación obtenida en un examen, pues existen muchos factores que pueden afectar los buenos o malos resultados y conocerlos ayuda a los responsables de la enseñanza-aprendizaje a planificar y personalizar sus programas educativos en función de la información recibida.

La EDM es uno de los enfoques de la minería de datos que puede proporcionar una ayuda eficaz para revelar las complejas relaciones que hay detrás de las calificaciones y el diagrama de la Figura 2 explica las diferentes etapas de la metodología para la extracción de información sobre datos educativos para la toma de decisiones académicas [9].

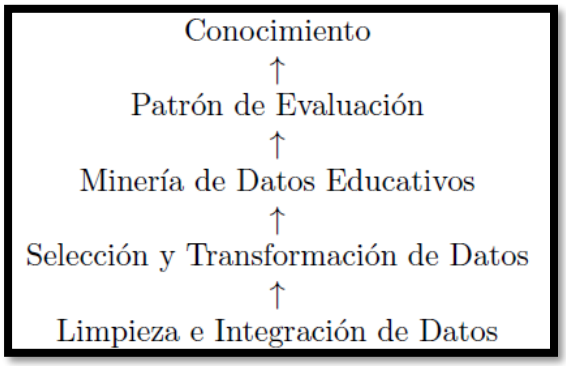

Figura 2. Etapas de la EDM [9]

Existen demasiados autores que aplicaron la minería de datos a la educación con diferentes propósitos como predecir el desempeño de los alumnos, la deserción de los alumnos en las diferentes áreas de la universidad, predicción de los resultados de los alumnos al final de un semestre o de una materia en específico, para descubrir patrones pedagógicos, dar un mejor servicio a los alumnos, etc. Algunos ejemplos de diferentes autores son:

Chamillard (2006) utilizó técnicas de análisis estadístico para predecir el desempeño de los estudiantes en un curso en particular. Las observaciones del análisis proporcionan información útil sobre las relaciones entre los cursos.

Superby et al . (2006) y Vandamme et al . (2007) estudiaron las correlaciones de varios parámetros como la asistencia, la probabilidad estimada de éxito, la experiencia académica previa y las habilidades de estudio. Descubrieron que cambiar los factores del proceso durante la estadía de un estudiante en la universidad juega un papel importante en el rendimiento académico. Además, experimentaron con la predicción del rendimiento de los

estudiantes utilizando árboles de decisión, redes neuronales y análisis discriminante lineal. Las tasas de predicción obtenidas no fueron especialmente buenas debido a la dificultad de clasificar a los estudiantes en 3 grupos, a saber, riesgo alto, riesgo medio y riesgo bajo, antes de los primeros exámenes universitarios.

Golding y Donaldson (2006) afirmaron que el uso del desempeño en el primer año del curso de ciencias de la computación es un posible factor que puede determinar el desempeño académico. También demostraron que el género y la edad no tienen una correlación significativa como factores predictivos.

McKenzie y Schweitzer (2001) investigaron predictores académicos, psicosociales, cognitivos y demográficos del desempeño académico para mejorar las intervenciones y los servicios de apoyo para estudiantes en riesgo de problemas académicos. Recomendaron implementar procedimientos estrictos de mantenimiento de registros a nivel universitario para permitir a los investigadores examinar completamente la relación entre la edad, el rendimiento académico anterior y el rendimiento universitario.

Merceron y Yacef (2005) utilizaron algoritmos de minería de datos para descubrir patrones pedagógicamente interesantes. Sus hallazgos se usaron para ayudar a los maestros a administrar su clase, comprender el comportamiento de sus alumnos y apoyar la reflexión de los alumnos a través de comentarios proactivos. Su conjunto de datos se recopila de una herramienta de tutoría basada en la web que se enfoca en ejercicios de pruebas lógicas.

Kotsiantis et al. (2003) predijeron abandonos en la mitad de un curso al comparar seis métodos de clasificación (Naive Bayes, árbol de decisiones, red neuronal de avance, máquina de vectores de soporte, 3 vecinos más cercanos y regresión logística). El conjunto de datos que constaba de 350 registros contenía datos demográficos, resultados de las primeras asignaciones de escritura y participación en reuniones de grupo. Sus mejores clasificadores, Naive Bayes y red neuronal, fueron capaces de predecir alrededor del 80 % de los abandonos.

Minaei-Bidgoli et al. (2003) predijeron los resultados finales del curso a partir de los datos de registro de un sistema de aprendizaje mediante la comparación de seis clasificadores (clasificador bayesiano cuadrático, 1-vecino más cercano, k-vecino más cercano, ventana de Parzen, red neuronal de avance y árbol de decisión). Los datos que consistieron en 250 registros contenían atributos relacionados con cada tarea resuelta y otras acciones como participar en el mecanismo de comunicación y leer material de apoyo. Su mejor clasificador, k-vecinos más cercanos, logró más del 80 % de precisión, cuando los resultados finales tenían solo dos clases (aprobado/no aprobado).

Minaei-Bidgoli et al. (2004) aplicaron clasificadores de minería de datos como medio de análisis y comparación del uso y desempeño de estudiantes que han tomado un curso técnico vía web. Sus resultados muestran que la combinación de múltiples clasificadores conduce a una mejora significativa de la precisión en el conjunto de datos dado.

Hamalainen y Vinni (2006) compararon métodos de aprendizaje automático para el sistema de tutoría inteligente. Abordaron el problema en el que los conjuntos de datos educativos son tan pequeños que los métodos de aprendizaje automático no se pueden aplicar directamente. Dieron esquemas generales y recomendaron variaciones de clasificadores bayesianos ingenuos que son robustos.

1.3 Introducción a Weka

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación, clasificación, regresión, agrupación, minería de reglas de asociación y visualización de datos. Weka es un software de código abierto emitido bajo la Licencia Pública General GNU. [21]

El proyecto WEKA fue financiado por el gobierno de Nueva Zelanda desde 1993 hasta hace poco. La solicitud de financiación original se presentó a finales de 1992 y establecía los objetivos del proyecto como: "El programa tiene como objetivo construir una instalación de vanguardia para desarrollar técnicas de aprendizaje automático e investigar su aplicación en esferas clave de la economía de Nueva Zelandia. Específicamente crearemos un banco de trabajo para el aprendizaje automático, determinaremos los factores que contribuyen a su aplicación exitosa en las industrias agrícolas y desarrollar nuevos métodos de aprendizaje automático y formas de evaluar su eficacia" [10]

WEKA tiene como objetivo proporcionar una colección completa de algoritmos de aprendizaje automático y herramientas de pre-procesamiento de datos para investigadores y profesionales por igual. Incluye algoritmos para regresión, clasificación, agrupación, minería de reglas de asociación y selección de atributos. [21]

Para instalar Weka en Windows se necesita tener una versión de java, la versión de java depende de la versión de Weka que se requiere instalar, normalmente se necesita de java 1.4 a posteriores. También existen versiones de Weka para el sistema operativo MAC OS con procesador Intel o con procesador ARM, y también se puede instalar en Linux. Para descargar el software de Weka se obtiene de este link: https://waikato.github.io/weka-wiki/downloading_weka/

WEKA tiene interfaces gráficas de usuario que permiten un fácil acceso a la funcionalidad y se basa en una interfaz de paneles, donde diferentes paneles corresponden a diferentes

tareas de minería de datos. La principal interfaz gráfica de usuario es el "Explorador". En el primer panel, llamado panel "Pre-procesamiento", los datos se pueden cargar y transformar utilizando las herramientas de pre-procesamiento de datos de WEKA, llamadas "filtros". Los datos se pueden cargar desde varias fuentes, incluidos archivos, URL y bases de datos. Los formatos de archivo admitidos incluyen el formato ARFF propio de WEKA, CSV, el formato de LibSVM y el formato de C4.5. También es posible generar datos usando una fuente de datos artificial y editar datos manualmente usando un editor de conjuntos de datos. [21]

El segundo panel en el Explorador da acceso a WEKA algoritmos de clasificación y regresión. El correspondiente panel se llama "Clasificar" porque las técnicas de regresión son vistos como predictores de “clases continuas”. También muestra una representación textual del modelo construido a partir del conjunto de datos completo. Sin embargo, otros modos de evaluación, establecido en un conjunto de prueba separado, también son compatibles. Si es aplicable, el panel también proporciona acceso a representaciones gráficas de modelos, p. árboles de decisión. Además, puede visualizar errores de predicción en diagramas de dispersión, y también permite la evaluación a través de curvas ROC y otras “curvas de umbral”. Los modelos pueden también se guardará y cargará en este panel. [21]

El panel "Cluster" que es el tercer panel, permite a los usuarios ejecutar un algoritmo de agrupamiento en los datos cargados en el Pre-proceso panel. El cuarto panel “Asociado” tiene más técnicas para el agrupamiento que para la minería de reglas de asociación. [21]

1.3.1 Algoritmos de clasificación en Weka

En este trabajo, se analizan técnicas de clasificación en minería de datos para predecir, a partir de las calificaciones promedio de los estudiantes en las asignaturas de habilidades del pensamiento, español y matemáticas, el puntaje en un examen estandarizado utilizado como prueba de admisión al nivel medio superior de una institución pública.

Para predecir el desempeño de los estudiantes se utilizaron los algoritmos más populares entre los investigadores: DecisionStump, J48, MultilayerPerceptron, RandomForest y RandomTree y se comparan sus desempeños utilizando las medidas: Precision, Recall, F-measure, Accuracy Scores y Kappa Statistic.

Con la finalidad de corroborar si el umbral de corte en el puntaje era el adecuado, se realizó un agrupamiento de datos. Los algoritmos utilizados para realizar el agrupamiento de los datos fue k-means clustering algorithm y NaiveBayes algorithm.

Árbol de decisión J48

La clasificación es el proceso de construir un modelo de clases a partir de un conjunto de registros que contienen etiquetas de clase. El algoritmo del árbol de decisión es para

averiguar la forma en que se comporta el vector de atributos para una serie de instancias. [22]

J48 es una extensión de ID3. Las características adicionales de J48 son la contabilidad de valores faltantes, la poda de árboles de decisión, los rangos de valores de atributos continuos, la derivación de reglas, etc. En la herramienta de minería de datos WEKA, J48 es una implementación Java de código abierto del algoritmo C4.5. La herramienta WEKA proporciona una serie de opciones asociadas con la poda de árboles. En caso de potencial sobreajuste, la poda se puede utilizar como herramienta para precisar. En otros algoritmos, la clasificación se realiza de forma recursiva hasta que cada hoja es pura, es decir, la clasificación de los datos debe ser lo más perfecta posible. Este algoritmo genera las reglas a partir de las cuales se genera la identidad particular de esos datos. El objetivo es la generalización progresiva de un árbol de decisión hasta que gane el equilibrio de flexibilidad y precisión. [22]

El clasificador J48 es un árbol de decisión C4.5 simple para la clasificación. Crea un árbol binario. El enfoque del árbol de decisión es más útil en el problema de clasificación. Con esta técnica, se construye un árbol para modelar el proceso de clasificación. Una vez que se construye el árbol, se aplica a cada tupla en la base de datos y da como resultado la clasificación para esa tupla. [23]

A continuación, se muestra el algoritmo de j48:

Algorithm J48 [23]:

```
INPUT:
        D //Training data
OUTPUT
        T //Decision tree
DTBUILD (*D)
{
        T=φ;
        T= Create root node and label with splitting attribute;
        T= Add arc to root node for each split predicate and
        label;
        For each arc do
                D= Database created by applying splitting
                predicate to D;
                If stopping point reached for this path, then
                        T'= create leaf node and label with
                        appropriate class;
```

Else

T'= DTBUILD(D);

T= add T' to arc;

}

Al construir un árbol, J48 ignora los valores que faltan, es decir, el valor de ese elemento se puede predecir en función de lo que se sabe sobre los valores de los atributos de los otros registros. La idea básica es dividir los datos en rangos según los valores de atributos para ese elemento que se encuentran en la muestra de entrenamiento. J48 permite la clasificación a través de árboles de decisión o reglas generadas a partir de ellos. [23]

k-means clustering algorithm

El proceso de k-means se ideó originalmente en un intento de encontrar un método factible para calcular dicha partición óptima. En general, el procedimiento de k-means no convergerá a una partición óptima, aunque hay casos especiales en los que sí lo hará. El procedimiento de k-means consiste simplemente en comenzar con k grupos, cada uno de los cuales consta de un solo punto aleatorio, y luego agregar cada nuevo punto al grupo cuya media es el punto nuevo más cercano. Después de agregar un punto a un grupo, la media de ese grupo se ajusta para tener en cuenta el nuevo punto. Así, en cada etapa, las k-means son, de hecho, las medias de los grupos que representan. [24]

El algoritmo propuesto por Hartigan y Wong (1979) [25] que define la variación total dentro de un clúster como la suma de las distancias al cuadrado sobre las distancias euclidianas entre elementos y el centroide, tal que: [26]

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Figura 3. Ecuación de k-means [26]

Donde, x_i diseña un punto de datos pertenecientes al clúster C_k y μ_k es el valor medio de los puntos asignados en el clúster C_k De esta manera, cada observación de x_i se asigna a un grupo dado de tal manera que la distancia de la suma de sus cuadrados de cada observación respecto del centro de cada grupo asignado μ_k sea minimizado. [26]

El enfoque propuesto comprende tres pasos principales: agrupar los nodos, encontrar la ruta óptima en cada grupo y volver a conectar los grupos. El primer paso usa el agrupamiento k-means para dividir los nodos en subproblemas, el segundo paso utiliza FA para encontrar la ruta óptima en cada grupo, finalmente reconectar todos los grupos y devolver la ruta entre ellos [27] .

El algoritmo K-means está basado en una función que minimiza la distancia entre el centroide y cada punto y, siempre usa el método de gradiente para obtener el extremo mínimo. La dirección de búsqueda en el método de gradiente siempre es a lo largo de la dirección en la que disminuye la función, lo que conducirá al hecho de que cuando el punto focal del clúster inicial no es adecuado, todo el algoritmo se hundirá fácilmente en el punto mínimo local (Li y Wu, 2012). El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo. El algoritmo lleva a cabo los pasos que se muestran en la figura 4. hasta alcanzar el criterio de convergencia (los objetos no se cambian de grupo, lo que significa que cada punto o nodo, encontró el centroide con distancia mínima) [27].

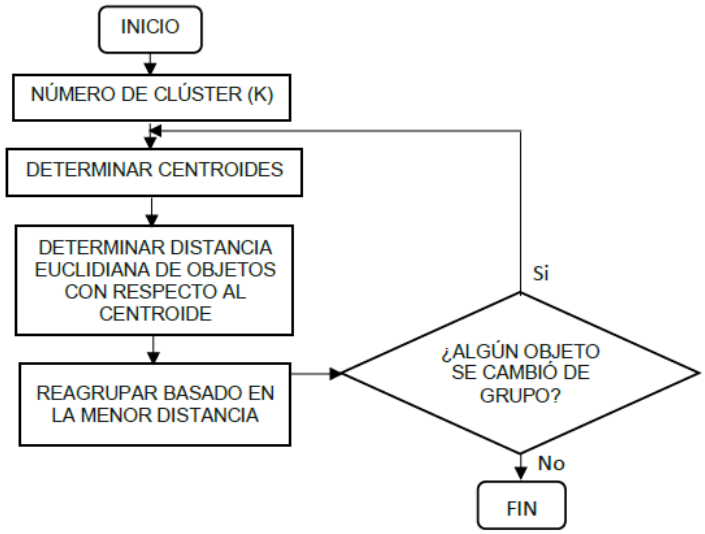

Figura 4. Diagrama de flujo del algoritmo K-means [27]

NaiveBayes algorithm

El algoritmo Naive Bayes es un clasificador probabilístico simple que calcula un conjunto de probabilidades contando la frecuencia y las combinaciones de valores en un conjunto de datos dado. El algoritmo usa el teorema de Bayes y asume que todos los atributos son independientes dado el valor de la variable de clase. Esta suposición de independencia condicional rara vez se cumple en aplicaciones del mundo real, por lo tanto, la caracterización como Ingenuo, pero el algoritmo tiende a funcionar bien y aprende rápidamente en varios problemas de clasificación supervisados. [23]

El teorema de Bayes y el teorema de la probabilidad total. La probabilidad de que un el documento d con el vector $\underset{x}{\rightarrow} =< x_1, \ldots, x_n >$ pertenece a la categoría c es [28].

$$P(C=c \mid \overrightarrow{X}=\vec{x}) = \frac{P(C=c)\cdot P(\overrightarrow{X}=\vec{x} \mid C=c)}{\prod_{k\in\{spam,legit\}} P(C=k)\cdot P(\overrightarrow{X}=\vec{x} \mid C=k)}$$

Figura 5. ecuación del teorema Bayes [28]

Sin embargo, los posibles valores de $\underset{X}{\rightarrow}$ son demasiados y también hay problemas de escasez de datos. Por lo tanto, el clasificador bayesiano Naïve asume que $x_1, \ldots, x_n$ son condicionalmente independientes dada la categoría c. Por lo tanto, en la práctica, la probabilidad de que un documento d con el vector $\underset{x}{\rightarrow} =< x_1, \ldots, x_n >$ pertenece a la categoría c es [28].

$$P(C=c \mid \overrightarrow{X}=\vec{x}) = \frac{P(C=c)\cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=c)}{\prod_{k\in\{spam,legit\}} P(C=k)\cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=k)}$$

Figura 6. ecuación del teorema Bayes [28]

$P(X_i|C)$ y $P(C)$ son fáciles de obtener a partir de las frecuencias del conjunto de datos de entrenamiento. Hasta ahora, muchas investigaciones han demostrado que el clasificador Naïve Bayesian es sorprendentemente efectivo [28].

Capítulo 2

Materiales y Procedimientos

2.1 Datos

Esta investigación está enfocada a estudiar los métodos de clasificación y de agrupamiento para predecir el desempeño de los estudiantes que desean ingresar al nivel medio superior al presentar su examen de admisión en una institución pública. Como se ha mencionado se busca identificar las características académicas en cada una de las asignaturas que conlleven a un mejor desempeño. Por esta razón, se describe la metodología y el procedimiento que ayude lograr el objetivo planteado.

Para esta investigación se utilizan los datos de 7118 estudiantes que aplicaron examen de admisión al nivel medio superior en el año 2020 a una institución pública. Se tomaron los promedios de calificaciones del nivel básico (Secundaria) en las asignaturas de habilidades, español y matemáticas junto con el puntaje alcanzado en la prueba de admisión. Es importante mencionar que únicamente se consideraron datos académicos, datos sensibles como nombre, número de identificación, lugar de procedencia, nivel socioeconómico, entre otros, no fueron considerados.

Los promedios de las asignaturas son datos numéricos enteros donde el mínimo es 6 y el máximo es 10. Por su parte, dado que el objetivo es predecir el desempeño del estudiante en la prueba de admisión, los puntajes logrados en el examen de admisión fueron agrupados de acuerdo con el estatus del estudiante en “Aceptado” (A), si el puntaje era superior a 600 y “No Aceptado” (NA), si el puntaje era inferior a 600. El umbral de 600 puntos como corte de aceptación fue tomado debido a que este número representó, en promedio, el puntaje mínimo necesario para ser aceptado en el nivel medio superior. La figura 7 muestra un ejemplo de los datos considerados en este trabajo de investigación.

Calif1_EM.arff

Relation: Calif1_EM

No.	1: P_HP Numeric	2: P_Esp Numeric	3: P_Mat Numeric	4: **Status** Nominal
1	9.0	9.0	9.0	A
2	7.0	8.0	7.0	NA
3	8.0	9.0	8.0	A
4	7.0	6.0	7.0	NA
5	9.0	9.0	10.0	A
6	8.0	8.0	9.0	A
7	8.0	8.0	8.0	A
8	8.0	8.0	8.0	A
9	7.0	7.0	7.0	NA
10	7.0	9.0	7.0	A
11	7.0	7.0	7.0	NA
12	7.0	7.0	8.0	NA
13	7.0	8.0	7.0	NA
14	8.0	8.0	8.0	A
15	7.0	7.0	8.0	NA

Figura 7. Muestra de la Base de Datos

La simbología utilizada se describe a continuación.

- **P_HP:** promedio en la asignatura especial *"Habilidades del Pensamiento"*
- **P_Esp:** promedio en la asignatura de *"Español"*
- **P_Mat:** promedio en la asignatura de *"Matemáticas"*
- **Status:** Estatus del estudiante Aceptado (A) y No Aceptado (NA)

2.2 Métodos de clasificación y agrupamiento

Con la finalidad de corroborar si el umbral del puntaje para la aceptación de un aspirante era el adecuado, se realizó un agrupamiento de datos. El agrupamiento o data clustering es el proceso mediante el cual se dividen los datos en grupos con características similares.

Los algoritmos utilizados para realizar el agrupamiento de los datos fue k-means clustering algorithm y NaiveBayes algorithm, ya que mostraron un mejor desempeño y de acuerdo con las referencias analizadas son los más utilizados. En el capítulo 1 se describe a más detalle estos algoritmos.

El el k-means clustering algorithm es un procedimiento de agrupación de una serie de ítems según criterios habitualmente de distancia, mientras que el NaiveBayes algorithm supone que, dada la variable clase, los valores de un rasgo particular no están relacionados con la presencia o ausencia de cualquier otro rasgo [29].

Con la investigación de diferentes algoritmos de clustering, el k-means clustering algorithm ofrece un método de agrupamiento, que tiene como objetivo la participación de un

conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor es más cercano.

Este agrupamiento fue corroborado a través del NaiveBayes algorithm mediante su matriz de confusión.

La clasificación es la técnica más utilizada en la minería de datos. Se trata de crear un modelo con la clasificación de un conjunto de datos. Hemos sometido a estudio los datos utilizando varios algoritmos de clasificación y, como veremos, algunos proporcionan excelentes resultados mientras que otros (como el DecisionStump) dan peores aproximaciones a priori.

En el capítulo 1 se ha explicado a más detalle los algoritmos de clasificación J48 y Naive Bayes. Adicionalmente, como se menciona en la sección 1.2 del capítulo 1 los diferentes autores citados utilizan estos algoritmos por la facilidad y la precisión de sus resultados.

En este trabajo los algoritmos utilizados, por su facilidad e importancia como se mencionó.

- J48: algoritmo de clasificación es el mejor modelo para predicción con mayor precisión. Además, utiliza la poda de error reducido.
- Random Tree: algoritmo de clasificación es eficiente en grandes bases de datos. En un gran número de conjuntos de datos da resultados precisos.
- Naïve Bayes: mejor algoritmo de clasificación para la predicción. Se utiliza cuando la dimensionalidad de la entrada es alta. Proporciona la mayor precisión y el menor error.

Algoritmos como Multilayer Perceptron y Ramdom Forrest fueron utilizados solo para contrastar los resultados obtenidos con los algortimos J48, Ramdom Tree y Naive Bayes.

Otro concepto fundamental que no podemos ignorar en los métodos de clasificación son los criterios para la evaluación de los clasificadores. Estos ayudan a estimar la bondad de un clasificador y se conoce como proceso de validación, permitiendo efectuar una eficaz medición sobre la capacidad de predicción del modelo generado a partir del clasificador.

Para verificar la bondad de los clasificadores empleados en este trabajo se utilizó la matriz de confusión. Esta matriz permite visualizar, mediante la tabla de contingencia, la distribución de errores cometidos por el clasificador y su forma para dos clases se puede observar en la Tabla 1.

	Clase 1	Clase 2
Clase 1	Verdaderos Positivos	Falsos Positivos
Clase 2	Falsos negativos	Verdaderos Negativos

Tabla 1. Matriz de Confusión

Otra alternativa utilizada es el Coeficiente Kappa (κ), el coeficiente estadístico propuesto por Cohen que permite medir la concordancia entre los resultados de dos o más variables cualitativas [30]. El índice κ, aplicado a la matriz de confusión permite evaluar si la clasificación observada es concordante con la clasificación predicha por el clasificador.

En la Tabla 2 se presentan las valoraciones de los valores de κ que utiliza la escala propuesta por Landis and Koch en [31].

Valor	Grado de Concordancia
$\kappa < 0.0$	No existe
$0.0 < \kappa \leq 0.2$	Insignificante
$0.2 < \kappa \leq 0.4$	Discreto
$0.4 < \kappa \leq 0.6$	Moderado
$0.6 < \kappa \leq 0.8$	Sustancial
$0.8 < \kappa \leq 1.0$	Casi perfecto

Tabla 2.Valoración del coeficiente Kappa

2.3 Procesamiento de datos

Para alcanzar el objetivo de este trabajo, se ha aplicado minería de datos usando el paquete de software Weka (Waikato Environment for Knowledge Analysis) que es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático sobre cualquier conjunto de datos del usuario [21,32]. En el capítulo 1 se detalla más acerca de este software utilizado.

Aunque Weka es una potente herramienta tanto en el desarrollo de algoritmos de clasificación y filtrado como en el preprocesado de los datos para que tengan una estructura adecuada, aquí utilizaremos datos ya formateados, por lo que no es necesario el tratamiento previo de los mismos.

Los primeros algoritmos aplicados fueron los de agrupamiento: k-means clustering algorithm y NaiveBayes algorithm. En el capítulo 1 se detalla más estos dos algoritmos.

La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional.

El fin era corroborar si el umbral de corte para los aspirantes aceptados y no aceptados era correcto o requería de algún ajuste.

Las instancias agrupadas del k-means clustering algorithm, así como la matriz de confusión del NaiveBayes algorithm se presentan en la Tabla 3.

k-means			NaiveBayes		
Clúster	Instancias			A	NA
A	2498		A	2183	315
NA	4621		NA	221	4400

Tabla 3.Instancias agrupadas y Matriz de confusión

Por otra parte, al aplicar los algoritmos de clasificación en Weka utilizamos Cross-validation, en donde Weka realiza una validación cruzada estratificada del número de particiones dado (Folds) [21,32].

La finalidad era dividir los datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir.

Los resultados obtenidos por los diferentes algoritmos se presentan en la Tabla 4.

Algoritmo	Clasificación Correcta(%)	Clasificación Incorrecta (%)	Kappa Statatistic (K)
DecisionStump	83.19%	16.82%	0.6482
J48	91.88%	8.12%	0.8212
NaiveBayes	92.47%	7.53%	0.8333
MultilayerPerceptron	92.64%	7.36%	0.8376
RandomForest	92.11%	7.89%	0.8260
RandomTree	89.56%	10.44%	0.7715

Tabla 4.Resultados de la ejecución de algoritmos

Como se puede apreciar, a partir de los resultados es posible establecer un ranking para determinar cuál de los algoritmos de clasificación basados en reglas, es más confiable.

La Tabla 4 también muestra los diferentes métodos de clasificación utilizados para crear los modelos y luego comparar el rendimiento con las medidas.

Como se mencionó anteriormente, en este trabajo se aplicó una división mediante una prueba de validación cruzada, Cross-validation (folds10), donde el conjunto de entrenamiento se utiliza para crear el modelo y el conjunto de prueba predice la clase, respectivamente.

La Figura 8 muestra el árbol generado, cada nodo-hoja se clasifica en dos clases: A (aceptado) y NA (no aceptado).

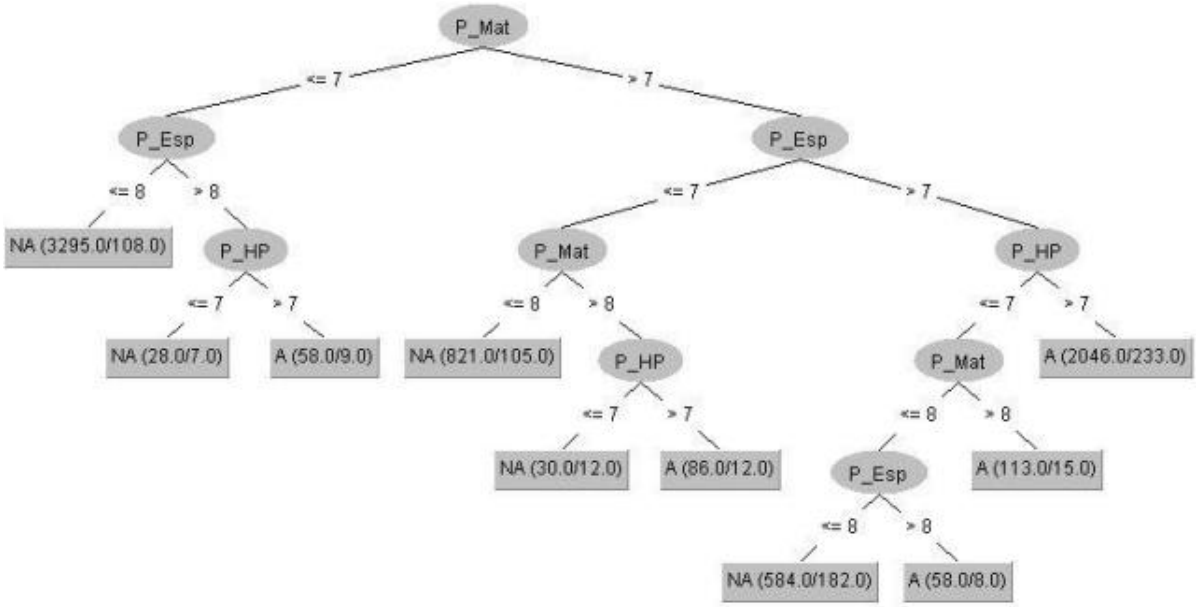

Figura 8. Árbol de clasificación

Una vez obtenidos los modelos, se extraen las medidas de desempeño utilizadas en este trabajo, que se pueden observar en la Tabla 5.

Algoritmo	Precision	Recall	F-measure	Accuracy
DecisionStump	0.847	0.832	0.835	0.834
J48	0.919	0.919	0.919	0.954
NaiveBayes	0.924	0.925	0.924	0.980
MultilayerPerceptron	0.926	0.926	0.926	0.980
RandomForest	0.921	0.921	0.921	0.973
RandomTree	0.896	0.896	0.896	0.887
k-means	1.000	1.000	1.000	1.000

Tabla 5.Desempeño de los modelos de clasificación

Capítulo 3

Análisis y Conclusiones

3.1 Análisis y Resultados

Este trabajo como se ha mencionado se está enfocado en el análisis de predicción del desempeño de los estudiantes que presentan un examen de admisión y se han aplicado varias técnicas de clasificación basadas en reglas que implementa el software Weka.

Un estudiante puede obtener un puntaje mínimo de 200 y un máximo de 1000 con una media de 600 puntos y una desviación estándar de 133.33.

Sin embargo, el ingreso de los aspirantes al nivel medio superior está condicionado por la infraestructura y recursos con los que cuenta la universidad, razón por la cual se establece un puntaje mínimo para poder ingresar (umbral de aceptación), en este caso es de 600 puntos y corresponde justamente a la media.

La aplicación de k-means clustering algorithm y NaiveBayes algorithm nos indica que el umbral de aceptación utilizado para determinar el corte en el puntaje de los alumnos que son aceptados y no aceptados es el correcto (Tablas 3 y 4).

De hecho, la clasificación dada por NaiveBayes algorithm proporciona una precisión del 92.47% y un error en falsos negativos y falsos positivos de apenas el 7.53%, siendo el algoritmo con mejor resultados.

Muy opuesto a lo que muestra el algoritmo DecisionStump, el cual tuvo una precisión de 83.19% y un error en falsos negativos y falsos positivos de 16.82%.

Por su parte, los algoritmos de clasificación, salvo DecisionStump y RandomTree, tienen un índice Kappa que supera el 0.8, esto es, muestran un grado de concordancia casi perfecto.

Además, el árbol de la Figura 8 se puede convertir en reglas que pueden utilizarse para predecir el desempeño de un nuevo estudiante al presentar el examen de admisión, simplemente recorriendo las ramas desde el nodo-raíz hasta llegar al nodo-hoja deseado.

A continuación, se describirá cada rama, recorriendo el árbol de forma preorden, iniciando en el nodo raíz P_Mat de izquierda a derecha.

1. Sí en P_Mat la calificación es menor o igual a 7 (7 o 6), y P_Esp es menor o igual a 8 (8, 7 o 6) el alumno no será aceptado.

2. Pero si en P_Esp obtiene calificación superior a 8 (9 o 10), la condición de aceptación recae en las habilidades del pensamiento P_HP. Si es menor o igual a 7 (7 o 6) no es aceptado, caso contrario si es mayor a 7 (8, 9 o 10).
3. Si en P_Mat la calificación es mayor a 7 (8, 9 o 10). Tenemos dos subárboles, donde P_Esp se convierte en el nodo raíz. Partiendo de este nodo.
4. Si P_Esp es menor o igual a 7 (7 o 6), y P_Mat es menor o igual a 8 (8, 7 o 6) el alumno no será aceptado.
5. Pero si en P_Mat es superior a 8 (9 o 10), la condición de ser aceptado recae en habilidades del pensamiento, si P_HP es menor o igual a 7 (7 o 6) el alumno no será aceptado, caso contrario si es mayor a 7 (8, 9 o 10).
6. Sin embargo, cuando P_Esp es superior a 7 (8, 9 o 10), hay dos subárboles donde P_HP se convierte en el nodo raíz. Partiendo de este nodo.
7. Si en P_HP es menor o igual a 7 (7 o 6), y en P_Mat es menor o igual a 8 (8, 7 o 6), la condición para ser aceptado recae en habilidades en español. Si P_Esp es menor o igual a 8 (8, 7 o 6) el alumno no será aceptado, caso contrario si en P_Esp es mayor a 8 (9 o 10).
8. Pero en el nodo P_HP es menor igual a 7 (7 o 6) y en P_Mat es mayor a 8 (9 o 10), es probable que el alumno sea aceptado.
9. Y el ultimo subárbol, donde P_Esp es mayor a 7 (8, 9 o 10) y e P_HP es mayor a 7 (8, 9 o 10) es probable que el alumno sea aceptado.

3.2 Conclusiones y Trabajo Futuro

Data mining se presenta como una tecnología innovadora, que ofrece una serie de beneficios: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas dedicadas a la educación; por otro, ahorra tiempo y recursos a las instituciones educativas, abriendo nuevas oportunidades en los procesos de aprendizaje de los estudiantes.

Además, no hay duda de que trabajar con esta tecnología implica cuidar un sin número de detalles debido a que el producto final involucra "toma de decisiones".

En este trabajo se ha se ha utilizado una alternativa de software de minería de datos, a saber, WEKA, que es una herramienta libre y muy interesante a la hora de aplicar diversas técnicas de minería de datos.

Además, es importante destacar, que con este trabajo se han utlizado los algoritmos de clasificación que tienen una confiablidad interesante, para poder aplicarlos en la predicción del desempeño académico de aspirantes a ingresar a cierto nivel de estudios.

Como se puede apreciar, a la hora de seleccionar un algoritmo de clasificación basada en reglas este trabajo concluye que los que tiene menos confianza son DecisionStump y RandomTree.

El trabajo desarrollado solo considera tres aspectos fundamentales en la formación de un estudiante del nivel básico: las habilidades, el lenguaje verbal y el lenguaje matemático.

Como se muestra en los análisis, estas tres áreas pueden predecir de forma casi perfecta el desempeño de los estudiantes en la prueba estandarizada que se emplea como examen de admisión al nivel medio superior.

Como se puede observar obtener una calificación baja en matemáticas (6 o 7) requiere un mejor desempeño en español (9 o 10) y en habilidades del pensamiento (8, 9 o 10) para tener posibilidades de ser aceptado. Caso contrario si la calificación obtenida es baja en habilidades del pensamiento (7 o 6) la posibilidad del alumno es no ser aceptado. Pero también el alumno puede no ser aceptado al obtener una calificación baja en español (7 o 6) y también una calificación baja en matemáticas (8, 7, 6).

Para que el alumno pueda tener posibilidades de ser aceptado se necesita que su calificación sea alta en matemáticas (9 o 10) y en una calificación alta en habilidades del pensamiento (8, 9, o 10). Si la calificación en habilidades del pensamiento es baja (7 o 6) el alumno posiblemente no será aceptado.

También el alumno puede ser rechazado al obtener calificación baja en habilidades del pensamiento (7 o 6) y en matemáticas una calificación baja (8, 7 o 6) y en español una calificación baja (8, 7 o 6). Pero si en español su calificación es alta (9 o 10) el alumno tiene posibilidad de ser aceptado a pesar de tener baja calificación en matemáticas y habilidades del pensamiento.

Pero si el alumno tiene una calificación baja en habilidades del pensamiento (7 o 6) pero alta calificación en matemáticas (9 o 10) el alumno tiene posibilidad de ser aceptado sin importar la calificación en español.

Al igual que si en español tiene una calificación alta (8, 9 o 10) y en habilidades del pensamiento alta (8, 9 o 10) el alumno tiene posibilidades de ser aceptado sin importar la calificación en matemáticas.

Y con esto agregaría que, si un alumno no es bueno en matemáticas, aun así, puede ser aceptado si sus habilidades en español y sus habilidades del pensamiento son altas. Pero si el alumno no es bueno ni en matemáticas ni en español sus posibilidades de ser aceptado bajan sin importar la calificación obtenida en habilidades del pensamiento.

Recordemos que hay otras asignaturas que un estudiante de nivel básico lleva al largo de su formación y que no fueron consideradas para este trabajo, las ciencias naturales (física, química y biología) y las ciencias sociales (formación cívica, historia y geografía). Una de las preguntas que surge es cómo estas asignaturas afectan (positiva o negativamente) la predicción del desempeño en la prueba estandarizada.

Otra área de oportunidad es investigar qué tanto una prueba estandarizada como la que se aplica en el examen de admisión predice el éxito académico de los estudiantes y minimiza la deserción escolar, sobre todo después de haber pasado por una pandemia.

Agradecimientos

Uno de los valores que me enseñaron es el decir gracias.

Por eso en este trabajo voy a agradecer a los sinodales por tomarse la molestia en leer y valorar el trabajo desarrollado.

También quiero agradecer a mis profesores que me trasmitieron su conocimiento tanto a académico como la preparación para el mundo laboral. Y que gracias a este conocimiento puedo desarrollarme en un mundo laboral competitivo. Gracias por la paciencia al momento de resolver mis dudas o de explicar de nuevo un tema en específico.

Bibliografía

[1] G. Piotesky-Shapiro and W. J. Frawley, Knowledge Discovery in Database, Cambridge, MA: AAA/MIT Press, (1991).

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

[3] Gil, H. S.; Rao, P. C. (1996). The Official Client / Server Computing Guide to Data Warehousing. Que Pub., Simon & Schuster Company. Grabmeier, J.; Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery, 6, 303-360.

[4] David L. La Red Martínez1, Marcelo Karanik1, Mirtha Giovannini1, María E. Báez1, Juliana Torre. Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. Universidad Tecnológica Nacional, Argentina http://uajournals.com/ojs/index.php/campusvirtuales/article/view/128/132

[5] Pimpa, C. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program, Proceedings of the International MultiConference of Engineers and Computer Scientists: http://www.iaeng.org/publication/IMECS2013/IMECS2013_pp332-336.pdf

[6] Bedregal-Alpaca, Norka; Aruquipa-Velazco, Danitza; Cornejo-Aparicio, Víctor. Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. Revista Ibérica de Sistemas e Tecnologias de Informação; Lousada N.º E27, (Mar 2020): 592-604. https://www.proquest.com/openview/06e5642b7afb32ac6c577d98eedb233d/1?pq-origsite=gscholar&cbl=1006393

[7] M. Al-Razgan, A. S. Al-Khalifa and M. S. Al-Khalifa, Educational data mining: A systematic review of the published literature 2006-2013, Proceeding the 1st International Conference on Advance Data and Information Engineering, 711-719, (2013).

[8] Pereira, R. T., Romero, A. C., & Toledo, J. J. (2013). DESCUBRIMIENTO DE PERFILES DE DESERCIÓN ESTUDIANTIL CON TÉCNICAS DE MINERÍA DE DATOS. Revista vínculos, 10(1), 373–383. https://doi.org/10.14483/2322939X.4687

[9] M. Durairaj and C. Vijitha, Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp. 5987-5991, (2014).

[10] L. S. Affendey, I.H.M. Paris , N. Mustapha, Md. Nasir Sulaiman and Z. Muda, "Ranking of Influencing Factors in Predicting Students' Academic Performance," International Technology Journal, vol. 9, no. 6 , pp. 832-837, 2010. ISBN 1812-5638

[11]Chamillard, A.T., 2006. Using student performance predictions in a computer science curriculum. Proceeding of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, June 26-28, Bologna, Italy, pp: 260-264

[12]Superby, J.F., J.P. Vandamme and N. Meskens, 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS`06), Jhongali, Taiwan, pp: 37-44

[13]Vandamme, J.P., N. Meskens and J.F. Superby, 2007. Predicting academic performance by data mining methods. Educ. Econ., 15: 405-419.

[14]Golding, P. and O. Donaldson, 2006. Predicting academic performance. Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference T1D-21, Oct. 28-31, San Diego, CA., pp: 1-6

[15] McKenzie, K. and R. Schweitzer, 2001. Who succeeds at University? Factors predicting academic performance in first year Australian university students. Higher Educ. Res. Dev., 20: 21-33.

[16] Merceron, A. and K. Yacef, 2005. Educational data mining: A case study. http://www.it.usyd.edu.au/~kalina/publis/merceron_yacef_aied05.pdf.

[17] Kotsiantis, S.B., C.J. Pierrakeas and P.E. Pintelas, 2003. Preventing student dropout in distance learning using machine learning techniques. Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Oct. 21, Springer Berlin, Heidelberg, pp: 267-274

[18] Minaei-Bidgoli, B., 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. Proceedings of the 33rd Annual Conference on Frontiers in Education, Nov. 5-8, IEEE Computer Society, Washington, DC, USA., pp: 13-18

[19] Minaei-Bidgoli, B., G. Kortemeyer and W.F. Punch, 2004. Enhancing online learning performance: An application of data mining method. Proceedings of the 7th IASTED

International Conference on Computers and Advanced Technology in Education, August 2004, Kauai, Hawaii, USA., pp: 173-178

[20] Hamalainen, W. and M. Vinni, 2006. Comparison of machine learning methods for intelligent tutoring systems. Proceedins of the 8th International Conference on Intelligent Tutoring Systems, June 2006, Jhongli, Taiwan, pp: 525-534

[21] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann e Ian H. Witten (2009). El software de minería de datos WEKA: una actualización. Exploraciones SIGKDD, volumen 11, número 1. https://www.kdd.org/exploration_files/p2V11n1.pdf https://www.cs.waikato.ac.nz/ml/weka/ , https://waikato.github.io/weka-wiki/downloading_weka/ , https://waikato.github.io/weka-wiki/requirements/

[22] Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications. 2014; 98(22):1–5

[23]Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013; 6(2):256–61

[24] MacQueen, J. B., Some methods for classification y analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281-297 (1967).

[25]Hartigan, J. A.; y Wong, M. A., Algorithm AS 136: A k-medias clustering clgorithm, https://doi.org/10.2307/2346830, Journal of the Royal Statistical Society, Series C (Applied Statistics), 28(1), 100-108 (1979).

[26] Valenzuela-Keller, Andrés A., Gálvez-Gamboa, Francisco A., Contreras, David R., & Parraguez, Felipe P.. (2021). Análisis del perfil emprendedor para la formación de las nuevas generaciones de jóvenes chilenos. Información tecnológica, 32(1), 209-216. https://dx.doi.org/10.4067/S0718-07642021000100209

[27] Jiménez-Carrión, Miguel, Sánchez-Candela, Luis, Keewong-Zapata, Roxani, & Bazán, José. (2020). Optimización de las rutas para la intervención de pozos de petróleo. Información tecnológica, 31(4), 71-84. https://dx.doi.org/10.4067/S0718-07642020000400071

[28] Youn, S., & McLeod, D. (2007). A comparative study for email classification. In Advances and innovations in systems, computing sciences and software engineering (pp. 387-391). Springer, Dordrecht.

[29]M. Durairaj and C. Vijitha, Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp. 5987-5991, (2014).

[30] Cohen, J., A coefficient of agreement for nominal scales, Educational and Psychological Measurement, Vol. 20(1), 37-46, (1960).

[31] Landis, J. R., and Koch, G. G., The measurement of observer agreement for categorical data, Biometrics, Vol. 33(1), 159-174, (1977).

[32] Eibe Frank, Mark A. Hall, and Ian H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition, (2016).