# UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN

## Facultad de Ingeniería

Escuela Profesional de Ingeniería en Informática y Sistemas

# DETERMINACIÓN DE LA RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO ESCOLAR Y EL RENDIMIENTO ACADÉMICO UNIVERSITARIO MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING EN INGRESANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN, AÑO 2023

## TESIS

Presentada por:

Bach. James Enrique Segovia Hinojosa

Para optar el Título Profesional de:

**INGENIERO EN INFORMÁTICA Y SISTEMAS**

TACNA – PERÚ

2024

**UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN**

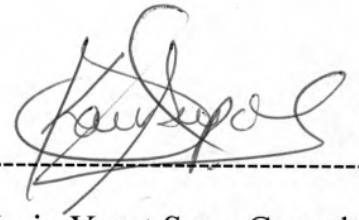
**FACULTAD DE INGENIERÍA**

**ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS**

# DETERMINACIÓN DE LA RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO ESCOLAR Y EL RENDIMIENTO ACADÉMICO UNIVERSITARIO MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING EN INGRESANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN, AÑO 2023

Tesis presentada y aprobada el 18 de diciembre de 2024 estando el jurado calificador integrado por:

Presidente :



---

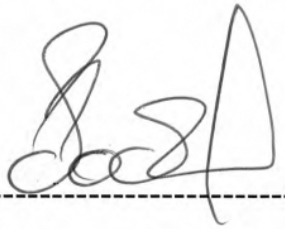
Dra. Karin Yanet Supo Gavacho

Secretario :

---

Dr. Edgar Aurelio Taya Acosta

Vocal :



---

Mag. Oliver Israel Santana Carbajal

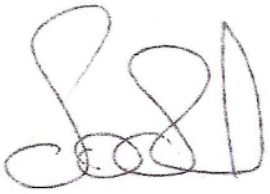
# CERTIFICADO DE SIMILITUD

Yo Oliver Israel Santana Carbajal, en mi condición de asesor acreditado por la Resolución de Facultad N° 07988-2023-FAIN/UNJBG de la tesis:

**"DETERMINACIÓN DE LA RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO ESCOLAR Y EL RENDIMIENTO ACADÉMICO UNIVERSITARIO MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING EN INGRESANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN, AÑO 2023".**

Presentado por James Enrique Segovia Hinojosa (2017-119003)". Presentado por el Bachiller James Enrique Segovia Hinojosa para optar el título profesional de Ingeniero en Informática y Sistemas.

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 9%. Por lo que, **CERTIFICO LA SIMILARIDAD** de la tesis enunciado líneas de arriba, la cual está expedita para continuar con los trámites para la obtención del Título Profesional según corresponda consiguientemente la publicación en el repositorio institucional.



Mgr. Oliver Israel Santana Carbajal
DNI: 00792755
ASESOR

Bach. James Enrique Segovia Hinojosa
DNI: 73857869
TESISTA

**Dedicatoria:**

A mi abuelita Agripina, cuyo amor infinito y apoyo incondicional han sido mi fortaleza en los momentos más difíciles. A mi madre, por su incansable aliento y por enseñarme a no rendirme nunca, incluso ante los desafíos más grandes. Y a toda mi familia, por su constante presencia y por ser mi pilar de apoyo en cada paso de este camino.

**Agradecimiento:**

Agradezco a Dios por sus bendiciones, por guiarme y brindarme la fortaleza necesaria para superar cada obstáculo en este camino.

A mi tío, el ingeniero Hinojosa Ramos, por su constante apoyo y orientación en mi desarrollo profesional. Sus consejos y experiencia han sido invaluables para mi crecimiento.

A mi co-asesor, el ingeniero Chaparro Cruz, por su inquebrantable apoyo y perseverancia en este trabajo. Su dedicación y compromiso han sido una fuente de inspiración y motivación para alcanzar este logro.

# ÍNDICE TEMÁTICO


ÍNDICE TEMÁTICO vi
ÍNDICE DE TABLAS viii
ÍNDICE DE FIGURAS ix
RESUMEN ixi
INTRODUCCIÓN 1
CAPÍTULO I PLANTEAMIENTO DEL PROBLEMA 2
1.1. Antecedentes del problema a investigar 2
1.2. Descripción del problema 2
1.3. Formulación del problema 3
1.3.1. Problema general 3
1.4. Objetivos de la investigación 4
1.4.1. Objetivo general 4
1.4.2. Objetivos específicos 4
1.5. Justificación e importancia de la investigación 5
1.6. Limitaciones 5
1.7. Viabilidad del estudio 6
1.8. Formulación de hipótesis 6
1.8.1. Hipótesis general 6
1.8.2. Hipótesis derivadas o secundarias 6
1.9. Variables 8
1.10. Operacionalización de variables 8
CAPÍTULO II MARCO TEÓRICO 9
2.1. Antecedentes del trabajo de investigación 9
2.2. Bases teóricas 11
2.2.1. Machine learning 11
2.2.1.1. Regresión lineal 11
2.2.1.2. Regresión lineal múltiple 11
2.2.2. Árboles de decisión 12
2.2.2.1. Redes neuronales 13

2.3. Definiciones conceptuales 14
2.3.1. Error cuadrático medio 14
2.3.2. Raíz del error cuadrático medio 15
2.3.3. Rendimiento escolar 16
2.3.4. Rendimiento académico 16
CAPÍTULO III MARCO METODOLÓGICO 18
3.1. Planteamiento metodológico 18
3.1.1. Nivel de la investigación 18
3.1.2. Diseño de la investigación 18
3.1.3. Tipo de investigación 19
3.2. Población y muestra 19
3.2.1. Población 19
3.2.2. Muestra 19
3.3. Equipos y materiales 19
3.4. Procedimiento de las pruebas experimentales 19
3.5. Técnicas de recolección de datos 20
3.5.1. Ficha de observación 20
3.6. Técnicas para el procesamiento de datos 21
CAPÍTULO IV RESULTADOS 23
4.1. Preparación del conjunto de datos 23
4.1.1. Datos de rendimiento escolar 23
4.1.2. Datos del rendimiento universitario. 43
4.1.3. Preparación del Dataset 46
4.2. Determinación de la relación usando Regresión Lineal 53
4.2.1. Regresión Ridge 54
4.2.2. Modelo de Regresión Lasso 61
4.3. Aplicación de Árbol de Decisión para Regresión 67
4.4. Modelo de redes neuronales 77
CAPÍTULO V DISCUSIÓN 82
CONCLUSIONES 84
RECOMENDACIONES 87
REFERENCIAS BIBLIOGRÁFICAS 88
ANEXOS 90

# ÍNDICE DE TABLAS

Tabla 1. Operacionalización de variables 8
Tabla 2. Cantidad de certificados por escuela 23
Tabla 3. Resumen de certificados por tipo y año 27
Tabla 4. Promedio de notas de arte por carrera 28
Tabla 5. Promedio de notas de ciencia y tecnología por carrera 30
Tabla 6. Promedio de notas de ciencias sociales por carrera 31
Tabla 7. Promedio de notas de comunicación por carrera 33
Tabla 8. Promedio de notas DPCC por carrera 34
Tabla 9. Promedio de notas educación física por carrera 36
Tabla 10. Promedio de notas de educación por el trabajo por carrera 37
Tabla 11. Promedio de notas de educación religiosa por carrera 39
Tabla 12. Promedio de notas de inglés por carrera 40
Tabla 13. Promedio de notas de matemática por carrera 42
Tabla 14. Cantidad de registros de notas utilizados por carrera 49
Tabla 15. Intercepto de las variables 57
Tabla 16. Métricas de desempeño obtenidas del método de Regresión Ridge 60
Tabla 17. Intercepto del modelo 64
Tabla 18. Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Regresión Lasso 65
Tabla 19. Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 3 75
Tabla 20. Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 4 76
Tabla 21. Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 5 77
Tabla 22. Valores obtenidos en las métricas de desempeño en base a la cantidad de neuronas 78

# ÍNDICE DE FIGURAS


Figura 1. Árbol de decisión 13
Figura 2. Esquema básico de una neurona artificial 14
Figura 3. Curvas de error en el entrenamiento y la prueba. 17
Figura 4. Gráfico de algoritmos de machine learning utilizados 20
Figura 5. Cantidad de certificados obtenidos 24
Figura 6. Tipos de certificados según campos 24
Figura 7. Equivalencia de cursos en el formato antiguo al formato moderno 25
Figura 8. Histograma del promedio de notas de arte por carrera. 29
Figura 9. Histograma del promedio de notas de ciencia y tecnología por carrera. 30
Figura 10. Histograma del promedio de notas de ciencias sociales por carrera. 32
Figura 11. Histograma del promedio de notas de comunicación por carrera 33
Figura 12. Histograma del promedio de notas de por carrera 35
Figura 13. Histograma del promedio de notas de por carrera 36
Figura 14. Histograma del promedio de notas de educación por el trabajo por carrera 38
Figura 15. Histograma del promedio de notas de educación religiosa por carrera 39
Figura 16. Histograma del promedio de notas de inglés por carrera 41
Figura 17. Histograma del promedio de notas de matemática por carrera 42
Figura 18. Promedio de notas en el primer semestre 43
Figura 19. Promedio de notas en el segundo semestre 44
Figura 20. Promedio de notas por cada escuela 45
Figura 21. Porcentaje de datos faltantes en registros de estudiante 47
Figura 22. Porcentaje de valores faltantes de cada columna 48
Figura 23. Gráfica de la relación de las variables independientes con la nota final 51
Figura 24. Matriz de correlación 52
Figura 25. Gráfico de dispersión de calificaciones con el modelo Bridge 55
Figura 26. Gráfico de líneas de predicción con el modelo Ridge 55
Figura 27. Gráfico de dispersión de valores reales y predichos 56

Figura 28. Gráfico de dispersión de valores reales y predicciones 62
Figura 29. Gráfico de líneas para valores reales y predicciones 63
Figura 30. Gráfico de Dispersión de Valores Predichos vs. Valores Reales 63
Figura 31. Gráfico de dispersión - Max_Depth 3 68
Figura 32. Gráfico de dispersión - Max_Depth 4 69
Figura 33. Gráfico de dispersión - Max_Depth 5 70
Figura 34. Gráfico de líneas - Max_Depth 3 71
Figura 35. Gráfico de líneas - Max_Depth 4 71
Figura 36. Gráfico de líneas - Max_Depth 5 71
Figura 37. Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth_3 72
Figura 38. Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth_4 72
Figura 39. Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth_5 73
Figura 40. Diagrama de Árbol de decisión - Max_Depth 3 74
Figura 41. Árbol de decisión - Max_Depth 4 74
Figura 42. Árbol de decisión - Max_Depth 5 74
Figura 43. Gráfico lineal de los valores obtenidos en las métricas de desempeño en base a la cantidad de neuronas 79

## RESUMEN

Este trabajo de investigación busca determinar la relación entre el rendimiento académico escolar y el universitario de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann en 2023, utilizando algoritmos de machine learning. Se emplearon modelos de regresión lineal, árboles de decisión y redes neuronales para desarrollar herramientas predictivas que identifiquen patrones significativos en los datos académicos, con el fin de mejorar los procesos de admisión y apoyo académico.

El estudio aborda la alta tasa de deserción y bajo rendimiento en los primeros años universitarios, subrayando la necesidad de predecir con mayor precisión qué estudiantes tienen mayor probabilidad de éxito. Se analizaron registros académicos de 403 estudiantes, preprocesados para garantizar su calidad, y se implementaron algoritmos de machine learning con herramientas como Scikit-learn y TensorFlow. La calidad de los modelos se evaluó mediante métricas como $R^2$, MSE, RMSE y MAE.

Los resultados indican que las redes neuronales tienen una mejor capacidad para identificar relaciones complejas entre las variables, aunque presentan problemas de sobreajuste. Por su parte, los árboles de decisión y la regresión lineal ofrecen resultados más interpretables, pero con menor precisión. El estudio concluye que el rendimiento escolar es un buen predictor del desempeño universitario, aunque su capacidad explicativa es limitada. Los algoritmos de machine learning permiten explorar nuevas perspectivas en el análisis educativo, sugiriendo la necesidad de incluir factores adicionales para mejorar las predicciones y estrategias de apoyo académico.



**Palabras clave**: rendimiento académico, machine learning, escuela secundaria, universidad, modelos predictivos, redes neuronales, regresión, árboles de decisión, sobreajuste

## ABSTRACT

This research aims to determine the relationship between high school academic performance and university academic performance of students entering the Faculty of Engineering at the Universidad Nacional Jorge Basadre Grohmann in 2023, using machine learning algorithms. Linear regression, decision trees, and neural networks were employed to develop predictive tools that identify significant patterns in academic data, aimed at improving admission processes and academic support.

The study addresses the high dropout rates and low academic performance in the first years of university, highlighting the need for more accurate predictions of which students are likely to succeed. Academic records from 403 students were analyzed, preprocessed to ensure quality, and machine learning algorithms were implemented using tools like Scikit-learn and TensorFlow. The model quality was evaluated using metrics such as $R^2$, MSE, RMSE, and MAE.

The results indicate that neural networks have better capacity to identify complex relationships between variables, though they present overfitting issues. In contrast, decision trees and linear regression provide more interpretable results but with lower accuracy. The study concludes that high school performance is a good predictor of university success, although its explanatory capacity is limited. Machine learning algorithms allow exploring new perspectives in educational analysis, suggesting the need to include additional factors to enhance predictions and academic support strategies.



**Keywords**: academic performance, machine learning, high school, university, predictive models, neural networks, regression, decision trees, overfitting

# INTRODUCCIÓN

El rendimiento académico es un indicador clave en la evaluación de estudiantes y su éxito en la educación superior. Comprender cómo el desempeño escolar influye en el rendimiento universitario es crucial, especialmente con el acceso a datos masivos y algoritmos de aprendizaje automático. Este estudio se enfoca en identificar factores predictivos del éxito académico universitario que puedan orientar decisiones en admisión, diseño curricular y apoyo estudiantil.

En la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, se analizan los datos de los ingresantes en 2023 para explorar la relación entre las calificaciones escolares y el desempeño en el primer año universitario. Utilizando algoritmos de machine learning como regresión lineal, árboles de decisión y redes neuronales, se busca determinar esta relación y evaluar el potencial predictivo de estas herramientas.

El estudio aborda la baja retención y éxito académico en los primeros años universitarios, destacando la importancia de los datos históricos de rendimiento escolar. Además, implementa y evalúa algoritmos mediante métricas como $R^2$, MSE, RMSE y MAE para validar los modelos.

La investigación contribuye al campo de la minería de datos educativa al integrar herramientas computacionales avanzadas para mejorar criterios de admisión, reducir deserción y optimizar el aprendizaje. Se espera que este enfoque aporte tanto a la teoría como a la práctica, promoviendo una gestión educativa más eficiente y alineada con las necesidades de los estudiantes y la institución.

# CAPÍTULO I
# PLANTEAMIENTO DEL PROBLEMA

## 1.1. Antecedentes del problema a investigar

Es ampliamente común observar cómo la cantidad inicial de estudiantes universitarios de una misma cohorte se reduce progresivamente a medida que transcurren los años de estudio en la institución educativa. Esta disminución se atribuye a diversos factores, tales como el rendimiento académico insatisfactorio, cambios en los intereses o metas profesionales, presiones financieras, factores personales y familiares, falta de apoyo académico y social, así como dificultades de adaptación al entorno universitario.

Chong González (1970) en su investigación titulada "Factores que inciden en el rendimiento académico de los estudiantes de la UPVT" llevada a cabo en la Universidad Politécnica del Valle de Toluca, se enfocó en examinar diversos aspectos que influyen en el desempeño académico de los estudiantes universitarios, incluyendo sus contextos universitarios y familiares, percepciones sobre el apoyo familiar, desafíos académicos y expectativas personales y familiares hacia sus carreras.

Los datos recopilados en esta investigación revelaron hallazgos significativos. En particular, se identificó una relación directa entre el apoyo percibido por parte de los estudiantes y su rendimiento académico en la universidad. Además, el estudio destacó la importancia de la participación activa de la familia en el proceso académico como un factor clave para promover y elevar el rendimiento académico de los estudiantes. Esta participación no solo contribuye a mejorar el desempeño escolar, sino que también puede desempeñar un papel fundamental en la prevención de la deserción y el abandono de los estudios superiores.

Estos resultados subrayan la relevancia de considerar factores más allá del entorno académico directo al abordar las cuestiones relacionadas con el rendimiento académico universitario. La inclusión de la familia como un elemento de apoyo puede tener un impacto significativo en el éxito estudiantil en la educación superior.

Oropeza Tena et al., (2017) en su investigación titulada "Comparación entre rendimiento académico, autoeficacia y práctica deportiva en universitarios" busca identificar posibles diferencias en el rendimiento académico y la autoeficacia entre estudiantes universitarios que practican deportes y aquellos que no lo hacen, con un enfoque particular en las diferencias de género.

Este estudio se llevó a cabo mediante una metodología cuantitativa, utilizando un diseño transversal y analítico. Participaron 331 estudiantes de la Facultad de Psicología de la institución académica, de los cuales 72 eran hombres y 259 mujeres. Los datos se recopilaron a través de una ficha de identificación que incluyó información sociodemográfica, así como inventarios que evaluaron las actividades académicas y extracurriculares, la autoeficacia general y se obtuvieron las boletas de calificaciones de los participantes.

Los hallazgos de esta investigación revelaron diferencias estadísticamente significativas en el rendimiento académico y la autoeficacia entre los estudiantes que practicaban actividades deportivas y aquellos que no lo hacían. Estos resultados indican que la práctica sistemática de actividades deportivas puede tener un impacto positivo tanto en el rendimiento académico como en la autoeficacia de los estudiantes. Se argumenta que esta influencia positiva se deriva de la promoción de la disciplina, un estilo de vida saludable y el fomento de un buen rendimiento académico.

## 1.2. Descripción del problema

El ingreso a la universidad está limitado por un número específico de vacantes, lo que impulsa a los aspirantes a someterse a exámenes que evalúen sus conocimientos acordes a la carrera que desean seguir. No obstante, es posible que este método no sea el más adecuado para determinar su aptitud académica, y que el historial de calificaciones obtenidas durante la educación secundaria constituya un indicador más idóneo para identificar si el estudiante posee los conocimientos necesarios para llevar a cabo, de manera satisfactoria, su trayectoria académica y profesional en la universidad.

En los últimos veinte años la cantidad de personas que está accediendo a la educación superior a través de procesos de admisión ha aumentado drásticamente. Según la UNESCO, (2020) la tasa bruta de matriculación en la educación superior a nivel mundial casi se duplicó, pasando del 19% al 38% entre 2000 y 2018.

En el contexto de la educación global, surge una preocupación fundamental: el bajo rendimiento académico de los estudiantes. El informe de la OCDE titulado *Low-Performing Students (2016)* arroja luz sobre este fenómeno y proporciona valiosas perspectivas para nuestra investigación.

Según los hallazgos de la OCDE, casi cuatro millones de estudiantes de 15 años en países miembros de la OCDE presentan bajo rendimiento en matemáticas, mientras que aproximadamente tres millones tienen bajo rendimiento en lectura y ciencias. Estos datos resaltan la magnitud del problema y su impacto en la calidad de la educación.

En el año 2019, se llevó a cabo una Evaluación Censal de Estudiantes dirigida a los alumnos de segundo de secundaria en los distintos departamentos del Perú. Los resultados de esta evaluación, proporcionados por el Ministerio de Educación (MINEDU), revelaron que el 27% de los estudiantes obtuvieron resultados satisfactorios en la evaluación de lectura, mientras que el 38% alcanzaron un desempeño satisfactorio en la evaluación de matemática.

En la Universidad Nacional Jorge Basadre Grohmann, la Facultad de Ingeniería ofrece cinco carreras que serán consideradas en nuestro estudio. Estas carreras son:

- Ingeniería Metalúrgica
- Ingeniería de Minas
- Ingeniería Mecánica
- Ingeniería en Informática y Sistemas
- Ingeniería Química

Estas disciplinas abarcan diversas áreas de la ingeniería y proporcionan un contexto relevante para analizar la relación entre el rendimiento académico escolar y el rendimiento académico universitario.

El presente estudio tiene como objetivo principal determinar la relación entre el rendimiento escolar y el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann.

## 1.3. Formulación del problema

### 1.3.1. Problema general

¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?

### 1.3.2. Problemas específicos

a) ¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?

b) ¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?

c) ¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?

## 1.4. Objetivos de la investigación

### 1.4.1. Objetivo general

Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

### 1.4.2. Objetivos específicos

a) Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023

b) Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023

c) Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023

## 1.5. Justificación e importancia de la investigación

La importancia de esta tesis radica en su enfoque en el rendimiento académico escolar como un predictor del éxito en la universidad, específicamente en la facultad de ingeniería. Al abordar esta relación, la investigación puede proporcionar información valiosa para el proceso de admisión universitaria, permitiendo una selección más precisa de estudiantes con mayores probabilidades de éxito.

Además, al utilizar algoritmos de Machine Learning para analizar patrones en los datos, la tesis puede identificar factores clave en el rendimiento escolar que puedan predecir el desempeño universitario. Esto podría informar la toma de decisiones y la mejora institucional, además de fomentar un enfoque más personalizado en la educación.

En última instancia, los resultados de esta investigación tienen el potencial de influir en la manera en que las instituciones educativas abordan la admisión de estudiantes y la promoción del éxito académico.

## 1.6. Limitaciones

Una de las principales limitaciones de esta investigación fue el proceso de recolección de datos de los certificados académicos. Se obtuvieron datos de 280 certificados, y el hecho de tener que transcribir manualmente las notas de cada uno representó un desafío significativo. Esta tarea resultó ser extremadamente laboriosa y propensa a errores, ya que cada nota debía ser tipeada individualmente. La gran cantidad de datos y el esfuerzo necesario para asegurarse de que cada entrada fuera correcta implicó una inversión considerable de tiempo y recursos.

Adicionalmente, solo se podía recolectar los datos de los certificados durante el horario de atención de la Dirección de Asuntos Académicos (DASA), lo que hizo que la

tarea de recolección de datos fuera mucho más laboriosa debido a las restricciones de tiempo disponibles.

## 1.7. Viabilidad del estudio

El presente estudio es viable debido a que los datos necesarios, como las calificaciones escolares y universitarias de los estudiantes, pueden obtenerse de manera relativamente sencilla mediante una solicitud formal a la oficina de DASA. Además, la implementación de algoritmos de aprendizaje automático para predecir el rendimiento académico resulta factible, ya que existen múltiples métodos ampliamente probados y herramientas accesibles para llevar a cabo este tipo de análisis. El uso de software de código abierto y el respaldo institucional para la investigación aseguran que los recursos tecnológicos y computacionales necesarios estén al alcance, lo que refuerza la factibilidad técnica y operativa.

## 1.8. Formulación de hipótesis

### 1.8.1. Hipótesis general

**HGo:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

**HGa:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning es de moderada a perfecta en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

### 1.8.2. Hipótesis derivadas o secundarias

**Hipótesis derivada 1**

- **H1o:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

- **H1a:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal es de moderada a perfecta en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

**Hipótesis derivada 2**

- **H2o:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.
- **H2a:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión es de moderada a perfecta en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

**Hipótesis derivada 3**

- **H3o:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.
- **H3a:** La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales es de moderada a perfecta en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.

## 1.9. Variables

- **Variable independiente:** Rendimiento Académico Escolar
- **Variable dependiente:** Rendimiento Académico Universitario

## 1.10. Operacionalización de variables

**Tabla 1.**

*Operacionalización de variables*

| VARIABLES | DIMENSIONES | INDICADORES |
|---|---|---|
| **Análisis del rendimiento académico escolar** | Rendimiento académico en matemática | Registro académico de notas en matemática |
| | Rendimiento académico en ciencia y tecnología | Registro académico de notas en ciencia y tecnología |
| | Rendimiento académico en comunicación | Registro académico de notas en comunicación |
| | Rendimiento académico en educación para el trabajo | Registro académico de notas en educación para el trabajo |
| | Rendimiento académico en ciencias sociales | Registro académico de notas en personal social |
| | Rendimiento académico en inglés | Registro académico de notas en ingles |
| | Rendimiento académico en EPT | Registro académico de notas en |
| | Rendimiento académico en arte y cultura | Registro académico de notas en arte y cultura |
| | Rendimiento académico en DPCC | Registro académico de notas en formación ciudadana y cívica |
| | Rendimiento académico en educación física | Registro académico de notas en educación física |
| **Análisis del rendimiento en pregrado** | Nivel académico | Promedio de notas de cursos generales |

# CAPÍTULO II
# MARCO TEÓRICO

## 2.1. Antecedentes del trabajo de investigación

Asor et al., (2023) en el estudio “Prediction of Senior High School Students’ Performance in a State University” se utilizó registros de 4 años de estudiantes de secundaria superior para predecir su rendimiento académico mediante algoritmos de aprendizaje automático. Se probaron modelos como árbol de decisiones, bayes ingenuo, bosque aleatorio, red neuronal y regresión lineal. Los resultados indicaron que el bayes ingenuo fue el más eficaz, seguido de las redes neuronales, mostrando un alto desempeño en la predicción.

El Guabassi et al., (2021) en su estudio “Forecasting Students’ Academic Performance Using Different Regression Algorithms” comparó siete algoritmos de aprendizaje automático para predecir el rendimiento académico de los estudiantes, incluyendo ANCOVA, regresión logística, regresión de vectores de soporte, regresión log-lineal, regresión de árbol de decisión, regresión de bosque aleatorio y regresión de mínimos cuadrados parciales. Los algoritmos fueron evaluados con diversas métricas de rendimiento. Los resultados revelaron que la regresión log-lineal fue la más precisa en la predicción, seguida de cerca por ANCOVA, destacando el potencial del aprendizaje automático como herramienta eficaz en el ámbito educativo.

Chavez et al., (2023) en su estudio titulado “Artificial neural network model to predict student performance using nonpersonal information” se utilizó un modelo de red neuronal artificial con datos de 32,000 estudiantes de The Open University of the United Kingdom, incluyendo el número de veces que tomaron el curso, evaluaciones, tasa de aprobación, uso de materiales virtuales y clics en aulas virtuales. El modelo logró una precisión del 93.81%, una precisión del 94.15%, un recall del 95.13% y un F1-score del 94.64%, ayudando a las autoridades educativas a prevenir el abandono escolar y mejorar el rendimiento académico.

Terán Montaño y Schulmeyer, (2022) en su artículo “Relación entre El Rendimiento Académico en Secundaria y el Rendimiento Académico Universitario”, se investigó si era posible predecir el rendimiento académico de los estudiantes de primer semestre de la Universidad Privada de Santa Cruz de la Sierra a partir de su rendimiento escolar. La muestra consistió en 1.935 estudiantes, 935 varones y 1.000 mujeres, de los primeros semestres de tres gestiones. Para establecer una relación entre el rendimiento escolar y el rendimiento universitario, se analizaron datos académicos que incluían la nota de matemáticas y el promedio de notas del último año de colegio, el índice de aprovechamiento académico en la Universidad, tipo de colegio, género, periodo de ingreso y facultad. Se encontró una correlación moderada entre las notas en la universidad y el colegio. No se encontraron diferencias por género ni por tipo de colegio. También se vio que la relación entre las notas en la universidad y el colegio es mayor en las facultades que involucran más materias numéricas (como ingeniería y administración). Sin embargo, la predictibilidad del rendimiento escolar es baja cuando se la toma como única variable de predicción.

Rico Páez (2022) en su estudio titulado "Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios" se abordó la necesidad de desarrollar modelos predictivos del rendimiento académico en estudiantes universitarios en México. El objetivo principal de esta investigación fue crear modelos de predicción de los resultados académicos utilizando técnicas de aprendizaje automático y evaluar su eficacia. La investigación se basó en la recopilación de calificaciones de actividades académicas de 260 estudiantes universitarios. Estos datos se utilizaron para construir modelos de predicción en diferentes etapas a lo largo del curso. La precisión de estos modelos se evaluó mediante la predicción de los resultados académicos de 112 estudiantes en un curso posterior. Los resultados revelaron una precisión de hasta el 70.5% en un tiempo que representaba el 21% de la duración total del curso. Este enfoque metodológico ofrece flexibilidad en cuanto a la elección de la etapa temporal en la que se realizan las predicciones, lo que lo hace adaptable a diversos tipos de cursos. Además, su capacidad para detectar problemas de rendimiento académico en etapas tempranas puede ayudar a prevenir la reprobación y la deserción estudiantil. Los hallazgos de esta investigación sugieren la utilidad de aplicar métodos de aprendizaje automático para

predecir el rendimiento académico de los estudiantes universitarios y brindar oportunidades para la intervención temprana y el apoyo académico.

En la investigación titulada “Rendimiento académico en estudiantes Vs factores que influyen en sus resultados: una relación a considerar” (Martínez Pérez et al., 2020) se examinaron los factores relacionados con el rendimiento académico de estudiantes de Medicina durante los primeros cinco años de la carrera. Se encontró que los estudiantes con bajo rendimiento académico obtuvieron un promedio de calificaciones 0,7 puntos inferior. Además, la motivación en este grupo fue notablemente baja, alcanzando solo un 39,13 %. También se observó que el 69,57 % de estos estudiantes dedicaba menos de 15 horas semanales al estudio. Por otro lado, los estudiantes con mayor rendimiento académico presentaron un índice académico superior en el preuniversitario, así como mejores resultados en las pruebas de ingreso a la universidad y en la asignatura de Morfofisiología. Estos hallazgos pueden ser útiles para diseñar estrategias de apoyo y mejora en el ámbito educativo.

## 2.2. Bases teóricas

### 2.2.1. Machine learning

Murphy K. (2012) menciona que el aprendizaje automático es el estudio y el desarrollo de algoritmos y modelos que pueden aprender de los datos y mejorar su rendimiento con la experiencia.

En el contexto de esta tesis, se explorarán y aplicarán diversos modelos de aprendizaje automático. Estos modelos desempeñan un papel fundamental en la resolución de problemas complejos y la toma de decisiones basada en datos. A continuación, presentamos los modelos específicos que se abordarán en este estudio.

#### 2.2.1.1. Regresión lineal

Esta técnica estadística ofrece una valiosa herramienta para analizar y comprender las relaciones subyacentes entre variables cuantitativas, lo que la convierte en un recurso esencial en la investigación y el análisis de datos en diversas disciplinas.

#### 2.2.1.2.Regresión lineal múltiple

La regresión lineal múltiple es una extensión de la regresión lineal simple que permite modelar la relación entre una variable dependiente y dos o más variables

independientes. A diferencia de la regresión lineal simple, que considera solo una variable independiente para predecir la variable dependiente, la regresión lineal múltiple incorpora múltiples variables independientes en el modelo.

La ecuación básica de la regresión lineal múltiple se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Donde:

- $Y$ es la variable dependiente que queremos predecir.
- $x_1, x_2, \ldots, x_p$ son las variables independientes que utilizamos para la predicción.
- $\beta_0$ es la ordenada al origen, que representa el valor de $Y$ cuando todas las variables independientes son iguales a cero.
- $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes de regresión que representan la contribución de cada variable independiente a la predicción de $Y$.
- $\varepsilon$ representa el error, que es la diferencia entre el valor observado y el valor predicho.

La regresión lineal múltiple es una herramienta poderosa para analizar la relación entre múltiples variables y predecir valores de la variable dependiente en función de las variables independientes. Se utiliza en una amplia gama de campos, desde la economía hasta la ciencia de datos, para comprender y predecir fenómenos complejos.

### 2.2.2. Árboles de decisión

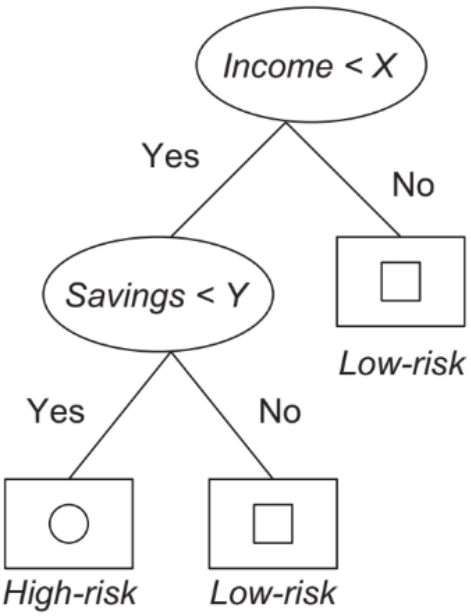
Los árboles de clasificación y regresión (CART), también conocidos como árboles de decisión, son modelos que dividen de manera recursiva el espacio de entrada en regiones específicas, asignando un modelo local a cada una de ellas. Estas divisiones se pueden representar como un árbol, donde cada hoja corresponde a una región del espacio de entrada definida por el modelo (Murphy Kevin, 2012).

Los árboles de decisión son ampliamente utilizados en machine learning debido a su capacidad para manejar tanto variables continuas como categóricas. Su estructura jerárquica facilita la identificación de patrones en los datos, permitiendo descomponer problemas complejos en decisiones más simples y comprensibles. Cada nodo interno del

árbol representa una condición sobre una variable de entrada, mientras que las hojas contienen el resultado asociado a esa ruta. Esto los convierte en modelos versátiles que encuentran aplicaciones en tareas de clasificación, regresión y segmentación de datos.

**Figura 1.**

*Árbol de decisión*




*Nota. Adaptado de "Machine Learning" de Ethem Alpaydin, 2016, p.78. Derechos de autor 2016 por Massachusetts Institute of Technology*

#### 2.2.2.1. Redes neuronales

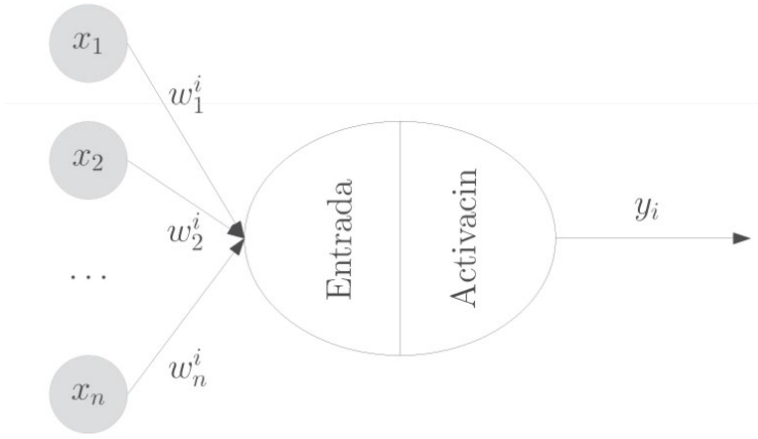
Las redes neuronales artificiales son herramientas versátiles que pueden llevar a cabo tareas como la clasificación en conjuntos de datos etiquetados, la regresión en datos continuos y la segmentación en datos no etiquetados, identificando similitudes entre los datos de entrada. Estas redes se utilizan principalmente para establecer relaciones entre datos de entrada y salida, lo que permite aproximar funciones con alta precisión, por lo que a menudo se consideran "aproximadores" (Bosch Rué et al., 2019).

Como se muestra en la Figura 2, en una red neuronal, múltiples neuronas artificiales trabajan juntas. Cada neurona procesa información y se conecta con otras

mediante sinapsis. Estas conexiones forman una estructura eficiente para el procesamiento de datos, el aprendizaje y el razonamiento.

**Figura 2.**

*Esquema básico de una neurona artificial*




*Nota. Reproducido de "Deep Learning: Principios y Fundamentos" de Bosch Rué, A., Casas Roma, J., & Lozano Bagén, T., 2019, p. 49. Editorial UOC. Derechos de autor 2019 por Editorial UOC*

La figura ilustra el funcionamiento básico de una neurona artificial. Cada neurona recibe un conjunto de entradas representadas como X =$\{X_1, X_2,...X_n\}$, cada una con un peso asociado $W^i$=$\{w_1^i, w_2^i, .., w_n^i\}$ , donde $w_j^i$ representa la importancia del valor de entrada $X_j$ que llega a la neurona i desde la neurona j.

Cada neurona combina estos valores de entrada aplicando una función de combinación o entrada. El valor resultante pasa por una función de activación que ajusta los valores de entrada para producir el valor de salida $y_i$. Este valor generalmente se transmite a las conexiones de la neurona i con otras neuronas o se utiliza como salida final de la red.

## 2.3. Definiciones conceptuales

### 2.3.1. Error cuadrático medio

El Error Cuadrático Medio (ECM) es una medida que se utiliza comúnmente en estadísticas y aprendizaje automático para evaluar la precisión de un modelo en relación

con los valores reales. Se calcula como la media de los cuadrados de las diferencias entre las predicciones del modelo y los valores reales. La fórmula matemática del ECM se expresa de la siguiente manera:

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

Donde:

- $n$ es el número total de observaciones o muestras.
- $y_i$ es el valor real de la observación $i$.
- $\hat{y}_i$ es la predicción del modelo para la observación $i$.

El error cuadrático medio penaliza más fuertemente las grandes desviaciones entre las predicciones y los valores reales, ya que las diferencias se elevan al cuadrado antes de tomar la media. Por lo tanto, es una medida útil para evaluar la calidad de las predicciones de un modelo y comparar diferentes modelos en términos de precisión.

### 2.3.2. Raíz del error cuadrático medio

La Raíz del Error cuadrático Medio (RMSE) es una métrica estadística utilizada para evaluar la precisión de los modelos predictivos, incluyendo aquellos empleados en el ámbito de Machine Learning. Su cálculo implica la obtención de la raíz cuadrada del Error Cuadrático Medio (ECM), que a su vez se define como la media de los cuadrados de las diferencias entre las predicciones del modelo y los valores reales. La fórmula para el cálculo del RMSE se expresa como:

$$RMSE = \sqrt{ECM}$$

Donde:

- *ECM* representa el Error Cuadrático Medio.
- La raíz cuadrada se aplica para proporcionar una medida en las mismas unidades que la variable de interés.

### 2.3.3. Rendimiento escolar

El rendimiento escolar se refiere a la evaluación del conocimiento adquirido por un estudiante durante su último año en el colegio. Este rendimiento se mide a través de las calificaciones obtenidas en exámenes y tareas a lo largo del año académico. Es un indicador clave que refleja la comprensión, habilidades y logros del estudiante en las materias cursadas.

### 2.3.4. Rendimiento académico

El rendimiento académico en la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann se refiere al promedio ponderado obtenido por un estudiante al finalizar su primer año de estudios. Este promedio considera las calificaciones de todas las materias cursadas durante ese período. Es un indicador crucial que refleja el desempeño y la comprensión del estudiante en su programa académico específico.

### 2.3.5. R cuadrado

El coeficiente de determinación, conocido como R-cuadrado ($R^2$), es una medida estadística utilizada para evaluar el ajuste de un modelo de regresión a los datos observados. El valor de $R^2$ oscila entre 0 y 1, donde un valor de 0 indica que el modelo no explica ninguna de las variaciones en la variable dependiente, mientras que un valor de 1 sugiere que el modelo explica completamente las variaciones observadas. En términos más específicos, $R^2$ mide la proporción de la variabilidad total de la variable dependiente que puede ser explicada por las variables independientes incluidas en el modelo. Este coeficiente se calcula como la razón entre la suma de los cuadrados explicados (SCE) y la suma total de los cuadrados (SCT), reflejando así qué tan bien los datos se ajustan al modelo teórico propuesto. Un $R^2$ elevado indica un buen ajuste del modelo, pero es crucial considerar también otros factores como la complejidad del modelo y la presencia de outliers, ya que un modelo excesivamente complejo puede sobre ajustar los datos y llevar a conclusiones erróneas.

### 2.3.6. Early stopping

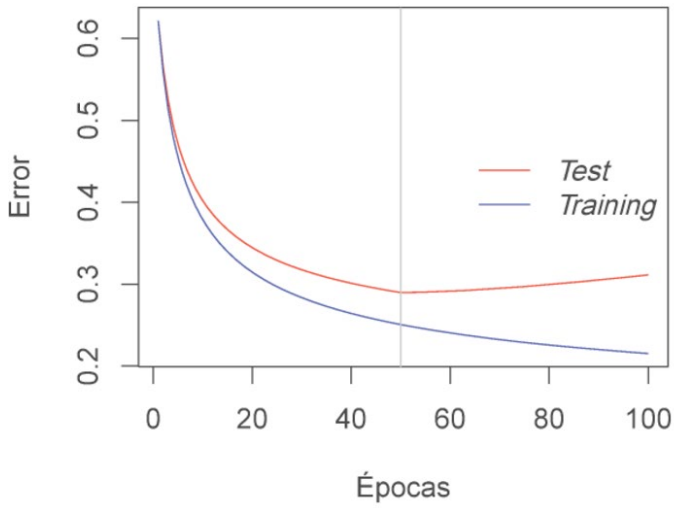
Es una técnica utilizada en el aprendizaje automático que detiene el entrenamiento del modelo cuando el error en el conjunto de validación comienza a aumentar, indicando

que el modelo está empezando a sobreajustarse a los datos de entrenamiento y perdiendo capacidad de generalización (Bosch Rué et al., 2019).

Esta técnica no solo ayuda a prevenir el sobreajuste, sino que también optimiza los recursos computacionales al detener el entrenamiento innecesario. Durante el entrenamiento, el modelo es evaluado periódicamente con el conjunto de datos de validación. Inicialmente, ambos errores, tanto el de entrenamiento como el de validación, disminuyen. Sin embargo, cuando el error de validación deja de disminuir y comienza a aumentar, es una señal de que el modelo ha aprendido demasiado los detalles y el ruido del conjunto de entrenamiento, perjudicando su rendimiento en datos nuevos. Al aplicar early stopping, se asegura que el modelo mantenga una buena capacidad de generalización al nuevo conjunto de datos, manteniendo así su eficacia predictiva en aplicaciones reales.

**Figura 3.**

*Curvas de error en el entrenamiento y la prueba.*




Nota. Reproducido de "Deep Learning: Principios y Fundamentos" de Bosch Rué, A., Casas Roma, J., & Lozano Bagén, T., 2019, p. 49. Editorial UOC. Derechos de autor 2019 por Editorial UOC.

# CAPÍTULO III
# MARCO METODOLÓGICO

## 3.1. Planteamiento metodológico

### 3.1.1. Nivel de la investigación

(Arias, 2012) en su libro "El proyecto de la investigación", menciona que la finalidad de una investigación correlacional es medir el grado de relación entre dos o más variables sin establecer causalidad directa. Aunque no define causas específicas, proporciona indicios útiles sobre posibles relaciones.

El nivel de esta investigación es de tipo correlacional, ya que tiene como objetivo determinar la relación existente entre el rendimiento académico escolar y el rendimiento académico universitario en los estudiantes. Para ello, se utilizarán algoritmos de machine learning que permitirán analizar y predecir estas relaciones a partir de datos históricos de rendimiento académico.

Este enfoque metodológico no solo facilita la identificación de patrones y tendencias significativas, sino que también proporciona una base sólida para la interpretación de los resultados y la elaboración de conclusiones fundamentadas sobre las posibles correlaciones entre las variables estudiadas.

### 3.1.2. Diseño de la investigación

En un diseño no experimental, las investigaciones se llevan a cabo sin manipular deliberadamente las variables, lo que permite observar los fenómenos en su contexto natural para analizar sus características y relaciones. (Hernández Sampieri et al., 2014).

Este enfoque es particularmente adecuado para esta investigación, ya que no se interviene en los procesos educativos, sino que se utilizan datos históricos, como las calificaciones escolares y universitarias, para analizar la relación entre estas variables mediante algoritmos de machine learning. El diseño no experimental garantiza que el análisis respete la naturaleza original de los datos, lo que permite identificar patrones y correlaciones sin alterar las condiciones bajo las cuales se produjeron.

### 3.1.3. Tipo de investigación

Tenemos dos tipos de investigación: la investigación básica, pura o fundamental y la investigación aplicada o tecnológica. Para el caso de la presente investigación se considera que es una investigación aplicada o tecnológica.

La investigación aplicada o tecnológica se enfoca en solucionar problemas relacionados con los procesos de producción, distribución, circulación y consumo de bienes y servicios en diversas actividades humanas. Este tipo de investigación busca mejorar, perfeccionar u optimizar el funcionamiento de sistemas, procedimientos y regulaciones tecnológicas, aprovechando los avances científicos y tecnológicos actuales.(Nicomedes, 2018).

## 3.2. Población y muestra

### 3.2.1. Población

La población de interés para este estudio está constituida por todos los estudiantes matriculados en la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann durante el año del 2023.

Dado que la intención es incluir a la totalidad de la población de estudiantes de la Facultad de Ingeniería, no se aplicará un proceso de muestreo. La totalidad de los alumnos matriculados, que asciende a 403 estudiantes, será considerada en este estudio.

### 3.2.2. Muestra

A efectos del presente estudio, la muestra será censal.

## 3.3. Equipos y materiales

- 1 computadora portátil ACER con procesador i-5

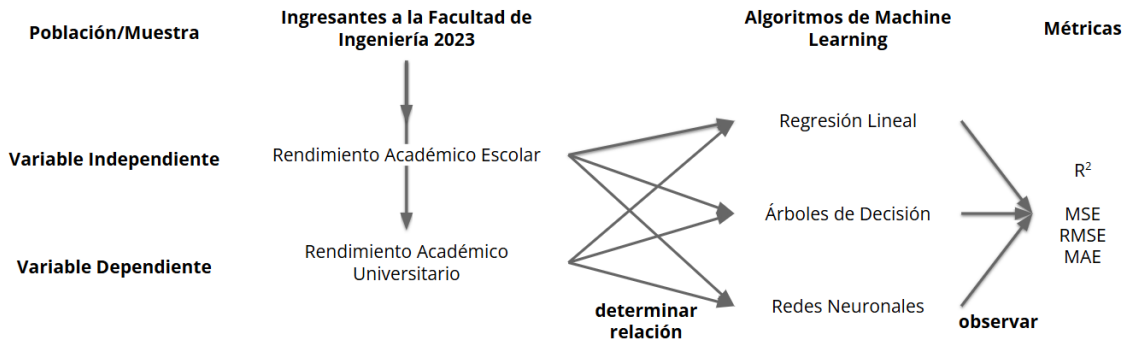
## 3.4. Procedimiento de las pruebas experimentales

En la presente investigación no se realizan experimentos manipulando variables (por ello la caracterizamos como descriptiva), sin embargo, si se realizan experimentos al utilizar algoritmos de machine learning para encontrar la relación entre el rendimiento académico escolar y el rendimiento académico universitario.

A continuación, se presenta un gráfico que resume los experimentos relacionados a los algoritmos de machine learning.

**Figura 4.**

*Gráfico de algoritmos de machine learning utilizados*




Los algoritmos de machine learning generarán una ecuación lineal (para el caso de la Regresión Lineal), un Árbol de Decisión (para el caso de Árboles de Decisión) y un conjunto de matrices de pesos (para el caso de Redes Neuronales) que determinarán la relación entre la variable independiente y la dependiente. Empero, es necesario medir la calidad de dicha relación encontrada por los algoritmos de Machine Learning, esto se realizar a través de las métricas R2, MSE, RMSE, MAE.

## 3.5. Técnicas de recolección de datos

### 3.5.1. Ficha de observación

En el presente estudio, se emplea una ficha de observación como herramienta metodológica para recopilar y sistematizar los datos generados durante la aplicación de los algoritmos de machine learning. Este instrumento permite registrar de manera estructurada los resultados obtenidos en cada experimento, lo que facilita el análisis comparativo de las métricas seleccionadas: coeficiente de determinación ($R^2$), error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y error absoluto medio (MAE). Estas métricas son fundamentales para evaluar la calidad y precisión de los modelos generados por los algoritmos de Regresión Lineal, Árboles de Decisión y Redes Neuronales.

La ficha de observación está diseñada para documentar los valores resultantes de cada métrica, asociados a las iteraciones y configuraciones específicas de los algoritmos. Por ejemplo, en el caso de la Regresión Lineal, se registra la ecuación lineal generada y los valores de $R^2$, MSE, RMSE y MAE correspondientes. Esto permite evaluar en qué medida el modelo lineal explica la relación entre el rendimiento académico escolar (variable independiente) y el rendimiento académico universitario (variable dependiente). Para los Árboles de Decisión, la ficha incluye información sobre la estructura del árbol y las métricas derivadas de la comparación entre los valores predichos y los valores observados en los datos de prueba. Finalmente, en el caso de las Redes Neuronales, se documentan las matrices de pesos finales, así como las métricas obtenidas durante el proceso de entrenamiento y validación del modelo.

## 3.6. Técnicas para el procesamiento de datos

A fin de determinar la relación, se realizarán los siguientes pasos:

**Preparación de datos**

Los datos necesarios para el estudio, que incluyen información sobre el rendimiento académico escolar y universitario, serán recopilados de las fuentes correspondientes. Estos datos serán organizados en conjuntos estructurados, generalmente en formato tabular, asegurando que cada registro contenga las variables necesarias (independientes y dependientes).

Antes de aplicar los algoritmos, se realiza una limpieza exhaustiva de los datos para manejar valores faltantes, duplicados o inconsistentes.

Los datos serán divididos en tres conjuntos principales:

- **Conjunto de entrenamiento**: Utilizado para ajustar los parámetros del modelo.
- **Conjunto de validación**: Empleado para evaluar el modelo durante el entrenamiento y evitar sobreajuste (*overfitting*).

**Determinación de la relación**

Se utilizarán las librerías de Scikit-Learn para procesar los datos con los algoritmos de Regresión Lineal y Árboles de Decisión, para las Redes Neuronales se utilizará Keras junto con Tensorflow. Ambas librerías están implementadas en Python. Se medirán las métricas $R^2$, MSE, RMSE, MAE.

# CAPÍTULO IV

# RESULTADOS

## 4.1. Preparación del conjunto de datos

### 4.1.1. Datos de rendimiento escolar

Se obtuvieron los certificados de los ingresantes del año 2023 a la Facultad de Ingeniería de la UNJBG. Para la obtención de los certificados, fue necesario acudir a las oficinas de la Dirección de Asuntos Académicos (DASA) de la universidad y presentar la solicitud correspondiente. Posteriormente, los datos de los 265 certificados fueron recopilados manualmente y transferidos a una hoja de cálculo para su análisis. En la Figura 5 se muestra el gráfico de barras con la cantidad de certificados obtenidos por escuela.

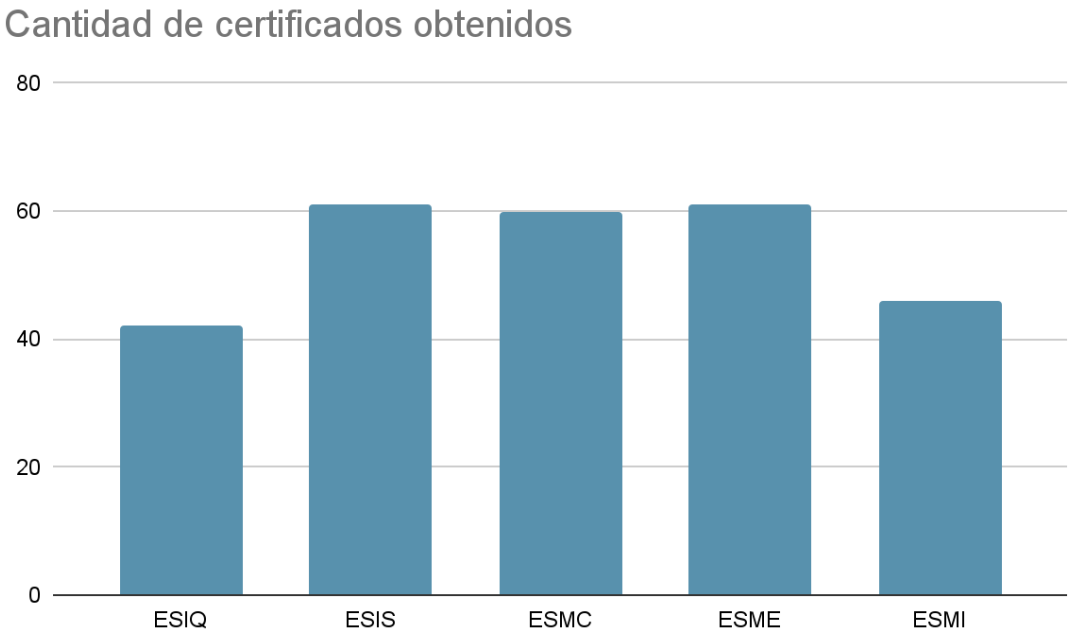
**Tabla 2.**

*Cantidad de certificados por escuela*

| CARRERAS | CERTIFICADOS |
|---|---|
| ESIQ | 41 |
| ESIS | 60 |
| ESMC | 59 |
| ESME | 60 |
| ESMI | 45 |
| TOTAL | 265 |

**Figura 5.**

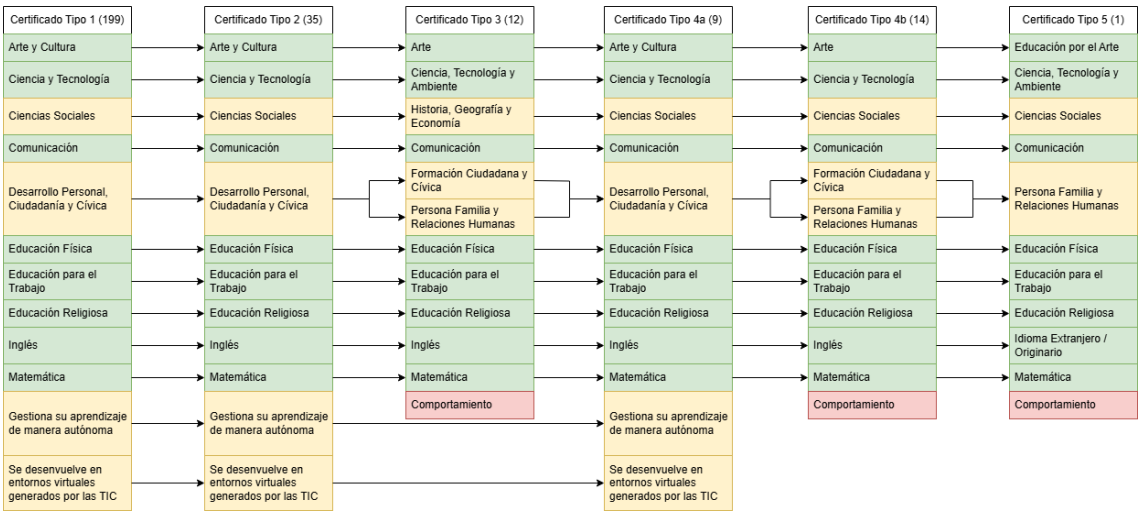
*Cantidad de certificados obtenidos*




En los certificados se identificaron seis tipos diferentes, cuya variación se detalla en secciones posteriores y depende tanto de la fecha de emisión como del formato, ya sea virtual o físico.

**Figura 6.**

*Tipos de certificados según campos*



Se optó por elegir las asignaturas en común que tenían los formatos, eliminando las que no estuvieran incluidas en gran parte de los certificados, quedando al final un grupo específico de columnas.

**Figura 7.**

*Equivalencia de cursos en el formato antiguo al formato moderno*


Se clasificaron los certificados en función de su formato y presentación. La clasificación detallada es la siguiente:

**Certificado Virtual - Formato Moderno - Con Separación por Curso (Tipo 1):**

Este formato es correspondiente a los años 2021 y 2022 y presenta una estructura en la que los cursos están organizados por áreas específicas.

Existen registros con campos vacíos o incompletos (por ejemplo, en las asignaturas de inglés y religión). Esta situación se debe a que los estudiantes fueron exonerados de dichos cursos durante su época escolar.

**Certificado Virtual - Formato Moderno - Sin Separación por Curso (Tipo 2):**

Este tipo de certificado corresponde a los años 2019 y 2020. Los cursos no están distribuidos en secciones por áreas.

También existen registros con campos vacíos.

**Certificado Virtual - Formato Antiguo - Sin Separación por Curso (Tipo 3):**

Este tipo de certificado corresponde a los años previos a 2018.

Certificados consistentes sin inconvenientes ni datos faltantes, todos con un formato uniforme.

**Certificado Físico - Formato Moderno - Sin Separación por Curso (Tipo 4):**

Este certificado cubre los cursos generales: Matemática, Comunicación, Inglés, Arte, Educación Física, Educación Religiosa, Educación para el Trabajo, Ciencias Sociales, Ciencia y Tecnología.

Existen dos subgrupos:

- Alumnos con cursos generales, calificación en DPCC, y dos competencias adicionales.
- Alumnos con cursos regulares, calificación en FCC y PFRH, con notas de comportamiento, pero sin competencias.

Observación en Tipo 4:

- Un estudiante presenta un certificado diferente: 2023-101052.

En los certificados se identificaron excepciones y anomalías, con siete registros eliminados debido a la falta de datos. Esto fue causado por factores como transferencia, educación alternativa y la ausencia de una carpeta física.

La clasificación de los certificados por año revela una clara transición de formato:

- Del 2021 al 2022: Todos los certificados registrados son Tipo 1 (Formato Moderno con Separación por Curso), con un total de 194 alumnos.
- Del 2019 al 2020: Predominan los certificados Tipo 2 y Tipo 4a, con dos excepciones que corresponden al formato antiguo (Tipo 4b). Se observó una discrepancia en dos registros: 2023-101043 y 2023-120052.
- 2018 y años anteriores: Se observan principalmente los certificados en formato antiguo, Tipo 3 o Tipo 4b, representando 25 alumnos.

**Tabla 3.**

*Resumen de certificados por tipo y año*

| Año | Tipo 1 (Moderno, Separado) | Tipo 2 (Moderno, No Separado) | Tipo 3 (Antiguo, No Separado) | Tipo 4a (Moderno) | Tipo 4b (Antiguo) | Tipo 5 (2008) | Total |
|---|---|---|---|---|---|---|---|
| **2021 - 2022** | 194 | 0 | 0 | 0 | 0 | 0 | 194 |
| **2019 - 2020** | 0 | 35 | 0 | 5 | 5 | 0 | 45 |
| **2018 y antes** | 0 | 0 | 12 | 0 | 13 | 1 | 26 |

**Promedio de notas de cursos por carrera**

En esta sección se presentan los resultados obtenidos del análisis de los promedios de notas de los estudiantes universitarios en sus cursos escolares. Se han elaborado diez gráficos de barras, uno para cada uno de los diez cursos tomados por los estudiantes durante su etapa escolar. En cada gráfico se muestran los promedios de notas de las cinco carreras universitarias estudiadas: ESME (Ingeniería Metalúrgica), ESMC (Ingeniería Mecánica), ESMI (Ingeniería Industrial), ESIQ (Ingeniería Química) y ESIS (Ingeniería de Sistemas).

Los gráficos permiten visualizar las diferencias y similitudes en el rendimiento académico de los estudiantes en función de su especialidad universitaria y proporcionan una perspectiva clara sobre las competencias adquiridas en cada curso durante su formación escolar.

**Arte**

En la Tabla 4 y en la Figura 8 se presentan los promedios obtenidos por los estudiantes de diferentes carreras de ingeniería en el curso de Artes durante su último año

escolar. Estos promedios reflejan el rendimiento académico en una materia de carácter general, proporcionando una visión inicial de las diferencias en los niveles de desempeño entre los futuros estudiantes universitarios. Las carreras incluidas son ESMC (Escuela de Ingeniería Mecánica), ESMI (Escuela de Ingeniería Industrial), ESME (Escuela de Ingeniería Eléctrica), ESIQ (Escuela de Ingeniería Química) y ESIS (Escuela de Ingeniería de Sistemas), cuyos valores promedio varían entre 14.70 y 16.30, indicando un rango de resultados heterogéneos entre los grupos evaluados.

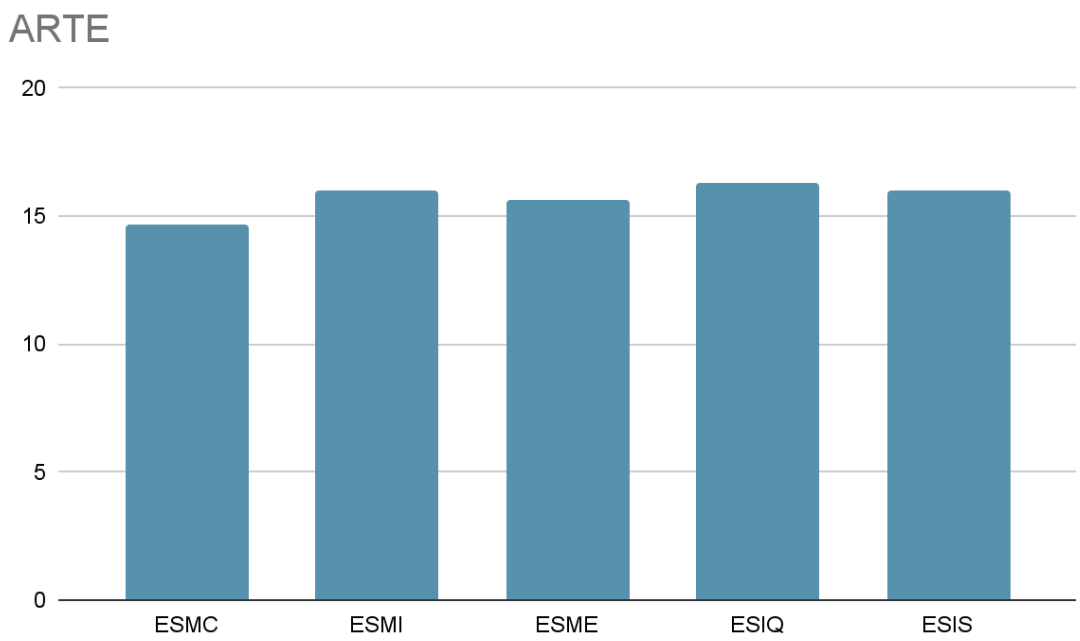
**Tabla 4.**

*Promedio de notas de arte por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.70192308 |
| ESMI | 16 |
| ESME | 15.6509434 |
| ESIQ | 16.3088235 |
| ESIS | 15.9727273 |

**Figura 8.**

*Histograma del promedio de notas de arte por carrera.*




**Ciencia y tecnología**

En la Tabla 5 y en la Figura 9 se presentan obtenidos por los estudiantes en el curso de Ciencia y Tecnología durante su último año escolar. Los resultados muestran un rango de promedios entre 14.25 y 15.65, lo que indica un desempeño relativamente homogéneo entre los grupos evaluados. Cabe destacar que los estudiantes de Ingeniería Química obtuvieron el promedio más alto (15.65), lo cual podría reflejar un interés o una preparación particular en áreas relacionadas con la ciencia y la tecnología.

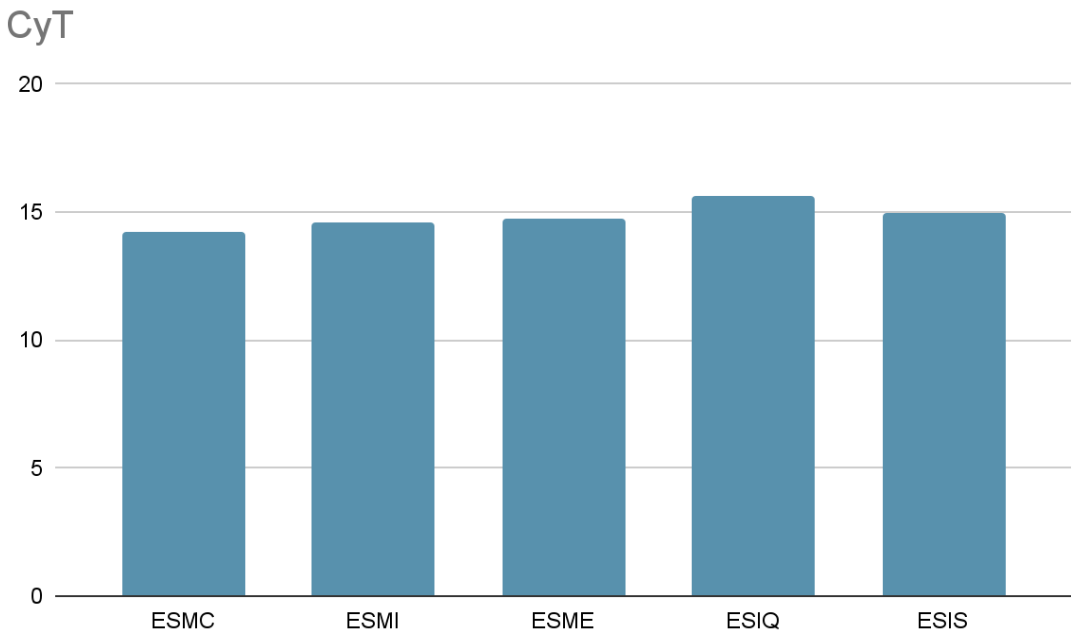
**Tabla 5.**

*Promedio de notas de ciencia y tecnología por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.25641026 |
| ESMI | 14.5833333 |
| ESME | 14.72327044 |
| ESIQ | 15.6568627 |
| ESIS | 14.9818182 |

**Figura 9.**

*Histograma del promedio de notas de ciencia y tecnología por carrera.*



**Ciencias Sociales**

En la Tabla 6 y en la Figura 10 se presentan los promedios obtenidos por los estudiantes en el curso de Ciencias Sociales durante su último año escolar. Este curso, orientado a desarrollar el análisis crítico y la comprensión de las dinámicas sociales, refleja un desempeño con cierta variación entre los diferentes grupos. Los promedios oscilan entre 14.56 y 15.87, destacando los estudiantes de Ingeniería Química como el grupo con el promedio más alto (15.87), mientras que los demás grupos mantienen valores cercanos, lo que sugiere un rendimiento relativamente uniforme en este ámbito.

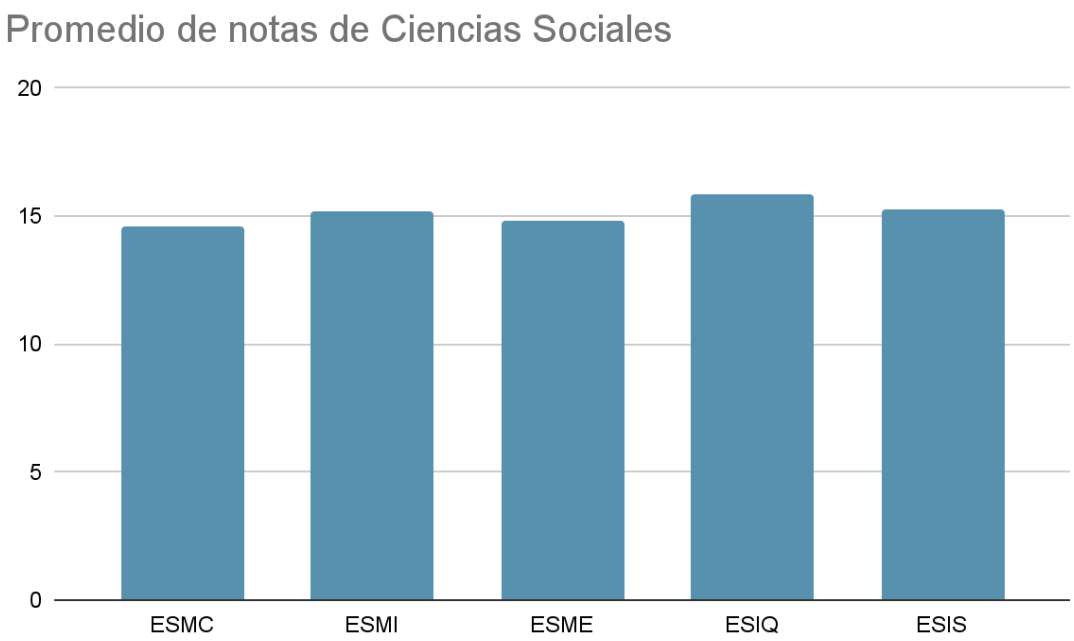
**Tabla 6.**

*Promedio de notas de ciencias sociales por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.56410256 |
| ESMI | 15.16666667 |
| ESME | 14.85534591 |
| ESIQ | 15.872549 |
| ESIS | 15.2848485 |

**Figura 10.**

*Histograma del promedio de notas de ciencias sociales por carrera.*




**Comunicación**

En la Tabla 7 y en la Figura 11 se presentan los promedios obtenidos por los estudiantes en el curso de Comunicación durante su último año escolar. Este curso, enfocado en el desarrollo de habilidades expresivas y comprensivas, evidencia un rango de promedios que varía entre 14.47 y 15.55. Los estudiantes de Ingeniería Química destacan nuevamente con el promedio más alto (15.55), mientras que los demás grupos muestran resultados cercanos entre sí, indicando un desempeño consistente en esta materia fundamental para el desarrollo académico y profesional.

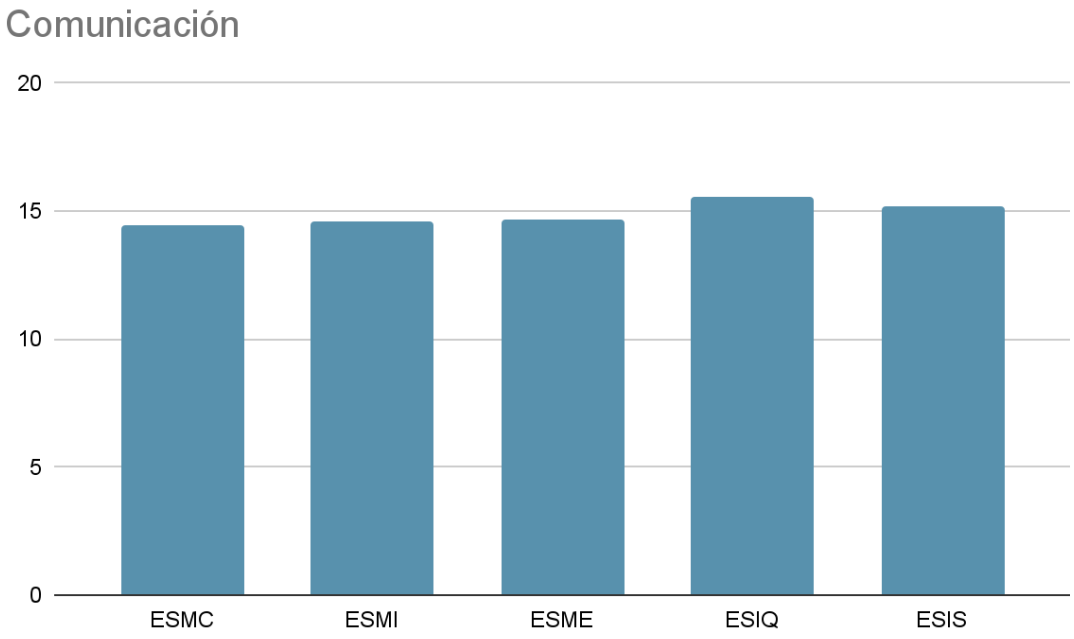
**Tabla 7.**

*Promedio de notas de comunicación por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.46794872 |
| ESMI | 14.63333333 |
| ESME | 14.66666667 |
| ESIQ | 15.5490196 |
| ESIS | 15.169697 |

**Figura 11.**

*Histograma del promedio de notas de comunicación por carrera*



**Desarrollo personal, ciudadanía y cívica**

En la Tabla 8 y en la Figura 12 se presentan los promedios obtenidos por los estudiantes en el curso de Desarrollo Personal, Ciudadanía y Cívica durante su último año escolar. Este curso, orientado a la formación en valores, ciudadanía activa y desarrollo integral, presenta un rango de promedios que varía entre 14.97 y 16.16. Destacan los estudiantes de Ingeniería Química con el promedio más alto (16.16), reflejando un desempeño sobresaliente en esta asignatura. Los demás grupos muestran resultados cercanos entre sí, lo que sugiere un nivel de rendimiento uniforme en esta materia de carácter formativo.

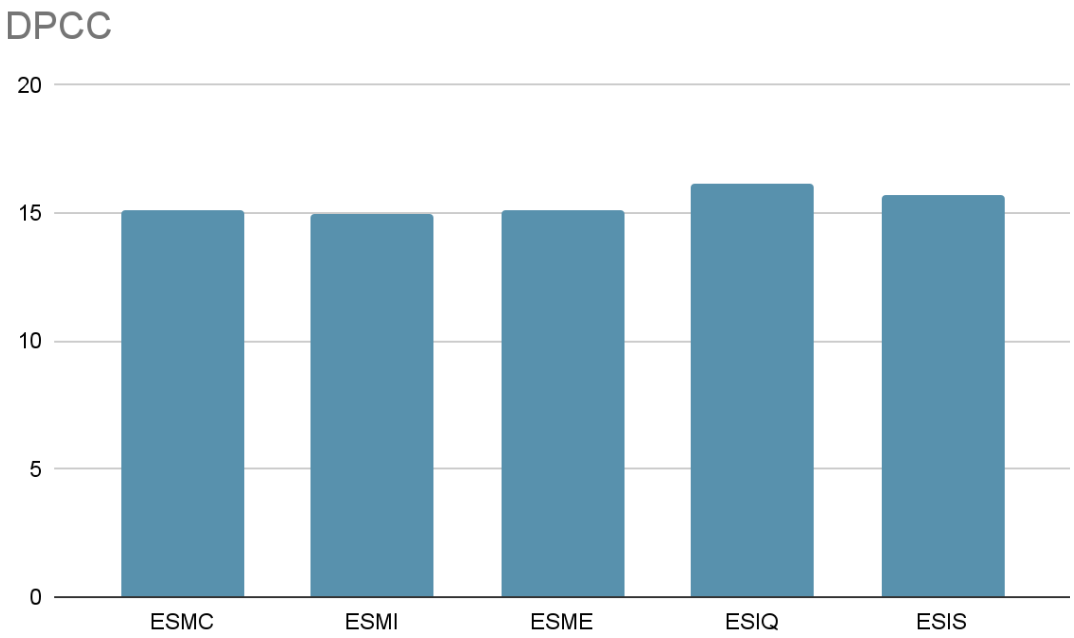
**Tabla 8.**

*Promedio de notas DPCC por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 15.09615385 |
| ESMI | 14.975 |
| ESME | 15.1320755 |
| ESIQ | 16.16176471 |
| ESIS | 15.71607491 |

**Figura 12.**

*Histograma del promedio de notas de por carrera*




**Educación física**

En la Tabla 9 y en la Figura 13 se presentan los promedios obtenidos por los estudiantes en el curso de Educación Física durante su último año escolar. Este curso, centrado en el fortalecimiento de la condición física y la promoción de hábitos saludables, muestra un rango de promedios que oscila entre 15.35 y 16.11. Los estudiantes de Ingeniería Industrial alcanzaron el promedio más alto (16.11), evidenciando un desempeño destacado en esta área, mientras que los demás grupos obtuvieron resultados cercanos, lo que sugiere un nivel de rendimiento uniforme en general.

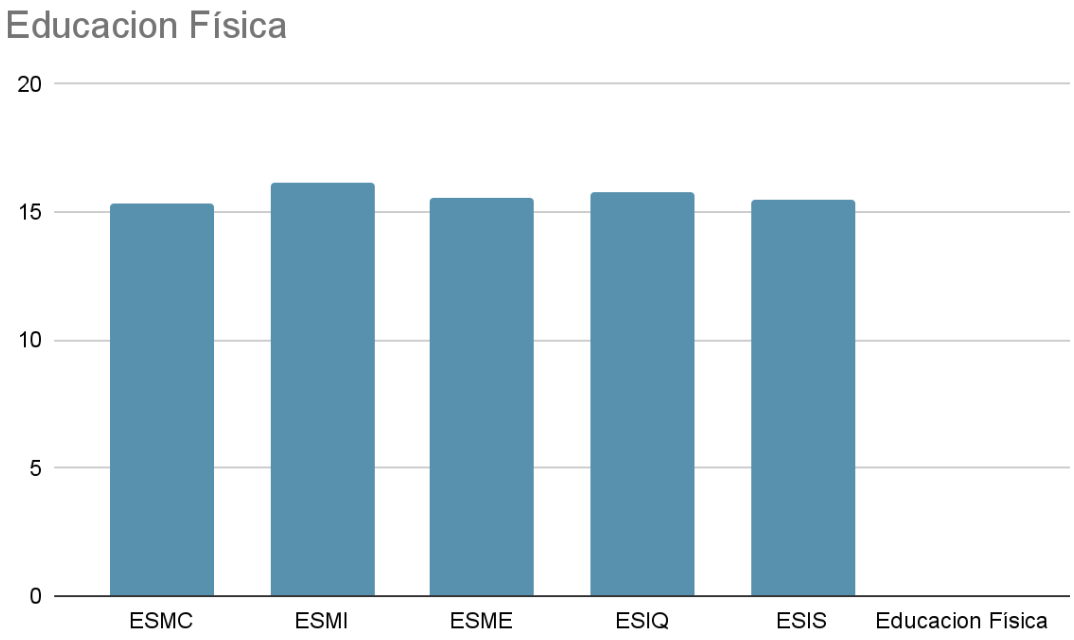
**Tabla 9.**

*Promedio de notas educación física por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 15.35897436 |
| ESMI | 16.11666667 |
| ESME | 15.55345912 |
| ESIQ | 15.7745098 |
| ESIS | 15.46060606 |

**Figura 13.**

*Histograma del promedio de notas de por carrera*



**Educación por el trabajo**

En la Tabla 10 y en la Figura 14 se presentan los promedios obtenidos por los estudiantes en el curso de Educación por el Trabajo durante su último año escolar. Este curso, orientado a desarrollar competencias técnicas y habilidades prácticas para el ámbito laboral, presenta un rango de promedios entre 15.21 y 16.17. Los estudiantes de Ingeniería Química obtuvieron el promedio más alto (16.17), destacándose en esta asignatura, mientras que los demás grupos presentan resultados relativamente uniformes, lo que refleja un desempeño sólido en esta materia de enfoque práctico.

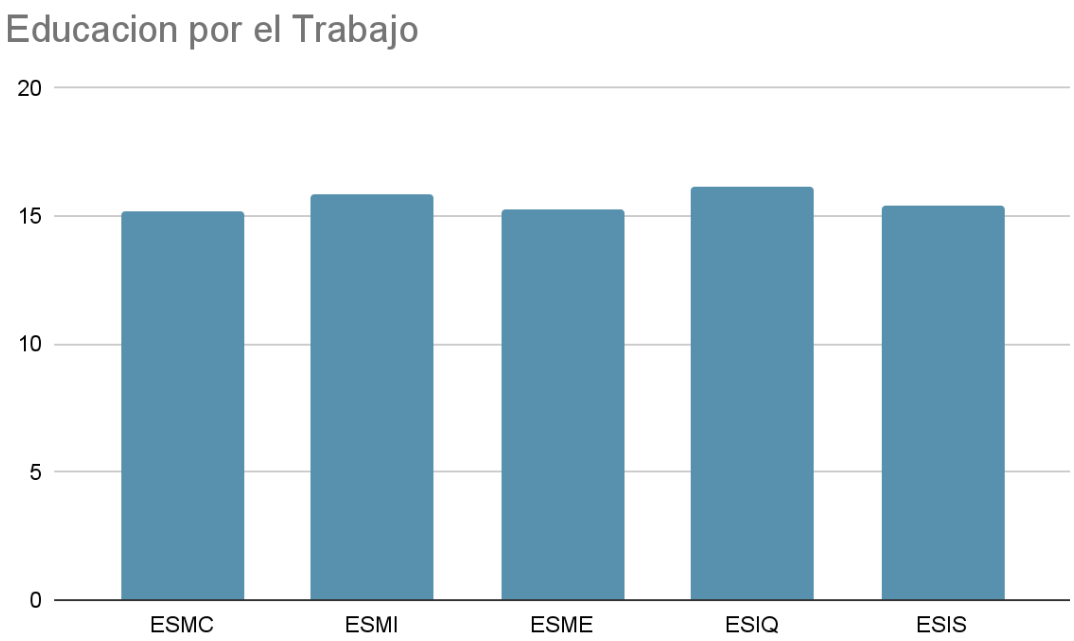
**Tabla 10.**

*Promedio de notas de educación por el trabajo por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 15.21153846 |
| ESMI | 15.85 |
| ESME | 15.22641509 |
| ESIQ | 16.17647059 |
| ESIS | 15.42754584 |

**Figura 14.**

*Histograma del promedio de notas de educación por el trabajo por carrera*




**Educación religiosa**

En la Tabla 11 y en la Figura 15 se presentan los promedios obtenidos por los estudiantes en el curso de Educación Religiosa durante su último año escolar. Este curso, enfocado en el desarrollo de valores espirituales y éticos, muestra un rango de promedios que varía entre 15.36 y 16.38. Los estudiantes de Ingeniería Química destacan con el promedio más alto (16.38), lo que sugiere un desempeño sobresaliente en esta asignatura. Los demás grupos también muestran resultados cercanos, lo que refleja un rendimiento homogéneo en esta materia de formación integral.

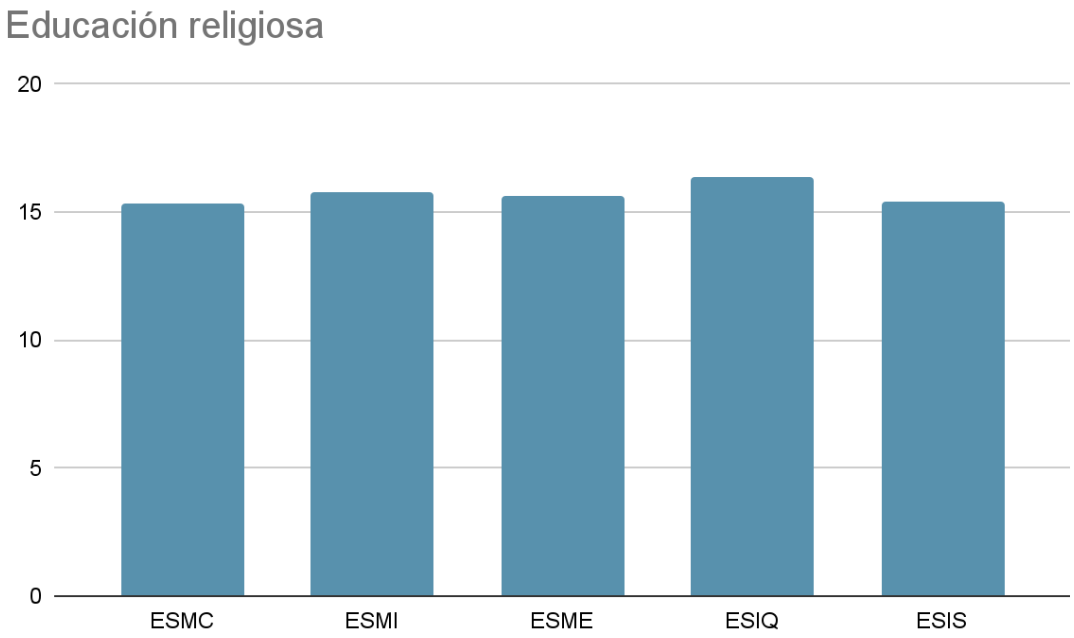
**Tabla 11.**

*Promedio de notas de educación religiosa por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 15.35897436 |
| ESMI | 15.8125 |
| ESME | 15.6132075 |
| ESIQ | 16.38338948 |
| ESIS | 15.42138566 |

**Figura 15.**

*Histograma del promedio de notas de educación religiosa por carrera*



**Ingles**

El cuadro muestra los promedios obtenidos por los estudiantes en el curso de Inglés durante su último año escolar. Este curso, clave para el desarrollo de habilidades lingüísticas y comunicativas, presenta un rango de promedios que varía entre 14.62 y 15.68. Los estudiantes de Ingeniería de Sistemas obtuvieron el promedio más alto (15.68), mientras que los demás grupos muestran resultados similares, con una ligera variación en los puntajes. Esto refleja un rendimiento relativamente uniforme entre los estudiantes en esta materia.

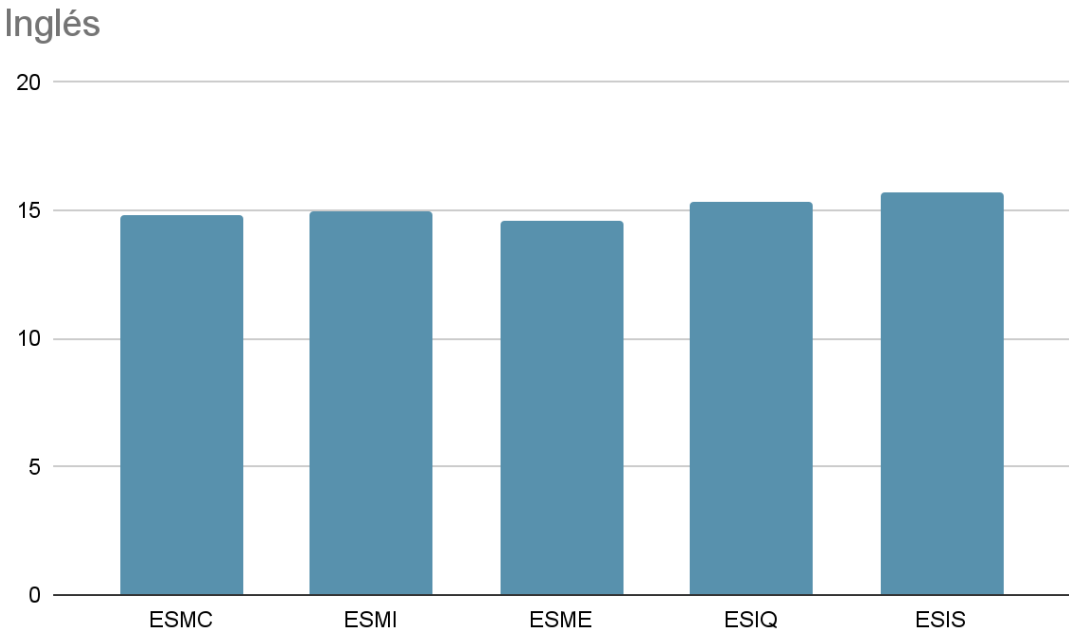
**Tabla 12.**

*Promedio de notas de inglés por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.84124209 |
| ESMI | 14.97694805 |
| ESME | 14.6178224 |
| ESIQ | 15.30392157 |
| ESIS | 15.67878788 |

**Figura 16.**

*Histograma del promedio de notas de inglés por carrera*




**Matemática**

En la Tabla 13 y en la Figura 17 se presentan los promedios obtenidos por los estudiantes en el curso de Matemática durante su último año escolar. Este curso, esencial para el fortalecimiento de las habilidades lógicas y analíticas, presenta una variabilidad de promedios entre 14.64 y 15.59. Los estudiantes de Ingeniería de Sistemas destacaron con el promedio más alto (15.59), mientras que los demás grupos obtuvieron resultados relativamente cercanos, lo que sugiere un rendimiento homogéneo en esta asignatura.

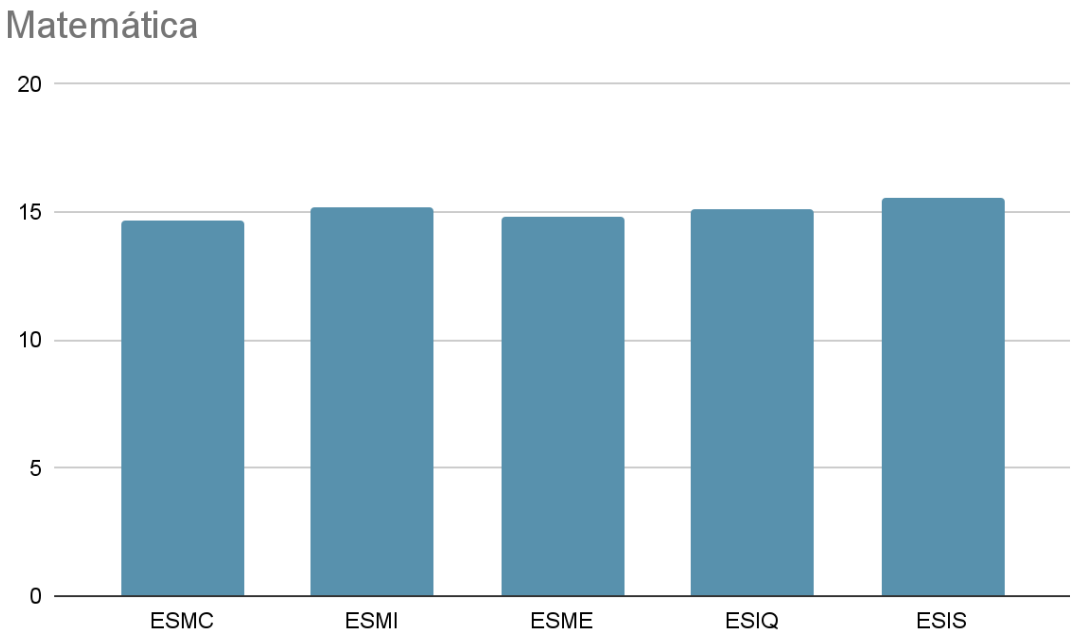
**Tabla 13.**

*Promedio de notas de matemática por carrera*

| Carrera | Promedio |
|---|---|
| ESMC | 14.64423077 |
| ESMI | 15.225 |
| ESME | 14.80188679 |
| ESIQ | 15.125 |
| ESIS | 15.59090909 |

**Figura 17.**

*Histograma del promedio de notas de matemática por carrera*



### 4.1.2. Datos del rendimiento universitario.

Se recopilaron un total de 300 registros de notas correspondientes a estudiantes de diversas carreras de ingeniería. Estos registros fueron proporcionados por la oficina de Dirección de Asuntos Académicos (DASA) y reflejan las calificaciones obtenidas por los estudiantes en su primer año. Cabe mencionar que no se lograron obtener todos los registros posibles, ya que algunos estudiantes se retiraron y no completaron ambos semestres.

**Notas de cada escuela**

Los resultados presentados en la Figura 18 y en la Figura 19 reflejan los promedios de las notas obtenidas por los estudiantes universitarios durante sus dos primeros semestres. Estos datos permiten analizar el rendimiento académico inicial de los estudiantes en la universidad, proporcionando una base comparativa con sus notas del último año de colegio.

**Figura 18.**

*Promedio de notas en el primer semestre*

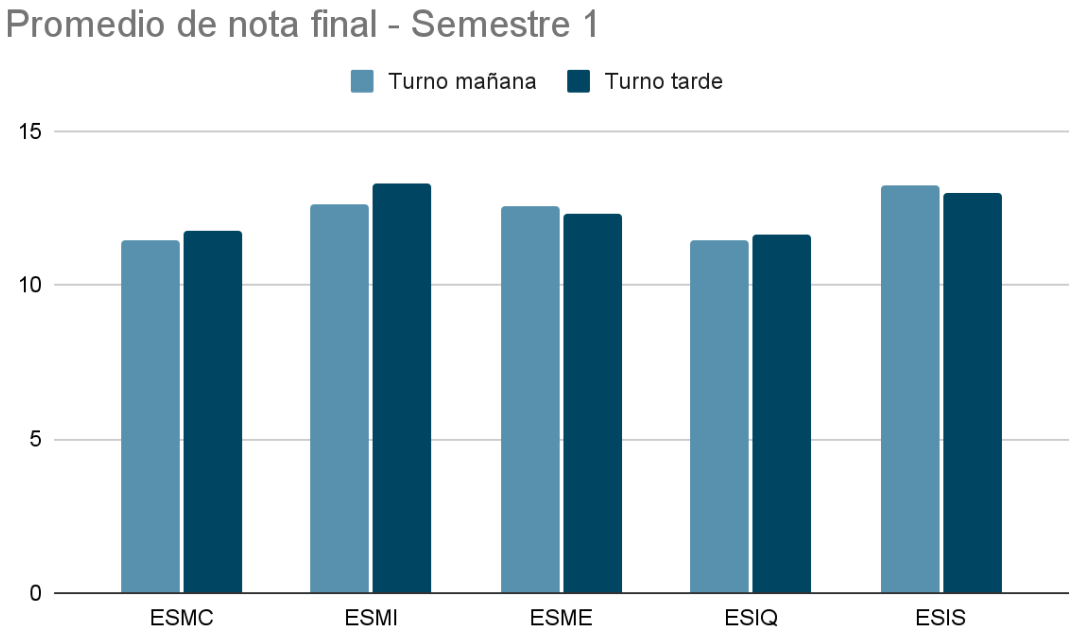
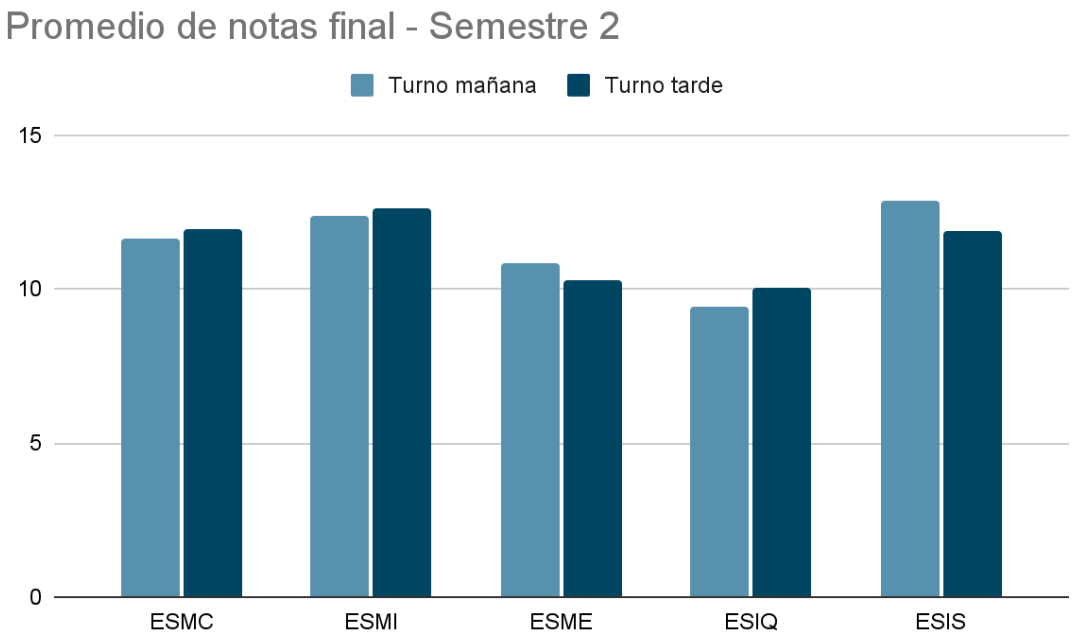


**Figura 19.**

*Promedio de notas en el segundo semestre*

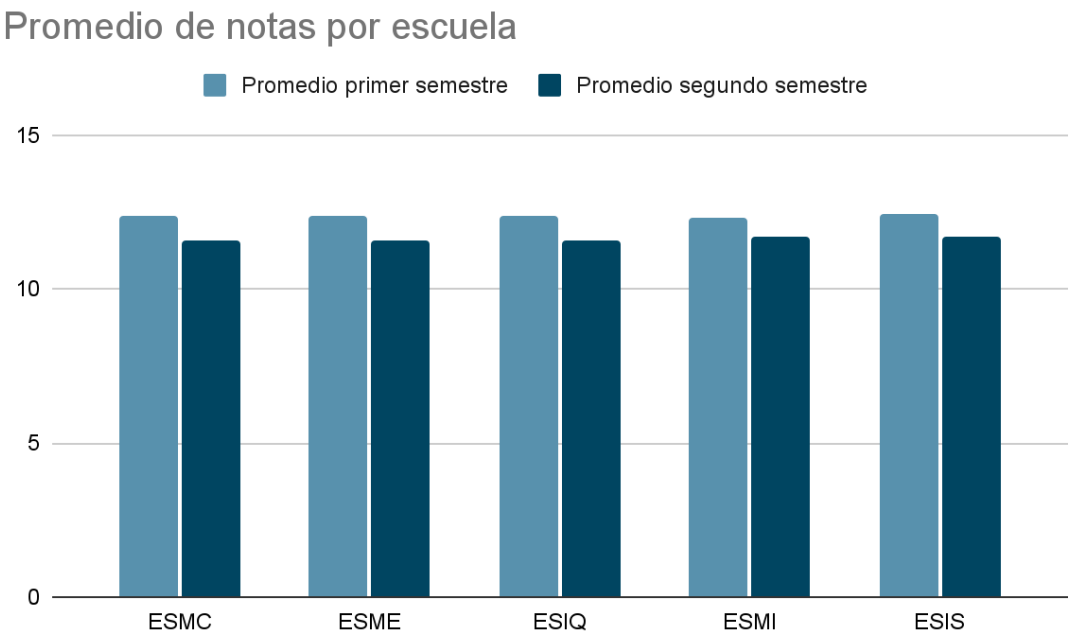



En el primer semestre, los estudiantes del turno de la tarde muestran un rendimiento ligeramente superior en comparación con los del turno de la mañana en la mayoría de las carreras. Este fenómeno podría atribuirse a varios factores, como diferencias en la disponibilidad de recursos académicos, la concentración de los estudiantes en diferentes momentos del día o las variaciones en la carga de trabajo personal.

En el segundo semestre, se observa una tendencia a la baja en los promedios de notas en ambos turnos, aunque esta disminución es más pronunciada en algunas carreras que en otras. Los estudiantes del turno de la mañana presentan una mayor disminución en sus promedios en comparación con los del turno de la tarde en ciertas carreras. Esto podría indicar que los estudiantes del turno de la tarde han desarrollado mejores estrategias de adaptación y manejo de la carga académica a lo largo del tiempo

**Figura 20.**

*Promedio de notas por cada escuela*




El análisis de los promedios de notas de los estudiantes de las cinco carreras de ingeniería revela varias tendencias significativas. En el primer semestre, los promedios de todas las carreras son bastante cercanos, lo que sugiere una homogeneidad inicial en el rendimiento académico entre los distintos programas. Sin embargo, en el segundo semestre se observa una tendencia general a la baja en los promedios de todas las carreras.

Esta disminución en el rendimiento académico podría atribuirse a varios factores. Una posible explicación es el aumento de la carga académica y la complejidad de los cursos en el segundo semestre, lo que podría afectar negativamente las calificaciones de los estudiantes. Además, el proceso de adaptación a la vida universitaria y a las demandas específicas de cada programa de ingeniería puede influir en el rendimiento de los estudiantes en esta etapa inicial de sus estudios.

Aunque los promedios generales disminuyen en el segundo semestre, el hecho de que las diferencias entre las carreras se mantengan relativamente pequeñas sugiere que estas variaciones pueden ser parte de un fenómeno más amplio que afecta a todos los estudiantes por igual, independientemente de la carrera específica. Esta observación

subraya la importancia de implementar estrategias de apoyo académico y personalizadas desde los primeros semestres para ayudar a los estudiantes a superar estas dificultades y mejorar su rendimiento académico a lo largo del tiempo.

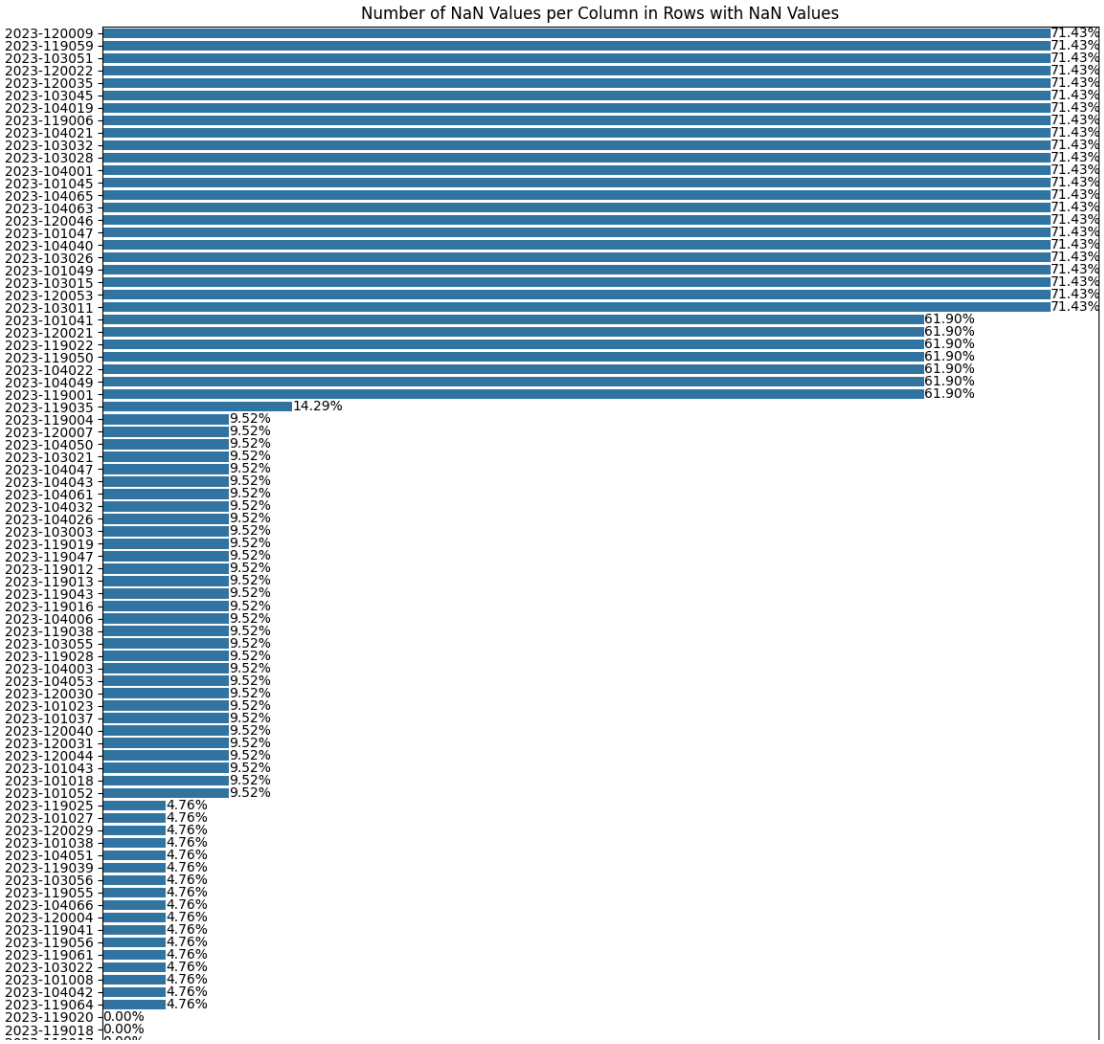
### 4.1.3. Preparación del Dataset

**Visualización de Valores Faltantes por Fila**

Para comprender mejor la distribución de los valores faltantes, se generó un gráfico de barras. En la Figura 21, cada barra representa un registro (identificado con su código único), y la longitud de la barra indica el porcentaje de datos faltantes. Visualizar los datos de esta forma ayuda a identificar rápidamente los registros con niveles críticos de falta de datos, haciendo evidente cuáles registros pueden necesitar una limpieza o tratamiento especial.

**Figura 21.**

*Porcentaje de datos faltantes en registros de estudiante*




## Selección de Filas con Valores Faltantes en un Nivel Aceptable

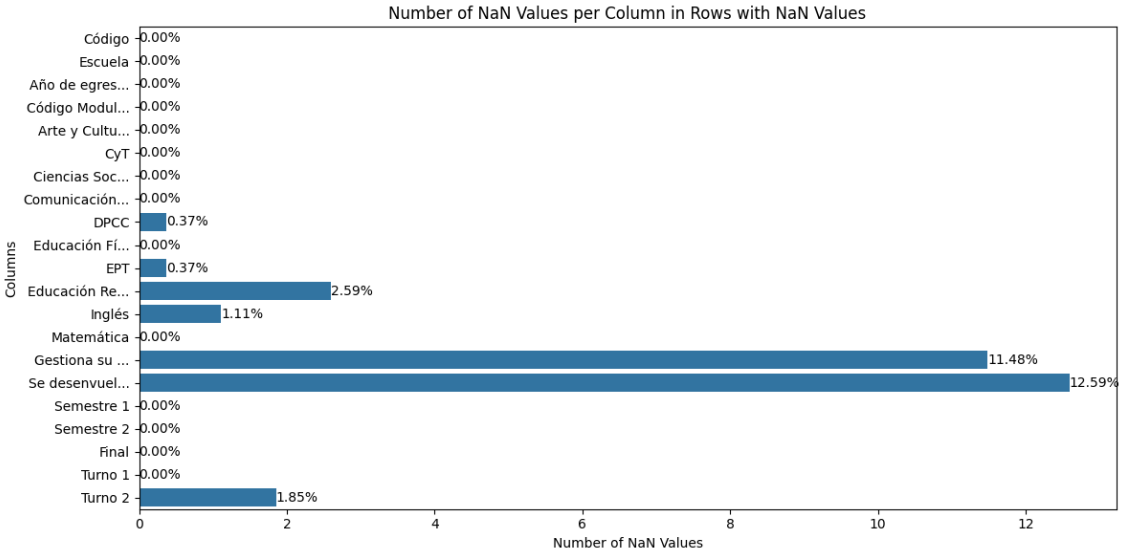
Dado que algunos registros mostraban un porcentaje alto de datos ausentes, se estableció un umbral del 20% como límite para definir qué registros serían útiles para análisis posteriores. Los registros que superaran ese porcentaje de datos faltantes fueron descartados, quedando así un subconjunto de datos más completo y confiable. Esta estrategia de filtrado asegura que los registros seleccionados sean representativos y cuenten con suficiente información para el análisis sin comprometer la precisión de los resultados.

**Visualización de Valores Faltantes por Columna en Filas Seleccionadas**

Una vez filtrados los registros, se generó un segundo gráfico de barras que muestra el número de valores faltantes en cada columna dentro del conjunto filtrado. Esta visualización permite identificar qué columnas, incluso en los registros menos afectados, tienen mayor tendencia a contener datos ausentes. Esta información es valiosa para decidir si se debe imputar o corregir estos datos o, en su defecto, si algunas columnas debiesen eliminarse por tener información insuficiente.

**Figura 22.**

*Porcentaje de valores faltantes de cada columna*




**Selección de Columnas con Valores Faltantes en un Nivel Aceptable**

Dado que algunos registros mostraban un porcentaje alto de datos ausentes, se estableció un umbral del 10% como límite para definir qué registros serían útiles para análisis posteriores. Los registros con más del 10% de datos faltantes fueron descartados, quedando así un subconjunto de datos más completo y confiable. Esta estrategia de filtrado asegura que los registros seleccionados sean representativos y cuenten con suficiente información para el análisis sin comprometer la precisión de los resultados.

**Relleno de Valores Faltantes con la Media**

Se creó una copia del dataset, donde los valores faltantes de las columnas seleccionadas se reemplazaron con la media de cada columna.

Este enfoque de imputación es útil para mantener la cantidad de registros sin introducir sesgos significativos, dado que se utiliza la media como estimación para los valores ausentes.

Al finalizar con la elaboración del dataset, se creó el dataframe de complete_df,

**Tabla 14.**

*Cantidad de registros de notas utilizados por carrera*

| Carrera | Cantidad |
|---|---|
| Ingeniería Mecánica | 52 |
| Ingeniería de Minas | 40 |
| Ingeniería Metalúrgica | 53 |
| Ingeniería Química | 34 |
| Ingeniería en Informática y Sistemas | 55 |
| TOTAL | 234 |

**Análisis de Correlación entre Variables Seleccionadas**

Este proceso se centra en seleccionar columnas numéricas para analizar correlaciones con la variable de calificación final y en rellenar valores faltantes para mejorar la completitud de los datos.

**Selección de Columnas Numéricas para el Análisis**

1. **Identificación de Columnas Numéricas:**
    - Se seleccionaron todas las columnas numéricas del DataFrame para enfocarse únicamente en aquellas que representan variables cuantitativas. Esta selección excluye las columnas de texto o categóricas, como los

códigos de estudiantes, ya que no son relevantes para el análisis de correlación.

2. **Exclusión de Variables Específicas:**
    - Se excluyeron ciertas columnas numéricas que no se desean considerar en el análisis, como Semestre 1, Semestre 2 y Final. Esta exclusión evita que las variables semestrales y la calificación final (variable dependiente) interfieran en el análisis de correlación de otras variables predictoras.

**Análisis de Correlación con la Variable "Final"**

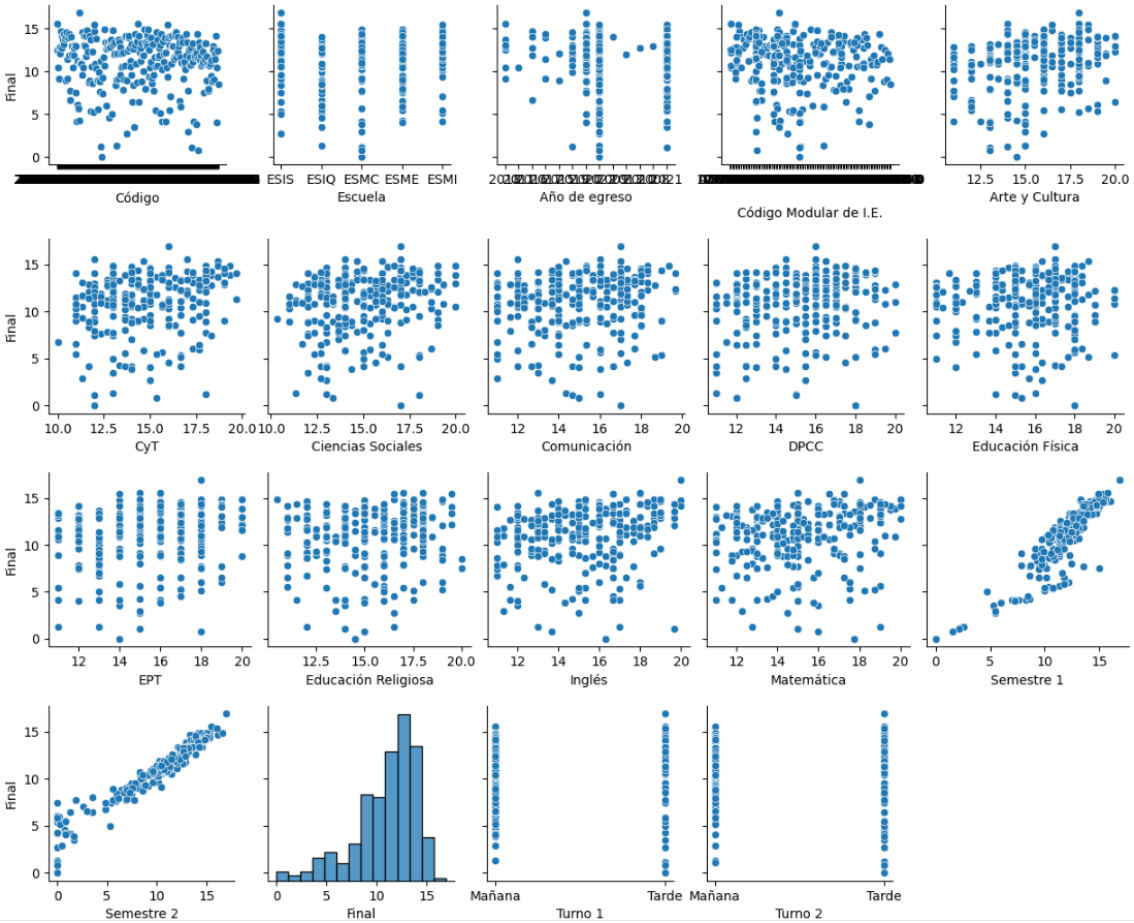
1. **Cálculo de Correlación:**
    - Se calcularon los coeficientes de correlación entre cada columna numérica y la calificación final ("Final"), ordenándolos de mayor a menor. Esto permite identificar qué variables están más fuertemente asociadas con el rendimiento final.
2. **Visualización de Correlación en Grupos de Variables:**
    - Para observar mejor estas relaciones, se generaron gráficos de dispersión (pairplots) entre cada grupo de cinco columnas y la variable "Final". Esta agrupación facilita una comparación gradual entre las diferentes variables predictoras y su relación con el rendimiento final.

**Análisis de Correlación entre Todas las Variables Numéricas**

1. **Cálculo de la Matriz de Correlación Completa:**
    - Se calculó una matriz de correlación para todas las columnas numéricas seleccionadas. Esta matriz muestra cómo cada variable se relaciona con las demás, proporcionando una visión más completa de las posibles interdependencias entre variables.

**Figura 23.**

*Gráfica de la relación de las variables independientes con la nota final*

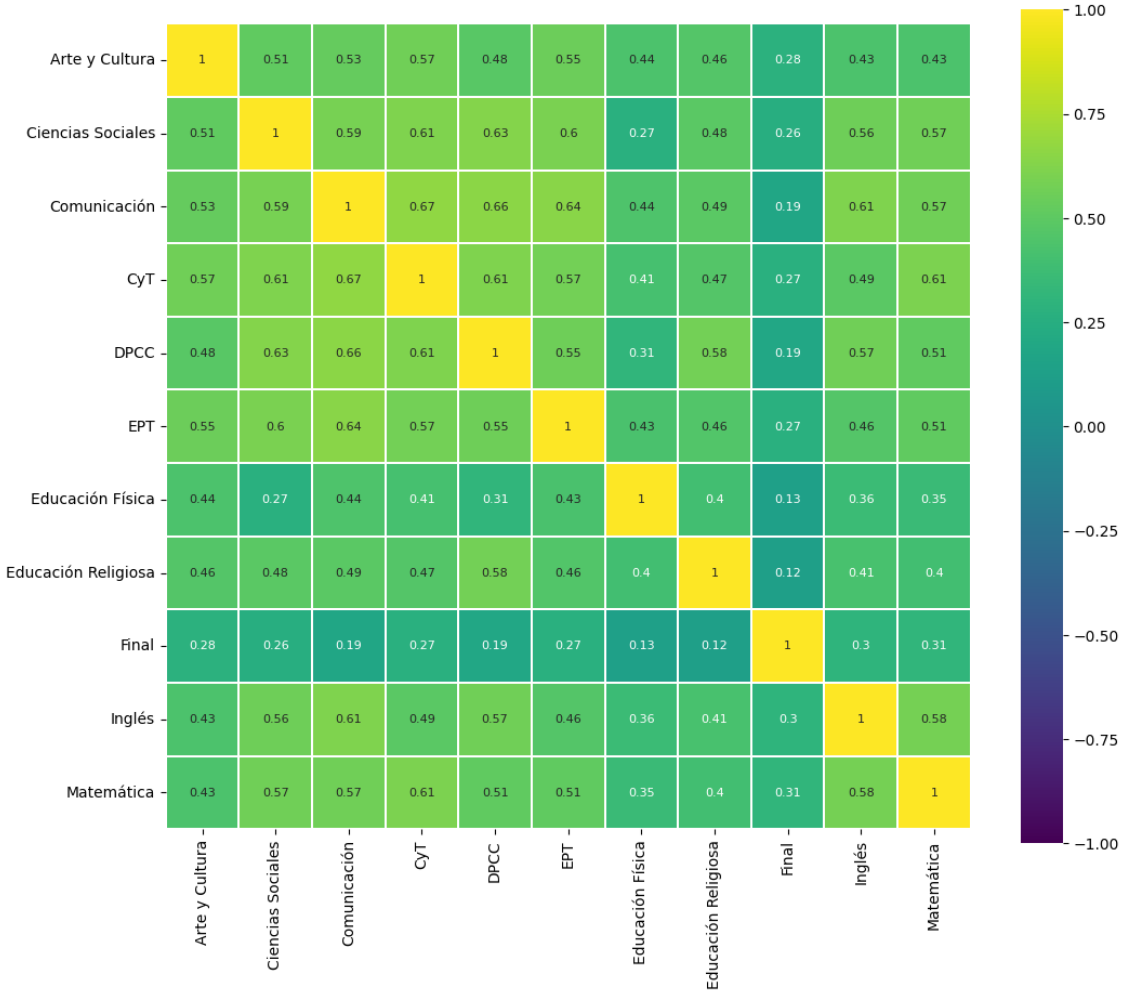



2. **Visualización de la Matriz de Correlación con un Heatmap:**

    - La matriz de correlación se visualizó usando un mapa de calor (heatmap) tal como se puede ver en la Figura 24. En este gráfico, los colores representan la intensidad de la correlación, y cada celda contiene el valor del coeficiente de correlación. Esto facilita la identificación de relaciones fuertes o débiles entre variables de una manera visualmente intuitiva.

**Figura 24.**

*Matriz de correlación*




## Creación del archivo CSV

Finalizado el análisis exploratorio de datos en conjunto con la construcción del dataframe finalizado, se creó un archivo CSV con dichos contenidos para la aplicación de los modelos de regresión.

## Conclusión

Este proceso de selección, imputación y análisis de correlación permite una exploración completa de las relaciones entre las variables cuantitativas y la calificación final. La visualización a través de pairplots y heatmaps facilita una comprensión profunda de cómo se comportan las variables numéricas y cuál podría ser su impacto en el

rendimiento final. Estos resultados serán útiles para identificar variables clave y potenciales factores predictivos que influyen en el desempeño estudiantil.

## 4.2. Determinación de la relación usando Regresión Lineal

Se presenta el uso de un modelo de regresión lineal para analizar la relación entre múltiples variables predictoras y la calificación final de los estudiantes. Este proceso permite explorar el grado en el cual las variables predictoras pueden estimar el rendimiento académico final de los estudiantes, proporcionando una base para análisis de regresión más avanzados.

### Carga y Exploración de los Datos

En primera instancia, se procedió a cargar el archivo complete_df.csv en un DataFrame utilizando la biblioteca pandas. A continuación, se revisaron las primeras filas y columnas del conjunto de datos con el objetivo de comprender su estructura básica y verificar que se encontraba en un formato adecuado para su análisis posterior.

### Selección de Variables

Se realizó una selección de las variables más relevantes para la predicción. En este proceso, se excluyeron aquellas que no contribuyen directamente, tales como Código, Año de egreso, Código Modular de I.E., Semestre 1, Semestre 2 y Turno 2. Estas variables fueron eliminadas de las predictoras (X). Por otro lado, se definió la columna Final como la variable objetivo (y), dado que esta representa la calificación final que el modelo deberá predecir.

### División de los Datos en Conjuntos de Entrenamiento y Prueba

Para entrenar y evaluar el modelo, se dividió el dataset en dos subconjuntos: un conjunto de entrenamiento, que corresponde al 70% de los datos, y un conjunto de prueba, correspondiente al 30%. Esta partición se realizó utilizando la función train_test_split de la biblioteca sklearn. Durante este proceso, se incluyeron las variables Escuela y Turno 1 como estratificadores para garantizar una representación equilibrada en ambos conjuntos.

Posteriormente, las columnas Escuela y Turno 1 fueron eliminadas de los conjuntos de entrenamiento y prueba. Esta eliminación tuvo como objetivo evitar que dichas variables influyan directamente en el desempeño del modelo, asegurando así que el proceso de predicción se basara únicamente en las variables seleccionadas previamente como relevantes.

### 4.2.1. Regresión Ridge

En esta sección, se explora el uso de la regresión Ridge, un modelo de regresión lineal que incluye una regularización para evitar el sobreajuste y mejorar la capacidad de generalización del modelo.

**Creación y Ajuste del Modelo Ridge:**

Se creó un modelo de regresión Ridge y se ajustó a los datos de entrenamiento (x_train y y_train). A diferencia de la regresión lineal simple, el modelo Ridge aplica una penalización a los coeficientes del modelo, lo cual ayuda a reducir la varianza en las predicciones y a mejorar su rendimiento en datos nuevos.

**Predicción de Calificaciones:**

Con el modelo entrenado, se generaron predicciones tanto para el conjunto de prueba (y_pred_ridge) como para el conjunto de entrenamiento (t_pred_ridge), lo cual permite evaluar el rendimiento del modelo en ambos contextos.

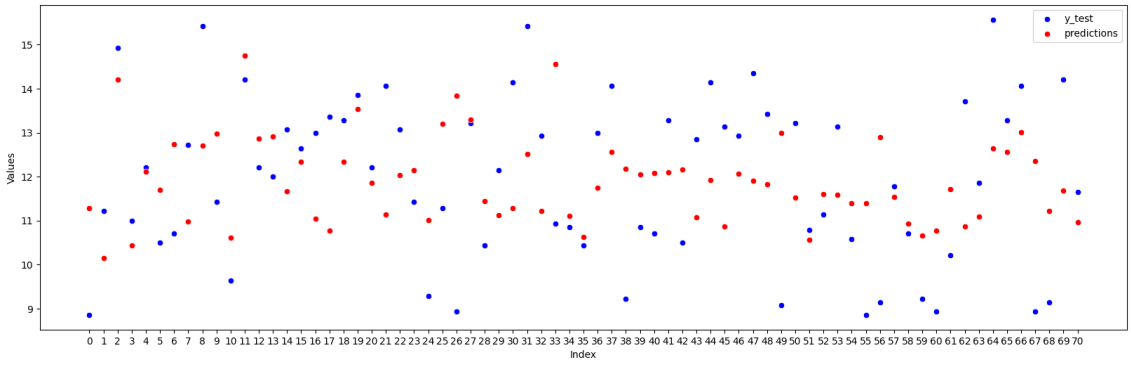
**Gráfico de Dispersión y Líneas de Predicción:**

Se generó un gráfico de dispersión que muestra tanto las calificaciones reales (y_test) como las predicciones (y_pred_ridge) tal como se puede ver en la Figura 25.

Además, se añadió una línea de tendencia a cada conjunto para visualizar cómo se ajustan las predicciones a los datos reales

**Figura 25.**

*Gráfico de dispersión de calificaciones con el modelo Bridge*




Esta comparación visual ayuda a ver si el modelo Ridge logra capturar la tendencia general de las calificaciones y observar cómo varían las predicciones respecto a los valores reales.

**Figura 26.**

*Gráfico de líneas de predicción con el modelo Ridge*

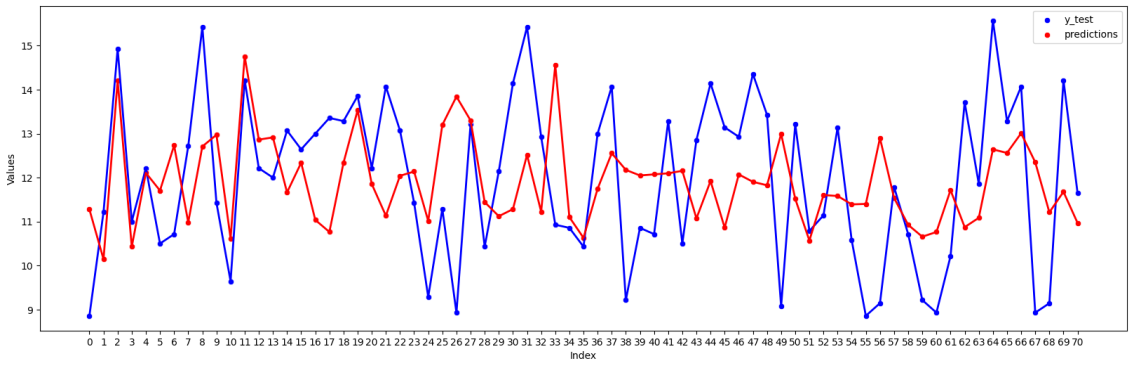
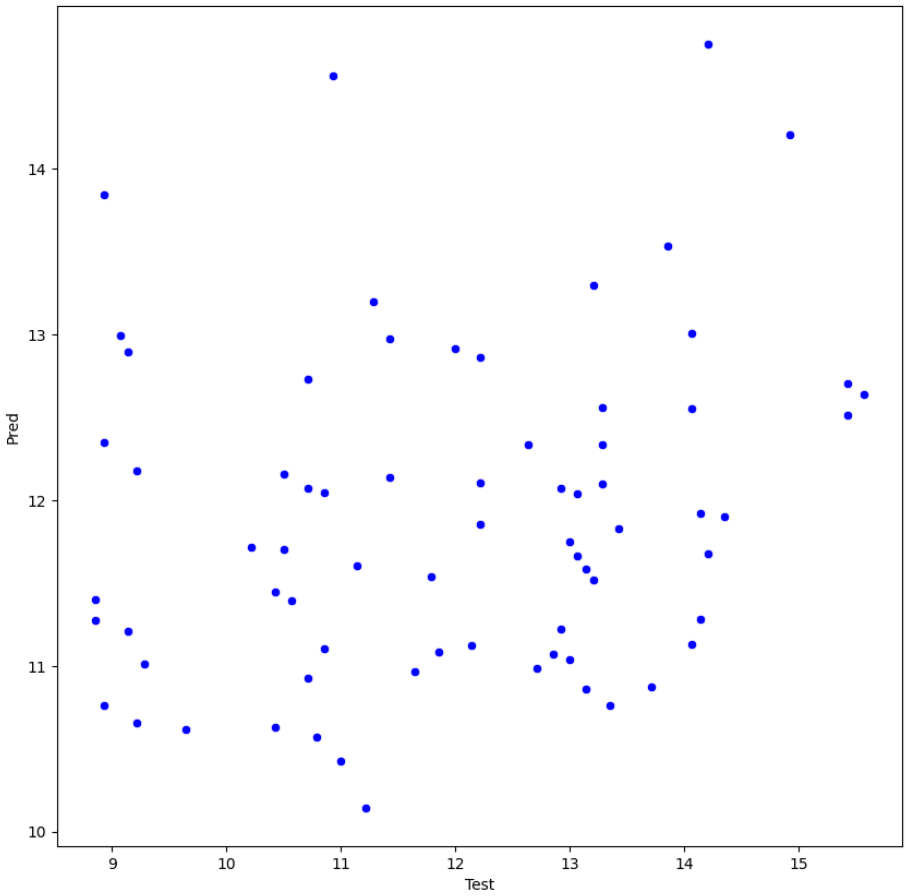



**Gráfico de Calificaciones Predichas vs. Reales:**

Se creó otro gráfico de dispersión, donde se comparan directamente los valores reales (y_test) con los predichos (y_pred_ridge, como se aprecia en la Figura 27. La proximidad de los puntos a la línea de identidad (y = x) indica qué tan bien se ajusta el modelo a las calificaciones reales.

**Figura 27.**

*Gráfico de dispersión de valores reales y predichos*




**Análisis de los Coeficientes del Modelo**

Se procedió a analizar los coeficientes obtenidos para cada variable predictora en el modelo Ridge. Estos coeficientes ofrecen una visión sobre la influencia relativa de cada variable en la calificación final, permitiendo identificar cuáles tienen mayor peso en las predicciones. En el caso de la regresión Ridge, los coeficientes suelen ser de menor magnitud en comparación con la regresión lineal estándar. Esto se debe al término de penalización introducido por Ridge, el cual regula los coeficientes para reducir la sensibilidad del modelo a las fluctuaciones en los datos. Este enfoque no solo mejora la estabilidad del modelo, sino que también facilita la identificación de las variables más significativas de manera más robusta.

**Intercepto del Modelo**

El análisis del intercepto del modelo Ridge reveló el valor base de la calificación final estimada cuando todas las variables predictoras tienen un valor de cero. Este valor representa un punto de referencia inicial sobre el cual se construyen las predicciones del modelo. Si bien este término puede no tener un significado directo en todos los contextos, proporciona una línea base útil para interpretar las salidas del modelo y evaluar cómo las variables predictoras ajustan las calificaciones en relación a esta referencia.

**Tabla 15.**

*Intercepto de las variables*

| **Variable** | **Coeficiente** |
|---|---|
| Arte y Cultura | 0.10282690887722312 |
| CyT (Ciencia y Tecnología) | 0.08963923883680835 |
| Ciencias Sociales | -0.03894562244082379 |
| Comunicación | -0.2756389607687126 |
| DPCC (Desarrollo Personal, Ciudadanía y Cívica) | 0.024980170744375446 |
| Educación Física | 0.07405626952956637 |
| EPT (Educación para el Trabajo) | 0.20769068572004748 |
| Educación Religiosa | -0.1069988877985563 |
| Inglés | 0.1978120379633508 |
| Matemática | 0.15541386943792465 |
| **Intercepto** | 5.285255384624042 |

A continuación, se realiza un análisis detallado de los coeficientes obtenidos para cada variable predictora en el modelo Ridge. Los valores de los coeficientes indican la magnitud y la dirección de la influencia de cada asignatura sobre la calificación final.

**Variables con Coeficientes Positivos**

a) **Arte y Cultura (0.1028):**

Tiene un impacto positivo moderado en la calificación final. Aunque no es el factor más influyente, contribuye al aumento de las predicciones de calificación.

b) **CyT (Ciencia y Tecnología) (0.0896):**

Presenta una influencia positiva leve en la calificación. Esto sugiere que un desempeño favorable en esta área contribuye ligeramente al incremento de la nota final.

c) **DPCC (Desarrollo Personal, Ciudadanía y Cívica) (0.0249):**

Su impacto es positivo pero muy pequeño, lo que indica una contribución marginal a la calificación final.

d) **Educación Física (0.0741):**

Tiene una influencia positiva moderada en la calificación final, pero no es de las variables más significativas.

e) **EPT (Educación para el Trabajo) (0.2077):**

Es la variable con el coeficiente más alto, lo que sugiere que tiene la mayor influencia positiva en la calificación final. Esto indica que el desempeño en esta área puede ser un fuerte indicador del éxito general.

f) **Inglés (0.1978):**

Al igual que EPT, tiene un impacto significativo positivo en las calificaciones finales, destacándose como una de las variables predictoras más relevantes.

g) **Matemática (0.1554):**

También tiene un coeficiente positivo considerable, lo que resalta su importancia en la predicción de la calificación final.

**Variables con Coeficientes Negativos**

a) **Ciencias Sociales (-0.0389):**

   Su coeficiente negativo indica una influencia leve pero adversa en la calificación final. Este impacto es menor en comparación con otras variables.

b) **Comunicación (-0.2756):**

   Este coeficiente negativo es el más pronunciado, indicando que el desempeño en esta área tiene el mayor impacto adverso en la calificación final.

c) **Educación Religiosa (-0.1070):**

   También tiene un impacto negativo en la calificación final, aunque menos significativo que Comunicación.

**Valor de Intercepto**

- **Intercepto (5.2853):**

  Este valor representa la calificación base estimada cuando todas las variables predictoras tienen un valor de cero. Sirve como referencia inicial para las predicciones del modelo y puede interpretarse como el nivel general de desempeño esperado en ausencia de contribuciones específicas de las variables predictoras.

## Interpretaciones Clave

- **EPT (Educación para el Trabajo)**, **Inglés** y **Matemática** son los predictores más influyentes positivamente, sugiriendo que el desempeño en estas áreas está altamente correlacionado con mejores calificaciones finales.
- **Comunicación** tiene el impacto negativo más fuerte, lo que podría indicar desafíos en esta área que afectan el desempeño global.
- La mayoría de las variables tienen coeficientes positivos, lo que indica que las calificaciones en estas asignaturas generalmente contribuyen a mejorar el resultado final. Sin embargo, los coeficientes negativos destacan áreas que

podrían estar asociadas con dificultades o que no están alineadas con los patrones de éxito general.

Los valores de $R^2$, MSE y RMSE obtenidos para el modelo Ridge se imprimieron para comprender su desempeño y compararlo con el modelo de regresión lineal simple. Idealmente, un modelo Ridge bien ajustado debería mostrar un equilibrio en el rendimiento de ambos conjuntos y reducir el riesgo de sobreajuste.

**Tabla 16.**

*Métricas de desempeño obtenidas del método de Regresión Ridge*

| Métrica | Train | Test |
|---|---|---|
| $R^2$ Score | 0.21621792981663845 | -0.05063121011124805 |
| MSE | 2.985858031152714 | 3.5620353264915448 |
| RMSE | 1.7279635502963349 | 1.887335509783977 |

**Coeficiente de Determinación ($R^2$)**

El modelo obtuvo un $R^2$ de 0.2162 en el conjunto de entrenamiento, lo que indica que aproximadamente el 21.62% de la variabilidad en los datos de entrenamiento es explicada por las variables predictoras. Este valor refleja un ajuste bajo y sugiere que el modelo no logra capturar plenamente las relaciones entre las variables. En el conjunto de prueba, el $R^2$ fue de -0.0506, lo que significa que el modelo tiene un desempeño inferior al de una predicción basada únicamente en la media de los valores observados. Este resultado evidencia una deficiente capacidad de generalización del modelo y refuerza la idea de que no captura adecuadamente los patrones subyacentes en los datos.

**Error Cuadrático Medio (MSE)**

El MSE obtenido en el conjunto de entrenamiento fue de 2.9859, lo que cuantifica el error promedio al cuadrado en las predicciones realizadas por el modelo. Aunque el

valor no es excesivamente alto, sigue siendo consistente con un ajuste insuficiente reflejado en el bajo $R^2$. En el conjunto de prueba, el MSE aumentó a 3.5620, lo que indica un rendimiento aún menor en datos no vistos. Esta discrepancia entre los conjuntos refuerza la necesidad de mejorar la capacidad del modelo para generalizar a nuevos datos.

**Raíz del Error Cuadrático Medio (RMSE)**

En términos de RMSE, el modelo mostró un error promedio de 1.7280 unidades en el conjunto de entrenamiento y de 1.8873 unidades en el conjunto de prueba. Aunque la diferencia entre estos valores no es drástica, el incremento en el conjunto de prueba refleja una pérdida de precisión. Este resultado, combinado con el bajo $R^2$ y el incremento del MSE, confirma que el modelo enfrenta dificultades para ajustarse adecuadamente a los datos y, al mismo tiempo, mantener un buen desempeño en escenarios desconocidos.

Los resultados obtenidos sugieren que el modelo Ridge presenta problemas de subadecuación, ya que no logra capturar de manera adecuada las relaciones entre las variables predictoras y la variable objetivo, tanto en el conjunto de entrenamiento como en el de prueba. Para mejorar el desempeño, es necesario revisar la selección de variables predictoras, ya que podrían no ser suficientes o adecuadas para explicar el comportamiento de la calificación final.

### 4.2.2. Modelo de Regresión Lasso

En esta parte del análisis, se evaluó el rendimiento del modelo de regresión Lasso en los datos de prueba y de entrenamiento, y se visualizaron los resultados para una interpretación más clara de las predicciones del modelo en comparación con los valores reales.

**Gráfico de Dispersión de Valores Reales vs. Predicciones**

Se generó un gráfico de dispersión en el cual los valores reales (y_test) y las predicciones (y_pred_lasso) se representaron con diferentes colores. Este gráfico proporciona una comparación visual directa de los valores predichos en relación con los valores reales.

**Figura 28.**

*Gráfico de dispersión de valores reales y predicciones*

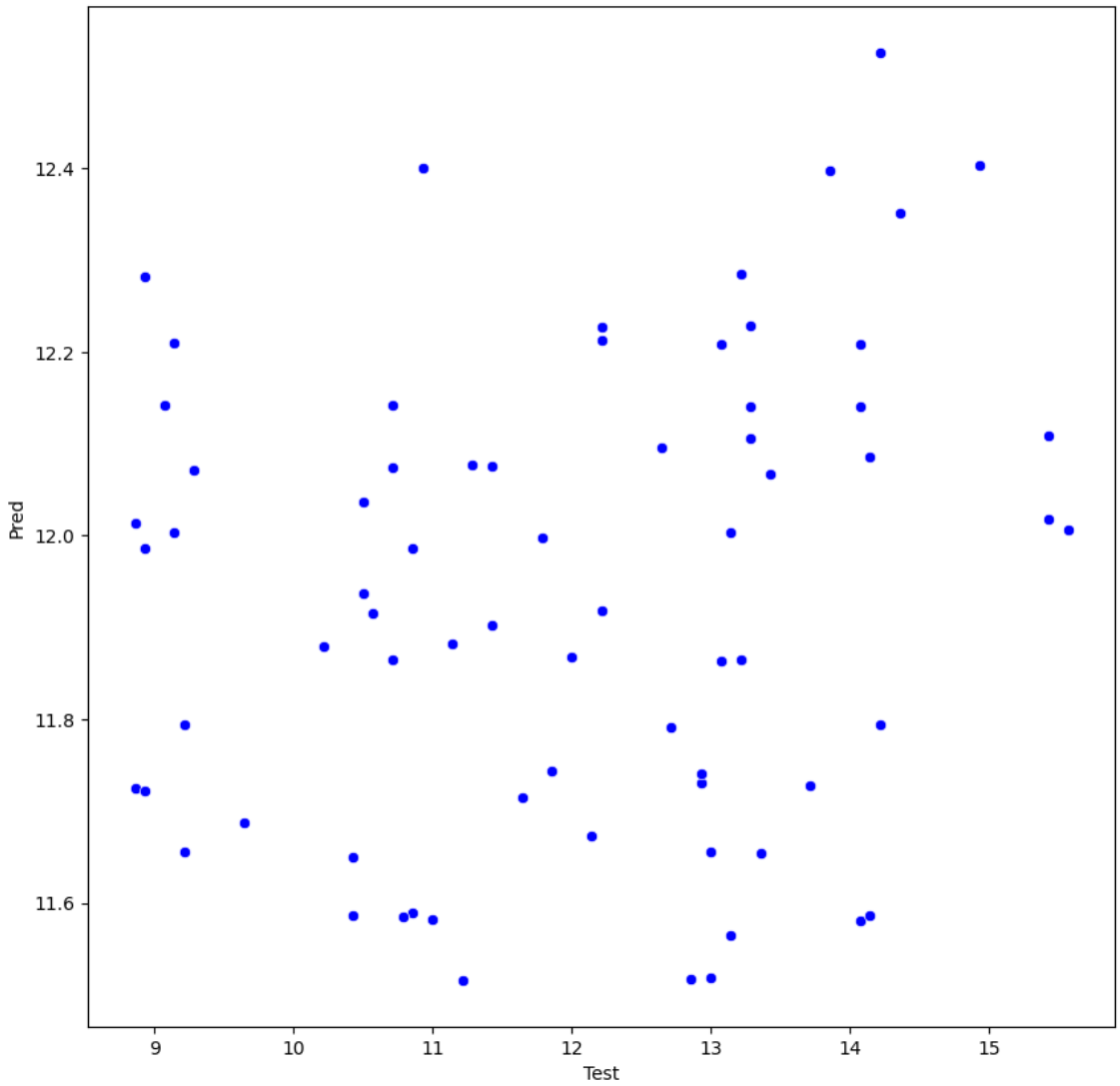



**Gráfico de Líneas para Valores Reales y Predicciones:**

En el gráfico de líneas que se muestra en la Figura 29, se representaron los valores reales y las predicciones, utilizando líneas continuas para reflejar la tendencia de ambos conjuntos. Este gráfico facilita la identificación de patrones de ajuste, mostrando cómo se alinean las predicciones con los valores reales en cada índice.

**Figura 29.**

*Gráfico de líneas para valores reales y predicciones*

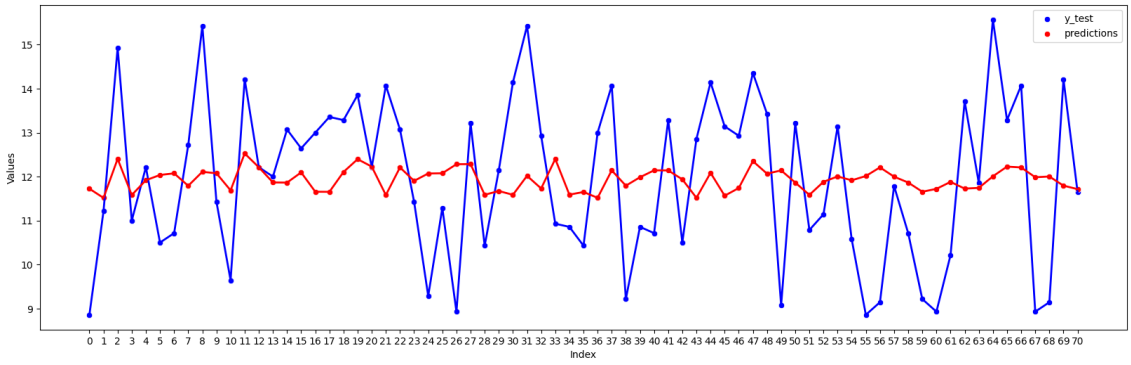
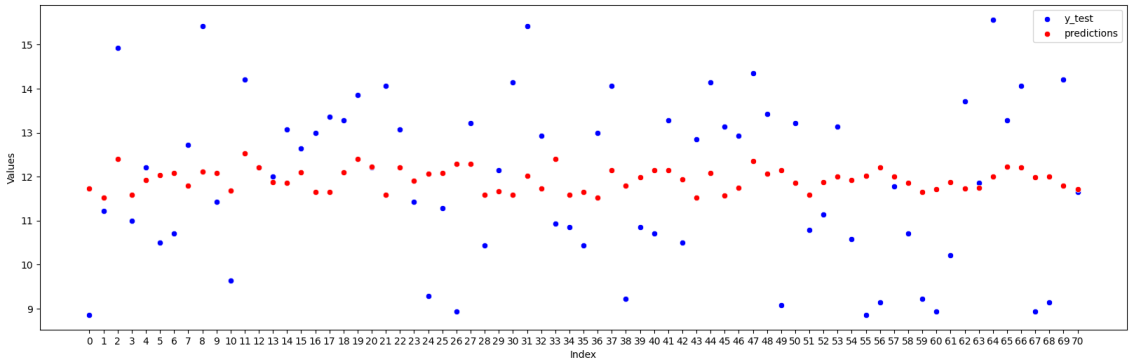



**Gráfico de Dispersión de Valores Predichos vs. Valores Reales:**

Se generó un gráfico adicional para comparar directamente los valores predichos (y_pred_lasso) con los valores reales (y_test). La cercanía de los puntos a la línea de identidad (y = x) indica qué tan precisas son las predicciones del modelo en comparación con los datos reales.

**Figura 30.**

*Gráfico de Dispersión de Valores Predichos vs. Valores Reales*




**Coeficientes de las Variables Predictoras:**

Los coeficientes de cada variable se imprimieron para interpretar el impacto de cada predictor en la calificación final. Los coeficientes que Lasso reduce a cero son aquellos que el modelo considera menos importantes, facilitando la identificación de las variables con mayor influencia en las predicciones.

**Intercepto del Modelo:**

Se imprimió el valor del intercepto, que representa la calificación final esperada cuando todas las variables predictoras son cero, proporcionando una referencia para las predicciones del modelo.

**Tabla 17.**

*Intercepto del modelo*

| **Variable** | **Coeficiente** |
|---|---|
| Arte y Cultura | 0.0 |
| CyT (Ciencia y Tecnología) | 0.0 |
| Ciencias Sociales | 0.0 |
| Comunicación | 0.0 |
| DPCC (Desarrollo Personal, Ciudadanía y Cívica) | 0.0 |
| Educación Física | 0.0 |
| EPT (Educación para el Trabajo) | 0.07013576407904516 |
| Educación Religiosa | 0.0 |
| Inglés | 0.001258222674734331 |
| Matemática | 0.06816302463462083 |
| **Intercepto** | 9.839365193050385 |

Para llevar a cabo un análisis riguroso, se procedió al entrenamiento del modelo de regresión Lasso con los datos disponibles. Este proceso permitió ajustar los parámetros del modelo de manera óptima para capturar las relaciones entre las variables predictoras y la variable objetivo.

El modelo de regresión Lasso fue ajustado utilizando los datos de entrenamiento (x_train y y_train). Una vez entrenado, se generaron predicciones tanto para el conjunto de prueba (y_pred_lasso) como para el conjunto de entrenamiento (t_pred_lasso). Este procedimiento permitió evaluar el desempeño del modelo en diferentes contextos, proporcionando una visión integral de su capacidad de generalización. La comparación entre los resultados en ambos conjuntos es fundamental para identificar posibles problemas de sobreajuste o subajuste, asegurando así la robustez del modelo.

El análisis conjunto de estas métricas proporciona una evaluación detallada del rendimiento del modelo Lasso y su capacidad para generalizar correctamente en datos no vistos, fortaleciendo la confianza en su aplicabilidad práctica.

**Tabla 18.**

*Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Regresión Lasso*

| Métrica | Train | Test |
|---|---|---|
| $R^2$ Score | 0.09595083690284056 | 0.027019124164858632 |
| MSE | 3.4440216954172507 | 3.2987714607854404 |
| RMSE | 1.8558075588318015 | 1.81625203669134 |

**Coeficiente de Determinación ($R^2$)**

El modelo obtuvo un $R^2$ de 0.0959 en el conjunto de entrenamiento, lo que indica que solo el 9.59% de la variabilidad en los datos de entrenamiento es explicada por las variables predictoras. Este valor es bajo, lo que sugiere que el modelo no logra capturar adecuadamente las relaciones entre las variables. Por otro lado, en el conjunto de prueba, el $R^2$ fue de 0.0270, un valor también bajo pero positivo, lo que indica una ligera mejora respecto a la simple predicción basada en la media de los valores observados. Sin embargo, ambos valores reflejan un ajuste deficiente del modelo, lo que apunta a una posible subadecuación.

**Error Cuadrático Medio (MSE)**

En términos de error cuadrático medio, el MSE para el conjunto de entrenamiento fue de **3.4440**, mientras que en el conjunto de prueba fue de **3.2988**. El MSE en el conjunto de prueba es ligeramente menor que en el conjunto de entrenamiento, lo que podría indicar que el modelo no está sobre ajustado. Sin embargo, estos valores aún son elevados y sugieren un margen significativo de error en las predicciones.

**Raíz del Error Cuadrático Medio (RMSE)**

El RMSE fue de **1.8558** en el conjunto de entrenamiento y de **1.8163** en el conjunto de prueba. Estas cifras indican que las predicciones tienen un error promedio de aproximadamente 1.8 unidades con respecto a los valores reales. La pequeña diferencia entre el RMSE de ambos conjuntos refuerza la idea de que el modelo tiene un ajuste equilibrado, pero el error promedio sigue siendo considerable.

**Interpretación de los Resultados**

**Ajuste del Modelo:**

El modelo Lasso presenta un ajuste limitado, como lo reflejan los valores bajos de $R^2$ tanto en el conjunto de entrenamiento como en el de prueba. Esto indica que el modelo no captura de manera efectiva las relaciones entre las variables predictoras y la variable objetivo. A pesar de ello, el valor positivo de $R^2$ en el conjunto de prueba sugiere que el modelo tiene una capacidad de generalización ligeramente mejor que en casos anteriores, aunque sigue siendo insuficiente.

**Generalización:**

La pequeña diferencia entre las métricas de entrenamiento y prueba (MSE y RMSE) indica que el modelo no muestra signos evidentes de sobreajuste o subajuste severo. Sin embargo, el error promedio sigue siendo alto en ambos conjuntos, lo que implica que el modelo no es adecuado para realizar predicciones precisas en este contexto.

**Eficacia Predictiva:**

Aunque el modelo no está sobre ajustado, su bajo $R^2$ y el elevado error promedio limitan su capacidad predictiva. Esto sugiere que las variables seleccionadas y/o la configuración del modelo no son óptimas para el problema planteado.

## 4.3. Aplicación de Árbol de Decisión para Regresión

En esta sección se utilizó un modelo de regresión basado en árboles de decisión para predecir la variable objetivo. Los árboles de decisión son algoritmos no lineales que dividen el conjunto de datos en subconjuntos basados en condiciones de decisión, facilitando la interpretación y permitiendo modelar relaciones complejas en los datos.

**Configuración del Modelo:**

Se utilizó el modelo DecisionTreeRegressor con una profundidad máxima (max_depth) de 3 a 5, limitando la cantidad de niveles que el árbol puede alcanzar. Este parámetro controla la complejidad del modelo, evitando que el árbol crezca demasiado y se sobreajuste a los datos de entrenamiento.

El parámetro min_impurity_decrease fue configurado en 0, lo que permite al árbol dividir los nodos sin requerir una mejora mínima en la calidad de las divisiones.

**Entrenamiento y Predicción:**

El modelo se ajustó usando los datos de entrenamiento (x_train y y_train), y se generaron predicciones para los datos de prueba (y_pred_tree) y para los datos de entrenamiento (t_pred_tree). Estas predicciones se utilizarán para evaluar la precisión y generalización del modelo.

**Gráfico de Dispersión de Valores Reales vs. Predicciones:**

En la Figura 31, en la Figura 32 y en la Figura 33 se muestran los gráficos de dispersión por cada nivel de profundidad, en los cuales se muestran las calificaciones reales (y_test) y las predicciones (y_pred_tree). Esta visualización permite observar cómo el modelo de árbol de decisión sigue la tendencia general de los valores reales y si hay discrepancias significativas.

**Figura 31.**

*Gráfico de dispersión - Max_Depth 3*

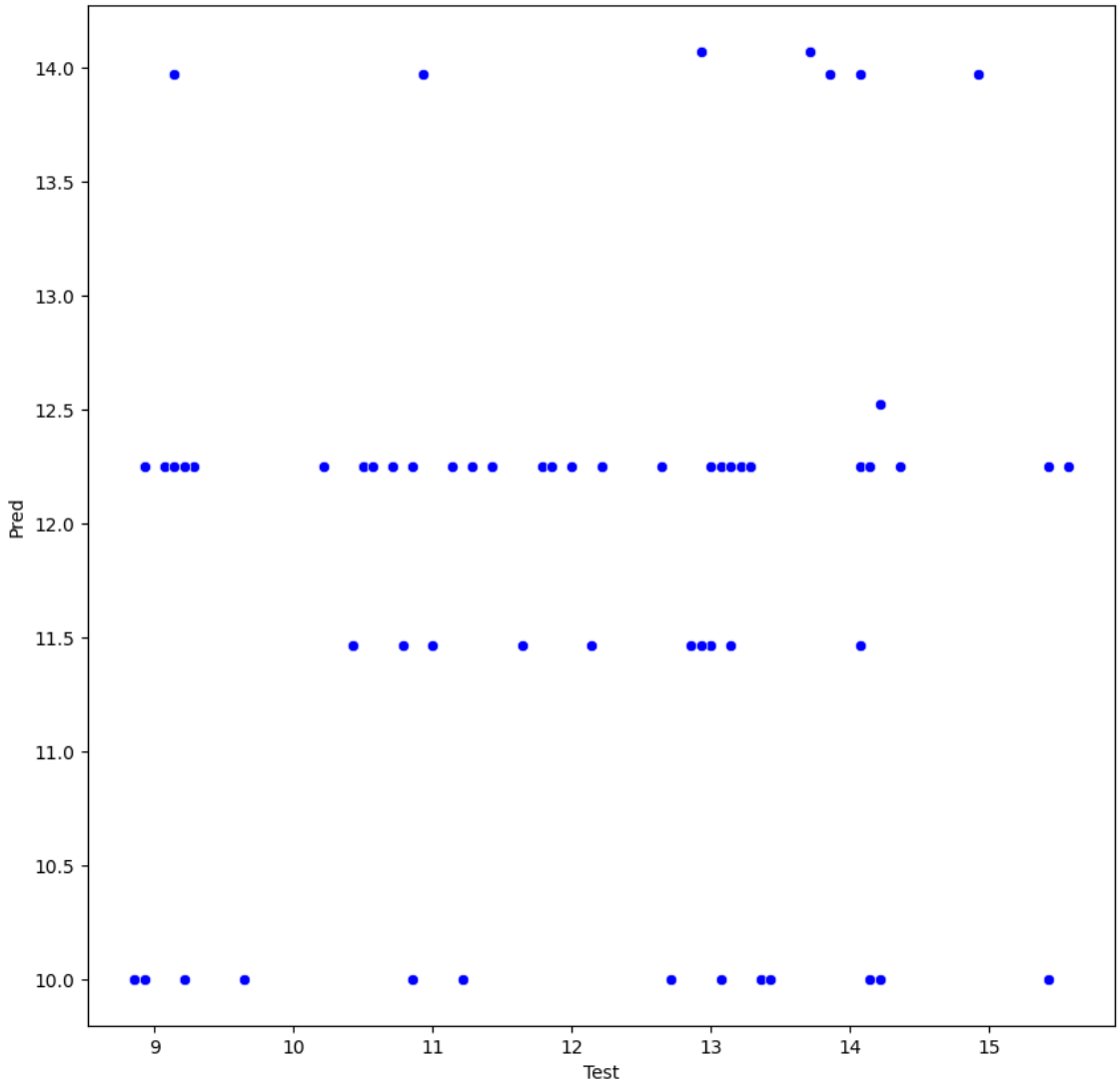


**Figura 32.**

*Gráfico de dispersión - Max_Depth 4*

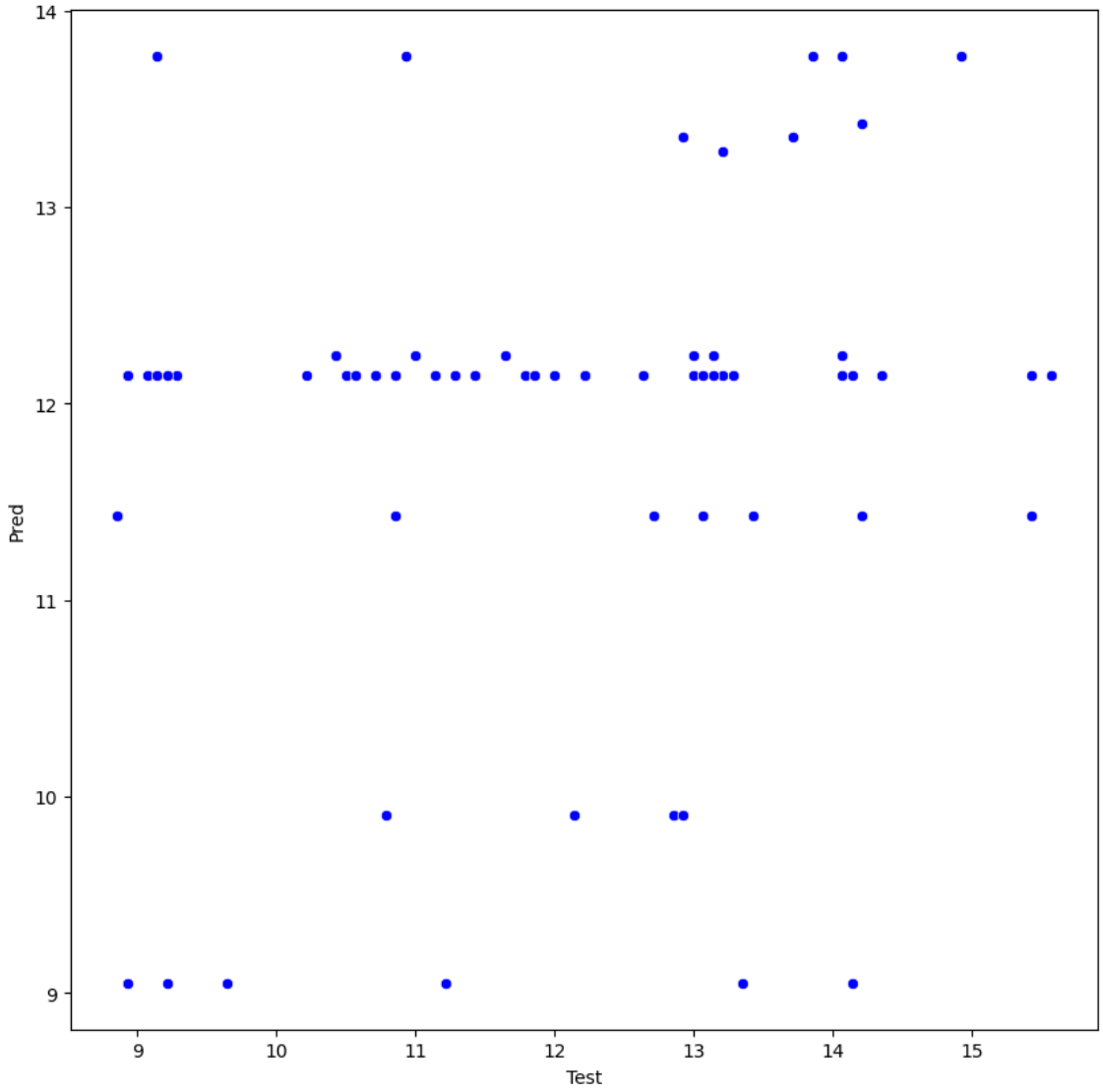


**Figura 33.**

*Gráfico de dispersión - Max_Depth 5*

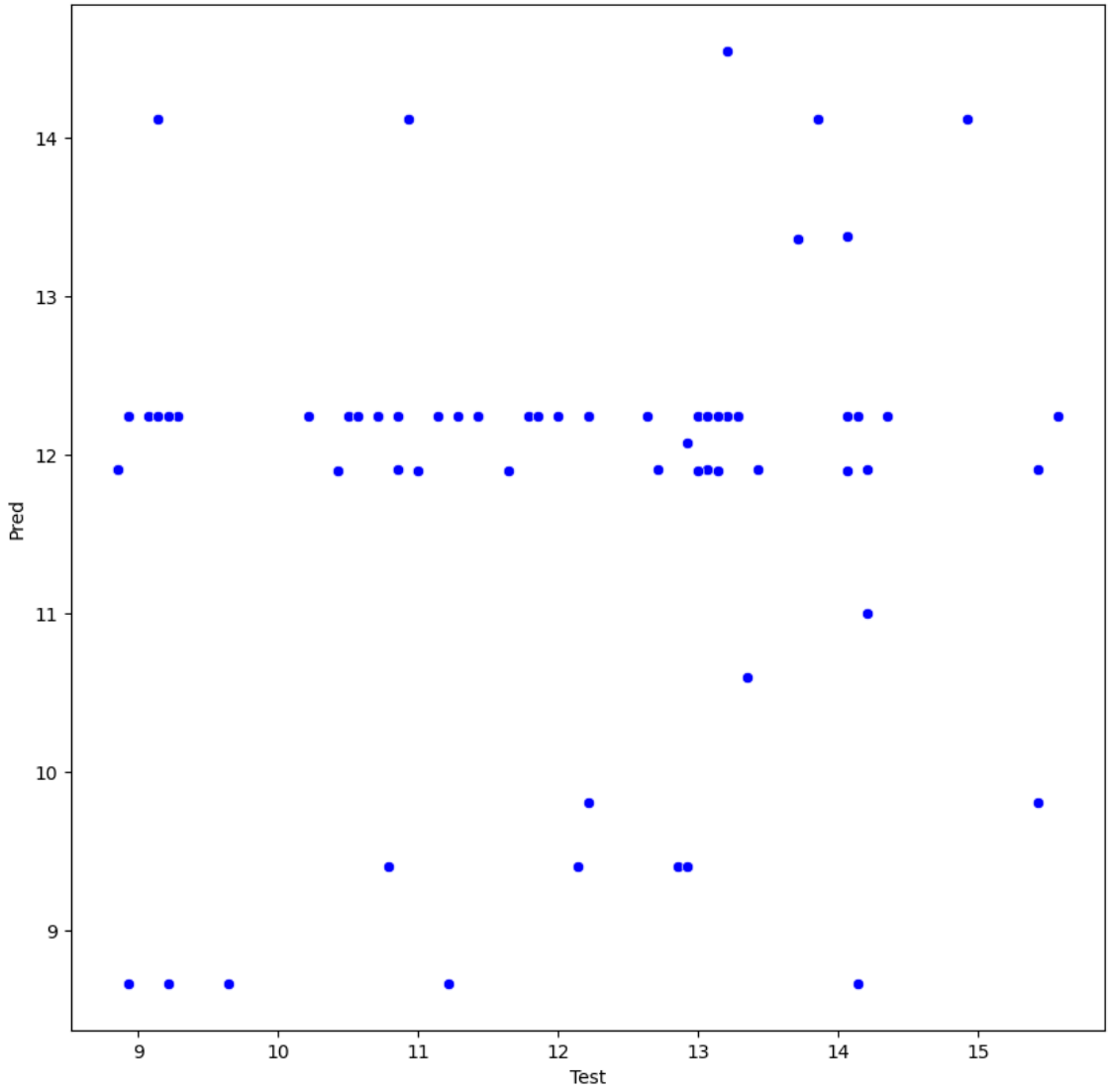



**Gráfico de Líneas para Comparación de Valores Reales y Predichos:**

Para visualizar mejor la alineación entre los valores reales y predichos, se añadió una línea de tendencia para ambos conjuntos en los gráficos de la Figura 34, Figura 35 y Figura 36. Este enfoque permite identificar patrones de ajuste y observar el comportamiento del modelo en los datos de prueba.

**Figura 34.**

*Gráfico de líneas - Max_Depth 3*

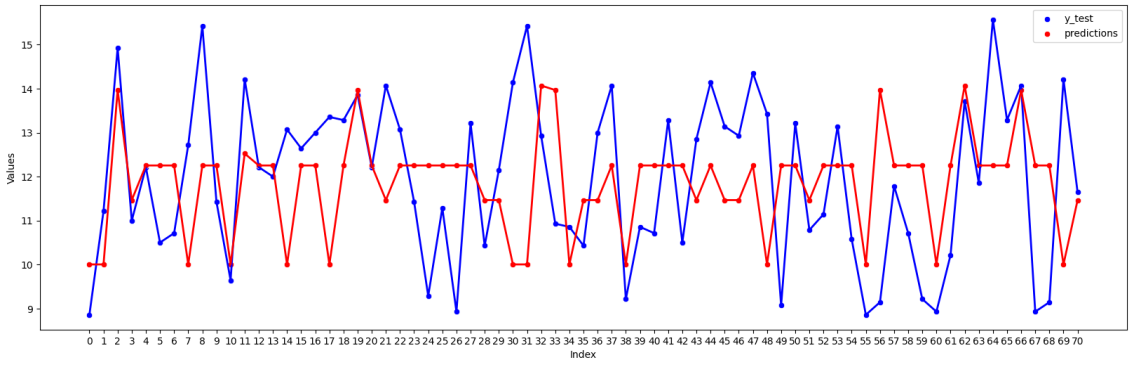



**Figura 35.**

*Gráfico de líneas - Max_Depth 4*

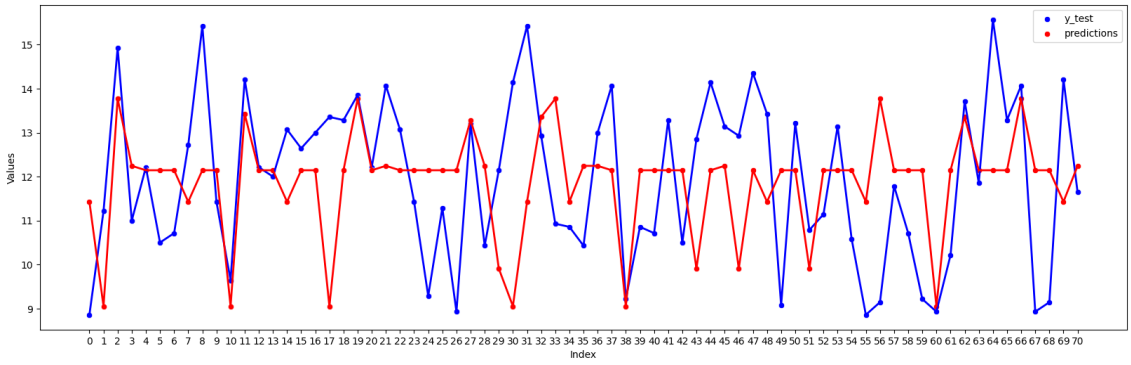



**Figura 36.**

*Gráfico de líneas - Max_Depth 5*

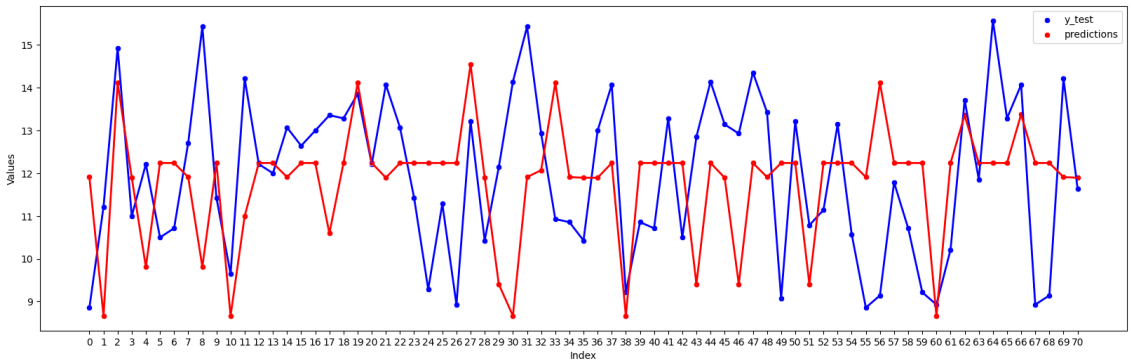


**Gráfico de Dispersión de Valores Predichos vs. Valores Reales:**

En la Figura 37, Figura 38 y Figura 39 se muestran gráficos que comparan directamente los valores reales y predichos, donde la cercanía de los puntos a la línea de identidad (y = x) indica la precisión del modelo.

**Figura 37.**

*Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth 3*

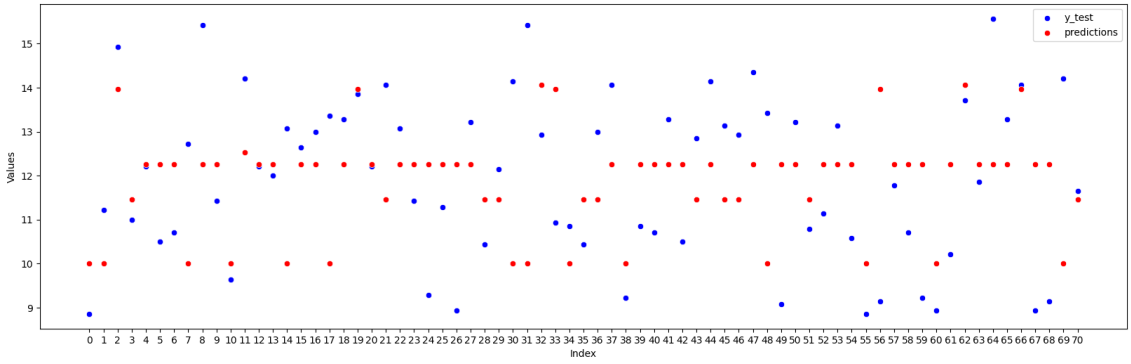



**Figura 38.**

*Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth 4*

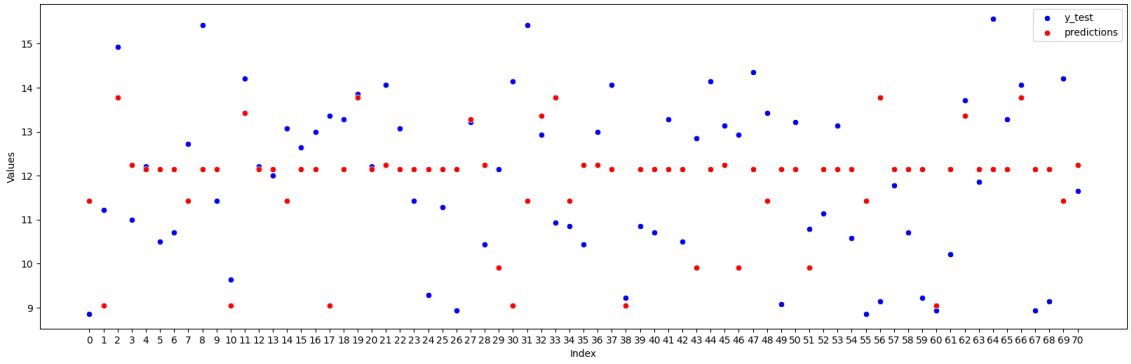


**Figura 39.**

*Gráfico de Dispersión de Valores Predichos vs. Valores Reales Max_Depth 5*

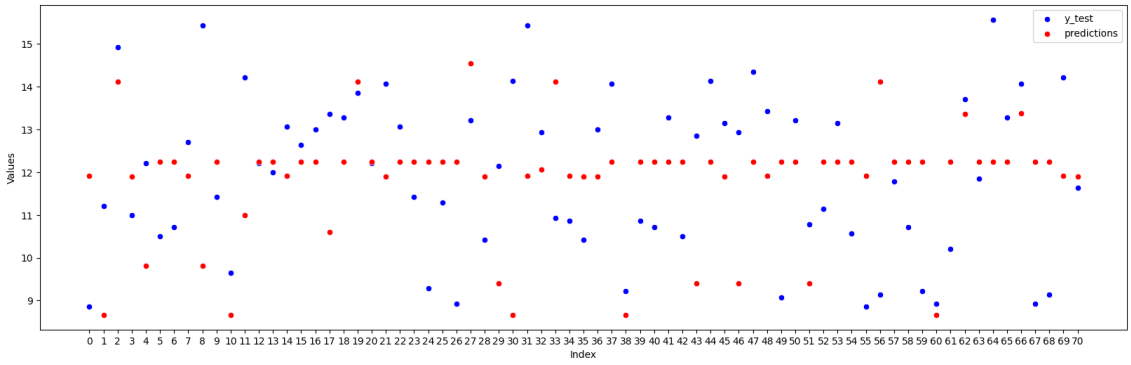



**Diagrama del Árbol de Decisión:**

Se visualizó la estructura completa del árbol de decisión usando plot_tree, lo cual muestra cómo el modelo segmenta los datos en diferentes niveles. Como se ve en la Figura 40, Figura 41 y en la Figura 42, cada nodo representa una condición de división basada en las características, mientras que las hojas del árbol indican las predicciones finales.

Esta visualización facilita la interpretación del modelo, ya que muestra las divisiones y decisiones que el árbol toma para llegar a una predicción. Además, permite identificar qué variables se consideran más importantes en cada nivel de decisión.

El árbol de decisión permite una interpretación visual clara de las reglas y divisiones que el modelo sigue para predecir la variable objetivo. Las métricas de rendimiento y las visualizaciones indican cómo el modelo generaliza en datos no vistos y resaltan las decisiones clave en el proceso de predicción. Este modelo es adecuado cuando se busca una representación gráfica del proceso de toma de decisiones y una interpretación detallada de los patrones en los datos.

**Figura 40.**

*Diagrama de Árbol de decisión - Max_Depth 3*

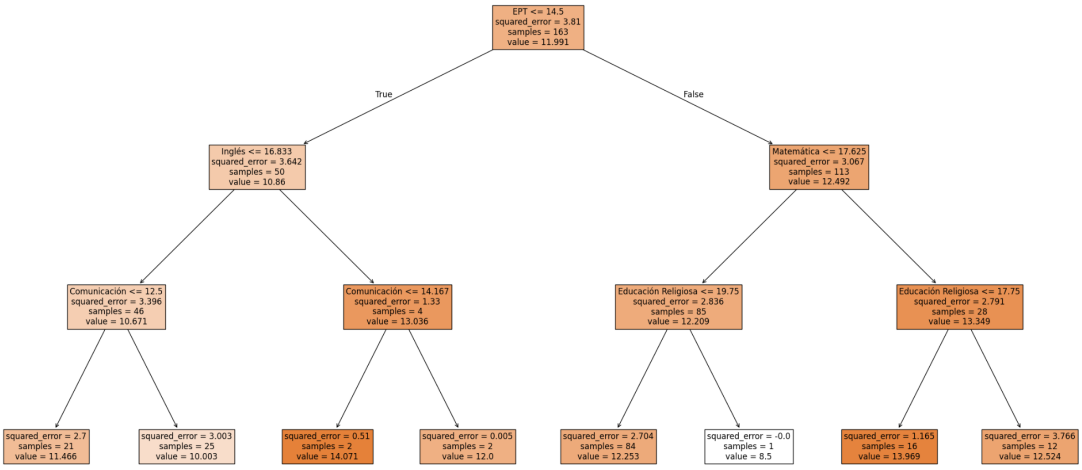



**Figura 41.**

*Árbol de decisión - Max_Depth 4*

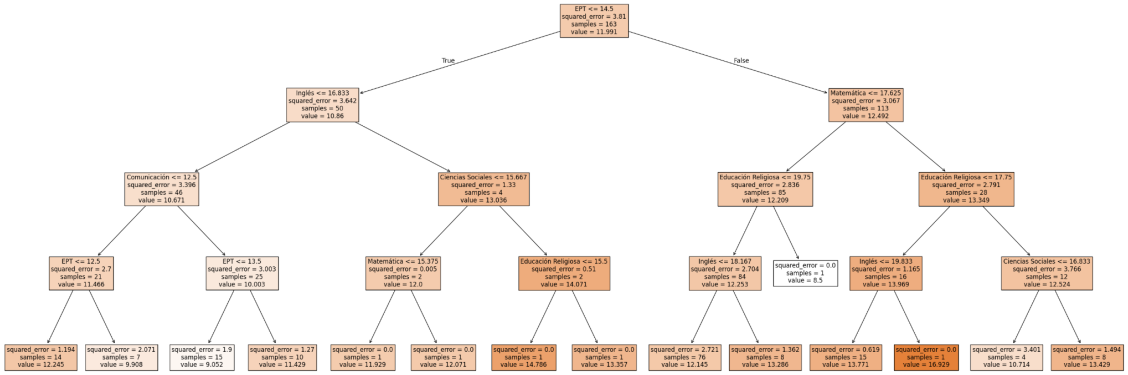
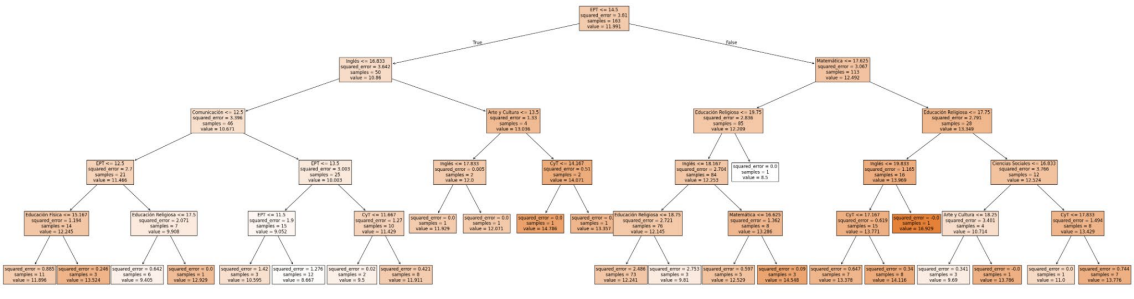


**Figura 42.**

*Árbol de decisión - Max_Depth 5*



**Interpretación de las Métricas:**

Para un max_depth de 3:

- $R^2$ Score:

  En el conjunto de entrenamiento, el modelo explica el 31.76% de la variabilidad en la variable objetivo, lo que sugiere un ajuste moderado.

  En el conjunto de prueba, el $R^2$ negativo indica que el modelo no captura adecuadamente las relaciones en datos no vistos y predice peor que un modelo promedio simple.

- Errores (MSE y RMSE):

  Los errores (MSE y RMSE) son significativamente mayores en el conjunto de prueba que en el de entrenamiento, lo que sugiere que el modelo podría estar sobreajustado a los datos de entrenamiento.

**Tabla 19.**

*Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 3*

| Métrica | Train | Test |
|---|---|---|
| $R^2$ Score | 0.31761329110422665 | -0.21985344840880972 |
| MSE | 2.599587197282594 | 4.135762420302187 |
| RMSE | 1.6123235398897438 | 2.0336573999329848 |

Para un max_depth de 4:

- **$R^2$ Score:**

  En el conjunto de entrenamiento, el modelo explica el 47.67% de la variabilidad en la variable objetivo, lo que indica un buen ajuste a los datos de entrenamiento.

  En el conjunto de prueba, el $R^2$ negativo (-0.20) sugiere que el modelo no generaliza bien, prediciendo incluso peor que un modelo promedio simple.

- **Errores (MSE y RMSE):**

  El MSE y RMSE son significativamente menores en el conjunto de entrenamiento que en el de prueba, lo que indica que el modelo está sobreajustado a los datos de entrenamiento, perdiendo capacidad de generalización.

**Tabla 20.**

*Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 4*

| Métrica | Train | Test |
|---|---|---|
| $R^2$ Score | 0.476728108299451 | -0.20086608532486494 |
| MSE | 1.9934311331529169 | 4.0713881114000445 |
| RMSE | 1.411889207109721 | 2.0177681014923503 |

Para un max_depth de 5:

- $R^2$ Score:

  En el conjunto de entrenamiento, el modelo explica el 60.74% de la variabilidad en la variable objetivo, lo que sugiere un ajuste fuerte a los datos de entrenamiento.

  En el conjunto de prueba, el $R^2$ negativo (-0.33) indica que el modelo no predice bien en datos no vistos, siendo menos efectivo que un modelo promedio simple.

- Errores (MSE y RMSE):

  Los errores en el conjunto de prueba (MSE y RMSE) son considerablemente mayores que en el conjunto de entrenamiento, lo que es un indicador claro de sobreajuste. El modelo ajusta demasiado los datos de entrenamiento, pero falla al generalizar. En la Tabla 21 se visualizan los resultados de estas 3 métricas

**Tabla 21.**

*Resultado de métricas de desempeño R2, MSE y RMSE del modelo de Árbol de Decisión con un max_depth de 5*

| Métrica | Train | Test |
|---|---|---|
| $R^2$ Score | 0.6074036773073012 | -0.3344973156469728 |
| MSE | 1.4956158448977774 | 4.524448289461454 |
| RMSE | 1.2229537378403883 | 2.127075054966668 |

## 4.4. Modelo de redes neuronales

La Tabla 22 muestra el rendimiento del modelo de red neuronal en términos de las métricas de desempeño, tanto en el conjunto de entrenamiento como en el de prueba para diferentes configuraciones de neuronas.

A medida que aumenta el número de neuronas, el MAE, MSE y RMSE en el conjunto de entrenamiento disminuyen, lo que indica que el modelo está aprendiendo mejor los datos de entrenamiento. El $R^2$ en el conjunto de entrenamiento también aumenta, alcanzando valores cercanos a 1, lo que sugiere un excelente ajuste a los datos de entrenamiento.

En el gráfico de la Figura 43 se observa como en el conjunto de prueba, inicialmente se observa una disminución en los valores de MAE, MSE y RMSE, lo que indica una mejora en el rendimiento del modelo con datos nuevos. Sin embargo, a partir de un cierto punto (alrededor de 16 neuronas), estas métricas comienzan a aumentar, lo que indica sobreajuste. El sobreajuste también se refleja en los valores negativos de $R^2$ en el conjunto de prueba, ya que el modelo no está explicando adecuadamente la variabilidad en los datos de prueba.

Un modelo con un número moderado de neuronas (por ejemplo, entre 8 y 16) logra un buen equilibrio entre el ajuste a los datos de entrenamiento y la capacidad de generalización en los datos de prueba. Esto sugiere que demasiadas neuronas pueden llevar a un sobreajuste del modelo, mientras que muy pocas neuronas pueden resultar en un modelo insuficientemente complejo.

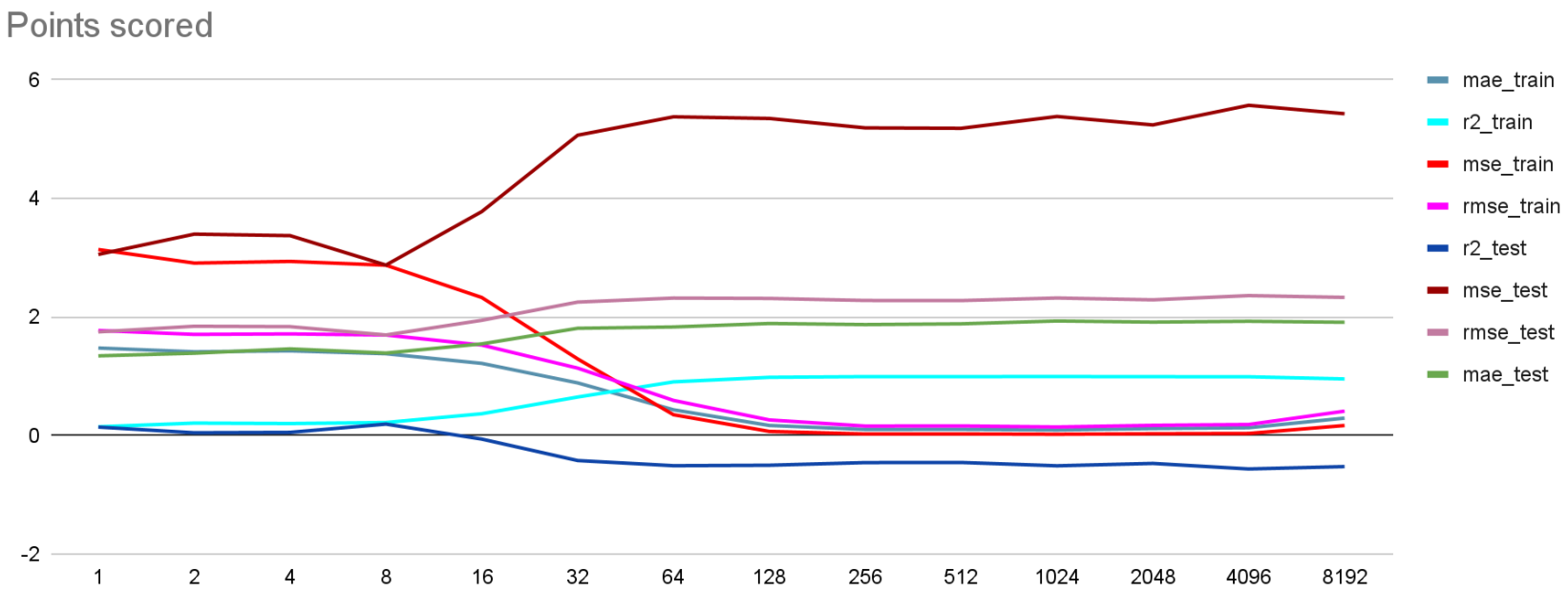
**Tabla 22.**

*Valores obtenidos en las métricas de desempeño en base a la cantidad de neuronas*

| neuronas | mae_train | r2_train | mse_train | rmse_train | mae_test | r2_test | mse_test | rmse_test |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.47578073 | 0.14705637 | 3.13742556 | 1.77127794 | 1.34356066 | 0.14199125 | 3.05505116 | 1.74787046 |
| 2 | 1.41222422 | 0.20913392 | 2.90908258 | 1.70560329 | 1.38803525 | 0.04581749 | 3.39749026 | 1.84322822 |
| 4 | 1.4278262 | 0.20147313 | 2.93726163 | 1.71384411 | 1.46138787 | 0.05293267 | 3.37215574 | 1.83634303 |
| 8 | 1.37947229 | 0.21886101 | 2.8733029 | 1.69508198 | 1.3911329 | 0.19242482 | 2.8754759 | 1.69572283 |
| 16 | 1.21648902 | 0.3670519 | 2.3282049 | 1.52584563 | 1.54580282 | -0.06105365 | 3.7780188 | 1.94371263 |
| 32 | 0.88735408 | 0.64881883 | 1.29176739 | 1.13655945 | 1.80887667 | -0.42288034 | 5.06634951 | 2.25085528 |
| 64 | 0.43330172 | 0.90486905 | 0.34992496 | 0.59154455 | 1.83108468 | -0.51005076 | 5.3767311 | 2.31877793 |
| 128 | 0.16934107 | 0.98133852 | 0.06864345 | 0.26199895 | 1.88974426 | -0.50219374 | 5.34875514 | 2.31273759 |
| 256 | 0.10211336 | 0.99310292 | 0.02536989 | 0.15927928 | 1.87112327 | -0.4575917 | 5.1899438 | 2.27814482 |
| 512 | 0.10329653 | 0.99289911 | 0.02611956 | 0.16161548 | 1.88401839 | -0.455628 | 5.1829518 | 2.27660972 |
| 1024 | 0.09303316 | 0.99437494 | 0.02069093 | 0.14384343 | 1.93239151 | -0.51158867 | 5.382207 | 2.3199584 |
| 2048 | 0.11685363 | 0.99200021 | 0.02942602 | 0.17154014 | 1.91325036 | -0.47159222 | 5.23979449 | 2.28905974 |
| 4096 | 0.13097747 | 0.99064458 | 0.03441252 | 0.18550613 | 1.92740823 | -0.56461577 | 5.57101685 | 2.36030016 |
| 8192 | 0.29279161 | 0.95413969 | 0.16869028 | 0.41071922 | 1.91014826 | -0.52469897 | 5.42888792 | 2.32999741 |

**Figura 43.**

*Gráfico lineal de los valores obtenidos en las métricas de desempeño en base a la cantidad de neuronas*



**Desempeño en el Conjunto de Entrenamiento**

A medida que aumenta el número de neuronas en la capa oculta, se observa una mejora significativa en las métricas del conjunto de entrenamiento. Los valores de MAE y MSE disminuyen progresivamente, lo que indica una capacidad creciente del modelo para ajustarse a los patrones presentes en los datos de entrenamiento. Por otro lado, el $R^2$ también muestra un aumento constante, alcanzando valores superiores a 0.99 cuando se utilizan configuraciones de 128 neuronas o más. Esto refleja que el modelo captura casi toda la variabilidad en los datos de entrenamiento. Sin embargo, este comportamiento sugiere un posible sobreajuste, ya que el modelo parece adaptarse demasiado a los datos de entrenamiento en detrimento de su capacidad para generalizar.

**Desempeño en el Conjunto de Prueba**

En el conjunto de prueba, el comportamiento es diferente. Para configuraciones simples, entre 1 y 8 neuronas, los errores (MAE, MSE, RMSE) son moderados, y el $R^2$ alcanza valores positivos, indicando un desempeño aceptable. Sin embargo, a partir de 16 neuronas, el $R^2$ en el conjunto de prueba se torna negativo y los errores comienzan a aumentar considerablemente. Esto implica que el modelo pierde la capacidad de generalizar, siendo menos efectivo que una simple predicción basada en la media de los valores observados. Este deterioro en las métricas del conjunto de prueba, a pesar de la mejora en el conjunto de entrenamiento, es un claro indicio de sobreajuste.

**Sobreajuste y Complejidad del Modelo**

El análisis muestra que configuraciones con más de 16 neuronas en la capa oculta generan modelos excesivamente complejos que se ajustan en exceso a los datos de entrenamiento. Esto se refleja en valores muy bajos de error y un $R^2$ cercano a 1 en este conjunto, pero con un rendimiento deficiente en el conjunto de prueba. El aumento de los errores y los valores negativos de $R^2$ en las configuraciones más complejas confirman que el modelo no es capaz de generalizar correctamente a datos nuevos, un fenómeno característico del sobreajuste.

**Mejor Configuración del Modelo**

Entre todas las configuraciones probadas, la red neuronal con 8 neuronas en la capa oculta ofrece el mejor equilibrio entre ajuste y generalización. En esta configuración, el $R^2$ en el conjunto de prueba es de 0.1924, y los errores se mantienen razonablemente bajos en comparación con configuraciones más complejas. Esto sugiere que esta configuración es adecuada para capturar patrones sin comprometer la capacidad del modelo para generalizar a datos nuevos.

# CAPÍTULO V
# DISCUSIÓN

La discusión de los resultados obtenidos en este estudio revela que, al igual que en investigaciones previas, la relación entre el rendimiento académico escolar y el rendimiento académico universitario es compleja y multifactorial. Por ejemplo, Asor et al. (2023) demostraron que los algoritmos de machine learning, como el bayes ingenuo y las redes neuronales, son eficaces en la predicción del rendimiento académico en contextos escolares, lo cual es consistente con el desempeño observado en este estudio para configuraciones intermedias de redes neuronales. Sin embargo, el limitado desempeño de las redes neuronales y otros algoritmos en los datos de prueba indica que estas técnicas no logran capturar completamente la variabilidad del rendimiento académico universitario, probablemente debido a la ausencia de variables explicativas adicionales.

Por otro lado, El Guabassi et al. (2021) destacaron la efectividad de algoritmos de regresión, como la log-lineal y ANCOVA, para predecir el rendimiento académico, pero también señalaron la importancia de ajustar los modelos según las características de los datos. En este estudio, aunque las métricas de error y $\mathbf{R^2}$ reflejan un ajuste aceptable en los datos de entrenamiento, los valores negativos de $\mathbf{R^2}$ en los datos de prueba evidencian problemas de generalización y posible sobreajuste en configuraciones más complejas, lo cual es un desafío común en la implementación de modelos predictivos en contextos educativos.

De manera similar, Terán Montaño y Schulmeyer (2022) encontraron que, aunque existe una correlación entre el rendimiento escolar y universitario, esta es moderada y la predictibilidad disminuye cuando el rendimiento escolar se considera como única variable. Estos hallazgos son congruentes con los resultados de este estudio, ya que la relación entre el rendimiento académico escolar y universitario no fue suficientemente fuerte para generar predicciones precisas.

Además, investigaciones como las de Martínez Pérez et al. (2020) y Gutiérrez Monsalve et al. (2021) subrayan la influencia de factores adicionales, como la motivación, el tipo de escuela secundaria y las horas dedicadas al estudio, en el rendimiento académico universitario. Estas variables, ausentes en este estudio, podrían proporcionar un contexto más rico para explicar la variabilidad observada. Asimismo, la investigación titulada "Tipo de Escuela Secundaria Predice el Rendimiento Académico en la Universidad" refuerza la idea de que el contexto educativo temprano tiene un impacto significativo en el rendimiento posterior, sugiriendo que la inclusión de factores relacionados con el tipo de institución escolar podría mejorar la capacidad predictiva de los modelos.

En resumen, aunque los resultados de este estudio muestran que los algoritmos de machine learning pueden identificar patrones en los datos de rendimiento académico, las limitaciones observadas en su capacidad de generalización y precisión en los datos de prueba son consistentes con hallazgos previos que destacan la necesidad de considerar variables contextuales y multifactoriales para lograr una predicción más efectiva del rendimiento académico universitario. Esto subraya la importancia de enriquecer los modelos con datos adicionales y de emplear enfoques integrales que tomen en cuenta no solo el rendimiento escolar, sino también factores personales, motivacionales y contextuales.

# CONCLUSIONES

## Primero

Se determinó la relación entre el rendimiento académico escolar y el rendimiento académico universitario de los ingresantes a la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann en 2023, empleando el algoritmo de machine learning Regresión Lineal con regularización Ridge y Lasso. Los resultados indican que ambos métodos tuvieron un desempeño limitado en la predicción del rendimiento académico universitario. En el caso de Ridge, el modelo logró un **$R^2$ score** de 0.216 en los datos de entrenamiento, pero presentó una caída significativa a -0.051 en el conjunto de prueba, indicando problemas de generalización. Los errores (MSE de 2.986 en entrenamiento y 3.562 en prueba; RMSE de 1.728 y 1.887, respectivamente) reflejan un mayor error en los datos de prueba. Por otro lado, el modelo Lasso mostró un **$R^2$ score** menor en el entrenamiento (0.096), pero con un mejor desempeño en el conjunto de prueba, logrando un **$R^2$ score** positivo de 0.027, además de menores errores en comparación con Ridge (MSE de 3.444 en entrenamiento y 3.299 en prueba; RMSE de 1.856 y 1.816, respectivamente). Estos resultados sugieren que, aunque Lasso tuvo un mejor desempeño relativo, ambos métodos fueron incapaces de capturar una relación significativa entre el rendimiento académico escolar y universitario, lo que indica que la variabilidad del rendimiento universitario no puede explicarse completamente con las variables utilizadas en el modelo.

## Segundo

Se determinó la relación entre el rendimiento académico escolar y el rendimiento académico universitario de los ingresantes a la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann en 2023, empleando el algoritmo de machine learning Árboles de Decisión considerando diferentes profundidades máximas. Para una profundidad máxima de 3, el modelo presentó un **$R^2$ score** de 0.318 en el conjunto de entrenamiento, lo que indica que logra explicar el 31.8% de la variabilidad en este conjunto. Sin embargo, en el conjunto de prueba, el **$R^2$ score** fue negativo (-0.220), mostrando que el modelo no generaliza bien. Además, los errores, con un MSE de 2.600 en entrenamiento y 4.136 en

prueba, junto con un RMSE de 1.612 y 2.034, respectivamente, indican un desempeño deficiente en predicciones. Al aumentar la profundidad máxima a 4, el **$R^2$ score** en entrenamiento mejora a 0.477, y aunque sigue siendo negativo en el conjunto de prueba (-0.201), presenta una ligera mejora en comparación con la profundidad 3. Los errores disminuyen en entrenamiento (MSE de 1.993 y RMSE de 1.412), pero se mantienen similares en prueba (MSE de 4.071 y RMSE de 2.018). Para una profundidad máxima de 5, el **$R^2$ score** en entrenamiento aumenta significativamente a 0.607, pero en prueba cae a -0.334, indicando sobreajuste. Los errores en entrenamiento (MSE de 1.496 y RMSE de 1.223) son los más bajos entre las profundidades evaluadas, pero en prueba aumentan (MSE de 4.524 y RMSE de 2.127), confirmando el deterioro en la capacidad de generalización del modelo, lo que indica que la variabilidad del rendimiento universitario no puede explicarse completamente con las variables utilizadas en el modelo.

**Tercero**

Se determinó la relación entre el rendimiento académico escolar y el rendimiento académico universitario de los ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann en 2023 mediante el uso del algoritmo de machine learning Redes Neuronales, presentándose una alta capacidad de ajuste en los datos de entrenamiento, evidenciada por el incremento del coeficiente de determinación (**$R^2$**) a medida que se aumenta el número de neuronas, alcanzando valores cercanos a 1.0 en configuraciones más complejas. No obstante, al evaluar los datos de prueba, las métricas reflejan una menor capacidad de generalización, con valores de **$R^2$** negativos a partir de las configuraciones más complejas (e.g., 32 neuronas o más), lo que indica un sobreajuste del modelo. Asimismo, los errores medios cuadráticos (**MSE**) y sus raíces cuadráticas (**RMSE**) en los datos de prueba son consistentemente mayores que en los datos de entrenamiento, con un aumento significativo en configuraciones excesivamente complejas. Por otro lado, redes con un número intermedio de neuronas (como 8) mostraron un mejor equilibrio entre las métricas de entrenamiento y prueba, destacándose como configuraciones más apropiadas. Estos resultados sugieren que, aunque las redes neuronales pueden captar patrones en los datos, el rendimiento académico universitario muestra una variabilidad que no puede explicarse completamente con las variables utilizadas en el modelo.

## Conclusión General

Se determinó la relación entre el rendimiento académico escolar y el rendimiento académico universitario de los ingresantes a la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann en 2023, utilizando diferentes algoritmos de machine learning: Regresión Lineal con regularización Ridge y Lasso, Árboles de Decisión y Redes Neuronales. Los resultados evidencian que, aunque los algoritmos evaluados lograron identificar ciertos patrones en los datos de entrenamiento, su capacidad para generalizar en los datos de prueba fue limitada, con valores de $\mathbf{R^2}$ frecuentemente negativos y errores más elevados en el conjunto de prueba, lo que señala problemas de generalización y sobreajuste en la mayoría de los modelos. Entre los algoritmos utilizados, las Redes Neuronales con una configuración intermedia (e.g., 8 neuronas) mostraron el mejor equilibrio entre el ajuste en los datos de entrenamiento y el desempeño en prueba. Sin embargo, los valores negativos o bajos de $\mathbf{R^2}$ en los datos de prueba y los altos errores en todos los modelos sugieren que la relación entre el rendimiento académico escolar y universitario es débil, indicando que la variabilidad del rendimiento universitario no puede explicarse completamente con las variables utilizadas en los modelos. Esto resalta que el rendimiento académico universitario no es completamente explicado por el rendimiento académico escolar.

# RECOMENDACIONES

Dado que los resultados de la investigación indican una relación limitada entre el rendimiento académico escolar y el universitario al emplear las variables actuales, se recomienda incluir factores adicionales que puedan influir significativamente en el desempeño académico. Variables como las características socioeconómicas, el nivel de motivación del estudiante, las horas dedicadas al estudio y el acceso a recursos educativos pueden ofrecer una perspectiva más completa y mejorar la capacidad predictiva de los modelos.

Para comprender mejor la relación entre el rendimiento escolar y el universitario, se sugiere realizar un estudio longitudinal que permita observar el desempeño de los estudiantes a lo largo de varios semestres. Esto proporcionaría información valiosa sobre la evolución académica y los factores que impactan en su rendimiento a lo largo del tiempo, lo que podría enriquecer las conclusiones y guiar estrategias más efectivas para el apoyo estudiantil.

Dado que algunos algoritmos de machine learning, como las redes neuronales, presentaron problemas de sobreajuste, es fundamental optimizar tanto los modelos como el preprocesamiento de los datos. Esto incluye aplicar técnicas como la validación cruzada, la regularización y la selección de características más relevantes. Además, se recomienda explorar algoritmos adicionales, como modelos de ensamblado (e.g., Random Forest o Gradient Boosting), que podrían ofrecer un mejor equilibrio entre ajuste y generalización en este tipo de análisis.

## REFERENCIAS BIBLIOGRÁFICAS


Arias, F. (2012). *El Proyecto de Investigación. Introducción a la metodología científica*.

Asor, J. R., Catedrilla, G. M. B., Buama, C. A. C., Malabayabas, M. E., & Malabayabas, C. E. (2023). Prediction of Senior High School Students' Performance in a State University: An Educational Data Mining Approach. *Volume 13, Issue 6, Pages 925 - 931*, *13*(6), 925–931. https://doi.org/10.18178/ijiet.2023.13.6.1888

Bosch Rué, A., Casas Roma, J., & Lozano Bagén, T. (2019). *Deep learning: principios y fundamentos*.

Chavez, H., Chavez-Arias, B., Contreras-Rosas, S., Alvarez-Rodríguez, J. M., & Raymundo, C. (2023). Artificial neural network model to predict student performance using nonpersonal information. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1106679

Chong González, E. G. (1970). Factores que inciden en el rendimiento académico de los estudiantes de la Universidad Politécnica del Valle de Toluca. *Revista Latinoamericana de Estudios Educativos*, *47*(1), 91–108. https://doi.org/10.48102/rlee.2017.47.1.159

El Guabassi, I., Bousalem, Z., Marah, R., & Qazdar, A. (2021). Forecasting Students' Academic Performance Using Different Regression Algorithms. *Volume 211 LNNS, Pages 221 - 231*, *211 LNNS*, Fez. https://doi.org/10.1007/978-3-030-73882-2_21

Hernández Sampieri, R., Fernández Collado, C., & del Pilar Baptista Lucio, M. (2014). *Metodología de la investigación, 5ta Ed*. www.FreeLibros.com

*Low-Performing Students*. (2016). OECD. https://doi.org/10.1787/9789264250246-en

Martínez Pérez, J. R., Ferrás Fernández, Y., Bermúdez Cordoví, L. L., Ortiz Cabrera, Y., & Pérez Leyva, E. H. (2020). Rendimiento académico en estudiantes Vs factores que influyen en sus resultados: una relación a considerar. *EDUMECENTRO*, *12*(4), 105–121. https://orcid.org/0000-0002-0415-6200

Murphy Kevin. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, MIT Press.

Nicomedes, E. N. (2018). *TIPOS DE INVESTIGACIÓN*. http://repositorio.usdg.edu.pe/handle/USDG/34

Oropeza Tena, R., Ávalos Latorre, M. L., & Ferreyra Murillo, D. A. (2017). Comparación entre rendimiento académico, autoeficacia y práctica deportiva en universitarios. *Actualidades Investigativas En Educación*, *17*(1). https://doi.org/10.15517/aie.v17i1.27271

Rico Páez, A. (2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *12*(24). https://doi.org/10.23913/ride.v12i24.1196

Terán Montaño, A., & Schulmeyer, M. K. (2022). *Relación entre El Rendimiento Académico en Secundaria y el Rendimiento Académico Universitario.* http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2306-86712022000100005&lng=es&tlng=es.

**ANEXOS**

# ANEXO 01:

# MATRIZ DE CONSISTENCIA

<table>
<tr><th colspan="6">DETERMINACIÓN DE LA RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO ESCOLAR Y EL RENDIMIENTO ACADÉMICO UNIVERSITARIO MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING EN INGRESANTESDE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN, AÑO 2023</th></tr>
<tr><th>Problemas</th><th>Objetivos</th><th>Hipótesis</th><th>Variables</th><th>Indicadores</th><th>Diseño de la investigación</th></tr>
<tr>
<td><u>Problema general</u><br>¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?</td>
<td><u>Objetivo general</u><br>Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.</td>
<td><u>Objetivo general</u><br>La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso de algoritmos de machine learning es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023.</td>
<td rowspan="4">Variable 1:<br><br>Rendimiento académico escolar<br><br>Variable 2:<br><br>Rendimiento Académico Universitario</td>
<td rowspan="4">Rendimiento académico escolar:<br>Seguridad<br>Funcionalidad<br>Registro académico de notas en matemática<br>Registro académico de notas en ciencia y tecnología<br>Registro académico de notas en comunicación<br>Registro académico de notas en educación para el trabajo<br>Registro académico de notas en personal social<br>Registro académico de notas en ingles<br>Registro académico de notas en<br>Registro académico de notas en arte y cultura<br>Registro académico de notas en formación ciudadana y cívica<br>Registro académico de notas en educación física<br><br>Rendimiento Académico Universitario:<br>Promedio de notas de cursos generales</td>
<td rowspan="4"><u>Población:</u><br>235 estudiantes<br><br><u>Tipo de investigación:</u><br>Aplicada o tecnológica<br><br><u>Diseño:</u><br>No-experimental<br><br><u>Nivel:</u><br>Correlacional<br><br><u>Instrumento:</u><br>Ficha de observación</td>
</tr>
<tr>
<td><u>Problema específico 1</u><br>¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?</td>
<td><u>Objetivo específico 1</u><br>Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023</td>
<td><u>Hipótesis específica 1</u><br>La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Regresión Lineal es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023</td>
</tr>
<tr>
<td><u>Problema específico 2</u><br>¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?</td>
<td><u>Objetivo específico 2</u><br>Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023</td>
<td><u>Hipótesis específica 2</u><br>La relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Árboles de Decisión es de nula a baja en ingresantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann, 2023</td>
</tr>
<tr>
<td><u>Problema específico 3</u><br>¿Cuál es la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023?</td>
<td><u>Objetivo específico 3</u><br>Determinar la relación entre el rendimiento académico escolar y el rendimiento académico universitario mediante el uso del algoritmo de machine learning Redes Neuronales en ingresantes de la Facultad de Ingeniería de la Universidad Jorge Basadre Grohmann, 2023</td>
<td><u>Hipótesis específica 3</u><br>Existe una diferencia significativa en la proporción de postulantes aceptados entre el Portal de Admisión y el sistema SYAM</td>
</tr>
</table>

## ANEXO 02:

## FICHA DE OBSERVACIÓN

| DETERMINACIÓN DE LA RELACIÓN ENTRE EL RENDIMIENTO ACADÉMICO ESCOLAR Y EL RENDIMIENTO ACADÉMICO UNIVERSITARIO MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING EN INGRESANTES DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN, 2023 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Datos de Entrenamiento de rendimiento académico escolar y rendimiento académico universitario | | | | Datos de Entrenamiento de rendimiento académico escolar y rendimiento académico universitario | | | |
| Algoritmos de Machine Learning | R2 | MAE | MSE | RMSE | R2 | MAE | MSE | RMSE |
| Regresión Lineal Ridge | | | | | | | | |
| Regresión Lineal Lasso | | | | | | | | |
| Árbol de Decisión de profundidad máxima 3 | | | | | | | | |
| Árbol de Decisión de profundidad máxima 4 | | | | | | | | |
| Árbol de Decisión de profundidad máxima 5 | | | | | | | | |
| Red Neuronal de 1 neurona | | | | | | | | |
| Red Neuronal de 2 neuronas | | | | | | | | |
| Red Neuronal de 4 neuronas | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Red Neuronal de 8 neuronas | | | | | | | | |
| Red Neuronal de 16 neuronas | | | | | | | | |
| Red Neuronal de 32 neuronas | | | | | | | | |
| Red Neuronal de 64 neuronas | | | | | | | | |
| Red Neuronal de 128 neuronas | | | | | | | | |
| Red Neuronal de 256 neuronas | | | | | | | | |
| Red Neuronal de 512 neuronas | | | | | | | | |
| Red Neuronal de 1024 neuronas | | | | | | | | |
| Red Neuronal de 2048 neuronas | | | | | | | | |
| Red Neuronal de 4096 neuronas | | | | | | | | |
| Red Neuronal de 8192 neuronas | | | | | | | | |

# ANEXO 03:
# VALIDEZ DEL INSTRUMENTO

## VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

### 1. DATOS DEL EXPERTO

| | |
|---|---|
| APELLIDO Y NOMBRES | CABANA YUPANQUI SILVANA BEATRIZ |
| GRADO ACADÉMICO | BACH. EN CS. EN ING. INFORMÁTICA Y SISTEMAS |
| TÍTULO PROFESIONAL | INGENIERO EN INFORMÁTICA Y SISTEMAS |
| EMPRESA / INSTITUCIÓN DONDE LABORA | UNJBG |
| INSTRUMENTO EVALUADO | FICHA DE OBSERVACIÓN |
| AUTOR DEL INSTRUMENTO | BACH. JAMES ENRIQUE SEGOVIA HINOJOSA |

### 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. | | | | | ✓ |
| OBJETIVIDAD | Nivel de imparcialidad y ausencia de opiniones personales | | | | ✓ | |
| ORGANIZACIÓN | Esta presentado de forma estructurada y secuencial. | | | | | ✓ |
| SUFICIENCIA | El contenido es adecuado y completo para cumplir los objetivos. | | | | | ✓ |
| PERTINENCIA | El contenido es relevante y apropiado para la investigación. | | | | | ✓ |
| COHERENCIA | Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo. | | | | ✓ | |
| RELEVANCIA | Evalúa la importancia del instrumento. | | | | | ✓ |
| ACTUALIDAD | Adecuado al avance de la tecnología | | | | | ✓ |

### 3. RESULTADOS DE VALIDACIÓN

| | |
|---|---|
| PROMEDIO DE VALIDACIÓN | 95 % |
| EVALUACIÓN DE FACTIBILIDAD | ES APLICABLE |
| FECHA DE EVALUACIÓN | 15/02/2024 |
| FIRMA DEL EXPERTO | SILVANA BEATRIZ CABANA YUPANQUI<br>ING. EN INFORMATICA Y SISTEMAS<br>CIP N° 274543 |

## VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

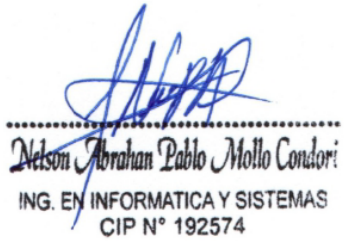
### 1. DATOS DEL EXPERTO

| | |
|---|---|
| APELLIDO Y NOMBRES | **MOLLO CONDORI NELSON ABRAHAN** |
| GRADO ACADÉMICO | **MSC. EN INGENIERÍA DE SISTEMAS E INFORMÁTICA** |
| TÍTULO PROFESIONAL | **INGENIERO EN INFORMÁTICA Y SISTEMAS** |
| EMPRESA / INSTITUCIÓN DONDE LABORA | **UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN** |
| INSTRUMENTO EVALUADO | **FICHA DE OBSERVACIÓN** |
| AUTOR DEL INSTRUMENTO | BACH. JAMES ENRIQUE SEGOVIA HINOJOSA |

### 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** | Esta 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** | **92.5%** |
| **EVALUACIÓN DE FACTIBILIDAD** | **ES APLICABLE** |
| **FECHA DE EVALUACIÓN** | **15/02/2024** |
| **FIRMA DEL EXPERTO** |  |

## VALIDACIÓN DEL INSTRUMENTO POR JUICIO DE EXPERTOS

### 1. DATOS DEL EXPERTO

| | |
|---|---|
| APELLIDO Y NOMBRES | Acero Mamani Nain Neptalí |
| GRADO ACADÉMICO | Bach. En C.S. En Ing. Informática y S. |
| TÍTULO PROFESIONAL | Ing. En Ingeniería y Sistemas |
| EMPRESA / INSTITUCIÓN DONDE LABORA | UNJBG - Informática |
| INSTRUMENTO EVALUADO | Ficha de Observación 01 |
| AUTOR DEL INSTRUMENTO | BACH. JAMES ENRIQUE SEGOVIA HINOJOSA |

### 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. | | | | | ✓ |
| OBJETIVIDAD | Nivel de imparcialidad y ausencia de opiniones personales | | | | ✓ | |
| ORGANIZACIÓN | Esta presentado de forma estructurada y secuencial. | | | | | ✓ |
| SUFICIENCIA | El contenido es adecuado y completo para cumplir los objetivos. | | | | | ✓ |
| PERTINENCIA | El contenido es relevante y apropiado para la investigación. | | | | | ✓ |
| COHERENCIA | Se tiene una relación lógica entre lo que se medirá con el contenido del trabajo. | | | | ✓ | |
| RELEVANCIA | Evalúa la importancia del instrumento. | | | | | ✓ |
| ACTUALIDAD | Adecuado al avance de la tecnología | | | | | ✓ |

### 3. RESULTADOS DE VALIDACIÓN

| | |
|---|---|
| PROMEDIO DE VALIDACIÓN | 91.25 |
| EVALUACIÓN DE FACTIBILIDAD | ES APLICABLE |
| FECHA DE EVALUACIÓN | 18/02/2024 |
| FIRMA DEL EXPERTO | NAIN NEPTALI ACERO MAMANI<br>ING. EN INFORMATICA Y SISTEMAS<br>CIP N° 316165 |