UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

MAESTRÍA EN INGENIERÍA DE SISTEMAS E INFORMÁTICA - ADMINISTRACIÓN DE TECNOLOGÍAS DE INFORMACIÓN

COMPARACIÓN DE ALGORITMOS DE MACHINE LEARNING EN LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO UNIVERSITARIO BASADO EN EL RENDIMIENTO EN EL EXAMEN DE ADMISIÓN DE LOS INGRESANTES A LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DEL AÑO 2023

TESIS

PRESENTADA POR:

NAIN NEPTALÍ ACERO MAMANI

Para optar el Grado Académico de

MAESTRO EN CIENCIAS (*MAGISTER SCIENTIAE*) CON MENCIÓN EN INGENIERÍA DE SISTEMAS E INFORMÁTICA - ADMINISTRACIÓN DE TECNOLOGÍAS DE INFORMACIÓN

TACNA – PERÚ

2025

UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

Escuela de Posgrado

MAESTRÍA EN INGENIERÍA DE SISTEMAS E INFORMÁTICA - ADMINISTRACIÓN DE TECNOLOGÍAS DE INFORMACIÓN

COMPARACIÓN DE ALGORITMOS DE MACHINE LEARNING EN LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO UNIVERSITARIO BASADO EN EL RENDIMIENTO EN EL EXAMEN DE ADMISIÓN DE LOS INGRESANTES A LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DEL AÑO 2023

Tesis sustentada y aprobada el 14 de julio de 2025; estando el jurado calificador integrado por:

PRESIDENTE : .....

Dr. Nataniel Mario Linares Gutiérrez

SECRETARIO : .....

Mgr. Luis Johnson Paúl Mori Sosa

MIEMBRO : .....

M.Sc. Israel Nazareth Chaparro Cruz

ASESOR : .....

M.Sc. Israel Nazareth Chaparro Cruz

CERTIFICADO DE SIMILITUD

Yo, MSc. Israel Nazareth Chaparro Cruz, en mi condición de asesor acreditado con Resolución de Escuela de Posgrado N° 14661-2024-ESPG/UNJBG del 18 de octubre del 2024, del trabajo de tesis titulado: *"Comparación de los algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento del examen de admisión de los estudiantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann en el año 2023"* presentado por el Sr. Nain Neptalí Acero Mamani, para optar el Grado Académico de Maestro en Ciencias (*Magíster Scientiae*) con mención en Ingeniería de Sistemas e Informática – Administración de Tecnologías de Información

Habiendo cumplido con lo establecido en el reglamento de originalidad y de similitud de trabajo de investigación y producción intelectual, considerando que según la revisión, evaluación y análisis realizado a través del software de similitud textual TURNITIN, cuenta con el nivel de similitud permitido cuyo porcentaje es 6%.

Por lo que CERTIFICO LA SIMILARIDAD de la tesis y está de acuerdo al nivel PERMITIDO, para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado a solicitud del interesado con fines de continuar con los trámites respectivos para la obtención del Grado Académico de Maestro en Ciencias (*Magíster Scientiae*) con mención en Ingeniería de Sistemas e Informática – Administración de Tecnologías de Información

Tacna, 29 mayo 2025

FIRMA ASESOR
Nombres y apellidos

.....
MSc. Israel Nazareth Chaparro Cruz
DNI N° 48584646

FIRMA TESISTA
Nombres y apellidos

.....
Sr. Nain Neptalí Acero Mamani
DNI N° 74575544

DEDICATORIA

A mi amado padre, Teofelo Acero Chambilla, cuya presencia sigue iluminando mi vida, aunque ya no esté físicamente con nosotros, cada día siento tu presencia, amor y guía en cada una de mis decisiones. Y aunque el tiempo y la distancia nos separe, siempre te llevo en el corazón.

AGRADECIMIENTO

A Dios, por ser mi fortaleza ante la adversidad.

A mi asesor, Msc. Israel Nazareth Chaparro Cruz, por compartir y guiar con sapiencia el desarrollo de la presente investigación.

A la plana docente de la Maestría en Ingeniería de Sistemas e Informática - Administración de Tecnologías de la Información de la UNJBG por fortalecer nuestro perfil profesional a través de sus orientaciones académicas.

Al proyecto de investigación del que soy miembro tesista: "OPTIMIZACIÓN DEL RENDIMIENTO ACADÉMICO UNIVERSITARIO APLICANDO CIENCIA DE DATOS AL EXAMEN DE ADMISIÓN EN LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN", aprobado por R. R. Nº 13627-2024-UNJBG. Agradezco a los miembros del proyecto por su apoyo y a la Universidad Nacional Jorge Basadre Grohmann por su financiación con fondos del Canon, Sobrecanon y Regalías Mineras.

A mi familia, quienes me inspiran a salir adelante cada día.

Y en especial a mi padre Teofelo Acero Chambilla, por sus valiosos consejos y por enseñarme a nunca rendirme ante los desafíos de la vida.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En el Perú, la Secretaría Nacional de la Juventud (SENAJU, 2023) menciona que los jóvenes entre los 15 a 29 años de edad, solo un 21,4 % han accedido a una educación universitaria en zonas urbanas y 24,7 % en zonas rurales. En el Perú, no se puede detectar con facilidad durante el proceso de admisión, quienes realmente deben seguir estudios universitarios, por lo cual, un número significativo de alumnos no han podido responder a las exigencias que conlleven a tener logros significativos.

El presente estudio aborda la necesidad de predecir el rendimiento académico universitario de los estudiantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann basado en el examen de admisión de los ingresantes en el año 2023. Se formuló el siguiente objetivo: "Comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023".

Se realizó un extenso estudio experimental comparativo utilizando modelos de machine learning como regresión lineal, árbol de decisión, bosques aleatorios y redes neuronales; obteniendo métricas como error absoluto medio (MAE), error cuadrático medio (MSE), raíz de error cuadrado medio (RMSE) y error de porcentaje medio absoluto (MAPE). Se concluyó que: "Se comparó los 2 algoritmos que tuvieron más precisión respecto a predecir el rendimiento académico universitario, luego de compararlos estadísticamente se observó que el modelo de redes neuronales sobresalió por su mayor precisión y capacidad predictiva, con un nivel de confianza al 95 %".

Finalmente, se publicaron el conjunto de datos como: las notas del rendimiento académico y el puntaje en los cursos de admisión, también los modelos de machine learning en la plataforma de gitLab para futuras investigaciones.

Palabras clave: machine learning, métricas, rendimiento académico universitario

ABSTRACT

In Peru, the National Youth Secretariat (SENAJU, 2023) mentions that only 21,4 % of young people between 15 and 29 years of age have access to university education in urban areas and 24,7 % in rural areas. In Peru, the admission process cannot easily detect those who really need to pursue university studies, so a significant number of students have not been able to respond to the demands that lead to significant achievements.

The study addresses the need to predict the university academic performance of the students of the Faculty of Engineering of the Universidad Nacional Jorge Basadre Grohmann based on the entrance exam of the entrants in the year 2023. The objective was formulated as 'To compare machine learning algorithms for predicting university academic performance based on the entrance exam of students entering the Faculty of Engineering of the UNJBG in the year 2023'.

An extensive comparative experimental study was conducted using machine learning models such as linear regression, decision tree, random forests and neural networks; obtaining metrics such as mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE) and mean absolute percentage error (MAPE). It was concluded that: 'We compared the 2 algorithms that had more accuracy with respect to predicting university academic performance, after comparing them statistically it was observed that the neural networks model stood out for its greater accuracy and predictive ability, with a level of confidence at 95 %'.

Finally, the dataset such as: academic performance grades and admission course scores, as well as machine learning models were published on the gitLab platform for future research.

Keywords: machine learning, metrics, university academic performance.

INTRODUCCIÓN

La presente investigación tiene como propósito hallar el mejor modelo de machine learning que pueda predecir el rendimiento académico universitario basado en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann en el año 2023.

La presente investigación se encuentra estructurada en cuatro capítulos, los cuales se detallan a continuación:

En el capítulo I, se describe la realidad problemática; así como se formula el problema, el objetivo y la hipótesis para esta investigación.

En el capítulo II, se ha revisado el estado del arte a través de publicaciones de alto impactó como artículos científicos, tesis relevantes, libros, etc.; se han descrito los antecedentes claves para esta investigación y las bases teóricas respectivas.

En el capítulo III, se ha descrito el tipo y diseño de la investigación, la población y muestra, así como los instrumentos de recolección de datos.

En el capítulo IV, se presentan los resultados que se obtuvo del experimento, las cuales son presentamos en gráficos y tablas informativas.

Finalmente, se presentan las discusiones, conclusiones, recomendaciones y referencias, respectivamente.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del problema

La educación desempeña un papel fundamental en el crecimiento económico de un país, y se considera una de las más importantes inversiones del capital humano, ya que aumenta las capacidades de la fuerza laboral y la capacidad innovadora de la economía, el conocimiento de nuevas tecnologías, productos, procesos y servicios (Benos & Zotou, 2014).

Sin embargo, el bajo rendimiento académico es un problema en todos los países de nuestro entorno cultural y económico, como es el caso de Europa, el tiempo que se invierte para que un estudiante termine la carrera o el abandono de los estudiantes son problemas comunes, este es un tema que no solo preocupa a las autoridades educativas, sino también a los responsables políticos, puesto que los gastos públicos que se realiza en educación no produce los resultados que se esperan obtener (Tejedor et al., 2007).

En el Perú, la secretaría nacional de la juventud (SENAJU, 2023) menciona que los jóvenes entre los 15 a 29 años de edad, solo un 21,4 % han accedido a una educación universitaria en zonas urbanas y 24,7 % en zonas rurales. Según el Instituto Nacional de Estadística e Informática (**INEI**) menciona que la tasa de deserción universitaria se incrementó en 16 puntos, paso de 39,2 % a 55,6 % entre los años 2019 y el 2020, y en el año 2021 la cifra de deserción estudiantes disminuyó en 7 puntos pasando a 48,6 %, según el informe nacional de juventudes 2021.

Según la Superintendencia Nacional de Educación Superior Universitaria (SUNEDU, 2023), el Perú a través de sus 49 universidades públicas y 48 universidades privadas licenciadas, recibe cada año nuevos ingresantes a través del proceso de admisión.

Por lo tanto, no se tiene claro si un estudiante que ingresa a la universidad puede desaprobar o abandonar un curso de especialidad o complementario cuantas veces quiera; como consecuencia, si no logran la nota mínima para aprobar, tienen que volver a llevar el curso, ya que no alcanzaron el objetivo de obtener una nota mínima aprobatoria, la ley

30220 (2014), en su artículo 102 detalla que los estudiantes que desaprueban un curso por tercera vez, son castigados y separados temporalmente de la universidad por 1 año sin derecho a realizar una matrícula, superado este lapso de tiempo pueden volverse a matricular pero solo en los curso que desaprobaron, finalmente se detalla que si el estudiante desaprueba el curso por cuarta vez procede a su retiro definitivo.

En nuestro país, no se puede detectar con facilidad en el proceso de admisión quienes realmente deben seguir estudios universitarios, por lo cual, un número significativo de alumnos no ha podido responder a las exigencias que le conduzcan a tener logros significativos. En la Universidad Nacional Jorge Basadre Grohmann, el rendimiento académico de los estudiantes se considera relativamente bajo, respecto a las 7 facultades: Facultad de Ciencias, Facultad de Ciencias de la Salud, Facultad de Ingeniería, Facultad de Ciencias Agrícolas, Facultad de Ciencias Jurídicas, Facultad de Educación, Comunicación y Humanidades, Facultad de Ingeniería Civil, Arquitectura y Geología. Cada año los estudiantes disminuyen su rendimiento académico y en algunos casos se tienen que bajar los niveles de exigencia (Chávez & Mendoza, 2019).

El machine learning se ha convertido en uno de los temas más importantes entre las organizaciones que buscan aprovechar sus datos para tener un nuevo nivel de comprensión, basado en una variedad de algoritmos que aprenden de los datos para describirlos, mejorarlos, y predecir resultados. Si se utiliza los datos adecuados, los modelos de machine learning tienen la oportunidad de predecir mejores resultados. Estos modelos son los resultados que se generan al entrenar los algoritmos de machine learning con datos, después haberlos entrenado, cuando se proporciona una entrada a un modelo, se le dará una salida (Hurwitz & Kirsch, 2018).

Los algoritmos de machine learning tienen muchas aplicaciones en el sector de educación, tales como la percepción docente, percepción estudiantil, rendimiento académico, pensamiento computacional, en otras aplicaciones, estos algoritmos dan buenos resultados a problemas complejos del sector educación de acuerdo a grandes volúmenes de datos, a través de una buena limpieza de datos y procesamiento de información pueden generar predicciones efectivas (Forero & Negre, 2023).

Diferentes tipos de investigaciones (Saire, 2023) que solucionaron problemas relacionados a la educación, centran la investigación en el sector universitario, con el objetivo de poder predecir el rendimiento académico universitario de los ingresantes a la educación superior.

En la Universidad Nacional Jorge Basadre Grohmann (**UNJBG**) la aplicación del machine learning podría ayudar a identificar patrones y/o factores relevantes en el proceso de admisión que impactan en el rendimiento académico universitario.

1.2. Formulación del problema

Arias (2012) menciona que la formulación del problema es: "la concreción del planteamiento en una pregunta precisa y delimitada en cuanto a espacio, tiempo y población (si fuere el caso)" (p.41).

1.2.1 Problema principal

¿Es posible comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?

1.2.2. Problemas secundarios

a) ¿Es posible preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?
b) ¿Es posible construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?
c) ¿Es posible evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?

1.3. Justificación e importancia de la investigación

1.3.1. Justificación tecnológica

Hoy en día las áreas de economía, el comercio y medicina, están generando información histórica y están presentando diversos cambios en la incursión de la analítica de datos, tanto ha sido el cambio que ha permeado en el sector de la educación, donde se empiezan a analizar grandes volúmenes de datos para poder solucionar problemas de la índole académica tales como la mejora de aprendizaje de los estudiantes, la deserción, el abandono y el rendimiento académico.

La incursión de la tecnología en el sector de educación para poder maximizar el aprendizaje de los estudiantes y aspectos relacionados, se han definido conceptos relacionados a la analítica en el sector de educación como la minería de datos, la analítica académica y aprendizaje, conceptos muy relacionados que van de la mano con el aprendizaje automático, con el fin de convertir datos en información que permita tomar acciones previas o fomentar la enseñanza (Contreras et al., 2020).

El machine learning es una herramienta tecnológica más actual que pueden ser capaz de gestionar y utilizar grandes volúmenes de datos para poder predecir el rendimiento académico universitario basado en el rendimiento del examen de admisión de los estudiantes de la Facultad de Ingeniería de la UNJBG, se aprovechará al máximo esta tecnología para poder comparar estos algoritmos y encontrar el mejor algoritmo predictor.

1.3.2. Justificación económica

Los responsables políticos invierten presupuesto en educación como gasto público, pero no se llega a los resultados que se desea obtener, el excesivo tiempo y dinero invertido en un estudiante para que pueda terminar la carrera y proceder a su titulación son problemas comunes en nuestro entorno cultural y económico (Tejedor et al., 2007).

Con la siguiente investigación busca predecir el rendimiento académico universitario que tendrá un estudiante ingresante a la Facultad de Ingeniería de la UNJBG,

en base del rendimiento del examen de admisión, la cual se anticipará a un posible abandono de la carrera profesional o una posible deserción del estudiante.

1.3.3. Justificación metodológica

La metodología CRISP-DM se utiliza en proyectos de análisis de datos dentro una organización, la cual proporciona un marco de trabajo riguroso a diferencia que los modelos utilizados para el desarrollo de software que son en cascada tradicional, se dividen en seis fases distintas (comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación, despliegue), se deben definir las tareas en cada una de estas fases y el resultado que se espera (Sharma et al., 2017).

Para poder comparar los algoritmos de machine learning la metodología para la siguiente investigación fue de CRISP-DM, esto ayudará a tener un mejor orden y poder llegar al objetivo planteado, para que finalmente las conclusiones de la presente investigación sirvan para futuras investigaciones.

1.4. Objetivos de la investigación

Asimismo, Arias (2012) menciona que el objetivo de investigación es: "un enunciado que expresa lo que se desea indagar y conocer para responder a un problema planteado" (p.43).

1.4.1. Objetivo general

Comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

1.4.2. Objetivos específicos

Según Arias (2012) indica que los objetivos específicos "indican con precisión los conceptos, variables o dimensiones que serán objeto de estudio. Se derivan del objetivo general y contribuyen al logro de este" (p.45).

a) Preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.
b) Construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.
c) Evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

1.5. Hipótesis

Según Hernández et al. (2014) indica que no todas las investigaciones cuantitativas llevan hipótesis, si se formulan o no va a depender del alcance que se tenga en el estudio.

Las investigaciones cuantitativas que deben formular hipótesis son las que tienen un alcance correlacional o explicativo, o las que tienen alcance descriptivo pero que intentan pronosticar una cifra o un hecho.

Existen diversas formas de poder clasificar una hipótesis, para esta investigación se utilizó la **hipótesis de investigación**, ya que estas hipótesis se definen como: "proposiciones tentativas..." (Hernández et al., 2014).

1.5.1. Hipótesis general

Si es posible comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

1.5.2. Hipótesis específicas

a) Si es posible preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de

admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

b) Si es posible construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.
c) Si es posible evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

1.6. Variables

1.6.1. Identificación de variables

Variable independiente: Algoritmos de machine learning.
Variable dependiente: Predicción del rendimiento académico universitario.

1.6.2. Definición conceptual de las variables

Variable Independiente: Algoritmos de machine learning

Por su naturaleza: Cuantitativa
Por su escala de medición: Numérica de razón

Variable Dependiente: Predicción del rendimiento académico universitario

Por su naturaleza: Cuantitativa
Por su escala de medición: Numérica de razón

1.6.3. Definición operacional de las variables

En la siguiente investigación se realizó la comparación de los algoritmos de machine learning (variable independiente) que puedan predecir mejor el rendimiento académico universitario (variable dependiente) basado en el examen de admisión.

Operacionalización de las variables

Variables	Indicador	Ítem	Técnica
Variable Independiente Algoritmos de machine learning			
Variable Dependiente Predicción del rendimiento académico universitario	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Técnica: Observación Instrumento: Ficha digital de observación
	MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	
	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	
	MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert x\ 100$	

Nota. Elaboración propia.

1.7. Limitaciones de la investigación

Entre las limitaciones para la siguiente investigación se mencionan a continuación:

- Se cuenta con datos sólo de ingresantes del año 2023 en las modalidades de FASE - I, FASE - II, CEPU OTOÑO (CEPU-I), CEPU INVIERNO (CEPU-II) y CEPU VERANO (CEPU-III) de la Facultad de Ingeniería para realizar la predicción.
- Solo se utilizó el promedio ponderado de los primeros años de los estudiantes de la Facultad de Ingeniería para realizar la predicción.

CAPÍTULO II

MARCO TEÓRICO

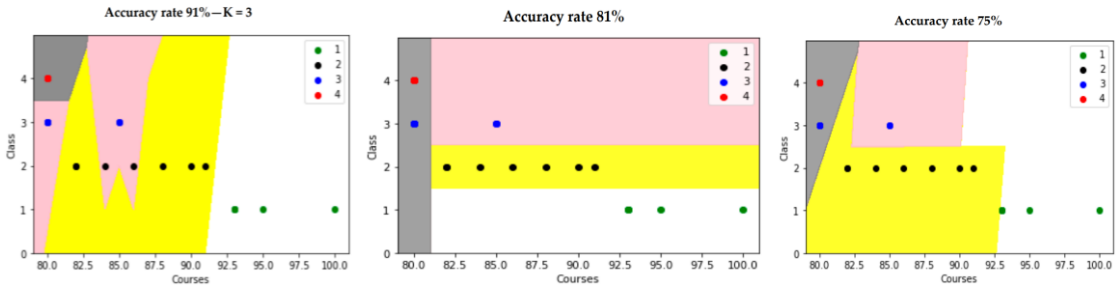
2.1. Antecedentes de la investigación

2.1.1. Antecedentes internacionales

Assiri et al. (2024). En el artículo publicado titulado: "Enhanced Student Admission Procedures at Universities Using Data Mining and Machine Learning Techniques". El estudio en mención realiza un análisis del coeficiente de correlación y un análisis de la distribución de datos para comprender las relaciones entre las características de admisión : promedio general de preparatoria (GPAH), el puntaje de aptitud general (GAT), el puntaje de logro (AT) y el rendimiento académico de los estudiantes mediante los algoritmos de machine learning como: k-nearest neighbor, decision tree, and support vector machine; la cual llega a la siguiente conclusión: "los resultados muestran que los procedimientos de admisión actuales no están ajustados para todas las carreras, ya que las relaciones varían de una carrera a otra"

Figura 1

Grafica de tasa de precisión de los modelos



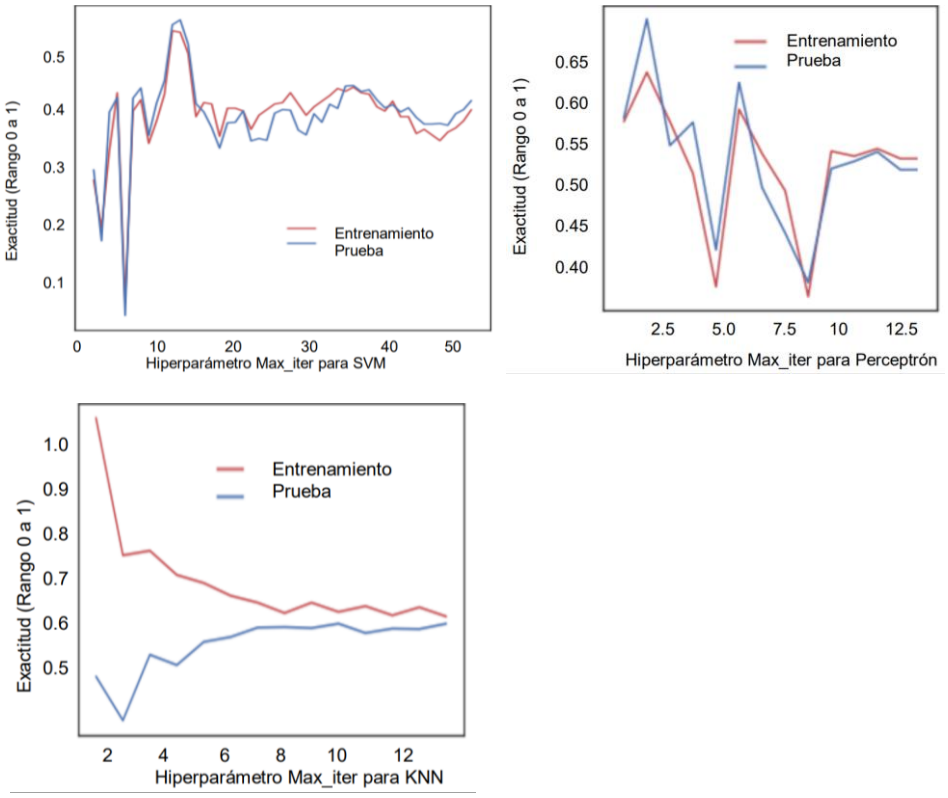
Nota. Por Assiri et al. (2024)

En el artículo se menciona que al usar el modelo k-nearest neighbors, se obtuvo una precisión del 100 % cuando el valor de K=1. Sin embargo, al aumentar K, la precisión disminuyó; por ejemplo, con K=3, la precisión del modelo bajó al 91 %, como también se entrenó con los mismos datos el modelo de decision tree donde se obtuvo una precisión del 81 % y para el modelo support vector machine su precisión fue del 75 %.

Contreras et al. (2020). En el artículo publicado titulado: "Academic performance prediction by machine learning as a success/failure indicator for engineering students"; la cual se utilizó un conjunto de datos de 1620 estudiantes que se matricularon en la escuela de Ingeniería Industrial (Colombia). En este estudio se plantea una propuesta para poder seleccionar las variables que más influyen en la predicción del rendimiento académico de los estudiantes, para ello, se implementaron algoritmos como: decision tree, k-nearest neighbors, perceptrón y otros, las cuales luego de ser comparados se concluye que: "las variables que más influyen en el rendimiento académico de los estudiantes de ingeniería son: edad, género, puntaje ICFES para aptitud matemática, puntaje global ICFES, valor de matrícula y puntaje ICFES para condición matemática y cohorte".

Figura 2

Búsqueda del mejor hiperparámetro de los algoritmos



Nota. Por Contreras et al. (2020)

Cada uno de los algoritmos que se usó en el artículo mencionado tienen su propio hiperparámetro que son valores que hay que identificar, probar y modificar para obtener mejores resultados. Por ejemplo, en el algoritmo de perceptrón el mejor hiperparámetro

es de max_iter = 2, o como también en el algoritmo de k-nearest neighbors (KNN) el hiperparámetro a evaluar es la cantidad de vecinos más cercanos, como se observa en la imagen (Figura 2) da buenos valores cuando el hiperparámetro es 10, 12 y 15.

Contreras et al. (2020) menciona que el mejor algoritmo que dio mejores resultados fue el de perceptrón para la determinación del rendimiento académico con una exactitud del 66,4 %, sin embargo, para mejorar los resultados de las métricas de evaluación de los algoritmos de machine learning es necesario utilizar otras variables que influyen al rendimiento académico como: factores de gestión académica universitaria, tecnológicos, de biblioteca, institucionales, pedagógicos e intelectuales; ya en dicho artículo se utilizan datos de entrada como: factores académicos pre-universidad, demográfico y socio-económicos.

Mengash (2020). En el artículo publicado titulado "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems"; la cual busco predecir el rendimiento académico de los estudiantes de una universidad pública basándose principalmente en la puntuación media ponderada de tres criterios de admisión.

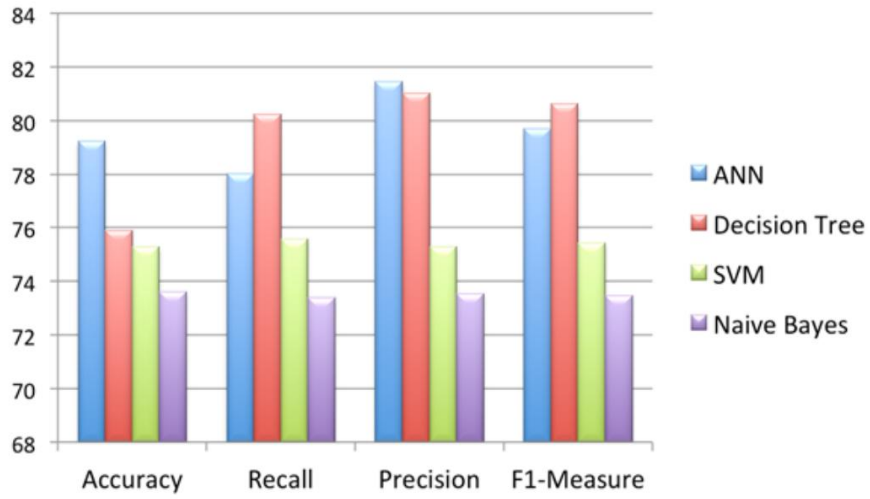
- **General aptitude test (GAT):** evalúa las habilidades matemáticas y verbales para medir la comprensión, el razonamiento lógico, la resolución de problemas y las habilidades inductivas/deductivas de los estudiantes.
- **Scholastic achievement admission test (SAAT):** evalúa la comprensión, la aplicación y la inferencia en cinco materias: biología, química, física, matemáticas e inglés.
- El promedio de calificaciones de la escuela secundaria. (HSGA)

En la investigación mencionada se utilizó un conjunto de datos de 2039 estudiantes que se matricularon en la Facultad de Ciencias de la Computación e Información en los años 2016 a 2019. Se utilizaron 4 algoritmos como redes neuronales artificial (ANN), decision tree, support vector machine (SVM) y naive bayes; se tuvo como conclusión que scholastic achievement admission test (SAAT) predice con mayor

precisión el rendimiento académico, por lo tanto, se debe darle más importancia en admisión.

Figura 3

Comparison of the classifier performance among all the applied classifiers



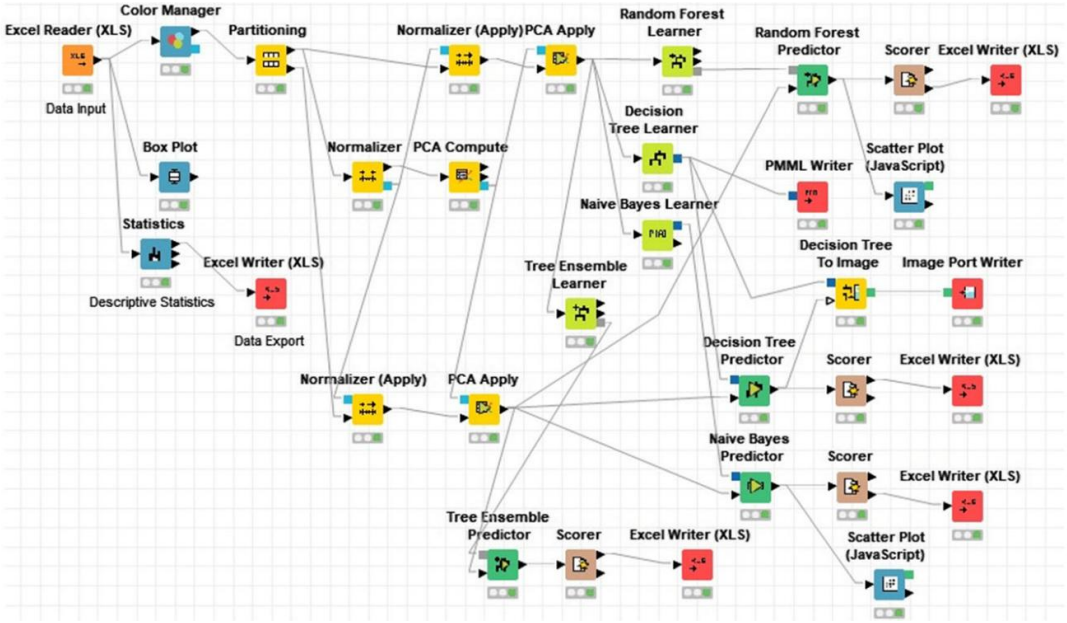
Nota. Por Mengash (2020).

En la figura 3 se observa como el algoritmo ANN tiene el mejor desempeño en exactitud (79 %) y precisión (81 %), sin embargo, el algoritmo de decision tree sobresale en las métricas de recall (80 %) y F1-measure (81 %), la cual estos algoritmos de machine learning son herramientas útiles para predecir el rendimiento académico temprano de los estudiantes en base al sistema de admisión.

Adekitan y Noma (2019). En el artículo publicado titulado: "Data mining approach to predicting the performance of first year student in a university using the admission requirements"; este estudio busco la relación mediante minería de datos el rendimiento académico de los estudiantes en el primer año medido por su clase de grado y el promedio de calificaciones real con las características cognitivas de ingreso de los estudiantes en el momento de la admisión, las cuales fueron: la edad de ingreso de los estudiantes, la puntuación total de WAEC, la puntuación de JAMB, la puntuación de CUSAS basada en la universidad. En esta investigación se analizaron 1445 registros de estudiantes de 2005 a 2009 de la Universidad Covenant en Nigeria. Para esta investigación las plataformas utilizadas fueron KNIME y orange como se describe a continuación:

Figura 4

The KNIME workflow

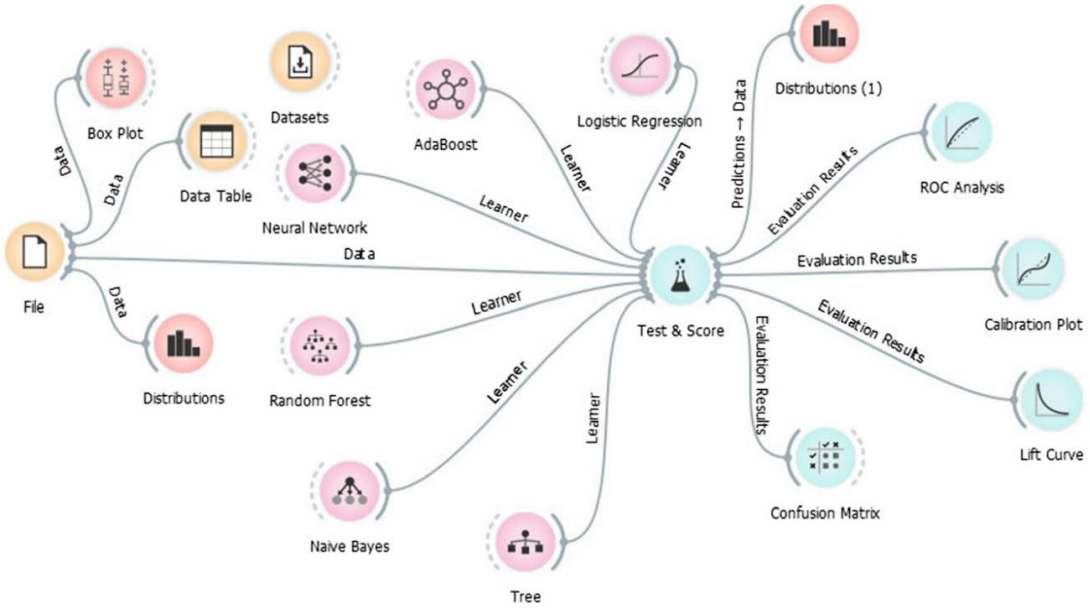


Nota. Por Adekitan y Noma (2019).

En la siguiente investigación se utilizó la plataforma KNIME con los siguientes algoritmos de machine learning: random forest, tree ensemble, decision tree, naive bayes, logistic regression, resilient backpropagation multi-layer perceptron (Rprop MLP); la cual la muestra se dividió en 70 % para entrenamiento y 30 % para evaluación teniendo en cuenta que se aplicó reducción de dimensiones con análisis de componentes principales para mejorar la precisión. Como resultado máximo se tubo al algoritmo de logistic regression con una precisión de 50,23 % y como mínimo el algoritmo de decision tree con una precisión del 39,63 %, dando a conocer la importancia de considerar factores adicionales no académicos en las decisiones de admisión.

Figura 5

The orange model workflow



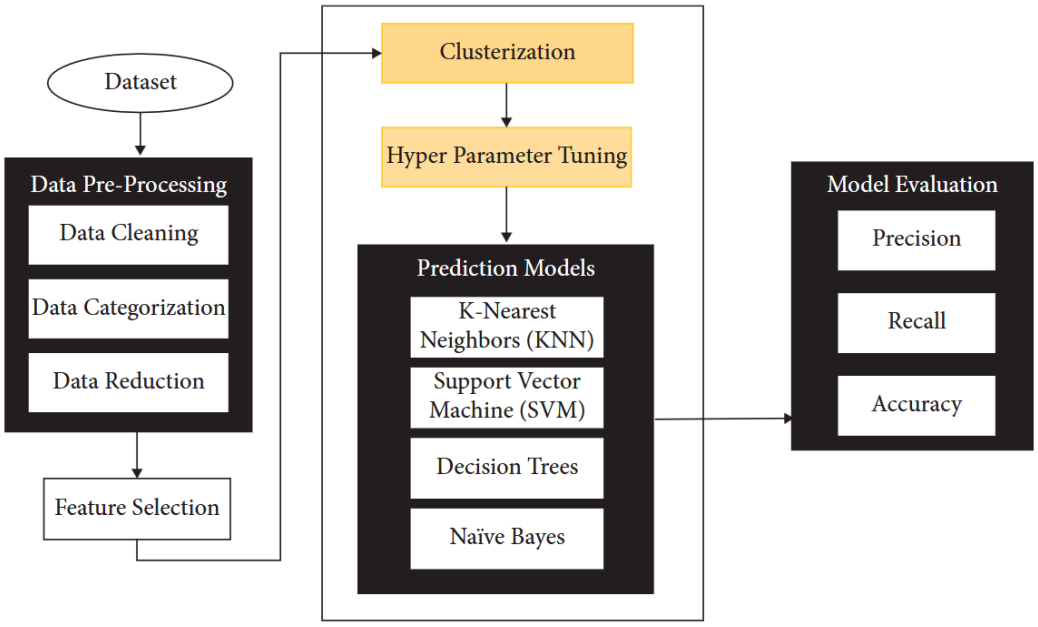
Nota. por Adekitan y Noma (2019).

La segunda plataforma que se utilizó en la investigación mencionada fue orange analytics, con los siguientes algoritmos de machine learning: decision tree, random forest, neural network, naive bayes, logistic regression, adaBoost; la cual se utilizó el muestreo estratificado con validación cruzada de 10 pliegues para la validación. Como resultado máximo se tubo al algoritmo de neural network con una precisión del 51,9 % y como mínimo el algoritmo de adaBoost con una precisión del 42,8 %, las cuales mostraron una relación muy débil entre los requisitos de admisión y el rendimiento académico.

Ahmed y Al (2024). En el artículo publicado titulado " Student Performance Prediction Using Machine Learning Algorithms"; la cual se utilizó un conjunto de datos de registros académicos de estudiantes de los años académicos 2017 a 2022, que consta de 32,582 registros de estudiantes como muestra para el estudio, tras eliminar datos faltantes y duplicados, se tiene como muestra final a 32,005 registros de la Universidad de Wollo y del Instituto de Tecnología Kombolcha en Etiopía. El objetivo de este estudio fue identificar factores influyentes en el rendimiento académico y predecir el éxito o fracaso estudiantil mediante modelos de aprendizaje automático.

Figura 6

Workfow of the proposed methodology



Nota. Por Ahmed y Al-Omari (2024).

El artículo propone un modelo con cuatro componentes principales: preprocesamiento de datos, ajuste de hiperparámetros, predicción del modelo y evaluación del modelo. Los datos fueron recolectados de la Universidad de Wollo y del Instituto de Tecnología Kombolcha en Etiopía, y se sometieron a un preprocesamiento dividido en tres etapas: limpieza, categorización y reducción, seguido de la extracción de características relevantes como género, región, resultados de admisión, intentos previos, créditos estudiados y discapacidades. Seguidamente, se utilizó el algoritmo de agrupamiento k-means para clasificar a los estudiantes y predecir su rendimiento académico.

También se utilizó la técnica grid search, que prueba todas las combinaciones posibles de los valores predefinidos para los hiperparámetros, con el objetivo de encontrar el parámetro ideal para los algoritmos de machine learning, aunque tiene como desventaja su elevado costo computacional para grandes volúmenes de datos.

Ahmed & Al (2024). Finalmente, se construyeron modelos de predicción/clasificación como support vector machine (SVM), árboles de decisión, k-

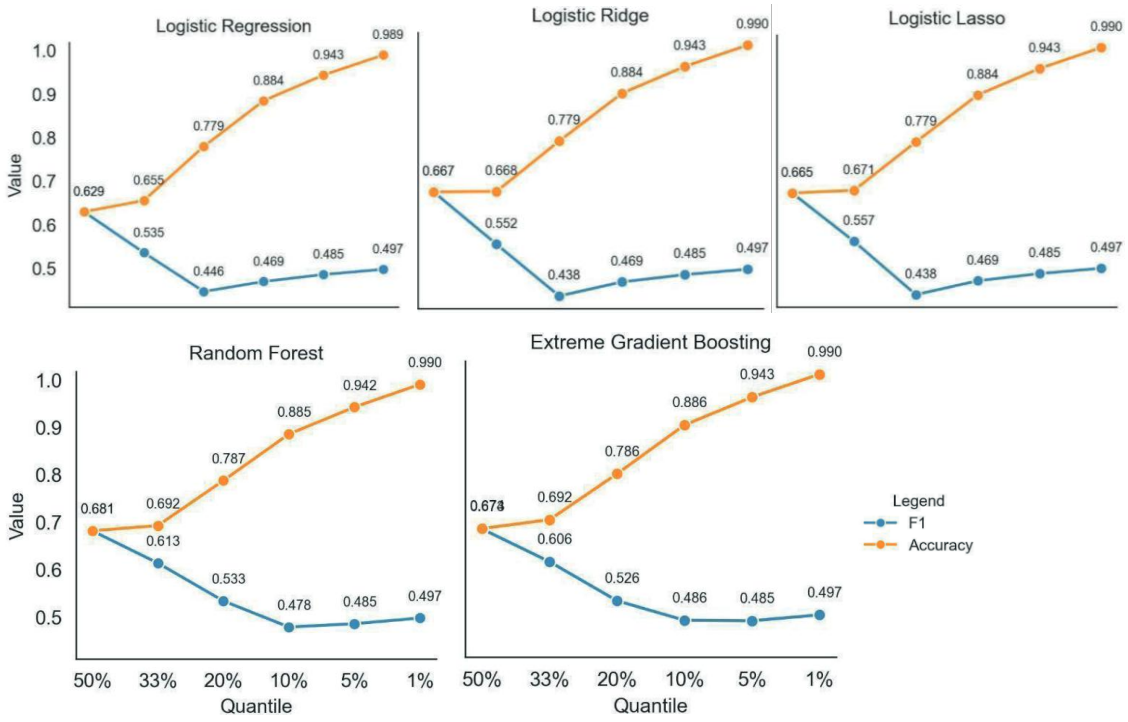
nearest neighbors (KNN) y naive bayes, y se evaluaron mediante las métricas de precisión, recall y accuracy. En este estudio se utilizó la validación cruzada k-fold repetida como técnica central para evaluar la eficacia de los modelos predictivos, aplicando 10-fold cross-validation al dividir los datos en 10 subconjuntos. El uso del hiperparámetro correcto ayudó a mejorar la precisión de los algoritmos, logrando que support vector machine mejorara de 95,4 % a 96,0 %, decision tree de 90,9 % a 93,4 %, naive bayes de 77,3 % a 83,3 % y k-nearest neighbors de 85,3 % a 87,3 %.

2.1.2. Antecedentes nacionales

Salas et al. (2024). En el artículo publicado titulado: "Predicting undergraduate academic performance in a leading Peruvian university: A machine learning approach"; la cual se utilizó un conjunto de datos de 3513 observaciones y 132 variables, enfocándose en el rendimiento académico de estudiantes matriculados en la PUCP desde 2018-1 hasta 2020-1. Además, la base de datos abarca información de desempeño universitario, resultados de pruebas de admisión, rendimiento en la educación secundaria e indicadores de desempeño histórico para sus respectivas escuelas; como también información demográfica, geográfica y socioeconómica, capturando los antecedentes y el contexto de los estudiantes. En este estudio se plantea crear un modelo predictivo basado en machine learning que identifique factores clave del rendimiento académico y permita predecir el desempeño futuro de los estudiantes para ello, se implementaron algoritmos como regresión logística, ridge regression, lasso regression, random forest y extreme gradient boosting (XGBoost).

Figura 7

Quantiles versus accuracy and F1 for trained models



Nota. Por Salas et al. (2024).

En la figura 7 se puede observar las métricas de f1 y accuracy cambian al predecir el rendimiento académico de los estudiantes según los distintos umbrales de cuantiles. Los cuantiles (50 %, 33 %, 20 %, 10 %, 5 %, 1 %) dividen a los estudiantes en grupos basados en su desempeño académico, y los modelos intentan predecir qué estudiantes están en el grupo de peor rendimiento. Finalmente, se puede observar que random forest y extreme gradient boosting muestran un rendimiento más robusto, con valores de f1 más altos, especialmente en el cuantil del 33 %, lo que los hace más efectivos para identificar estudiantes con bajo rendimiento académico en un contexto más equilibrado de datos.

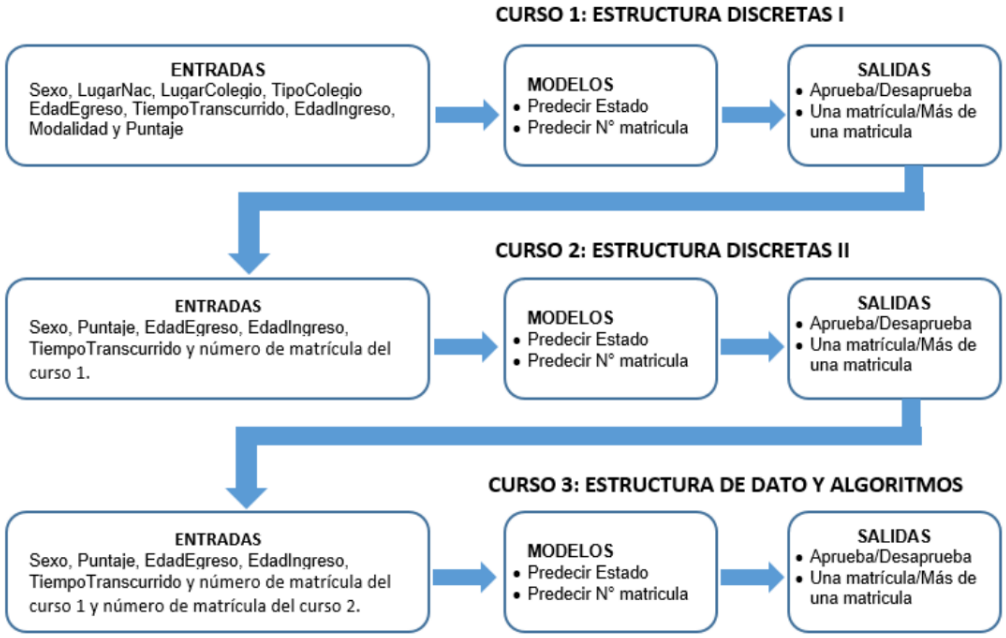
Candia (2019). En la tesis de maestría titulada: "Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático"; tuvo como objetivo general " Predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático"; la metodología utilizada se desarrolló en base al enfoque cuantitativo, de tipo correlacional y no experimental; la muestra estuvo constituida por 12,698 alumnos ingresantes a la

UNSAAC por las diferentes modalidades; se utilizaron 5 de los algoritmos más importantes para este tipo de casos de estudio como son: decision tree J-48, random forest, vecinos más cercanos, función logística, perceptrón multicapa; tuvo como conclusión que los factores claves de los datos de ingreso que determinan el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de los datos de ingreso son: la nota de ingreso, la escuela profesional que se estudia, el semestre, el género y la modalidad de ingreso.

Siare (2023). En la tesis de doctorado titulada: “Predicción de la ruta de rendimiento académico con algoritmos de clasificación”; tuvo como objetivo general “Predecir la ruta del rendimiento académico de los universitarios ingresantes utilizando algoritmos de clasificación”; el enfoque fue cuantitativo, teniendo como alcance o nivel correlacional, el diseño fue no experimental y de tipo transversal; la muestra estuvo constituida por 778 estudiantes de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín de Arequipa, los cuales fueron recolectados del año 2011 al 2020 en la oficina de admisión y la Escuela de Sistemas; los algoritmos de clasificación que mostraron los mejores resultados fueron random forest, xgboost y regresión logística con un promedio de precisión del 89 %; el análisis se desarrolló en base a los cursos de carrera los cuales fueron estructuras discretas I, estructuras discretas II, estructura de datos y algoritmos. Se ha concluido que se ha podido predecir la ruta del rendimiento académico de los estudiantes ingresantes a la Escuela de Sistema con una precisión aproximada del 89 %, las cuales se ha determinado que las variables más influyentes son sexo, puntaje de ingreso, edad de egreso, tiempo transcurrido y edad de ingreso.

Figura 8

Flujo de datos para la predicción



Nota. Por Saire (2023).

En la figura 8 se observa la predicción del rendimiento académico universitario, pero en base a cursos a diferencia de otras investigaciones que se utilizaba el promedio ponderado, cursos como estructuras discretas I, estructuras discretas II y estructura de datos & algoritmos; donde las entradas incluyen características personales y académicas de los estudiantes, como sexo, puntaje, edad de ingreso y egreso, entre otros. A partir de estas variables, los modelos predictivos estiman dos salidas: el estado del curso (aprobado/desaprobado) y el número de matrículas necesarias para aprobar. Este proceso permite identificar tempranamente patrones de desempeño académico, facilitando estrategias de apoyo para evitar el bajo rendimiento y fomentar el éxito estudiantil. En la tesis doctoral mencionada, se llega al siguiente resultado: el modelo integrador representa a los modelos clasificadores del estado (aprobado/desaprobado) y los modelos que clasificadores del número de intentos de los estudiantes, se han logrado precisiones en promedio mayor a 86 %, por lo tanto, el modelo predice el rendimiento académico de los estudiantes ingresantes a la carrera de Ingeniería de Sistemas en base a los datos de admisión y datos académicos.

Yamao (2018). En la tesis de maestría titulada: “Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de las Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú”; tuvo como objetivo general “Predecir el 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”; el enfoque fue cuantitativo, teniendo como tipo de investigación explicativa y correlacional; el diseño de la investigación es transeccional del tipo correlacional causal, ya que el autor se centró sólo en describir la relación que existe entre el rendimiento académico y los factores social, económico y académico de los ingresantes. La muestra estuvo constituida por los estudiantes ingresantes de los periodos 2010-I a 2015-II a la carrera de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres, que hacen un total de 1304 ingresantes admitidos. En la tesis de maestría mencionada, se realizaron predicciones a través de tres técnicas: regresión lineal, decision tree y support vector machine (SVM), y el mejor resultado fue del algoritmo c5,0 de decision tree con una exactitud de predicción de 82,87 %. Finalmente se menciona que, “los que más influyeron en el rendimiento académico fueron los siguientes: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios”.

Arana (2021). En la tesis de doctorado titulada: “Modelo de predicción del éxito académico de los procesos de admisión con criterios múltiples empleando herramientas de Machine Learning”; tuvo como objetivo general “Determinar la influencia del proceso de admisión con criterios múltiples en el éxito académico de los postulantes de la UNCP mediante un modelo predictivo basado en herramientas de machine learning”; el enfoque fue cuantitativo, teniendo como tipo de investigación aplicada y de nivel explicativo-predictivo; el diseño de investigación fue no experimental, longitudinal con diseño de cohorte. La muestra estuvo constituida por todos los postulantes ingresantes a la Universidad Nacional del Centro del Perú durante los periodos 2020-I, 2020-II, 2021-I y 2021-II; los cuales ascienden a 11 466,00 ingresantes. En esta investigación utilizó los siguientes indicadores para entrenar los modelos las cuales fueron: la nota de ingreso, todas las modalidades de ingreso, la preparación académica, edad del ingresante, participaciones deportivas y artísticas al ingresar a la universidad, los créditos aprobados

y el promedio obtenido al finalizar el primer semestre académico. La predicción fue realizada mediante tres algoritmos de machine learning: regresión lineal, árbol de decisiones y clasificación; donde el resultado fue 95,44 % de predicción con el modelo de clasificación.

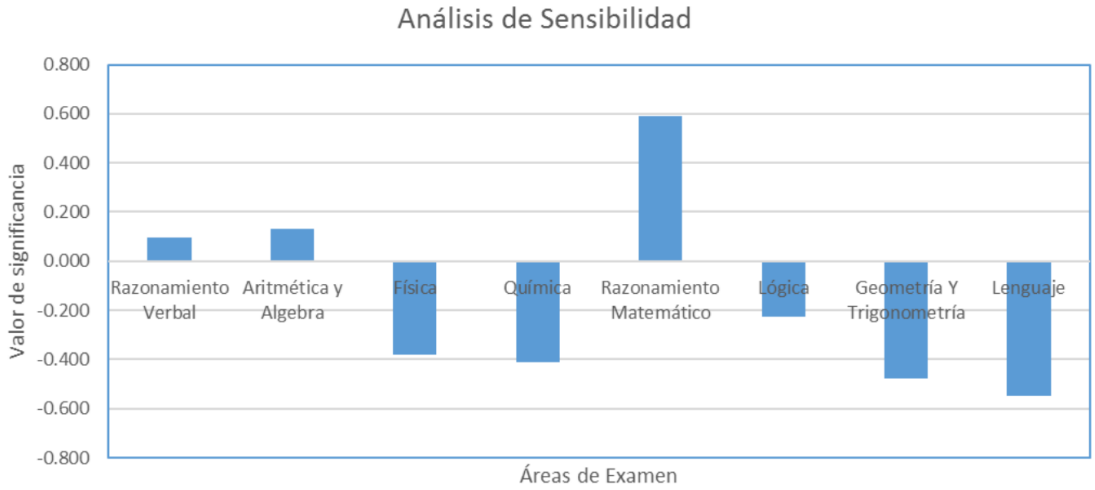
Aronés (2021). En la tesis titulada: “Predicción del rendimiento académico basado en Machine Learning, Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021”; el tipo de la siguiente investigación es observacional, esta investigación es también del tipo retrospectivo y transaccional, y de diseño no experimental. La muestra estuvo constituida por los registros de información de los alumnos de la Escuela Profesional de Ingeniería de Sistemas desde los semestres académicos 2016-II hasta el 2019-I. Se estableció como variables predictoras a las columnas: colegio, curso, créditos, modalidad curso y docente; las cuales se utilizaron algoritmos de machine learning como regresión logística, random forest, n-nearest neighbors, support vector machine y decision tree. Finalmente se llega a la siguiente conclusión: “el algoritmo de Regresión Logística brindó una probabilidad predictiva del 73,6 % según la métrica de validación de la curva ROC, dando como resultado un modelo que predice el rendimiento académico de los alumnos de la Escuela Profesional de Ingeniería de Sistemas”.

2.1.3. Antecedentes locales

Yupanqui (2018). En la tesis titulada: "Análisis Predictivo del Rendimiento Académico en los Alumnos de la Escuela Profesional de Ingeniería en Informática y Sistemas de la UNJBG, utilizando Redes Neuronales, Semestre 2017-I”, tuvo como objetivo general "Realizar un análisis predictivo del rendimiento académico de los alumnos de la Escuela Profesional de Ingeniería en Informática y Sistemas"; el enfoque fue cuantitativo, el diseño no experimental transaccional descriptiva; la muestra estuvo constituida por toda la población la cual estuvo conformada por 69 registros de ingresantes a la Escuela Profesional de Ingeniería en Informática y Sistemas cuya nota promedio en su primer semestre fue mayor a 7,00; el algoritmo que se utilizó fue el retropropagación aplicado en una red neuronal multicapa para el análisis predictivo del rendimiento académico.

Figura 9

Análisis de sensibilidad en la red neuronal de topología 8:3:3:4



Nota. Por Yupanqui (2018).

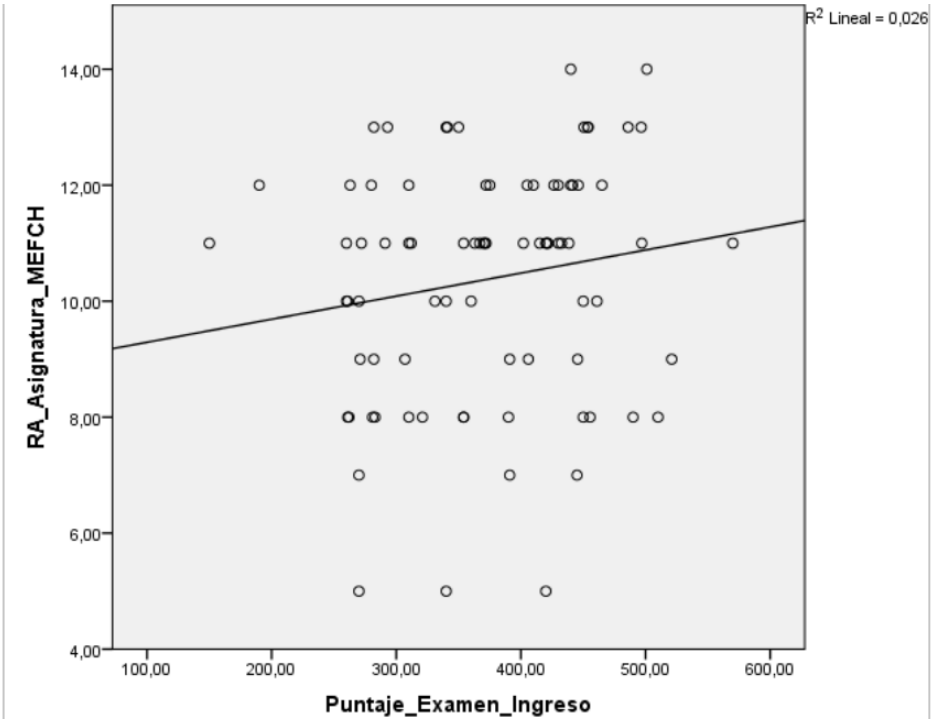
En la figura anterior se destaca la importancia de ciertas áreas evaluadas en el examen de admisión sobre el rendimiento académico de los estudiantes durante su primer semestre. Los resultados indican que responder correctamente en las áreas de razonamiento verbal, aritmética y álgebra, y razonamiento matemático tiene una influencia positiva en el desempeño académico, con pesos de 0,09, 0,13 y 0,59, respectivamente, la cual son los cursos que debe darle más importancia la oficina de admisión.

Contreras (2020). En la tesis titulada: " Relación entre Puntaje del examen de admisión o CEPU, y el rendimiento académico en la asignatura de Morfología, estructura y función del cuerpo humano de estudiantes del Primer año de Odontología de la Universidad Nacional Jorge Basadre Grohmann de Tacna, Año 2016-2017", tuvo como objetivo general " Determinar la relación que existe entre el puntaje de examen de admisión o CEPU, y el rendimiento académico en la asignatura de Morfología, estructura y función del cuerpo humano de estudiantes del primer año de Odontología de la Universidad Nacional Jorge Basadre Grohmann, año 2016-2017"; el tipo de investigación fue básica, así como el diseño fue no experimental, transaccional, descriptivo, correlacional y retrospectivo. Para la muestra se incluyeron como criterios de selección a todos los estudiantes matriculados en el curso de MEFCH durante los años 2016 y 2017. Por otro lado, se excluyeron aquellos estudiantes que no asistieron regularmente, se

retiraron, abandonaron el curso o ingresaron a través de modalidades distintas al examen de admisión o CEPU. Como resultado, la muestra quedó conformada por 82 estudiantes que cuentan con calificaciones registradas en el curso de MEFCH.

Figura 10

Diagrama de dispersión entre el puntaje de examen de admisión o CEPU, y rendimiento académico en la asignatura de morfología, estructura y función del cuerpo humano



Nota. Por Contreras (2020).

Según la figura 10 el puntaje de examen de admisión o CEPU, y rendimiento académico en la asignatura de morfología, estructura y función del cuerpo humano comparten el 2,6 % de la varianza. La cual se llega a la siguiente conclusión "No existe una relación entre puntaje de examen de admisión o CEPU y el rendimiento académico en la asignatura de morfología, estructura y función del cuerpo humano en estudiantes del primer año de odontología de la Universidad Nacional Jorge Basadre Grohmann, en los años 2016 - 2017".

Palacios y Pajares (2019). En el artículo publicado titulado: "Relación entre el rendimiento del examen de admisión y el académico. Tacna, 2001- 2005"; en el cual se

utilizaron un conjunto de datos de ingresantes a la Facultad de Ciencias de la Educación de la UNJBG el año 2001, compuesta por 254 estudiantes distribuidos en cinco especialidades: Especialidad de Lengua, Literatura y Gestión Educativa (LEGE), Especialidad de Idioma Extranjero, Traductor e Intérprete (IETI), Especialidad Ciencias Sociales y Promoción Sociocultural (SPRO), Especialidad de Matemática, Computación e Informática (MACI) y Especialidad de Ciencias Naturales Tecnología y Ambiente (NATA). La cual arribó al siguiente resultado: “Al relacionar estadísticamente el puntaje de ingreso y el rendimiento académico, puede visualizarse que existe relación directa ($p<0,05$) con una baja correlación ($R=0,34$), esto, con escaso significado estadístico, lo que implicaría que no necesariamente el estudiante que ingresa con alto puntaje ha tenido un rendimiento alto. Esto lleva a inferir que el proceso de admisión no cumple con el rol predictor que teóricamente debe tener”.

En el artículo publicado titulado: "Modelo matemático de los factores que influyen en el rendimiento académico de estudiantes universitarios ingresantes"; la cual se utilizó un conjunto de datos en el año 2018, se consideró una población de 1754 estudiantes ingresantes a la UNJBG y una muestra de 1526 estudiantes; como también para el año académico 2019, se consideró una población de 1756 estudiantes ingresantes a la UNJBG y una muestra de 1 455. La cual llega a la conclusión de que la edad no es un factor que influye significativamente en el rendimiento académico, en cambio el puntaje de examen de ingreso si es un factor que influye significativamente en el rendimiento académico, detectado con la correlación de Pearson $r = 0,380$; $r = 0,306$ y una significancia al 99 % para los años 2018 y 2019 respectivamente.

2.2. Bases teóricas

2.2.1. Machine learning

El machine learning según Géron (2019) es: "la ciencia (y el arte) de programar ordenadores para que aprendan a partir de datos", se puede utilizar para:

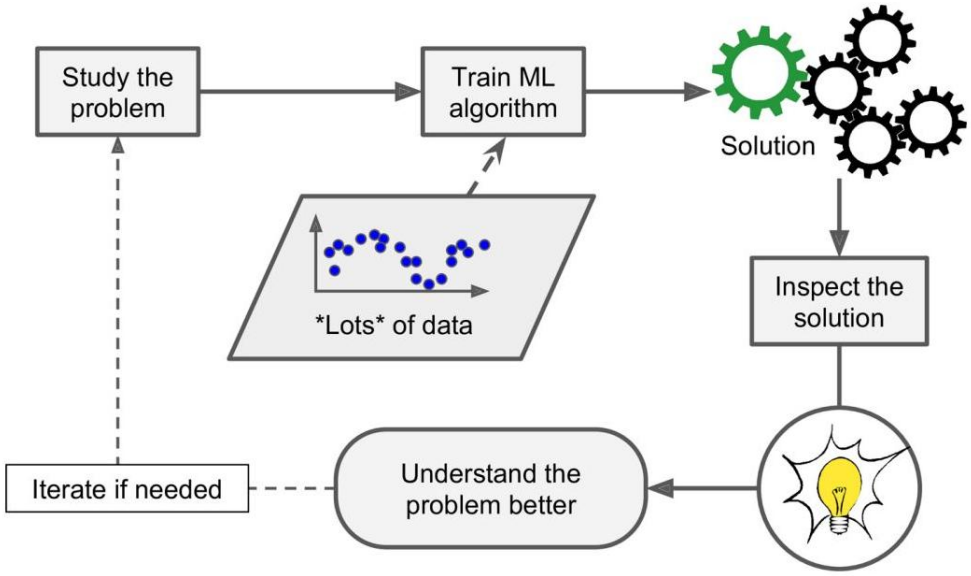
- Problemas grandes y complejos donde el uso de los enfoques tradicionales no ofrece una buena solución.

- Problemas cuyas soluciones requieran una larga lista de reglas o muchos ajustes; a menudo, los algoritmos de machine learning nos ayudan a simplificar código y tener un mejor rendimiento superior al enfoque tradicional.
- Entornos cambiantes, donde el modelo de machine learning puede adaptarse a los nuevos datos.
- Poder entender mejor los problemas complejos donde se tenga grandes cantidades de datos.

El machine learning es una ayuda a los seres humanos a comprender mejor el problema que se está estudiando, nos puede llevar a revelar correlaciones entre variables insospechadas o alguna tendencia nueva que surgieron en los datos, ya que explorar grandes cantidades de datos aplicando técnicas de machine learning puede ayudar a descubrir patrones que no eran perecibles de manera inmediata, por lo tanto, esto llevara a comprender mejor el problema (Géron, 2019)

Figura 11

El machine learning puede ayudar a los humanos a aprender



Nota. Por Géron (2019).

2.2.2. Aprendizaje supervisado

Hurwitz y Kirsch (2018) mencionan que el aprendizaje supervisado empieza con un conjunto de datos establecidos y una comprensión de cómo estos datos se van a

clasificar; tiene como objetivo encontrar ciertos patrones en los datos que puedan aplicarse a un proceso de análisis.

Entre los algoritmos supervisados que se estudian y utilizan en la siguiente investigación son:

- Regresión lineal.
- Árbol de decisión.
- Bosques aleatorios.
- Redes Neuronales.

2.2.2.1. Regresión lineal

Un modelo de regresión lineal hace predicciones simples calculando una suma ponderada de las características de entrada, conjuntamente con una constante ("término de sesgo" o "término de intercepción"), como se detalla a continuación:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

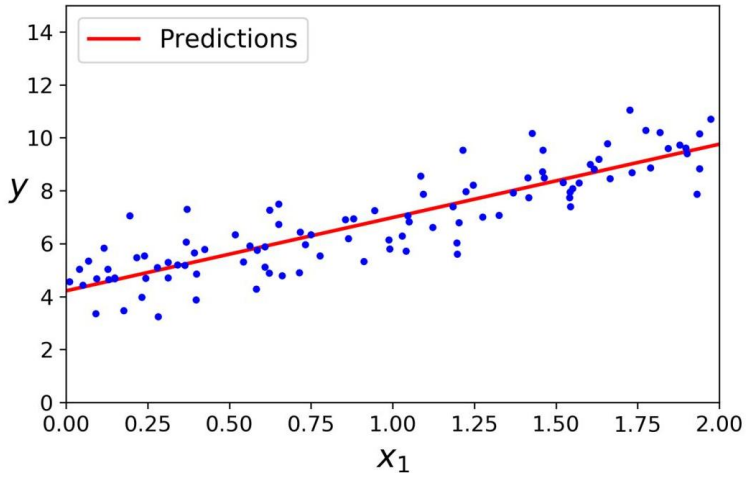
Donde:

- $\hat{y}$ es el valor predicho.
- n es el número de características.
- x_i es el valor de la i^a características.
- θ_i es el j^0 parámetro del modelo (incluyendo el término de sesgo θ_0 y los pesos de características $\theta_1, \theta_2, \ldots, \theta_n$).

El objetivo de la regresión lineal es determinar los parámetros $(\theta_1, \theta_2, \ldots, \theta_n)$ a partir de las observaciones de la variable $\hat{y}$ correspondientes a cada x_i. Se forman rectas que pasen por la nube de puntos formadas por las variables (x_i, y_i), la cual posteriormente se escogerá aquella recta que minimice la suma de los errores, como se visualiza en la Figura 12.

Figura 12

Predicciones del modelo de regresión lineal



Nota. Por Géron (2019).

2.2.2.2. Regresión ridge

La regresión ridge (también conocida como regularización Tikhonov), es una versión regulada de la regresión lineal, la cual incorpora un término de regularización igual a $\alpha \sum_{i=1}^{n} \theta_i^2$ a la función de pérdida. Esto ayuda a que el modelo no solo se ajuste a los datos, sino también a que los "pesos" del modelo sean lo más pequeños posible. Es importante recordar que este ajuste adicional solo se debe hacer durante el entrenamiento del modelo.

A continuación, se presenta la ecuación de la función de pérdida de la regresión de ridge.

$$J(\theta) = ECM(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$$

El hiperparámetro α determina cuánto deseamos regularizar el modelo. Si $\alpha = 0$, entonces la regresión de ridge se convierte simplemente en regresión lineal. Si α es muy alto, todos los pesos se acercan a cero, lo que da como resultado una línea plana que pasa por el promedio de los datos.

Nota que el término de sesgo, θ_0, no está regularizado, por lo que la suma comienza desde i = 1 y no desde 0, Si consideramos **w** como el vector de pesos asociados a las características (desde θ_1 hasta θ_n), el término de regularización se define como $\frac{1}{2}(||\boldsymbol{w}||_2)^2$, donde $||\boldsymbol{w}||_2$ representa la norma l_2 del vector de pesos. Esta norma l_2 es básicamente la raíz cuadrada de la suma de los cuadrados de los valores de los pesos. Cuando usamos el descenso de gradiente para optimizar el modelo, lo único que necesitas hacer es agregar un término extra, $\alpha\boldsymbol{w}$, al vector de gradiente calculado con el error cuadrático medio (ECM) (Géron, 2019).

2.2.2.3. Regresión lasso

La regresión least absolute shrinkage and selection operator (Lasso) es una variante regularizada de la regresión lineal. Al igual que la regresión ridge, lasso agrega un término de regularización a la función de pérdida, pero utiliza la normal l_1 del vector de peso en vez de la mitad del cuadrado de la normal l_2.

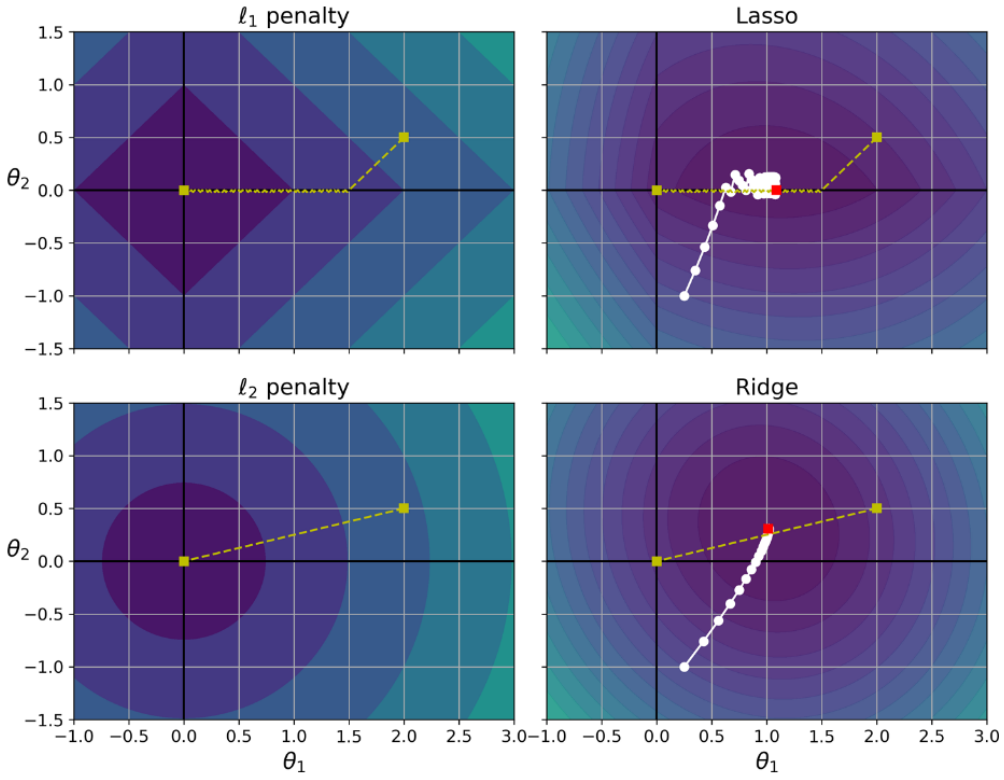
A continuación, se presenta la ecuación de la función de pérdida de la regresión de lasso.

$$J(\theta) = EMC(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

Una de las principales características de la regresión de lasso es que tiende a eliminar los pesos de las características menos importantes (es decir, los pone a cero).

Figura 13

Lasso versus ridge regularization



Nota. Por Géron (2019).

En el gráfico en la parte superior derecha, los contornos representan la función de pérdida de lasso (que es una función de pérdida de ECM más un término de penalización l_1). Los pequeños círculos blancos indican el camino seguido por el descenso de gradiente al optimizar algunos parámetros del modelo, los cuales se inicializan en torno a $\theta_1 = 0{,}25\ y\ \theta_2 = -1$. Observa que la trayectoria llega cerca de $\theta_2 = 0$, luego baja y termina oscilando alrededor del óptimo global (representado por el cuadrado rojo). Si incrementamos el valor de α, el óptimo global se moverá hacia la izquierda a lo largo de la línea amarilla discontinua, mientras que, si reducimos α, el óptimo global se desplazará hacia la derecha. En este caso, los parámetros óptimos para el ECM sin regularización son $\theta_1 = 2\ y\ \theta_2 = 0{,}5$ (Géron, 2019).

2.2.2.4. Árbol de decisión (decision tree)

Los árboles de decisión son algoritmos de machine learning que se pueden hacer para hacer tareas de regresión o clasificación, capaces de ajustar conjunto de datos complejos (Géron, 2019).

Un árbol de decisión es un modelo para poder realizar predicciones utilizada en diferentes ramas de la ciencia, su objetivo principal es el aprendizaje inductivo a partir de construcciones lógicas y observaciones. El árbol es representado por nodos, donde el nodo principal raíz es el atributo a partir del cual se inicia el proceso, los nodos hijos son preguntas acerca del atributo o problema, y los nodos hoja corresponden a una decisión la cual debe coincidir con una de las variables clase del problema a resolver (Charris et al., 2018).

El atributo **gini** de un nodo mide su impureza: un nodo es "puro"(gini=0) si todas las instancias de entrenamiento a las que se aplica pertenecen a la misma clase.

A continuación, se presenta la ecuación de la impureza de gini.

$$G_i = 1 - \sum_{k=1}^{n} p^2{}_{i,k}$$

$p_{i,k}$ es la ratio de instancias de clase k entre las instancias de entrenamiento del i^0 nodo.

El concepto de entropía nació en la termodinámica, donde se usaba para medir qué tan desordenadas estaban las moléculas. Cuando las moléculas están completamente quietas y ordenadas, la entropía es casi cero. Con el tiempo, esta idea se empezó a aplicar en muchos otros campos, como la teoría de la información de Shannon, donde la entropía indica cuánta información promedio hay en un mensaje. Si todos los mensajes son iguales, entonces la entropía también es cero. En el mundo del machine learning, la entropía se usa mucho para medir qué tan "impuro" es un conjunto de datos. Por ejemplo, si todas las instancias pertenecen a una misma clase, la entropía también será cero, porque no hay incertidumbre (Charris et al., 2018).

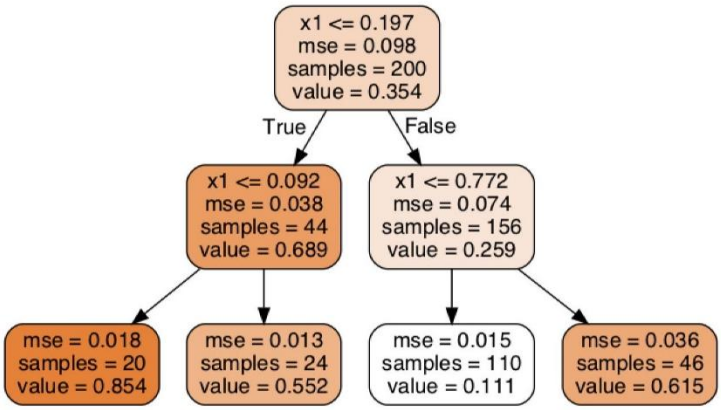
A continuación, se presenta la ecuación de entropía.

$$H_i = -\sum_{k=1}^{n} p_{i,k} log_2(p_{i,k})$$

$$p_{i,k} \neq 0$$

Figura 14

Árbol de decisión para regresión



Nota. Por Géron (2019).

Los árboles de regresión son parecidos a los árboles de clasificación, la diferencia principal es que, en lugar de predecir una clase en cada nodo, este predice un valor.

2.2.2.5. Bosques aleatorios (random forest)

El algoritmo de random forest o bosques aleatorios es un ensamblaje de los modelos de árboles de regresión o clasificación (CART). Los modelos de CART son usados para poder realizar regresiones o clasificaciones, el algoritmo CART selecciona ese umbral de decisión óptimo para una variable basándose en la partición recursiva para llegar a un resultado que se le conoce como árbol de decisión (Rigatti, 2017).

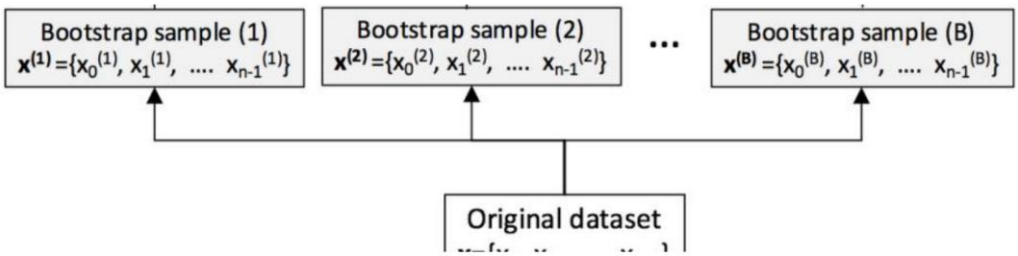
Los modelos CART de un solo árbol no suelen ser adecuados para analizar datos complejos con muchos predictores que interactúan entre sí. Sin embargo, un conjunto de muchos de estos árboles que utilicen diferentes conjuntos de predictores puede, al agregarse, producir resultados de predicción muy buenos (Rigatti, 2017).

Según García (2018), cada árbol de decisión se construirá de la siguiente manera:

- Se comenzará con un conjunto de N observaciones distintas y se seleccionará aleatoriamente una muestra de tamaño N, con reemplazo. Esta técnica, conocida como "bootstrapping", es común en diversos algoritmos de aprendizaje automático. Lo que hace especial a esta técnica es que introduce un grado de aleatoriedad en el proceso, lo que resulta en que cada árbol se construya de manera ligeramente diferente.

Figura 15

Técnica bootstrapping



Nota. Por García (2018)

- Dado un conjunto de M variables de entrada, en cada nodo del árbol se elegirá de manera aleatoria un número p de estas variables, con la condición de que p<<M. Este número p será el mismo en todo el proceso de construcción del árbol y aporta una segunda capa de aleatoriedad al algoritmo.
- El árbol continuará creciendo sin limitaciones, es decir, se permitirá que crezca hasta alcanzar su máxima extensión posible.

2.2.2.6. Redes neuronales artificiales

Las redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico (Matich, 2001).

Perceptrón: es una de las arquitecturas de redes neuronales más simples, desarrollada en 1957 por Frank Rosenblatt. Se basa en una neurona artificial con una

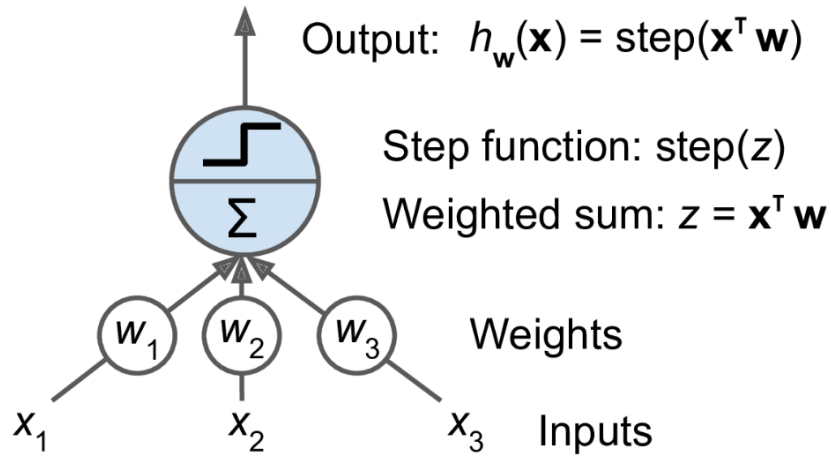
estructura algo diferente llamada threshold logic unit (TLU), o también llamada como linear threshold unit (LTU). Las entradas y salidas son números, en vez de valores binarios de activación/desactivación, y cada conexión de entrada tiene asignado un peso. TLU realiza un cálculo de una suma ponderada de las entradas ($z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \boldsymbol{x}^T \boldsymbol{w}$) y, a continuación, aplica una función escalonada a esa suma para generar el resultado final. Esta función es: $h_w(x) = step(z)$, donde $z = \boldsymbol{x}^T \boldsymbol{w}$ (Géron, 2019).

A continuación, se presentan las funciones escalonadas heaviside y signo, las cuales se utilizan en los perceptrones.

$$Heaviside(z) = \begin{cases} 0 \; si \; z < 0 \\ 1 \; si \; z \geq 0 \end{cases} \quad Signo(z) = \begin{cases} -1 \; si \; z < 0 \\ 0 \; si \; z = 0 \\ +1 \; si \; z > 0 \end{cases}$$

Figura 16

Una neurona artificial que calcula una suma ponderada de sus entradas y luego aplica una función escalonada



Nota. Por Géron (2019).

Un perceptrón es básicamente una red neuronal muy simple, compuesta por una sola capa de unidades llamadas TLU. Cada una de estas unidades está conectada a todas las entradas. Cuando cada neurona de una capa está conectada a todas las neuronas de la capa anterior (es decir, las neuronas que reciben las entradas), se dice que es una capa totalmente conectada o densa. Las señales que recibe el perceptrón pasan primero por unas neuronas especiales llamadas neuronas de entrada, que simplemente transmiten los

datos tal como llegan. Todas estas neuronas forman lo que se conoce como la capa de entrada (Géron, 2019).

A continuación, se presentan la ecuación para poder calcular las salidas de una capa completamente conectada.

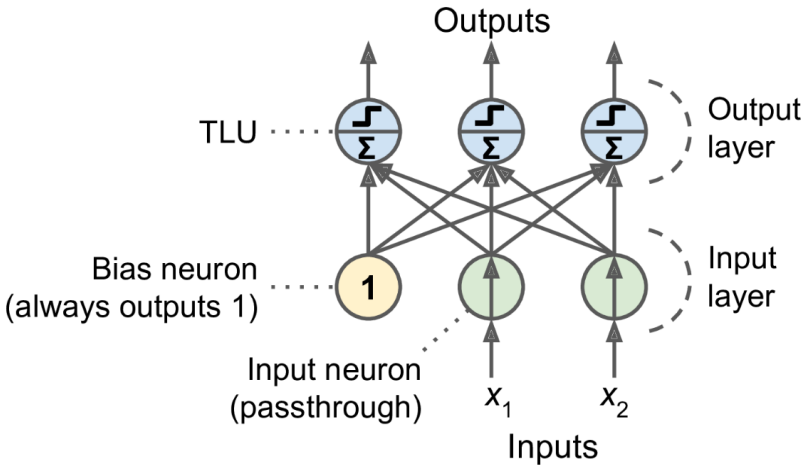
$$h_{W,b}(X) = \emptyset(XW + b)$$

Donde:

- El símbolo $\boldsymbol{X}$, representa la matriz de características de entrada.
- El símbolo $\boldsymbol{W}$, contiene todos los pesos de conexión excepto los de la neurona de sesgo.
- El símbolo $\boldsymbol{b}$, contiene todos los pesos de conexión entre la neurona de sesgo y las neuronas artificiales.
- El símbolo $\emptyset$, es una función de activación: cuando las neuronales artificiales son TLU, es una función escalonada.

Figura 17

Arquitectura de un perceptrón con dos neuronas de entrada, una neurona de sesgo y tres neuronas de salida



Nota. Por Géron (2019).

El perceptrón ajusta sus conexiones para minimizar los errores. Específicamente, analiza cada ejemplo del entrenamiento uno por uno y hace una predicción para cada

caso. Si alguna de sus neuronas de salida se equivoca, el modelo refuerza las conexiones relacionadas con las entradas que habrían llevado a una predicción correcta (Géron, 2019).

A continuación, se presentan la ecuación de la regla de aprendizaje del perceptrón (ajuste de peso).

$$W_{i,j}{}^{(siguiente\ paso)} = W_{i,j} + n(y_j - \hat{y}_j)x_i$$

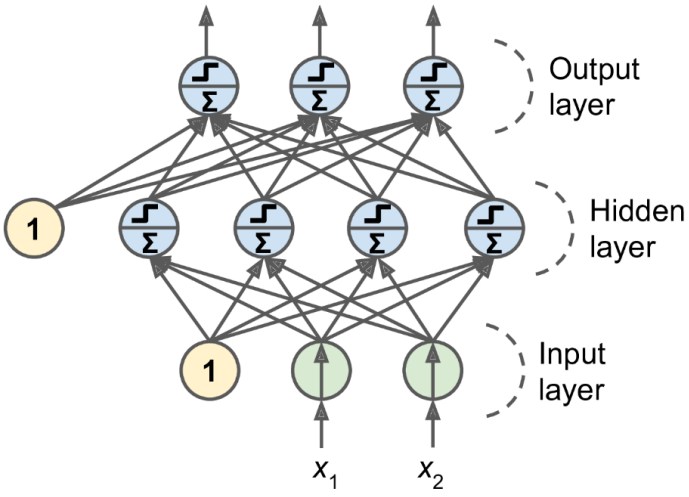
Donde:

- $W_{i,j}$ es el peso de conexión entre la i^a neurona de entrada y la j^a neurona de salida.
- x_i es el i^0 valor de entrada de la instancia de entrenamiento actual.
- $\hat{y}_j$ es la salida de la j^a neurona de salida para la instancia de entrenamiento actual.
- y_j es la salida objetivo de la j^a neurona de salida para la instancia de entrenamiento actual.
- n es la tasa de aprendizaje.

Perceptrón multicapa: Un MLPs está compuesto por una capa de entrada, una o más capas TLU conocidas como "capas ocultas", y una capa TLU final que actúa como capa de salida. Las capas que se encuentran cerca de la entrada se denominan "capas inferiores", mientras que las que están más cerca de la salida se llaman "capas superiores". Todas las capas, excepto la de salida, tienen una neurona de sesgo y están completamente conectadas con la capa siguiente (Géron, 2019).

Figura 18

Arquitectura de un perceptrón multicapa con dos entradas, una capa oculta de cuatro neuronas y tres neuronas de salida



Nota. Por Géron (2019).

MLPs de regresión: Los MLPs se utilizan principalmente en tareas de regresión. Si se desea predecir un único valor, se requiere una sola neurona de salida, la cual proporciona el valor estimado. En el caso de una regresión multivariable, se necesita una neurona de salida para cada una de las dimensiones que se desean predecir (Géron, 2019).

Tabla 1

Arquitectura típica de un MLPs de regresión

Hiperparámetro	Valor típico
# neuronas de entrada	Una por característica de salida (por ejemplo, 28 x 28 = 784 para MNIST)
# capas ocultas	Depende del problema, pero normalmente entre 1 y 5
# neuronas por capa oculta	Depende del problema, pero normalmente entre 10 y 100
# neuronas de salida	1 por dimensión de predicción
Activación de oculta	ReLU
Activación de salida	Ninguna, o reLU/softplus (para salidas positivas) o logística/tanh (para salidas limitadas)
Función de pérdida	ECM o EAM/Huber (si hay valores atípicos)

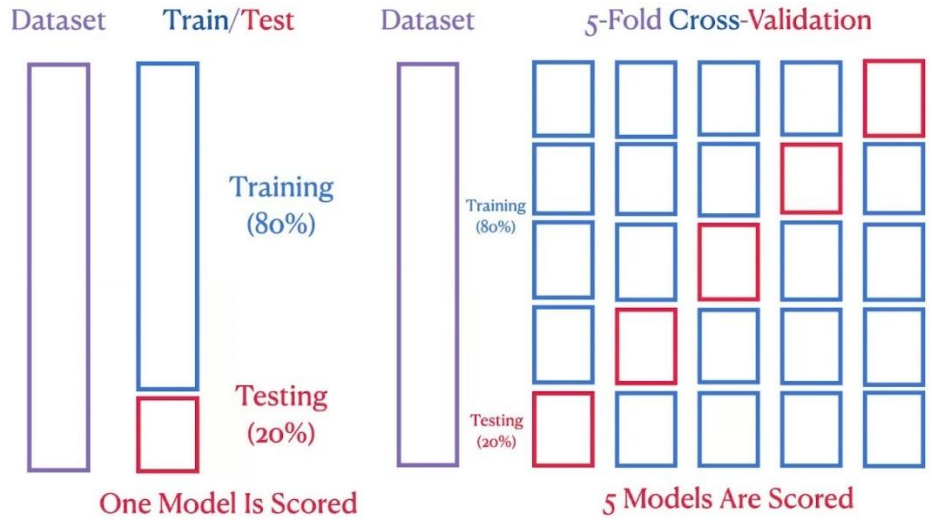
Nota. Por Géron (2019).

2.2.3. K-fold cross-validation

La validación cruzada k-fold es una técnica robusta que se utiliza para evaluar el rendimiento de los modelos de aprendizaje automático. Básicamente, divide los datos en varias partes y prueba el modelo en diferentes combinaciones de entrenamiento y prueba. Así se asegura de que el modelo no solo funcione bien con los datos que ya conoce, sino que también pueda adaptarse a datos nuevos que no ha visto antes (Géron, 2019).

Figura 19

Ilustración de la división entrenamiento/prueba



Nota. Por Vinod (2024).

One model is scored: es el método clásico de división entre entrenamiento y prueba, donde el conjunto de datos se separa en dos partes: el 80 % se usa para entrenar el modelo y el 20 % restante se reserva para evaluarlo (Vinod, 2024).

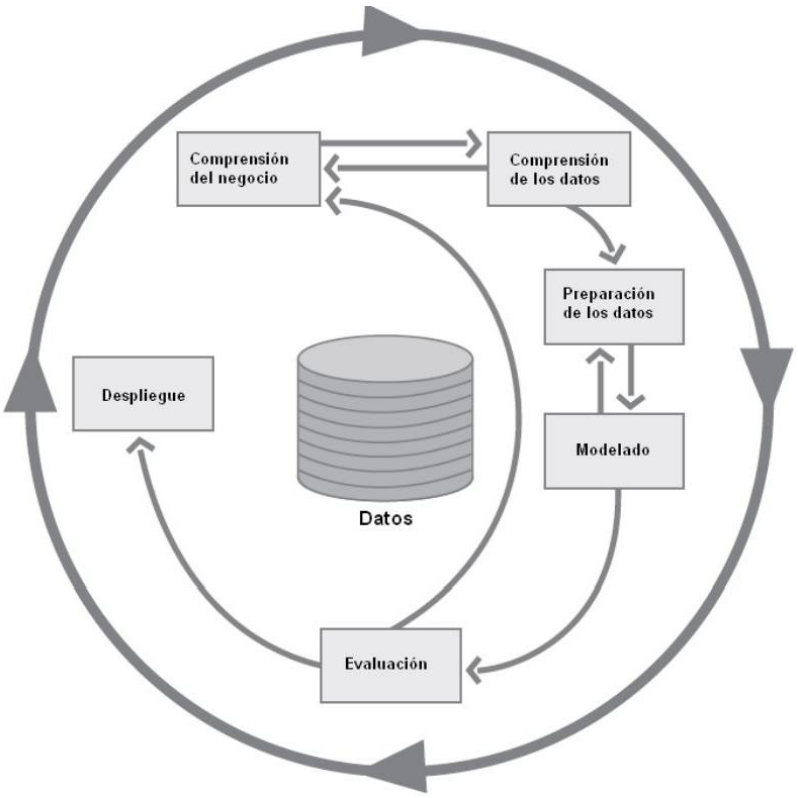
5 model are scored: validación cruzada quíntuple, donde los datos se dividen en cinco segmentos. En cada una de las cinco iteraciones, uno de estos segmentos se usa como conjunto de prueba, mientras que los otros cuatro se emplean para entrenar, garantizando que cada parte se utilice una vez para probar y el resto para entrenar (Vinod, 2024).

2.2.4. CRISP-DM

El modelo CRISP-DM ofrece una visión general del ciclo de vida de un proyecto de minería de datos. El autor Chapman et al. (1999) menciona que este modelo tiene 6 fases como se muestra en la Figura 20, y cada fase tiene sus respectivas tareas y las relaciones entre estas tareas.

Figura 20

Modelo CRISP-DM



Nota. Por Chapman et al. (1999).

2.2.4.1. Comprensión del negocio (business understanding)

Según Chapman et al. (1999) menciona que la fase comprensión del negocio se centra en "comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, para luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para alcanzar los objetivos" (p.10).

Tareas de la fase comprensión del negocio: determinar los objetivos del negocio, evaluar la situación, determinar los objetivos del proceso de minería de datos y producir un plan de proyecto.

2.2.4.2. Comprensión de los datos (data understanding)

Según Chapman et al. (1999) menciona que la fase comprensión de los datos se realiza una "recolección inicial de los mismos y continúa con actividades que te permiten familiarizarte con los datos, identificar problemas de calidad de los datos, descubrir los primeros conocimientos sobre ellos y/o detectar subconjuntos interesantes para formular hipótesis sobre información oculta" (p.10).

Tareas de la fase comprensión de los datos: recopilar datos iniciales, describir los datos, explorar los datos y verificar la calidad de los datos.

2.2.4.3. Preparación de los datos (data preparation)

Según Chapman et al. (1999) menciona que la fase preparación de los datos "se abarca todas las actividades necesarias para construir el conjunto de datos final [datos que se introducirán en la(s) herramienta(s) de modelado] a partir de los datos brutos iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para las herramientas de modelado" (p.11).

Tareas de la fase preparación de los datos: conjunto de datos, seleccionar los datos, limpiar los datos, construir los datos, integrar los datos y formatear los datos.

2.2.4.4. Modelado (modeling)

Según Chapman et al. (1999) menciona que la fase de modelado "se seleccionan y aplican diversas técnicas de modelado, y sus parámetros se calibran a valores óptimos. Por lo general, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos" (p.11).

Tareas de la fase modelado: seleccionar técnica de modelado, generar diseño de prueba, construir el modelo y evaluar el modelo.

2.2.4.5. Evaluación (evaluation)

Según Chapman et al. (1999) menciona que antes de proceder al despliegue final del modelo "es importante evaluarlo minuciosamente y revisar los pasos ejecutados para crearlo, para asegurarse de que el modelo logra adecuadamente los objetivos comerciales. Un objetivo clave es determinar si hay algún problema comercial importante que no se haya considerado suficientemente" (p.11).

Tareas de la fase evaluación: evaluar los resultados, revisar el proceso y determinar los siguientes pasos.

2.2.4.6. Despliegue (deployment)

Según Chapman et al. (1999) menciona que la fase despliegue puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos a toda la organización, va a depender mucho de los requisitos, "en muchos casos, es el cliente, no el analista de datos, quien lleva a cabo los pasos de despliegue. Sin embargo, incluso si el analista lleva a cabo el esfuerzo de despliegue, es importante que el cliente comprenda de antemano qué acciones deben realizarse para poder usar realmente los modelos creados" (p.11).

Tareas de la fase despliegue: planificar el despliegue, planificar el monitoreo y mantenimiento, producir el informe final y revisar el proyecto.

2.2.5. Examen de admisión

El examen de admisión es una prueba que busca definir si una persona puede acceder o no a la universidad, no es un método universal. En muchos países, se aplica una evaluación a nivel nacional, y las universidades toman en cuenta los resultados de estas pruebas como parte del proceso para elegir a sus futuros estudiantes (Ocaña, 2014).

En el Perú, el examen de admisión sigue siendo una herramienta clave para definir quién accede a la universidad, aunque no hay un único modelo que se aplique en todas las instituciones. Algunas universidades optan por una prueba general para todos los postulantes, mientras que otras, como San Marcos, aplican exámenes específicos según

la carrera elegida. También hay casos, como en la UPC, donde el proceso de admisión incluye otras modalidades, como entrevistas personales o la presentación de cartas de recomendación de docentes. Incluso, en universidades como la de Lima, pertenecer al tercio superior en la etapa escolar puede asegurar el ingreso directo, sin necesidad de pasar por un examen (Ocaña, 2014).

2.2.6. Rendimiento académico universitario

El rendimiento académico en estudiantes universitarios según Garbanzo (2012) es la "suma de diferentes y complejos factores que actúan en la persona que aprende, y ha sido definido con un valor atribuido al logro del estudiante en las tareas académicas. Se mide mediante las calificaciones obtenidas, con una valoración cuantitativa, cuyos resultados muestran las materias ganadas o perdidas, la deserción y el grado de éxito académico".

El rendimiento académico se entiende como la capacidad del estudiante para alcanzar los objetivos, metas o logros que se plantean en una asignatura o programa de estudios. Diversos autores coinciden en que este rendimiento no depende de un solo factor, sino que es el resultado de una compleja combinación de aspectos. Entre ellos destacan el contexto familiar y social del estudiante, la calidad de la relación que establece con sus docentes y compañeros, así como los métodos de enseñanza, el ambiente institucional y las políticas educativas que lo rodean. En otras palabras, aprender y rendir bien académicamente va más allá del esfuerzo individual: también influyen profundamente el entorno y las condiciones en las que se desarrolla el aprendizaje (D. et al., 2006)

2.3. Definición de términos

2.3.1. Error absoluto medio (MAE)

Según Chang (2023), menciona que la función MAE nos permite identificar el pronóstico y los resultados posibles, su fórmula es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Donde:

- y_i : valor real de la i-ésima observación.
- $\hat{y}_i$: valor predicho de la i-ésima observación.
- n : número total de observaciones.

2.3.2. Error cuadrático medio (MSE)

Según (Chang Hidalgo, 2023), menciona que esta métrica calcula el valor promedio de los errores elevados al cuadrado, su fórmula es:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde:

- y_i : valor real de la i-ésima observación
- $\hat{y}_i$: valor predicho de la i-ésima observación
- n : número total de observaciones

2.3.3. Raíz de error cuadrado medio (RMSE)

Según Chang (2023), menciona que la función RMSE calcula la raíz de la métrica MSE (conocida como desviación media cuadrática - RMSD), su fórmula es:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Donde:

- y_i : valor real de la i-ésima observación
- $\hat{y}_i$: valor predicho de la i-ésima observación

- n : número total de observaciones

2.3.4. Error porcentual absoluto medio (MAPE)

Según Chang (2023), menciona esta métrica permite calcular la dimensión del error de tipo absoluto expresado en porcentajes, siendo su fórmula:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| x\ 100$$

Donde:

- y_i : valor real de la i-ésima observación
- $\hat{y}_i$: valor predicho de la i-ésima observación
- n : número total de observaciones

2.3.5. Hiperparámetro

Según Luo (2016), los hiperparámetros son configuraciones que se ajustan antes de entrenar un modelo y desempeñan un papel clave en la dirección del proceso de aprendizaje. Para lograr que el modelo sea más preciso, existen métodos automáticos que ayudan a encontrar los valores óptimos de estos hiperparámetros, como la búsqueda exhaustiva, la búsqueda aleatoria, los algoritmos genéticos, entre otros.

2.3.6. Grid Search

Grid search cross-validation (GSCV) es una técnica que se utiliza para encontrar la mejor combinación de hiperparámetros que optimicen el rendimiento de un modelo. Este proceso implica entrenar múltiples versiones del modelo, cada una con diferentes configuraciones de hiperparámetros, y evaluar su desempeño mediante validación cruzada para identificar la combinación más efectiva.

2.3.7. Overfitting

Según García (2018) es un concepto comúnmente utilizado en estadísticas y aprendizaje automático. Este problema surge cuando un algoritmo logra hacer buenas

predicciones con los datos de entrenamiento, pero tiene dificultades para generalizar y pierde precisión cuando se enfrenta a nuevos datos diferentes a los que usó inicialmente.

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Tipo y diseño de la investigación

Behar (2008) menciona que el tipo de investigación aplicada: "es el estudio y aplicación de la investigación a problemas concretos, en circunstancias y características concretas. Esta forma de investigación se dirige a su aplicación inmediata y no al desarrollo de teorías" (p.20). La presente investigación busca resolver el problema de predicción del rendimiento académico universitario de los estudiantes de la facultad de Ingeniería, por lo tanto, se caracteriza como **aplicada**.

Según Hernández et al. (2014) indica que los experimentos "manipulan tratamientos, estímulos, influencias o intervenciones (denominadas variables independientes) para observar sus efectos sobre otras variables (las dependientes) en una situación de control" (p.129), y se caracterizan por la:

- Manipulación intencional de variables (independientes).
- Medición de variables (dependientes).
- Control y validez.
- Dos o más grupos de comparación.
- Participantes asignados al azar o emparejados.

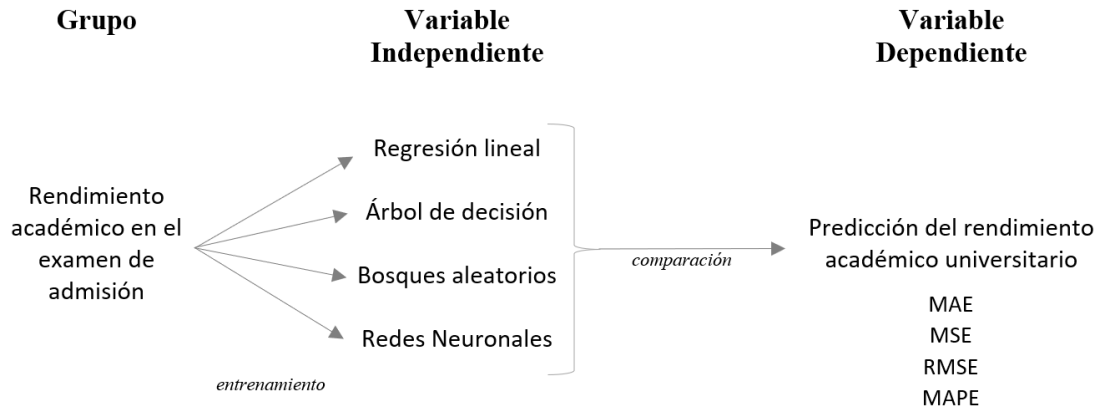
Según Hernández et al. (2014) indica que los diseños de investigación experimentales puros utilizan "prepruebas y pospruebas para analizar la evolución de los grupos antes y después del tratamiento experimental. Desde luego, no todos los diseños experimentales puros utilizan preprueba; aunque la posprueba sí es necesaria para determinar los efectos de las condiciones experimentales" (p.141).

De acuerdo a las definiciones anteriores, esta investigación se ubica como diseño de investigación de tipo **experimental** y de **experimentos puros**, pues se manipula de manera intencional la variable independiente que son los algoritmos de machine learning para examinar los efectos que la manipulación dispone en el rendimiento académico universitario.

Seguidamente, se presenta el diseño experimental puro en la Figura 21.

Figura 21

Diseño experimental



Nota. Elaboración propia.

Finalmente, la siguiente investigación presenta un enfoque cuantitativo, porque según Hernández et al. (2014) se utilizó: "la recolección de datos para probar hipótesis con base en la medición numérica y el análisis estadístico, con el fin establecer pautas de comportamiento y probar teorías" (p.4).

3.2. Población y muestra de estudio

3.2.1. Población

La población es la totalidad de un fenómeno de estudio, incluye la totalidad de entidades o unidades de análisis de población que integran dicho fenómeno la cual debe cuantificarse para un determinado estudio (Tamayo, 2014).

Para la siguiente investigación se tomó como población a los ingresantes en las modalidades de FASE-I, FASE-II, CEPU OTOÑO (CEPU-I), CEPU INVIERNO (CEPU-II) y CEPU VERANO (CEPU-III) de la Facultad de Ingeniería, en las escuelas profesionales de: Ingeniería de Minas (ESMI), Ingeniería Metalúrgica (ESME), Ingeniería Mecánica (ESMC), Ingeniería en Informática y Sistemas (ESIS) e Ingeniería Química (ESIQ) de la Universidad Nacional Jorge Basadre Grohmann en el año 2023.

3.2.2. Muestra

Luego de definirse la población de investigación, se debe determinar la muestra, según Tamayo (2014) menciona que "cuando no es posible medir cada una de las entidades de la población; esta muestra, se considera, es representativa de la población" (p.176).

Para el siguiente estudio, la muestra es de 311 alumnos ingresantes a la Universidad Nacional Jorge Basadre Grohmann, en la Facultad de Ingeniería por las diferentes modalidades, y estará compuesta por la totalidad de la población, la cual se detalla en la Tabla 3. En este caso, la muestra es de tipo censal, ya que se considerará al 100 % de los integrantes de la población de estudio.

Tabla 2

Ingresantes en el año 2023

Escuela	FaseI	FaseII	CepuI	CepuII	CepuIII	Total
ESMI	14	23	5	5	4	51
ESME	13	35	7	7	5	68
ESMC	14	32	7	7	6	50
ESIS	11	15	10	10	10	87
ESIQ	11	24	6	3	6	55

Nota. Elaboración propia.

3.3. Acciones y actividades para la ejecución del proyecto

Luego de haber realizado la operacionalización de las variables (independiente y dependiente) y definido los instrumentos de medición, se realizarán acciones y actividades para poder cumplir los objetivos planteados.

Para esta investigación se utilizó la validación de juicio de experto para poder validar el instrumento de medición.

3.4. Materiales e instrumentos

Arias (2012) menciona que la observación es "una técnica que consiste en visualizar o captar mediante la vista, en forma sistemática, cualquier hecho, fenómeno o situación que se produzca en la naturaleza o en la sociedad, en función de unos objetivos de investigación preestablecidos" (p.69).

Para la siguiente investigación se utilizó la observación como técnica, la cual fue de tipo observación estructurada, ya que según Arias (2012) indica que este tipo de observación se realiza "en correspondencia con unos objetivos, utiliza una guía diseñada previamente, en la que se especifican los elementos que serán observados" (p.70).

La siguiente ficha se adjunta en el ANEXO 1.

3.5. Tratamiento de datos

3.5.1. Procedimiento de recolección de datos

Se recolectó información detallada sobre el proceso de admisión de los estudiantes que ingresaron en el año 2023. Este paso fue importante para contar con una base clara de datos.

Los datos relacionados con el rendimiento académico fueron recopilados a través de la oficina de Dirección Académica de Actividades y Servicios Académicos (DASA) de la UNJBG de los periodos (2023-I, 2023-II, 2024-I, 2024-II). Esta información corresponde específicamente a los estudiantes ingresantes a la Facultad de Ingeniería.

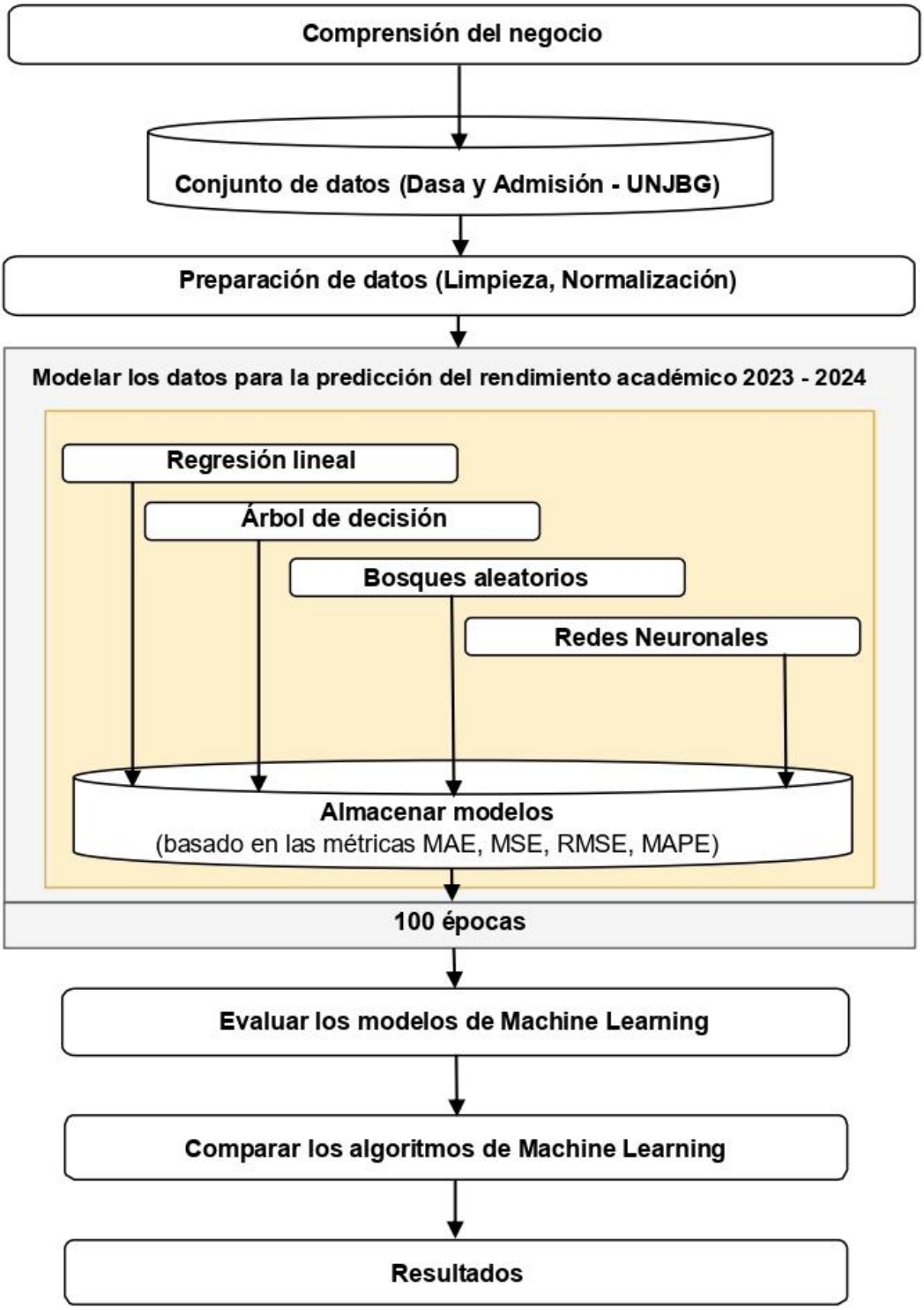
Los datos fueron recolectados en formato excel, donde luego se tuvo que hacer una limpieza de la información y ser procesada en python, para poder predecir el rendimiento académico universitario en base del rendimiento del examen de admisión.

3.5.2. Análisis y procesamiento de datos

A continuación, se presenta el flujo de trabajo para poder predecir el rendimiento académico de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023:

Figura 22

Flujo de trabajo experimental



Nota. Elaboración propia.

CAPÍTULO IV

RESULTADOS DE LA INVESTIGACIÓN

4.1. Presentación y análisis de los resultados

4.1.1. Preparar los datos para construir modelos de machine learning

Los datos relacionados con el rendimiento académico fueron recopilados a través de la oficina de Dirección Académica de Actividades y Servicios Académicos (DASA) de la UNJBG de los periodos (2023-I, 2023-II, 2024-I, 2024-II). Esta información corresponde específicamente a los estudiantes ingresantes a la Facultad de Ingeniería.

El propósito de esta recopilación es analizar el desempeño académico inicial de los alumnos, permitiendo comprender mejor las características de su formación durante los primeros ciclos de estudio. A continuación, se detalla:

Tabla 3

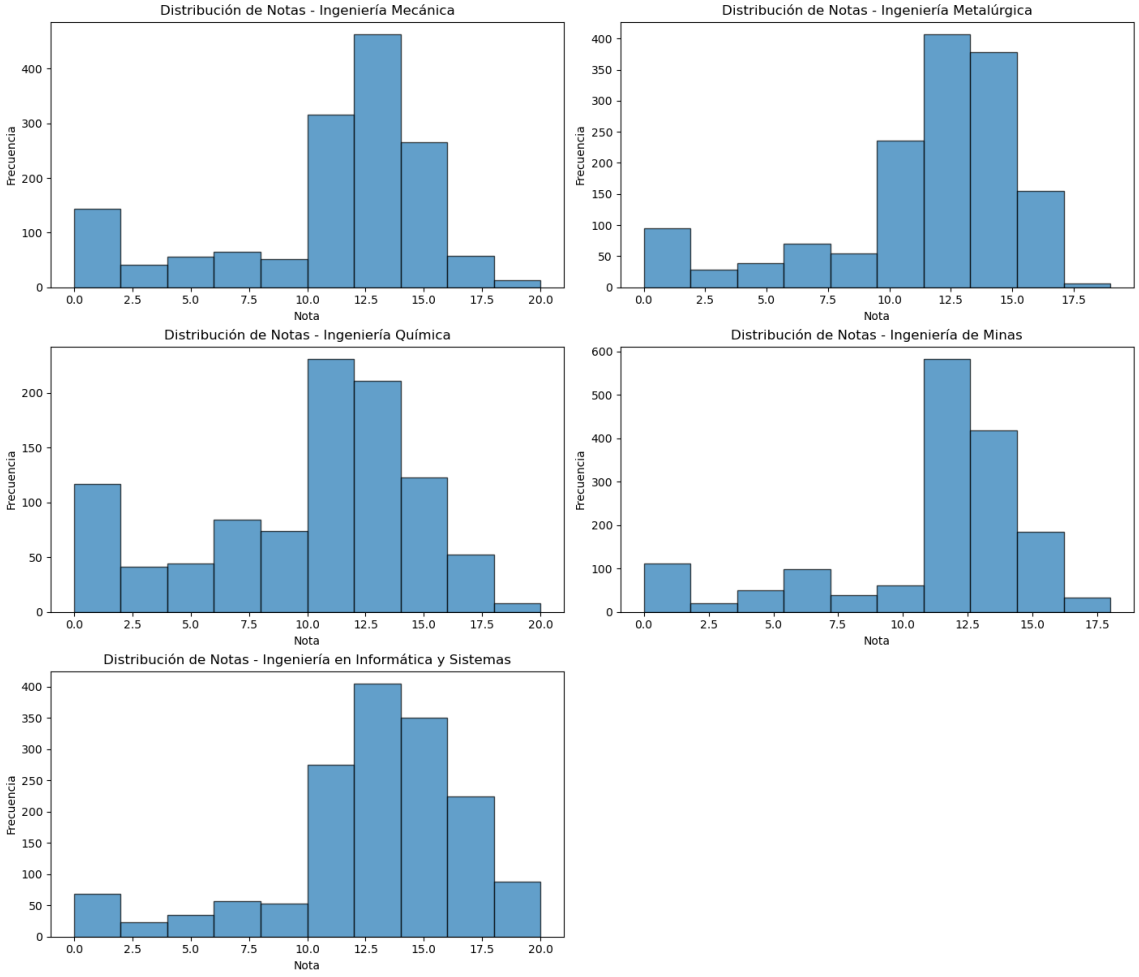
Distribución de promedios por especialidad

Carrera	Mean	Std	Min	25 %	50 %	75 %	Max
Ingeniería Mecánica	9,0	4,2	0,0	7,3	10,1	12,4	14,3
Ingeniería Metalúrgica	10,6	4,4	0,1	10,0	11,6	13,4	14,8
Ingeniería Química	8,0	4,0	0,2	7,0	8,3	10,7	15,0
Ingeniería de Minas	11,5	2,7	0,0	10,4	11,8	13,3	15,0
Ingeniería en Informática y Sistemas	11,4	3,4	0,1	10,5	12,4	13,5	16,8

Nota. Elaboración propia.

Figura 23

Frecuencia de promedios por especialidad



Nota. Elaboración propia.

La gráfica presentada muestra la distribución de las notas obtenidas por los estudiantes ingresantes en las diferentes especialidades de la Facultad de Ingeniería. Cada gráfico representa una especialidad, permitiendo identificar patrones de rendimiento académico según las calificaciones obtenidas. En algunas especialidades, la mayoría de los alumnos tienen calificaciones similares, mientras que en otras hay más variedad. Esto podría indicar que los estudiantes llegan con distintos niveles de preparación según la carrera que eligen.

Después de analizar la distribución de notas y frecuencias, se identificó que algunos estudiantes presentaban una carga académica atípica. Es decir, algunos adelantaron cursos de ciclos superiores, mientras que otros llevaron menos asignaturas de

las que correspondían a su plan de estudios en el ciclo en que se matricularon. Esta situación generaba una diferencia en la evaluación del rendimiento académico, ya que los estudiantes regulares, que cursaban todas las materias de su ciclo, enfrentaban una carga completa, mientras que aquellos con menos asignaturas podían obtener mejores promedios sin la misma exigencia.

Para corregir esta desigualdad y garantizar una evaluación más justa, se decidió aplicar un ajuste en las calificaciones. Primero, se eliminaron del análisis los cursos adelantados que no pertenecían al ciclo correspondiente. Luego, para los estudiantes que no llevaron algunas materias de su plan de estudios en el ciclo respectivo, se asignó una calificación de 0 en esos cursos.

Tabla 4

Ejemplo de matrícula del III ciclo

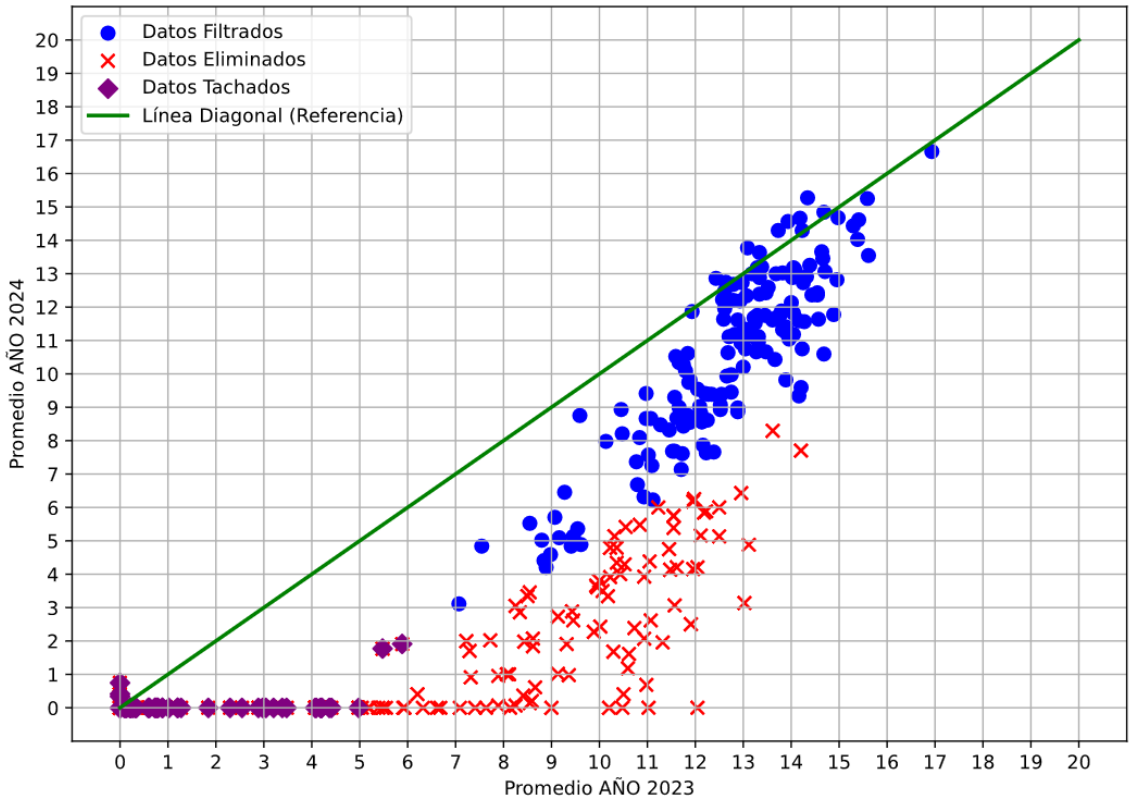
Código	Materia	Sección	Nota
19,03415	Teoría general de sistemas	A	12
19,03416	Sistemas eléctricos y electrónicos	A	18
19,03419	Matemática II	A	9
19,03420	Estadística y probabilidades	B	14
19,05430	Base de datos I	B	0
19,05431	Investigación operativa I	B	0

Nota. Elaboración propia.

Para este estudiante debió llevar los siguientes cursos: 19,03415, 19,03416, 19,03419, 19,03420, 19,03417, 19,03418; sin embargo, los cursos 19,05430, 19,05431 no son de su ciclo son del V ciclo por lo tanto está adelantando cursos, por lo que se considera una nota de 0 en dichos cursos, asegurando una evaluación más equitativa de su rendimiento académico.

Figura 24

Comparación de promedios ponderados: 2023 vs 2024



Nota. Elaboración propia.

La gráfica muestra que la mayoría de los estudiantes mantienen una relación consistente entre ambos promedios, pero algunos presentan casos extremos, como promedios ponderados de nota cero en el año 2024 (abandono) o malas notas (ej. pasar de 12 a 3), posiblemente por problemas personales o académicos. Estos casos atípicos se excluyen del análisis para esta investigación, con la idea que no existe mucha diferencia entre el promedio de cada semestre, donde el criterio es exclusión es el valor absoluto de la diferencia del promedio2023 - promedio2024 sea menor que 5 (en el grafico representa la parte roja eliminados) y también los promedios del promedio2023 y promedio2024 sean menores que 3 (en el grafico representa la parte morada los datos tachados), la parte azul del grafico son los alumnos que cumplen los criterios para esta investigación.

Respecto al examen de **FASE I, FASE II**, los cursos para ambos exámenes son los siguientes: razonamiento verbal (14), razonamiento matemático (14), realidad

nacional (8), aritmética y álgebra (6), geometría y trigonometría (4), lógica (4), lenguaje (2), física (4), química (4).

De la misma manera, para el examen de **CEPU I, CEPU II**, los cursos para ambos exámenes son los siguientes: razonamiento verbal (7), razonamiento matemático (7), aritmética y álgebra (8), geometría y trigonometría (7), lógica (2), lenguaje (2), física (7), química (5).

Finalmente, para al examen de **CEPU III**, se tuvo 2 exámenes, el primer examen estuvo conformada por los siguientes cursos: razonamiento verbal (14), aritmética y álgebra (18), física (14), química (14); el segundo examen estuvo conformada por los siguientes cursos: razonamiento matemático (14), lógica (14), geometría y trigonometría (20), lenguaje (12).

Después de haber hecho un análisis, se ha determinado eliminar la columna del curso de realidad nacional, ya que esta materia solo está presente en el examen de FASE I -FASE II y no en CEPU I, CEPU II Y CEPU III.

Además, se ha decidido excluir de este análisis a los estudiantes que ingresaron mediante el examen extraordinario. Esto se debe a que no se logró obtener acceso a sus calificaciones de admisión, lo que impide una comparación equitativa con el resto de los ingresantes. La ausencia de estos datos podría generar sesgos en el estudio, afectando la interpretación de los resultados. Por ello, para garantizar un análisis más preciso y representativo, se ha optado por no incluir a este grupo en la evaluación.

Como resultado, luego de haber aplicado los criterios de exclusión, obtuvimos un conjunto final de datos completamente limpia para poder ya modelar los algoritmos de machine learning, compuesto por 143 estudiantes que ingresaron a la Facultad de Ingeniería bajo las modalidades de FASE I, FASE II, CEPU I, CEPU II y CEPU III.

Tabla 5

Ingresantes considerados para el presente estudio

Escuela	FaseI	FaseII	CepuI	CepuII	CepuIII	Total
ESMI	7	11	4	3	1	26
ESME	8	20	4	5	3	40
ESMC	4	9	5	2	2	22
ESIS	8	12	6	9	9	44
ESIQ	1	7	0	2	1	11

Nota. Elaboración propia.

La tabla 5 refleja las cifras de los estudiantes que fueron considerados para el estudio después de aplicar los filtros de exclusión.

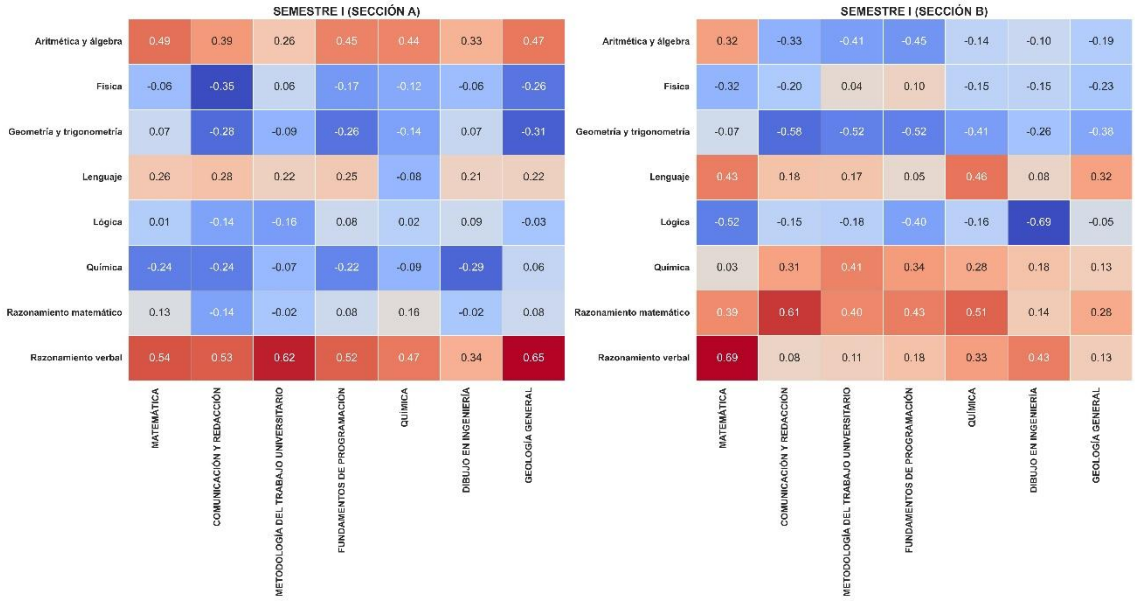
Seguidamente, se aplicó una matriz de correlación con el objetivo de ver cómo se relacionan los cursos del examen de admisión, como razonamiento verbal, razonamiento matemático, realidad nacional, aritmética, álgebra, geometría, trigonometría, lógica, lenguaje, física y química, con los cursos que los estudiantes de Ingeniería tomarán en la universidad. Esto nos ayuda a entender si un buen desempeño en los exámenes de admisión está relacionado con un buen rendimiento en asignaturas importantes de la carrera.

Se optado por abreviar los cursos del examen de admisión para tener una mejor visualización de los datos de la siguiente manera: aritmética y álgebra (AA), física (FIS), geometría y trigonometría (GT), lenguaje (LEN), lógica (LOG), química (QUI), razonamiento matemático (RM) y razonamiento verbal (RV). Asimismo, se han asignado los códigos correspondientes a cada curso, los cuales están detallados en el plan curricular, que se encuentra en los anexos.

Matriz de correlación de Ingeniería de Minas

Figura 25

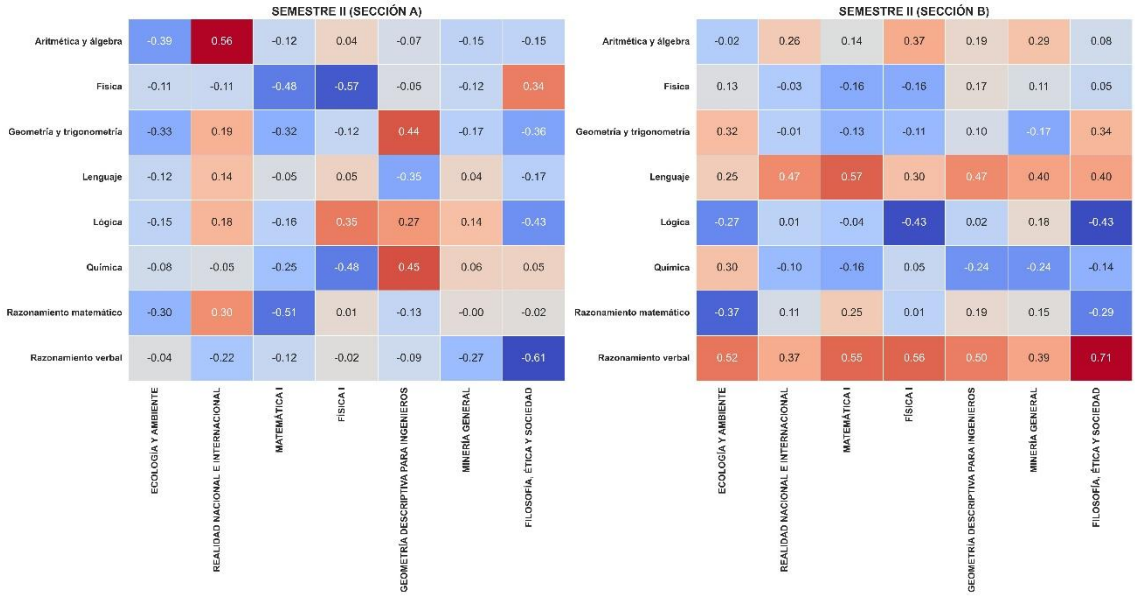
Matriz de correlación por cursos y secciones del I SEMESTRE - ESMI



Nota. Elaboración propia.

Figura 26

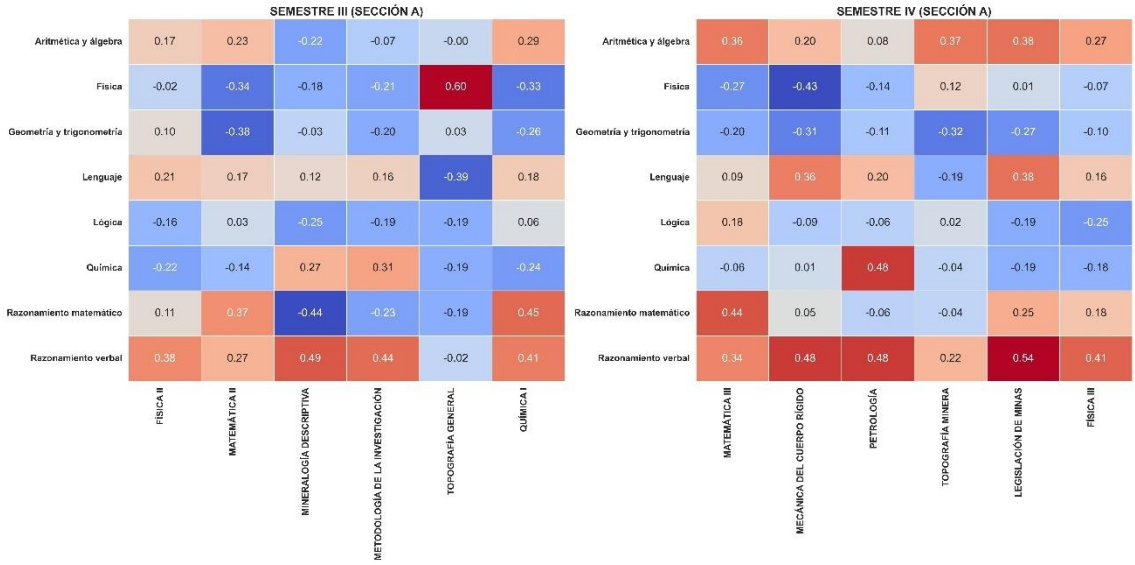
Matriz de correlación por cursos y secciones del II SEMESTRE - ESMI



Nota. Elaboración propia.

Figura 27

Matriz de correlación por cursos y secciones del III – IV SEMESTRE - ESMI

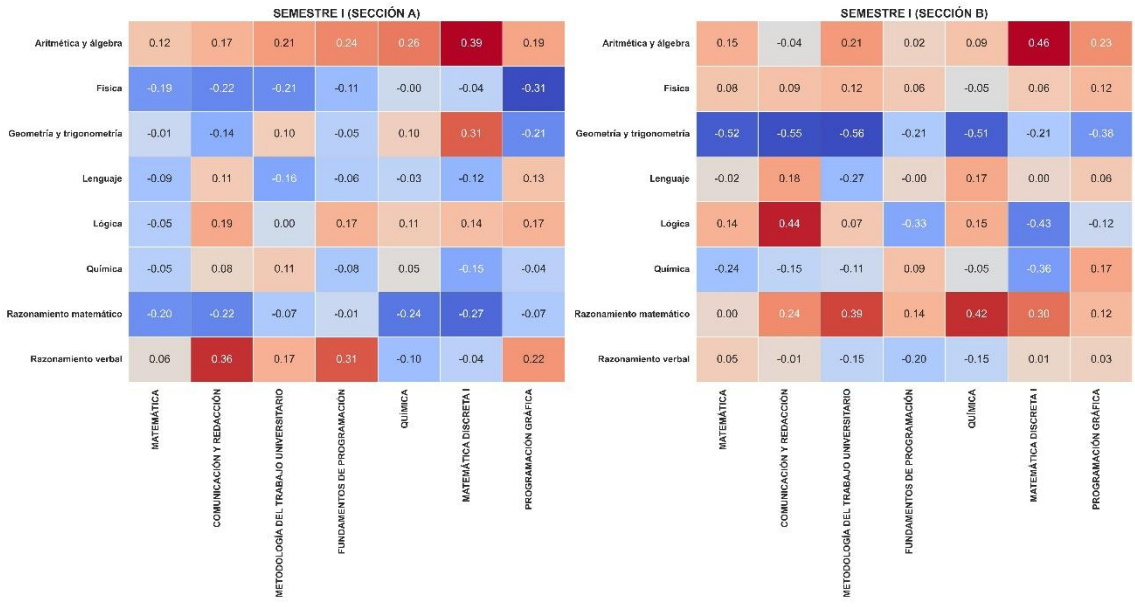


Nota. Elaboración propia.

Matriz de correlación de Ingeniería en Informática y Sistemas

Figura 28

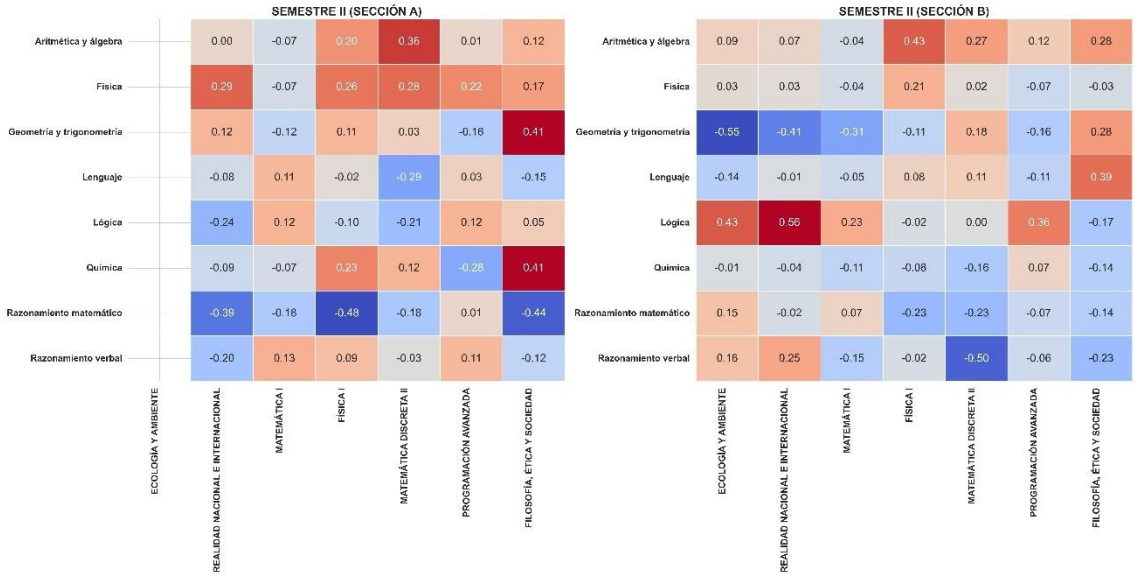
Matriz de correlación por cursos y secciones del I SEMESTRE - ESIS



Nota. Elaboración propia.

Figura 29

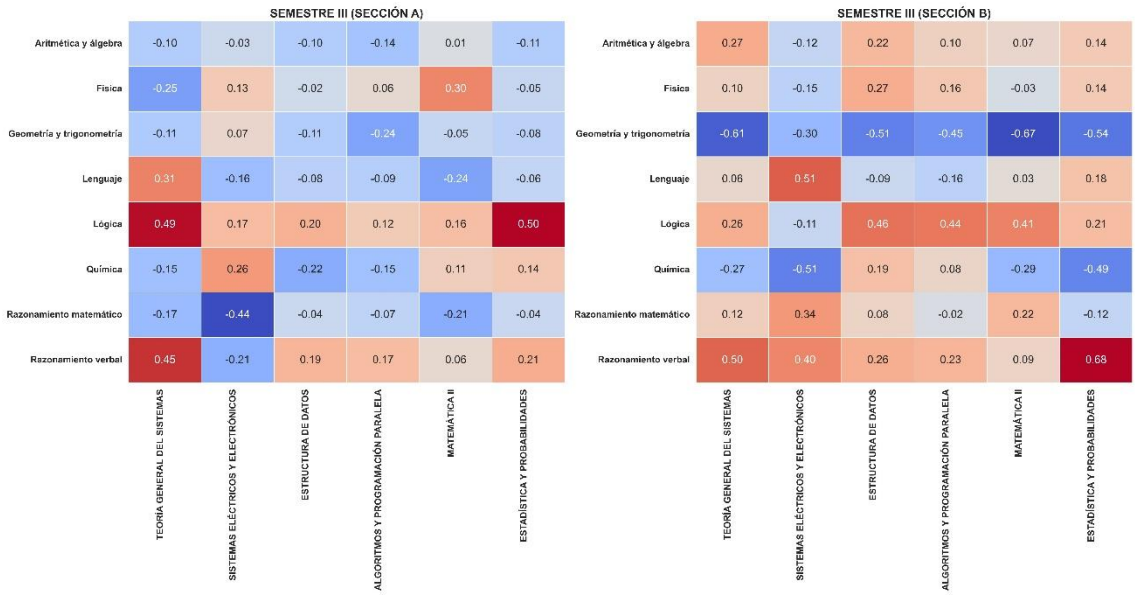
Matriz de correlación por cursos y secciones del II SEMESTRE - ESIS



Nota. Elaboración propia.

Figura 30

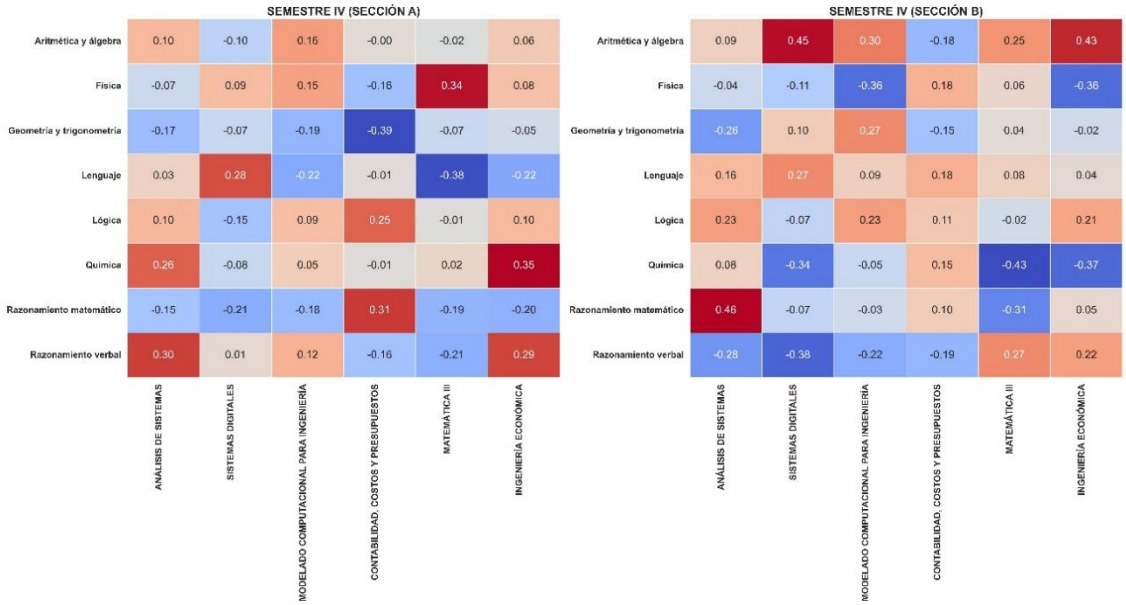
Matriz de correlación por cursos y secciones del III SEMESTRE - ESIS



Nota. Elaboración propia.

Figura 31

Matriz de correlación por cursos y secciones del IV SEMESTRE - ESIS

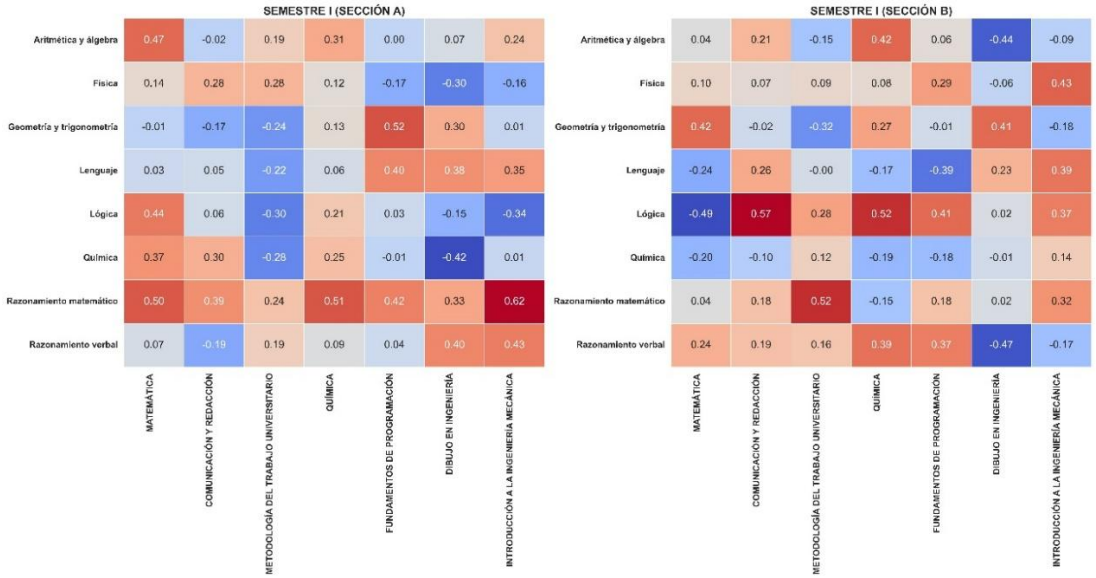


Nota. Elaboración propia.

Matriz de correlación de Ingeniería Mecánica

Figura 32

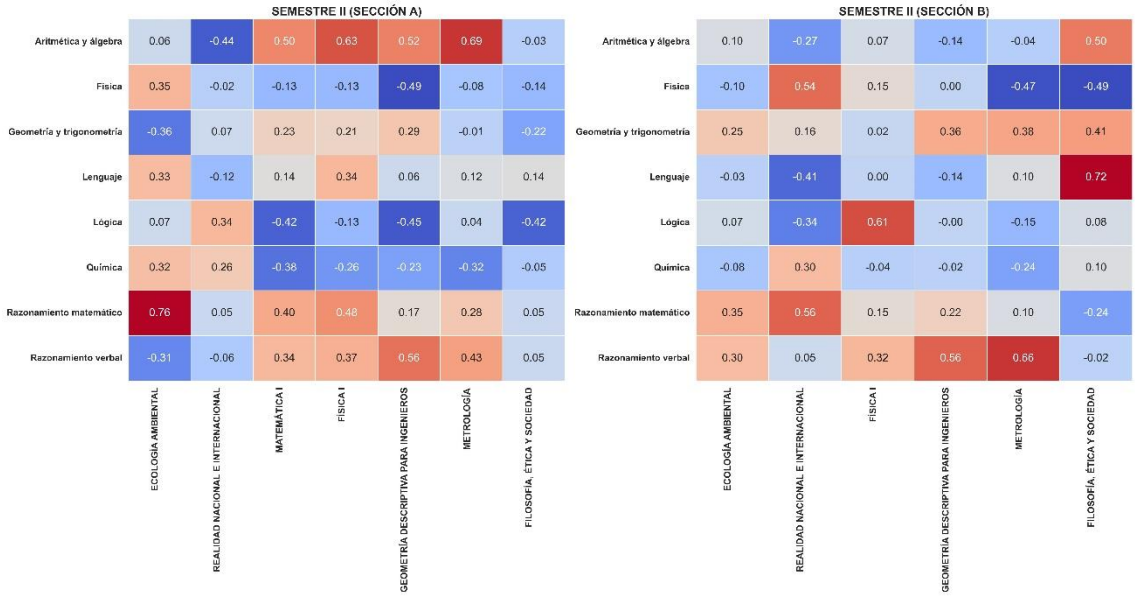
Matriz de correlación por cursos y secciones del I SEMESTRE - ESMC



Nota. Elaboración propia.

Figura 33

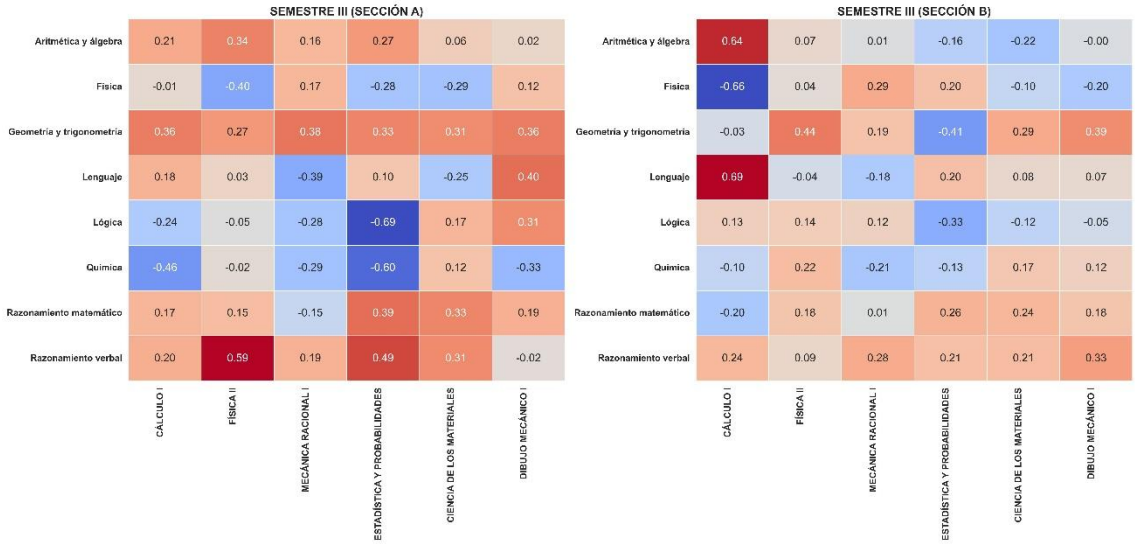
Matriz de correlación por cursos y secciones del II SEMESTRE - ESMC



Nota. Elaboración propia.

Figura 34

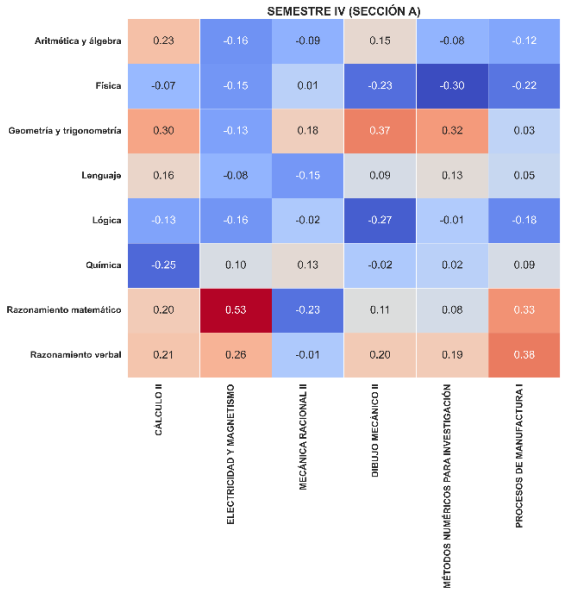
Matriz de correlación por cursos y secciones del III SEMESTRE - ESMC



Nota. Elaboración propia.

Figura 35

Matriz de correlación por cursos y secciones del IV SEMESTRE - ESMC

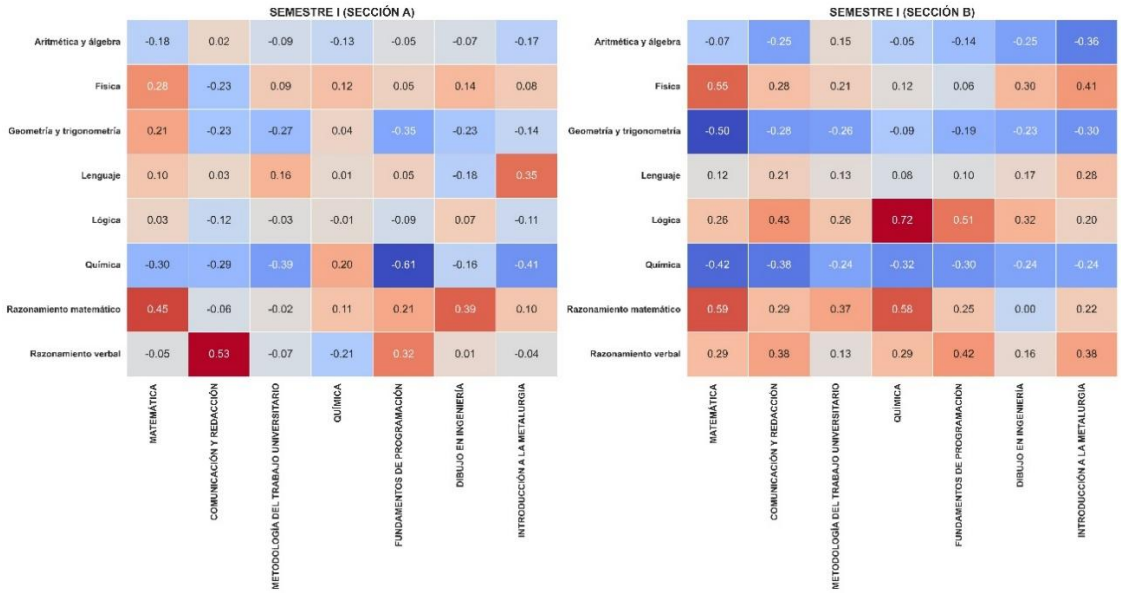


Nota. Elaboración propia.

Matriz de correlación de Ingeniería Metalúrgica

Figura 36

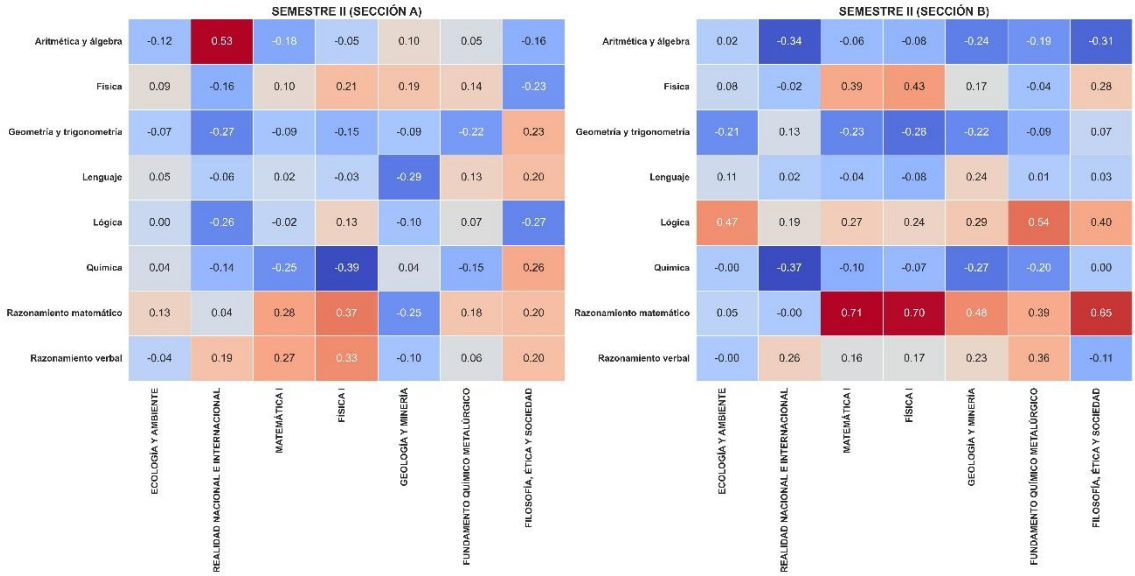
Matriz de correlación por cursos y secciones del I SEMESTRE - ESME



Nota. Elaboración propia.

Figura 37

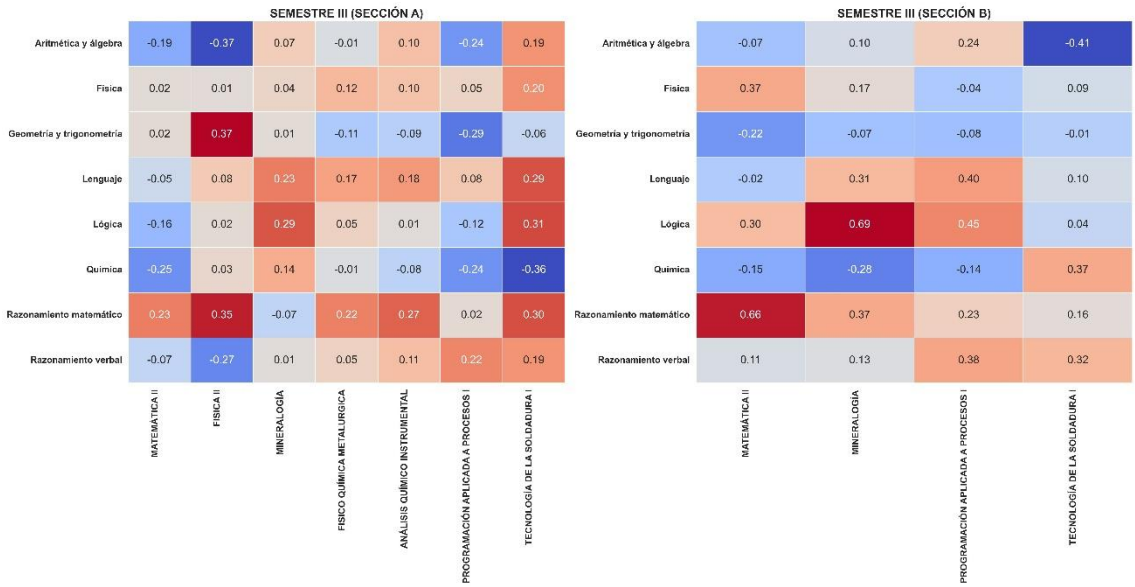
Matriz de correlación por cursos y secciones del II SEMESTRE - ESME



Nota. Elaboración propia.

Figura 38

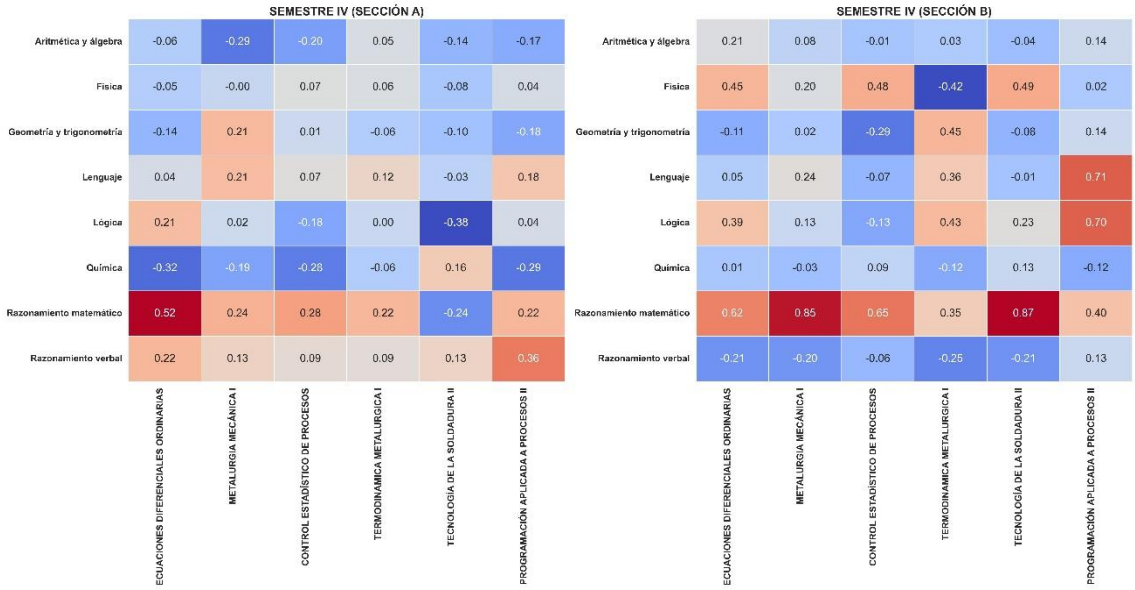
Matriz de correlación por cursos y secciones del III SEMESTRE - ESME



Nota. Elaboración propia.

Figura 39

Matriz de correlación por cursos y secciones del IV SEMESTRE - ESME

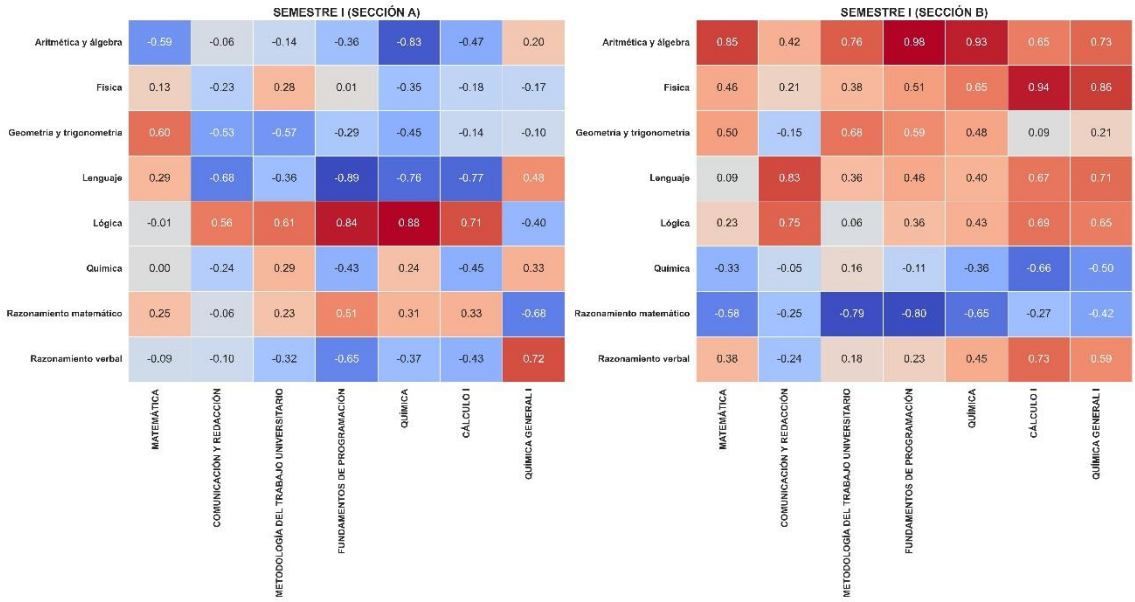


Nota. Elaboración propia.

Matriz de correlación de Ingeniería Química

Figura 40

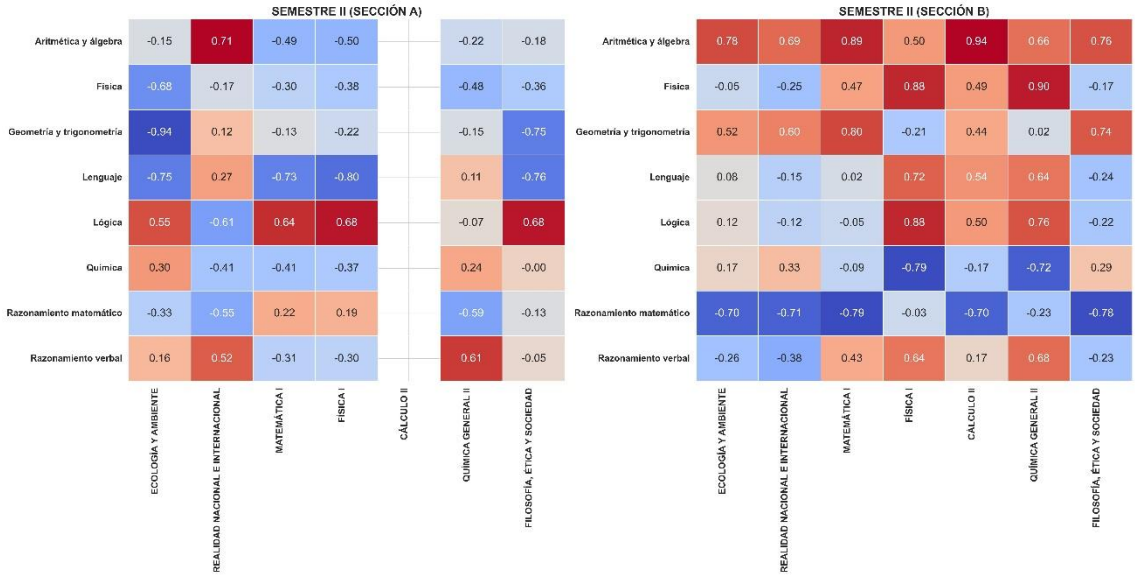
Matriz de correlación por cursos y secciones del I SEMESTRE - ESIQ



Nota. Elaboración propia.

Figura 41

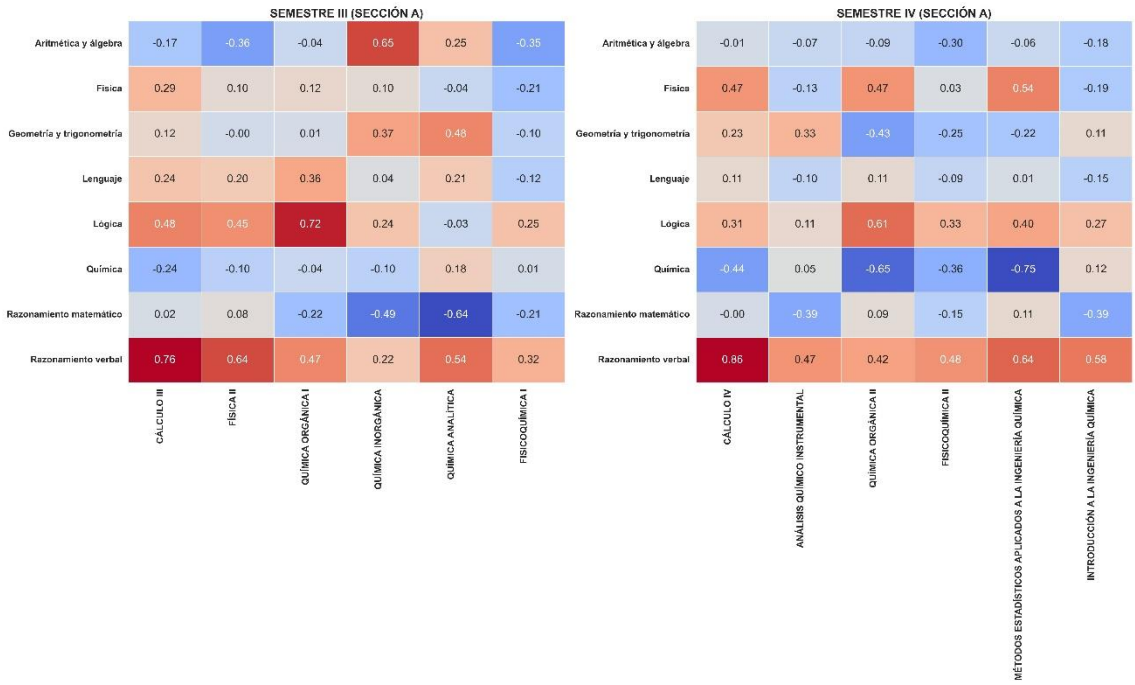
Matriz de correlación por cursos y secciones del II SEMESTRE - ESIQ



Nota. Elaboración propia.

Figura 42

Matriz de correlación por cursos y secciones del III - IV SEMESTRE - ESIQ



Nota. Elaboración propia.

4.1.2. Construir modelos de machine learning

Después de haber limpiado los datos, pasamos a construir los modelos de machine learning que nos ayudarán a predecir el rendimiento académico de los ingresantes a la Facultad de Ingeniería de la UNJBG. Para lograrlo, utilizamos el método de grid search para buscar el mejor hiperparámetro, y luego construir el modelo de machine learning. Hemos seleccionado cuatro enfoques distintos de modelado: regresión lineal (regresión ridge y regresión lasso), árboles de decisión**,** bosques aleatorios y redes neuronales artificiales**.** Cada uno de estos algoritmos aporta una perspectiva diferente al problema, lo que nos permite evaluar distintos escenarios y mejorar la capacidad predictiva del modelo.

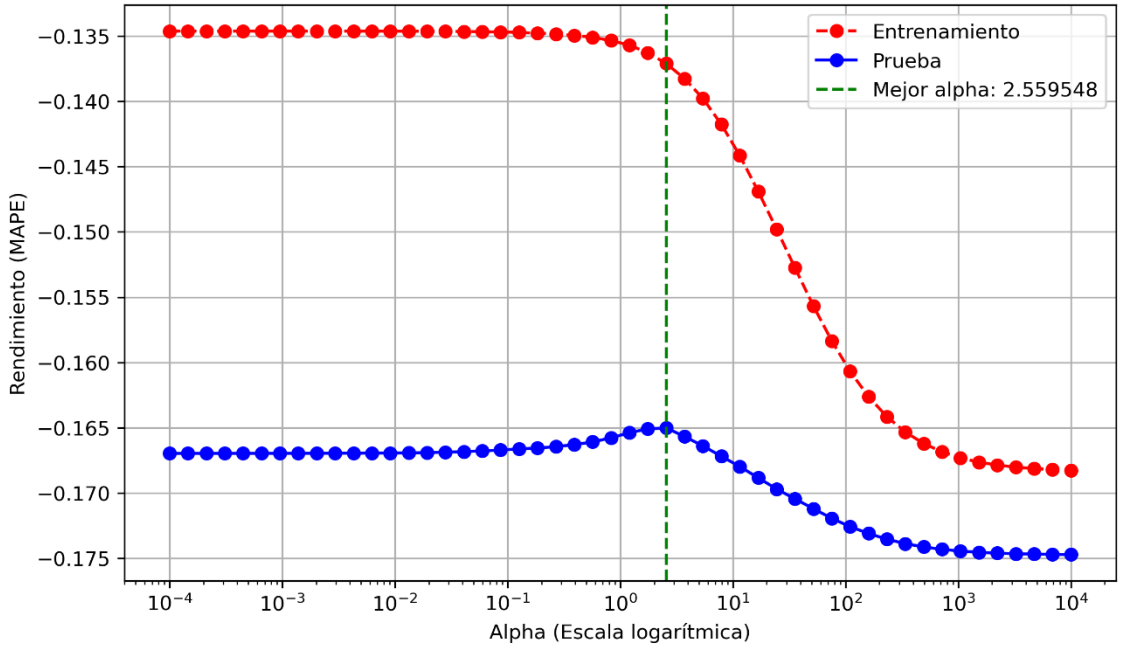
En este proceso, no solo consideramos los puntajes obtenidos en los cursos del examen de admisión: aritmética y álgebra, física, geometría y trigonometría, lenguaje, lógica, química, razonamiento matemático y razonamiento verbal; sino también otros factores relevantes como la sección en la que se matriculó cada estudiante y la escuela profesional a la que ingresó.

La regresión lineal es una técnica usada para predecir una variable a partir de otras. Sin embargo, cuando los datos son complicados o hay demasiadas variables, se utilizan los métodos ridge y lasso para mejorar el modelo.

4.1.2.1. Regresión ridge

Figura 43

Grid search para alpha de la regresión ridge



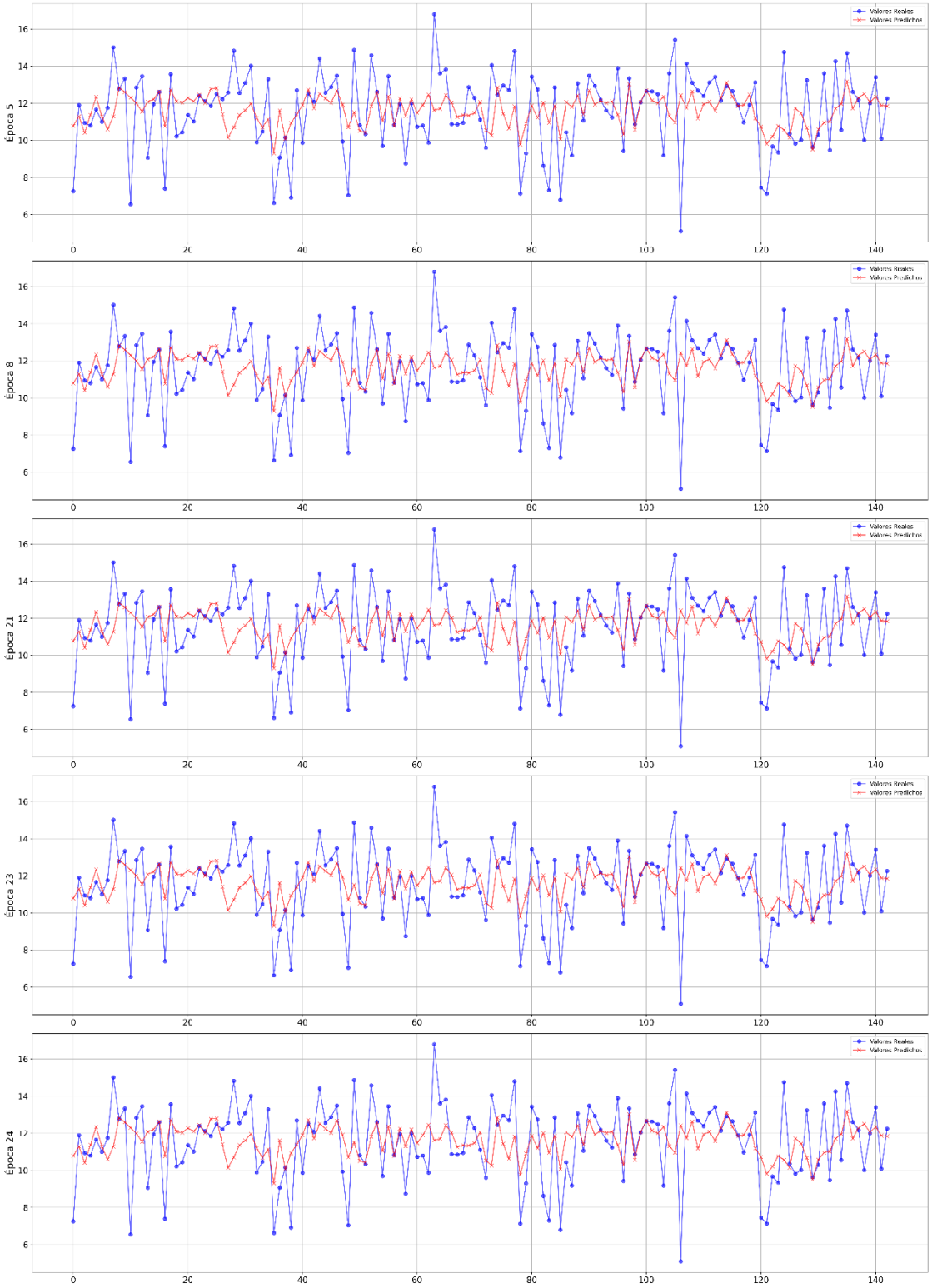
Nota. Elaboración propia.

En la figura 43, se puede visualizar un valor de α =2,559548, lo que indica que el modelo ha alcanzado su máximo rendimiento en datos no vistos (conjunto de prueba) mientras mantiene una adecuada precisión en el conjunto de entrenamiento. Esto demuestra que la regularización aplicada con este valor es la más adecuada para evitar el sobreajuste, garantizando que el modelo sea robusto y eficiente al enfrentar datos nuevos.

Se utilizó el método de grid search la cual se probó con 50 valores logarítmicos en el rango de 10^{-4} a 10^{4} valores para α, basándose en la minimización del MAPE**.** Este enfoque permitió encontrar el valor óptimo de α, maximizando así el rendimiento del modelo y garantizando una adecuada generalización sin sobreajuste.

Figura 44

Gráfico de líneas de predicción de las 5 mejores épocas de la regresión ridge



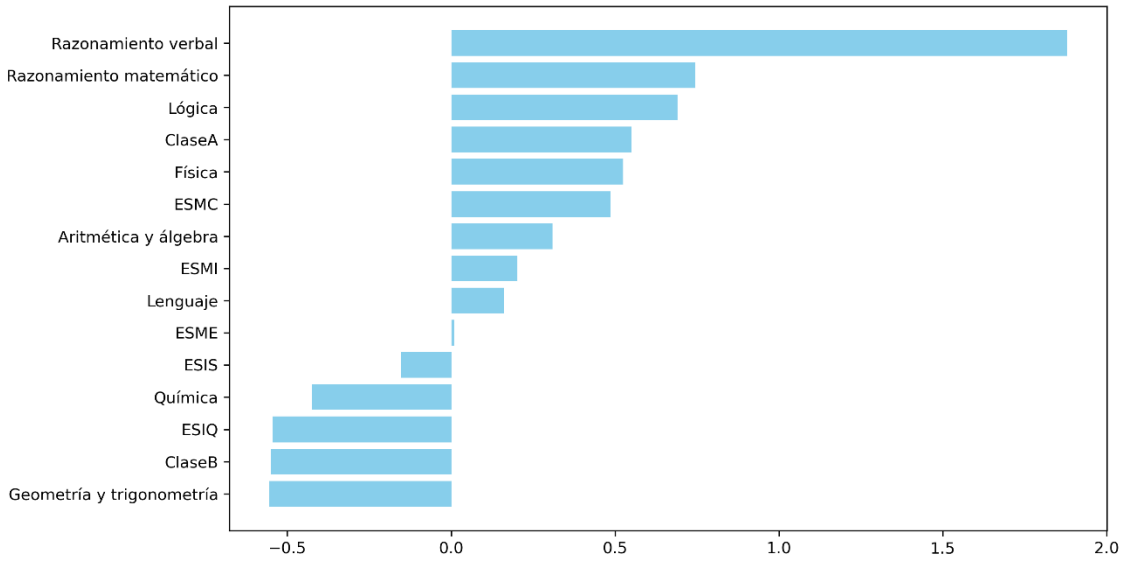
Nota. Elaboración propia.

En la figura 44 se comparan los valores reales (representados por los puntos azules) con los valores predichos (representados por las cruces rojas), pero utilizando líneas para ilustrar la tendencia de ambos conjuntos de datos. A pesar de que la regresión ridge es un modelo generalmente confiable para predecir tendencias, en este caso, podemos observar que los puntos reales (azules) están bastante alejados de las cruces rojas que representan los valores predichos en muchas partes de la gráfica. Esto indica que el modelo no ha logrado capturar correctamente el comportamiento de los datos.

El motivo de esta falta de precisión podría ser que el modelo no está ajustado correctamente a los datos. En particular, cuando los valores reales (puntos azules) muestran cambios bruscos o fluctuaciones, las cruces rojas no logran seguir esas oscilaciones, lo que indica que el modelo no puede capturar bien esos movimientos. Esto puede suceder si el modelo no tiene la capacidad suficiente para adaptarse a las variaciones de los datos.

Figura 45

Gráfico de feature importance de la regresión ridge



Nota. Elaboración propia.

En la figura 45 se aprecia la importancia de las características del algoritmo de regresión ridge, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción

del desempeño académico. Este curso tiene una fuerte relación con el éxito académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

A continuación, razonamiento matemático ocupa el segundo lugar en términos de importancia. Este curso, al igual que razonamiento verbal, tiene un gran impacto en las predicciones, lo que sugiere que la habilidad en matemáticas también juega un papel crucial en el rendimiento académico universitario. Lógica es otro de los cursos relevantes, aunque su influencia es algo menor en comparación con los anteriores, sigue siendo un factor importante para predecir el éxito académico.

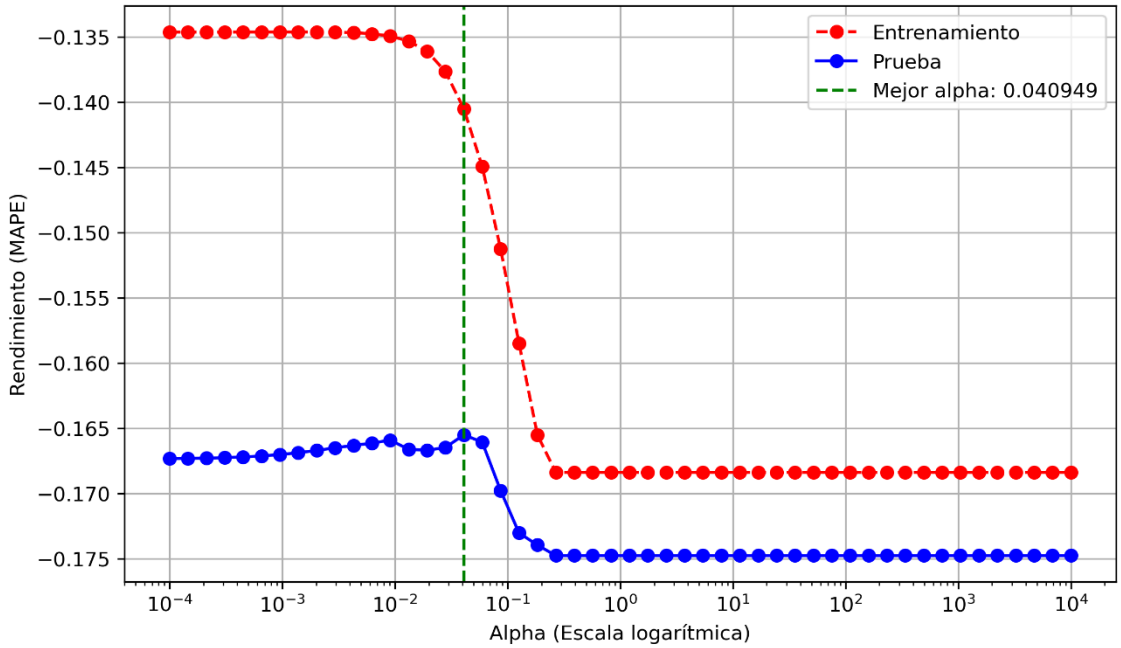
En términos de impacto moderado, física y aritmética y álgebra también son cursos significativos. Aunque no son tan determinantes como razonamiento verbal, su relevancia es considerable en el modelo de predicción. Estos cursos influyen en el desempeño académico de los estudiantes, pero su peso no es tan alto como el de los cursos más esenciales.

Por último, geometría y trigonometría**,** lenguaje y química son los cursos que tienen una influencia menor en la predicción del rendimiento. Aunque estos cursos siguen siendo relevantes, su impacto es relativamente bajo en comparación con los demás. Esto sugiere que, aunque son parte importante del proceso de admisión y tienen su lugar en la formación de los estudiantes, no son tan cruciales para predecir el rendimiento académico universitario.

4.1.2.2. Regresión lasso

Figura 46

Grid search para alpha de la regresión lasso



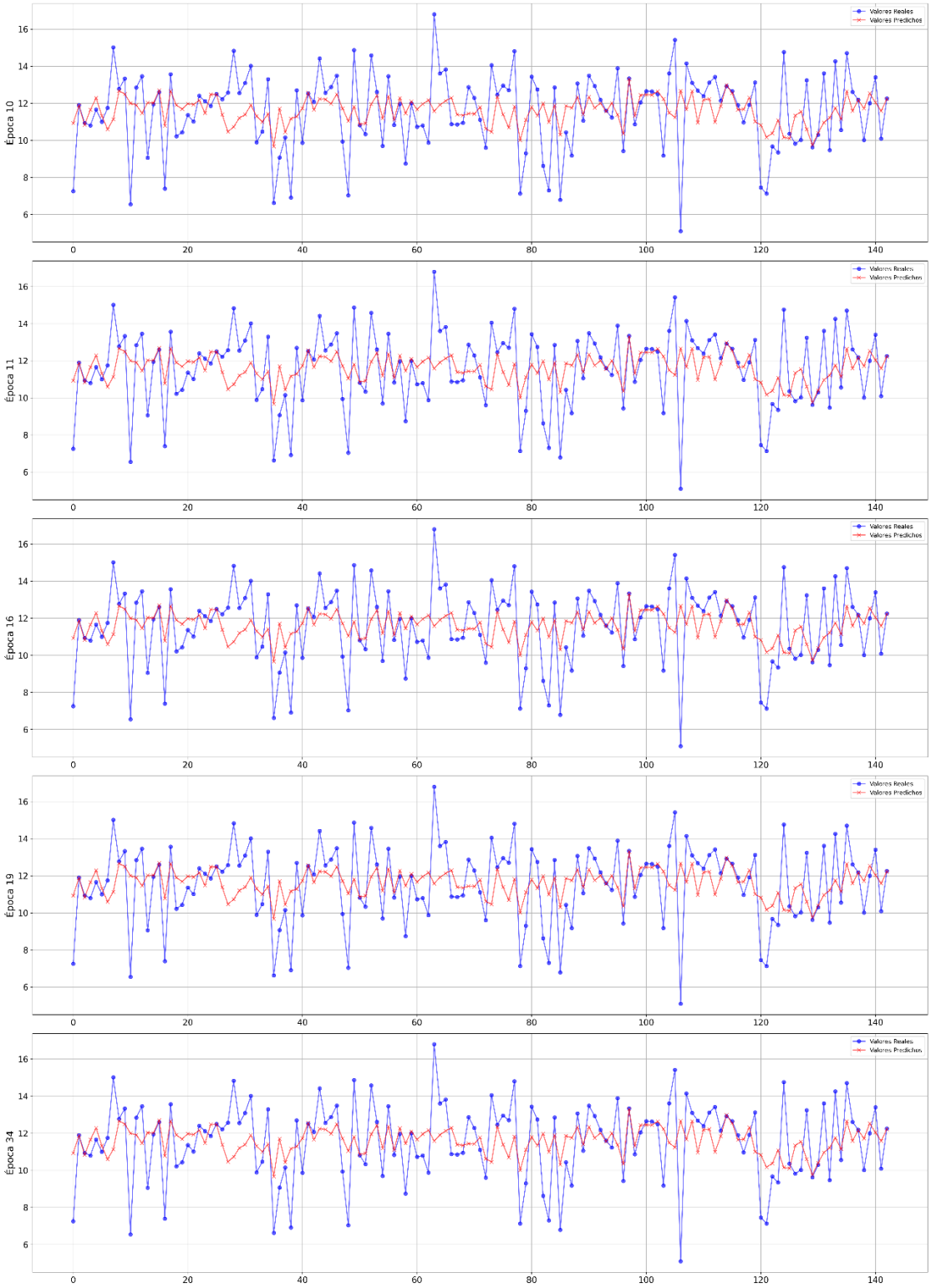
Nota. Elaboración propia.

En la figura 46 se puede visualizar un valor de α =0,040949, lo que indica que el modelo ha alcanzado su máximo rendimiento en datos no vistos (conjunto de prueba) mientras mantiene una adecuada precisión en el conjunto de entrenamiento. Esto demuestra que la regularización aplicada con este valor es la más adecuada para evitar el sobreajuste, garantizando que el modelo sea robusto y eficiente al enfrentar datos nuevos.

Se utilizó el método de grid search la cual se probó con 50 valores logarítmicos en el rango de 10^{-4} a 10^{4} valores para α, basándose en la minimización del MAPE**.** Este enfoque permitió encontrar el valor óptimo de α, maximizando así el rendimiento del modelo y garantizando una adecuada generalización sin sobreajuste.

Figura 47

Gráfico de líneas de predicción de las 5 mejores épocas de la regresión lasso



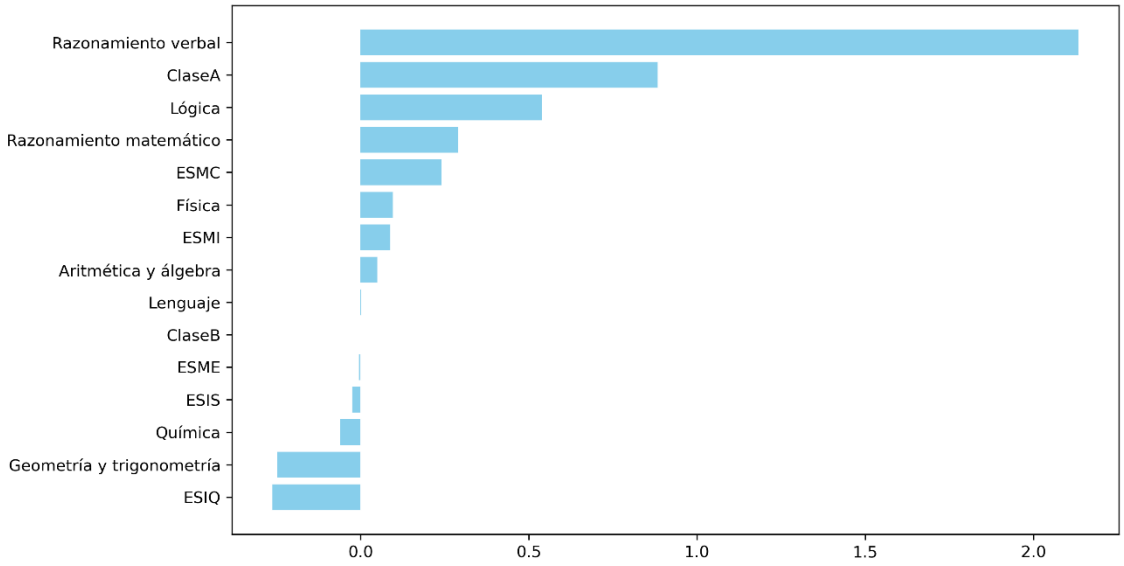
Nota. Elaboración propia.

En la figura 47 se comparan los valores reales (representados por los puntos azules) con los valores predichos (representados por las cruces rojas), pero utilizando líneas para ilustrar la tendencia de ambos conjuntos de datos. A pesar de que la regresión lasso es un modelo generalmente confiable para predecir tendencias, en este caso, podemos observar que los puntos reales (azules) están bastante alejados de las cruces rojas que representan los valores predichos en muchas partes de la gráfica. Esto indica que el modelo no ha logrado capturar correctamente el comportamiento de los datos.

El motivo podría ser que la regresión lasso impone una penalización en los coeficientes del modelo, lo que lleva a una regularización más fuerte, reduciendo la capacidad del modelo para adaptarse a todas las variaciones de los datos.

Figura 48

Gráfico de feature importance de la regresión lasso



Nota. Elaboración propia.

En la figura 48 se aprecia la importancia de las características del algoritmo de regresión lasso, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Este curso tiene una fuerte relación con el éxito académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

A continuación, lógica ocupa el segundo lugar en términos de importancia. Este curso, al igual que razonamiento verbal, tiene un gran impacto en las predicciones, lo que sugiere que la habilidad en pensamiento crítico y resolución de problemas también juega un papel crucial en el rendimiento académico universitario. Razonamiento matemático es otro de los cursos relevantes, aunque su influencia es algo menor en comparación con los anteriores, sigue siendo un factor importante para predecir el éxito académico.

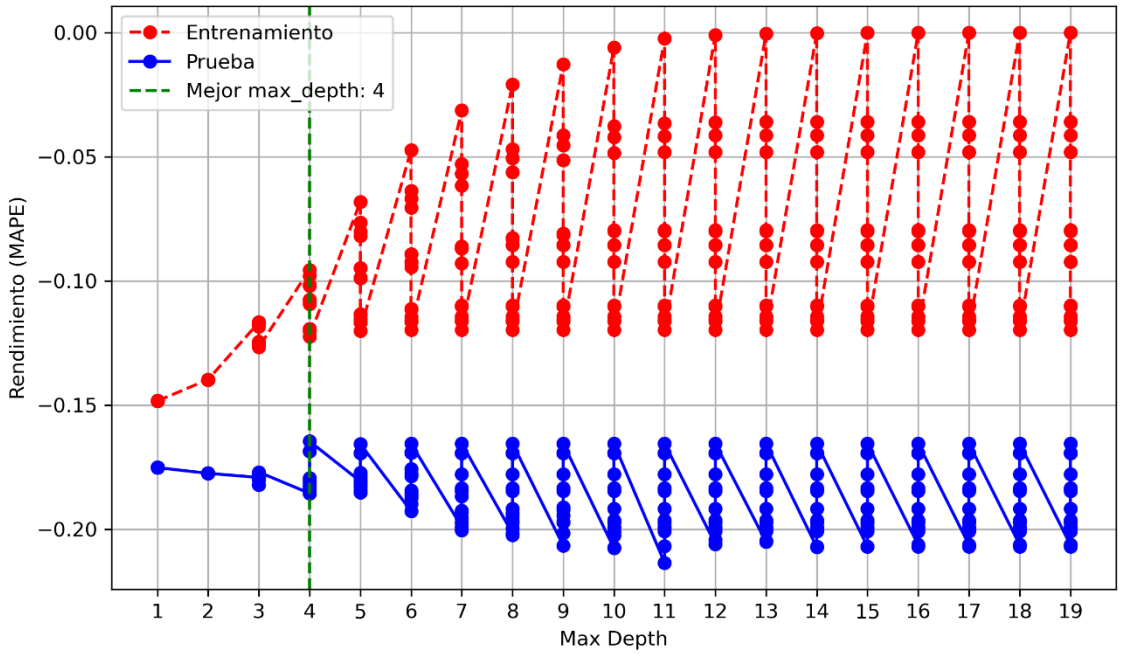
En términos de impacto moderado, física y aritmética y álgebra también son cursos significativos. Aunque no son tan determinantes como razonamiento verbal, su relevancia es considerable en el modelo de predicción. Estos cursos influyen en el desempeño académico de los estudiantes, pero su peso no es tan alto como el de los cursos más esenciales.

Por último, geometría y trigonometría**,** lenguaje y química son los cursos que tienen una influencia menor en la predicción del rendimiento. Aunque estos cursos siguen siendo relevantes, su impacto es relativamente bajo en comparación con los demás. Esto sugiere que, aunque son parte importante del proceso de admisión y tienen su lugar en la formación de los estudiantes, no son tan cruciales para predecir el rendimiento académico universitario.

4.1.2.3. Árbol de decisión

Figura 49

Grid Search para depth de árbol de decisión

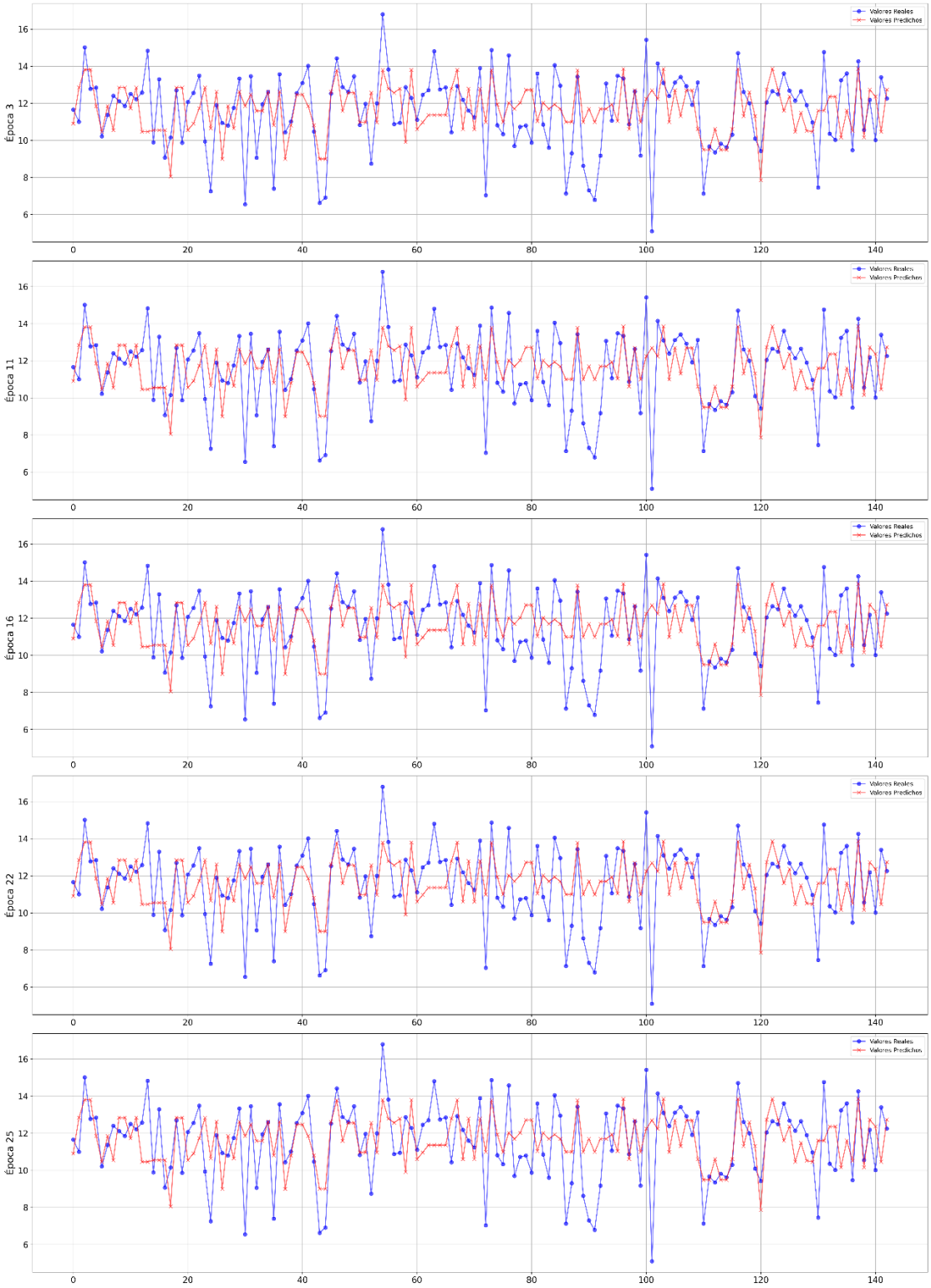


Nota. Elaboración propia.

En la figura 49 se puede ver que la mejor profundidad para el árbol de decisión es 4, lo que indica que el modelo logra un equilibrio ideal entre aprender lo suficiente sin volverse demasiado detallado. Si la profundidad fuera menor, el modelo sería demasiado simple, como si intentara predecir el rendimiento académico de un estudiante basándose solo en uno o dos cursos del examen de admisión, ignorando información valiosa. En cambio, si fuera demasiado profundo, el modelo intentaría ajustarse a cada pequeña variación en los puntajes de los cursos, incluso aquellas que no tienen un impacto real en el rendimiento universitario, perdiendo la capacidad de hacer buenas predicciones. Con una profundidad de 4, el modelo aprende lo necesario sin caer en excesos.

Figura 50

Gráfico de líneas de predicción de las 5 mejores épocas de árbol de decisión



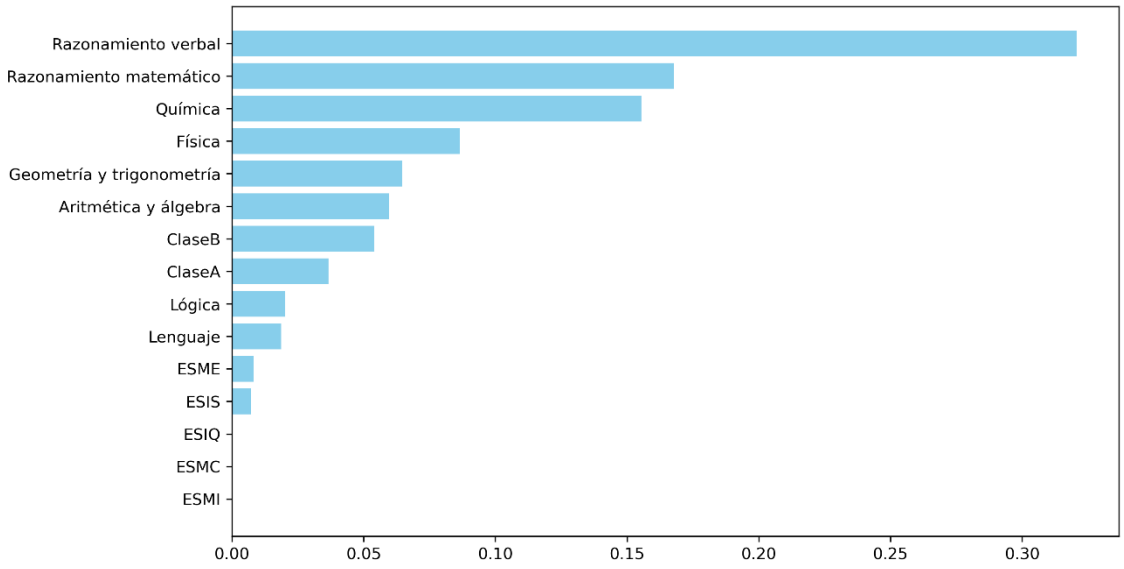
Nota. Elaboración propia.

En la figura 50 se comparan los valores reales (representados por los puntos azules) con los valores predichos (representados por las cruces rojas), pero utilizando líneas para ilustrar la tendencia de ambos conjuntos de datos. A pesar de que la regresión árbol de decisión son conocidos por ser efectivos en la captura de relaciones no lineales en los datos, en este caso, podemos observar que las cruces rojas (predicciones) se desvían significativamente de los puntos azules (valores reales) en varias ocasiones. Este alejamiento indica que el modelo no ha logrado capturar correctamente algunas variaciones en los datos.

El motivo puede ser que el modelo está haciendo predicciones muy precisas para los datos que ya ha visto, pero cuando se le presentan datos nuevos o diferentes, no es capaz de hacer buenas predicciones. En este caso, el modelo de árbol de decisión no está siendo lo suficientemente flexible para adaptarse correctamente a los nuevos datos sin perder precisión.

Figura 51

Gráfico de feature importance de árbol de decisión



Nota. Elaboración propia.

En la figura 51 se aprecia la importancia de las características del algoritmo de árbol de decisión, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la

predicción del desempeño académico. Este curso tiene una fuerte relación con el éxito académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

A continuación, razonamiento matemático ocupa el segundo lugar en términos de importancia. Este curso, al igual que razonamiento verbal, tiene un gran impacto en las predicciones, lo que sugiere que la habilidad en matemáticas también juega un papel crucial en el rendimiento académico universitario. Química es otro de los cursos relevantes, aunque su influencia es algo menor en comparación con los anteriores, sigue siendo un factor importante para predecir el éxito académico.

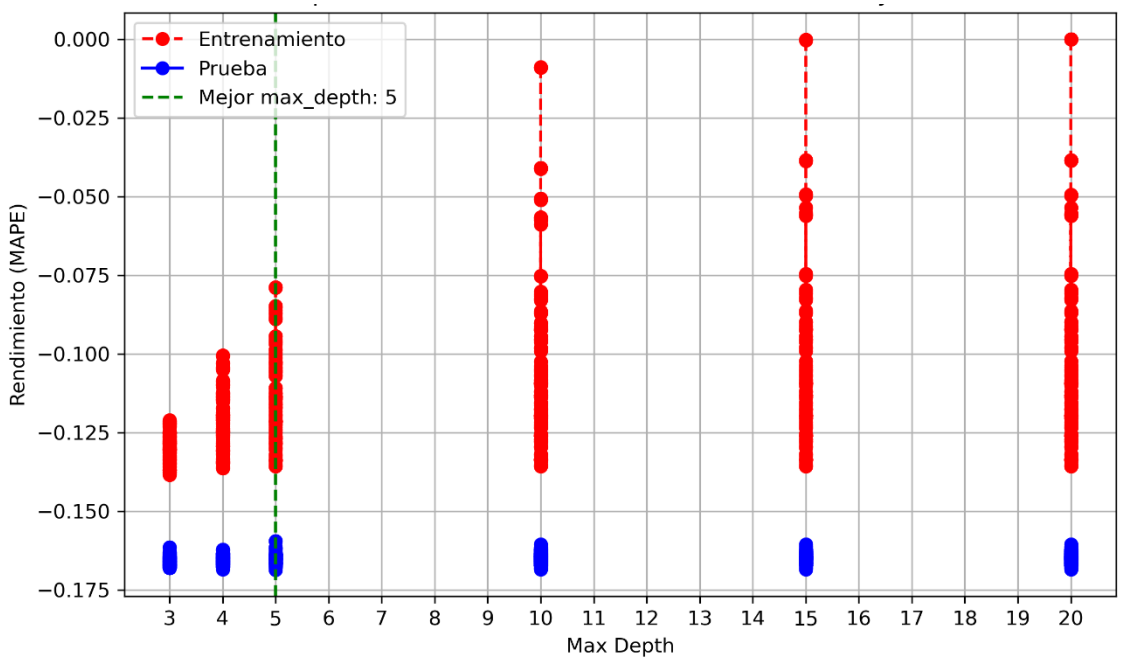
En términos de impacto moderado, física**,** geometría y trigonometría y aritmética y álgebra también son cursos significativos. Aunque no son tan determinantes como razonamiento verbal, su relevancia es considerable en el modelo de predicción. Estos cursos influyen en el desempeño académico de los estudiantes, pero su peso no es tan alto como el de los cursos más esenciales.

Por último, lenguaje y lógica son los cursos que tienen una influencia menor en la predicción del rendimiento. Aunque estos cursos siguen siendo relevantes, su impacto es relativamente bajo en comparación con los demás. Esto sugiere que, aunque son parte importante del proceso de admisión y tienen su lugar en la formación de los estudiantes, no son tan cruciales para predecir el rendimiento académico universitario.

4.1.2.4. Bosques aleatorios

Figura 52

Grid search para depth de bosques aleatorios



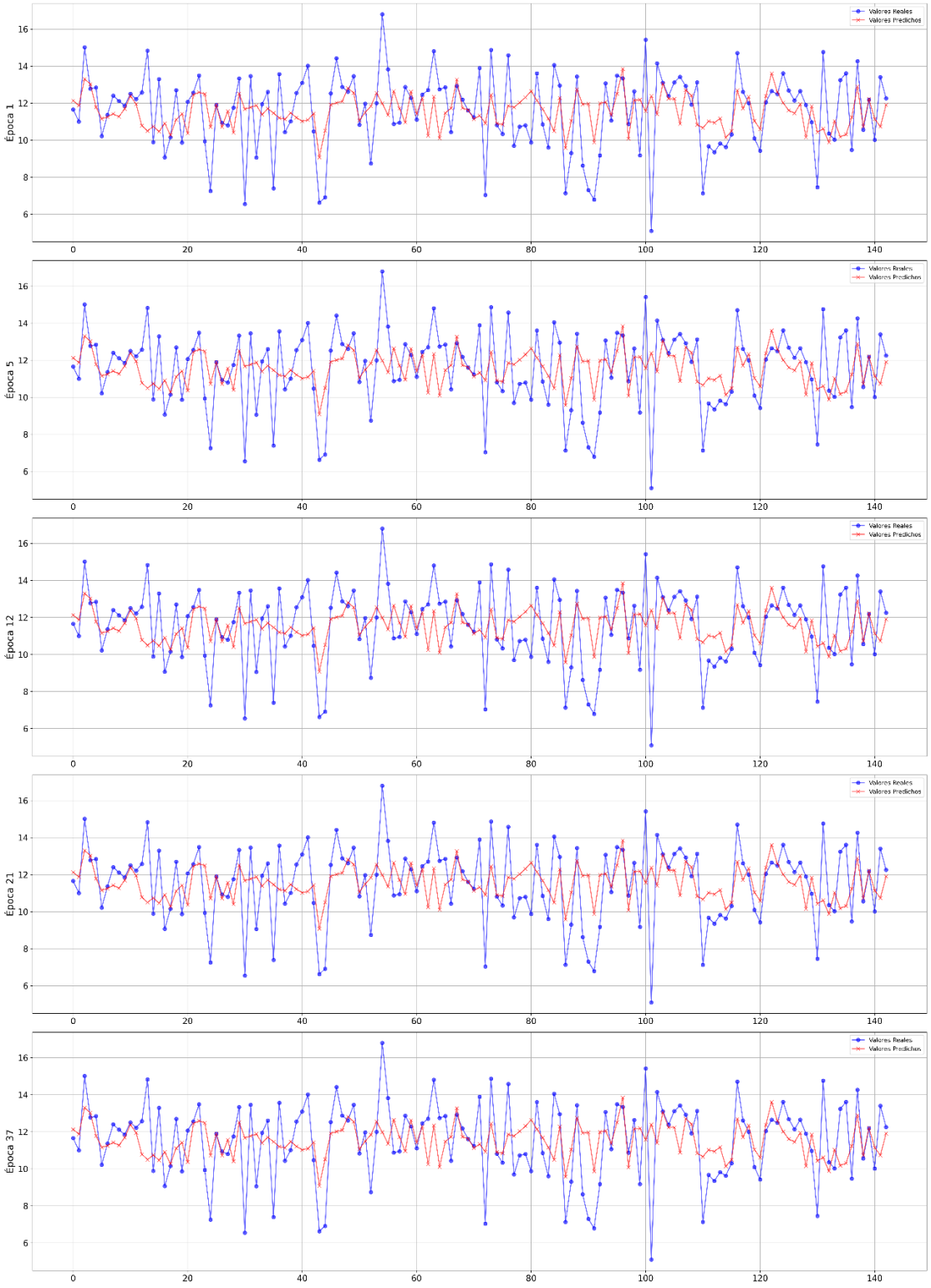
Nota. Elaboración propia.

En la figura 52 se observa que la mejor profundidad para el modelo de bosques aleatorios es 5, lo que indica que los árboles dentro del bosque aprenden lo suficiente de los cursos del examen de admisión sin caer en patrones demasiado específicos. Si la profundidad fuera menor, el modelo podría perder información valiosa sobre la relación entre los cursos y el rendimiento académico universitario. En cambio, si la profundidad fuera demasiado grande, cada árbol se ajustaría demasiado a los datos de entrenamiento, lo que haría que el modelo fuera menos capaz de hacer buenas predicciones para nuevos estudiantes.

En el gráfico, cuando max_depth = 10 se observa una mayor variabilidad en los puntos, lo que indica que el modelo no es tan estable en ese nivel de profundidad; se aprecia que el modelo parece empezar a ajustarse demasiado a los datos de entrenamiento, lo que podría afectar su capacidad de predecir correctamente con nuevos datos.

Figura 53

Gráfico de líneas de predicción de las 5 mejores épocas de bosques aleatorios



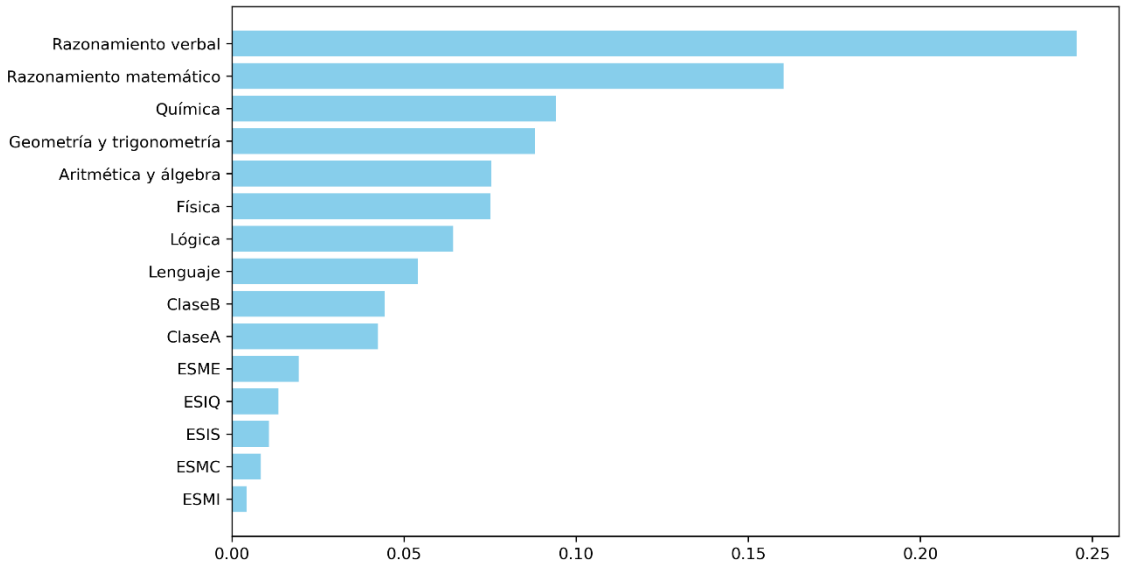
Nota. Elaboración propia.

En la figura 53 se comparan los valores reales (representados por los puntos azules) con los valores predichos (representados por las cruces rojas), pero utilizando líneas para ilustrar la tendencia de ambos conjuntos de datos. Aunque bosques aleatorios es un modelo robusto, preciso, y fácil de usar, especialmente para reducir el sobreajuste y manejar grandes volúmenes de datos; en este caso, podemos observar que las cruces rojas (predicciones) se desvían significativamente de los puntos azules (valores reales) en varias ocasiones. Este alejamiento indica que el modelo no ha logrado capturar correctamente algunas variaciones en los datos.

El motivo puede ser que el modelo está haciendo predicciones muy precisas para los datos que ya ha visto, pero cuando se le presentan datos nuevos o diferentes, no es capaz de hacer buenas predicciones. En este caso, el modelo de bosques aleatorios no está siendo lo suficientemente flexible para adaptarse correctamente a los nuevos datos sin perder precisión.

Figura 54

Gráfico de feature importance de bosques aleatorios



Nota. Elaboración propia.

En la figura 54 se aprecia la importancia de las características del algoritmo de bosques aleatorios, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la

predicción del desempeño académico. Este curso tiene una fuerte relación con el éxito académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

A continuación, razonamiento matemático ocupa el segundo lugar en términos de importancia. Este curso, al igual que razonamiento verbal, tiene un gran impacto en las predicciones, lo que sugiere que la habilidad en matemáticas también juega un papel crucial en el rendimiento académico universitario. Química es otro de los cursos relevantes, aunque su influencia es algo menor en comparación con los anteriores, sigue siendo un factor importante para predecir el éxito académico.

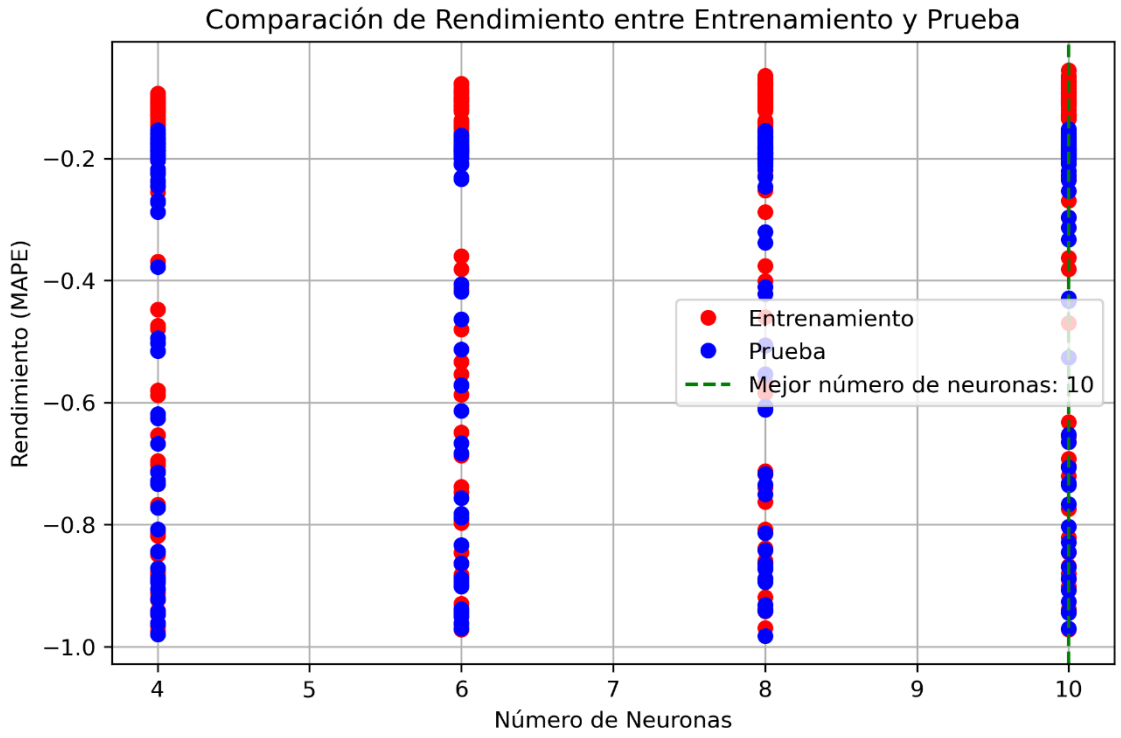
En términos de impacto moderado, geometría y trigonometría, aritmética y álgebra y física también son cursos significativos. Aunque no son tan determinantes como razonamiento verbal, su relevancia es considerable en el modelo de predicción. Estos cursos influyen en el desempeño académico de los estudiantes, pero su peso no es tan alto como el de los cursos más esenciales.

Por último, lenguaje y lógica son los cursos que tienen una influencia menor en la predicción del rendimiento. Aunque estos cursos siguen siendo relevantes, su impacto es relativamente bajo en comparación con los demás. Esto sugiere que, aunque son parte importante del proceso de admisión y tienen su lugar en la formación de los estudiantes, no son tan cruciales para predecir el rendimiento académico universitario.

4.1.2.5. Redes neuronales

Figura 55

Grid search para el número de neuronas de redes neuronales

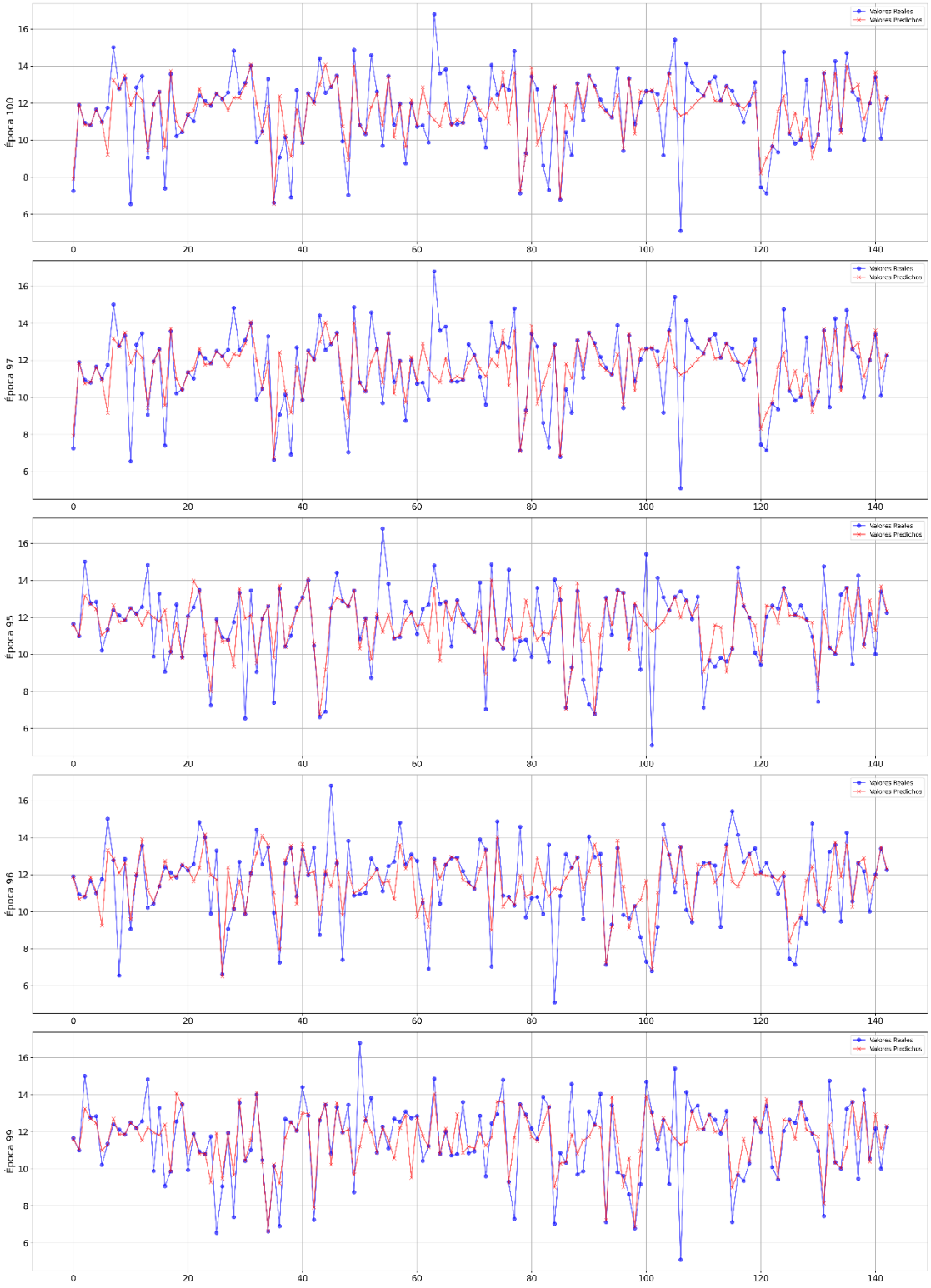


Nota. Elaboración propia.

En la figura 55 se observa que el mejor número de neuronas en la capa oculta es 10, Si el número de neuronas fuera menor, la red no tendría suficiente capacidad para identificar patrones en los datos, resultando en predicciones poco precisas. Por otro lado, si el número de neuronas aumentara demasiado, el modelo podría volverse más complejo de lo necesario, lo que no garantiza una mejora en la precisión y podría llevar a problemas como el sobreajuste. Este comportamiento se debe a que una red neuronal con demasiadas neuronas tiene mayor capacidad de memoria y puede aprender incluso patrones irrelevantes o ruido en los datos, en lugar de enfocarse en las relaciones generales y útiles. Por eso, el punto óptimo de 10 neuronas asegura que el modelo generalice bien sin volverse demasiado simple o propenso a errores.

Figura 56

Gráfico de líneas de predicción de las 5 mejores épocas de redes neuronales



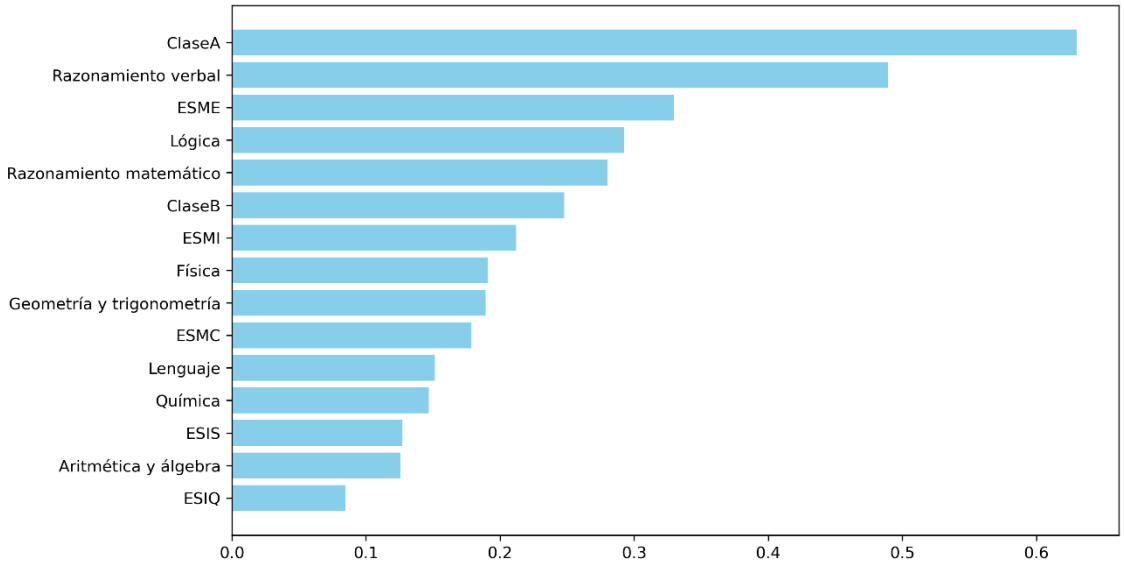
Nota. Elaboración propia.

En la figura 56 se comparan los valores reales (representados por los puntos azules) con los valores predichos (representados por las cruces rojas), pero utilizando líneas para ilustrar la tendencia de ambos conjuntos de datos. Se puede observar una alta precisión entre las líneas azul y roja, lo que indica que la red neuronal fue capaz de aprender el patrón entre los datos del examen de admisión y el rendimiento académico.

A simple vista se puede apreciar que la red neuronal fue la mejor alternativa para predecir el rendimiento académico en función de los datos del examen de admisión, ya que logró acercarse con precisión a los valores reales, incluso con variabilidad natural en los datos. El modelo logró mantener buenos resultados en diferentes etapas del entrenamiento, lo que demuestra que entendió bien los patrones en los datos. Por eso, se puede decir que la red neuronal es una herramienta útil y confiable para este tipo de predicciones.

Figura 57

Gráfico de feature importance de redes neuronales



Nota. Elaboración propia.

En la figura 57 se aprecia la importancia de las características del algoritmo de redes neuronales, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Este curso tiene una fuerte relación con el éxito

académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

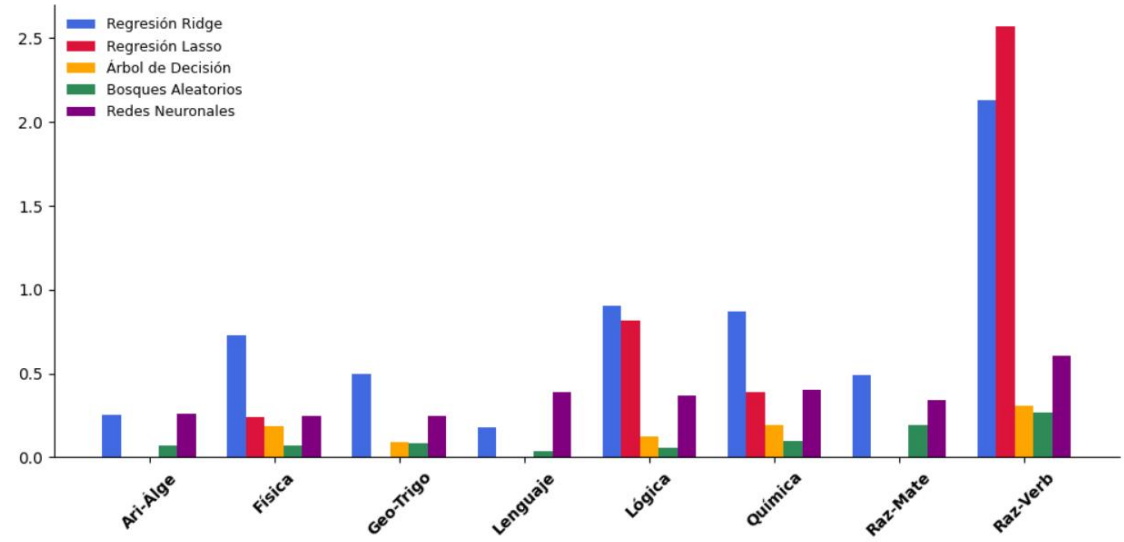
A continuación, lógica ocupa el segundo lugar en términos de importancia. Este curso, al igual que razonamiento verbal, tiene un gran impacto en las predicciones, lo que sugiere que la habilidad en matemáticas también juega un papel crucial en el rendimiento académico universitario. Razonamiento matemático es otro de los cursos relevantes, aunque su influencia es algo menor en comparación con los anteriores, sigue siendo un factor importante para predecir el éxito académico.

En términos de impacto moderado, física y geometría y trigonometría y lenguaje también son cursos significativos. Aunque no son tan determinantes como razonamiento verbal, su relevancia es considerable en el modelo de predicción. Estos cursos influyen en el desempeño académico de los estudiantes, pero su peso no es tan alto como el de los cursos más esenciales.

Por último, química y aritmética y álgebra son los cursos que tienen una influencia menor en la predicción del rendimiento. Aunque estos cursos siguen siendo relevantes, su impacto es relativamente bajo en comparación con los demás. Esto sugiere que, aunque son parte importante del proceso de admisión y tienen su lugar en la formación de los estudiantes, no son tan cruciales para predecir el rendimiento académico universitario. Lo que llama la atención es que la red neuronal revela que, si un estudiante pertenece a la sección A, es la más influyente en la predicción del rendimiento académico universitario. Su peso supera considerablemente al de las demás variables, incluidas las notas de cursos del examen de admisión, la cual se evidencia que el hecho de haber estudiado en la sección A tiene una relación fuerte con el desempeño académico posterior.

Figura 58

Gráfico de feature importance de los modelos de machine learning



Nota. Elaboración propia.

En la figura 58 se aprecia la importancia de las características de los algoritmos de machine learning, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Como también se observa que el algoritmo de regresión lasso alcanza la mayor importancia absoluta en el curso de razonamiento verbal, sin embargo, lasso también presenta valores nulos en varias asignaturas (aritmética y álgebra, geometría y trigonometría, lenguaje y razonamiento matemático), lo que sugiere que, debido a su naturaleza de regularización L1, eliminó completamente la influencia de estas variables en la predicción.

Por otro lado, redes neuronales asigna su mayor importancia a razonamiento verbal, seguido de química, lenguaje y lógica, mostrando un patrón más diversificado que lasso, pero con menor magnitud de pesos que ridge.

4.1.3. Evaluar los modelos de machine learning

Después de haber construido los algoritmos de machine learning y entrenado nuestros modelos con los datos disponibles durante 100 épocas, el siguiente paso crucial es evaluar su rendimiento para asegurar que las predicciones generadas sean lo más

precisas posible. Para realizar esta evaluación, empleamos una serie de métricas que nos permiten cuantificar la precisión de los modelos. En este caso, nos centramos en cuatro métricas fundamentales: el error absoluto medio (MAE), el error cuadrático medio (MSE), la raíz de error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE).

La evaluación de los algoritmos de machine learning a través de estas cuatro métricas proporciona una visión integral del desempeño de nuestros modelos, permitiéndonos analizar su rendimiento desde distintas perspectivas y bajo diferentes enfoques.

Para evaluar el rendimiento de los algoritmos como regresión lineal**,** regresión ridge**,** regresión lasso**,** árbol de decisión, bosques aleatorios y redes neuronales artificiales, se utilizaron varias métricas, pero le dimos más peso al error cuadrático medio (MSE). El MSE es especialmente útil porque resalta los errores más grandes, lo que nos ayuda a detectar cuando el modelo está haciendo predicciones muy alejadas de los valores reales. Al ser una métrica que eleva al cuadrado las diferencias, un solo error grande afecta mucho su valor, lo que permite que el modelo se ajuste para minimizar esos errores. Aunque también se usaron otras métricas como el MAE**,** RMSE y MAPE para tener una visión más completa, el MSE es la que nos da una idea más clara de cómo el modelo está manejando los errores en general.

4.1.3.1. Regresión ridge

Tabla 6

Lista de los 5 mejores folds de las métricas de desempeño de la regresión ridge

#		MAE	MSE	RMSE	MAPE
5 Fold	Train	1,427277	3,479814	1,864893	0,138793
	Test	1,478906	3,918222	1,979200	0,145207
8 Fold	Train	1,427277	3,479814	1,864893	0,138793
	Test	1,478906	3,918222	1,979200	0,145207
21 Fold	Train	1,427277	3,479814	1,864893	0,138793
	Test	1,478906	3,918222	1,979200	0,145207
23 Fold	Train	1,427277	3,479814	1,864893	0,138793
	Test	1,478906	3,918222	1,979200	0,145207
24 Fold	Train	1,427277	3,479814	1,864893	0,138793
	Test	1,478906	3,918222	1,979200	0,145207

Nota. Elaboración propia.

Tabla 7

Intervalos de confianza al 95 % de la regresión ridge

#		MAE	MSE	RMSE	MAPE
Train	Intervalo1	1,420964	3,481970	1,864902	0,138296
	Intervalo2	1,422431	3,484803	1,865752	0,138412
Test	Intervalo1	1,519847	4,062101	2,003866	0,148735
	Intervalo2	1,531213	4,097736	2,010485	0,149675

Nota. Elaboración propia.

Las métricas de desempeño proporcionadas en la Tabla 6 reflejan cómo el modelo de ridge ajusta las predicciones en relación con los valores reales de las notas obtenidas. En cuanto al error absoluto medio (MAE), se observa que el modelo tiene un desempeño bastante preciso, con un valor cercano a 1,42 para el conjunto de entrenamiento y 1,48 para el conjunto de prueba. Esto indica que, en promedio, la predicción de la nota está a menos de 2 puntos de la nota real, lo cual es un error aceptable para un modelo de predicción de rendimiento académico.

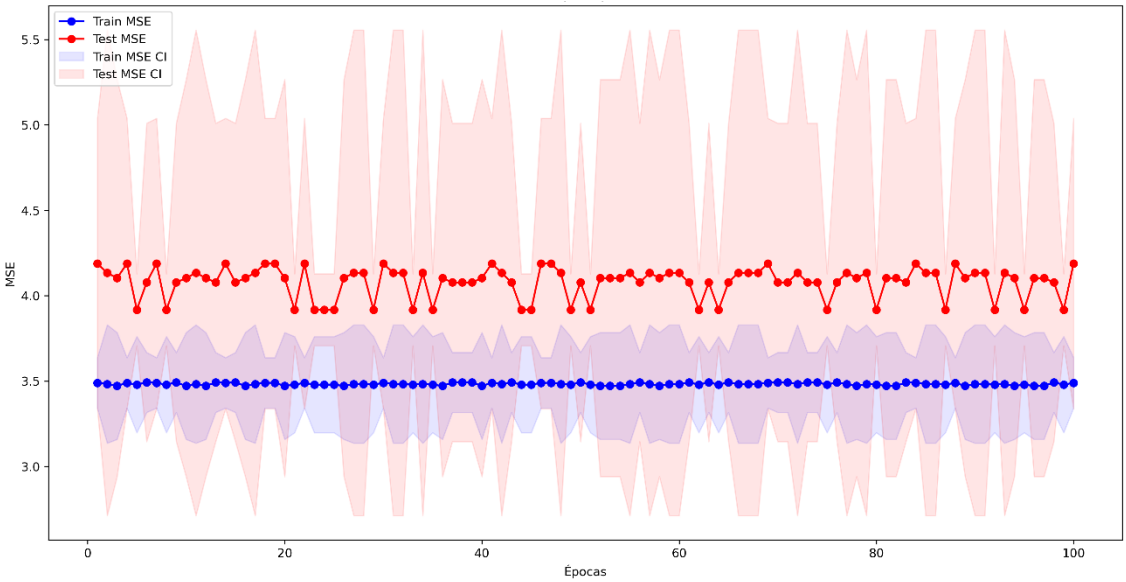
En cuanto al error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), los valores observados son 3,48 y 1,86 respectivamente para el entrenamiento, y 3,92 y 1,98 para el conjunto de prueba. Estos valores también indican que el modelo presenta un desempeño razonable, con un error moderado en las predicciones. Aunque el RMSE es un poco mayor en el conjunto de prueba, este valor sigue siendo relativamente bajo considerando que las notas se encuentran en una escala de 0 a 20,

El error porcentual absoluto medio (MAPE) muestra un valor del 13,8 % para el conjunto de entrenamiento y 14,5 % para el conjunto de prueba. Esto indica que, en términos relativos, el modelo predice las notas con un margen de error porcentual un poco alto, siendo un poco mayor en el conjunto de prueba.

Los intervalos de confianza al 95 % de la Tabla 7 muestran un comportamiento estable para las métricas de regresión ridge en el conjunto de entrenamiento, con intervalos estrechos, lo que indica una alta consistencia en el desempeño del modelo. Sin embargo, en el conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor variabilidad en el rendimiento del modelo al predecir datos nuevos.

Figura 59

Gráfico de MSE de épocas de la regresión ridge



Nota. Elaboración propia.

En la Figura 59 se puede apreciar el comportamiento del error cuadrático medio (MSE) a lo largo de las épocas durante el entrenamiento de la regresión ridge, donde el MSE para el conjunto de entrenamiento (representado por los puntos azules) permanece relativamente constante, fluctuando alrededor de un valor bajo. Esto indica que el modelo está aprendiendo de manera estable y no presenta grandes variaciones en su rendimiento durante las épocas. Por otro lado, el MSE para el conjunto de prueba (representado por los puntos rojos) muestra una variabilidad significativa. Las fluctuaciones grandes en el MSE para el conjunto de prueba indican que el modelo tiene un rendimiento menos estable cuando se evalúa con datos no vistos. Esto podría sugerir que el modelo está sobreajustado a los datos de entrenamiento, es decir, no generaliza bien para nuevos datos.

Las sombras alrededor de las líneas de MSE indican los intervalos de confianza al 95 % para las métricas. En el conjunto de entrenamiento, los intervalos son más estrechos, lo que indica una mayor certeza sobre las predicciones del modelo. En el conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor incertidumbre y variabilidad en el rendimiento del modelo en los datos no vistos.

4.1.3.2. Regresión lasso

Tabla 8

Lista de los 5 mejores folds de las métricas de desempeño de la regresión lasso

#		MAE	MSE	RMSE	MAPE
10 Fold	Train	1,458481	3,581764	1,892064	0,142268
	Test	1,487168	3,999087	1,999263	0,146553
11 Fold	Train	1,458481	3,581764	1,892064	0,142268
	Test	1,487168	3,999087	1,999263	0,146553
16 Fold	Train	1,458481	3,581764	1,892064	0,142268
	Test	1,487168	3,999087	1,999263	0,146553
19 Fold	Train	1,458481	3,581764	1,892064	0,142268
	Test	1,487168	3,999087	1,999263	0,146553
34 Fold	Train	1,458481	3,581764	1,892064	0,142268
	Test	1,487168	3,999087	1,999263	0,146553

Nota. Elaboración propia.

Tabla 9

Intervalos de confianza al 95 % de la regresión lasso

#		MAE	MSE	RMSE	MAPE
Train	Intervalo1	1,448119	3,608100	1,898602	0,141567
	Intervalo2	1,450348	3,616608	1,900838	0,141758
Test	Intervalo1	1,539706	4,157641	2,024487	0,151226
	Intervalo2	1,552167	4,191772	2,030876	0,152297

Nota. Elaboración propia.

Las métricas de desempeño proporcionadas en la Tabla 8 reflejan cómo el modelo de lasso ajusta las predicciones en relación con los valores reales de las notas obtenidas. En cuanto al error absoluto medio (MAE), se observa que el modelo tiene un desempeño consistente, con un valor de 1,46 para el conjunto de entrenamiento y 1,49 para el conjunto de prueba. Esto indica que, en promedio, la predicción de la nota está a menos de 2 puntos de la nota real. Este error es aceptable, considerando que las notas se encuentran en una escala de 0 a 20,

En cuanto al error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), los valores observados son 3,58 y 1,89 respectivamente para el entrenamiento, y 3,99 y 1,99 para el conjunto de prueba. Este comportamiento muestra que el modelo penaliza de manera razonable los errores más grandes, con una ligera diferencia entre los conjuntos de entrenamiento y prueba.

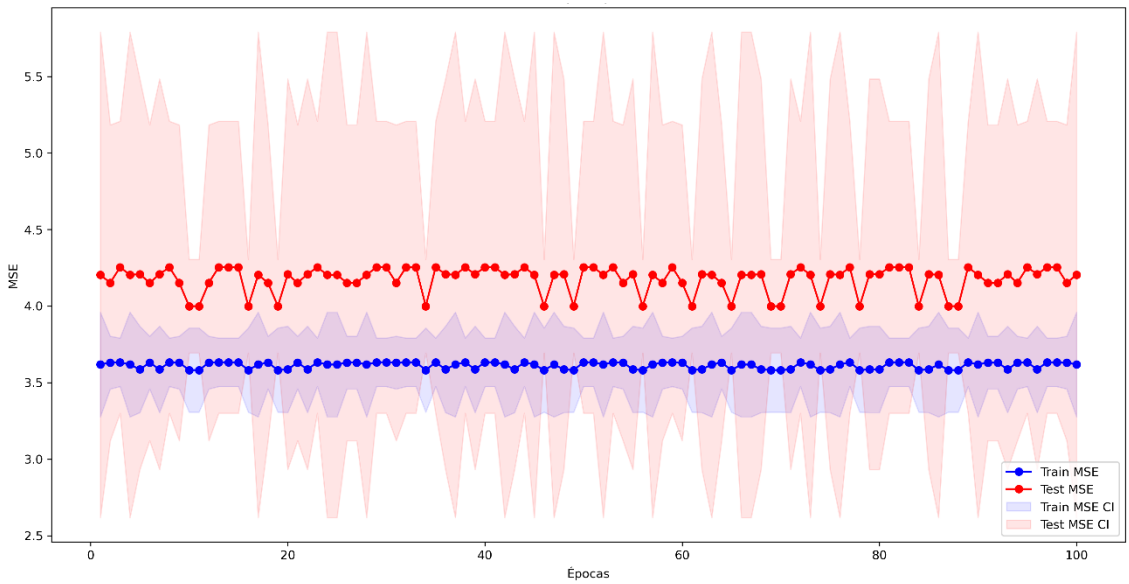
El error porcentual absoluto medio (MAPE) muestra un valor del 14,2 % para el conjunto de entrenamiento y 14,6 % para el conjunto de prueba. Esto indica que, en términos relativos, el modelo predice las notas con un margen de error porcentual un poco alto**,** siendo un poco mayor en el conjunto de prueba**.** Este margen de error sugiere que las predicciones pueden estar desviándose en un 14 % respecto a las notas reales, lo cual es un error considerable cuando se trata de predecir el rendimiento académico.

Los intervalos de confianza al 95 % de la Tabla 9 muestran un comportamiento estable en el conjunto de entrenamiento, con intervalos estrechos para las métricas de desempeño, lo que indica una alta consistencia en el desempeño del modelo durante el

entrenamiento. Sin embargo, en el conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor variabilidad en el rendimiento del modelo al predecir datos nuevos.

Figura 60

Gráfico de MSE de épocas de la regresión lasso



Nota. Elaboración propia.

En la figura 60 se puede apreciar el comportamiento del error cuadrático medio (MSE) a lo largo de las épocas durante el entrenamiento de la regresión lasso, donde el MSE para el conjunto de entrenamiento (representado por los puntos azules) permanece relativamente constante, fluctuando alrededor de un valor bajo. Esto indica que el modelo está aprendiendo de manera estable y no presenta grandes variaciones en su rendimiento durante las épocas. Por otro lado, el MSE para el conjunto de prueba (representado por los puntos rojos) muestra una variabilidad significativa. Las fluctuaciones grandes en el MSE para el conjunto de prueba indican que el modelo tiene un rendimiento menos estable cuando se evalúa con datos no vistos. Esto podría sugerir que el modelo está sobreajustado a los datos de entrenamiento, es decir, no generaliza bien para nuevos datos.

Las sombras alrededor de las líneas de MSE indican los intervalos de confianza al 95 % para las métricas. En el conjunto de entrenamiento, los intervalos son más estrechos, lo que indica una mayor certeza sobre las predicciones del modelo. En el

conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor incertidumbre y variabilidad en el rendimiento del modelo en los datos no vistos.

4.1.3.3. Árbol de decisión

Tabla 10

Lista de los 5 mejores folds de las métricas de desempeño de árbol de decisión

#		MAE	MSE	RMSE	MAPE
3 Fold	Train	1,267244	2,651709	1,627740	0,123052
	Test	1,571148	3,909093	1,966379	0,155078
11 Fold	Train	1,267244	2,651709	1,627740	0,123052
	Test	1,571148	3,909093	1,966379	0,155078
16 Fold	Train	1,267244	2,651709	1,627740	0,123052
	Test	1,571148	3,909093	1,966379	0,155078
22 Fold	Train	1,267244	2,651709	1,627740	0,123052
	Test	1,571148	3,909093	1,966379	0,155078
25 Fold	Train	1,267244	2,651709	1,627740	0,123052
	Test	1,571148	3,909093	1,966379	0,155078

Nota. Elaboración propia.

Tabla 11

Intervalos de confianza al 95 % de árbol de decisión

#		MAE	MSE	RMSE	MAPE
Train	Intervalo1	1,283265	2,734729	1,651515	0,124824
	Intervalo2	1,292416	2,776025	1,663656	0,125751
Test	Intervalo1	1,642596	4,386369	2,078043	0,160665
	Intervalo2	1,668980	4,540891	2,115043	0,162720

Nota. Elaboración propia.

Las métricas de desempeño proporcionadas en la Tabla 10 reflejan cómo el modelo de árbol de decisión se ajusta las predicciones en relación con los valores reales de las notas obtenidas. En cuanto al error absoluto medio (MAE), se observa que el modelo tiene un error relativamente pequeño en el conjunto de entrenamiento, donde en promedio, las predicciones están a 1,27 puntos de las notas reales. Sin embargo, al

predecir sobre datos nuevos (el conjunto de prueba), el modelo aumenta su error a 1,57, lo que indica que las predicciones se desvían un poco más de las notas reales cuando se enfrentan a ejemplos no entrenados. Esto es típico en modelos de predicción que no logran generalizar completamente a nuevos datos.

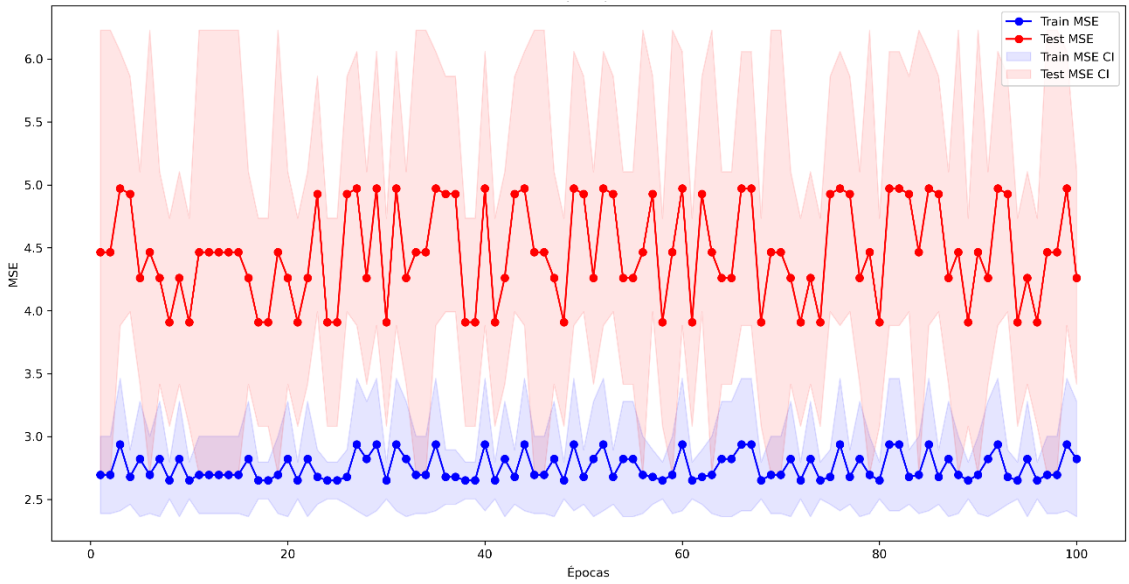
En cuanto al error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), los valores observados son 2,65 y 1,63 respectivamente para el entrenamiento, y 3,91 y 1,97 para el conjunto de prueba. Este aumento sugiere que, aunque el modelo tiene un buen desempeño al predecir en el conjunto de entrenamiento, en el conjunto de prueba comete errores mayores, lo que se refleja en el aumento del MSE y RMSE.

El error porcentual absoluto medio (MAPE) muestra un valor del 12,3 % para el conjunto de entrenamiento y 15,5 % para el conjunto de prueba. Esto sugiere que, en términos relativos, las predicciones del modelo en promedio se desvían un 12,3 % de las notas reales en el conjunto de entrenamiento y un 15,5 % en el conjunto de prueba. Este margen de error es bastante alto, especialmente considerando que las notas están en una escala de 0 a 20, lo que implica que el modelo no está proporcionando predicciones suficientemente precisas, particularmente en el conjunto de prueba.

Los intervalos de confianza al 95 % de la Tabla 11 muestran una diferencia significativa entre el rendimiento del modelo en el conjunto de entrenamiento y el de prueba. En el conjunto de entrenamiento, los intervalos son estrechos, con un MAE entre 1,28 y 1,29, lo que indica que el modelo presenta un desempeño constante y confiable al predecir los datos con los que fue entrenado. Sin embargo, en el conjunto de prueba, los intervalos son más amplios, con un MAE entre 1,64 y 1,67, lo que sugiere una mayor variabilidad en las predicciones del modelo cuando se enfrenta a nuevos datos.

Figura 61

Gráfico de MSE de épocas de árbol de decisión



Nota. Elaboración propia.

En la figura 61 se puede apreciar el comportamiento del error cuadrático medio (MSE) a lo largo de las épocas durante el entrenamiento del árbol de decisión, los puntos azules representan el MSE para el conjunto de entrenamiento la cual permanece relativamente constante, pero con pequeñas fluctuaciones. Aunque el valor se mantiene en un rango bajo durante todas las épocas, las pequeñas oscilaciones sugieren que el modelo está aprendiendo de manera estable, pero con una ligera variabilidad en su rendimiento a lo largo del tiempo. Por otro lado, el MSE para el conjunto de prueba (representado por los puntos rojos) muestra una variabilidad significativa. Las fluctuaciones grandes en el MSE para el conjunto de prueba indican que el modelo tiene un rendimiento menos estable cuando se evalúa con datos no vistos.

Las sombras alrededor de las líneas de MSE indican los intervalos de confianza al 95 % para las métricas. En el conjunto de entrenamiento, los intervalos son más estrechos, lo que indica una mayor certeza sobre las predicciones del modelo. En el conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor incertidumbre y variabilidad en el rendimiento del modelo en los datos no vistos.

4.1.3.4. Bosques aleatorios

Tabla 12

Lista de los 5 mejores folds de las métricas de desempeño de bosques aleatorios

#		MAE	MSE	RMSE	MAPE
1 Fold	Train	0,930702	1,417792	1,190265	0,089473
	Test	1,525464	3,994598	1,977846	0,149010
5 Fold	Train	0,930702	1,417792	1,190265	0,089473
	Test	1,525464	3,994598	1,977846	0,149010
12 Fold	Train	0,930702	1,417792	1,190265	0,089473
	Test	1,525464	3,994598	1,977846	0,149010
21 Fold	Train	0,930702	1,417792	1,190265	0,089473
	Test	1,525464	3,994598	1,977846	0,149010
37 Fold	Train	0,930702	1,417792	1,190265	0,089473
	Test	1,525464	3,994598	1,977846	0,149010

Nota. Elaboración propia.

Tabla 13

Intervalos de confianza al 95 % de bosques aleatorios

#		MAE	MSE	RMSE	MAPE
Train	Intervalo1	0,916218	1,356821	1,163433	0,088150
	Intervalo2	0,920509	1,372332	1,170135	0,088546
Test	Intervalo1	1,541875	4,100156	2,010394	0,150263
	Intervalo2	1,553136	4,143180	2,019846	0,151304

Nota. Elaboración propia.

Las métricas de desempeño proporcionadas en la Tabla 12 reflejan cómo el modelo de bosques aleatorios se ajusta a las predicciones en relación con los valores reales de las notas obtenidas. En cuanto al error absoluto medio (MAE), se observa que, en el conjunto de entrenamiento, el MAE es 0,93, y en el conjunto de prueba es 1,53. Esto sugiere que el modelo tiene un rendimiento muy preciso en el conjunto de entrenamiento, con un error relativamente pequeño (menos de 1 punto). Sin embargo, al predecir sobre

datos nuevos, el error aumenta considerablemente a 1,53, lo que indica que las predicciones en el conjunto de prueba están más alejadas de los valores reales.

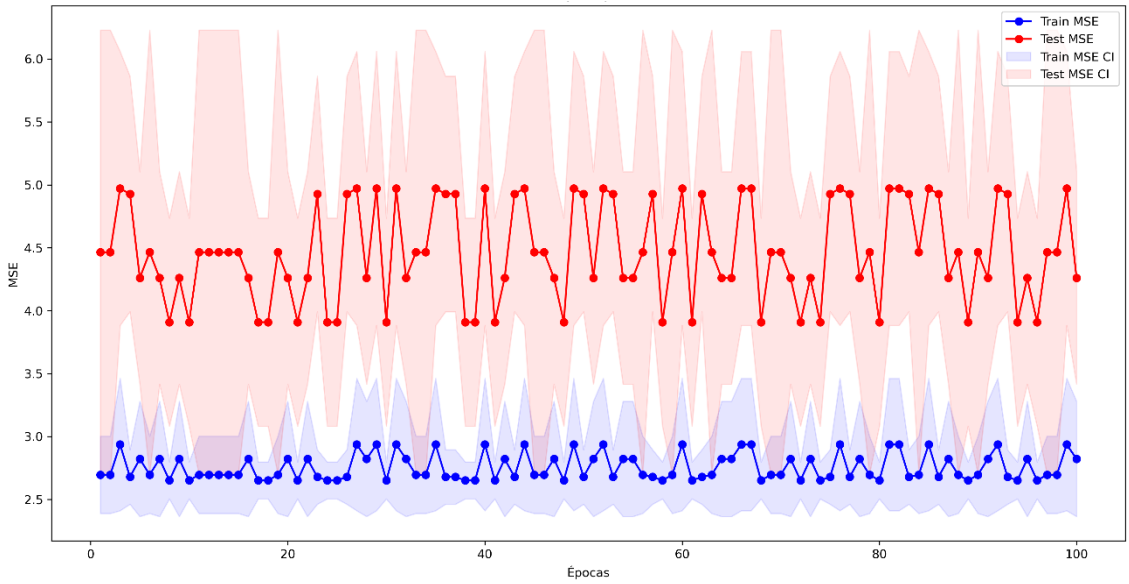
En cuanto al error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), los valores observados son en el conjunto de entrenamiento, el MSE es 1,42 y el RMSE es 1,19. Estos valores son relativamente bajos, lo que indica que el modelo tiene un buen ajuste en el conjunto de entrenamiento. No obstante, en el conjunto de prueba, el MSE aumenta a 3,99 y el RMSE a 1,98, lo que refleja una mayor discrepancia entre las predicciones y los valores reales en el conjunto de prueba. Esto sugiere que el modelo no generaliza tan bien a datos no vistos.

El error porcentual absoluto medio (MAPE) muestra un valor en el conjunto de entrenamiento es 8,95 %, y en el conjunto de prueba es 14,90 %. Este aumento en el MAPE del conjunto de prueba sugiere que, en términos relativos, el modelo tiene un margen de error porcentual considerablemente mayor al predecir datos no entrenados. Aunque el modelo es bastante preciso en el conjunto de entrenamiento, las predicciones sobre el conjunto de prueba presentan un mayor margen de error.

Los intervalos de confianza al 95 % de la Tabla 13 muestran una diferencia notable entre el rendimiento del modelo en el conjunto de entrenamiento y en el conjunto de prueba. En el conjunto de entrenamiento, los intervalos son estrechos, con un MAE entre 0,92 y 0,93, lo que indica un desempeño estable y confiable al predecir los datos de entrenamiento. Sin embargo, en el conjunto de prueba, los intervalos son más amplios, con un MAE entre 1,54 y 1,55, lo que refleja una mayor variabilidad en las predicciones del modelo al enfrentarse a datos nuevos.

Figura 62

Gráfico de MSE de épocas de bosques aleatorios



Nota. Elaboración propia.

En la figura 62 se puede apreciar el comportamiento del error cuadrático medio (MSE) a lo largo de las épocas durante el entrenamiento de los bosques aleatorios, los puntos azules representan el MSE para el conjunto de entrenamiento la cual permanece relativamente constante, pero con pequeñas fluctuaciones. Aunque el valor se mantiene en un rango bajo durante todas las épocas, las pequeñas oscilaciones sugieren que el modelo está aprendiendo de manera estable, pero con una ligera variabilidad en su rendimiento a lo largo del tiempo. Por otro lado, el MSE para el conjunto de prueba (representado por los puntos rojos) muestra una variabilidad significativa. Las fluctuaciones grandes en el MSE para el conjunto de prueba indican que el modelo tiene un rendimiento menos estable cuando se evalúa con datos no vistos.

Las sombras alrededor de las líneas de MSE indican los intervalos de confianza al 95 % para las métricas. En el conjunto de entrenamiento, los intervalos son más estrechos, lo que indica una mayor certeza sobre las predicciones del modelo. En el conjunto de prueba, los intervalos son más amplios, lo que refleja una mayor incertidumbre y variabilidad en el rendimiento del modelo en los datos no vistos.

4.1.3.5. Redes neuronales

Tabla 14

Lista de los 5 mejores folds de las métricas de desempeño de redes neuronales

#		MAE	MSE	RMSE	MAPE
100 Fold	Train	0,876611	2,132590	1,459377	0,086390
	Test	0,885267	2,157038	1,464414	0,087272
97 Fold	Train	0,886409	2,156469	1,467391	0,087259
	Test	0,893666	2,169761	1,468575	0,087958
95 Fold	Train	0,893367	2,167244	1,470975	0,088146
	Test	0,897801	2,176396	1,443031	0,088461
96 Fold	Train	0,900630	2,163698	1,469748	0,088937
	Test	0,913076	2,180031	1,468637	0,090079
99 Fold	Train	0,882569	2,137421	1,461054	0,087034
	Test	0,899493	2,181271	1,444568	0,088583

Nota. Elaboración propia.

Tabla 15

Intervalos de confianza al 95 % de redes neuronales

#		MAE	MSE	RMSE	MAPE
Train	Intervalo1	1,513142	7,019041	2,078800	0,140700
	Intervalo2	2,503714	17,914777	3,040873	0,223436
Test	Intervalo1	1,524108	7,050909	2,069016	0,141754
	Intervalo2	2,513716	17,949587	3,032387	0,224387

Nota. Elaboración propia.

Las métricas de desempeño proporcionadas en la Tabla 14 reflejan cómo el modelo de redes neuronales ajusta las predicciones en relación con los valores reales de las notas obtenidas. En cuanto al error absoluto medio (MAE), se observa que el modelo en el conjunto de entrenamiento, el MAE promedio es 0,88, con un ligero aumento en el conjunto de prueba, donde el MAE es de 0,89. Esto indica que el modelo tiene un buen desempeño, con un error muy pequeño en promedio, tanto en los datos de entrenamiento

como en los de prueba. Las predicciones del modelo están muy cerca de los valores reales, con una diferencia de menos de 1 punto.

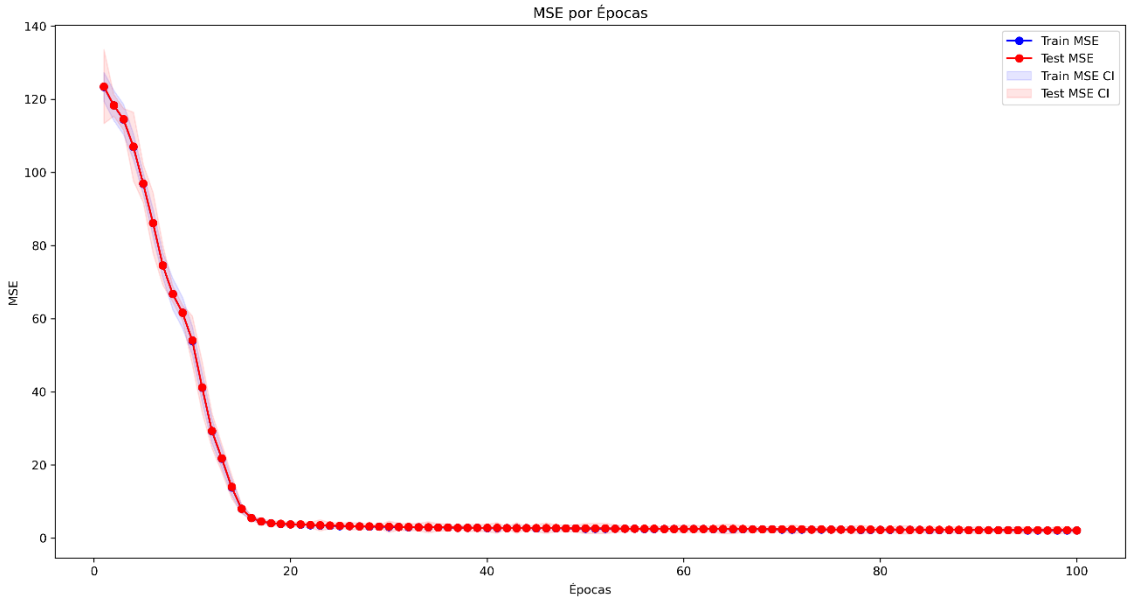
En cuanto al error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), los valores observados son 2,13 y 1,46 respectivamente para el entrenamiento, y 2,16 y 1,46 para el conjunto de prueba. Estos valores indican que el modelo tiene un buen ajuste y baja discrepancia entre las predicciones y los valores reales en ambos conjuntos de datos. Las diferencias entre los conjuntos de entrenamiento y prueba son mínimas, lo que sugiere que el modelo generaliza bien a datos no vistos.

El error porcentual absoluto medio (MAPE) muestra un valor en el conjunto de entrenamiento es 8,64 %, y en el conjunto de prueba es 8,73 %. Este es un margen de error porcentual bajo, lo que indica que el modelo realiza predicciones con una desviación porcentual pequeña respecto a las notas reales. Aunque el MAPE es ligeramente mayor en el conjunto de prueba, sigue siendo un valor aceptable.

Los intervalos de confianza al 95 % de la Tabla 15 muestran un MAE que varía entre 1,51 y 2,50, lo que sugiere que el modelo puede tener cierta variabilidad en su desempeño en algunos casos, con valores más altos que indican un rendimiento menos consistente. De manera similar, en el conjunto de prueba, el MAE oscila entre 1,52 y 2,51, lo que refleja una variabilidad comparable en las predicciones sobre datos no entrenados. Aunque el modelo muestra un desempeño generalmente bueno, estos intervalos amplios indican que algunas predicciones pueden ser menos confiables, especialmente las primeras que se realizaron donde mostraban valores malos, pero mientras pasaban las épocas los resultados fueron mejorando.

Tabla 16

Gráfico de MSE de épocas de redes neuronales



Nota. Elaboración propia.

En la figura 63 se puede apreciar el comportamiento del error cuadrático medio (MSE) a lo largo de las épocas durante el entrenamiento de redes neuronales, al principio, el MSE para el conjunto de entrenamiento y el MSE para el conjunto de prueba son altos, pero rápidamente el MSE disminuye a medida que avanzan las épocas, alcanzando valores muy bajos al final del proceso. Esta rápida disminución indica que el modelo está aprendiendo eficazmente, ajustando las predicciones a los datos de entrenamiento y prueba, lo que demuestra una buena capacidad de ajuste inicial. La estrecha proximidad entre las curvas de entrenamiento y prueba sugiere que el modelo tiene un buen ajuste a los datos tanto en el conjunto de entrenamiento como en el de prueba, sin mostrar signos de sobreajuste (overfitting). Esto es un indicativo de que el modelo está generalizando bien y no se está quedando únicamente con las características específicas de los datos de entrenamiento.

Los intervalos de confianza (sombras alrededor de las líneas de MSE) son bastante pequeños, lo que indica que el modelo tiene una alta certeza en sus predicciones para ambos conjuntos de datos, tanto de entrenamiento como de prueba. La pequeña dispersión en los intervalos refleja un buen nivel de estabilidad en las predicciones.

4.2. Contrastación de hipótesis

4.2.1. Análisis estadístico hipótesis general

Si es posible comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Para poder comparar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023, se tomó 20 muestras aleatorias de las 100 épocas que se entrenó a los modelos de machine learning. Las métricas que se analizaron fueron: error absoluto medio (MAE), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE); las cuales se pudo observar que los mejores resultados fueron del modelo de redes neuronales; como también el modelo de bosques aleatorios presento resultados aceptables, para poder determinar si existe una diferencia significativa entre estos 2 modelos, se plantea la siguiente hipótesis de investigación:

Primeramente, se muestra los datos aleatorios de los modelos de machine learning en la Tabla 15, seguidamente se realizó la prueba de normalidad, la cual se utilizó Shapiro-Wilk ya que se utiliza una muestra menor a 50 ($n < 50$). Las pruebas se realizaron en el programa SPSS con un nivel de confiabilidad de 95 %.

Tabla 17

Lista de los modelos de machine learning analizando el MAPE para datos de prueba

Regresión Ridge	Regresión Lasso	Árbol de decisión	Bosques aleatorios	Redes neuronales
0,149460	0,154223	0,169440	0,152073	0,093953
0,151661	0,152129	0,155078	0,154181	0,092130
0,151661	0,146553	0,155078	0,152083	0,111428
0,149460	0,154223	0,155078	0,149010	0,092868
0,149460	0,150381	0,159737	0,154181	0,090007
0,151312	0,150381	0,169440	0,152083	0,141565
0,148511	0,153583	0,159737	0,149010	0,107893
0,148511	0,154223	0,155078	0,152073	0,092728
0,148511	0,146553	0,155078	0,146502	0,095759
0,151661	0,153583	0,159755	0,149010	0,098777
0,151661	0,152129	0,155078	0,149010	0,089304
0,149460	0,154223	0,167930	0,152083	0,106905
0,151661	0,153583	0,167930	0,149010	0,100904
0,149460	0,154223	0,167930	0,152083	0,101484
0,149460	0,153583	0,155078	0,152073	0,102937
0,145207	0,154223	0,159737	0,152073	0,113047
0,151661	0,154223	0,167930	0,152083	0,114437
0,145207	0,153583	0,155078	0,149010	0,105112
0,151661	0,152129	0,167930	0,154181	0,093714
0,151661	0,146553	0,155078	0,152073	0,491241

Nota. Elaboración propia.

Aplicación de la prueba de normalidad

H0: La muestra sigue una distribución normal.

H1: La muestra no sigue una distribución normal.

Tabla 18

Prueba de normalidad

	Estadístico	gl	Sig.
Redes neuronales	0,343	20	0,00000002
Bosques aleatorios	0,848	20	0,00498879

Nota. Elaboración propia.

Como los valores son menores que 0,05 entonces se rechaza la hipótesis nula y afirmamos que la muestra no sigue una distribución normal. Entonces se usa la prueba no paramétrica **wilcoxon**.

Aplicación prueba de hipótesis

H0: No existen diferencias significativas entre los modelos de redes neuronales y bosques aleatorios respecto al error porcentual absoluto medio para la predicción del rendimiento académico de los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

H1: Si existen diferencias significativas entre los modelos de redes neuronales y bosques aleatorios respecto al error porcentual absoluto medio para la predicción del rendimiento académico de los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Tabla 19

Estadístico de la prueba de wilcoxon

Z	-3,173[b]
Sig. asin. (bilateral)	0,0015073

Nota. Elaboración propia.

Decisión estadística

En base a los resultados se observa un p-valor bilateral de 0,0015 < 0,05, por la cual se rechaza la hipótesis nula y se afirma que: si existen diferencias significativas entre los modelos de redes neuronales y bosques aleatorios respecto al error porcentual absoluto

medio para la predicción del rendimiento académico de los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Respecto a la hipótesis general, se comparó los 2 algoritmos que tuvieron más precisión respecto a predecir el rendimiento académico universitario, las cuales fueron redes neuronales y bosques aleatorios; luego de compararlos estadísticamente se observó que el modelo de redes neuronales sobresalió por su mayor precisión y capacidad predictiva, demostrando un desempeño superior frente a otros algoritmos empleados, fue el mejor algoritmo que pudo predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023; con un nivel de confianza al 95 %.

4.2.2. Análisis estadístico hipótesis específica 1

Si es posible preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Para la construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023, se realizó un proceso exhaustivo de limpieza y preparación de datos. A continuación, se detalla en 3 puntos clave:

- **Eliminación de datos faltantes**

Se identificaron y se eliminaron los registros de estudiantes que tenían datos faltantes, ya que la ausencia de información relevante podría afectar la precisión de los modelos.

- **Ajuste en las calificaciones de los estudiantes que adelantaron cursos**

Se asignó una calificación de cero a las materias que deberían haber cursado en su ciclo, pero que no lo hicieron; además se eliminaron del análisis aquellos cursos que los estudiantes habían adelantado y que no correspondían al ciclo en el que debían estar matriculados.

- **Exclusión de estudiantes con notas atípicas**

Se excluyeron del análisis a los estudiantes con resultados extremadamente bajos, como aquellos con promedios de 0 o aquellos que presentaban una diferencia notable entre los promedios de los años 2023 y 2024 (por ejemplo, pasar de un promedio de 12 a un 3), ya que se consideró que estos casos podrían haber sido resultado de problemas personales o académicos.

Finalmente, si se pudo preparar los datos para para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023, ya que en un principio se tenía las siguientes cantidades de ingresantes: ESMI (51), ESME (68), ESMC (50), ESIS (87) y ESIQ (55); teniendo un total de 311 alumnos ingresantes a la Universidad Nacional Jorge Basadre Grohmann, en la Facultad de Ingeniería por las diferentes modalidades. Luego de realizar la limpieza se datos se tiene lo siguiente: ESMI (26), ESME (40), ESMC (22), ESIS (44) y ESIQ (11); teniendo un total de 143 alumnos ingresantes a la Universidad Nacional Jorge Basadre Grohmann, en la Facultad de Ingeniería por las diferentes modalidades; disminuyendo en un 54,02 % de la data inicial.

4.2.3. Análisis estadístico hipótesis específica 2

Si es posible construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Si fue posible construir los modelos de machine learning, el primer modelo fue el de regresión lineal, la cual se utilizó los métodos de ridge con un α =2,559548, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico; el otro método fue lasso con un α =0,040949, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico.

El segundo modelo fue el de árbol de decisión con una profundidad de 4, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Este curso tiene una fuerte relación con el éxito académico de los estudiantes, lo que significa que su rendimiento en esta área es un predictor clave de cómo les irá en la universidad.

El tercer modelo fue de bosques aleatorios la cual la mejor profundidad fue de 5, lo que indica que los árboles dentro del bosque aprenden lo suficiente de los cursos del examen de admisión sin caer en patrones demasiado específicos. La importancia de las características del algoritmo de bosques aleatorios, la cual muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico.

El cuarto y último modelo fue de redes neuronales, la cual se encontró que el mejor número de neuronas en la capa oculta fue de 10, Este modelo muestra que, entre los cursos de admisión evaluados en el examen de ingreso, razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Lo que llama la atención es que la red neuronal revela que, si un estudiante pertenece a la sección A, es la más influyente en la predicción del rendimiento académico universitario. Su peso supera considerablemente al de las demás variables, incluidas las notas de cursos del examen de admisión, la cual se evidencia que el hecho de haber estudiado en la sección A tiene una relación fuerte con el desempeño académico posterior.

4.2.4. Análisis estadístico hipótesis específica 3

Si es posible evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Si fue posible evaluar los modelos de machine learning en base a las 4 métricas (error absoluto medio, error cuadrático medio, raíz del error cuadrático medio, error porcentual absoluto medio), el primer modelo fue el de regresión lineal, la cual se utilizó los métodos de ridge; teniendo en el MAE un valor cercano a 1,42 para el conjunto de

entrenamiento y 1,48 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 3,48 y 1,86 respectivamente para el entrenamiento, y 3,92 y 1,98 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 13,8 % para el conjunto de entrenamiento y 14,5 % para el conjunto de prueba. El otro método fue lasso; teniendo en el MAE un valor cercano a 1,46 para el conjunto de entrenamiento y 1,49 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 3,58 y 1,89 respectivamente para el entrenamiento, y 3,99 y 1,99 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 14,2 % para el conjunto de entrenamiento y 14,6 % para el conjunto de prueba.

El segundo modelo fue el de árbol de decisión**,** teniendo en el MAE un valor cercano a 1,27 para el conjunto de entrenamiento y 1,57 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 2,65 y 1,63 respectivamente para el entrenamiento, y 3,91 y 1,97 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 12,3 % para el conjunto de entrenamiento y 15,5 % para el conjunto de prueba.

El tercer modelo fue de bosques aleatorios**,** teniendo en el MAE un valor cercano a 0,93 para el conjunto de entrenamiento y 1,53 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 1,42 y 1,19 respectivamente para el entrenamiento, y 3,99 y 1,98 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 8,95 % para el conjunto de entrenamiento y 14,90 % para el conjunto de prueba.

El cuarto y último modelo fue de redes neuronales, teniendo en el MAE un valor cercano a 0,88 para el conjunto de entrenamiento y 0,89 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 2,13 y 1,46 respectivamente para el entrenamiento, y 2,16 y 1,46 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 8,64 % para el conjunto de entrenamiento y 8,73 % para el conjunto de prueba.

DISCUSIONES

En la presente investigación se comparó los 2 algoritmos que tuvieron más precisión respecto a predecir el rendimiento académico universitario, las cuales fueron redes neuronales y bosques aleatorios; luego de compararlos estadísticamente se observó que el modelo de redes neuronales sobresalió por su mayor precisión y capacidad predictiva, demostrando un desempeño superior frente a otros algoritmos empleados, fue el mejor algoritmo que pudo predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023; con un nivel de confianza al 95 %.

En este sentido, se tuvo discrepancia con los resultados de Assiri et al. (2024), ya que en su investigación menciona que el mejor modelo fue k-nearest neighbors con una precisión del 100 %, seguido por decision tree donde se obtuvo una precisión del 81 % y el modelo support vector machine con una precisión del 75 %. En la investigación de Contreras et al. (2020) se seleccionaron los algoritmos de SVM y perceptrón las cuales obtuvieron los mejores resultados en cuanto a métricas de evaluación, determinando que el modelo de perceptrón muestra ser exitoso para la determinación del rendimiento académico con una exactitud del 66,4 %.

Por otro lado, en la investigación de Mengash (2020) mencionan que, el modelo de ANN logró alcanzar el mejor rendimiento alcanzando un 79,22 % de presión. Como también en la investigación de Ahmed y Al-Omari (2024) menciona que, support vector machine (SVM) logró la mayor precisión con un 96,0 % en comparación con otros algoritmos de machine learning como: decision tres, naïve bayes y k-nearest neighbors. Como también en la investigación de Salas et al. (2024) menciona que, los modelos basados en árboles de conjuntos, específicamente random forest y extreme gradient boosting, son altamente efectivos para predecir el rendimiento académico de los estudiantes de pregrado. Entre estos, el modelo random forest supera ligeramente al modelo extreme gradient boosting. En la investigación de Candia (2019), se menciona que el algoritmo random forest tuvo mejor desempeño en la predicción del rendimiento académico de los ingresantes durante los primeros semestres en la UNSAAC alcanzó una precisión del 69 %. En este estudio de caso, el segundo mejor rendimiento lo obtuvo el

algoritmo de regresión logística, con un 68 % de precisión. En la investigación de Yupanqui (2018) menciona que, la red neuronal de topología 8:3:4:1 la cual fue óptima para la predicción del rendimiento académico.

Se prepararon los datos para para construir modelos de machine learning para la predicción del rendimiento académico universitario, ya que en un principio se tenía un total de 311 alumnos ingresantes a la Universidad Nacional Jorge Basadre Grohmann, en la Facultad de Ingeniería por las diferentes modalidades. Luego de realizar la limpieza se tiene un total de 143 alumnos ingresantes a la Universidad Nacional Jorge Basadre Grohmann, en la Facultad de Ingeniería por las diferentes modalidades; disminuyendo en un 54,02 % de la data inicial.

En este sentido, se tuvo coincidencia con los resultados de Assiri et al. (2024) la cual hace una investigación realiza una limpieza de datos una vez recolectada la información, para corregir cualquier valor faltante y eliminar datos con ruido. En la investigación de Contreras et al. (2020) se tuvo una data inicial de 1620 registros, la cual se depuró los registros de datos erróneos de la base de datos, quedando un conjunto de datos con 1571 registros, realizando un proceso de limpieza de datos. Por otro lado, en la investigación de Mengash (2020), se utilizó un conjunto de datos de 2039 estudiantes matriculados en una Facultad de Ciencias de la Computación e Información de una universidad pública saudí de 2016 a 2019, la cual se depuró los registros de datos eliminando atributos irrelevantes, eliminación de registros con valores faltantes y la eliminación de duplicados. Luego de haber realizado la limpieza de datos, el total de registros estudiantiles restantes fue de 1430 registros para 828 estudiantes del curso académico 2016-2017 y 602 estudiantes del curso académico 2017-2018.

Como también en la investigación de Ahmed y Al-Omari (2024) menciona que, el conjunto de datos usada en dicha investigación fue recopilado de la Universidad de Wollo y el Instituto Tecnológico Kombolcha, incluyendo información de estudiantes de los años académicos 2017-2022, con un total de 32,582 registros. Sin embargo, debido a la presencia de valores faltantes y registros duplicados, se realizó un proceso de limpieza de datos utilizando python. Como resultado, el conjunto de datos limpio quedó conformado por 32,005 registros. En la investigación de Candia (2019), se menciona que

los datos fueron un total de 12,698 alumnos ingresantes a la UNSAAC por las diferentes modalidades, de los cuales se eliminaron 89 registros de la base de datos original debido a inconsistencias en su información. Estas irregularidades incluían, por ejemplo, cantidades inusuales de créditos matriculados (hasta 46), datos incoherentes relacionados con el género, errores en los nombres de las escuelas profesionales, fechas de nacimiento incorrectas, procedencia dudosa y problemas en las encuestas. Para llevar a cabo esta depuración se utilizaron las herramientas WEKA y microsoft excel 2016. El autor menciona que luego de haber hecho esta limpieza, se ha trabajado con 12609 registros, los cuales serán utilizados para la predicción del rendimiento académico.

Así mismo la investigación de Saire (2023) menciona que se trabajó en un principio con una data de 778 estudiantes de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín de Arequipa, los cuales fueron recolectados del año 2011 al 2020 en la oficina de admisión y la Escuela de Sistemas, luego el autor realiza una exploración de datos llegando a realizar posteriormente una limpieza de datos eliminando registros de estudiantes que han ingresado a la universidad, sin embargo, nunca se matricularon en ningún curso, el cual es un total de 4 estudiantes. En la investigación de Yamao (2018), se menciona que los datos fueron un total de 1304 ingresantes de los periodos 2010-I a 2015-II a la carrera de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres, de los cuales se eliminaron algunos alumnos que sí estaban matriculados, pero aparentemente nunca asistieron a clases, ya que en todos los cursos figuraba la nota cero. En la investigación de Yupanqui (2018), se excluyó de la investigación a los estudiantes cuya nota promedio en su primer semestre fue mayor a 7, aplicando este filtro se trabajó con 69 registros de ingresantes a la Escuela Profesional de Ingeniería en Informática y Sistemas de la Universidad Nacional Jorge Basadre Grohmann.

En la siguiente investigación, se construyó los modelos de machine learning con los mejores hiperparámetros gracias al método de grid search, el primer modelo fue el de regresión lineal, la cual se utilizó los métodos de ridge con un α =2,559548, el otro método fue lasso con un α =0,040949**;** el segundo modelo fue el de árbol de decisión con una profundidad de 4, el tercer modelo fue de bosques aleatorios la cual la mejor profundidad

fue de 5, el cuarto y último modelo fue de redes neuronales, la cual se encontró que el mejor número de neuronas en la capa oculta fue de 10, Respecto a las características más importantes, los 4 modelos indican que razonamiento verbal se destaca como el factor más influyente en la predicción del desempeño académico. Cabe recalcar en redes neuronales revela que, si un estudiante pertenece a la sección A, es la más influyente en la predicción del rendimiento académico universitario. Su peso supera considerablemente al de las demás variables, incluidas las notas de cursos del examen de admisión, la cual se evidencia que el hecho de haber estudiado en la sección A tiene una relación fuerte con el desempeño académico posterior.

En este sentido, se tuvo coincidencia con los resultados de Assiri et al. (2024) la cual hace una investigación usa el mejor hiperparámetro para su modelo de k-nearest neighbors con un valor de k=1, sin embargo, al aumentar k, la precisión disminuyó; por ejemplo, con k=3, la precisión del modelo bajó. El autor hace mención también que los algoritmos de machine learning utilizados como k-nearest neighbor otorga mayor importancia a los cursos de admisión de: inglés, química y biología**.** En la investigación de Contreras et al. (2020) se seleccionaron y construyeron diversos algoritmos de aprendizaje automático para tareas de clasificación, los cuales son: SVC obteniendo mejores resultados con el hiperparámetro max_iter = 6, perceptrón con un hiperparámetro max_iter = 2, KNN la cual se observa que el mejor hiperparámetro a evaluar es n_neighbors = 10, 12 y 15, y finalmente árbol de decisión con un hiperparámetro de max_depth =7 (profundidad del árbol). Luego de haber construido los algoritmos de machine learning Contreras et al. (2020) menciona que, las variables que más influyen en el rendimiento académico de los estudiantes de ingeniería a partir de los factores analizados son: edad, genero, puntaje ICFES para aptitud matemática, puntaje global ICFES, valor de matrícula, puntaje ICFES para condición matemática y cohorte.

Por otro lado, en la investigación de Mengash (2020), se desarrolló cuatro modelos de predicción aplicando cuatro técnicas de clasificación de minería de datos reconocidas: redes neuronales artificiales (RNA), árboles de decisión, máquinas de vectores de soporte (SVM) y bayesiano naive; las cuales estos modelos se ejecutaron utilizando los valores de los parámetros definidos por defecto en el software WEKA;

finalmente el autor concluye que la puntuación **SAAT** (biología, química, física, matemáticas e inglés) es el criterio que predice con mayor precisión el rendimiento académico, por lo que se le debe dar más importancia. Como también en la investigación de Ahmed y Al (2024) menciona que, se construyeron modelos de predicción/clasificación con los mejores hiperparámetros encontrados, las cuales son: decision trees obteniendo mejores resultados con el hiperparámetro max_depth = 7 y min_split = 2, naïve bayes obteniendo mejores resultados con el hiperparámetro smoothing_parameter = 1, k-nearest neighbors obteniendo mejores resultados con el hiperparámetro number_neighbors = 10 y weight_function = "distance", y finalmente support vector machine (SVM) obteniendo mejores resultados con el hiperparámetro regularization_parameter_(C) = 10 y gamma = 0,01. Como también en la investigación de Salas et al. (2024) menciona que, se construyeron modelos de clasificación con los mejores hiperparámetros encontrados, las cuales son: logistic regression no fue necesario encontrar el mejor hiperparámetro, ridge obteniendo mejores resultados con el hiperparámetro regularization_strength = 100, lasso obteniendo mejores resultados con el hiperparámetro regularization_strength = 100, random forest obteniendo mejores resultados con el hiperparámetro number_trees = 500 y max_depth_trees = 10, y finalmente gradient boosting trees obteniendo mejores resultados con el hiperparámetro number_trees = 250 y max_depth_trees = 1. Así mismo la investigación de (Saire, 2023) menciona que, se construyeron modelos de machine learning con los mejores hiperparámetros encontrados, las cuales son: random forest obteniendo mejores resultados con el hiperparámetro max_depth = 14,34, max_features = 0,5692, max_samples = 0,8324, n_estimators = 70,62; xgboost obteniendo mejores resultados con el hiperparámetro eta = 0,4933, gamma = 2,014, max_depth = 11,91, n_estimators = 168,1, reg_alpha = 0,3947, reg_lambda = 0,4499. Finalmente, el autor concluye que las variables más influyentes son: sexo, puntaje, edad de egreso, tiempo transcurrido y edad de ingreso.

En la investigación de Yupanqui (2018) menciona que, la red que tubos mejores resultados fue la topología 8:3:4:1, la cual se tuvo 8 neuronas en la capa de entrada, 3 neuronas en la primera capa oculta, 4 neuronas en la segunda capa oculta y 1 neurona en la capa de salida. El autor menciona que las áreas de razonamiento matemático, aritmética

y algebra y razonamiento verbal influyen positivamente en un 0,59; 0,13 y 0,09 respectivamente al rendimiento académico en los alumnos del primer semestre de la Escuela Profesional de Ingeniería en Informática y Sistemas.

En la siguiente investigación se han evaluado 4 los modelos de machine learning las cuales son: regresión lineal, árbol de decisión, bosques aleatorios y redes neuronales, las cuales fueron 2 algoritmos los que mostraron mejores resultados, el primero es bosques aleatorios teniendo en el MAE un valor cercano a 0,93 para el conjunto de entrenamiento y 1,53 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 1,42 y 1,19 respectivamente para el entrenamiento, y 3,99 y 1,98 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 8,95 % para el conjunto de entrenamiento y 14,90 % para el conjunto de prueba. El segundo modelo que mostró buenos resultados fue de redes neuronales con un MAE un valor cercano a 0,88 para el conjunto de entrenamiento y 0,89 para el conjunto de prueba. Respecto al MSE y RMSE los valores observados son 2,13 y 1,46 respectivamente para el entrenamiento, y 2,16 y 1,46 para el conjunto de prueba. Y la última métrica evaluada el MAPE, la cual muestra un valor del 8,64 % para el conjunto de entrenamiento y 8,73 % para el conjunto de prueba.

En la investigación de Contreras et al. (2020) se evalúan 4 modelos de machine learning, las cuales son: SVC con una exactitud de 0,66, precisión de 0,61, Recall de 0,66, y F1 de 0,54; KNN con una exactitud de 0,64, precisión de 0,55, Recall de 0,64, y F1 de 0,57; perceptrón con una exactitud de 0,6624, precisión de 0,61, Recall de 0,66, y F1 de 0,54; árbol de decisión con una exactitud de 0,65, precisión de 0,55, Recall de 0,64, y F1 de 0,57. Por otro lado, en la investigación de (Mengash, 2020), se evalúan 4 modelos de machine learning, las cuales son: naive bayes con un accuracy de 73,61, recall de 73,38, precisión de 73,54, y f1-measure de 73,46; svm con un accuracy de 75,28, recall de 75,58, precisión de 75,30, y f1-measure de 75,44; decision tree con un accuracy de 75,91, recall de 80,24, precisión de 81,02, y f1-measure de 80,63; ann con un accuracy de 79,22, recall de 78,03, precisión de 81,44, y f1-measure de 79,70, Como también en la investigación de (Ahmed & Al-Omari, 2024), se evalúan 4 modelos de machine learning, las cuales son: decision tres con una precisión de 0,9018, recall de 0,9043, accuracy de 0,9341 y

kappa statistic de 0,9004462; naïve bayes con una precisión de 0,8918, recall de 0,8943, accuracy de 0,8332 y kappa statistic de 0,7428229; k-nearest neighbors con una precisión de 0,8941, recall de 0,8984, accuracy de 0,8738 y kappa statistic de 0,7738735; support vector machine (svm) con una precisión de 0,9418, recall de 0,9843, accuracy de 0,9603 y kappa statistic de 0,9398497.

En la investigación de Salas et al. (2024) se evalúan 5 modelos de machine learning, las cuales son: logistic regression con un accuracy de 0,676, AUC ROC de 0,703 y f1 de 0,568; ridge con un accuracy de 0,679, AUC ROC de 0,698 y f1 de 0,556; lasso con un accuracy de 0,678, AUC ROC de 0,695 y f1 de 0,551; random forest con un accuracy de 0,684, AUC ROC de 0,718 y f1 de 0,604; extreme gradient boosting con un accuracy de 0,683, AUC ROC de 0,716 y f1 de 0,595. En la investigación de Candia Oviedo (2019), se menciona que se evalúan 5 modelos de machine learning, las cuales son: árboles de decisión j-48 con un porcentaje de predicción acertada de 67,27 %, acertó un total de 3 898 de 6 119 para malo y un total de 4 323 de 6 490 para bueno; random forest con un porcentaje de predicción acertada de 69,35 %, acertó un total de 4 094 de 6 119 para malo y un total de 4 399 de 6 490 para bueno; algoritmo de vecino más cercano con un porcentaje de predicción acertada de 63,85 %, acertó un total de 3 866 de 6 119 para malo y un total de 3 933 de 6 490 para bueno; algoritmo de función logística con un porcentaje de predicción acertada de 68,33 %, acertó un total de 4 009 de 6 119 para malo y un total de 4 359 de 6 490 para bueno; algoritmo de perceptrón multicapa con un porcentaje de predicción acertada de 64,80 %, acertó un total de 3 657 de 6 119 para malo y un total de 4 262 de 6 490 para bueno.

Así mismo la investigación de Saire (2023) menciona que, se evalúa las métricas de los algoritmos pero por curso, donde se pretende evaluar si un estudiante va a aprobar o desaprobar un curso, teniendo como primer curso estructuras discretas I, donde se tiene los siguientes algoritmos: xgboost con una precisión de 0,870, recall de 0,980 y f1 de 0,920; random forest con una precisión de 0,916, recall de 0,918 y f1 de 0,880, El segundo curso fue estructuras discretas II, donde solo están presentes los estudiantes que aprobaron el curso estructuras discretas I, por lo cual se trata de predecir si un estudiante aprueba o desaprueba un curso, teniendo los siguientes algoritmos: random forest con una

precisión de 0,92, recall de 0,99 y f1 de 0,96; xGBoots con una precisión de 0,93, recall de 1,00 y f1 de 0,96. El tercer curso fue estructura de datos y algoritmo, donde solo están presentes los estudiantes que aprobaron el curso estructuras discretas II, por lo cual se trata de predecir si un estudiante aprueba o desaprueba un curso, teniendo los siguientes algoritmos: random forest con una precisión de 0,89, recall de 0,97 y f1 de 0,93; xGBoots con una precisión de 0,52, recall de 1,00 y f1 de 0,67. En la investigación de (Yupanqui, 2018), se evaluó el modelo de redes neuronales con el algoritmo de retropropagación en perceptrón multicapa donde se tuvo un error de 5 % en entrenamiento y un error cuadrático medio de 6,2 % de error en validación.

CONCLUSIONES

1. Se comparó los 2 algoritmos que tuvieron más precisión respecto a predecir el rendimiento académico universitario, luego de compararlos estadísticamente se observó que el modelo de redes neuronales sobresalió por su mayor exactitud y capacidad predictiva, con un nivel de confianza al 95 %.

2. Se logró preparar la base de datos necesaria para construir modelos de machine learning orientados a la predicción del rendimiento académico universitario de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann (UNJBG) en el año 2023, se contó con un total de 311 alumnos ingresantes como data inicial, luego de haber realizado el proceso de limpieza de datos aplicando los criterios de exclusión, se tuvo un total de 143 alumnos ingresantes, disminuyendo en un 54,02 % de la data inicial.

3. Se construyeron modelos de machine learning con sus mejores hiperparámetros, las cuales son: regresión lineal en sus modelos de ridge con un α =2,559548 y lasso con un α =0,040949, árbol de decisión con una profundidad de 4, bosques aleatorios con una profundidad fue de 5 y redes neuronales con 10 neuronas en la capa oculta.

4. Se evaluaron 4 modelos de machine learning, las cuales son: regresión lineal en sus modelos de ridge y lasso, árbol de decisión, bosques aleatorios y redes neuronales; las cuales fueron 2 algoritmos que mostraron mejores resultados, el primero es bosques aleatorios teniendo un MAPE igual a 8,95 % para el conjunto de entrenamiento y 14,90 % para el conjunto de prueba; el segundo modelo que mostró buenos resultados fue Redes neuronales con un MAPE igual a 8,64 % para el conjunto de entrenamiento y 8,73 % para el conjunto de prueba.

RECOMENDACIONES

1. Comparar estadísticamente más métricas en los modelos de machine learning, para saber si hay diferencias significativas.

2. Ampliar el conjunto de datos incorporando registros académicos de estudiantes de años anteriores.

3. Construir otros modelos aparte de las que ya se hicieron en esta investigación, como: xGBoost, bayesian regression, elastic net, etc; así como se sugiere a la oficina de Admisión otorgar una mayor importancia al curso de **razonamiento verbal**, ya que se destaca como el factor más influyente en la predicción del desempeño académico.

4. Evaluar otras métricas aparte de las que ya se hicieron en esta investigación, como: R^2, SMAPE, HUBER LOSS, etc.

REFERENCIAS BIBLIOGRÁFICAS

Adekitan, I., & Noma, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies, 24*, 1527–1543. https://doi.org/10,1007/s10639-018-9839-7

Ahmed, E., & Al-Omari, A. (2024). Student Performance Prediction Using Machine Learning Algorithms. *Applied Computational Intelligence and Soft Computing, 2024*(1), 4067721. https://doi.org/10.1155/2024/4067721

Arana, M. (2021). *Modelo de predicción del éxito académico de los procesos de admisión con criterios múltiples empleando herramientas de machine learning.* [Tesis Doctoral, Universidad Nacional del Centro del Perú]. http://hdl.handle.net/20.500.12894/10359

Arias, F. (2012). *El proyecto de investigación. Introducción a la metodología científica.* Editorial Episteme. Caraca-República Bolivariana de Venezuela.

Aronés, E. (2021). *Predicción del rendimiento académico basado en Machine Learning, Escuela Profesional de Ingeniería de Sistemas, Ayacucho 2021.* [Tesis de pregrado, Universidad Nacional de San Cristóbal de Huamanga]. http://repositorio.unsch.edu.pe/handle/UNSCH/4694

Assiri, B., Bashraheel, M., & Alsuri, A. (2024). Enhanced Student Admission Procedures at Universities Using Data Mining and Machine Learning Techniques. *Applied Sciences, 3*, 14. https://doi.org/10,3390/app14031109

Behar, D. (2008). *Metodología de la investigación.* Editorial Shalom.

Benos, N., & Zotou, S. (2014). Education and Economic Growth: A Meta-Regression Analysis. *World Development, 64*, 669–689. https://doi.org/10.1016/J.WORLDDEV.2014.06.034

Caballero, D., Abello, R., & Palacio, J. (2007). Relación del burnout y el rendimiento académico con la satisfacción frente a los estudios en estudiantes universitarios. *Avances en Psicología Latinoamericana, 25*(2), 98-111. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-47242007000200007

Cabana, S. (2018). *Análisis predictivo del rendimiento académico en los alumnos de la escuela profesional de ingeniería en informática y sistemas de la UNJBG, utilizando redes neuronales semestre 2017-I.* [Tesis de pregrado, Universidad Nacional Jorge Basadre Grohmann]. http://repositorio.unjbg.edu.pe/handle/UNJBG/3200

Candia, D. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático.* [Tesis de pregrado, Universidad Nacional de San Antonio Abad del Cusco]. http://hdl.handle.net/20,500,12918/4120

Chang, & H. (2023). *Comparación de técnicas de estimación basadas en machine learning para predecir costos en los planes de adquisiciones de las entidades públicas del Perú.* [Tesis de pregrado, Universidad Señor de Sipan]. https://hdl.handle.net/20.500.12802/10566

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *Step-by-step data mining guide.*

Charris, L., Henriquez, C., Hernandez, S., Jimeno, L., Guillen, O., & Moreno, S. (2018). Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos. *Investigación y Desarrollo en TIC, 9*(1). https://revistas.unisimon.edu.co/index.php/identic/article/view/3158

Chávez, D., & Mendoza, J. (2019). Rendimiento académico de secundaria y su relación con el de primer año de educación universitaria en la UN/JBG de Tacna, en el año 2006. *Ciencia & Desarrollo, 10*, 47–50. https://doi.org/10.33326/26176033.2006.10.197

Contreras, L., Fuentes, H., & Rodríguez, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación universitaria, 13*(5), 233-246. https://doi.org/10.4067/S0718-50062020000500233

Contreras, R. (2020). *Relación entre Puntaje de Examen de Admisión o CEPU, y el Rendimiento Académico en la Asignatura de Morfología, Estructura y Función del Cuerpo Humano de Estudiantes del Primer Año de Odontología de la Universidad Nacional Jorge Basadre Grohmann de Tacna,.* [Tesis de Maestría, Universidad Privada de Tacna]. http://hdl.handle.net/20.500.12969/1374

Forero, W., & Negre, F. (2023). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia, 27*(1), 209–253. https://doi.org/10.5944/ried.27.1.37491

Garbanzo, G. (2012). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación, 31*(1), 43. https://doi.org/10,15517/revedu.v31i1.1252

García, M. (2018). *Análisis de Sensibilidad Mediante Random Forest.* [Trabajo de grado, Universidad Politécnica de Madrid]. https://oa.upm.es/53368/

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly.

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación.* McGraw-Hill.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning for dumies.* IBM limited edition.

Ley 30220. (2014). *Ley Universitaria.* Ministerio de Educación.

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics, 5*(1), 1–16. https://doi.org/10.1007/S13721-016-0125-6/METRICS

Mengash, A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. *IEEE Access, 8*, 55462–55470. https://doi.org/10,1109/ACCESS.2020,2981905

Ocaña, Y. (2014). Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. *Revista del Instituto de Investigaciones Educativas, 27*(15). https://revistasinvestigacion.unmsm.edu.pe/index.php/educa/article/view/6473

Palacios, B., & Pajares, M. (2019). Relación entre el rendimiento del examen de admisión y el académico. Tacna, 2001-2005. *Ciencia & Desarrollo, 10*, 15-18. https://doi.org/10.33326/26176033.2006.10.190

Rigatti, J. (2017). Random Forest. *Journal of Insurance Medicine, 47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Saire, E. (2023). *Predicción de la ruta de rendimiento académico con algoritmos de clasificación.* [Tesis Doctoral, Universidad Nacional de San Agustín de Arequipa]. https://hdl.handle.net/20.500.12773/16154

Salas, F., Caldas, J., Salas, M., Borja, N., Daniel, J., & Velásquez, C. (2024). Predicting undergraduate academic performance in a leading Peruvian university: A machine learning approach. *Educación, 33*(64), 55–85. https://doi.org/10,18800/EDUCACION.202401.M003

SENAJU. (31 de Marzo de 2023). *Día Mundial de la Educación: Más del 90% de jóvenes de 15 a 29 años accede a la educación secundaria y menos del 40% transita a la educación superior*. https://juventud.gob.pe/2023/03/dia-mundial-de-la-educacion-mas-del-90-de-jovenes-de-15-a-29-anos-accede-a-la-educacion-secundaria-y-menos-del-40-transita-a-la-educacion-superior/

Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., & Martin, L. (2017). An Agile Group Aware Process beyond CRISP-DM: A Hospital Data Mining Case Study. *Proceedings of the International Conference on Compute and Data Analysis*, 109–113. https://doi.org/10,1145/3093241.3093273

SUNEDU. (2023). *Universidades Licenciadas*. https://www.sunedu.gob.pe/lista-de-universidades-licenciadas/#:~:text=Lista%20%20de%20%20universidades%20%20licenciadas,se%20%20han%20%20otorgado%20%2097%20%20licenciamientos

Tamayo, M. (2014). *El proceso de la investigación científica.* Limusa.

Tejedor, F., & García, A. (2007). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos): propuestas de mejora en el marco del EEES. *Revista de Educación, 342*, 443–474. https://dialnet.unirioja.es/servlet/articulo?codigo=2254218

Vinod, C. (30 de Julio de 2024). *Guía completa para la validación cruzada K-Fold*. datacamp: https://www.datacamp.com/es/tutorial/k-fold-cross-validation

Yamao, E. (2018). *Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de las Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú.* [Tesis de Maestría, Universidad de San Martín de Porres]. https://hdl.handle.net/20.500.12727/3555

ANEXOS

ANEXO 1: FICHA DIGITAL DE OBSERVACIÓN

Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023

	Datos de entrenamiento del rendimiento en el examen de admisión y rendimiento académico universitario	Datos de validación del rendimiento en el examen de admisión y rendimiento académico universitario
Algoritmos de machine learning	MAE MSE RMSE MAPE	MAE MSE RMSE MAPE
Regresión lineal ridge		
Regresión lineal lasso		
Árbol de decisión		
Bosques aleatorios		
Redes neuronales		

ANEXO 2: VALIDACIÓN DE EXPERTOS

VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

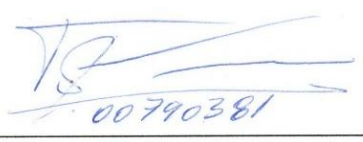
1. DATOS DEL EXPERTO

APELLIDO Y NOMBRES	Polar Fuentes Jaime Freddy
GRADO ACADÉMICO	Msc. Ing. de Sistemas e Informatica - Adm. Tec. Inf.
TÍTULO PROFESIONAL	Ingeniero de Sistemas
EMPRESA / INSTITUCIÓN DONDE LABORA	Docente - Posgrado UNJBG
INSTRUMENTO EVALUADO	**FICHA DE OBSERVACIÓN 01**
AUTOR DEL INSTRUMENTO	**BACH. NAIN NEPTALÍ ACERO MAMANI**
TITULO DE TESIS	Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023.

2. CRITERIOS DE VALIDACIÓN

INDICADOR	MÉTRICAS	CALIFICACIÓN				
		DEFICIENTE 01-20%	MALO 21-40%	REGULAR 41-60%	BUENA 61-80%	EXCELENTE 81-100%
CLARIDAD	Facilidad con la que se entiende el contenido.					X
OBJETIVIDAD	Nivel de imparcialidad y ausencia de opiniones personales.				X	
ORGANIZACIÓN	Está presentado de forma estructurada y secuencial.					X
SUFICIENCIA	El contenido es adecuado y completo para cumplir los objetivos.					X
PERTINENCIA	El contenido es relevante y apropiado para la investigación.					X
COHERENCIA	Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo.				X	
RELEVANCIA	Evalúa la importancia del instrumento.					X
ACTUALIDAD	Adecuado al avance de la tecnología.					X

3. RESULTADOS DE VALIDACIÓN

PROMEDIO DE VALIDACIÓN	95%
EVALUACIÓN DE FACTIBILIDAD	Es Aplicable
FECHA DE EVALUACIÓN	20/12/2024
FIRMA DEL EXPERTO	 00790381

VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

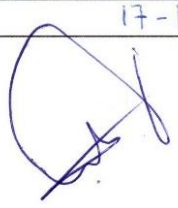
1. DATOS DEL EXPERTO

APELLIDO Y NOMBRES	Mori Sosa Luis Johnson Paul
GRADO ACADÉMICO	Magister en Administración
TÍTULO PROFESIONAL	Ingeniero de sistemas
EMPRESA / INSTITUCIÓN DONDE LABORA	Docente - UNJBG
INSTRUMENTO EVALUADO	**FICHA DE OBSERVACIÓN 01**
AUTOR DEL INSTRUMENTO	**BACH. NAIN NEPTALÍ ACERO MAMANI**
TITULO DE TESIS	Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023.

2. CRITERIOS DE VALIDACIÓN

INDICADOR	MÉTRICAS	CALIFICACIÓN				
		DEFICIENTE 01-20%	MALO 21-40%	REGULAR 41-60%	BUENA 61-80%	EXCELENTE 81-100%
CLARIDAD	Facilidad con la que se entiende el contenido.					X
OBJETIVIDAD	Nivel de imparcialidad y ausencia de opiniones personales.					X
ORGANIZACIÓN	Está presentado de forma estructurada y secuencial.					X
SUFICIENCIA	El contenido es adecuado y completo para cumplir los objetivos.					X
PERTINENCIA	El contenido es relevante y apropiado para la investigación.					X
COHERENCIA	Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo.					X
RELEVANCIA	Evalúa la importancia del instrumento.					X
ACTUALIDAD	Adecuado al avance de la tecnología.					X

3. RESULTADOS DE VALIDACIÓN

PROMEDIO DE VALIDACIÓN	100%
EVALUACIÓN DE FACTIBILIDAD	Es Aplicable
FECHA DE EVALUACIÓN	17-12-2024
FIRMA DEL EXPERTO	

VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

1. DATOS DEL EXPERTO

APELLIDO Y NOMBRES	Mollo Condori Nelson Abraham Pablo
GRADO ACADÉMICO	Msc. Ing. de Sistemas e Informática - Adm. tec. Inf.
TÍTULO PROFESIONAL	Ingeniero en Informatica y Sistemas
EMPRESA / INSTITUCIÓN DONDE LABORA	IDSAR - UNJBG
INSTRUMENTO EVALUADO	**FICHA DE OBSERVACIÓN 01**
AUTOR DEL INSTRUMENTO	**BACH. NAIN NEPTALÍ ACERO MAMANI**
TITULO DE TESIS	Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023.

2. CRITERIOS DE VALIDACIÓN

INDICADOR	MÉTRICAS	CALIFICACIÓN				
		DEFICIENTE 01-20%	MALO 21-40%	REGULAR 41-60%	BUENA 61-80%	EXCELENTE 81-100%
CLARIDAD	Facilidad con la que se entiende el contenido.					X
OBJETIVIDAD	Nivel de imparcialidad y ausencia de opiniones personales.					X
ORGANIZACIÓN	Está presentado de forma estructurada y secuencial.				X	
SUFICIENCIA	El contenido es adecuado y completo para cumplir los objetivos.					X
PERTINENCIA	El contenido es relevante y apropiado para la investigación.					X
COHERENCIA	Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo.					X
RELEVANCIA	Evalúa la importancia del instrumento.					X
ACTUALIDAD	Adecuado al avance de la tecnología.					X

3. RESULTADOS DE VALIDACIÓN

PROMEDIO DE VALIDACIÓN	95%
EVALUACIÓN DE FACTIBILIDAD	Es Aplicable
FECHA DE EVALUACIÓN	16-12-2024
FIRMA DEL EXPERTO	

VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

1. DATOS DEL EXPERTO

APELLIDO Y NOMBRES	Cori Moron Ana Silvia
GRADO ACADÉMICO	Doctora en Ingeniería de Sistemas
TÍTULO PROFESIONAL	Ingeniero en Informatica y Sistemas
EMPRESA / INSTITUCIÓN DONDE LABORA	Docente - UNJBG
INSTRUMENTO EVALUADO	**FICHA DE OBSERVACIÓN 01**
AUTOR DEL INSTRUMENTO	**BACH. NAIN NEPTALÍ ACERO MAMANI**
TITULO DE TESIS	Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023.

2. CRITERIOS DE VALIDACIÓN

INDICADOR	MÉTRICAS	CALIFICACIÓN				
		DEFICIENTE 01-20%	MALO 21-40%	REGULAR 41-60%	BUENA 61-80%	EXCELENTE 81-100%
CLARIDAD	Facilidad con la que se entiende el contenido.				X	
OBJETIVIDAD	Nivel de imparcialidad y ausencia de opiniones personales.					X
ORGANIZACIÓN	Está presentado de forma estructurada y secuencial.					X
SUFICIENCIA	El contenido es adecuado y completo para cumplir los objetivos.					X
PERTINENCIA	El contenido es relevante y apropiado para la investigación.					X
COHERENCIA	Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo.					X
RELEVANCIA	Evalúa la importancia del instrumento.					X
ACTUALIDAD	Adecuado al avance de la tecnología.					X

3. RESULTADOS DE VALIDACIÓN

PROMEDIO DE VALIDACIÓN	97%
EVALUACIÓN DE FACTIBILIDAD	Es Aplicable
FECHA DE EVALUACIÓN	20/12/2024
FIRMA DEL EXPERTO	DRA. ANA CORI MORON DNI 41620155

VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

1. DATOS DEL EXPERTO

APELLIDO Y NOMBRES	Loaiza Fabian Arnold Christian
GRADO ACADÉMICO	Msc. Informatica, con Mención en Tec. Inf.
TÍTULO PROFESIONAL	Ingeniero en Informatica y Sistemas
EMPRESA / INSTITUCIÓN DONDE LABORA	Docente - UNJBG
INSTRUMENTO EVALUADO	**FICHA DE OBSERVACIÓN 01**
AUTOR DEL INSTRUMENTO	**BACH. NAIN NEPTALÍ ACERO MAMANI**
TITULO DE TESIS	Comparación de algoritmos de machine learning en la predicción del rendimiento académico universitario basado en el rendimiento en el examen de admisión de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann del año 2023.

2. CRITERIOS DE VALIDACIÓN

INDICADOR	MÉTRICAS	CALIFICACIÓN				
		DEFICIENTE 01-20%	MALO 21-40%	REGULAR 41-60%	BUENA 61-80%	EXCELENTE 81-100%
CLARIDAD	Facilidad con la que se entiende el contenido.					X
OBJETIVIDAD	Nivel de imparcialidad y ausencia de opiniones personales.					X
ORGANIZACIÓN	Está presentado de forma estructurada y secuencial.				X	
SUFICIENCIA	El contenido es adecuado y completo para cumplir los objetivos.					X
PERTINENCIA	El contenido es relevante y apropiado para la investigación.				X	
COHERENCIA	Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo.				X	
RELEVANCIA	Evalúa la importancia del instrumento.				X	
ACTUALIDAD	Adecuado al avance de la tecnología.					X

3. RESULTADOS DE VALIDACIÓN

PROMEDIO DE VALIDACIÓN	90%
EVALUACIÓN DE FACTIBILIDAD	Es Aplicable
FECHA DE EVALUACIÓN	19-12-2024
FIRMA DEL EXPERTO	 45674849

ANEXO 3: ARCHIVOS EN GITLAB

https://gitlab.com/nain.acero24/modelosunjbg

Name	Last commit	Last update
.ipynb_checkpoints	modelos machine learning	36 seconds ago
DATASET	modelos machine learning	36 seconds ago
REPORTES	modelos machine learning	36 seconds ago
10_REDES_NEURONALES.ipynb	modelos machine learning	36 seconds ago
1_LIMPIEZA_DATA_RENDIMIENTO_...	modelos machine learning	36 seconds ago
2_LIMPIEZA_DATA_ADMISION.ipynb	modelos machine learning	36 seconds ago
3_EXCLUSION.ipynb	modelos machine learning	36 seconds ago
4_UNIR_DATA_PROCESADA.ipynb	modelos machine learning	36 seconds ago
5_ESIQ_MATRIZ_CORRELACION.ip...	modelos machine learning	36 seconds ago
5_ESIS_MATRIZ_CORRELACION.ipy...	modelos machine learning	36 seconds ago
5_ESMC_MATRIZ_CORRELACION.i...	modelos machine learning	36 seconds ago
5_ESME_MATRIZ_CORRELACION.i...	modelos machine learning	36 seconds ago
5_ESMI_MATRIZ_CORRELACION.ip...	modelos machine learning	36 seconds ago
5_MATRIZ_CORRELACION.ipynb	modelos machine learning	36 seconds ago
5_Z_PARTIR.ipynb	modelos machine learning	36 seconds ago

ANEXO 4: MATRIZ DE CONSISTENCIA

<table>
<tr><th>Problema</th><th>Objetivos</th><th>Hipótesis</th><th colspan="2">Variables e indicadores</th><th>Metodología</th></tr>
<tr>
<td rowspan="5">Problema general
¿Es posible comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?

Problemas específicos
¿Es posible preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?</td>
<td rowspan="5">Objetivo general
Comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Objetivos específicos
Preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.</td>
<td rowspan="5">Hipótesis general
Si es posible comparar los algoritmos de machine learning para predecir el rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.

Hipótesis específicas
Si es posible preparar los datos para construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.</td>
<td colspan="2">Variable independiente: Algoritmos de machine learning</td>
<td rowspan="5">Enfoque la Investigación: Cuantitativo

Diseño de investigación: Experimental

Tipo de investigación: Aplicada

Técnicas de recolección de datos: observación estructurada
Instrumentos de recolección de datos: Ficha digital</td>
</tr>
<tr>
<td colspan="2">Variable independiente: Predicción del rendimiento académico universitario</td>
</tr>
<tr>
<td>Dimensiones</td>
<td>Indicadores</td>
</tr>
<tr>
<td rowspan="3">Rendimiento</td>
<td>Error absoluto medio (MAE)</td>
</tr>
<tr>
<td>Error cuadrático medio (MSE)</td>
</tr>
<tr>
<td>Raíz de error cuadrado medio (RMSE)</td>
</tr>
</table>

¿Es posible construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023? ¿Es posible evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023?	Construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023. Evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.	Si es posible construir modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023. Si es posible evaluar los modelos de machine learning para la predicción del rendimiento académico universitario basado en el examen de admisión en los ingresantes a la Facultad de Ingeniería de la UNJBG en el año 2023.		Error porcentual absoluto medio (MAPE)	de observación Población: 311 estudiantes Muestra: 311 estudiantes