TELOS: Revista de Estudios Interdisciplinarios en Ciencias Sociales
ISSN 1317-0570 / ISSN (e) 2343-5763
Vol. 23 (3): 616-639. Septiembre-Diciembre 2021

RENDIMIENTO ACADÉMICO DE ESTUDIANTES EN EDUCACIÓN SUPERIOR: PREDICCIONES DE FACTORES INFLUYENTES A PARTIR DE ÁRBOLES DE DECISIÓN

Brenda Díaz-Landa*
https://orcid.org/0000-0002-2596-1778
Rosana Meleán-Romero**
https://orcid.org/0000-0001-8779-738X
William Marín-Rodriguez***
https://orcid.org/0000-0002-0861-9663

RECIBIDO: Mayo 2021 / **ACEPTADO**: Julio 2021 / **PUBLICADO**: Septiembre 2021

Como citar: Díaz-Landa, Brenda; Meleán-Romero, Rosana; Marín-Rodriguez, William. (2021) Rendimiento académico de estudiantes en educación superior: predicciones de factores influyentes a partir de árboles de decisión. **Telos: revista de Estudios Interdisciplinarios en Ciencias Sociales**, 23 (3), Venezuela. (Pp. 616-639).
DOI: www.doi.org/10.36390/telos233.08

RESUMEN

El artículo tuvo como objetivo predecir el rendimiento académico de estudiantes de maestrías en educación, teniendo como autores principales a Camborda Zamudio (2014), Candia Oviedo (2019), Castrillón et al. (2020), Hussain et al. (2018), Yarlequé Wong (2019). Se empleó la técnica de árbol de decisión y minería de datos y herramientas que provee la inteligencia artificial para construir un modelo con el algoritmo J48 del software WEKA, teniendo en cuenta factores educacionales, familiares, socioeconómicos, de hábitos y costumbres. La muestra estuvo constituida por 237 estudiantes de una universidad pública en Perú, obteniendo mediante el coeficiente Kappa de Cohen un nivel de acierto del 66%. Los resultados dan cuenta de una metodología capaz de entrenar un sistema para clasificar a un estudiante, a partir de una de las categorías del rendimiento académico. Esta clasificación puede identificar a priori a los estudiantes con posibles problemas de rendimiento académico. Como resultado de ello, las medidas de acompañamiento y mitigación se pueden implementar de inmediato.

* Ingeniero de Sistemas, Universidad Nacional Jose Faustino Sanchez Carrión, Huacho, Perú. E-mail: diaz.landa18@gmail.com

** Posdoctora en Agronegocios Universidad Nacional Toribio Rodríguez Mendoza (UNTRM). Doctora en Ciencias Sociales, mención Gerencia. Magíster en Gerencia de empresas, Mención Gerencia Industrial. Licenciada en Administración. Universidad del Zulia (LUZ), Maracaibo, Venezuela. Profesora Titular e Investigadora del Centro de Estudios de la Empresa (CEE) de la Facultad de Ciencias Económicas y Sociales (FCES) de la Universidad del Zulia. Docente e investigadora. Scopus Author ID: 22954427400. E-mail: rosanamelean@gmail.com

*** Maestro en Administración Estratégica, Ingeniero Informático. Docente e investigador en la Universidad Nacional José Faustino Sánchez Carrión y Universidad San Pedro, Perú. Scopus Author ID: 57222415761. E-mail: wmarin@unjfsc.edu.pe

Palabras clave: rendimiento académico; árbol de decisión; minería de datos; predicción; software WEKA.

Academic performance of higher education students: Predictions of influencing factors from decision trees

ABSTRACT

The article aimed to predict the academic performance of students of master's degrees in education, having as main authors Camborda Zamudio (2014), Candia Oviedo (2019), Castrillón et al (2020), Hussain et al. (2018), Yarlequé Wong (2019). The decision tree technique and data mining and tools provided by artificial intelligence were used to build a model with the J48 algorithm of the WEKA software, considering educational, family, socioeconomic, habits, and customs factors. The sample consisted of 237 students from a public university in Peru, obtaining a level of success through Cohen's Kappa coefficient of 66%. The results show a methodology capable of training a system to classify a student based on one of the academic performance categories. This classification can a priori identify students with possible academic performance problems. As a result, accompanying and mitigation measures can be implemented immediately.

Keywords: decision tree; data mining; prediction; academic performance; WEKA software.

Introducción

En la actualidad la búsqueda permanente de la calidad educativa es uno de los objetivos centrales de las instituciones de educación superior. “Con este fin, se ha implementado estrategias y programas diseñados para mejorar el desempeño y la perseverancia de los estudiantes” (Helal et al, 2018). Considerando diferentes factores de influencia, no es fácil determinar la combinación correcta de acciones y decisiones estratégicas que pueden maximizar el desempeño de los estudiantes (Castrillón et al., 2020), que se encuentran orientados por la facilidad y disponibilidad de información con la que se cuenta. Con el transcurso del tiempo se han diseñado sistemas de información para los diferentes niveles de la organización, desde los transaccionales hasta los de soporte para la toma de decisiones, siendo de mayor valor el diseño de sistemas de información que permitan predecir futuros escenarios, lo cual aún es muy reducido.

En el ámbito de la educación superior universitaria específicamente en las escuelas de postgrado, es evidente la necesidad de cumplir con estándares de calidad en el servicio educativo ofertado. Por ello, es clave la autoevaluación de manera que esta permita mejorar sus procesos institucionales con fines de mejora; identificar sus debilidades y potenciar sus fortalezas. En función de lo anterior, surge la necesidad de diseñar un modelo para predecir el rendimiento académico de los estudiantes que permita ajustar las variables de los cursos o las condiciones que ofrecen. Al respecto, es preciso destacar que el rendimiento académico está asociado a factores como sociales, personales e institucionales. A partir de la inquietud

presentada, se plantea como objetivo de la investigación, predecir el rendimiento académico de los estudiantes de las maestrías de los programas en educación en la Escuela de Posgrado de la Universidad Nacional José Faustino Sánchez Carrión Perú, empleando para ello la técnica de minería de datos y árboles de decisión.

Para Redondo Rojo (1997), el éxito está relacionado con características diferentes, como métodos y valores. "Sin embargo, la mayoría de estudios de predicción del rendimiento académico están relacionados con la educación básica y la educación primaria, existen pocas aplicaciones en la educación superior" (Mandelman et al. 2016).

Desde esta perspectiva y teniendo en cuenta el objetivo del artículo, el mismo se estructura en secciones que permitirán responder de manera organizada al fin propuesto. En primera instancia, se realiza un análisis retrospectivo de la variable rendimiento académico como punto de partida de la investigación, de manera que se precisen elementos centrales de su estado del arte. Posterior a ello, se aborda lo referente a la minería de datos, técnica empleada en la investigación, para finalmente apoyado en árboles de decisión como modelo predictivo y determinados softwares para proyectar resultados y el modelo construido como propuesta del estudio. También, se precisa la metodología empleada. Por último, se plantean las conclusiones como síntesis lógica del análisis realizado.

Rendimiento académico: Análisis retrospectivo y comparaciones previas

En esta sección se presenta información relevante, obtenida de estudios previos y que servirán como punto de partida de la investigación realizada. Se cita en esta oportunidad a Timarán-Pereira et al. (2019), autores que determinaron factores asociados con el rendimiento académico, aplicando el método CRISP-DM a partir de un diseño no experimental, apoyados en la base de datos del ICFES para obtener información socioeconómica, académica e institucional, que representó la base para diseñar un modelo predictivo con la herramienta de minería de datos del software WEKA, empleando árboles de decisión, logrando identificar patrones asociados al desempeño (bueno o malo). El estudio fue aplicado a 1,061.680 estudiantes entre los años 2015 y 2016, los resultados refieren que fue posible generar el modelo de predicción.

Otro estudio importante fue propuesto por Candia Oviedo (2019), quien planteó predecir el rendimiento académico de 12,968 alumnos a partir de los datos obtenidos del proceso de ingreso en sus diferentes modalidades, utilizó el aprendizaje automático para desarrollar un modelo predictivo a través de la metodología CRISP-DM y el software WEKA. Los resultados demostraron que es factible predecir el rendimiento en un 69 % de efectividad; además de identificar diversos factores asociados como el promedio de ingreso, carrera, semestre académico, género y modalidad de ingreso. Otra investigación importante fue la desarrollada por Hussain et al. (2018), con el propósito de determinar factores influyentes del rendimiento académico. Los autores, diseñaron un instrumento para analizar 24 factores

conjuntamente con el software WEKA y su algoritmo J48, PART, árboles de decisión y redes bayesianas, logrando determinar la precisión de cada algoritmo diseñado. A partir de datos socioeconómicos y demográficos de 300 estudiantes de tres universidades en la ciudad de Assam – India, determinaron que el algoritmo de árboles de decisión tuvo una precisión del 99%, PART (74.33%), J48 (73%) y las redes bayesianas (65.33%) respectivamente.

De la misma manera, Amaya Torrado, Barrientos Avendaño y Heredia Vizcaíno (2014), con el diseño de un modelo predictivo para la deserción de los estudiantes según condiciones personales, aplicaron un procedimiento cuyas fases se orientaron a la recopilación de información, caracterización de datos personales e información académica, utilizaron el software WEKA para construir y probar el modelo de deserción, además de validar las razones y factores que generan deserción de los estudiantes, generaron un modelo de predicción a más de la mitad de los 201 estudiantes matriculados en el segundo semestre con posibilidades de deserción académica.

Finalmente, Camborda Zamudio (2014), demuestra que la especialidad Ingeniería Civil presenta un bajo rendimiento académico que obedece a factores académicos, demográficos, actitudinales e institucionales. El autor diseñó un modelo de predicción identificando variables influyentes, el proceso de construcción del modelo utilizó la técnica de árbol de decisión, con el algoritmo J48 WEKA, considera que el árbol de decisión se forma a partir de la combinación de atributos provenientes de los registros del I al III Ciclo, pudiendo determinar con una precisión mayor al 80% que las variables académicas son las que definen el rendimiento académico de los estudiantes.

Ahora bien al abordar el rendimiento académico, desde la perspectiva de comparaciones previas, es necesario referenciar el MINEDU – Ministerio de Educación (2009), ente que define el rendimiento académico como el grado en el que se desarrollan las habilidades, los conocimientos y las actitudes, expresados mediante criterios literalmente definidos que el estudiante sabe hacer y evidenciar, refiriéndose a los logros de aprendizaje de los estudiantes otorgándole un calificativo respectivo y que expresa su situación académica. Himmel quien es citado por Chávez Uribe (2006), indica que el rendimiento académico es el grado en el que puede alcanzar las metas establecidas en un programa de estudios sea en el nivel básico o superior. A su vez Yarlequé Wong (2019), indica que en el rendimiento académico intervienen factores como el nivel intelectual, la personalidad, motivación, aptitudes, intereses, hábitos de estudio, autoestima o la relación profesor alumno.

El Ministerio de Educación del Perú (2016), “consigna en el Currículo Nacional de Educación Básica que el rendimiento académico se mide a través de competencias, capacidades, estándares de aprendizaje y desempeño que los estudiantes logran al egresar y que forman parte de su perfil”, definiendo las competencias, como la facultad de una persona en combinar un conjunto de capacidades para lograr un propósito específico en una situación determinada actuando de manera pertinente y con sentido ético.

"El ser competente supone comprender la situación que debe afrontar y evaluar las posibilidades para resolverlo, identifica conocimientos y habilidades inherentes que se encuentren disponibles en el entorno, analizar las combinaciones pertinentes a la situación y propósito, posteriormente tomar decisiones" (Ministerio de Educación, 2016). Las capacidades son recursos para actuar de manera competente y son conocimientos, habilidades y actitudes que los estudiantes utilizan para afrontar una situación determinada; suponen operaciones menores implicadas en las competencias y respecto a los estándares de aprendizaje, son descripciones del desarrollo de las competencias de inicio a fin en la educación básica, siendo descripciones holísticas que hacen referencia a las capacidades que se ponen en acción al resolver o enfrentar situaciones auténticas (Yarlequé Wong, 2019).

El desempeño se considera como la descripción del estudiante respecto al nivel de desarrollo de las competencias que son observables en una diversidad de situaciones. No presentan carácter exhaustivo, ilustran algunas actuaciones que los demuestra cuando están en proceso de alcanzar el nivel esperado de la competencia o cuando lo logra (Yarlequé Wong, 2019). Para el Ministerio de Educación (2016):

> "el desempeño se presenta en los programas curriculares de los niveles o modalidades por edades (inicial o en otras modalidades y niveles de la educación básica), en apoyo a los docentes en la planificación y evaluación, reconociendo que dentro de un grupo de estudiantes hay una diversidad de niveles de desempeño, que pueden estar por encima o por debajo del estándar normal".

Utilizando los fundamentos de la ciencia de datos así como factores socioeconómicos y educativos asociados, Orihuela Maita (2019), predijo el rendimiento académico de 2796 estudiantes, para ello aplica técnicas de extracción, limpieza de datos, explorando y aplicando modelos de machine learning (aprendizaje automático), con modelos de aprendizaje supervisado como la regresión logística y el random forest, en este sentido, los resultados muestran que fue posible generar el modelo predictivo en una de sus fases con el coeficiente de determinación R2, obteniendo como promedio de precisión del 80% en los modelos para la data de entrenamiento y 76% para la data de validación.

En relación a la minería de datos, Yamao (2018), quien utilizó dicha técnica, logró realizar la predicción del rendimiento académico para un total de 1304 estudiantes, su investigación describe la relación que existe entre el rendimiento académico y los factores sociales, económicos y académicos de los ingresantes, los resultados obtenidos indicaron predicciones a través de las técnicas de regresión lineal, árbol de decisiones y support vector machines. En la investigación, el 82.87% se obtuvo utilizando árbol de decisiones, los factores de mayor influencia fueron la nota de examen de admisión, el género, la edad, la modalidad de ingreso y la distancia desde su casa hasta el centro de estudios. Por su parte, Cuji, Gavilanes y Sánchez (2017), utilizaron las técnicas de clasificación basada en árboles de decisión, construyendo un modelo predictivo de deserción estudiantil para 378 estudiantes que pronostique la probabilidad que un estudiante abandone su programa académico, emplearon la

metodología Knowledge Discovery in Database (KDD) con cinco etapas: selección, procesamiento, transformación, minería de datos y evaluación, aplicaron en su estudio el algoritmo Classification and Regression Tree (CART) de la herramienta R, lograron construir un modelo predictivo con un árbol de decisión de cuatro niveles de profundidad y con el mismo número reglas evaluaron a los posibles desertores, concluyendo que las variables nivel y notas tienen mayor influencia en la deserción de estudiantes.

Menacho Chiok (2017), aplica árboles de decisión, regresión logística, redes neuronales y bayesianas con la data de 914 estudiantes para predecir su clasificación final (aprobado o desaprobado), con un modelo predictivo de aprendizaje supervisado y la técnica de minería de datos y así obtener un modelo para predecir el resultado de los alumnos, demostrando que con la técnica de red (Naive de Bayes) es de mayor precisión con el 71,0%, frente a cuatro técnicas analizadas (redes neuronales, regresión logística, árbol de decisión y red bayesiana) con mayor porcentaje para la clase aprobados y menor para los desaprobados.

De la misma manera, García Tinisaray (2015), generó un modelo para determinar el rendimiento académico de 23583 estudiantes basado en learning analytics con técnicas multivariantes, seleccionando para ello variables demográficas, académicas y tecnológicas, las variables personales las considero de nivel inferior, la variable docente y asignatura intermedio y las variables institucionales como superior, concluye que se mantiene asociación positiva con los factores analizados en la predicción del estudiante, para luego eliminar factores no asociados al rendimiento académico no encontrando una significancia estadística.

En función de la discusión previa generada sobre la variable rendimiento académico, se precisa como hipótesis: "los factores asociados que permiten predecir el rendimiento académico de los estudiantes de los programas de maestría en educación"

Data mining (minería de datos)

En lo que respecta a la minería de datos, esta herramienta es propicia para el tratamiento de grandes cantidades de datos, con la finalidad de generar conocimiento. Utiliza bases de datos, de donde se extrae información de forma automatizada, utiliza el análisis matemático y estadístico para deducir patrones y tendencias. Típicamente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos ya que las relaciones son demasiado complejas o porque existe demasiada data (Microsoft, 2019). Su desarrollo utiliza técnicas estadísticas e inteligencia artificial para detectar patrones y anomalías en grandes cantidades de datos.

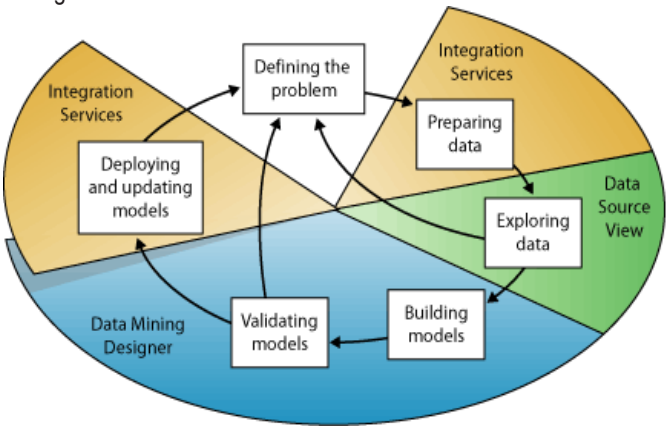
Con relación al proceso de la minería de datos, Camborda Zamudio (2014), "realiza un resumen de principales características: descubre hechos y relaciones de datos, necesita poca intervención humana, encuentra patrones, determina, almacena, reutiliza y establece reglas, presenta información al usuario final capaz de analizar en los siguientes pasos:

a. **Selección del conjunto de datos.** Referida a la variable objetivo a predecir, como la variable independiente (cálculo) y el muestreo de los registros disponibles.
b. **Análisis propiedades de datos.** Como histogramas, diagramas de dispersión, presencia de valores atípicos y valores nulos.

c. **Transformación de datos de entrada.** Llevado a cabo de diversas formas en función del análisis previo, el objetivo es aplicar la técnica de minería de datos que mejor se adapte al conjunto de datos y problema, conocido como pre procesamiento de la data.
d. **Elección y aplicación de la técnica de minería de datos**. Aquí se construye el modelo predictivo, para clasificar y segmentar.
e. **Extracción de conocimiento**. Con la técnica de minería de datos se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o sus relaciones. Se utilizan métodos diferentes al mismo tiempo para crear modelos diferentes.
f. **Interpretación y evaluación de datos**. Obtenido el modelo, se procede a validarlos comprobando que las conclusiones son válidas y satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos" (p. 14-15).

Según la aproximación de Microsoft, la minería de datos utiliza estos pasos cuya interacción es cíclica, según se muestra en la figura 1.

Figura 1. Proceso de minería de datos



Fuente: Microsoft (2019)

La técnica de minería de datos, es estructurada en el tiempo y deriva del desarrollo entre la inteligencia artificial y la estadística, se trata de algoritmos sofisticados que, aplicados sobre una estructura de datos se obtienen resultados. Las técnicas más representativas según la agrupación realizada por Camborda Zamudio (2014) son redes neuronales, árboles de decisión, regresión lineal, modelos estadísticos, agrupamiento (clustering) y reglas de asociación.

De esta manera, se tiene como hipótesis principal que guía la investigación: **"**Es posible aplicar la técnica de minería de datos mediante árboles de decisión para crear un modelo

de predicción del comportamiento del rendimiento académico de los estudiantes". En esta oportunidad, se enfatizan los árboles de decisión, como modelo predictivo utilizado en el campo de la inteligencia artificial (IA). Según plantea Kumar Yadav, Baharadwaj y Pal (2012), esta herramienta representa "uno de los algoritmos más utilizados en los métodos de aprendizaje supervisado para la exploración de la data basada en la técnica divide y vencerás" (p. 13); desarrollada en el año 1963 en el aprendizaje de modelos de decisión elaborados desde una muestra de datos, permite construir un "modelo" o "representación" de la regularidad existente en los datos. Los diagramas de estructura lógica similares a los sistemas de predicción basados en reglas se construyen en una base de datos y se utilizan para representar y clasificar una serie de condiciones que ocurren continuamente para resolver un problema.

Según Mitchell (2000), "los árboles de decisión son una técnica de clasificación fácil de interpretar y utilizar, que generan reglas del tipo Si…entonces", para Urbina-Nájera (2021): "son reglas en forma de árbol, donde el conjunto de datos se divide en ramas hasta obtener segmentos de similar comportamiento en función de la variable objetivo y se utilizan en la toma de decisiones dado que son de fácil interpretación" (p. 42).

El árbol de decisión desarrolla un test a medida que recorre hacia las hojas logra una decisión, contiene nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un test sobre algún valor de una de las propiedades, un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada. Para el diseño de aplicaciones de software, el árbol indica las acciones a realizar en función del valor obtenido en una o más variables. De otro lado, se debe analizar un algoritmo para determinar el desempeño al momento de realizar una tarea como: clasificar, reconocer, identificar, agrupar, categorizar, entre otros, algunas métricas que evalúan dicho desempeño son: la precisión, exactitud, recuperación, Medida-F, matriz de confusión y similares (Witten, Frank, Hall, Pal, 2016).

Figura 2. Árbol de decisión: características

Fuente: elaboración propia.

Con el enfoque de la inteligencia artificial (IA) empleando árboles de decisión, se muestran una diversidad de trabajos en realizar el análisis del rendimiento académico.

Bravo et al. (2015):

> (…) aplicaron dos técnicas (bosques aleatorios y árboles de regresión), en una base de datos de estudiantes chilenos de octavo grado, identificaron y caracterizaron el perfil de los estudiantes con base al rendimiento académico obtenido en matemáticas. Encontraron que las expectativas educativas de los padres, el tipo de escuela y el índice de habilidades matemáticas, eran los factores de mayor influencia (p. 1).

En función de lo anterior, se plantea como segunda hipótesis: "La exactitud de la técnica de árboles de decisión para predecir el rendimiento académico de los estudiantes de los programas de maestría en educación es muy buena", precisando como tercera hipótesis: "La medida de concordancia de la técnica de árboles de decisión para predecir el rendimiento académico es muy buena".

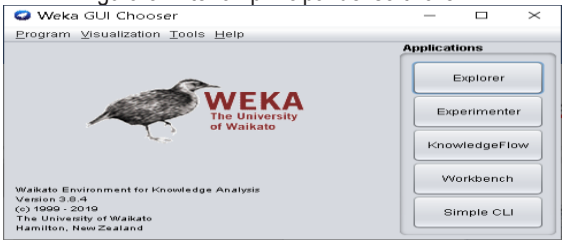
Miguéis et al (2018), "utilizaron técnicas de minería de datos para clasificar a los estudiantes según su potencial académico con el fin de disminuir los fracasos, mejorar los resultados, conseguir más recursos, logrando un nivel de efectividad del 95%"

Software WEKA

WEKA es un software que se utiliza en el aprendizaje automático y la minería de datos (Córdoba Fallas, 2011), el software fue diseñado en Java y fue desarrollado en la Universidad de Waikato en Nueva Zelanda en el año 1993. WEKA (Waikato Environment for Knowledge Analysis) es una herramienta de distribución de licencia GNU-GLP y contiene una colección de algoritmos para realizar el análisis de datos y modelado predictivo, cuenta con herramientas para visualizar los datos, además de proveer una interfaz gráfica que integra sus herramientas para a una mejor disposición.

El WEKA, según indica Candia Oviedo (2019), es una herramienta de software muy versátil, soporta varias tareas de la minería de datos en especial los de procesamiento de datos, regresión, clasificación, clustering entre otras. Todas las técnicas del software WEKA están basadas en la asunción de datos disponibles en un fichero plano, donde cada registro está descrito por un número fijo de atributos nominales o numéricos, a la vez que permite acceder a otras instancias de la base de datos mediante consultas SQL.

Figura 3. Interfaz principal del software WEKA



Fuente: autoría propia

Metodología de la investigación

La investigación es de tipo aplicada, se desarrolló mediante técnicas de clasificación inteligente, en el algoritmo J48 ejecutado con WEKA (https://www.cs.waikato.ac.nz/ml/weka/), para predecir de forma a priori el rendimiento académico de los estudiantes y, a partir de ello, "establecer estrategias personalizadas, programas de ayuda, seguimiento, en aquellos estudiantes cuyo rendimiento proyectado arroje resultados no satisfactorios, siendo posible determinar los factores incidentes en el rendimiento de un estudiante con miras a implementar programas de mejora continua a nivel institucional" (Castrillón et al, 2020).

Con la información obtenida, se busca analizar si es posible asociar la aplicación de la técnica de árbol de decisión utilizando la minería de datos con el rendimiento académico de los estudiantes y así elaborar un modelo predictivo. La muestra de estudio fue de 237 estudiantes matriculados en el semestre académico 2019-II, en los programas de Docencia Superior e Investigación Universitaria, Gerencia de la Educación y Ciencias de la Gestión Educativa con mención en Pedagogía de la Universidad Nacional José Faustino Sánchez Carrión Perú, teniendo en cuenta la totalidad de casos se utilizó un muestreo poblacional no probabilístico, como un "subconjunto de la población en la que la elección de los elementos no depende de la probabilidad, si no de las características del estudio" (Hernández Sampieri, Fernández Collado y Baptista Lucio, 2014, p. 176). Se tuvo acceso a la data de los estudiantes de la base de datos de registros académicos de la Escuela de Posgrado.

Cuadro 1. Variables, dimensiones e indicadores

VARIABLES	DIMENSIONES	INDICADORES
Rendimiento académico.	Nota de promedio ponderado	Datos de la Oficina de Registros académicos
	Identificación	
	Académica	Cuestionario
	Actitudinal	
Técnica de árbol de decisión	Confianza institucional Aplicación de la técnica a través de un modelo de simulación	Modelo de simulación mediante la técnica de árbol de decisión

Fuente: elaboración propia.

Con la finalidad de tener acceso a la data, se utilizó la técnica de revisión bibliográfica, el análisis documental y los récords académicos de los estudiantes, como instrumento de medición se empleó un cuestionario con la finalidad de obtener los factores respectivos con el modelo de predicción, con datos demográficos, académicos, actitudinales e institucionales. De manera similar Camborda Zamudio (2014) adaptó las características de la población de estudio, validando a través de juicio de expertos.

El cuestionario diseñado de 26 ítems distribuidos y agrupados en las variables de identificación, académico, actitudinal, deserción y confianza institucional, fue aplicado a estudiantes seleccionados. Las respuestas obtenidas se les asignó un valor en escala de 1 al 5, posteriormente se sometieron a pruebas estructuradas e ingresados al software WEKA, obteniendo la data para el análisis en la base de datos de entrenamiento, culminado este

proceso se creó una base de datos en MS. Excel vinculando aspectos influyentes en el rendimiento académico, categorizando conforme al promedio ponderado obtenidos mediante la revisión documental (registros de notas). La tabla 2 muestra la categorización del rendimiento académico que requiere el algoritmo J48 del software WEKA, siendo la clase a predecir de tipo categórico y no numérico como la variable promedio.

Tabla 2. Categorización del rendimiento académico según promedio ponderado

Promedio ponderado	Rendimiento académico
De 0 a 10	Malo
De 11 a 15	Bueno
De 16 a 20	Muy bueno

Fuente: autoría propia

Con la tabulación de datos en MS Excel se ingresó al software WEKA, convirtiéndose a texto simple en formato plano de codificación ANSI con extensión .ARFF, analizando los 26 ítems como base de entrenamiento del software, posteriormente se procedió a generar un nuevo archivo plano con la variable promedio, reemplazada por un carácter de incógnita (?) que es utilizado por el software para generar predicciones; el nuevo archivo plano con extensión .ARFF generado se ingresó al software WEKA, obteniéndose una predicción del atributo promedio en base al entrenamiento de la base de datos original, se comprobó el nivel alcanzado de la predicción por el software, se procedió a analizar el nivel de acierto a través del índice Kappa de Cohen con el estadístico correspondiente a la fuerza de concordancia de la predicción que brinda como resultado el software WEKA (Tabla 3).

Tabla 3. Valor de interpretación de la fuerza de concordancia según el valor del coeficiente Kappa de Cohen

Coeficiente Kappa	Fuerza de concordancia
0	Pobre
0.01 – 0.20	Leve
0.21 – 0.40	Aceptable
0.41 – 0.60	Moderada
0.61 – 0.80	Considerable
0.81 – 1.00	Casi perfecta

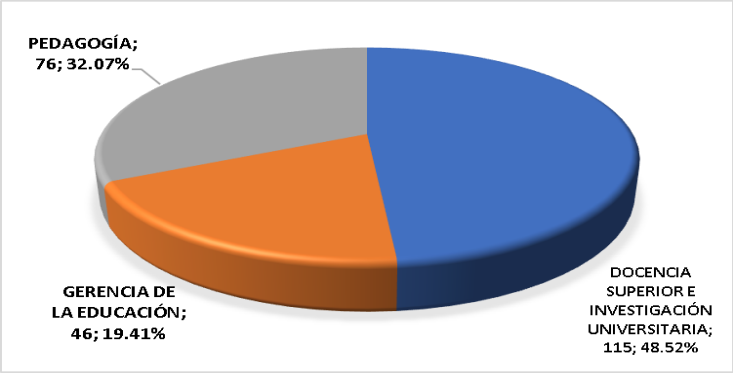
Fuente: Cerda y Villarroel (2008)

Rendimiento académico de estudiantes de maestrías en educación: Resultados

Análisis descriptivo

A continuación, se presentan los resultados obtenidos del análisis estadístico realizado a los 237 estudiantes matriculados en los programas de maestría en educación, del total de estudiantes encuestados la figura 4, 115 cursan la maestría en Docencia Superior e Investigación Universitaria representando el 48,52%, 76 estudiantes cursan la maestría de Pedagogía representando el 32,07% y a 46 estudiantes que cursan la maestría en Gerencia de la Educación representando el 19,41%

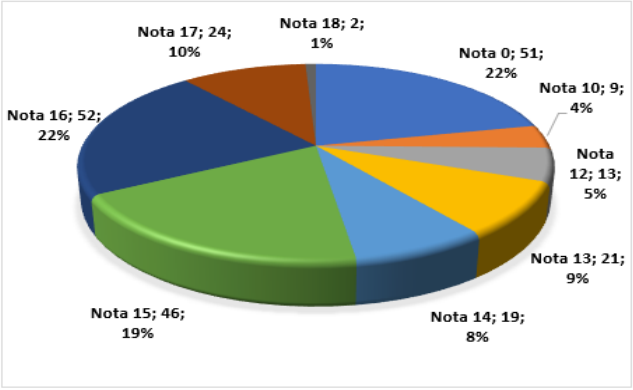
Figura 4. Distribución de encuestados según maestría que cursan



Fuente: elaboración propia.

La figura 5, muestra la distribución del promedio ponderado de los estudiantes, 46 estudiantes obtuvieron el promedio ponderado 15 equivalente al 19%, 52 obtuvieron un promedio de 16 equivalente al 22%, 24 estudiantes obtuvieron un promedio de 17 equivalente al 10% y así sucesivamente, en la figura se resalta que 51 estudiantes se retiraron de los cursos obteniendo el promedio de 0 equivalente al 22%.

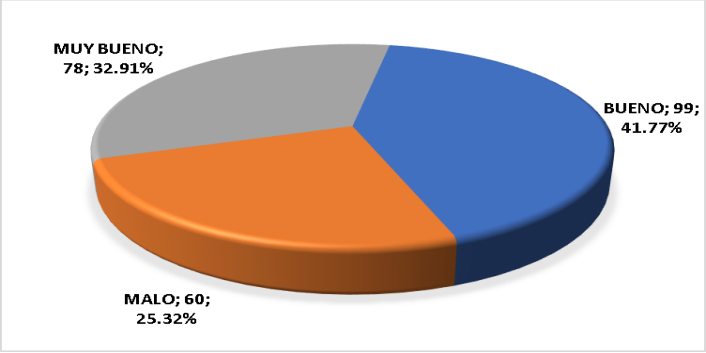
Figura 5. Distribución de promedios ponderados alcanzados por los estudiantes



Fuente: elaboración propia.

La figura 6, muestra la distribución del rendimiento académico por categorías con respecto al promedio ponderado de los estudiantes (ver tabla 2), el análisis respectivo será a través del algoritmo J48 del software WEKA utilizando árboles de decisión.

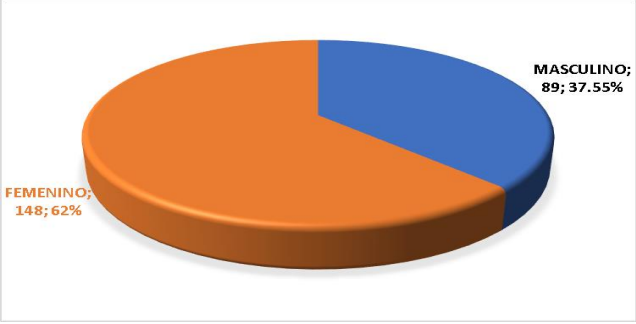
Figura 6. Distribución del rendimiento académico por categorías



Fuente: elaboración propia.

Los resultados en la figura 7, muestra información correspondiente al género de los estudiantes, 148 son mujeres representando el 62% y 89 son varones lo que representa el 37.55%, el análisis se realizará con el software WEKA.

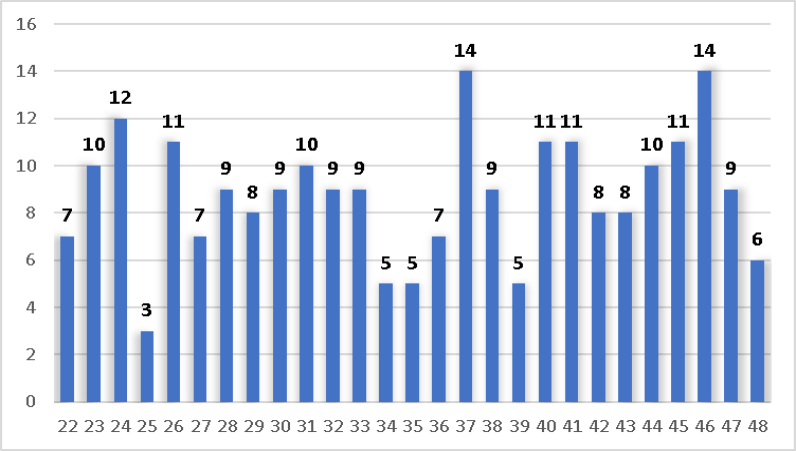
Figura 7. Distribución de estudiantes por sexo



Fuente: elaboración propia.

La figura 8, muestra la distribución de frecuencia de estudiantes por edad, la mayor cantidad de estudiantes están comprendidos entre los 37 y 46 años, corresponde a una edad adulto maduro, etapa que regularmente ya se cuenta con un oficio o trabajo definido y con carga familiar, el software debe analizar si es un atributo que contribuye también en determinar el rendimiento académico.

Figura 8. Distribución de estudiantes por edad



Fuente: elaboración propia.

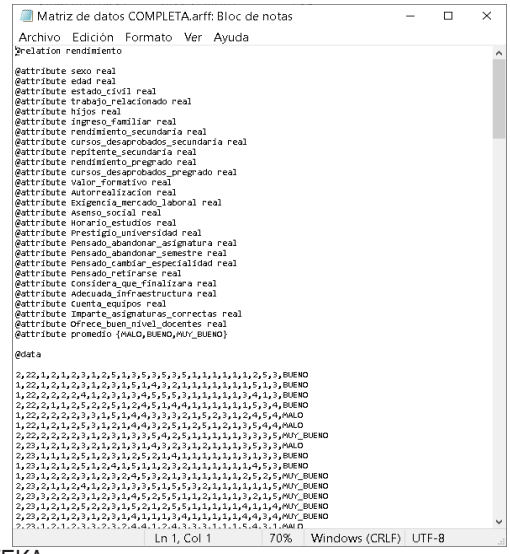
Prueba de hipótesis

Se procedió a determinar las pruebas de hipótesis usadas para la investigación.

Primera prueba de hipótesis:

"Sí es posible determinar los factores asociados que permiten predecir el rendimiento académico de los estudiantes de los programas de maestría en educación". Se elaboró la matriz de datos resultante en el software WEKA, con la finalidad de determinar la relación estadística de los valores de los atributos como resultado de la clase utilizada (rendimiento académico).

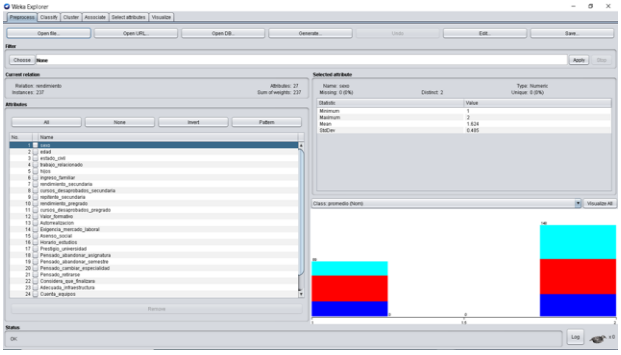
Figura 9. Matriz de datos que se va a procesar con el software WEKA



Fuente: software WEKA

Se procedió con la carga de la data en la base de datos de entrenamiento, WEKA identificó las respuestas en 26 atributos que se definen es la clase (promedio), corresponde al rendimiento académico categorizado de tipo nominal.

Figura 10. Carga de datos en la base de datos de entrenamiento (interfaz de WEKA)

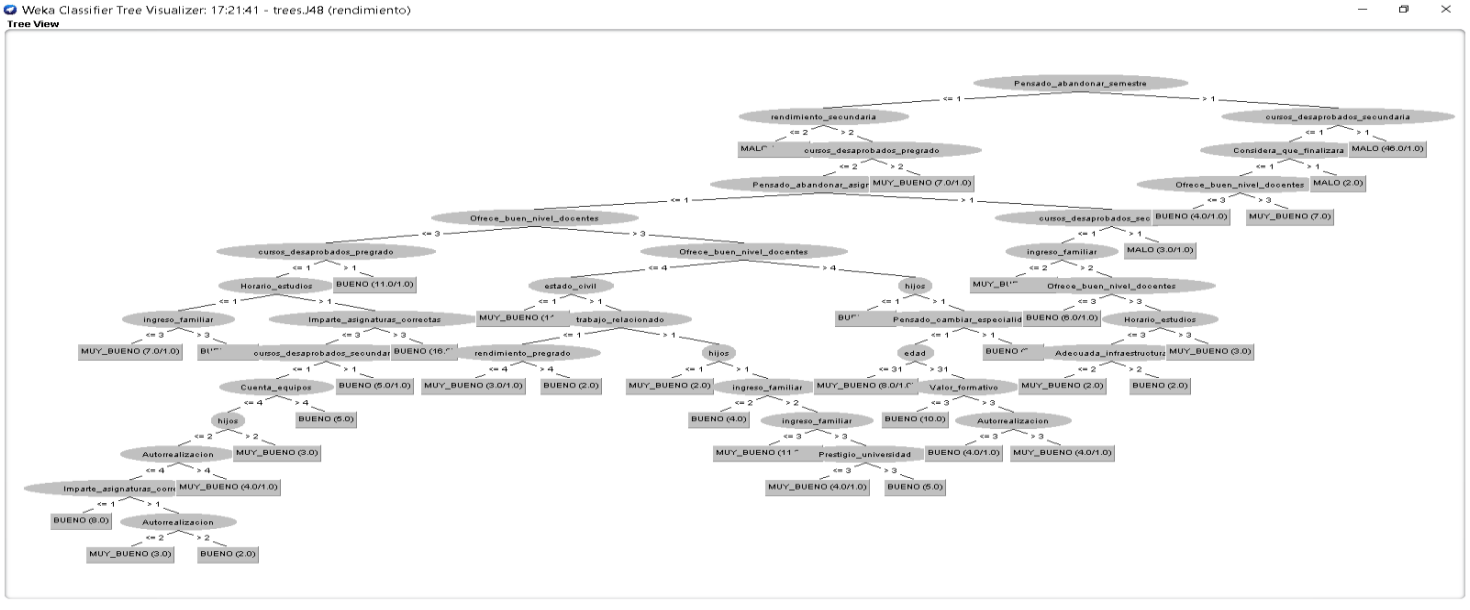


Fuente: software WEKA

Utilizando el algoritmo de árboles de decisión J48 se realizó el análisis estadístico para efectuar el entrenamiento con los datos, las relaciones entre los atributos y la clase que busca

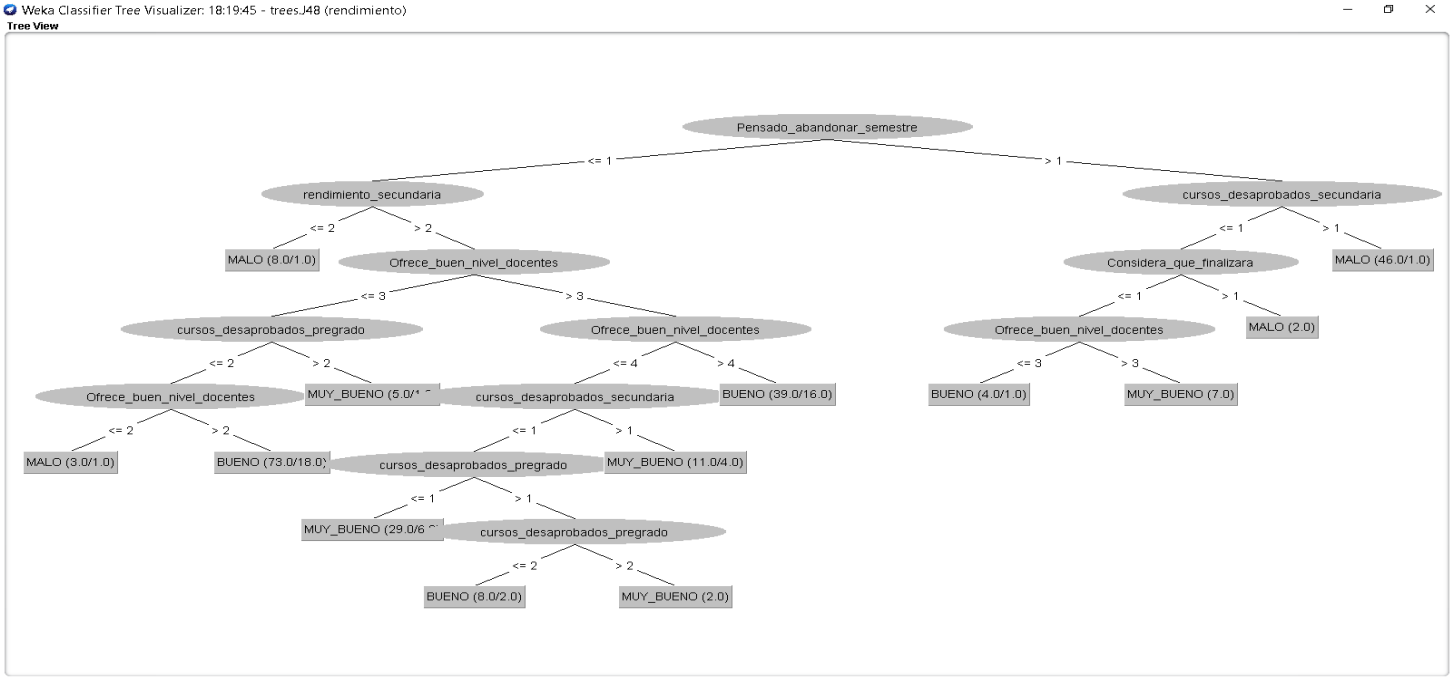
predecir, el software fue capaz de clasificar correctamente las instancias en un 62,45%, el análisis de la base de entrenamiento fue capaz de determinar las reglas entre los atributos que se definieron previamente en la clase buscada (promedio) y así realizar la predicción respectiva correspondiendo a la exactitud del modelo, por lo que se acepta la primera prueba de hipótesis. Con relación al análisis inferencial para las predicciones se realizó con el coeficiente Kappa de Cohen obteniendo un valor de 0,4199 que corresponde a una concordancia moderada con respecto a la exactitud del modelo predictivo creado. La base de entrenamiento analizada por el software WEKA y el algoritmo J48 generó un árbol de decisión con los 26 atributos ingresados (figura 11).

Figura 11. Árbol de decisión generado por del análisis de la base de entrenamiento con el algoritmo J48 software WEKA



Fuente: software WEKA

Figura 12. Árbol de decisión generado con el algoritmo J48 software WEKA priorizando 5 primeros niveles a partir del modelo inicial



Fuente: software WEKA

Segunda prueba de hipótesis:

"La exactitud de la técnica de árboles de decisión para predecir el rendimiento académico de los estudiantes de los programas de maestría en educación es muy buena".

En este caso, la precisión se calcula dividiendo el número total de registros clasificados por el número total de registros y se expresa como porcentaje. El resultado obtenido fue de 0,623 no logrando llegar a una exactitud del 0,75; por lo tanto, se rechaza la segunda prueba de hipótesis.

Tercera prueba de hipótesis:

"La medida de concordancia de la técnica de árboles de decisión para predecir el rendimiento académico es muy buena".

Para este caso la concordancia se relaciona el coeficiente Kappa de Cohen con dos medidas realizadas con un instrumento distinto, para analizar el acierto de uno de ellos con relación al otro. El coeficiente Kappa de Cohen obtenido fue de 0,412 que corresponde a una concordancia moderada, por lo tanto, se rechaza la segunda prueba de hipótesis puesto que el valor esperado debió ser mayor a 0,81.

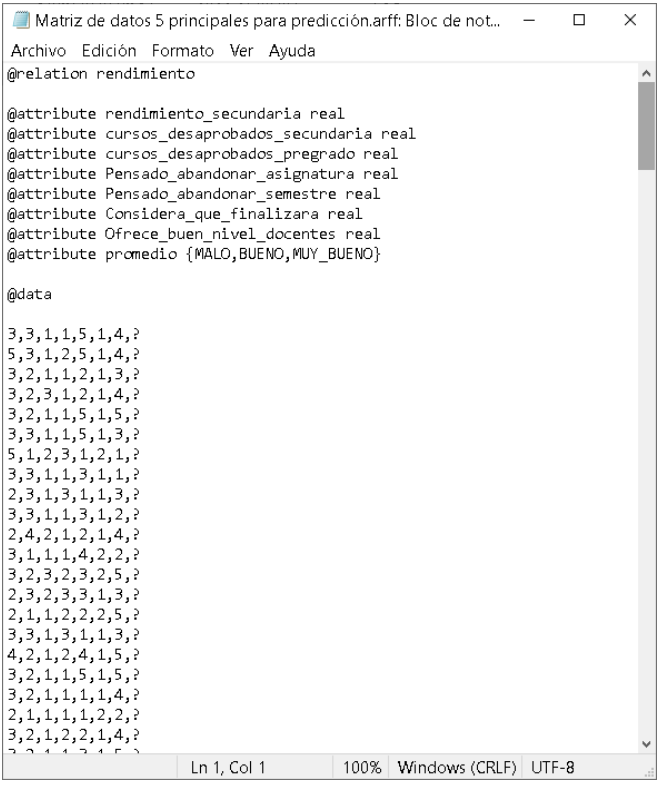
Prueba de hipótesis principal:

"Es posible aplicar la técnica de minería de datos mediante árboles de decisión para crear un modelo de predicción del comportamiento del rendimiento académico de los estudiantes".

Al aceptarse la primera prueba de hipótesis fue posible determinar los factores que permiten predecir el rendimiento académico de los estudiantes, a pesar que se tuvo como resultado un nivel de "concordancia moderada". Para esta prueba de hipótesis se procedió a construir una nueva base de datos de entrenamiento y su correspondiente prueba de predicción, considerando los 5 niveles iniciales del árbol de decisión generado por el software WEKA mediante el algoritmo J48, la figura 12 muestra el árbol de decisión reajustado priorizando 5 niveles.

En las predicciones generadas por este segundo modelo se utilizó un nuevo archivo de texto para el ingreso de datos a través de la opción "supply test set" desde la pestaña "classify" del software WEKA, como se puede visualizar en la figura 13.

Figura 13. Matriz de datos para ser procesado por el software WEKA utilizando los 5 niveles principales del árbol de decisión



Fuente: software WEKA

Los resultados obtenidos con este nuevo modelo de predicción se estructuraron y fueron comparados con los valores reales del rendimiento académico en el software SPSS V. 25, para el análisis de coeficiente Kappa de Cohen se utilizaron los primeros 5 niveles del árbol de decisión, en comparación con los datos reales del rendimiento académico y la predicción obtenida con el software WEKA, el valor obtenido del coeficiente Kappa de Cohen fue de 0,666 corroborando la fuerza de esta relación, concluyendo que la concordancia de este modelo reajustado permite adjudicar una mayor confiabilidad a las predicciones realizadas por el modelo, demostrándose la hipótesis principal de la investigación.

Discusión

Los resultados obtenidos demuestran que el modelo de predicción generado con el algoritmo J48, las variables rendimiento académico y técnicas de árbol de decisión presentan relación significativa, muy similar a las investigaciones realizadas por Hussain et al. (2018), Amaya Torrado, Barrientos Avendaño y Heredia Vizcaíno (2014), Timarán-Pereyra, Caicedo-

Zambrano y Hidalgo-Troya (2019), Camborda Zamudio (2014); así mismo es preciso resaltar que no se empleó un modelo basado en la técnica de multivariado tal como lo realizó en su estudio García Tinisaray (2015). En relación a la exactitud del modelo para la clase rendimiento académico, el valor obtenido fue de 62,45% menor al valor obtenido en la investigación realizada por Hussain et al. (2018) quienes en su modelo de predicción con J48 algoritmo del software WEKA, obtuvieron el valor de 73%, muy cercano al resultado de efectividad obtenido por Candia Oviedo (2019) que fue del 69%. El valor que se obtuvo en la investigación fue menor al obtenido en la investigación de Camborda Zamudio (2014) quien para predecir el rendimiento académico utilizó el algoritmo J48 de WEKA, obteniendo un valor superior al 80%.

Comparando con la investigación realizada por Orihuela Maita (2019) obtuvo como resultado un valor menor al 80% de exactitud en el modelo para la data de entrenamiento y el 76% para la data de validez. A su vez Yamao (2018) en su investigación obtuvo el 82.87% de acierto utilizando la técnica de árbol de decisión, representado con un valor mayor al resultado de nuestra investigación, existiendo la posibilidad de mejorar la exactitud del modelo al priorizar los ítems, atributos y su respectivo análisis. Teniendo en consideración la construcción de un modelo de predicción ajustado y con respecto a la concordancia este fue estimado utilizando el coeficiente Kappa de Cohen, obteniéndose el nivel de acierto con un valor de 41% que le corresponde a una concordancia moderada. Teniendo en cuenta que se generó un nuevo modelo con los cinco primeros niveles del árbol de decisión generado a raíz del modelo inicial siendo este modelo de mejor concordancia, ya que fue medida por el coeficiente Kappa de Cohen obteniéndose un nivel de acierto de 66% correspondiendo la asociación “considerable”.

“Las técnicas inteligentes y, en especial, las técnicas bayesianas, permiten obtener muy buenos resultados con un conjunto pequeños de datos, es decir, aunque la muestra obtenida sea pequeña los resultados son fiables” (Valencia Cárdenas et al., 2015). Se inicia con una muestra estadística válida, a la vez se infiere para un promedio excelente o muy bueno, va a depender del empeño y dedicación de los docentes y alumnos demuestren en el desarrollo de clases, incluso mucho más que los recursos que tenga la institución; sin embargo, si se logra lo anterior y el trabajo u los demás factores de los estudiantes disminuye, el promedio también puede bajar de manera significativa, siendo malo inclusive. En resumen, de acorde con los resultados, a partir de la muestra el rendimiento académico este se asocia más a factores humanos que los propios recursos materiales.

Conclusiones

La investigación pone en evidencia que es posible aplicar la metodología de minería de datos empleando árboles de decisión para generar el modelo de predicción del rendimiento académico, se demuestra la asociación en la predicción del rendimiento académico empleando árbol de decisión conjuntamente con J48 algoritmo bayesiano del software WEKA. a su vez fue posible determinar factores que permitieron realizar la predicción del rendimiento académico clasificándose en un árbol de decisión. Contrariamente no se logró evidenciar la exactitud con el modelo inicial al obtener el valor de 62,45%, tampoco se pudo demostrar la concordancia del modelo mediante el coeficiente Kappa de Cohen que alcanzó el valor de 42% siendo de concordancia “moderada”, con el modelo de predicción reajustado se obtuvo un nivel de acierto de 66%, se identificaron técnicas inteligentes con respecto a los principales factores influyentes

en el rendimiento académico, estos factores fueron la pedagogía, adecuados horarios de clase, buena relación interpersonal docente-estudiante, calidad académica, lográndose con en el algoritmo de clasificación bayesiano J48, generando el árbol con todos los atributos influyentes del rendimiento académico, con la técnica utilizada se logró una efectividad del 66.6%.

Teniendo en cuenta que "predecir el rendimiento académico ayuda a las universidades a reducir su tasa de deserción y mejorar el rendimiento académico de sus estudiantes. Se sigue investigando para averiguar qué algoritmo es mejor utilizar y qué características tener en cuenta" (Katarya, Gaba, Garg y Verma, 2021). Es por ello que resulta necesario a futuro seguir profundizando en próximas investigaciones sobre la predicción del rendimiento académico tomando como base el estudio realizado, sin embargo, es necesario avanzar en nuevas investigaciones para observar otros factores que pueden afectar el rendimiento académico en los estudiantes.

Referencias bibliográficas

Amaya Torrado, Yegny Karina; Barrientos Avendaño, Edwin; Heredia Vizcaíno, Diana Judith. (2014). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Extraído de: https://dspace.redclara.net/handle/10786/759

Bravo Sanzana, Mónica; Salvo Garrido, Sonia; Muñoz Poblete, Carlos. (2015). Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests. **Studies in Educational Evaluation**, 44, Uk. (Pp. 50-59). https://doi.org/10.1016/j.stueduc.2015.01.002

Castrillón, Omar D.; Sarache, William; Ruiz-Herrera, Santiago. (2020). Prediction of academic performance using artificial intelligence techniques. **Formación universitaria**, 13(1), Chile. (Pp. 93-102). https://doi.org/10.4067/S0718-50062020000100093

Camborda Zamudio, Maria (2014). **Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de ingeniería civil de la Universidad Continental**. Tesis para optar por el grado académico de Magíster en Ingeniería de Sistemas, Universidad Nacional del Centro del Perú, Escuela de Posgrado, Huancayo. Perú. Extraído de: http://repositorio.uncp.edu.pe/handle/20.500.12894/1477

Candia Oviedo, Dennis Iván (2019). **Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático**. Tesis para obtener el grado académico de Maestro en Informática, Universidad Nacional de San Antonio Abad del Cusco, Escuela de Posgrado, Cusco, Perú. Extraído de: http://repositorio.unsaac.edu.pe/handle/20.500.12918/4120

Cerda, Jaime; Villarroel, Luis (2008). Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. **Revista chilena de pediatría**, 79(1), Chile. (Pp. 54-58). https://dx.doi.org/10.4067/S0370-41062008000100008

Chávez Uribe, Alfonso. (2006). **Bienestar psicológico y su influencia en el rendimiento académico de estudiantes de nivel medio superior**. Tesis de maestría. Facultad de Psicología, Universidad de Colima, México.

Córdoba Fallas, Luis (2011). Weka. Minería de Datos. Recuperado de: http://cor-mineriadedatos.blogspot.com/2011/06/weka.html

Cuji, Blanca; Gavilanes, Wilma; Sanchez, Rina. (2017). Modelo predictivo de deserción estudiantil basado. **Espacios**, *55*(38), Venezuela (Pp. 17). Obtenido de Espacios: https://www.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

García Tinisaray, Daysi (2015). **Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en learning analytics (análisis de aprendizaje) mediante el uso de técnicas multivariantes**. Tesis doctoral, Universidad de Sevilla, Sevilla, España. Extraído de: https://idus.us.es/handle/11441/40436

Helal, Sumyea, Li, Jiuyong; Liu, Lin; Ebrahimie, Esmaeil; Dawson, Shane; Murray, Duncan J.; Long, Qi. (2018). Predicting academic performance by considering student heterogeneity. **Knowledge-Based Systems**, 161, Holanda. (Pp. 134-146). https://doi.org/10.1016/j.knosys.2018.07.042

Hernández Sampieri, Roberto; Fernández Collado, Carlos y Baptista Lucio, Pilar (2018). **Metodología de la investigación**. (Sexta ed.). McGraw Hill. México.

Hussain, Sadiq; Abdulaziz Dahan, Neama; Ba-Alwi, Fadl Mutaher; Ribata, Najoua. (2018). Educational data mining and analysis of students' academic performance using weka. **Indonesian Journal of Electrical Engineering and Computer Science**, 9(2), Indonesia. (Pp. 447-459). https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Katarya, Rahul; Gaba, Jalaj; Garg, Aryan; Verma, Varsha. (2021). A review on machine learning based student's academic performance prediction systems. **2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)**, (Pp. 254-259). Coimbatore, India: IEEE. https://doi.org/10.1109/ICAIS50930.2021.9395767

Kumar Yadav, Surjeet; Baharadwaj, Brijesh; Pal, Saurabh. (2012). Data Mining Applications: A comparative Study for Predicting Student's performance. **International Journal of Innovative Technology & Creative Engineering,** 1 (1), UK (Pp. 13-19). https://arxiv.org/abs/1202.4815v2

Mandelman, Samuel D.; Barbot, Baptiste; Grigorenko, Elena L. (2016). Predicting academic performance and trajectories from a measure of successful intelligence. **Learning and Individual Differences**, 51, UK (Pp. 387-393). https://doi.org/10.1016/j.lindif.2015.02.003

Menacho Chiok, Cesar Higinio (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. **Anales Científicos**, 78(1), Perú. (Pp. 26-33). http://dx.doi.org/10.21704/ac.v78i1.811

Microsoft. (2019). *Data Mining Concepts*. Recuperado el 27 de diciembre de 2019. https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-concepts

Miguéis, Vera; Freitas, Ana; Garcia, Paulo; Silva, André. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. **Decision Support Systems**, 115, Holanda. (Pp. 36-51). https://doi.org/10.1016/j.dss.2018.09.001

Ministerio de Educación. (2009). Cómo rinden los estudiantes peruanos en comunicación y matemática: Resultados de la evaluación nacional 2009 informe descriptivo. Lima.

Ministerio de Educación. (2016). Currículo Nacional de la Educación Básica. Lima.

Mitchell, Tom (2000). Decision Tree Learning. Extraído de https://bit.ly/2GqqYNq

Orihuela Maita, Gerson Yovanni. (2019). **Aplicación de Data Science para la predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú**. Tesis de pregrado. Universidad Nacional del Centro del Perú, Facultad de Ingeniería de Sistemas, Huancayo, Perú. Extraído de: http://repositorio.uncp.edu.pe/handle/20.500.12894/5837

Redondo Rojo, Jesus M. (1997). La dinámica escolar: de la diferencia a la desigualdad. **Revista de Psicología**, 6, Chile. (Pp. Pág. 7-18). https://doi.org/10.5354/0719-0581.1997.18656

Timarán-Pereira, Ricardo; Caicedo-Zambrano, Javier; Hidalgo-Troya, Arsenio. (2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. **Revista de Investigación, Desarrollo e Innovación**, 9 (2), Colombia. (Pp. 363-378). https://doi.org/10.19053/20278306.v9.n2.2019.9184

Urbina-Nájera, Argelia. (2021). Variables que influyen en el rendimiento de los estudiantes de posgrado: Una perspectiva desde la analítica del aprendizaje. **Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales**, 23 (1), Venezuela. (Pp.36-50). DOI: www.doi.org/10.36390/telos231.04

Valencia Cárdenas, Marisol; Correa Morales, Juan Carlos; Díaz Serna, Francisco Javier. (2015). Métodos estadísticos clásicos y bayesianos para el pronóstico de demanda. Un análisis comparativo. **Revista de la Facultad de Ciencias**, 4 (1), Colombia. (Pp. 52-67). https://doi.org/10.15446/rev.fac.cienc.v4n1.49775

Witten, Ian; Frank, Eibe; Hall, Mark; Pal, Christopher. (2016). Data mining: Practical Machine Learning Tools and Techniques. 4th ed. Morgan Kauffman. USA.

Yamao, Eiriku. (2018). **Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la escuela profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres, Lima-Perú**. Tesis de maestría. Universidad de San Martin de Porres, Lima, Perú. Extraído de https://repositorio.usmp.edu.pe/handle/20.500.12727/3555

Yarlequé Wong, Rocio. (2019). **Estilos de aprendizaje en el rendimiento académico de los estudiantes del primer grado de primaria de la institución educativa N° 20320 Domingo Mandamiento Sipan, Huacho, 2018**. Tesis de maestría. Universidad Nacional José Faustino Sánchez Carrión, Huacho, Perú. Extraído de: http://repositorio.unjfsc.edu.pe/handle/UNJFSC/3104