PAPEL • **ACCESO ABIERTO**

Modelo de Predicción del Rendimiento Estudiantil basado en Algoritmos de aprendizaje automático supervisado

Para citar este artículo: Ali Salah Hashim et al 2020 IOP Conf. Ser.: Mater. ciencia Ing. **928** 032019

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

Modelo de Predicción del Desempeño Estudiantil basado en Supervisado Algoritmos de aprendizaje automático

Ali Salah Hashim1
[1]*Facultad de Ciencias de la Computación y Tecnología de la Información / Universidad de Basora, Basora, Irak.*
alishashim2009@gmail.com

Wid Akeel Awadh2
[2]*Facultad de Ciencias de la Computación y Tecnología de la Información / Universidad de Basora, Basora, Irak.*
umzainali@gmail.com

Alaa Khalaf Hamoud3
[3]*Facultad de Ciencias de la Computación y Tecnología de la Información / Universidad de Basora, Basora Irak*
Alaak7alaf@gmail.com

Abstracto

Las instituciones de educación superior tienen como objetivo pronosticar el éxito de los estudiantes, que es un tema de investigación importante. Pronosticar el éxito de los estudiantes puede permitir a los profesores evitar que los estudiantes abandonen los estudios antes de los exámenes finales, identificar a aquellos que necesitan ayuda adicional y aumentar la clasificación y el prestigio de la institución. Las técnicas de aprendizaje automático en la minería de datos educativos tienen como objetivo desarrollar un modelo para descubrir patrones ocultos significativos y explorar información útil de entornos educativos. Las características tradicionales clave de los estudiantes (demografía, antecedentes académicos y características de comportamiento) son los principales factores esenciales que pueden representar el conjunto de datos de entrenamiento para los algoritmos de aprendizaje automático supervisado. En este estudio, comparamos el rendimiento de varios algoritmos de aprendizaje automático supervisado, como Árbol de decisión, Naïve Bayes, Regresión logística, Máquina de vectores de soporte, K-Vecino más cercano, Optimización mínima secuencial y Red neuronal. Entrenamos un modelo utilizando conjuntos de datos proporcionados por cursos en los programas de estudio de licenciatura de la Facultad de Ciencias de la Computación y Tecnología de la Información de la Universidad de Basora, para los años académicos 2017-2018 y 2018-2019 para predecir el rendimiento de los estudiantes en los exámenes finales. Los resultados indicaron que el clasificador de regresión logística es el más preciso para predecir las calificaciones finales exactas de los estudiantes (68,7% para aprobado y 88,8% para reprobado).

Palabras clave: aprendizaje automático supervisado, minería de datos educativos, árbol de decisión, Naive Bayes, Regresión logística, K-vecino más cercano, perceptrón multicapa, red neuronal

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

1. Introducción

El rápido desarrollo de la tecnología de la información (TI) ha aumentado considerablemente la cantidad de datos en diferentes instituciones. Los enormes almacenes contienen una gran cantidad de datos y constituyen una valiosa mina de oro de información. Esta dramática inflación en la cantidad de datos en las instituciones no ha seguido el ritmo de las formas eficientes de invertir estos datos. Así, recientemente ha surgido un nuevo desafío, es decir, pasar de las bases de datos tradicionales que almacenan y buscan información solo a través de preguntas formuladas por un investigador a técnicas utilizadas en la extracción de conocimiento mediante la exploración de patrones de datos predominantes para la toma de decisiones, la planificación y la visión de futuro. Una de estas técnicas es la tecnología de minería de datos (DM) [1].

DM descubre correlaciones útiles entre atributos, tendencias y patrones ocultos mediante el análisis de grandes cantidades de conjuntos de datos almacenados en almacenes. También se utiliza como técnica de reconocimiento de patrones y método matemático y estadístico para reducir costos y aumentar los ingresos. Además, DM es un campo de descubrimiento de conocimiento en bases de datos [2].

El proceso de gestión de las instituciones educativas es una de las dificultades que enfrentan los administradores debido a la complejidad de la estructura de datos, las múltiples fuentes y el gran tamaño de los datos.
Las instituciones educativas se enfrentan a muchos otros problemas administrativos, financieros y educativos al gestionar los procedimientos educativos. Todos estos problemas deben ser analizados para generar recomendaciones y conclusiones que apoyen a los decisores en la toma de decisiones para la coordinación y gestión del proceso educativo [3, 4].

La minería de datos educativos (EDM) es un DM utilizado en instituciones educativas y académicas. Está orientado a la teoría y tiene como objetivo desarrollar enfoques computacionales que combinen teoría y datos para ayudar y mejorar la calidad del rendimiento académico de los estudiantes y graduados y la información de la facultad de estas instituciones [5, 6]. EDM utiliza diferentes técnicas, como árboles de decisión (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Association Rule Mining y Neural Networks. Muchos tipos de conocimiento, como la predicación, las reglas de asociación, las clasificaciones y los agrupamientos, se pueden descubrir utilizando estas técnicas.

EDM es una herramienta útil para las instituciones académicas. Las universidades pueden usar EDM para predecir qué estudiantes aprobarán o reprobarán y tendrán un bajo rendimiento educativo, para saber quién aprobará los exámenes en materias particulares y para obtener la proporción de graduados. EDM también se utiliza para otra información estratégica. Estas universidades pueden entonces desarrollar y mejorar sus políticas educativas para ayudar a los estudiantes reprobados a elevar su nivel educativo o guiarlos hacia especializaciones que se adapten a sus preparaciones, preferencias y habilidades. Las medidas y políticas mejoradas resultantes de EDM se pueden utilizar para mejorar el rendimiento académico de las instituciones [7, 8].

Además, EDM es un área de investigación común para explorar datos de campos educativos mediante el uso de técnicas de DM y enfoques de aprendizaje automático [9]. La investigación sobre el aprendizaje automático tiene como objetivo aprender a reconocer automáticamente patrones ocultos complejos y crear datos inteligentes para la toma de decisiones [10, 11]. La ruta predictiva de DM es un proceso especial de DM que realiza predicciones en

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

datos actuales [12]. Obtener una máquina para adaptar la acción es el principal interés en el enfoque de aprendizaje automático, y este interés puede mejorar la precisión de ciertas acciones o experiencias.
La expectativa en el enfoque de clasificación es que las computadoras deben aprender a clasificar técnicas de ejemplos de observación, mientras que en el enfoque de regresión, la salida debe continuar teniendo una cantidad numérica en lugar de una cantidad discreta [13]. El aprendizaje automático supervisado se utiliza para resolver problemas de clasificación y regresión [14].

Este documento discutió la efectividad de los algoritmos de aprendizaje automático supervisado en la predicción del éxito y el rendimiento académico de los estudiantes en la educación superior. Específicamente, los algoritmos de aprendizaje automático supervisado midieron el rendimiento de los estudiantes en función de la calificación o el estado real (aprobado o reprobado). Se aplicaron diferentes algoritmos de aprendizaje automático supervisado y se evaluaron los criterios de rendimiento. Los experimentos mostraron que el algoritmo clasificador de regresión logística se desempeñó mejor. Se utilizó el entorno Waikato para el análisis del conocimiento (Weka) 3.8.0 (un entorno de software DM de código abierto) para implementar los algoritmos de aprendizaje automático supervisado.

2. Trabajos relacionados

Se revisaron varios trabajos que utilizaron algoritmos de aprendizaje automático, como DT, NB, Logistic Regression, Support Vector Machine (SVM), KNN, Sequential Minimal Optimization (SMO) y Neural Network, para predecir los resultados de los estudiantes. Los detalles se muestran en la Tabla 1.

Tabla 1. Revisión de la literatura

No.	*Algoritmo*	*Referencia*
1	DT	[12], [13], [15], [16]
2	nota *3*	[14], [15], [16], [17]
	Regresión logística	[15], [16] [15], [16]
4	MVS	[14] [17] [14], [15],
5	KNN	[16], [17]
6	SOMOS	
7	Red neuronal	

S. Natek y M. Zwilling [15] realizaron un estudio sobre DM con conjuntos de datos de estudiantes de pequeño tamaño comparando dos métodos diferentes de DM. Sus conclusiones fueron positivas y revelaron que la integración de herramientas de DM es una parte importante de los sistemas de gestión de información en las instituciones de educación superior (IES). El conjunto de datos contenía tres años de datos: 2010-2011 (42 estudiantes), 2011-2012 (32 estudiantes) y 2012-2013 (32 estudiantes). Los datos recopilados cubrieron varios aspectos de las historias de los estudiantes, incluidos los registros académicos anteriores, los antecedentes familiares y la demografía. Se aplicaron tres clasificadores, a saber, los modelos Rep Tree, J48 y M5P, para obtener el rendimiento académico de los estudiantes. Los experimentos mostraron que J48 era menos preciso pero más sensible que Rep Tree. Sin embargo, la cantidad de clasificadores utilizados para comparar el desempeño de los estudiantes fue menor que la cantidad de algoritmos en los enfoques de aprendizaje automático supervisado y no supervisado.

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

Alaa Khalaf et. Alabama. [16] utilizaron Weka para evaluar el desempeño de los estudiantes universitarios y obtener los factores que afectan el éxito/fracaso de los estudiantes. Se escribieron un total de 161 cuestionarios en formularios de Google y se utilizó una aplicación de código abierto (LimeSurvey) para realizar una encuesta a estudiantes en la Facultad de Ciencias de la Computación y Tecnología de la Información de la Universidad de Basora. Los autores utilizaron técnicas de clasificación (J48, Random Tree y Rep Tree) en los cuestionarios cumplimentados por los estudiantes. En términos de precisión, J48 superó a los otros dos. Los DT utilizados en este documento produjeron resultados sobresalientes y precisos, pero muchos otros campos del aprendizaje automático pueden lograr resultados de predicción más precisos. Además, el modelo solo predijo el estado de los estudiantes como "aprobado o reprobado" y no predijo sus calificaciones reales.

Erman Yukselturk et. al [17] se centró en identificar a los estudiantes que abandonaron la escuela mediante el uso de enfoques de DM en una aplicación en línea. Aplicaron cuatro enfoques de DM, a saber, KNN, DT, NB y Neural Network. KNN se desempeñó mejor entre todos los clasificadores, con un 87 % de precisión.
Sin embargo, el modelo solo examinó cuatro algoritmos para predecir las deserciones y no las calificaciones reales de los estudiantes.

Estudios anteriores analizaron el rendimiento de cinco algoritmos populares de aprendizaje automático que clasifican a los estudiantes en riesgo por adelantado y predicen las dificultades que enfrentan en la educación superior a distancia [18, 19]. Estos algoritmos fueron Redes Neuronales Artificiales (ANNs), SVM, Regresión Logística, clasificadores NB y DTs. Las ANN y SVM son más precisas (57 %) cuando solo usan datos demográficos que otros algoritmos [18], mientras que NB tiene una precisión adecuada pero no tan prometedora como los otros modelos [19].

Acharya y Sinha [20] utilizaron el aprendizaje automático para predecir el rendimiento de los estudiantes. Las características de entrada en su estudio incluyeron género, ingresos, notas en la pizarra y asistencia. Las técnicas aplicadas fueron C4.5, SMO, NB, 1-Nearest Neighborhood y Multi-layer Perceptron (MLP). Los investigadores revelaron que SMO es ideal para mejorar el rendimiento del modelo para todos los estudiantes en un curso, con una precisión de prueba promedio más alta (66 %) que otros enfoques. Sin embargo, la precisión es menos sobresaliente que el rendimiento de otros modelos.

3. electroerosión

Este término se extendió durante el primer taller sobre el concepto de EDM en 2005, y este taller se ha convertido en una conferencia internacional en Montreal desde 2008 [21]. Las sociedades periódicas han mostrado interés en publicar las últimas investigaciones sobre EDM. Las sociedades más populares creadas en 2011 y 2012 son la International EDM Society (http://www.educationaldatamining.org/) y el IEEE Task Force of EDM (http://datamining.it.uts.edu.au/edd/), respectivamente. EDM utiliza métodos DM para estudiar los datos extraídos de los sistemas educativos (estudiantes e instructores) y analizar los procesos de aprendizaje de los estudiantes en las instituciones educativas.

Los métodos EDM a menudo tienen múltiples niveles de jerarquías significativas, que a menudo deben decidirse sobre la base de las propiedades de los datos, en lugar de por adelantado, ya sea que se tomen de la universidad.

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

datos administrativos, datos de aprendizaje colaborativo basados en computadoras o uso de entornos de aprendizaje interactivo por parte de los estudiantes [22]. De manera similar a los métodos tradicionales de DM, EDM debe identificar el objetivo principal del estudio y los datos requeridos, extraer datos del entorno educativo, preprocesar los datos (limpiar y organizar una selección de técnicas que se pueden aplicar), interpretar los resultados y verificar las técnicas aplicadas. Los objetivos y técnicas utilizadas en EDM se derivan de la especificidad del entorno educativo y el propósito de la exploración [23]. Las aplicaciones que adoptan EDM siguen varios pasos, como se muestra en la Figura 1.

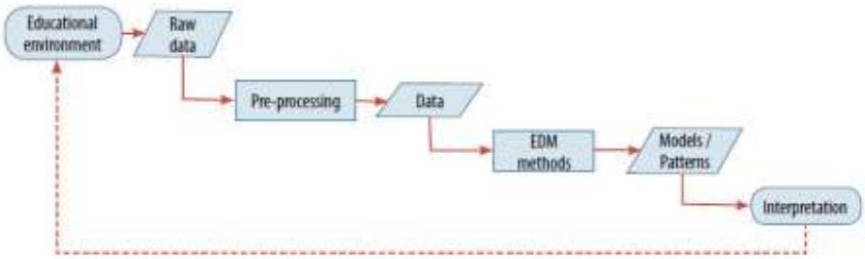


Figura 1. Diagrama de flujo de implementación del modelo [24]

4. Herramienta Weka

Una de las aplicaciones de aprendizaje automático más comunes es Weka, que es una herramienta escrita y desarrollada en lenguaje Java en la Universidad de Waikato, Nueva Zelanda. Weka es un software gratuito de código abierto bajo la licencia pública general GNU. El banco de trabajo Weka proporciona una colección de algoritmos y herramientas para analizar datos e implementar modelos predictivos.

Weka proporciona una interfaz gráfica de usuario para facilitar el acceso y realiza todos los algoritmos de DM [25].

La versión de Weka que no es Java se desarrolla utilizando TCL/TK, con algoritmos de modelado y preprocesamiento de datos utilizando lenguaje C con un sistema basado en Makefile para ejecutar experimentos de aprendizaje automático. Weka se usa originalmente para analizar los datos de los dominios agrícolas, mientras que la versión Weka Java lanzada en 1997 se usa en muchos campos, como los campos educativos y de investigación [26].

Weka admite diferentes tareas estándar de DM, como preprocesamiento de datos, visualización de datos, agrupación, clasificación, regresión y selección de funciones. Todas las técnicas de Weka se basan en la suposición de que se puede acceder a los datos como un único archivo plano o relación, donde cada punto de datos está representado por un número fijo de atributos (numéricos, nominales o algún otro tipo de atributos). Weka puede acceder a muchos archivos de conjuntos de datos a través de conexiones o puertas de enlace utilizando la conectividad de la base de datos Java, la base de datos SQL y los valores separados por comas y muchos otros tipos de conjuntos de datos.

El DM multirelacional es imposible en Weka, donde se debe usar un software separado para convertir una colección de tablas de bases de datos vinculadas en una sola tabla adecuada para el procesamiento de Weka. El modelado de secuencias es otra área importante que actualmente no están cubiertas por los algoritmos de distribución de Weka [27, 28].

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

5. Técnicas de aprendizaje automático supervisado

El campo del aprendizaje automático ha llamado la atención de los investigadores de informática y TI. El campo del análisis de datos se ha vuelto más esencial que antes, debido a la creciente cantidad de datos enormes que se procesan todos los días. Los tres tipos básicos de aprendizaje automático son aprendizajes supervisados, no supervisados y semisupervisados [29]. En el aprendizaje supervisado, el conjunto de datos de entrenamiento solo consta de datos etiquetados. Se entrena una función supervisada durante el proceso de aprendizaje, con el objetivo de predecir las etiquetas futuras de los datos no vistos. Los dos problemas supervisados básicos son la regresión y la clasificación, especialmente para la clasificación de funciones discretas y la regresión continua [30]. El aprendizaje no supervisado tiene como objetivo encontrar patrones regulares significativos sin intervención humana en datos no etiquetados. Su conjunto de entrenamiento se compone de datos no etiquetados y no hay ningún instructor presente para ayudar a identificar estos patrones. Algunos métodos supervisados populares incluyen el agrupamiento, la identificación de novedades y la reducción de la dimensionalidad [4, 31]. El aprendizaje semisupervisado es una combinación de procesos de aprendizaje supervisados y no supervisados. Se utiliza para lograr resultados mejorados con pocos ejemplos etiquetados. Su conjunto de datos de entrenamiento consta de datos etiquetados y no etiquetados. DT, NB, Logistic Regression, SVMs, KNN, SMO y Neural Network son técnicas supervisadas bien conocidas con resultados precisos en diferentes campos científicos [8, 28, 32, 33].

5.1 DT es un algoritmo de aprendizaje automático supervisado que utiliza una metodología de ramificación para mostrar todos los resultados posibles de una decisión de acuerdo con ciertos parámetros. La estructura de árbol consta de conjuntos de reglas organizadas jerárquicamente, comenzando con atributos raíz y terminando con nodos hoja; cada rama del árbol representa uno o más resultados del conjunto de datos original [32, 33]. El nodo raíz es el nodo superior del árbol sin ramas entrantes, y todas las ramas salientes representan todas las filas según el conjunto de datos. El nodo interno del árbol es el nodo con ramas entrantes y salientes y se puede utilizar para probar el atributo. El nodo terminal o la hoja es el nodo descendente con solo una rama entrante. Este nodo representa el nodo final del árbol, que puede tener muchos nodos hoja que representan los cálculos finales [34].

5.2 NB es un algoritmo construido sobre la base del teorema de Bayes. Esta hipótesis es formulada por Thomas Bayes. Este modelo es fácil de construir y se usa principalmente para conjuntos de datos muy grandes [35]. NB tiene como objetivo calcular el proceso de distribución de probabilidad condicional de cada característica. La probabilidad condicional de que un vector se clasifique en la clase C es igual al producto de probabilidad de cada una de las características del vector en la clase C. Este algoritmo se denomina "ingenuo" debido a sus suposiciones básicas de independencia condicional. Se cree que todas las características de entrada son independientes entre sí. Si la suposición condicional de independencia realmente se cumple, entonces un clasificador NB puede converger más rápido que otros modelos, como la regresión logística [36, 37].

5.3 La regresión logística se utiliza principalmente para analizar y explicar la relación entre una variable binaria (p. ej., 'aprobado' o 'fallido') y una serie de variables previstas [38]. Su objetivo es encontrar el mejor modelo que se ajuste para explicar la relación entre los conjuntos de variables dependientes e independientes. La regresión logística se desarrolla junto con la regresión lineal, pero difieren en la respuesta variable binaria y variable continua [39].

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

5.4 SVM se basa en el principio de aprendizaje teórico de Vapnik. SVM encarna el concepto de minimización del riesgo sistémico [40]. Las SVM se han aplicado a muchos campos de regresión, clasificación y detección de valores atípicos. El espacio de entrada original en una SVM se mapea a través de un kernel en un espacio de producto de puntos de alta dimensión. El nuevo espacio se denomina espacio de características, donde se define un hiperplano óptimo para optimizar la capacidad de generalización. Unos pocos puntos de datos llamados vectores de soporte pueden decidir el hiperplano óptimo. Una SVM puede proporcionar un fuerte resultado de generalización para problemas de clasificación, aunque no implementa el conocimiento del dominio del problema [41].

5.5 KNN es un algoritmo de aprendizaje automático simple donde un objeto es calificado por sus vecinos Voto mayoritario. El objeto se asigna a la clase más común entre sus vecinos más cercanos. K representa un número positivo y suele ser pequeño. Si k es igual a 1, entonces el objeto se asigna a la clase de su vecino más cercano. Elegir k como un número impar en problemas de clasificación binaria (dos clases) es bueno para eliminar los votos empatados. Seleccionar el parámetro k en este algoritmo puede ser importante [34, 42, 43].

5.6 SMO es un nuevo algoritmo de entrenamiento para SVM. En 1998, John Platt propuso un método fácil y rápido llamado algoritmo SMO para entrenar una SVM. La idea clave es resolver el problema de la optimización cuadrática dual optimizando el subconjunto mínimo en cada iteración, incluidos dos componentes. SMO separa el gran problema de la programación cuadrática en una colección de problemas más pequeños resueltos analíticamente. Cuando SMO se gestiona con una pequeña cantidad de conjuntos de entrenamiento, la cantidad de memoria necesaria es lineal. Dado que se evita el cálculo matricial, SMO escala en algún lugar entre lineal y cuadrático en el tamaño del conjunto de entrenamiento, mientras que el algoritmo SVM de fragmentación regular escala en algún lugar entre lineal y cúbico. Por lo tanto, SMO es el más rápido entre SVM lineales [43, 44].

5.7 Red neuronal es otra técnica común utilizada en EDM. Una red neuronal multicapa se compone de varias unidades (neuronas) unidas entre sí en un patrón. Las unidades de una red se dividen en tres clases: unidades de entrada, de salida y ocultas [8, 45, 46]. El beneficio de la red neuronal es su capacidad para detectar todas las interacciones posibles entre las variables. También puede realizar una detección completa sin ninguna duda, incluso en la relación no lineal entre variables dependientes e independientes [47].

6. Metodología

Este estudio utilizó varios algoritmos de aprendizaje automático supervisado para predecir el rendimiento académico de los estudiantes en sus exámenes finales, y se compararon los resultados. La metodología propuesta constó de dos etapas. La primera etapa involucró el preprocesamiento de datos, en el cual los datos se preparan, consolidan y limpian para prepararse para la segunda etapa. La segunda etapa involucró el rendimiento de clasificación del algoritmo más común y frecuentemente utilizado por la técnica de aprendizaje automático mencionada.

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

6.1 Preprocesamiento de datos

La adquisición de datos se centró en los cursos para programas de estudio de licenciatura en la Facultad de Ciencias de la Computación y Tecnología de la Información (CSIT), Universidad de Basora para los años académicos 2017-2018 y 2018-2019. Los datos fueron importados del Sistema de Comité de Examen a las herramientas de tabla de Microsoft (MS) Excel con el complemento DM, que se instaló en una computadora portátil.

El primer paso en el preprocesamiento de datos es preparar los datos eliminando registros con valores vacíos y convirtiendo los datos para su procesamiento. Un total de 50 registros con valores vacíos estaban en una o más columnas. Después de eliminar estos registros, se obtuvieron un total de 499 registros. Luego, los valores de registro se convirtieron para el procesamiento de datos en Weka 3.8 con sus clasificadores integrados. La muestra de investigación (499 estudiantes) fue una muestra aceptable de población CSIT con un margen de error del 10% [48]. El segundo paso implica medir la consistencia del conjunto de datos al encontrar el alfa de Cronbach [49, 50], como se muestra en la tabla 2. La fórmula se calcula de la siguiente manera:

$$= (1 \quad \ddot{y} \quad) \ldots\ldots. (1)$$

Donde k representa el número de elemento, es la varianza del i-ésimo elemento y representa la varianza puntuada total de todos los elementos del conjunto de datos.

Tabla 2. Alfa de Cronbach del conjunto de datos

Número de características	*Suma de características variaciones*	*Suma de todos los registros' variaciones*	*Alfa de Cronbach*
8	38	155.3452	0.863295

El alfa de Cronbach calculado (0,86) mostró una consistencia muy satisfactoria y una alta confiabilidad interna entre los elementos del conjunto de datos. Los conjuntos de datos del modelo son los pasos clave para crear el modelo DM con las siguientes columnas de conjuntos de datos de estudiantes enumerados en la Tabla 3. La Tabla 3 muestra el conjunto de datos de los estudiantes, donde los atributos tomados son el número de estudiantes (1–499), año de estudio (2017–2018, 2018–2019), género (femenino o masculino), año de nacimiento de los estudiantes (por ejemplo, 1997), registro (primero o repetido), empleo (sí o no), puntos de actividad (0–50), puntos de examen (0–50) y puntos finales (0–50). La última columna se estableció como el atributo predecible que es la calificación ('F', 'P', 'M', 'G', 'V' y 'E').

Tabla 3. Ejemplos de conjuntos de datos de estudiantes

Estudiante Número	*Estudio Año Género Nacimiento Año*	*Matrícula Curso Empleo Actividad*			*Punto (40)*	*examinar ación Punto (60)*	*Final Punto (100)*	*Calificación*
1	2016–2017 Mujer 1997	1	P1	sí	32	31	63	METRO
2	2016–2017 Mujer 1997	1	P1	sí	20	19	39	F
3	2016–2017 Mujer 1997		P1	No	20	8	28	F
4	2016–2017 Mujer 1997	1 1	P1	sí	15	6	21	F
5	2016–2017 Mujer 1997	1	P1	No	diecisiete	6	22	F

6	2016–2017 Mujer 1997				P1	No	34	26	60	METRO
7	2016–2017 Mujer 1997			1 1	P1	No	26	13	39	F
8	2016-2017	Masculino	1997	1	P1	No	34	14	48	F
9	2016-2017	Masculino	1997	1	P1	No	43	49	92	Y
.										
.										
499	2017–2018 Mujer 1998			1	P2	No	25	13	38	

La tecnología DM fue seleccionada para el siguiente paso. MS proporciona tres opciones analíticas para el nivel de DM. Los niveles básico, intermedio y experto incluyen herramientas de tablas de MS Excel, características complementarias de MS Excel DM y capacidades de MS SQL Server DM, respectivamente. En esta investigación se seleccionó el nivel básico.

6.3 Evaluación de atributos

Se utilizaron cuatro criterios para medir la eficiencia de los siete algoritmos y el rendimiento de los algoritmos de aprendizaje automático supervisado del modelo. Estos criterios fueron la tasa de verdaderos positivos (TP), la tasa de falsos positivos (FP), la precisión y los atributos de recuperación. Las ecuaciones (2)–(5) muestran estos atributos. La tasa de TP (a veces llamada sensibilidad) indica la proporción del número de predicciones verdaderas en la predicción positiva:

= = — ———, …… (2)

Donde TP y FN son los números de errores verdaderos detectados y no detectados, respectivamente. La tasa de FP (a veces llamada especificidad) es la proporción del número de negativos esperados:

= = — ——— …….. (3)

La precisión es el porcentaje de coincidencias de TP completas de todas las coincidencias de TP:

Precisión = ——— …….. (4)

Si la precisión es cercana a uno, las expectativas se vuelven poco a poco precisas. Recall es el porcentaje de coincidencias de TP de todas las posibles coincidencias positivas:

recordar = ——— ……... (5)

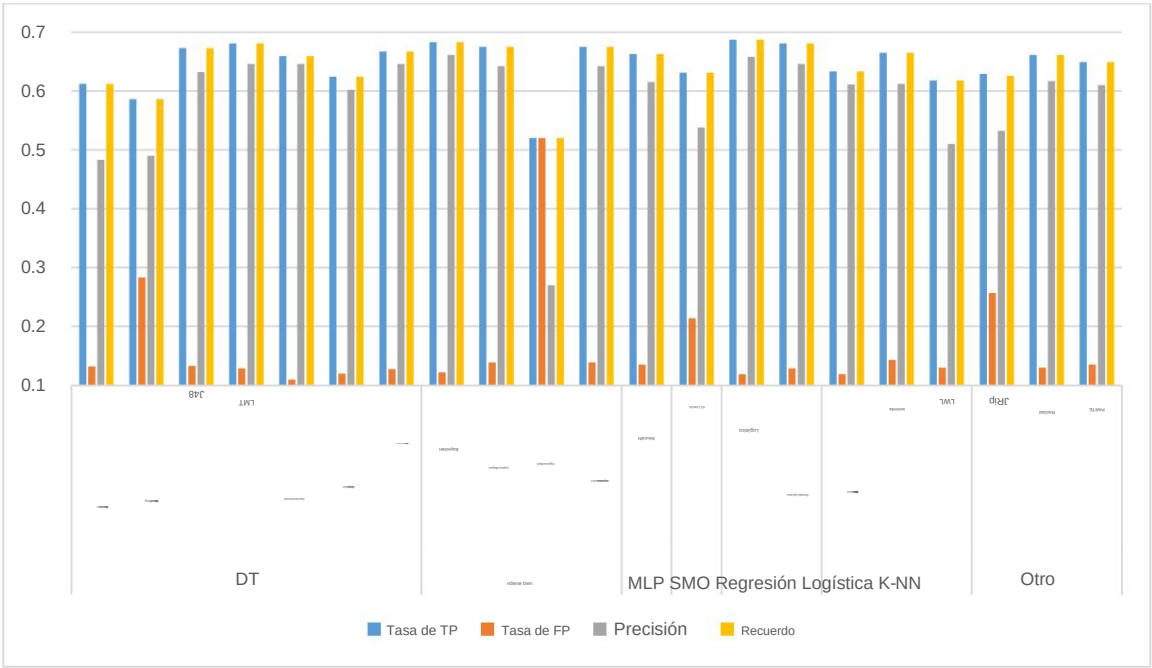
7. Resultado

Esta investigación exploró la posibilidad de predecir la calificación exacta, el éxito y el fracaso de los estudiantes a partir de diferentes variables de entrada obtenidas en las IES. El modelo se desarrolló utilizando varios algoritmos de aprendizaje automático supervisado y se compararon los resultados. Weka se instaló y cargó en estos algoritmos. Los clasificadores se usaron para probar opciones cruzadas.

validación, y el tamaño de los datos fue de 499, que se dividió en un 70 % de datos de entrenamiento (349 instancias) y un 30 % de datos de prueba (150 instancias) para implementar todos los algoritmos. La Tabla 4 enumera los criterios de desempeño de diferentes algoritmos de aprendizaje automático supervisado después de implementar el modelo para predecir la calificación final exacta de los estudiantes. El clasificador de regresión logística fue el más preciso (66 %) entre otros algoritmos.

Tabla 4. Criterios de rendimiento del modelo de predicción de calificaciones reales

No.	*Categoría*	*Algoritmo*	*Tasa TP*	*Tasa FP* 0,612	*Recuperación de*	
					precisión 0,483 0,612	
1	DT	Tocón de decisión	0,132			
		Árbol de Hoeffding	0.586	0.283	0.490	0.586
		J48	0.673	0.133	0.632	0.673
		LMT	0,681	0,129	0,646	0,681
		Bosque aleatorio	0,659	0,110	0,646	0,659
		Árbol aleatorio	0,624	0,120	0,602	0,624
		Árbol de representantes	0,667	0,128	0,646	0,667
2	nota	red bayesiana	0.683	0.122	0.661	0.683
		IngenuoBayes	0,675	0,139	0,642	0,675
		IngenuoMulti	0,520	0,520	0,270	0,520
		actualización ingenua	0,675	0,139	0,642	0,675
3	MPL	NeuralN	0,663	0,135	0,615	0,663
4	SOMOS	ESTAMOS	0.631	0.214	0.538	0.631
5	Regresión logística	Logístico	0.687	0.119	0.658	0.687
		SimpleLogística	0,681	0,129	0,646	0,681
6	KNN	IBK (K más cercano)	0,633	0,119	0,611	0,633
		kestrella	0,665	0,143	0,612	0,665
		LWL	0,618	0,130	0,510	0,618
7	otros	JRip	0.629	0.257	0.532	0.626
		Raridad	0.661	0.130	0.617	0.661
		PARTE	0.649	0.135	0.610	0.649



Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

Figura 2. Criterios de rendimiento del modelo de predicción de calificaciones reales

La figura 2 ilustra los criterios de rendimiento de los siete algoritmos de aprendizaje automático supervisado. En el apartado DT, LMT presentó la puntuación más alta en el índice TP (68,1%), seguido de J48 (67,3%), Rep Tree (66,7%), Random Forest (65,9%), Random Tree (62,4%), Decision Stump (61,2%) y Hoeffding Tree (58,6%). En términos de BN, BayesNet demostró la puntuación de tasa de TP más alta (68,3 %), seguida de NaïveBayes y Naïveupdate (67,5 %) y NaïveMulti (52 %). NeuralN obtuvo un 66,3 % en el campo MLP y SMO obtuvo un 63,1 %. En el campo Regresión logística, Logistic obtuvo la puntuación más alta entre todos los algoritmos con un 68,7 %, mientras que SimpleLogistic obtuvo un 68,1 %. En el campo KNN, KStar, K Nearest y LWL obtuvieron 66,5 %, 63,3 % y 61,8 %, respectivamente. Otros algoritmos, como OneR, PART y JRip, obtuvieron 66,1 %, 64,9 % y 62,9 %, respectivamente.

Una tasa de FP baja reduce la tasa de predicción falsa. Random Forest obtuvo el valor más bajo (11 %), seguido de Random Tree (12 %), Rep Tree (12,8 %), LMT (12,9 %), Decision Stump (13,2 %), J48 (13,3 %) y Hoeffding Tree (28,3 %). %). En términos de NB, BayesNet obtuvo un 12,2 %, seguido de NaïveBayes y Naïveupdate (13,9 %) y NaïveMulti (52 %). En el campo MLP, NeuralN obtuvo un 13,5 %, seguido de SMO con un 21,4 %. Para Logistic Regression, Logistic y SimpleLogistic obtuvieron una puntuación de 11,9 % y 12,9 %, respectivamente. En el campo KNN, K Nearest, LWL y KStar obtuvieron 11,9 %, 13 % y 14,3 %, respectivamente. OneR, PART y JRip obtuvieron 13%, 13,5% y 25,7%, respectivamente.

Para el criterio de precisión en la sección DT, LMT, Rep Tree y Random Forest mostraron la puntuación más alta con un 64,6 %, seguidos de J48 (63,2 %), Random Tree (60,2 %), Decision Stump a (61,2 %), Hoeffding Tree (49%). En términos de BN, el puntaje de precisión más alto para todos los algoritmos lo obtuvo BayesNet (66,1 %), seguido de NaïveBayes y Naïveupdate (64,2 %) y NaïveMulti (27 %). NeuralN obtuvo un 61,5 % en el campo MLP, mientras que SOM obtuvo un 53,8 %. En términos de regresión logística, Logistic y SimpleLogistic obtuvieron un 65,8 % y un 64,6 %, respectivamente. En el campo KNN, KStar, K Nearest y LWL obtuvieron 61,2 %, 61,1 % y 51 %, respectivamente. Otros algoritmos, como OneR, PART y JRip, obtuvieron 61,7 %, 61 % y 53,2 %, respectivamente. El campo de criterio de recuerdo obtuvo los mismos valores que el de la tasa de TP.

Tabla 5. Criterios de desempeño del modelo de predicción del estado de los estudiantes

Nº Categoría 1 DT	***Algoritmo***	***Tasa TP Tasa FP*** 0.884		Recuperación de	
	Tocón de decisión	0.120		***precisión*** 0,886 0,884	
	Árbol de Hoeffding	0.845	0.156	0.845	0.845
	J48	0.876	0.127	0.877	0.876
	LMT	0,880	0,122	0,880	0,880
	Bosque aleatorio	0,876	0,127	0,876	0,876
	Árbol aleatorio	0,851	0,149	0,851	0,851
	Árbol de representantes	0.878	0.125	0.879	0.878
2 nota	BayesNet	0.884	0.120	0.886	0.884
	IngenuoBayes	0,876	0,127	0,877	0,876
	IngenuoMulti	0,520	0,520	0,270	0,520

Conferencia de la OIO. Serie: Ciencia e ingeniería de materiales **928** (2020) 032019 doi: 10.1088 / 1757-899X / 928/3/032019

		actualización ingenua	0.876	0.127	0.877	0.876
3 MPL		NeuralN	0,873	0,129	0,874	0,873
4 SOMOS		ESTAMOS	0,876	0,128	0,878	0,876
5	Logístico	Logístico	0,888	0,114	0,888	0,888
	Regresión	SimpleLogística	0,880	0,122	0,880	0,880
6 KNN		IBK (K más cercano)	0.855	0.145	0.855	0.855
		kestrella	0,884	0,120	0,886	0,884
		LWL	0,882	0,122	0,884	0,882
7 Otros		JRip	0,871	0,132	0,873	0,871
		Raridad	0,878	0,125	0,878	0,878
		PARTE	0.871	0.130	0.872	0.871

La Tabla 5 muestra los criterios de desempeño de diferentes algoritmos de aprendizaje automático después de implementar el modelo para predecir si los estudiantes aprobaron o reprobaron; la precisión de la regresión logística fue del 89%. La Figura 3 muestra el cuadro de criterios de desempeño para todas las categorías y algoritmos utilizados para predecir el estado del estudiante (aprobado o reprobado).

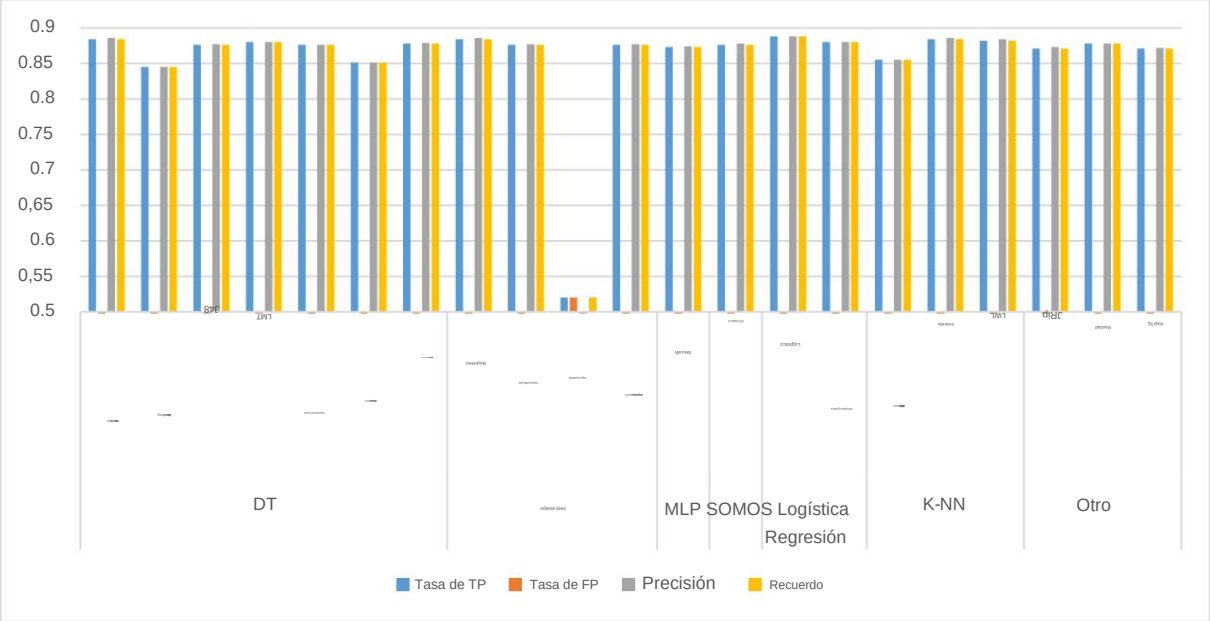


Figura 3. Criterios de rendimiento del modelo de predicción del estado de los estudiantes

La figura 3 muestra los criterios de rendimiento de los siete algoritmos de aprendizaje automático supervisado. En el apartado DT, Decision Stump presentó la puntuación más alta en el índice TP (88,4%), seguida de LMT (88%), Random Forest y J48 (87,6%), Random Tree (85,1%) y Hoeffding Tree (84,5%). . Para BN, BayesNet logró la puntuación de tasa de TP más alta (88,4 %), mientras que NaïveBayes y Naïveupdate obtuvieron un 87,6 % y NaïveMulti obtuvo un 52 %. SMO obtuvo un 87,6 %, mientras que NeuralN obtuvo un 87,3 % en el campo MLP. Para Logistic Regression, Logistic obtuvo la puntuación más alta entre todos los algoritmos con un 88,8 %, mientras que SimpleLogistic

anotó 88%. En el campo KNN, KStar, LWL y K Nearest obtuvieron 88,4 %, 88,2 % y 85,5 %, respectivamente. OneR obtuvo un 87,8 %, y PART y JRip obtuvieron un 87,1 %.

En el campo de la tasa de FP con algoritmos DT, Decision Stump obtuvo el valor más bajo (12 %), seguido de LMT (12,2 %), Rep Tree (12,5 %), J48 y Random Forest (12,7 %), Random Tree (14,9 %). %) y Hoeffding Tree (15,6%). En el campo de NB, BayesNet obtuvo una puntuación del 12 %, NaïveBayes y Naïveupdate obtuvieron una puntuación del 12,7 % y NaïveMulti obtuvo una puntuación del 52 %. NeuralN obtuvo un 12,9 % en el campo MLP, mientras que SOM obtuvo un 12,8 %. El campo Regresión logística obtuvo la puntuación más baja con un 11,4 % para Logística y un 12,2 % para Logística simple. En el campo KNN, KStar, LWL y k más cercano obtuvieron un 12 %, 12,2 % y 14,5 %, respectivamente. Otros algoritmos, como OneR, PART y JRip obtuvieron un 12,5 %, 13 % y 13,2 %, respectivamente.

Para el criterio de precisión en la sección DT, Decision Stump obtuvo la puntuación más alta entre todos los algoritmos DT con un 88,6 %, seguido de LMT (88 %), Rep Tree (87,9 %), J48 (87,7 %), Random Forest (87,6 %). , Random Tree (85,1%) y Hoeffding Tree (84,5%). En términos de BN, BayesNet obtuvo la precisión más alta entre todos los algoritmos con un 88,6 %, mientras que NaïveBayes y Naïveupdate obtuvieron un 87,7 % y NaïveMulti obtuvo un 27 %. En el campo MLP, NeuralN obtuvo un 87,4 %, mientras que SMO obtuvo un 87,8 %. En el campo Regresión logística, Logistic obtuvo la puntuación más alta entre todos los algoritmos con un 88,8 %, mientras que SimpleLogistic obtuvo un 88 %. Para KNN, KStar, LWL y K Nearest obtuvieron 88,6 %, 88,4 % y 85,5 %, respectivamente.
Otros algoritmos, como OneR, JRIP y PART obtuvieron un 87,8 %, 87,3 % y 87,2 %, respectivamente.
El campo de criterio de recuerdo obtuvo los mismos valores que el de la tasa de TP.

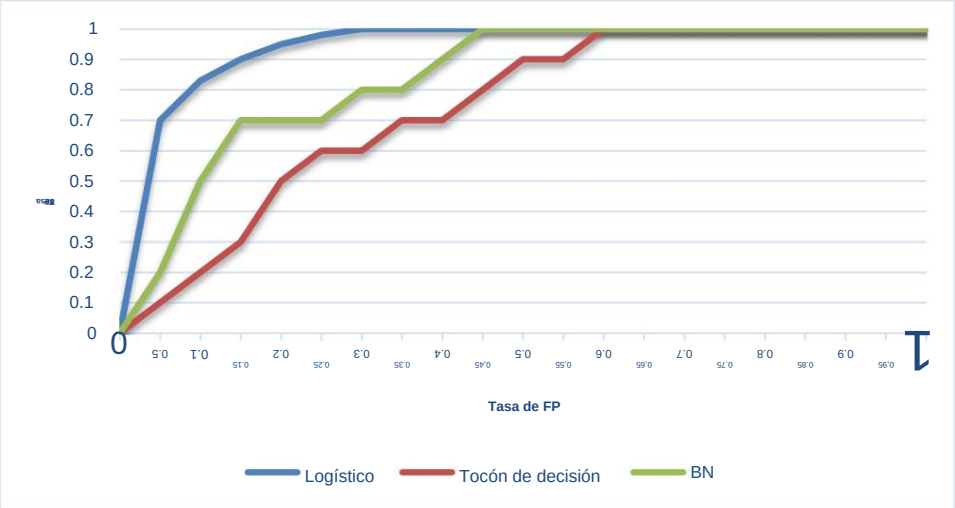


Figura 4. Curva característica operativa del receptor (ROC) de la predicción del estado de los estudiantes

ROC es una herramienta útil para evaluar el rendimiento de los clasificadores. ROC se utiliza para la teoría de detección de señales y el análisis de imágenes de radar. ROC básicamente presenta la compensación entre las tasas de TP y FP como un diagrama de trama. ROC puede ayudar a analizar y reconocer cómo el modelo puede

realizar o identificar erróneamente los casos negativos como positivos. Cuando la curva ROC alcanza el valor de 1, el modelo es preciso para predecir los casos positivos. Cuando la curva ROC alcanza la línea diagonal o 0,5, el modelo tiene poca precisión para predecir los valores [51, 52]. La Figura 4 muestra el ROC de los tres mejores algoritmos supervisados precisos (Logística, Decision Stump DT y BN) para predecir el estado del estudiante como 'aprobado'. Dado que la curva logística alcanzó el valor de 1 en la mayoría de los casos, la curva ROC del primer modelo mostró que el algoritmo Logística fue el mejor en la predicción del estado del estudiante, seguido de Decision Stump DT y BN. La curva ROC del algoritmo Logístico comenzó desde el valor por encima de 0,5 y avanzó hasta alcanzar el valor de 1. El área bajo la curva (AUC) mide toda el área bajo la curva ROC. El valor AUC puede reflejar el rendimiento del modelo; si el valor es cercano a 1, entonces el modelo funciona bien. Los valores AUC de los algoritmos Logistic, BN y Decision Stump fueron 0,9541, 0,9346 y 0,8627, respectivamente.

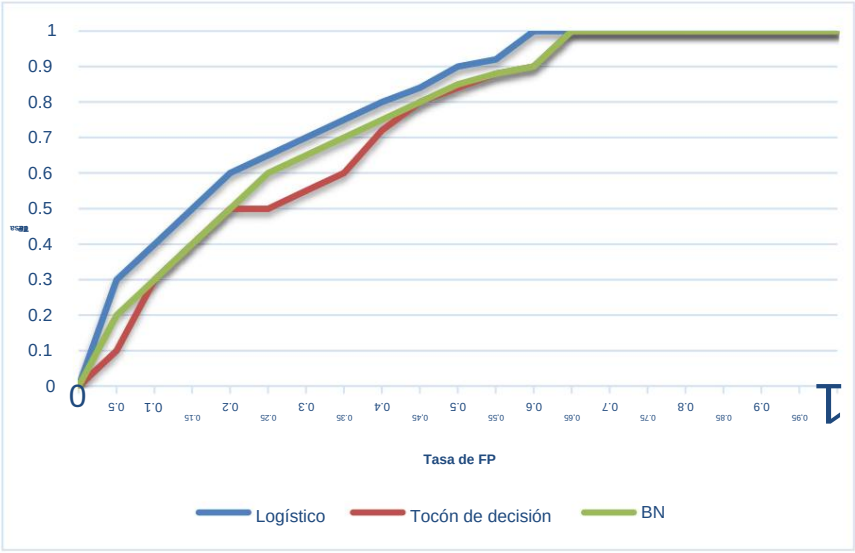


Figura 5. Curva ROC de la predicción de calificación real

La Figura 5 muestra las curvas ROC de los tres algoritmos supervisados más precisos (Logistic, Decision Stump DT y BN) en el segundo modelo para predecir la calificación real 'M' de los estudiantes. La curva ROC del segundo modelo mostró que el mejor modelo para predecir el estado del estudiante fue el algoritmo Logístico porque su curva alcanzó el valor de 1 en la mayoría de los casos, seguido por BN y Decision Stump DT. Los valores AUC de los algoritmos Logistic, BN y Decision Stump fueron 0,9012, 0,8893 y 07984, respectivamente. El valor AUC del algoritmo logístico fue el mejor entre los tres, lo que indica la precisión de la predicción del modelo. La ROC del primer modelo fue mejor que la del segundo porque el número de ítems en la clase final fue de dos,

mientras que la de los elementos predichos en el segundo modelo fue seis, lo que sugiere que el segundo modelo fue más preciso que el primer modelo.

8. Conclusión

Predecir el desempeño de los estudiantes es importante en el ámbito educativo porque el análisis del estado de los estudiantes ayuda a mejorar el desempeño de las instituciones. Las diferentes fuentes de información, como las bases de datos tradicionales (demográficas, académicas y de comportamiento) y las bases de datos multimedia, suelen ser accesibles en las instituciones educativas. Estas fuentes ayudan a los administradores a encontrar información (p. ej., requisitos de admisión), predecir la escala de horarios de la inscripción a clases y ayudar a los estudiantes a decidir cómo elegir cursos dependiendo de qué tan bien les irá en los cursos elegidos. El modelo propuesto predijo el desempeño de los estudiantes. Este modelo se entrenó y los datos se probaron con los datos de los estudiantes durante dos semestres mediante el uso de varios algoritmos de aprendizaje automático supervisados, como Decision Tree, NB, Logistic Regression, KNN, MLP, SMO, Neural Network, PART, JRip y OneR. Se examinaron los criterios de rendimiento de todos los algoritmos para predecir dos grupos de resultados (la calificación real y el estado final de los estudiantes). Se obtuvieron los resultados más perfectos y precisos. La calificación final exacta y el estado previsto del estudiante (aprobado o reprobado) se mostraron mediante regresión logística, con precisiones del 68,7 % y 88,8 %, respectiva Muchos factores afectaron la precisión de los resultados obtenidos después de implementar los algoritmos. Estos factores incluyeron los datos limpios, el dominio de las características, el número de características, el tamaño del conjunto de datos y el dominio de la clase final. La precisión aumentó cuando se redujo el número de valores predichos. Cuando el número de valores en la clase final era seis, la precisión no superaba el 68,7 %, pero cuando este número se convertía en dos, la precisión superaba el 88,8 %. El tamaño del conjunto de datos también afectó la precisión, es decir, la precisión aumentó cuando aumentó el tamaño. ROC y AUC pueden ayudar a determinar la precisión del modelo al observar los valores de la curva. Después de observar la ROC del primer modelo al predecir el estado del estudiante como 'aprobado', se encontró que la regresión logística es el mejor algoritmo, con un AUC de 0,9541, que se considera una buena puntuación. La observación de ROC para el segundo modelo en la predicción de la calificación real del estudiante 'M' también mostró que la regresión logística fue el mejor algoritmo para la predicción, con el AUC de 0,9012. El AUC y ROC pueden ayudar a evaluar el rendimiento del modelo y reflejar la precisión real en la predicción del caso determinado.

REFERENCIAS

[1] J. Luan, "Doctorado en Jefe de Planificación e Investigación", *Cabrillo College Fundador, Knowledge Discovery Laboratories "Aplicaciones de Minería de Datos en la Educación Superior*.

[2] RS Baker, "Minería de datos educativos: un avance para los sistemas inteligentes en la educación", *IEEE Intelligent Systems,* vol. 29, págs. 78-82, 2014.

[3] A. Hamoud, A. Humadi, WA Awadh y AS Hashim, "Predicción del éxito de los estudiantes basada en algoritmos de Bayes", *Revista internacional de aplicaciones informáticas,* vol. 178, págs. 6-12, 2017.

[4] AK HAMOUD, "CLASIFICACIÓN DE LAS RESPUESTAS DE LOS ESTUDIANTES MEDIANTE ALGORITMOS DE GRUPOS BASADOS EN EL ANÁLISIS DE COMPONENTES PRINCIPALES", *Journal of Theoretical & Applied Information Technology,* vol. 96, 2018.

[5] SK Mohamad y Z. Tasir, "Extracción de datos educativos: una revisión", *Procedia-Social and Behavioral Sciences,* vol. 97, págs. 320-324, 2013.

[6] M. Berland, RS Baker y P. Blikstein, "Minería de datos educativos y análisis de aprendizaje: aplicaciones a la investigación construccionista", *Tecnología, conocimiento y aprendizaje,* vol. 19, págs. 205-220, 2014.

[7] Hamoud, Alaa, Ali Salah Hashim y Wid Akeel Awadh. "Almacén de datos clínicos: una revisión". *Revista iraquí de computadoras e informática* 44.2 (2018).

[8] AK Hamoud y AM Humadi, "Modelo de predicción del éxito del estudiante basado en redes neuronales artificiales (ANN) y una combinación de métodos de selección de características", *Journal of Southwest Jiaotong University,* vol. 54, 2019.

[9] B. Guo, R. Zhang, G. Xu, C. Shi y L. Yang, "Predicción del rendimiento de los estudiantes en la minería de datos educativos", en *Simposio internacional sobre tecnología educativa (ISET) de* 2015, 2015, págs. 125-128.

[10] IA Najm, AK Hamoud, J. Lloret e I. Bosch, "Machine Learning Prediction Approach to Enhance Congestion Control in 5G IoT Environment", *Electronics,* vol. 8, pág. 607, 2019.

[11] G. MeeraGandhi, "Enfoque de aprendizaje automático para la predicción y clasificación de ataques mediante algoritmos de aprendizaje supervisado", *Int. J. Cómputo. ciencia Común,* vol. 1, págs. 11465-11484, 2010.

[12] J. Han, J. Pei y M. Kamber, *Minería de datos: conceptos y técnicas:* Elsevier, 2011.

[13] M. Gopal, *Aprendizaje automático aplicado:* McGraw-Hill Education, 2018. [14]

C. Verma, Z. Illés y V. Stoffová, "Modelos predictivos de grupos de edad para la predicción en tiempo real de los estudiantes universitarios que utilizan el aprendizaje automático: resultados preliminares", en *la Conferencia internacional IEEE sobre tecnologías eléctricas, informáticas y de comunicación (ICECCT)* de 2019 , 2019, págs. 1-7.

[15] S. Natek y M. Zwilling, "Solución de minería de datos de estudiantes: sistema de gestión del conocimiento relacionado con instituciones de educación superior", *Sistemas expertos con aplicaciones,* vol. 41, págs. 6400-6407, 2014.

[dieciséis] A. Hamoud, AS Hashim y WA Awadh, "Predicción del rendimiento de los estudiantes en instituciones de educación superior mediante el análisis de árboles de decisión", *Revista internacional de multimedia interactiva e inteligencia artificial,* vol. 5, págs. 26-31, 2018.

[17] E. Yukselturk, S. Ozekes y YK Türel, "Predicción de la deserción estudiantil: una aplicación de métodos de minería de datos en un programa de educación en línea", *Revista europea de aprendizaje abierto, a distancia y electrónico,* vol. 17, págs. 118-133, 2014.

[18] M. Hussain, W. Zhu, W. Zhang, SMR Abidi y S. Ali, "Uso del aprendizaje automático para predecir las dificultades de los estudiantes a partir de los datos de la sesión de aprendizaje", *Artificial Intelligence Review,* vol. 52, págs. 381-407, 2019.

[19] F. Marbouti, HA Diefes-Dux y K. Madhavan, "Modelos para la predicción temprana de estudiantes en riesgo en un curso mediante calificaciones basadas en estándares", *Computers & Education,* vol. 103, págs. 1-15, 2016.

[20] A. Acharya y D. Sinha, "Predicción temprana del rendimiento de los estudiantes mediante técnicas de aprendizaje automático", *Revista internacional de aplicaciones informáticas,* vol. 107, 2014.

[21] RSJ de Baker, T. Barnes y JE Beck, "Minería de datos educativos 2008", en *la 1.ª Conferencia internacional sobre minería de datos educativos Montreal, Québec, Canadá,* 2008.

[22] AS Hashima, AK Hamoud y WA Awadh, "Análisis de las respuestas de los estudiantes mediante la minería de reglas de asociación basada en la selección de funciones", *Journal of Southwest Jiaotong University,* vol. 53, 2018.

[23] E. Costa, RS Baker, L. Amorim, J. Magalhães y T. Marinho, "Minería de datos educativos: conceptos, técnicas, herramientas y aplicaciones", *Journal of Computing in Education,* vol. 1, págs. 1-29, 2013.

[24] RS Baker y PS Inventado, "Minería de datos educativos y análisis de aprendizaje", en *Análisis de aprendizaje,* ed: Springer, 2014, pp. 61-75.

[25] SB Jagtap, "Extracción de datos del censo y análisis de datos mediante WEKA", *versión preliminar de arXiv arXiv:1310.4647,* 2013.

[26] IH Witten, "Minería de datos con weka", *Departamento de Ciencias de la Computación de la Universidad de Waikato, Nueva Zelanda,* 2013.

[27] F. Akter, MA Hossin, GM Daiyan y MM Hossain, "Clasificación de datos hematológicos mediante técnicas de minería de datos para predecir enfermedades", *Journal of Computer and Communications,* vol. 6, pág. 76, 2018.

[28] A. Hamoud, "Selección del mejor algoritmo de árbol de decisión para la predicción y clasificación de la acción de los estudiantes", *American International Journal of Research in Science, Technology, Engineering & Mathematics,* vol. 16, págs. 26-32, 2016.

[29] G. Kostopoulos, S. Kotsiantis y P. Pintelas, "Predicción del rendimiento de los estudiantes en la educación superior a distancia mediante técnicas semisupervisadas", en *Ingeniería de datos y modelos,* ed: Springer, 2015, págs. 259-270.

[30] KP Murphy, *Aprendizaje automático: una perspectiva probabilística:* MIT press, 2012.

[31] X. Zhu y AB Goldberg, "Introducción al aprendizaje semisupervisado", *Conferencias de síntesis sobre inteligencia artificial y aprendizaje automático,* vol. 3, págs. 1-130, 2009.

[32] P. Guleria, N. Thakur y M. Sood, "Predicción del rendimiento de los estudiantes mediante clasificadores de árboles de decisión y ganancia de información", en *Conferencia internacional sobre computación paralela, distribuida y en cuadrícula de* 2014, 2014, págs. 126-129.

[33] A. Hamoud, "Aplicación de reglas de asociación y algoritmos de árboles de decisión con datos de diagnóstico de tumores" *Revista Internacional de Investigación de Ingeniería y Tecnología,* vol. 3, págs. 27-31, 2017.

[34] K. Basu, T. Basu, R. Buckmire y N. Lal, "Modelos predictivos de las decisiones de compromiso de los estudiantes con la universidad mediante el aprendizaje automático", *Data,* vol. 4, pág. 65, 2019.

[35] E. Osmanbegovic y M. Suljic, "Enfoque de minería de datos para predecir el desempeño de los estudiantes" *Revista Económica: Revista de Economía y Negocios,* vol. 10, págs. 3-12, 2012.

[36] A. Géron, *Aprendizaje automático práctico con Scikit-Learn, Keras y TensorFlow: conceptos, herramientas y técnicas para crear sistemas inteligentes:* O'Reilly Media, 2019.

[37] B. Godsey, *Piense como un científico de datos: aborde el proceso de ciencia de datos paso a paso:* Manning Publications Co., 2017.

[38] RB Millar, *Estimación e inferencia de máxima verosimilitud: con ejemplos en R, SAS y ADMB* vol. 111: John Wiley & Sons, 2011.

[39] G. Fitzmaurice y N. Laird, "Análisis multivariante: variables discretas (regresión logística)", 2001.

[40] Y. Liu y H. Huang, "Máquinas de vectores de soporte difusos para reconocimiento de patrones y minería de datos". *Revista internacional de sistemas difusos,* vol. 4, págs. 826-835, 2002.

[41] J. Grus, *Ciencia de datos desde cero: primeros principios con python:* O'Reilly Media, 2019.

[42] T. Hastie, R. Tibshirani y J. Friedman, *Los elementos del aprendizaje estadístico: extracción de datos, inferencia y predicción:* Springer Science & Business Media, 2009.

[43] V. Chaurasia y S. Pal, "Un enfoque novedoso para la detección del cáncer de mama mediante técnicas de minería de datos", *Revista internacional de investigación innovadora en ingeniería informática y de comunicaciones (una organización certificada ISO 3297: 2007) vol,* vol. 2, 2017.

[44] J. Platt, "Optimización mínima secuencial: un algoritmo rápido para entrenar máquinas de vectores de soporte", 1998.

[45] PM Arsad y N. Buniyamin, "Modelo de predicción del rendimiento de los estudiantes de una red neuronal (NNSPPM)", en *la Conferencia internacional IEEE de 2013 sobre instrumentación, medición y aplicaciones inteligentes (ICSIMA),* 2013, págs. 1-5.

[46] MA Ulkareem, WA Awadh y AS Alasady, "Un estudio comparativo para obtener un modelo adecuado en la predicción de los requisitos de electricidad para un período futuro determinado", en *la Conferencia Internacional sobre Tecnología de Ingeniería y sus Aplicaciones (IICETA) de* 2018, 2018, págs. 30-35.

[47] Ulkareem, Maysaa Abd, Wid Akeel Awadh y Ali Salah Alasady. "Estudio comparativo para obtener un modelo adecuado en la predicción de los requerimientos de energía eléctrica para un período futuro dado". *2018 Congreso Internacional de Tecnologías de la Ingeniería y sus Aplicaciones (IICETA).* IEEE, 2018.

[48] GD Israel, "Determinación del tamaño de la muestra", 1992.

[49] U. Sekaran y R. Bougie, *Métodos de investigación para los negocios: un enfoque de desarrollo de habilidades:* John Wiley & Sons, 2016.

[50] M. Tavakol y R. Dennick, "Dar sentido al alfa de Cronbach", *Revista internacional de medicina educación,* vol. 2, pág. 53, 2011.

[51] AP Bradley, "El uso del área bajo la curva ROC en la evaluación de algoritmos de aprendizaje automático", *Reconocimiento de patrones,* vol. 30, págs. 1145-1159, 1997.

[52] J. Han, M. Kamber y J. Pei, "Conceptos y técnicas de minería de datos, tercera edición", *Morgan Kaufmann,* 2011.