UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018-2022)

TESIS

PRESENTADA POR:

EDILBERTO CALLATA COLLANQUI

JOSUÉ CALIP OJEDA MAMANI

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO - PERÚ

2025

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

EDILBERTO CALLATA COLLANQUI JOSUÉ CALIP O...

ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS ...

My Files
My Files
Universidad Nacional del Altiplano

Detalles del documento

Identificador de la entrega
trn:oid:::8254:553217923

Fecha de entrega
5 feb 2026, 7:10 a.m. GMT-5

Fecha de descarga
5 feb 2026, 7:14 a.m. GMT-5

Nombre del archivo
TESIS-JOSUE-OJEDA-04.docx

Tamaño del archivo
6.1 MB

164 páginas
27.428 palabras
160.310 caracteres

turnitin Página 1 de 184 - Portada Identificador de la entrega trn:oid:::8254:553217923

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

20% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Coincidencias menores (menos de 10 palabras)

Fuentes principales

13% Fuentes de Internet
6% Publicaciones
15% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alerta de integridad para revisión

Texto oculto
26 caracteres sospechosos en N.° de páginas
El texto es alterado para mezclarse con el fondo blanco del documento.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

UNA PUNO
Firmado digitalmente por: JIMENEZ CHURA Adolfo Carlos FAU 20145496170 soft
Motivo: Doy V.B.
Fecha: 05/02/2026 07:15:45
Firmado con FirmaUNA

UNA PUNO
Firmado digitalmente por: JIMENEZ CHURA Adolfo Carlos FAU 20145496170 soft
Motivo: Doy V.B.
Fecha: 05/02/2026 07:16:21
Firmado con FirmaUNA

turnitin Página 2 de 184 - Descripción general de integridad Identificador de la entrega trn:oid:::8254:553217923

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018-2022)

TESIS PRESENTADA POR:

EDILBERTO CALLATA COLLANQUI

JOSUÉ CALIP OJEDA MAMANI

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

APROBADO POR EL JURADO REVISOR CONFORMADO POR:

PRESIDENTE: .....
Dr. ROBERT ANTONIO ROMERO FLORES

PRIMER MIEMBRO: .....
D.Sc. FIDEL ERNESTO TICONA YANQUI

SEGUNDO MIEMBRO: .....
M.Sc. WILDO SUCASAIRE MONROY

ASESOR DE TESIS: .....
D.Sc. ADOLFO CARLOS JIMENEZ CHURA

ÁREA: Sistemas, Computación e Informática

TEMA: Inteligencia Artificial y Sistemas bio-inspirados

FECHA DE SUSTENTACIÓN: 30 de mayo de 2025

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

DEDICATORIA

A Dios, por la vida, la salud y por su guía.

A mis queridos padres Wenceslao Callata y Eusebia Collanqui, se los dedico con todo mi amor y gratitud. No hay palabras suficientes para agradecer su apoyo inquebrantable y el amor incondicional que siempre me han brindado, ustedes son mi ejemplo de perseverancia y mi mayor fuente de inspiración que me motiva superarme día a día. Sin su guía y fortaleza, nada esto sería posible.

A mis hermanos, por su confianza y motivación constante. Este logro también es de ustedes.

Edilberto Callata Collanqui

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

DEDICATORIA

A Dios por darme la voluntad de seguir en esta etapa y me bendice en cada paso que doy.

También se lo dedico especialmente con mucho respeto a mis padres Francisco Ojeda y Florencia Mamani, quienes me apoyaron incondicionalmente, ellos son quienes me dieron grandes enseñanzas y los principales protagonistas para la construcción de mi vida profesional, haciéndome una persona de bien, con todo mi amor, aprecio y afecto se los dedico.

A mis hermanos Yola, Hernán, Hilda y Vilma, que más me han motivado en mi vida, dándome los mejores consejos y guiándome para que siguiera adelante y cumpla mis ideales.

Josué Calip Ojeda Mamani

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

AGRADECIMIENTOS

A Dios por la vida y la bondad de la existencia.

A nuestra alma mater, Universidad Nacional del Altiplano, en especial a la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas, a la plana docente quienes contribuyeron en nuestra formación profesional.

A nuestro asesor de tesis D.Sc. Adolfo Carlos Jimenez Chura, por su invaluable apoyo, orientación constante y paciencia durante el desarrollo de esta investigación.

A nuestros distinguidos jurados, cuyos comentarios y orientaciones han sido pilares fundamentales en el desarrollo y mejora de esta investigación.

En especial, a nuestra familia por su apoyo incondicional y constante motivación, los cuales fueron esenciales para lograr el objetivo propuesto.

Edilberto Callata Collanqui

Josué Calip Ojeda Mamani

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE GENERAL

Pág.

DEDICATORIA

AGRADECIMIENTOS

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

ACRÓNIMOS

RESUMEN **19**

ABSTRACT **20**

CAPÍTULO I

INTRODUCCIÓN

1.1. PLANTEAMIENTO DEL PROBLEMA 23

1.2. FORMULACIÓN DEL PROBLEMA 24

1.2.1. Problema general 24

1.3. JUSTIFICACIÓN DEL ESTUDIO 24

1.4. OBJETIVOS DE LA INVESTIGACIÓN 25

1.4.1. Objetivo general 25

1.4.1.1. Objetivos específicos 25

1.5. HIPÓTESIS DE LA INVESTIGACIÓN 26

1.5.1. Hipótesis general 26

1.5.1.1. Hipótesis nula 26

1.5.1.2. Hipótesis alternativa 26

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

1.6. VARIABLES DE ESTUDIO 27

1.7. OPERACIONALIZACIÓN DE VARIABLES 28

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES 29

2.1.1. Nivel internacional 29

2.1.2. Nivel nacional 39

2.2. MARCO TEÓRICO 43

2.2.1. Minería de datos 43

2.2.2. Inteligencia Artificial (IA) 44

2.2.2.1.Modelos de entrenamiento de IA 45

2.2.3. Análisis predictivo 47

2.2.4. Funcionamiento del análisis predictivo 48

2.2.4.1.Recopilación de datos 49

2.2.4.2.Limpieza y combinación de datos 49

2.2.4.3.Desarrollar el modelo 49

2.2.4.4.Integración de la analítica con sistemas 50

2.2.4.5.Validación de los resultados 50

2.2.5. Tipos de modelos predictivos 50

2.2.5.1.Modelos de clasificación 50

2.2.5.2.Modelos de regresión 58

2.2.6. Métricas en Machine Learning 59

2.2.6.1.Exactitud (Accuracy) 59

2.2.6.2.Precisión (precision) 59

2.2.6.3.Sensibilidad (Recall o tasa de verdaderos positivos) 60

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.6.4.F1 Score 60

2.2.6.5.Especificidad (Specificity o Tasa de Verdaderos Negativos) ..60

2.2.6.6.Matriz de confusión.....60

2.2.6.7.La curva ROC 62

2.2.6.8.Área bajo la curva ROC - AUC 62

2.2.7. Ingeniería de características 63

2.2.7.1.Procesos de la ingeniería de características 64

2.2.7.2.Técnicas de ingeniería de características 67

2.2.8. Norma técnica de ruralidad 71

2.2.9. Factor de Inflación de la Varianza 72

2.3. GLOSARIO DE TÉRMINOS BÁSICOS 73

2.3.1. Python 73

2.3.2. Colaboratory.....74

2.3.3. librerias de python.....74

2.3.4. Scikit-learn 75

2.3.5. Mathplotlib.....76

2.3.6. Dataset.....77

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. ÁREA DE ESTUDIO.....78

3.2. POBLACIÓN 78

3.3. MUESTRA.....80

3.3.1. Postulantes por las modalidades General y CEPREUNA.....80

3.4. TIPO DE INVESTIGACIÓN 81

3.5. DISEÑO DE LA INVESTIGACIÓN 82

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

3.6. TÉCNICAS E INSTRUMENTOS 82

3.6.1. Las técnicas de recolección de datos 82

3.6.1.1. Solicitud de información 82

3.6.2. Lenguaje de programación 83

3.6.3. Herramientas de análisis 83

3.7. TRATAMIENTO DE DATOS 83

3.7.1. Exploración de datos 83

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. OBJETIVO ESPECÍFICO 1 85

4.1.1. Solicitud y acceso a datos 85

4.1.2. Variables recopiladas 85

4.1.2.1. Carácterísticas de las variables 86

4.1.3. Criterio de exclusión 87

4.2. OBJETIVO ESPECÍFICO 2 89

4.2.1. Análisis inicial de las variables detectadas - preprocesamiento 90

4.2.1.1. Aplicando ingeniería de características 94

4.2.1.2. Eliminación de variables innecesarias 96

4.2.1.3. Eliminación de postulantes e ingresantes por discapacidad 97

4.2.1.4. Análisis de postulantes por modalidad y género 98

4.2.1.5. Análisis de ingresantes por género y fecha de proceso 102

4.2.1.6. Análisis del número de postulaciones 107

4.2.1.7. Relación entre la variable edad y puntaje 110

4.2.1.8. Análisis de postulantes e ingresantes por edad 112

4.2.1.9. Análisis de postulante/ingresantes por provincia 114

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.2.1.10.Análisis de post. e ingres. según tipo de colegio.....117

4.2.1.11.Análisis de la distribución de ruralidad de ingresantes.....119

4.2.1.12.Análisis de correlación entre edad y años de egreso.....121

4.2.1.13.Codificación de variables categóricas.....123

4.2.1.14.Codificación de variables categóricas - One-Hot Encodig ..125

4.2.1.15.Análisis de multicolinealidad para la predicción 126

4.2.1.16.Normalización de datos.....130

4.3. OBJETIVO ESPECÍFICO 3.....131

4.3.1. Regresión logística.....132

4.3.2. Árbol de decisión 135

4.3.3. Random Forest 137

4.3.4. XGBoost.....139

4.3.5. K-Nearest Neighbors.....140

4.3.6. Support Vector Machines.....142

4.4. OBJETIVO ESPECÍFICO 4.....144

V. CONCLUSIONES.....148

VI. RECOMENDACIONES 151

VII. REFERENCIAS BIBLIOGRÁFICAS.....153

ANEXOS.....157

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE TABLAS

Pág.

Tabla 1 Operacionalización de variables 28
Tabla 2 Matriz de confusión 61
Tabla 3 Representación ideal para la matriz de confusión 61
Tabla 4 Representación errónea para la matriz de confusión 61
Tabla 5 Precio de propiedades por ciudad 64
Tabla 6 Agregando una columna costo por pie cuadrado 65
Tabla 7 Transformación dato continuo a categórico 65
Tabla 8 Se extrae una nueva característica basada en otras 66
Tabla 9 Cantidad de postulantes por modalidad 78
Tabla 10 Cant. de postulantes por fecha y modalidad examen GENERAL 79
Tabla 11 CEPREUNA Cantidad de postulantes por fecha y modalidad 79
Tabla 12 Cant. de postulantes por modalidad para la muestra 80
Tabla 13 Cant. de postulantes examen General para la muestra 81
Tabla 14 Cant. de postulantes examen CEPREUNA para la muestra 81
Tabla 15 Estructura del dataset de postulantes 86
Tabla 16 Parte de la estructura del dataset 88
Tabla 17 Estructura y tipo de datos de las variables 89
Tabla 18 Resumen de las variables numéricas 91
Tabla 19 Vista de la variable año de egreso 95
Tabla 20 Vista de variables procesadas 96
Tabla 21 Cantidad de postulantes por modalidad 97
Tabla 22 Porcentaje de postulantes por edades 113

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 23 Cantidad de ingresantes por ruralidad 121

Tabla 24 Matriz de correlación de Pearson 122

Tabla 25 Valores de VIF de forma descendente..... 127

Tabla 26 Conjunto de datos normalizado 131

Tabla 27 Métricas obtenidas con el modelo de regresión logística 133

Tabla 28 Métricas obtenidas con el modelo de árbol de decisión 135

Tabla 29 Métricas obtenidas con el modelo de Random Forest..... 137

Tabla 30 Métricas obtenidas con el modelo XGBoost 139

Tabla 31 Métricas obtenidas con el modelo KNN 141

Tabla 32 Métricas obtenidas con el modelo SVM 143

Tabla 33 Resumen de métricas de modelos empleados 145

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE FIGURAS

Pág.

Figura 1 Tipos de aprendizaje automático 47

Figura 2 Proceso de análisis de datos 48

Figura 3 Representación de la función logística 53

Figura 4 Representación del árbol de decisión 54

Figura 5 Predicción de Random Forest 55

Figura 6 Definición del margen entre clases - SVM 57

Figura 7 Representación del algoritmo KNN 57

Figura 8 La curva ROC y su efectividad 63

Figura 9 El lugar de la ingeniería de características 64

Figura 10 Boxplot de edades en el preprocesamiento 92

Figura 11 Boxplot de edades corregidas 93

Figura 12 BoxPlot de año de egreso de postulantes 94

Figura 13 Cant. de postulantes por género, modalidad CEPREUNA 98

Figura 14 Cant. de ingresantes por género, modalidad CEPREUNA 99

Figura 15 Cant. de postulantes por género, modalidad General 100

Figura 16 Cant. de ingresantes por género, modalidad General 101

Figura 17 General - cantidad de ingresantes por género y fecha 103

Figura 18 General - cant. de postulantes por género y fecha de proceso 104

Figura 19 CEPREUNA – cant. de ingresantes por género y fecha 104

Figura 20 CEPREUNA - cant. de post. por género y fecha de proceso 105

Figura 21 Número de intentos para obtener una vacante 108

Figura 22 Intentos de postulación modalidad General 108

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 23 Intentos de postulación modalidad CEPREUNA 109

Figura 24 Ingresantes según intentos de postulación 110

Figura 25 Relación edad y puntaje obtenido 111

Figura 26 Distribución etaria de postulantes por edad 112

Figura 27 Distribución de ingresantes por edad 114

Figura 28 Cantidad de postulantes por provincia 116

Figura 29 Cantidad de ingresantes por provincia 117

Figura 30 Ingresantes de otras provincias 117

Figura 31 Postulantes según tipo de colegio 118

Figura 32 Ingresantes según tipo de colegio 119

Figura 33 Tipos de ruralidad de ingresantes por provincia 121

Figura 34 Correlación entre edad y años de egreso 123

Figura 35 Valores de VIF de cada una de las variables de estudio 129

Figura 36 Valores de VIF excluido la variable edad 130

Figura 37 Matriz de confusión del modelo de regresión logística 133

Figura 38 Visualización de la curva ROC de la regresión logística 134

Figura 39 Matriz de confusión del modelo de árbol de decisión 136

Figura 40 Matriz de confusión del modelo de Random Forest 138

Figura 41 Matriz de confusión del modelo de XGBoost 140

Figura 42 Matriz de confusión del modelo de KNN 141

Figura 43 Matriz de confusión del modelo de SVM 143

Figura 44 Gráfico de barras de modelos empleados 146

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE ANEXOS

Pág.

Anexo 1 Documento presentado a la Dirección de Admisión 157

Anexo 2 Funciones para el cálculo de la curva ROC..... 158

Anexo 3 Función para graficar la matriz de confusión 159

Anexo 4 Cantidad de postulantes por modalidad y fecha 160

Anexo 5 Declaración jurada de autenticidad de tesis..... 161

Anexo 6 Autorización para el depósito de tesis en el Repositorio Institucional.... 163

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ACRÓNIMOS

IA:	Inteligencia Artificial
SVM:	Máquina de Soporte Vectorial
KNN:	K-Nearest Neighbors
ROC:	Receiver Operating Characteristic
AUC:	Area Under Curve

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

RESUMEN

La presente investigación tiene como objetivo desarrollar, entrenar y evaluar modelos de aprendizaje automático para predecir el ingreso de postulantes a la Escuela Profesional de Ingeniería de Sistemas, utilizando métricas de rendimiento como precisión, recall, F1-score y la curva ROC (AUC). Se recopiló un conjunto de datos de los procesos de admisión de los años 2018 al 2021, que se utilizó para el entrenamiento y validación de los modelos, mientras que los datos del año 2022 se emplearon para realizar predicciones y evaluar el desempeño real de los modelos en datos no vistos. Se exploraron técnicas de preprocesamiento de datos, como normalización, selección de características y balanceo de clases, con el fin de optimizar la calidad del dataset. Se implementaron y compararon seis modelos de clasificación: Regresión Logística, Árboles de decisión, Random Forest, XGBoost, K-Nearest Neighbors (KNN) y Support Vector Machines (SVM). El modelo SVM presentó la mayor precisión global con un 92% de exactitud, mientras que la Regresión Logística obtuvo un desempeño equilibrado con un AUC de 72%. Finalmente, se validaron los modelos con los datos de admisión del año 2022, observándose una disminución en el rendimiento predictivo, lo que sugiere la necesidad de una actualización constante de los modelos con datos recientes.

Palabras clave: Análisis exploratorio, Aprendizaje supervisado, Inteligencia artificial, Métricas de evaluación, Procesos de admisión.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ABSTRACT

The present research aims to develop, train and evaluate machine learning models to predict the admission of applicants to the Professional School of Systems Engineering, using performance metrics such as accuracy, recall, F1-score and the ROC curve (AUC). A dataset was collected from the 2018 to 2021 admissions processes, which was used for training and validation of the models, while the 2022 data was used to make predictions and evaluate the actual performance of the models on unseen data. Data preprocessing techniques, such as normalization, feature selection, and class balancing, were explored to optimize dataset quality. Six classification models were implemented and compared: Logistic Regression, Decision Trees, Random Forest, XGBoost, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). The SVM model presented the highest overall accuracy with 92% accuracy, while Logistic Regression obtained a balanced performance with an AUC of 0.72. Finally, the models were validated with the 2022 intake data, showing a decrease in predictive performance, suggesting the need for constant updating of the models with recent data.

Keywords: Exploratory analysis, Supervised learning, Artificial intelligence, Evaluation metrics, Admission processes.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPÍTULO I

INTRODUCCIÓN

La admisión a programas universitarios altamente competitivos, representa un desafío significativo tanto para la diversidad de postulantes como para las instituciones universitarias estatales en nuestro país (Chiny et al., 2008). Tradicionalmente, los procesos de selección se han basado en métodos que, aunque bien establecidos, pueden no capturar adecuadamente la información completa de los postulantes para realizar un proceso de análisis de datos. En este contexto, el uso de técnicas avanzadas de análisis de datos (Arnau, 1996) y aprendizaje automático ofrece una oportunidad sin precedentes para mejorar la precisión y objetividad en la predicción de los resultados de los exámenes de admisión.

El uso de técnicas de Machine Learning en los procesos de admisión se ha vuelto esencial para mejorar la eficiencia y equidad en la selección de postulantes. Al aplicar modelos predictivos y análisis avanzados, las instituciones pueden identificar patrones y tendencias en los datos demográficos, lo que permite una evaluación más precisa y justa de los ingresantes (Brasca et al., 2022), por ejemplo, se pueden analizar atributos históricos para entender las características de los estudiantes que tienen más probabilidades de tener éxito, permitiendo a las universidades tomar decisiones más informadas y personalizadas sobre los postulantes.

Esta investigación se centra en el análisis predictivo de los resultados de los exámenes de admisión a la Escuela de Ingeniería de Sistemas durante el período comprendido entre 2018 y 2022. El objetivo principal es desarrollar y evaluar modelos de aprendizaje automático que, utilizando variables demográficas de los postulantes, permitan predecir con alta precisión los resultados de ingresantes de los exámenes. Estos

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

modelos serán evaluados específicamente con los datos del año 2022, con la finalidad de validar su rendimiento y precisión.

La metodología de investigación incluye la recopilación y preparación de los datos, un análisis exploratorio inicial para identificar relaciones significativas entre las variables, y el desarrollo y entrenamiento de modelos de aprendizaje automático. La efectividad de estos modelos se evaluó utilizando métricas de precisión y rendimiento (Sicilia Gómez, 2024), y los resultados se validaron con los datos del año 2022.

En el primer capítulo se plantea el problema de la investigación, que busca entender los motivos detrás de la admisión a la Escuela Profesional de Ingeniería de Sistemas mediante modelos de aprendizaje automático. Se establecen los objetivos generales y específicos y variables de estudio. También se describen las limitaciones encontradas durante el desarrollo de la investigación.

En el segundo capítulo se revisa los antecedentes a nivel nacional e internacional sobre el uso de técnicas de aprendizaje automático para predecir resultados académicos. Se destacan las metodologías y hallazgos relevantes que aportan al desarrollo de esta investigación. Además, se definen los conceptos teóricos esenciales para la comprensión de los modelos de aprendizaje automático y su aplicación en la predicción de resultados de exámenes de admisión.

En el tercer capítulo se describe el área de estudio, la población y muestra utilizada, así como el tipo y diseño de la investigación. Se detallan las técnicas e instrumentos de recolección de datos, el tratamiento de los datos y el análisis exploratorio realizado. También se explican las herramientas de análisis y el lenguaje de programación utilizado para implementar los modelos predictivos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

En el cuarto capítulo se presentan los resultados obtenidos, acompañados de su interpretación y análisis. Se utilizan tablas y gráficos para ilustrar los hallazgos y facilitar su comprensión. Los resultados se organizan de acuerdo con cada objetivo específico planteado, mostrando el desarrollo de los modelos predictivos y su evaluación.

En el quinto y sexto capítulo se resumen los hallazgos más importantes de la investigación, destacando la utilidad de los modelos de aprendizaje automático para predecir los resultados de los exámenes de admisión. Se discuten las implicaciones de los resultados y se proponen recomendaciones para mejorar la precisión y eficiencia de los modelos.

1.1. PLANTEAMIENTO DEL PROBLEMA

Los procesos de admisión a las universidades estatales periódicamente se dan y contemplan una serie de etapas y muchas de ellas comparten actividades en común y criterios propios de admisión donde se obtiene información demográfica del postulante y con su respectivo puntaje de ingreso para cada universidad (UNA, 2024).

Los procesos de admisión en la Universidad Nacional del Altiplano de Puno se realizan periódicamente y constan de varias etapas. Actualmente los métodos tradicionales de recojo de información pueden no considerar adecuadamente las diversas variables que influyen en el desempeño de los postulantes, como sus antecedentes demográficos (Gay et al., 2018). Por otro lado, la falta de un enfoque basado en datos completos y exhaustivos limita la capacidad de los modelos actuales para generalizar y predecir con precisión los resultados de los exámenes de admisión. Finalmente, la rápida evolución de las tecnologías de aprendizaje automático y la disponibilidad creciente de datos ofrecen una oportunidad sin precedentes para mejorar estos procesos, sin embargo, los modelos predictivos podrían identificar patrones y tendencias emergentes,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

proporcionando información valiosa para futuras investigaciones y ajustes en los procesos de admisión (Sekeroglu et al., 2019).

La presente tesis se centra en el análisis predictivo de los resultados de ingresantes de los exámenes de admisión a la Escuela Profesional de Ingeniería de Sistemas, empleando modelos de aprendizaje automático a las variables demográficas durante el período comprendido entre 2018 y 2022 y evaluarlos con los datos del año 2022.

1.2. FORMULACIÓN DEL PROBLEMA

1.2.1. Problema general

¿Cómo pueden los modelos de aprendizaje automático ser utilizados para predecir con precisión los resultados de los exámenes de admisión a la Escuela de Ingeniería de Sistemas con datos demográficos comprendidos entre 2018 y 2022?

1.3. JUSTIFICACIÓN DEL ESTUDIO

El análisis predictivo de los resultados de los exámenes de admisión a la Escuela de Ingeniería de Sistemas mediante modelos de aprendizaje automático entre los años 2018 y 2022 es crucial por varias razones, en primer lugar, permite a la institución identificar patrones y tendencias en los datos demográficos de los postulantes. Al comprender mejor estos factores, se garantiza que los postulantes seleccionados no sólo tengan un alto potencial académico, sino que también reflejen la diversidad de la sociedad.

En segundo lugar, los modelos de aprendizaje automático ofrecen una precisión y eficiencia superiores en la predicción de los resultados de admisión en comparación con los métodos tradicionales. Además, el uso de estas tecnologías puede proporcionar una retroalimentación valiosa para los postulantes y las políticas educativas. Por ejemplo, al

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

identificar qué características demográficas están más fuertemente asociadas con el éxito en el proceso de admisión, la Escuela Profesional de Ingeniería de Sistemas puede ofrecer mejores orientaciones y preparaciones para futuros ingresantes, contribuyendo así a elevar el nivel general de los ingresantes.

Finalmente, el período 2018-2022 incluye varios eventos y cambios sociales significativos, como la pandemia de COVID-19, que pueden haber afectado los perfiles y desempeños de los postulantes. Analizar estos años permite a la institución entender cómo tales eventos impactan en la admisión y adaptar sus estrategias para futuros desafíos.

Este estudio permite comprender cómo las características demográficas afectan la retención estudiantil y puede ayudar a desarrollar intervenciones personalizadas para mejorar la experiencia académica y reducir en un futuro la tasa de deserción usando tecnología avanzada en educación superior, promoviendo la cultura de innovación y mejora continua.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. Objetivo general

Desarrollar y analizar modelos de aprendizaje automático que permitan predecir los resultados de los exámenes de admisión a la Escuela Profesional de Ingeniería de Sistemas, utilizando para ello variables demográficas de los postulantes.

1.4.1.1. Objetivos específicos

- Recopilar datos demográficos de los postulantes entre 2018 y 2022.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- Realizar un análisis exploratorio para identificar patrones y tendencias relevantes.
- Desarrollar, entrenar y evaluar modelos de aprendizaje automático utilizando métricas de rendimiento.
- Validar los modelos con los datos del año 2022 y proporcionar recomendaciones basadas en los resultados obtenidos.

1.5. HIPÓTESIS DE LA INVESTIGACIÓN

1.5.1. Hipótesis general

El uso de modelos de aprendizaje automático permite identificar de forma significativa las características demográficas más influyentes en los resultados de los exámenes de admisión a la Escuela Profesional de Ingeniería de Sistemas, mejorando así la capacidad predictiva y la toma de decisiones en el proceso de admisión entre 2018 y 2022.

1.5.1.1. Hipótesis nula

El uso de modelos de aprendizaje automático no permite identificar con precisión las características demográficas más influyentes en los resultados de los exámenes de admisión a la Escuela Profesional de Ingeniería de Sistemas, y no mejora la capacidad predictiva ni la toma de decisiones en el proceso de admisión entre 2018 y 2022.

1.5.1.2. Hipótesis alternativa

El uso de modelos de aprendizaje automático permite identificar con precisión las características demográficas más influyentes en los resultados de los exámenes de admisión a la Escuela Profesional de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Ingeniería de Sistemas, mejorando así la capacidad predictiva y la toma de decisiones en el proceso de admisión entre 2018 y 2022.

1.6. VARIABLES DE ESTUDIO

Variable dependiente

Resultado de los exámenes de admisión

Variable independiente

Características demográficas

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

1.7. OPERACIONALIZACIÓN DE VARIABLES

Tabla 1

Operacionalización de variables

Variable	Tipo	Dimensión	Indicadores	Escala de medición	Técnicas e instrumentos
Modelos de aprendizaje automático	Independ.	Tipo de modelo aplicado Métricas de validación	- Regresión logística - Árbol de decisión - Random forest - SVM - KNN - XGBoost	Categórica nominal	Programación con Python (Scikit-learn, mathplotlib, pandas)
Resultados de exámenes de admisión	Depend.	Resultado del proceso Calidad de predicción	- ingresó = 1/ no ingresó = 0 - Métricas: accuracy, precisión, recall, F1-Score, AUC	Categórica binaria proporcional (0 a 1)	Base de datos de postulantes, scripts de validación, métricas sklearn

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES

En este capítulo se revisa los antecedentes a nivel nacional e internacional que han empleado técnicas de aprendizaje automático para predecir resultados académicos, destacando las metodologías y hallazgos relevantes que aportan al desarrollo de esta investigación. Además, se definirán los conceptos teóricos esenciales para la comprensión de los modelos de aprendizaje automático y su aplicación en la predicción de resultados de exámenes de admisión, proporcionando un marco conceptual robusto que sustente la investigación.

2.1.1. Nivel internacional

Los investigadores Contreras et al., (2020) propusieron una metodología sobre las variables indicando que:

> “Esta propuesta plantea la selección de variables que influyen en la predicción del rendimiento en estudiantes de ingeniería industrial de la Universidad Distrital (Colombia) por diferentes metodologías: filtro, envoltura e integrados. Se implementaron algoritmos de clasificación a través del lenguaje de programación Python como árbol de decisión, K vecinos más cercanos, perceptrón y otros, los cuales son comparados para conocer el mejor resultado de predicción. El género y el puntaje ICFES (examen de estado en Colombia) para condición matemática se encuentran en el rango superior de todos los métodos de selección de características, y el algoritmo perceptrón arroja mejor exactitud con respecto a los otros

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

algoritmos usados. Se concluye que las variables que más influyen en el rendimiento académico de los estudiantes de ingeniería son: edad, género, puntaje ICFES para aptitud matemática, puntaje global ICFES, valor de matrícula y puntaje ICFES para condición matemática y cohorte".

La investigación de Zaffar et al. (2018) sobre todas las organizaciones educativas tuvo como objetivo el mejoramiento de la calidad de la educación y elevar el rendimiento académico de los estudiantes. La minería de datos educativos (EDM) es un campo de investigación en auge que ayuda a las instituciones académicas a mejorar el rendimiento de sus estudiantes. Las instituciones académicas suelen ser juzgadas por las calificaciones obtenidas por los estudiantes en los exámenes. EDM ofrece diferentes prácticas para predecir el rendimiento académico de los estudiantes. En EDM, la selección de características desempeña un papel fundamental en la mejora de la calidad de los modelos de predicción para conjuntos de datos educativos. Los algoritmos de FS eliminan los datos no relacionados de los repositorios educativos y, por lo tanto, aumentan el rendimiento de la precisión del clasificador utilizado en diferentes prácticas de EDM para apoyar la toma de decisiones en entornos educativos. La buena calidad del conjunto de datos educativos puede producir mejores resultados y, por lo tanto, las decisiones basadas en dicho conjunto de datos de calidad pueden aumentar la calidad de la educación mediante la predicción del rendimiento de los estudiantes. A la luz de este hecho, es necesario elegir cuidadosamente un algoritmo de selección de características. Este artículo presenta un análisis del rendimiento de algoritmos de selección de características de filtrado y algoritmos de clasificación en dos conjuntos de datos de estudiantes diferentes. Los resultados obtenidos a partir de diferentes algoritmos de FS y clasificadores en dos conjuntos de datos de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

estudiantes con diferente número de características también ayudarán a los investigadores a encontrar las mejores combinaciones de algoritmos de selección de características de filtro y clasificadores. Es muy necesario arrojar luz sobre la relevancia de la selección de características para la predicción del rendimiento de los estudiantes, ya que las estrategias educativas constructivas pueden derivarse a través del conjunto relevante de características. Los resultados de nuestro estudio muestran que existe una diferencia del 10% en la precisión de la predicción entre los resultados de los conjuntos de datos con distinto número de características.

García et al., (2019) realizan el "análisis de predicción y resultados de la prueba de calidad de la educación superior en Colombia - Saber Pro. Se utilizó la metodología de extracción de conocimiento en bases de datos KDD, sobre la cual se construyó una base de datos del rendimiento académico de los estudiantes en áreas asociadas a los contenidos de la prueba Saber Pro, y como técnica de minería de datos se utilizó redes neuronales. Las redes neuronales permitieron la predicción de los resultados de la prueba Sabre Pro con alta exactitud tanto en rangos cualitativos como cuantitativos. También se encontró una correlación entre el rendimiento académico y los resultados de Saber Pro. Los hallazgos sugieren que la metodología utilizada es una excelente guía para el descubrimiento de patrones ocultos en los datos, y permite establecer estrategias para mejorar los resultados de las pruebas Saber Pro que involucren el rendimiento académico del estudiante".

Un sistema de admisión basado en criterios de admisión válidos y confiables es muy importante para seleccionar candidatos que puedan tener un buen desempeño académico en las instituciones de educación superior. Este trabajo aborda estrategias para respaldar a las universidades en sus decisiones de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

admisión, aplicando técnicas de minería de datos para anticipar el desempeño académico de los aspirantes. Se empleó un conjunto de datos de 2039 estudiantes inscritos entre 2016 y 2019 en una Facultad de Ciencias de la Computación e Información de una universidad pública de Arabia Saudita para validar la metodología propuesta. Los hallazgos indican que el desempeño inicial de los estudiantes en la universidad puede ser estimado antes de su admisión, considerando ciertos factores previos, como el promedio de calificaciones en la escuela secundaria, la puntuación en la prueba de logros académicos y el resultado en el examen de aptitud general. Los resultados también muestran que el puntaje de la prueba de admisión de rendimiento académico es el criterio previo a la admisión que predice con mayor precisión el rendimiento futuro de los estudiantes. Por lo tanto, este puntaje debería tener más peso en los sistemas de admisión. También encontramos que la técnica de Redes Neuronales Artificiales tiene una tasa de precisión superior al 79%, lo que la hace superior a otras técnicas de clasificación consideradas como árboles de decisión, máquinas de soporte vectorial y Naïve Bayes. (Mengash, 2020)

Predecir el desempeño de los estudiantes se vuelve más desafiante debido al gran volumen de datos en las bases de datos educativas. Actualmente en Malasia, no se está abordando la falta de un sistema existente para analizar y monitorear el progreso y el desempeño de los estudiantes. Hay dos razones principales de por qué esto está sucediendo. En primer lugar, la investigación actual sobre los métodos de predicción sigue siendo insuficiente para determinar cuáles son los más efectivos para estimar el rendimiento académico de los estudiantes en las instituciones educativas de Malasia. En segundo lugar, esto se debe a la escasez de estudios que analicen los factores que influyen en el éxito de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

los estudiantes en cursos específicos dentro del contexto malasio. Por ello, se plantea una revisión sistemática de la literatura sobre la predicción del desempeño estudiantil, empleando técnicas de minería de datos con el objetivo de optimizar los logros académicos. El principal objetivo de este artículo es proporcionar una visión general de las técnicas de minería de datos que se han utilizado para predecir el rendimiento de los estudiantes. Este documento también se enfoca en cómo se puede usar el algoritmo de predicción para identificar los atributos más importantes en los datos de los estudiantes. De hecho, podríamos mejorar los logros y el éxito de los estudiantes de manera más efectiva y eficiente utilizando técnicas educativas de minería de datos. Podría traer beneficios e impactos a estudiantes, educadores e instituciones académicas. (Shahiri et al., 2015)

El objetivo principal de este estudio es aplicar técnicas de minería de datos para predecir y analizar el rendimiento académico de los estudiantes y la participación en foros. La minería de datos educativos (EDM) es una herramienta emergente para la intervención académica. Las instituciones educativas pueden utilizar EDU para un amplio análisis de las características de los estudiantes. En este estudio hemos recogido datos de estudiantes de dos licenciaturas. Tres algoritmos de clasificación de minería de datos de datos (Naïve Bayes, red neuronal y árbol de decisión). La predicción de los tres clasificadores es comparada y se observa que el clasificador Naïve Bayes supera a los otros dos clasificadores alcanzando una precisión de predicción del 86%. Este estudio ayudará a los profesores a mejorar el rendimiento académico de los alumnos. (Mueen et al., 2016)

En el contexto de la creciente necesidad de científicos e ingenieros calificados, el flujo heterogéneo de estudiantes que ingresan a los programas de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ciencia e ingeniería es particularmente desafiante en las universidades con un sistema de admisión abierta. El objetivo principal del presente estudio es determinar los principales determinantes académicos y no académicos del éxito de los estudios en un programa de estudios STEM en la universidad más grande de Flandes (Bélgica). El Inventario de Estrategias de Estudio y Aprendizaje (LASSI), complementado con preguntas de antecedentes adicionales, fue completado pqor 1521 estudiantes de ciencias e ingeniería de primer año al comienzo del año académico. Para evaluar el valor incremental de un predictor particular para explicar el GPA del primer año, se evaluaron una serie de modelos de regresión anidados. El nivel de matemáticas y el GPA en matemáticas/ciencias en la escuela secundaria estaban fuertemente relacionados con el GPA del primer año. El análisis del cuestionario LASSI mostró que la motivación/persistencia, la concentración y las habilidades de gestión del tiempo de los estudiantes al principio influyeron significativamente en el rendimiento de los estudiantes al final del primer año, aunque el valor incremental sobre el rendimiento anterior fue pequeño. En conjunto, nuestros resultados muestran que la capacidad de los estudiantes entrantes para regular sus esfuerzos de estudio tiene consecuencias beneficiosas en términos de rendimiento. Además, una recomendación negativa por parte de la junta de maestros de la escuela secundaria fue un claro indicador para identificar a los estudiantes en riesgo. En las universidades de admisión abierta en las que a los nuevos estudiantes no se les puede negar formalmente el acceso en base a un bajo rendimiento previo en matemáticas y ciencias, un enfoque en el entrenamiento de habilidades de autorregulación relacionadas con el esfuerzo (por ejemplo, sesiones de gestión del tiempo) ofrece oportunidades valiosas para intervenciones correctivas. (Pinxten et al., 2019)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Gupta et al., (2016) en su investigación, estudian el proceso de admisión en las universidades estadounidenses desde la perspectiva de los estudiantes. Su objetivo es construir un modelo de apoyo a la toma de decisiones que proporcione a los candidatos información pertinente, así como la capacidad de evaluar sus opciones durante el proceso de solicitud. Este modelo se basa en un análisis exhaustivo basado en el aprendizaje automático de grandes cantidades de datos históricos disponibles en la web. Su análisis tiene en cuenta factores como las puntuaciones de los exámenes estandarizados y el GPA, así como el conocimiento mundial, como la reputación de la universidad. El problema de aprendizaje se modela como un problema de clasificación binaria con variables latentes que dan cuenta de información oculta, como múltiples programas de posgrado dentro de la misma institución. Una aportación adicional de este trabajo es la recopilación de un nuevo conjunto de datos de más de 25.000 estudiantes, con una media de 6 solicitudes por estudiante y, por tanto, más de 150.000 solicitudes que abarcan más de 3.000 instituciones de origen. El conjunto de datos abarca cientos de universidades de destino a lo largo de varios años, y nos permite desarrollar modelos que proporcionan información sobre el comportamiento de solicitud de los estudiantes y los patrones de decisión de las universidades. Nuestro estudio experimental revela algunos factores clave en el proceso de decisión de los programas que proporcionan a los solicitantes la capacidad de tomar una decisión informada durante la solicitud, con una alta confianza de ser aceptados.

La investigación de Acharya et al., (2019) sobre futuros estudiantes de posgrado siempre se enfrentan a un dilema al decidir las universidades de su elección al postularse para los programas de maestría. Si bien hay un buen número de predictores y consultorías que guían a un estudiante, no siempre son confiables

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ya que la decisión se toma sobre la base de admisiones pasadas seleccionadas. En este artículo, presentan un método basado en Machine Learning donde se compara diferentes algoritmos de regresión, como Regresión lineal, Regresión de vectores de soporte, Árboles de decisión y Bosque aleatorio, dado el perfil del estudiante. Luego se calcula las funciones de error para los diferentes modelos y comparamos su rendimiento para seleccionar el modelo con mejor rendimiento. Los resultados indican entonces si la universidad elegida es ambiciosa o segura.

Según la investigación de Khan et al., (2021) indican que una de las aplicaciones clave importantes de la analítica del aprendizaje es ofrecer una oportunidad a las instituciones de realizar un seguimiento de las actividades académicas de los estudiantes y brindarles consultas adaptativas en tiempo real si el rendimiento académico del estudiante se desvía hacia un resultado inadecuado. Aun así, existen numerosas barreras al desarrollar e implementar este tipo de aplicaciones de análisis de aprendizaje. Los algoritmos de aprendizaje automático emergen como herramientas útiles para respaldar la analítica del aprendizaje mediante la construcción de modelos capaces de pronosticar el resultado final de los estudiantes en función de sus atributos disponibles. El rendimiento del algoritmo de aprendizaje automático disminuye con el uso de todos los atributos y, por lo tanto, una selección atenta de atributos de predicción aumenta el rendimiento del modelo producido. Aunque existen varias técnicas constructivas que facilitan la identificación del subconjunto de atributos productivos, la tarea desafiante es evaluar si los atributos de predicción son significativos, explícitos y controlables por los estudiantes. Este artículo revisa la literatura existente para encontrar los atributos de los estudiantes utilizados en el desarrollo de modelos de predicción. Proponemos un marco conceptual que demuestra la clasificación de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

atributos como latentes o dinámicos. Los atributos latentes pueden parecer significativos pero el estudiante no es capaz de controlarlos; por otro lado, el estudiante tiene el mando para restringir los atributos dinámicos. Cada una de las clases principales se clasifica aún más para presentar una oportunidad a los investigadores de elegir atributos constructivos para el desarrollo del modelo.

Según la investigación de Bolaño y Duarte (2024) el uso de la inteligencia artificial (IA) en la educación ha cobrado relevancia en los últimos años, especialmente para mejorar la personalización del aprendizaje, la retroalimentación y la evaluación de los estudiantes. Por otro lado, existen desafíos y limitaciones asociados a su implementación. Este trabajo se propuso identificar las principales tendencias y áreas de aplicación de la IA en la educación, además de analizar sus beneficios y limitaciones. Mediante una revisión sistemática que siguió una metodología de investigación en cinco etapas, se utilizaron Scopus como fuente primaria y VOSviewer para analizar los resultados. Los resultados indican que la inteligencia artificial optimiza de manera notable la adaptación del aprendizaje, ofreciendo sugerencias y evaluaciones ajustadas a cada persona. No obstante, aún existen retos como la fiabilidad de los datos empleados por la IA, la formación requerida para docentes y alumnos, además de inquietudes sobre la protección y confidencialidad de la información académica. Es esencial continuar analizando el uso de la IA en la enseñanza para garantizar su implementación adecuada y ética.

La investigación de Bravo et al. (2021) destacan que, en los últimos años, el ámbito educativo ha incorporado modelos y algoritmos de análisis de datos con el propósito de extraer conocimiento útil que contribuya a mejorar el rendimiento académico y otros indicadores relevantes. El estudio se centró en la predicción del

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

desempeño académico de estudiantes mediante técnicas de aprendizaje automático. Para ello, se analizaron 324 variables utilizando métodos de selección de características, con el fin de identificar aquellas con mayor influencia. El modelo predictivo fue evaluado mediante algoritmos supervisados como KNN, SVC, Naive Bayes y árboles de decisión, optimizados a través del lenguaje de programación Python. Asimismo, se implementaron técnicas de ensamble para mejorar la precisión de los clasificadores, incluyendo métodos Bagging (como CART, Random Forest y ExtraTreesClassifier), Boosting (AdaBoost, GBM, XGBoost, CatBoost y LightBoost) y Voting (Blending y Stacking). Los resultados más destacados se obtuvieron con los algoritmos de Stacking y Blending, alcanzando niveles de precisión aproximados del 85 % en entrenamiento y 75 % en pruebas.

En el artículo científico de Menchaca et al. (2024) examinan la predicción del rendimiento académico universitario mediante técnicas de aprendizaje automático, destacando investigaciones previas que han identificado diversos factores y algoritmos relevantes en este campo. El estudio detalla el conjunto de datos utilizado, el proceso de preprocesamiento y los algoritmos aplicados, entre los que se incluyen regresión logística, árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM), redes neuronales y modelos de ensamblaje. Se subraya la importancia de la validación cruzada y de una evaluación rigurosa del desempeño de los modelos. Los resultados revelan que tanto los modelos de ensamblaje como las redes neuronales ofrecieron los mejores niveles de precisión y generalización. Entre las variables predictoras más significativas se identificaron el promedio general (GPA), las calificaciones previas y la asistencia a clases. Las conclusiones del estudio resaltan la aplicabilidad práctica de los

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

modelos desarrollados, al tiempo que reconocen sus limitaciones y proponen líneas futuras de investigación, como el desarrollo de modelos específicos por disciplina académica y la mejora en la interpretabilidad de los algoritmos.

2.1.2. Nivel nacional

En el trabajo realizado por Gutiérrez et al. (2022), se analizaron diversas técnicas de aprendizaje automático —incluyendo árboles de decisión, bosques aleatorios, redes neuronales y máquinas de soporte vectorial— con el objetivo de predecir el rendimiento académico estudiantil. Los resultados indicaron que el algoritmo de bosques aleatorios obtuvo la mayor precisión en las predicciones. Esta investigación resulta significativa para las instituciones educativas, ya que puede contribuir a la formulación de políticas orientadas al seguimiento y apoyo de estudiantes en riesgo académico. Además, se evaluó el potencial de aplicación de estos algoritmos en distintas carreras de la Facultad de Ciencias Empresariales y Económicas de la Universidad de Lima.

La investigación de Candia (2019) desarrolló un estudio cuyo objetivo principal fue anticipar el rendimiento académico de los estudiantes de primer semestre en la Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), utilizando como base los datos obtenidos durante el proceso de admisión. Esta predicción temprana resulta valiosa para que las instituciones universitarias puedan implementar estrategias que favorezcan tanto a los docentes como a los propios estudiantes en el proceso de enseñanza-aprendizaje. Para ello, se aplicaron técnicas de aprendizaje automático, una rama de la inteligencia artificial, siguiendo la metodología CRISP-DM, ampliamente utilizada en proyectos de minería de datos. El análisis se llevó a cabo mediante la plataforma de software

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

libre WEKA, que permitió comparar el desempeño de distintos algoritmos de clasificación supervisada. Entre los algoritmos evaluados se encuentran J48 (árboles de decisión), Random Forest, K-Vecinos más cercanos (KNN), regresión logística y perceptrón multicapa. De estos, el algoritmo Random Forest obtuvo el mejor desempeño predictivo con una precisión del 69.35 %, seguido por la regresión logística con un 68.33 %. Además, se identificaron como factores influyentes en el rendimiento académico la nota de ingreso, la escuela profesional, el semestre, la modalidad de ingreso y el número de cursos matriculados.

El trabajo de investigación de Perez y Rojas (2020) se enfocó en la predicción de la deserción estudiantil en la sede central de la Universidad Tecnológica del Perú (UTP), ubicada en Lima. Utilizando técnicas de aprendizaje automático, el estudio se basó en los datos de estudiantes que abandonaron sus estudios en semestres anteriores para identificar patrones de comportamiento asociados a la deserción. A partir de estos patrones, fue posible anticipar qué estudiantes presentaban un alto riesgo de abandonar sus estudios, lo que permitiría a la universidad intervenir oportunamente. Este aporte resulta significativo, ya que facilita la formulación de estrategias institucionales orientadas a retener a los estudiantes y brindarles orientación académica. La metodología empleada fue CRISP-DM, y el algoritmo de máquinas de soporte vectorial (SVM) fue utilizado para modelar y predecir la probabilidad de deserción con base en los comportamientos observados.

La investigación de Ponce Guizabalo (2024) tuvo como finalidad desarrollar un modelo de aprendizaje automático para predecir el rendimiento académico de los estudiantes ingresantes a la Facultad de Ingeniería en Industrias Alimentarias de la UNAS. La hipótesis planteada sostenía que dicha predicción,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

basada en modelos de aprendizaje automático, sería significativa. Para la recolección de datos se utilizó una ficha de análisis documental, obteniéndose un total de 204 registros. El análisis se llevó a cabo mediante el software WEKA 3.9.5, evaluando cinco modelos: Vote, k-Vecinos más Cercanos (IBK), Random Forest, Naive Bayes y Bagging. Aplicando la metodología CRISP-DM y técnicas de minería de datos, se logró establecer un modelo predictivo eficaz. Los resultados indicaron que las variables más influyentes en el rendimiento académico fueron la opción de ingreso, la nota de ingreso (siendo esta la de mayor peso), el sexo del estudiante y el número de hermanos. Mediante la prueba de Friedman ANOVA, se determinó que los modelos Random Forest e IBK alcanzaron una precisión del 98.4 %, siendo Random Forest el más preciso para predecir el rendimiento de los estudiantes aprobados (98.34 %).

En la investigación realizada por Ayrton et al. (2021) señalan que el análisis de datos en el ámbito educativo representa un desafío considerable, ya que requiere investigaciones profundas para comprender la diversidad del estudiantado que accede a la educación superior. Este conocimiento es esencial para diseñar estrategias pedagógicas que fortalezcan el proceso de enseñanza-aprendizaje. El objetivo de su estudio fue caracterizar el perfil de los estudiantes ingresantes a una universidad pública, considerando variables sociodemográficas, económicas y académicas. Para ello, se empleó el algoritmo K-prototypes, utilizando datos obtenidos del examen de admisión, la ficha del ingresante y los certificados escolares de los estudiantes de la Universidad Nacional Agraria La Molina (Lima, Perú). El análisis permitió identificar cinco perfiles distintos entre los ingresantes, cada uno con características particulares. Esta segmentación contribuye a la formulación de políticas de acompañamiento más efectivas,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

promoviendo una educación personalizada y mejoras en la calidad de los entornos de aprendizaje, adaptados a las necesidades de cada perfil estudiantil.

Chavez et al. (2021) destacan que, en el contexto de la educación superior, la gestión de datos se ha vuelto esencial para la toma de decisiones académicas y la mejora continua de los procesos educativos. En este sentido, la analítica y la estadística han sido integradas al entorno tecnológico, priorizando la automatización y el manejo de grandes volúmenes de datos mediante algoritmos de aprendizaje automático. Entre estos, los algoritmos de agrupamiento (clustering) son ampliamente utilizados para clasificar datos según similitudes. El objetivo de su estudio fue identificar distintos tipos de estudiantes universitarios a partir de variables sociodemográficas, económicas y académicas, utilizando el algoritmo K-medoid. El análisis se realizó con datos de ingresantes a la Universidad Nacional Agraria La Molina (Lima, Perú), recolectados durante el proceso de admisión. Los resultados permitieron segmentar a los estudiantes en tres grupos con características diferenciadas, lo que facilita la implementación de estrategias educativas personalizadas y la mejora de la calidad de los espacios de enseñanza en función del perfil de cada estudiante.

Paredes y Zuazo (2014) señalan que el interés por acceder a la educación superior en el Perú ha ido en aumento, lo que ha intensificado la competencia entre las instituciones que la ofrecen. Esta tendencia también se refleja en la carrera de Medicina Veterinaria y Zootecnia, lo que hace necesario enfocar los esfuerzos de marketing hacia el público objetivo con mayor probabilidad de optar por esta carrera, optimizando así los recursos destinados a la promoción. En este contexto, el estudio tuvo como propósito identificar las características de los estudiantes ingresantes a la Facultad de Medicina Veterinaria y Zootecnia de la Universidad

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Peruana Cayetano Heredia (FAVEZ-UPCH) durante el periodo 2010–2013. Para ello, se construyó una base de datos en Microsoft Excel a partir de la información proporcionada por la Oficina de Admisión, Matrícula y Registro Académico (OFAD). Se aplicaron técnicas de estadística descriptiva y se utilizó la prueba de Chi cuadrado para analizar la asociación entre variables. Los resultados revelaron una tendencia creciente en el número de ingresantes, con predominancia del género femenino (74%), una mayoría nacida en Lima (92.4%) y de nacionalidad peruana (98.2%). Además, el 30.3% residía en Lima Centro, la edad promedio de ingreso fue de 18.12 años, el 86.6% provenía de colegios particulares y el 56.3% ingresó por la modalidad de factor excelencia, mientras que el 29.2% no contaba con ningún tipo de beca. Estos hallazgos constituyen una base útil para la caracterización del perfil de los futuros postulantes a dicha carrera.

2.2. MARCO TEÓRICO

2.2.1. Minería de datos

La definición de minería de datos puede variar según el enfoque de cada investigador, ya sea desde la estadística, el análisis de datos u otras disciplinas afines. Existen múltiples interpretaciones que reflejan la diversidad de perspectivas sobre este campo, todas coincidiendo en que se trata de un proceso orientado a descubrir patrones útiles, relaciones significativas y conocimientos ocultos dentro de grandes volúmenes de datos. Algunas definiciones:

“La minería de datos es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma autorizada

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" (Piatetsky-Shapiro y Frawley, 1991, como se citó en Barrera Rodríguez, 2002).

"La minería de datos combina técnicas de la estadística, inteligencia artificial, bases de datos, visualización y otras áreas, para descubrir, de forma automática o semiautomática, modelos de series de datos" (Siebes, 2000 como se citó en Barrera Rodríguez, 2002).

"La minería de datos utiliza técnicas de reconocimiento de patrones para identificar y extraer patrones y tendencias significativas de los datos. Las técnicas más comunes empleadas en este proceso son las descriptivas y las predictivas" (Barrera Rodríguez, 2002).

2.2.2. Inteligencia Artificial (IA)

La inteligencia artificial (IA) es una disciplina científica enfocada en el desarrollo de sistemas computacionales capaces de razonar, aprender y ejecutar tareas que normalmente requerirían inteligencia humana, o que implican el procesamiento de volúmenes de datos que exceden la capacidad de análisis manual.

Este campo abarca múltiples áreas del conocimiento, como la informática, la estadística, la ingeniería de software y hardware, la lingüística, la neurociencia, así como disciplinas humanísticas como la filosofía y la psicología.

Desde una perspectiva operativa en el ámbito empresarial, la IA se compone de un conjunto de tecnologías, principalmente el aprendizaje automático y el aprendizaje profundo, que permiten realizar análisis de datos, generar

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

predicciones, clasificar información, procesar lenguaje natural, ofrecer recomendaciones y recuperar datos de forma inteligente, entre otras aplicaciones. (Google Cloud, s.f.)

2.2.2.1. Modelos de entrenamiento de IA

Cuando las empresas mencionan la IA, a menudo se refieren a los "datos de entrenamiento". ¿Pero qué implica esto? La inteligencia artificial con memoria limitada es una IA que se mejora con el tiempo al ser entrenada con nuevos datos. El aprendizaje automático, una rama de la inteligencia artificial, utiliza algoritmos para entrenar datos y producir resultados.

Suelen usarse tres tipos de modelos de aprendizaje en el aprendizaje automático:

El aprendizaje supervisado

Un modelo de aprendizaje automático supervisado se entrena utilizando datos estructurados que ya han sido etiquetados, lo que le permite asociar entradas específicas con resultados determinados. Por ejemplo, para que un algoritmo aprenda a identificar imágenes de gatos, se le proporciona un conjunto de imágenes previamente clasificadas como tales, lo que le permite reconocer patrones y realizar predicciones precisas en nuevos datos.

El aprendizaje no supervisado

El aprendizaje automático no supervisado se basa en el análisis de datos no etiquetados o no estructurados, lo que significa que el modelo no

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

cuenta con resultados predefinidos durante el proceso de entrenamiento. A diferencia del aprendizaje supervisado, este enfoque permite que el algoritmo identifique patrones y relaciones dentro de los datos por sí mismo, agrupándolos según similitudes en sus atributos. Este tipo de aprendizaje es especialmente útil para tareas de modelado descriptivo y para descubrir estructuras ocultas en grandes conjuntos de datos.

Aprendizaje semisupervisado

Además de los enfoques supervisado y no supervisado, existe una modalidad intermedia conocida como aprendizaje semisupervisado. En este método, solo una parte de los datos cuenta con etiquetas, mientras que el resto permanece sin clasificar. Aunque se dispone de un resultado final conocido, el algoritmo debe aprender a estructurar y organizar los datos de forma autónoma para alcanzar dicho resultado. Este enfoque combina las ventajas de ambos métodos y es especialmente útil cuando el etiquetado completo de los datos resulta costoso o impráctico.

El aprendizaje por refuerzo

El aprendizaje por refuerzo es un enfoque dentro del aprendizaje automático que puede describirse, en términos generales, como un proceso de “aprender mediante la experiencia”. En este modelo, un agente interactúa con su entorno a través de un ciclo de prueba y error, ajustando su comportamiento en función de los resultados obtenidos. El agente recibe recompensas positivas cuando ejecuta correctamente una tarea y penalizaciones cuando su desempeño es deficiente. Un ejemplo típico de este tipo de aprendizaje es entrenar a una mano robótica para que recoja

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

una pelota, donde el agente mejora progresivamente su rendimiento a medida que aprende de sus acciones. (Google Cloud, s.f.)

Figura 1

Tipos de aprendizaje automático

Nota: aprendizajes automáticos. Tomado de Aprender Libre (2024)

2.2.3. Análisis predictivo

El análisis predictivo forma parte de la minería de datos y se enfoca en obtener conocimiento a partir de la información existente para prever tendencias y patrones de conducta. Su aplicación abarca cualquier suceso desconocido, ya sea pasado, presente o futuro. Se fundamenta en reconocer vínculos entre variables en eventos previos, con el propósito de aprovechar estas relaciones para estimar resultados en escenarios futuros. Sin embargo, es importante considerar que la precisión de los resultados depende en gran medida de cómo se ha llevado a cabo el análisis de los datos y de la calidad de las suposiciones realizadas. (IBM, 2024)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

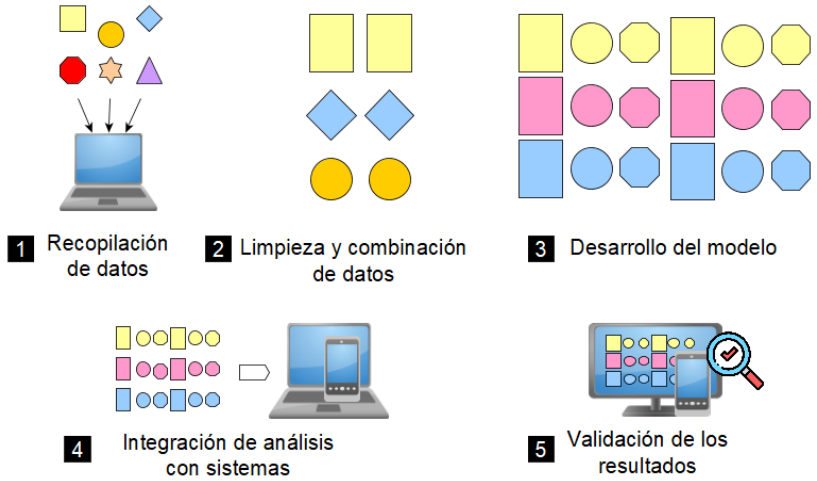
UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.4. Funcionamiento del análisis predictivo

El análisis predictivo se basa en gran medida en técnicas avanzadas de machine learning. Los científicos de datos emplean el aprendizaje profundo y complejos algoritmos para examinar diversas variables y crear modelos que puedan prever comportamientos futuros a partir de grandes volúmenes de datos. Similar a muchas aplicaciones de Machine Learning, el análisis predictivo es una actividad dinámica que actualiza constantemente sus predicciones con nuevos datos. Esto implica usar herramientas clásicas de Machine Learning para limpiar datos, entrenar modelos, desplegarlos, revisarlos, reentrenarlos y volver a desplegarlos, además de la capacidad de procesar datos casi en tiempo real.

Figura 2

Proceso de análisis de datos



Nota: El gráfico representa los 5 pasos para realizar un análisis de datos. Tomado de SafetyCulture (2024)

Algunas de estas técnicas incluyen análisis de regresión, árboles de decisión, análisis de series temporales y redes neuronales profundas. Los modelos predictivos, que influyen en las decisiones empresariales futuras, tienden a ser complejos y consideran múltiples factores. Su desarrollo y validación suelen ser

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

procesos largos, y requieren revisiones constantes para adaptarse a cambios en el entorno económico y empresarial. (AWS, s.f.)

2.2.4.1. Recopilación de datos

Importar datos de fuentes relevantes, como hojas de cálculo y bases de datos, y asegurarse de que estos conjuntos de datos se adaptan a los requisitos del objetivo.

2.2.4.2. Limpieza y combinación de datos

Es fundamental depurar los datos eliminando cualquier elemento irrelevante, verificar la integridad de la información y detectar posibles inconsistencias que puedan comprometer su precisión. Asimismo, es necesario identificar y combinar las fuentes de datos pertinentes, seleccionándolas en función del objetivo que se desea alcanzar con el análisis.

2.2.4.3. Desarrollar el modelo

Una vez formulada la hipótesis, se procede a construir el modelo utilizando los datos recopilados, aplicando técnicas estadísticas. Es fundamental considerar todas las variables y factores relevantes durante este proceso. Posteriormente, el modelo se valida mediante datos históricos, con el fin de evaluar su precisión y verificar si los resultados obtenidos respaldan la hipótesis planteada.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.4.4. Integración de la analítica con sistemas

Implica incorporar el modelo desarrollado dentro del entorno de producción y los procesos operativos de la organización. Esto se logra mediante su implementación en el software o dispositivos utilizados, permitiendo que el modelo forme parte activa del flujo de trabajo y contribuya a la toma de decisiones en tiempo real.

2.2.4.5. Validación de los resultados

Es fundamental monitorear el desempeño del modelo para verificar si continúa generando los resultados esperados. Dado que diversos factores pueden influir en su eficacia con el paso del tiempo, resulta necesario evaluar periódicamente su rendimiento y realizar ajustes o actualizaciones cuando sea pertinente. SafetyCulture (2024)

2.2.5. Tipos de modelos predictivos

Los tipos de modelo predictivos que existen son fundamentalmente dos: el de clasificación y el de regresión.

2.2.5.1. Modelos de clasificación

Los modelos de clasificación son una categoría de modelos predictivos en el campo del aprendizaje automático y la estadística que se utilizan para asignar datos a una o más categorías o clases discretas. El objetivo de un modelo de clasificación es predecir la etiqueta de clase de nuevas observaciones basándose en datos de entrenamiento previamente etiquetados (Universitat, 2021). A continuación, se detallan los conceptos clave y algoritmos de modelos de clasificación:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Conceptos clave:

Variable dependiente (o de respuesta):

- La variable que se quiere predecir, que es categórica.
- Ejemplos: tipo de enfermedad (cáncer/no cáncer), aprobación de préstamo (aprobado/rechazado).

Variables Independientes (o predictoras):

- Las variables de entrada utilizadas para predecir la clase de la variable dependiente.
- Ejemplos: edad, ingresos, resultados de pruebas médicas.

Etiquetas de clase:

- Las categorías discretas que la variable dependiente puede tomar.
- Ejemplos: spam/no spam, clase de riesgo (alto/medio/bajo).

Regresión logística

Un modelo de regresión con variable dependiente binaria, conocido como modelo logístico o de regresión logística, es un enfoque que permite analizar si dicha variable discreta está condicionada por una o varias variables. Si una variable binaria con parámetro p es independiente de otra variable X, se cumple que (p|X=x) =p, para cualquier valor x de la variable X. (Pérez López, 2004)

Por lo tanto, un modelo de regresión con variable dependiente binaria y una única variable independiente X se expresa mediante una

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

función donde **p** varía en función de **X** ciertos coeficientes cuyo análisis permite abordar la relación de dependencia. En el caso de una única variable independiente **X**, el modelo de regresión logística adopta la siguiente forma:

$\mathbf{ln}\left(\frac{p}{q}\middle|\ \boldsymbol{X}=\boldsymbol{x}\right)=\boldsymbol{\alpha_0}+\boldsymbol{\alpha_1 X}$, o de forma simplificada: $\mathbf{ln}\ (\boldsymbol{p/q})=\boldsymbol{\alpha_0}+\boldsymbol{\alpha_1 X}$

$$\mathbf{ln}\left(\frac{\boldsymbol{p}}{\boldsymbol{p-1}}\right)=\boldsymbol{\alpha_0}+\boldsymbol{\alpha_1 X}\leftrightarrow\frac{\boldsymbol{p}}{\boldsymbol{p-1}}=\boldsymbol{e^{\alpha_0+\alpha_1 X}}$$

$$\boldsymbol{p}=\frac{\boldsymbol{e^{\alpha_0+\alpha_1 X}}}{\mathbf{1}+\boldsymbol{e^{\alpha_0+\alpha_1 X}}}$$

$$\boldsymbol{p}=\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e^{-(\alpha_0+\alpha_1 X)}}}$$

Si se conocen los coeficientes α_0 y α_1, estas dos últimas expresiones permiten directamente la probabilidad del proceso binomial para los diferentes valores de la variable X.

A la función: $f_{(z)}=\frac{1}{1+e^{-z}}$ se le denomina función logística. En un modelo de regresión logística, la probabilidad de un proceso binomial se representa mediante la función logística aplicada a una combinación lineal de la variable dependiente.

Este modelo puede extenderse de manera sencilla para considerar k variables independientes, dando lugar al modelo logístico múltiple, cuya formulación se expresa de la siguiente manera:

$$\ln\left(\frac{p}{q}\right)=\alpha_0+\alpha_1X_1+\alpha_2X_2+\alpha_3X_3+\cdots+\alpha_kX_k$$

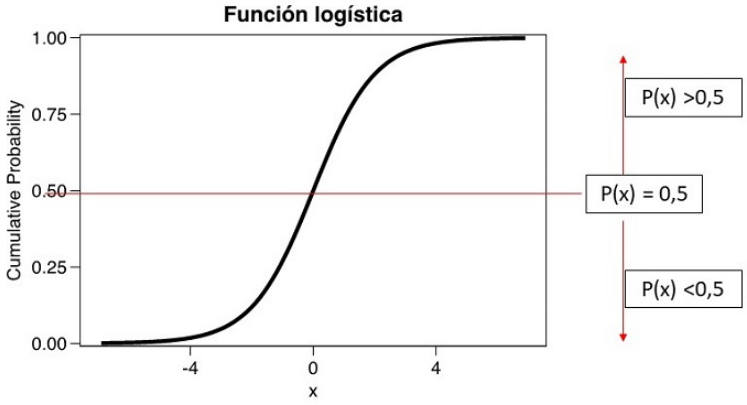
repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

$$p = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k)}}$$

Figura 3

Representación de la función logística



Nota: valores de probabilidad para la función logística. Tomado de Ortega Páez et al. (2022)

Árboles de decisión

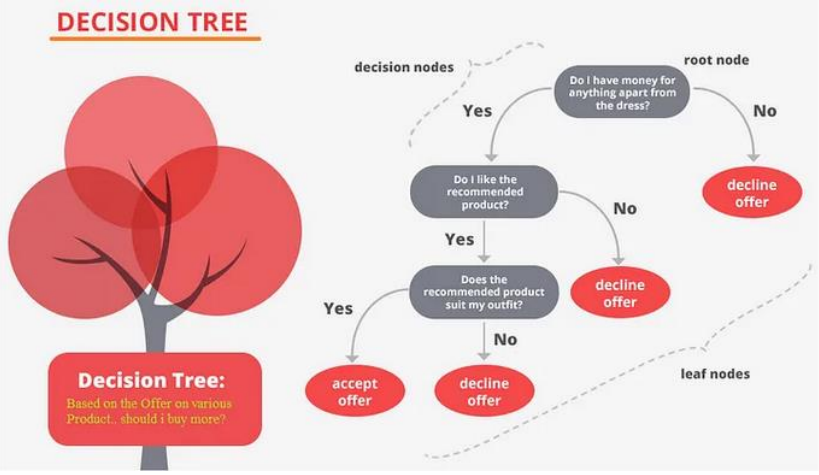
En el ámbito del aprendizaje automático, un árbol de decisión es una organización jerárquica comparable a un diagrama de flujo. En esta estructura, los nodos internos simbolizan atributos o características, las ramas representan criterios de decisión, y las hojas muestran los resultados finales. El nodo inicial se conoce como nodo raíz. Este tipo de modelo de clasificación organiza los datos en función de reglas asociadas a etiquetas de clase específicas. El árbol aprende a dividir los datos según los valores de los atributos, realizando particiones de forma recursiva. Este proceso se guía por criterios como la entropía, con el objetivo de minimizar la impureza en cada división, utilizando un conjunto amplio de datos de entrenamiento. (Oday y Suguiura, 2022)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 4

Representación del árbol de decisión



Nota: imagen tomada de Bootcamp AI (2020)

Bosques Aleatorios (Random Forest)

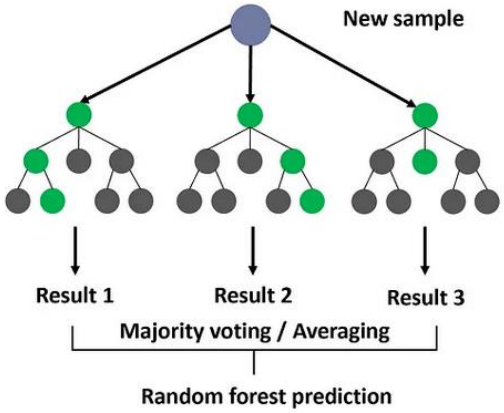
El método Random Forest es una técnica de aprendizaje automático que integra múltiples árboles de decisión para fortalecer la precisión y estabilidad de las predicciones. Fue presentado por Leo Breiman en 2001. En un Random Forest, cada árbol de decisión es entrenado en un subconjunto aleatorio de los datos de entrenamiento y de las características disponibles. Las predicciones de todos los árboles son combinadas (generalmente mediante votación para clasificación o promedio para regresión) para producir una predicción final más precisa y menos susceptible a sobreajuste en comparación con un único árbol de decisión. (Breiman, 2001)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 5

Predicción de Random Forest



Nota: la figura muestra una representación del modelo de Random Forest. Tomado de Yehoshua (2023)

Máquinas de Soporte Vectorial (SVM)

Una Máquina de Soporte Vectorial (SVM) es un algoritmo supervisado de machine learning utilizado principalmente para problemas de clasificación y regresión. Su objetivo es encontrar el hiperplano que mejor separa los datos en diferentes clases. Algunos conceptos que se deben tener en cuenta son:

- **Hiperplano:** en un espacio de características de ‘n’ dimensiones, un hiperplano es una subvariedad de ‘n−1’ dimensiones que divide el espacio en dos partes. Por ejemplo, en un espacio bidimensional, un hiperplano es una línea; en un espacio tridimensional, es un plano.
- **Margen:** es la distancia entre el hiperplano y los puntos más cercanos de cada clase. SVM busca maximizar este margen, lo que ayuda a mejorar la generalización del modelo.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- **Vectores de soporte:** son los puntos de datos que están más cerca del hiperplano y que determinan su posición. Estos puntos son cruciales para la construcción del hiperplano óptimo. (Cobos et al., 2020)

Funcionamiento

Selección del Hiperplano Óptimo

SVM busca el hiperplano que separa las clases de manera que el margen entre las clases sea lo más grande posible. Este proceso se llama maximización del margen.

Kernel Trick

Para problemas no lineales, donde los datos no pueden ser separados por un hiperplano lineal, SVM utiliza funciones de kernel. Estas funciones transforman los datos en un espacio de mayor dimensión donde es posible encontrar un hiperplano lineal que separe las clases. Los kernels comunes incluyen:

- Lineal, adecuado para datos que son linealmente separables.
- Polinómico, adecuado para datos con relaciones polinómicas.
- Radial Basis Function (RBF) o Gaussiano, adecuado para datos no linealmente separables.

Optimización

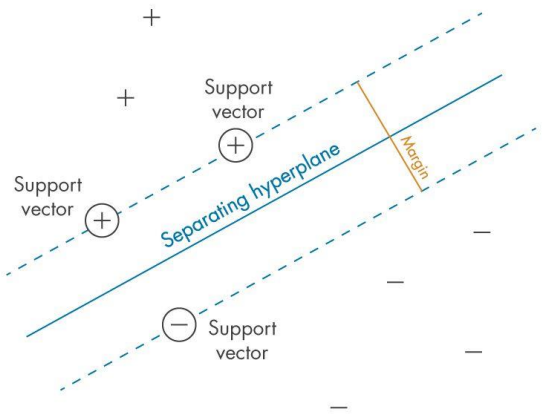
SVM utiliza métodos de optimización convexa, como el método de los multiplicadores de Lagrange, para encontrar el hiperplano óptimo.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 6

Definición del margen entre clases - SVM



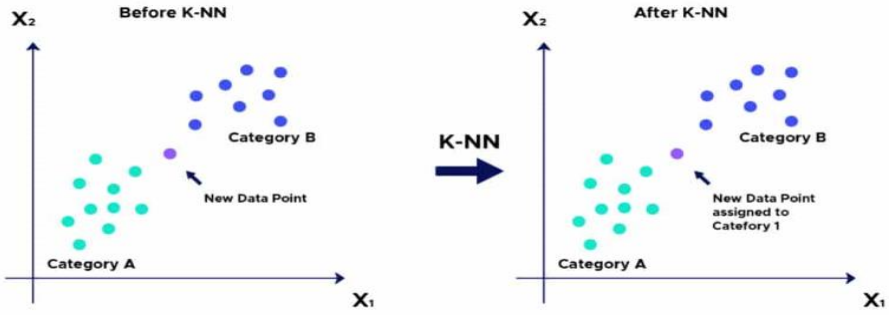
Nota: vista del hiperplano y margen por clases. Tomado de MathWorks en Support Vector Machine - MATLAB y Simulink, (s.f.)

K-Vecinos Más Cercanos (K-Nearest Neighbors, KNN)

Es un algoritmo de Machine Learning que clasifica un dato o realiza una predicción basándose en las K muestras de entrenamiento más cercanas en el espacio de características. Para determinar la cercanía entre los datos, comúnmente se utiliza la distancia euclidiana, aunque también pueden emplearse otras métricas de distancia como la Manhattan o la distancia de Minkowski. (DataScientest, 2021)

Figura 7

Representación del algoritmo KNN



Nota: la figura muestra la asignación de un dato a una determinada categoría. Tomado de DataScientest (2021)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.5.2. Modelos de regresión

Los modelos de regresión son un conjunto de técnicas estadísticas y de aprendizaje automático utilizadas para predecir el valor de una variable dependiente (o de respuesta) contínua basada en el valor de una o más variables independientes (o predictoras). A continuación, se describen los conceptos clave, tipos y aplicaciones de los modelos de regresión:

Conceptos clave

Variable Dependiente (o de Respuesta):

- La variable que se quiere predecir.
- Ejemplos: precio de una casa, ventas mensuales, temperatura.

Variables Independientes (o Predictoras):

- Las variables de entrada utilizadas para predecir la variable dependiente.
- Ejemplos: tamaño de la casa, gasto en marketing, hora del día.

Relación Funcional:

- Los modelos de regresión tratan de establecer una relación matemática entre las variables independientes y la variable dependiente.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.6. Métricas en Machine Learning

En Machine Learning, las métricas de evaluación son esenciales para medir el rendimiento y la efectividad de los modelos predictivos. Estas métricas permiten comparar diferentes modelos y seleccionar el más adecuado para una tarea específica. Las métricas varían según el tipo de problema (clasificación, regresión, clustering, etc.) y las características del conjunto de datos.

En problemas de clasificación las métricas más comunes son:

2.2.6.1. Exactitud (Accuracy)

Representa la proporción de predicciones correctas sobre el total de predicciones. La fórmula para obtener el resultado es:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

TP : es True Positive (verdadero positivo)

TN : es True Negative (verdadero negativo)

FP : es False Positive (falso positivo)

FN : es False Negative (falso negativo)

2.2.6.2. Precisión (precision)

Representa la proporción de predicciones positivas correctas sobre el total de predicciones positivas.

$$Precision = \frac{TP}{TP + FP}$$

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.6.3. Sensibilidad (Recall o tasa de verdaderos positivos)

Representa la proporción de verdaderos positivos (TP) sobre el total de casos reales positivos.

$$Recall = \frac{TP}{TP + FN}$$

2.2.6.4. F1 Score

Representa la media armónica de precisión y recall. Es útil cuando hay un desequilibrio en las clases.

$$F1 = 2.\frac{Precision\,.Recall}{Precision + Recall}$$

2.2.6.5. Especificidad (Specificity o Tasa de Verdaderos Negativos)

Representa la proporción de negativos reales que son correctamente identificados como negativos por el clasificador.

$$Especificidad = \frac{TN}{TN + FP}$$

2.2.6.6. Matriz de confusión

Es una tabla de doble entrada que muestra las predicciones correctas e incorrectas desglosadas por cada clase. Ayuda a identificar dónde el modelo está cometiendo errores.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 2

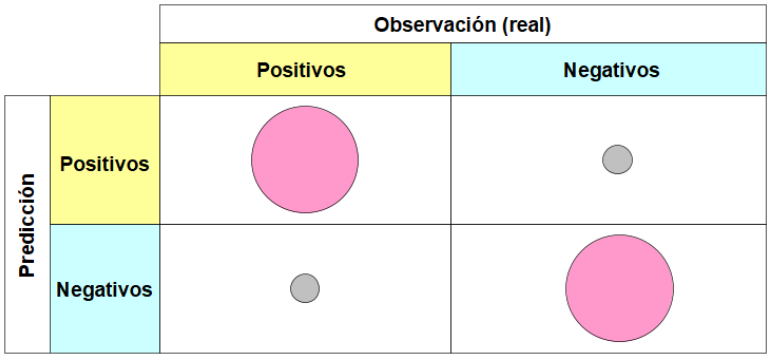
Matriz de confusión

		Observación (real)	
		Positivos	**Negativos**
Predicción	**Positivos**	**Verdaderos Positivos (VP)** Cantidad de positivos que fueron clasificados correctamente como positivos	**Falsos Positivos (FP)** Cantidad de negativos que fueron clasificados incorrectamente como positivos
	Negativos	**Falsos Negativos (FN)** Cantidad de positivos que fueron clasificados incorrectamente como negativos	**Verdaderos Negativos (VN)** Cantidad de negativos que fueron clasificados correctamente como negativos

Nota: la tabla muestra la parte teórica de las observaciones reales con las observaciones predichas.

Tabla 3

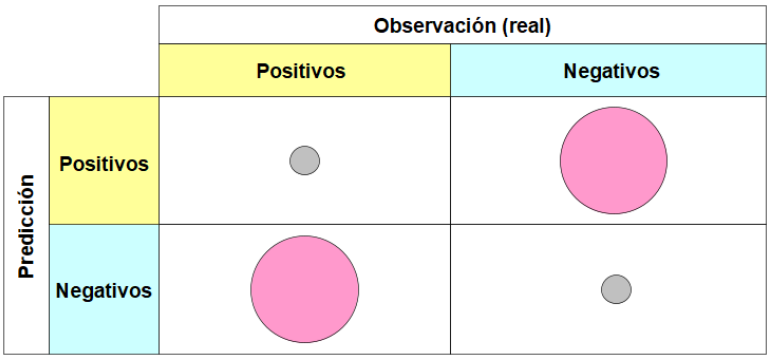
Representación ideal para la matriz de confusión



Nota: los círculos representan valores numéricos

Tabla 4

Representación errónea para la matriz de confusión



Nota: los círculos representan valores numéricos

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.6.7. La curva ROC

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada para evaluar el rendimiento de un clasificador binario a través de diferentes umbrales de decisión. La curva ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR), proporcionando una visión completa del comportamiento del modelo a medida que varía el umbral de clasificación.

2.2.6.8. Área bajo la curva ROC - AUC

La AUC-ROC mide la capacidad del modelo, es decir, que tan bueno es el modelo para distinguir entre la clase positiva y la clase negativa. (Inteligencia Artificial, 2023)

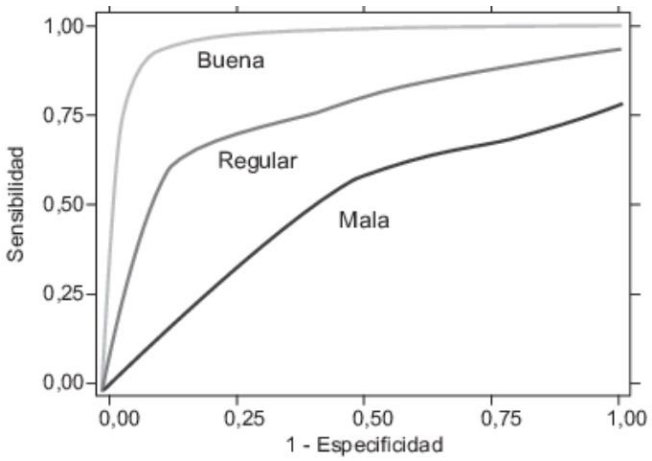
Una AUC igual a 1 indica que el modelo tiene un rendimiento perfecto, es capaz de clasificar correctamente todas las instancias positivas y negativas. Una AUC > 0.5 indica que el modelo tiene un rendimiento mejor que el azar, es capaz de clasificar correctamente las instancias en un porcentaje mayor que al azar. Una AUC igual a 0.5 indica que el modelo tiene un rendimiento similar al azar, no es capaz de distinguir entre las clases y sus predicciones son equivalentes a lanzar una moneda al aire. Una AUC menor que 0.5 indica que el modelo tiene un rendimiento peor que el azar, clasifica incorrectamente las instancias en un porcentaje mayor que al azar.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 8

La curva ROC y su efectividad



Nota: figura tomada del artículo de Valero Galán et al. (2021).

2.2.7. Ingeniería de características

La ingeniería de características, en la ciencia de datos, se refiere a la manipulación (adición, eliminación, combinación, mutación) de su conjunto de datos para mejorar la formación del modelo de aprendizaje automático, lo que lleva a un mejor rendimiento y una mayor precisión. Una ingeniería de características eficaz se basa en un buen conocimiento del problema empresarial y de las fuentes de datos disponibles.

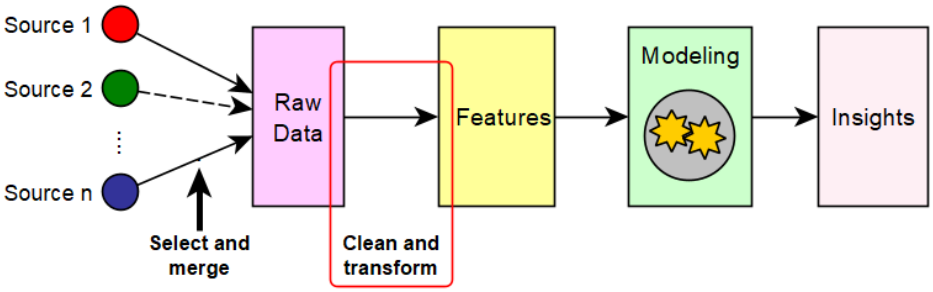
La creación de nuevas características permite comprender mejor los datos y obtener información más valiosa. Cuando se realiza correctamente, la ingeniería de características es una de las técnicas más valiosas de la ciencia de datos, pero también una de las más difíciles. Un ejemplo común de ingeniería de características es cuando su médico utiliza su índice de masa corporal (IMC). El IMC se calcula a partir del peso corporal y la altura, y sirve como sustituto de una característica que es muy difícil de medir con precisión: la proporción de masa corporal magra. (Domino, 2024)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 9

El lugar de la ingeniería de características



Nota: Flujo de trabajo de Machine Learning, tomado del libro de Zheng y Casari (2018).

2.2.7.1. Procesos de la ingeniería de características

La ingeniería de características consta de varios procesos:

Creación de características

La creación de características implica la creación de nuevas variables que serán más útiles para nuestro modelo. Esto puede ser añadiendo o eliminando algunas características. Como se ve en las siguientes tablas, la columna de costo por pie cuadrado es una creación de características.

Tabla 5

Precio de propiedades por ciudad

Área	Precio
2400	9 millones
3200	15 millones
2500	10 millones
2100	1.5 millones
2500	8.9 millones

Nota: Se muestra el área en pies cuadrados de la casa con su precio total

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 6

Agregando una columna costo por pie cuadrado

Área	Precio	Costo por pie cuadrado
2400	9 millones	4150
3200	15 millones	4944
2500	10 millones	3950
2100	1.5 millones	510
2500	8.9 millones	3600

Nota: Se muestra el área de la casa, su precio total y una nueva característica

Transformaciones

La transformación de características es simplemente una función que transforma características de una representación a otra. El objetivo es representar y visualizar los datos. Si algo no cuadra con las nuevas características, podemos reducir el número de características utilizadas, acelerar el entrenamiento o aumentar la precisión de un determinado modelo.

Tabla 7

Transformación dato continuo a categórico

Ingreso anual	Categoría ingreso
15000	Bajo
30000	Bajo
50000	Medio
70000	Medio
120000	Alto
25000	Bajo
85000	Alto

Nota: Esta transformación se da bajo ciertos intervalos

Extracción de características

La extracción de características es el proceso de extraer características de un conjunto de datos para identificar información útil. Sin distorsionar las relaciones originales ni la información significativa,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

esto comprime la cantidad de datos en cantidades manejables para que los procesen los algoritmos.

Tabla 8

Se extrae una nueva característica basada en otras

Producto	Precio	Cantidad	Descuento	Ingresos
A	10	100	0.10	900.0
B	20	50	0.15	850.0
C	15	80	0.00	1200.0
A	12	200	0.05	2280.0
B	18	40	0.20	576.0
C	16	60	0.00	960.0
A	14	150	0.10	1890.0

Nota: A partir de un conjunto de características se genera una nueva característica sustituta (ingresos)

Análisis exploratorio de datos

El análisis exploratorio de datos (AED) es una herramienta potente y sencilla que puede utilizarse para mejorar la comprensión de los datos, explorando sus propiedades. La técnica suele aplicarse cuando el objetivo es crear nuevas hipótesis o encontrar patrones en los datos. Suele utilizarse con grandes cantidades de datos cualitativos o cuantitativos que no se han analizado antes.

Benchmark

Un modelo de referencia es el modelo más fácil de usar, fiable, transparente e interpretable con el que puede medir el suyo. Es una buena idea ejecutar conjuntos de datos de prueba para ver si su nuevo modelo de aprendizaje automático supera a un punto de referencia reconocido. Estos puntos de referencia suelen utilizarse para comparar el rendimiento de distintos modelos de aprendizaje automático, como redes neuronales y

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

máquinas de vectores de soporte, clasificadores lineales y no lineales o distintos enfoques, como bagging y boosting. (Harshil, 2024)

2.2.7.2. Técnicas de ingeniería de características

Algunas de las técnicas enumeradas pueden funcionar mejor con determinados algoritmos o conjuntos de datos, mientras que otras pueden ser útiles en todas las situaciones.

Imputación

Los valores perdidos son uno de los problemas más comunes a la hora de preparar los datos para el aprendizaje automático. Los errores humanos, las interrupciones en el flujo de datos, los problemas de privacidad y otros factores pueden contribuir a que falten valores. Los valores perdidos afectan al rendimiento de los modelos de aprendizaje automático (Amarilla, 2023). El objetivo principal de la imputación es tratar estos valores perdidos. Existen dos tipos de imputación:

Imputación numérica

```
#Llenando todos los valores faltantes con 0
data = data.fillna(0)
```

Imputación categórica

```
#Rellena valores faltantes con la de mayor frecuencia (moda)
data['column'].fillna(data['column'].value_counts().idxmax(), inplace=True)
```

Tratamiento de valores atípicos

El tratamiento de valores atípicos es una técnica para eliminar los valores atípicos de un conjunto de datos. Este método puede utilizarse en

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

diversas escalas para obtener una representación más precisa de los datos. Esto repercute en el rendimiento del modelo. Dependiendo del modelo, el efecto puede ser grande o mínimo. Por ejemplo, la regresión lineal es especialmente susceptible a los valores atípicos. Este procedimiento debe completarse antes del entrenamiento del modelo. Los distintos métodos para tratar los valores atípicos incluyen:

Eliminación

Las entradas que contienen valores atípicos se eliminan de la distribución. Sin embargo, si hay valores atípicos en numerosas variables, esta estrategia puede hacer que se pierda una gran parte del conjunto de datos.

Sustitución de valores

Como alternativa, los valores atípicos podrían tratarse como valores perdidos y sustituirse con una imputación adecuada.

Capping

Utilización de un valor arbitrario o de un valor de una distribución variable para sustituir los valores máximo y mínimo.

Discretización

La discretización es el proceso de convertir variables, modelos y características continuas en discretas. Para ello, se construye una serie de intervalos continuos (o intervalos) que abarcan el rango de la variable/modelo/característica deseada.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Transformación logarítmica

La transformación logarítmica es la técnica más utilizada entre los científicos de datos. Se utiliza sobre todo para convertir una distribución sesgada en una distribución normal o menos sesgada. Se toma el logaritmo de los valores de una columna y se usa esos valores como columna en esta transformación. Se utiliza para manejar datos confusos, y los datos se vuelven más aproximados a las aplicaciones normales. (Amarilla, 2023)

One -Hot Encoding

Una codificación de un solo bit es un tipo de codificación en el que un elemento de un conjunto finito se representa por el índice en ese conjunto, donde sólo un elemento tiene su índice fijado en “1” y a todos los demás elementos se les asignan índices dentro del intervalo [0, n-1]. A diferencia de los esquemas de codificación binaria, en los que cada bit puede representar dos valores (es decir, 0 y 1), este esquema asigna un único valor para cada caso posible.

Label Encoding

El Label Encoding o Codificación de Etiquetas es una técnica utilizada en el preprocesamiento de datos para convertir variables categóricas ordinales en valores numéricos. Esta técnica es especialmente útil cuando las categorías de una variable tienen un orden inherente o una jerarquía, y es importante que el modelo de aprendizaje automático capture esta relación ordinal. En el Label Encoding, a cada categoría única se le asigna un valor entero único. La asignación se realiza de manera secuencial, generalmente comenzando desde 0 o 1, y siguiendo el orden

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

de las categorías. De esta forma, se preserva la información ordinal de las variables categóricas. (Dahouda & Joe, 2021)

Scaling

El escalado de características es uno de los problemas más omnipresentes y difíciles en el aprendizaje automático, y sin embargo es una de las cosas más importantes que hay que hacer bien. Para entrenar un modelo predictivo, necesitamos datos con un conjunto conocido de características que deben ampliarse o reducirse según convenga. Tras una operación de escalado, las características continuas se vuelven similares en términos de rango. Aunque este paso no es necesario para muchos algoritmos, sigue siendo una buena idea hacerlo. Por otra parte, los algoritmos basados en la distancia, como K-Nearest Neighbor y K-Means, requieren características continuas escaladas como entrada del modelo (DATOS, 2024). Existen dos formas habituales de escalado:

Normalización

Todos los valores se escalan en un intervalo especificado entre 0 y 1 mediante normalización (o normalización mín-máx). Esta modificación no influye en la distribución de la característica, pero agrava los efectos de los valores atípicos debido a las desviaciones estándar más bajas. Por ello, se recomienda tratar los valores atípicos antes de la normalización.

Estandarización

La estandarización, o normalización de la puntuación z, es el proceso de escalar valores teniendo en cuenta la desviación estándar. Si la

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

desviación estándar de las características es diferente, el rango de esas características también lo será. En consecuencia, se reduce el efecto de los valores atípicos en las características. Para obtener una distribución con media 0 y varianza 1, a todos los puntos de datos se les resta su media y el resultado se divide por la varianza de la distribución.

2.2.8. Norma técnica de ruralidad

La norma técnica denominada "Norma que regula la metodología para determinar el ámbito urbano o rural de los servicios educativos, y su grado de ruralidad" fue aprobada mediante Resolución Ministerial Nº 564-2023 MINEDU y establecen los criterios de clasificación de mediante la gradiente de ruralidad de los centros poblados, identificando tres grupos al interior de la categoría rural, de mayor a menor ruralidad. (Norma Técnica, 2023)

Rural 1: "Todos los centros o núcleos poblados rurales que cuentan con un número menor o igual a quinientos (500) habitantes, y que puedan acceder a la capital provincial urbana más cercana en un tiempo mayor a ciento veinte (120) minutos".

Rural 2: "Todos los centros o núcleos poblados rurales que cuentan con un número mayor a quinientos (500) habitantes y que puedan acceder a la capital provincial urbana más cercana en un tiempo mayor a ciento veinte (120) minutos. Asimismo, todos los centros o núcleos poblados rurales que cuentan con un número de habitantes menor o igual a quinientos (500) habitantes, y que puedan acceder a la capital provincial urbana más cercana en un tiempo mayor a treinta (30) minutos y menor o igual a ciento veinte (120) minutos".

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Rural 3: "Todos los centros o núcleos poblados rurales que cuentan con un número de habitantes mayor a quinientos (500) habitantes y que puedan acceder a la capital provincial urbana más cercana en un tiempo menor o igual a ciento veinte (120) minutos. Asimismo, todos los centros o núcleos poblados rurales que cuentan con un número de habitantes menor o igual a quinientos (500) habitantes y que puedan acceder a la capital provincial urbana más cercana en un tiempo menor o igual a treinta (30) minutos".

2.2.9. Factor de Inflación de la Varianza

El Factor de Inflación de la Varianza (VIF) es una medida estadística que cuantifica el grado de multicolinealidad presente en un modelo de regresión múltiple. La multicolinealidad se refiere a la existencia de correlaciones elevadas entre dos o más variables independientes (predictoras) en el modelo.

El VIF mide cuánto se incrementa la varianza de los coeficientes de regresión estimados debido a la multicolinealidad. Un VIF alto indica que la varianza de un coeficiente se ha inflado, lo que sugiere que la estimación de dicho coeficiente es menos precisa y confiable. (Hair et al., 2019)

Cálculo del VIF

El VIF para la variable predictora *i* se calcula de la siguiente manera:

$$VIF = \frac{1}{R_i^2}$$

Donde R_i^2 es el coeficiente de determinación de la regresión de la variable *i* contra todas las demás variables predictoras en el modelo.

Interpretación del VIF

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

En general, se considera que un VIF superior a 5 o 10 indica un problema de multicolinealidad significativo. Sin embargo, no existe un valor de corte universalmente aceptado, y la interpretación del VIF puede depender del contexto específico de la investigación.

Implicaciones de la multicolinealidad

La multicolinealidad puede tener varias consecuencias negativas en el análisis de regresión, incluyendo:

Estimaciones inestables de los coeficientes: Los coeficientes de regresión pueden volverse muy sensibles a pequeños cambios en los datos, lo que dificulta la interpretación de sus efectos individuales.

Dificultad para identificar variables importantes: La multicolinealidad puede enmascarar la verdadera importancia de algunas variables predictoras, ya que sus efectos se "mezclan" con los de otras variables correlacionadas.

Reducción de la potencia estadística: La multicolinealidad puede dificultar la identificación de relaciones estadísticamente significativas entre las variables predictoras y la variable dependiente.

2.3. GLOSARIO DE TÉRMINOS BÁSICOS

2.3.1. Python

Python es un lenguaje de programación de alto nivel, interpretado y de propósito general que es conocido por su simplicidad y legibilidad. Fue creado por Guido van Rossum y su primera versión fue lanzada en 1991. Python ha ganado una popularidad significativa en los últimos años debido a su facilidad de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

uso, versatilidad y la amplia gama de bibliotecas y frameworks disponibles. A continuación, se presenta una descripción detallada de Python, incluyendo sus características, ventajas y aplicaciones. (AWS, s.f.)

2.3.2. Colaboratory

Es una herramienta gratuita proporcionada por Google que permite escribir y ejecutar código en Python directamente en el navegador, similar a Jupyter Notebooks. Google Colab es especialmente popular en la comunidad de ciencia de datos y aprendizaje automático debido a su facilidad de uso y acceso a recursos computacionales potentes. (Google Colab, s.f.)

2.3.3. librerias de python

Pandas

Es una biblioteca de Python especializada en la manipulación y el análisis de datos. Es una de las herramientas más importantes y ampliamente utilizadas en el campo de la ciencia de datos debido a su capacidad para manejar grandes volúmenes de datos y realizar operaciones complejas de manera eficiente y con una sintaxis intuitiva. Pandas proporciona estructuras de datos flexibles, principalmente los DataFrames, que permiten trabajar con datos tabulares de una manera similar a como se hace en hojas de cálculo o bases de datos. (DataScientes, 2022)

```
import pandas as pd
# Crear el dataframe a partir de un diccionario
data = {
‘Nombre’: [’Ana’, ‘Luis’, ‘Carla’, ‘Pedro’, ‘María’],
‘Edad’: [18, 19, 20, 18, 21],
‘Horas de estudio por semana’: [15, 10, 20, 8, 12],
‘Calificacion final’: [15, 12, 18, 8, 11]
}
df = pd.DataFrame(data)
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

```
#Mostrar el dataframe
print(df)
```

Numpy

Es una biblioteca fundamental para la computación científica en Python. Proporciona soporte para arrays y matrices multidimensionales, junto con una colección de funciones matemáticas de alto nivel para realizar operaciones con estos datos. NumPy es la base sobre la que se construyen muchas otras bibliotecas de Python para ciencia de datos, como Pandas, SciPy y Scikit-learn, debido a su eficiencia y capacidad para manejar grandes volúmenes de datos numéricos. (DataScientest, 2023)

```
import numpy as np
# Crear un array unidimensional
a = np.array([1, 2, 3, 4, 5])
# Crear un array bidimensional
b = np.array([1,2,3], [4,5,6])
# operaciones aritméticas
c = a +10          #sumar 10 a cada elemento del array
d = a * 2          #multiplicar por 2 a cada elemento del array
# Calcular la media y la desviación estándar
mean_a = np.mean(a)
std_a = np.std(a)
# Producto matricial
e = np.dot(b, b.T)
# Crear un array con valores aleatorios
random_array = np.random.rand(3, 3)
print("Array unidimensional a:", a)
print("Array bidimensional b:", b)
print("Array después de sumar 10:", c)
print("Array después de multiplicar por 2:", d)
print("Media de:", mean_a)
print("Desviación estándar de:", std_a)
print("Producto matricial de b y su transpuesta:", e)
print("Array aleatorio:", random_array)
```

2.3.4. Scikit-learn

Es una biblioteca de Python ampliamente utilizada para el aprendizaje automático y el análisis de datos. Proporciona una variedad de herramientas

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

eficientes y fáciles de usar para modelado predictivo, incluyendo algoritmos de clasificación, regresión, clustering, reducción de dimensionalidad, selección de modelos y preprocesamiento de datos. (Canle Fernandez, 2023)

A continuación, se presenta un ejemplo simple de cómo utilizar Scikit-Learn para entrenar y evaluar un modelo de clasificación usando el algoritmo de k-vecinos más cercanos (K-NN):

```
import numpy as np
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
# Cargar el dataset Iris
iris = load_iris()
X, y = iris.data, iris.target
# Dividir el dataset en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# Escalar las características
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
# Crear y entrenar el modelo k-NN
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)
# Hacer predicciones
y_pred = knn.predict(X_test)
# Evaluar el modelo
accuracy = accuracy_score(y_test, y_pred)
print("Precisión del modelo k-NN:", accuracy)
```

2.3.5. Mathplotlib

Matplotlib es una biblioteca de visualización de datos en Python ampliamente utilizada para crear gráficos 2D de alta calidad. Ofrece una interfaz sencilla (similar a MATLAB) y es compatible con numerosos tipos de gráficos como líneas, barras, histogramas, dispersión (scatter), diagramas de caja, entre otros. El submódulo más comúnmente usado es pyplot, que proporciona funciones para construir gráficos paso a paso.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

A continuación, se muestra un ejemplo:

```
import matplotlib.pyplot as plt
# Datos para el gráfico
x = [0, 1, 2, 3, 4]
y = [0, 1, 4, 9, 16]
# Crear el gráfico
plt.plot(x, y, label=”Cuadrados”, color=’blue’, marker=’o’)
# Agregar etiquetas y título
plt.xlabel(‘X’)
plt.ylabel(‘Y’)
plt.title(‘Ejemplo de gráfico de línea’)
plt.legend()  # Mostrar leyenda
# Mostrar el gráfico
plt.grid()  # Agrega una cuadrícula para facilitar la lectura
plt.show()
```

2.3.6. Dataset

El conjunto de datos (dataset) es una colección organizada de datos que se utiliza para análisis, entrenamiento de modelos de aprendizaje automático, investigación y toma de decisiones. Los datasets pueden estar compuestos por diversos tipos de datos, incluyendo números, textos, imágenes, audios, videos y más. Estos datos se estructuran generalmente en forma de tablas, donde las filas representan observaciones individuales y las columnas representan características o variables específicas de esas observaciones. (Information Lab, 2023)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. ÁREA DE ESTUDIO

En la presente investigación, no se tuvo ningún tipo de comunicación con personas, ya que el objetivo principal es realizar un modelo predictivo basado en datos demográficos de los procesos de admisión, por lo que, los datos usados son información demográfica obtenido de los procesos de admisión correspondiente a los años 2018 y 2022 de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano del departamento de Puno.

3.2. POBLACIÓN

La población de este estudio está conformada por todos los postulantes que participaron en los exámenes de admisión ordinario (General y CEPREUNA) a la Escuela Profesional de Ingeniería de Sistemas durante el período 2018-2022. Esta población incluye un conjunto amplio y diverso de individuos que provienen mayoritariamente de la provincia de Puno y San Román y de otras provincias del departamento. La cantidad de postulantes entre ambas modalidades es de 5385. El dataset contiene 35 variables y ambas modalidades contienen la misma estructura.

Tabla 9

Cantidad de postulantes por modalidad

Nº	Proceso	Cantidad
1	General	3814
2	Cepreuna	1571
	Total	5385

Nota: Esta tabla muestra el total de postulantes por modalidad.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

A continuación, se muestra el código para calcular el total de postulantes por modalidad:

```
# cantidad por proceso (general - cepreuna)
# data['modalidad'].value_counts()
gener = data['modalidad'].isin([1,9]).sum()
cepre = data['modalidad'].isin([2]).sum()
print(f'Cantidad general: {gener}')
print(f'Cantidad cepreuna: {cepre}')
```

Tabla 10

Cant. de postulantes por fecha y modalidad examen GENERAL

Nº	Fecha de proceso	Cantidad
1	14/08/2022	448
2	19/03/2022	611
3	01/10/2021	387
4	11/06/2021	618
5	18/08/2019	629
6	13/01/2019	353
7	05/08/2018	334
8	27/05/2018	434

Nota: Esta tabla muestra la cantidad de postulantes ordenados por fecha de proceso de forma descendente.

A continuación, se muestra el código del resultado de la tabla anterior:

```
# Contar registros por modalidad y fecha
conteo = data.groupby(['modalidad', 'fecha']).size().reset_index(name='Cantidad')
# Ordenar por fecha y modalidad en orden descendente
conteo = conteo.sort_values(by=['modalidad', 'fecha'], ascending=[True, False])
conteo
```

Tabla 11

CEPREUNA Cantidad de postulantes por fecha y modalidad

Nº	Fecha de proceso	Cantidad
1	07/08/2022	244
2	12/03/2022	180
3	07/08/2021	205
4	14/11/2020	86
5	23/10/2020	115
6	23/06/2019	174
7	17/03/2019	185
8	16/12/2018	114
9	16/09/2018	121

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Nº	Fecha de proceso	Cantidad
10	24/06/2018	147

Nota: Esta tabla muestra la cantidad de postulantes ordenados por fecha de proceso de forma descendente.

3.3. MUESTRA

A diferencia de diferentes estudios que emplean un proceso de muestreo tradicional y suelen basarse en muestras representativas (López, 2004), esta investigación utilizó la totalidad de los datos disponibles de los postulantes de las dos modalidades, lo que garantizó una visión completa y exhaustiva de las tendencias y patrones en los exámenes de admisión (James et al., 2023). La ausencia de un proceso de muestreo minimiza la posibilidad de error asociada a la selección de una muestra por lo que se justifica la intención de maximizar la representatividad y la aplicabilidad de los resultados, aprovechando al máximo la disponibilidad de datos completos. Sin embargo, en el proceso de exploración y entendimiento de los datos se tuvo que depurar los datos disponibles, con el fin de garantizar la calidad y relevancia de los registros incluidos en el análisis. Como parte de este proceso, se seleccionó una muestra por conveniencia, siguiendo las recomendaciones metodológicas de Hernández et al., (2014).

3.3.1. Postulantes por las modalidades General y CEPREUNA

A continuación, se muestra los resúmenes de postulantes por modalidad y fecha de los procesos de admisión.

Tabla 12

Cant. de postulantes por modalidad para la muestra

Nº	Proceso	Cantidad
1	General	3353
2	Cepreuna	1501
	Total	4854

Nota: Esta tabla muestra el total de postulantes por modalidad para la muestra. Ver código Python en anexo 3.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 13

Cant. de postulantes examen General para la muestra

Nº	Fecha de proceso	Cantidad
1	14/08/2022	446
2	19/03/2022	608
3	01/10/2021	385
4	11/06/2021	617
5	18/08/2019	389
6	13/01/2019	321
7	05/08/2018	237
8	27/05/2018	350

Nota: Esta tabla muestra la cantidad de postulantes por fecha, estos datos serán usados para la muestra. Ver código Python en anexo 3.

Tabla 14

Cant. de postulantes examen CEPREUNA para la muestra

Nº	Fecha de proceso	Cantidad
1	07/08/2022	244
2	12/03/2022	180
3	07/08/2021	205
4	14/11/2020	84
5	23/10/2020	113
6	23/06/2019	169
7	17/03/2019	155
8	16/12/2018	100
9	16/09/2018	109
10	24/06/2018	142

Nota: Esta tabla muestra la cantidad de postulantes por fecha, estos datos serán usados para la muestra. Ver código Python en anexo 3.

3.4. TIPO DE INVESTIGACIÓN

La investigación aplicada tiene como objetivo principal encontrar soluciones a problemas reales y mejorar procesos existentes mediante la aplicación de teorías y técnicas científicas. Por lo tanto, esta investigación es de **tipo aplicada**, centrada en el uso práctico del conocimiento teórico para resolver problemas prácticos (Arias y Covinos, 2021), tal como indica Canales et al. (2020), en el campo de las ingenierías se busca aplicar estos conocimientos para dar lugar a la construcción de nuevos objetos (modelos).

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

El **enfoque es cuantitativo** dado que se basa en el análisis de datos demográficos de postulantes, utilizando técnicas estadísticas y cuya cuantificación de las variables y el uso de métricas específicas permiten medir la precisión y efectividad de los modelos predictivos.

3.5. DISEÑO DE LA INVESTIGACIÓN

El diseño de la presente investigación es no experimental dado que no realiza manipulación de variables, ya que se trabaja con datos históricos recopilados de los exámenes de admisión de los años 2018 – 2022; es **descriptivo** porque describe la población de postulantes, con características demográficas lo que permitió identificar patrones y comportamientos relevantes para el análisis predictivo y, **correlacional** debido a que se analizan las relaciones entre las variables independientes (demográficas) y la variable dependiente (resultado de los exámenes de admisión). (Hernández et al., 2014)

3.6. TÉCNICAS E INSTRUMENTOS

3.6.1. Las técnicas de recolección de datos

Para la recolección de datos se realizó la revisión del dataset proporcionado por la Dirección de Admisión de la Universidad Nacional del Altiplano, donde se contempla información demográfica de los postulantes.

3.6.1.1. Solicitud de información

La técnica utilizada en este estudio es la obtención de datos mediante la solicitud y uso de un conjunto de datos (dataset) presentada a la Dirección de Admisión de la UNA. Estos datos fueron obtenidos en formato Excel para el análisis, con el propósito de analizar y predecir los

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

resultados de los exámenes de admisión a partir de diversas características demográficas de los postulantes (ver anexo A).

3.6.2. Lenguaje de programación

Python es el lenguaje de programación principal para la implementación del modelo de aprendizaje automático.

3.6.3. Herramientas de análisis

Una vez obtenidos los datos, se utilizaron herramientas de análisis de datos, principalmente Python y bibliotecas como pandas, Numpy y Scikit-Learn, para preparar y analizar los datos. A través de este proceso, se logró un modelo predictivo de los resultados de los exámenes de admisión basado en características demográficas.

3.7. TRATAMIENTO DE DATOS

Los datos utilizados en el presente estudio fueron inicialmente obtenidos en formato Excel, lo que facilitó la organización preliminar de la información. Posteriormente, se empleó el lenguaje de programación Python para la carga y procesamiento de los datos, aprovechando su capacidad para manejar grandes volúmenes de información de manera eficiente y realizar análisis avanzados.

3.7.1. Exploración de datos

En este proceso, se realizó un análisis descriptivo de las variables demográficas de los postulantes, como edad, género, lugar de procedencia, tipo de institución educativa de origen, entre otros.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Inicialmente, se verificó la integridad y calidad de los datos, identificando valores faltantes, inconsistentes, los cuales fueron tratados mediante técnicas de limpieza y preprocesamiento. Posteriormente, se aplicaron herramientas estadísticas y visuales, como histogramas, diagramas de caja y gráficos de barras, para identificar patrones, distribuciones y posibles relaciones entre las variables y todo ello se realizó usando la librería Pandas y matplotlib.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En este capítulo se presentan los resultados obtenidos, acompañados de su respectiva interpretación y análisis. Para ello, se emplearon tablas y gráficos de diferentes tipos, entre otros recursos visuales, con el fin de ilustrar los hallazgos y facilitar su comprensión, organizados de acuerdo con cada objetivo específico planteado.

4.1. OBJETIVO ESPECÍFICO 1

Recopilar datos demográficos de post. entre 2018 y 2022

Para llevar a cabo la recopilación de datos demográficos de postulantes, se realizaron las siguientes acciones:

4.1.1. Solicitud y acceso a datos

Se gestionó la obtención de información a través de un requerimiento institucional formal dirigido a la Dirección de Admisión de la Universidad Nacional del Altiplano obteniendo información oficial sobre los postulantes de los años 2018 al 2022 (General y CEPREUNA). Los datos fueron proporcionados en formatos Excel, asegurando su compatibilidad con herramientas de análisis como Python.

4.1.2. Variables recopiladas

Los datos recopilados incluyeron las siguientes variables demográficas:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.1.2.1. Carácterísticas de las variables

El dataset incluye información sobre las características demográficas de los postulantes como los que se muestra en la siguiente tabla:

Tabla 15

Estructura del dataset de postulantes

Nº	Variable/característica	Tipo dato	Descripción
1	fecha	Cadena	Fecha de proceso
2	modalidad	Entero	Tipo de proceso de admisión
3	id_carrera	Entero	Código de la carrera
4	nombre	Cadena	Nombre de la carrera
5	dni	Cadena	Doc. Nac. del postulante
6	paterno	Cadena	Apellido paterno
7	materno	Cadena	Apellido materno
8	nombres	Cadena	Nombres del postulante
9	fec_nac	Fecha	Fecha nacimiento del postulante
10	genero	Cadena	Género del postulante
11	ubi_nac	Cadena	Ubigeo de nacimiento
12	iddep_nac	Cadena	Código de depart. Nac.
13	idprov_nac	Cadena	Código de prov. Nac.
14	iddist_nac	Cadena	Código de dist. Nac.
15	dep_nac	Cadena	Nombre depart. Nac.
16	prov_nac	Cadena	Nombre prov. Nac.
17	dist_nac	Cadena	Nombre dist. Nac.
18	ecivil	Entero	Estado civil
19	tipo_colegio	Cadena	Tipo de colegio
20	id_colegio	Entero	Identificador de colegio
21	nombre_col	Cadena	Nombre del colegio
22	id_departamento	Cadena	Código depart. colegio
23	id_provincia	Cadena	Código prov. Colegio
24	id_distrito	Cadena	Código dist. Colegio
25	depart_col	Cadena	Nombre depart. Colegio
26	prov_col	Cadena	Nombre prov. Colegio
27	dist_col	Cadena	Nombre dist. Colegio
28	tipo_ruralidad	Cadena	Tipo de ruralidad
29	egreso	Entero	Año de egreso del colegio
30	puntajetotal	Flotante	Puntaje obtenido del post.
31	est_resultado	Entero	Valor que indica post. apto
32	grado	Cadena	Grado estudios (4to/5to/normal)
33	edad	Entero	Edad del postulante

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Nº	Variable/característica	Tipo dato	Descripción
34	postulaciones	Entero	Cantidad de postulaciones
35	Ingreso (1/0)	Entero	Estado del postulante

Nota: Datos obtenidos de la oficina de la Dirección de Admisión de la UNA.

4.1.3. Criterio de exclusión

La eliminación de algunos registros en este análisis se basa en la necesidad de asegurar la consistencia y la validez de los datos examinados.

Postulantes de 4to y 5to grado de secundaria

Los registros de postulantes que aún no habían culminado sus estudios secundarios al momento del proceso de admisión fueron excluidos, debido a que sus características podrían no ser comparables con los de los postulantes regulares.

Postulantes con discapacidad

Los registros de postulantes que declararon alguna discapacidad (EXAMEN General) fueron excluidos del análisis. Esto responde a la intención de evitar sesgos en el modelo predictivo, dado que estos postulantes suelen tener procesos diferenciados de evaluación o adaptación en el examen de admisión.

Estructura y almacenamiento de datos

La estructura del archivo es tal como se muestra en la siguiente tabla y, fue almacenado en el Google Drive para su posterior análisis. La etapa de lectura consistió en leer el archivo del drive con la biblioteca Pandas y almacenar su contenido en un dataFrame. Esto facilitó la exploración, limpieza y preparación de la información. Este enfoque permitió garantizar la integridad y consistencia de los datos desde el inicio del proceso analítico.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 16

Parte de la estructura del dataset

fecha	modalidad	id_carrera	Nombre
2018-05-27	1	34	Ingeniería de Sistemas
2018-05-27	1	34	Ingeniería de Sistemas
2018-05-27	1	34	Ingeniería de Sistemas
2018-05-27	1	34	Ingeniería de Sistemas
2018-05-27	1	34	Ingeniería de Sistemas
2018-05-27	1	34	Ingeniería de Sistemas

Nota: Captura de pantalla del archivo almacenado en el drive.

A continuación, se muestra el código desarrollado en Python para la lectura de datos en un dataFrame.

```
# Importamos la biblioteca “drive” para montar la unidad
from google.colab import drive
# Montamos la unidad
drive.mount(‘/content/drive’)
data = pd.read_excel(“/content/drive/MyDrive/data.xlsx”, engine=’openpyxl’)
```

Confidencialidad y ética

Se implementaron medidas para proteger la confidencialidad de la información. Los datos fueron anonimizados y tratados según las normativas vigentes de protección de datos personales (Ley Nº 29733), asegurando la integridad y privacidad de los postulantes.

Este proceso permitió obtener un conjunto de datos limpio, estructurado y listo para ser utilizado en el análisis exploratorio y en el desarrollo de modelos de aprendizaje automático.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.2. OBJETIVO ESPECÍFICO 2

Realizar un análisis exploratorio para identificar patrones y tendencias relevantes

Para empezar el análisis exploratorio se empezó con el método **data.info ()** para realizar una revisión rápida de la estructura y características generales del dataset, esto para facilitar la preparación y limpieza de los datos, se tuvo un total de 5385 observaciones. Además, la presencia de variables categóricas y continuas permitió la aplicación de múltiples técnicas estadísticas y algoritmos de aprendizaje automático para identificar patrones significativos y realizar predicciones.

También, se identificaron valores nulos en las columnas **puntaje total** (5323) y **grado** (520), los cuales se trataron mediante estrategias adecuadas como imputación o eliminación durante el preprocesamiento de los datos.

Tabla 17

Estructura y tipo de datos de las variables

Nº	Column	Non-Null	Count	Dtype
1	Fecha	5385	non-null	object
2	modalidad	5385	non-null	int64
3	id_carrera	5385	non-null	int64
4	nombre	5385	non-null	object
5	dni	5385	non-null	int64
6	paterno	5385	non-null	object
7	materno	5385	non-null	object
8	nombres	5385	non-null	object
9	fec_nac	5385	non-null	datetime64[ns]
10	genero	5385	non-null	object
11	ubi_nac	5385	non-null	int64
12	iddep_nac	5385	non-null	int64
13	idprov_nac	5385	non-null	int64
14	iddist_nac	5385	non-null	int64
15	dep_nac	5385	non-null	object
16	prov_nac	5385	non-null	object
17	dist_nac	5385	non-null	object

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Nº	Column	Non-Null	Count	Dtype
18	ecivil	5385	non-null	int64
19	tipo_colegio	5385	non-null	int64
20	id_colegio	5385	non-null	int64
21	nombre_col	5385	non-null	object
22	IdDepartamento	5385	non-null	int64
23	IdProvincia	5385	non-null	int64
24	IdDistrito	5385	non-null	int64
25	depart_col	5385	non-null	object
26	prov_col	5385	non-null	object
27	dist_col	5385	non-null	object
28	tipo_ruralidad	5385	non-null	object
29	anno_egreso	5385	non-null	int64
30	puntajetotal	5323	non-null	float64
31	est_resultado	5385	non-null	int64
32	grado	520	non-null	object
33	edad	5385	non-null	int64
34	postulacion	5385	non-null	int64
35	ingreso	5385	non-null	int64

Nota: Resultado con el método info() del DataFrame

De la tabla anterior se observa la cantidad de tipos de datos del dataset: un (1) datetime (fecha), un (1) float64 (flotante) y quince (15) objects (string).

4.2.1. Análisis inicial de las variables detectadas - preprocesamiento

Formato incorrecto en la variable fecha:

La variable 'fecha', tabla 16, se encuentra representada como un tipo de dato object en lugar de datetime. Esto indica que los valores no están almacenados en un formato reconocido como fecha, lo que podría impedir la manipulación y análisis temporal.

Código para la conversión a datetime de la variable fecha:

```
# convirtiendo la variable 'fecha' (del proceso) en datetime
data['fecha'] = pd.to_datetime(data['fecha'])
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Valores inconsistentes en la variable ‘edad’:

La variable 'edad' presenta valores negativos, lo cual es incoherente en el contexto de esta variable, ya que se espera que represente edades en años positivos. En la tabla 17 se observa como edad mínima el valor de -2. Dado que la información la registra los postulantes, estos valores se debieron a errores de digitación.

Tabla 18

Resumen de las variables numéricas

	Anno_egreso	Puntajetotal	Edad
count	5385.000000	5323.000000	5385.000000
mean	2017.971588	1173.083570	17.885794
min	1991.000000	0.000000	-2.000000
25%	2017.000000	720.000000	17.000000
50%	2018.000000	1104.000000	17.000000
75%	2020.000000	1579.000000	19.000000
max	2021.000000	3419.250000	45.000000
std	2.395539	574.782411	2.376471

Nota: captura de pantalla del notebook

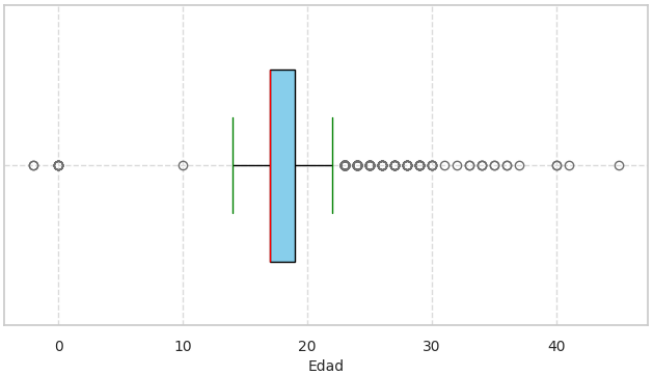
Para explorar los valores atípicos presentes en la variable **'edad'**, se generó un diagrama de boxplot utilizando la librería seaborn de visualización de datos en Python. Este gráfico permitió identificar de manera clara los valores extremos que se encuentran fuera del rango intercuartil (IQR), definido por los límites superior e inferior del conjunto de datos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 10

Boxplot de edades en el preprocesamiento



Nota: captura de pantalla del notebook

Código usado para mostrar el boxplot de edades de postulantes:

```
# Estilo Seaborn
sns.set(style="whitegrid")  # Estilo de fondo blanco con rejilla
# Crear el boxplot con estilo
plt.figure(figsize=(8, 4))  # Tamaño del gráfico
sns.boxplot(x='edad', data=data, color='skyblue',
            fliersize=6,  # Tamaño de los puntos fuera de los bigotes
            linewidth=1,  # Grosor de las líneas
            width=0.6,  # Ancho de la caja
            boxprops=dict(facecolor='skyblue', edgecolor='black'),  # Colores de la caja
            whiskerprops=dict(color='black', linewidth=1),  # Estilo de los bigotes
            capprops=dict(color='green', linewidth=1),  # Estilo de las tapas
            medianprops=dict(color='red', linewidth=1))  # Estilo de la mediana
# Títulos y etiquetas
plt.title('Boxplot de Edades', fontsize=13, fontweight='bold')
plt.xlabel('Edad', fontsize=10)
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
# # Rejilla de fondo y muestra del gráfico
plt.grid(True, linestyle='--', alpha=0.7)
plt.show()
```

En los procesos de admisión se observa postulantes con edades que superan los 15 años de edad, para reemplazar las edades negativas con el promedio de las edades válidas se usó la técnica de **imputación mediante promedio**. Esta técnica es útil cuando el número de valores negativos es bajo y no introduce un sesgo significativo en los datos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

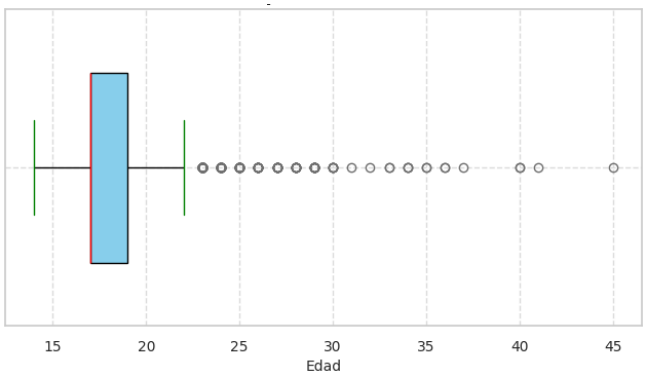
Calculando la edad promedio de los postulantes que son mayores de 14 años para imputar datos erróneos.

```
# Cálculo del promedio de edades
promedio = data.loc[data['edad']>14, 'edad'].mean()
# Filtra y reemplaza la columna edad bajo la condición indicada
data.loc[data['edad']<14,'edad'] = int(promedio)
```

Viendo los nuevos cambios

Figura 11

Boxplot de edades corregidas



Nota: captura de pantalla del notebook

Análisis de la variable años de egreso (anno_egreso)

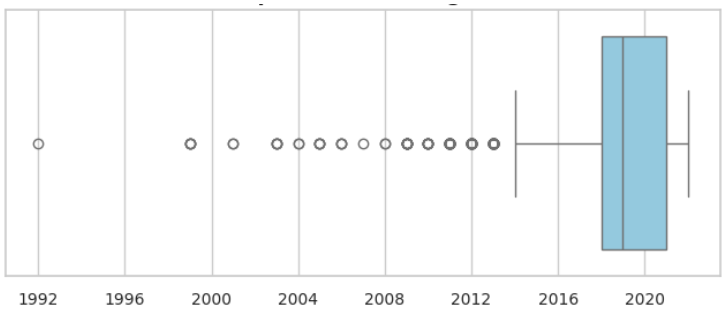
Para evaluar la distribución de los años de egreso de los postulantes, se generó un diagrama de caja y bigotes (*boxplot*), el cual permite identificar la dispersión de los datos, la presencia de valores atípicos y la tendencia central de la variable.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 12

BoxPlot de año de egreso de postulantes



Nota: captura de pantalla del notebook

El análisis gráfico muestra que los valores de "Año de Egreso" se encuentran en un rango que va desde 1992 hasta 2022. Sin embargo, se observa que la mayoría de los datos están concentrados en los años más recientes. En particular, el límite inferior del bigote se encuentra entre el año 2013, lo que indica que la mayoría de los postulantes de los procesos General y CEPREUNA egresaron del colegio a partir de esta fecha. Los valores anteriores a este punto pueden ser considerados como posibles valores atípicos, representando casos excepcionales en el conjunto de datos.

4.2.1.1. Aplicando ingeniería de características

El marco teórico menciona la ingeniería de características, por lo que en este apartado se describe el procedimiento técnico llevado a cabo para transformar la variable **año de egreso** (convertir a tipo fecha) y calcular la nueva variable "**dif_anios**", que representa el tiempo transcurrido entre la fecha de egreso del colegio y la fecha de postulación, esto con la finalidad de mejorar los modelos de Machine Learning y predecir un mejor resultado.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 19

Vista de la variable año de egreso

	Anno_egreso
0	2018
1	2012
2	2018
3	2017
4	2017

Nota: captura de pantalla del notebook

Conversión de la variable año de egreso

La variable año de egreso (anno_egreso), originalmente representada como un valor numérico (por ejemplo, 2018), fue transformada a un formato de fecha completo 30-12-AAAA. Esto se realizó utilizando la función pd.to_datetime de Python, donde cada valor fue completado con el día y mes estándar 30-12. Esta transformación asegura que la variable sea del tipo datetime, permitiendo su uso en cálculos posteriores.

```
# años de pasaron desde que salio del colegio hasta el momento de postular.
data['anno_egreso'] = pd.to_datetime(data['anno_egreso'].astype(str) + '-12-30')
# Convertir la columna 'Fecha_Examen' a datetime
data['anno_egreso'] = pd.to_datetime(data['anno_egreso'])
```

Cálculo de la variable diferencia de años (dif_anios)

Para determinar el tiempo transcurrido entre el egreso del colegio y la fecha de postulación a la universidad, se realizó la resta directa entre las variables fecha de postulación (fecha) y fecha de egreso (anno_egreso). El resultado de esta operación se almacenó como una **nueva variable** llamada “dif_anios”, calculada como la diferencia en años entre ambas fechas. Tal como se muestra en el siguiente código en Python.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

```
# Cálculo de difierencia de años desde que salio del colegio
data['dif_anios'] = np.where(data['grado'] =='Normal', (data['fecha'] - data['anno_egreso']).dt.total_seconds() / (365.25 * 24 * 3600), np.nan)
```

Visualización de datos procesados

Tabla 20

Vista de variables procesadas

	Anno_egreso	dif_anios
1913	1991-12-30	27.633128
4352	1998-12-30	23.216975
148	1998-12-30	19.405886
5105	1998-12-30	23.622177
5073	1998-12-30	23.622177

Nota: captura de pantalla del notebook

4.2.1.2. Eliminación de variables innecesarias

Durante el procesamiento de datos, es fundamental garantizar que las variables incluidas en el análisis aporten valor a los objetivos de este estudio y no generen ruido que pueda afectar los resultados. Para ello, se evaluaron todas las variables presentes en el conjunto de datos, y aquellas consideradas innecesarias fueron eliminadas. A continuación, se detallan las variables eliminadas:

- **orden**: Representa un identificador secuencial de los registros, sin valor analítico.
- **dni**: Información personal sensible que no aporta al análisis.
- **paterno, materno, nombres**: Información irrelevante para el análisis de patrones y tendencias.
- **id_carrera y nombre_col**: Variables redundantes cuando ya se cuenta con descripciones categóricas consolidadas y únicas.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- **est_resultado**: Contiene valor categórico si logró cumplir todas las etapas para obtener el a la escuela que postuló, además, esta variable ya se encuentra considerada como objetivo en otra.
- **ubi_nac, iddep_nac, idprov_nac, iddist_nac, id_colegio, IdDepartamento, IdProvincia, IdDistrito**: variables numéricas considerados como identificador de departamento, provincia, distrito y código de colegio.
- **ecivil**: Variable numérica y única.

El código para eliminar las columnas innecesarias es el siguiente:

```
# Eliminando columnas innecesarias
data.drop(columns=['id_carrera', 'nombre', 'paterno', 'materno', ‘nombres', 'est_resultado', 'ubi_nac', 'iddep_nac', 'idprov_nac', 'iddist_nac', 'id_colegio', 'IdDepartamento', 'IdProvincia', 'IdDistrito', 'ecivil'], inplace=True)
```

4.2.1.3. Eliminación de postulantes e ingresantes por discapacidad

La eliminación de estos registros se realizó con el objetivo de garantizar la homogeneidad de la muestra y evitar sesgos de selección que pudieran afectar la interpretación de los resultados. Este filtrado de datos contribuye a preservar la validez interna del estudio, asegurando que los modelos analíticos reflejen patrones generales de desempeño en postulantes sin discapacidad. En la siguiente tabla se puede visualizar la cantidad de discapacitados.

Tabla 21

Cantidad de postulantes por modalidad

	Modalidad	Edad
1	3803	17.924270
2	1571	17.879058
3	11	20.454545

Nota: captura de pantalla del notebook

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

La eliminación se realizó con la siguiente línea:

```
# Eliminando
 data = data[ data['modalidad'] != 9 ]
```

4.2.1.4. Análisis de postulantes por modalidad y género

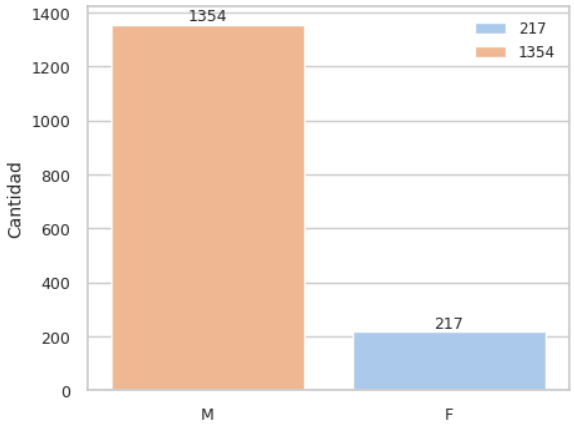
Se realizó un análisis de la cantidad de postulantes según la modalidad de admisión, diferenciando entre CEPREUNA Y GENERAL, así como su distribución por género.

Distribución de postulantes en la modalidad CEPREUNA

En la modalidad CEPREUNA, el total de postulantes es 1,571, de los cuales **1,354 (86.2%)** son varones y **217 (13.8%)** son mujeres. Esto indica que la gran mayoría de los postulantes en esta modalidad son de género masculino, lo que sugiere una brecha de participación en los procesos de admisión.

Figura 13

Cant. de postulantes por género, modalidad CEPREUNA



Nota: captura de pantalla del notebook

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

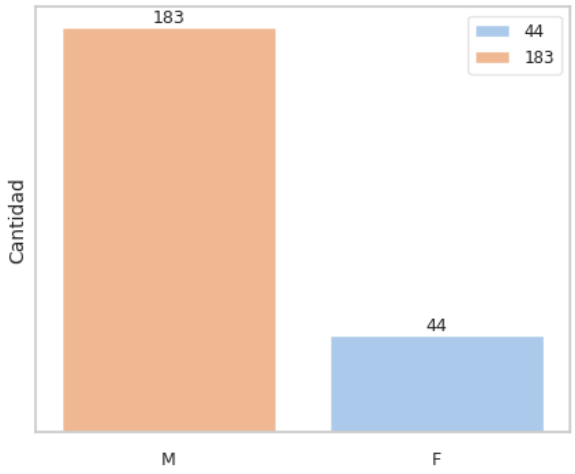
UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Distribución de ingresantes en la Modalidad CEPREUNA

En la modalidad **CEPREUNA**, de un total de **227** ingresantes, 183 **(80.6%)** son varones y 44 **(19.4%)** son mujeres. Si bien la proporción de varones sigue siendo mayoritaria, se observa un incremento relativo en la participación femenina en comparación con los postulantes en esta misma modalidad, donde representaban solo el **13.8%**. Esto sugiere que la tasa de ingreso de las mujeres en CEPREUNA es relativamente más alta que la de los varones.

Figura 14

Cant. de ingresantes por género, modalidad CEPREUNA



Nota: captura de pantalla del notebook

Distribución de postulantes en la modalidad examen General

En la modalidad examen General, el total de postulantes asciende a **3,597**, con una distribución de **3,380 (94.0%)** varones y **217 (6.0%)** mujeres. Se observa que la proporción de postulantes masculinos es aún mayor en comparación con la modalidad CEPREUNA.

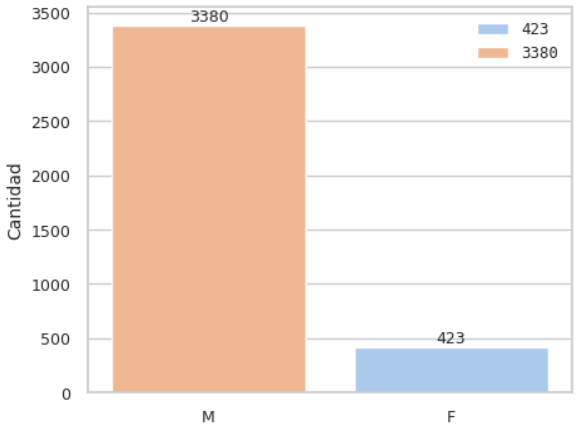
repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

El análisis comparativo entre ambas modalidades refleja una tendencia general en la que los postulantes varones superan ampliamente a las mujeres en ambas modalidades de admisión. Este resultado puede ser relevante para estudios sobre la equidad de género en el acceso a la educación superior y en la toma de decisiones para la implementación de políticas de inclusión.

Figura 15

Cant. de postulantes por género, modalidad General



Nota: captura de pantalla del notebook

Distribución de ingresantes en la modalidad examen General

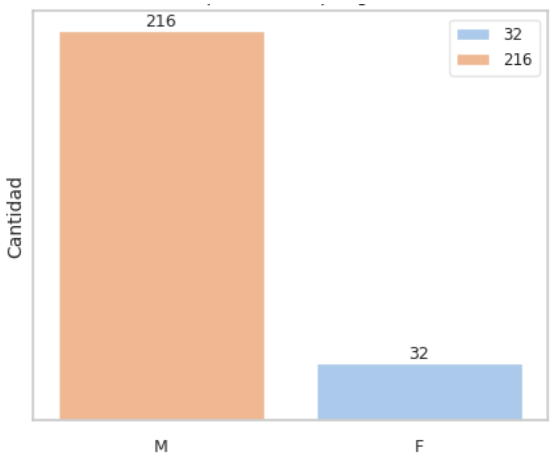
En la modalidad Examen General, el número total de ingresantes es **248**, con **216 (87.1%)** varones y **32 (12.9%)** mujeres. Aquí se mantiene la tendencia observada en los postulantes, donde los varones tienen una participación significativamente mayor. Sin embargo, la proporción de mujeres ingresantes sigue siendo menor en comparación con la modalidad CEPREUNA, lo que indica que su tasa de éxito en el examen General es más baja.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 16

Cant. de ingresantes por género, modalidad General



Nota: captura de pantalla del notebook

Comparación general

Al comparar ambas modalidades, se observa que la modalidad CEPREUNA presenta una mayor tasa de ingreso femenino (19.4%) en comparación con el examen General (12.9%). Esto podría indicar que las estudiantes mujeres tienen un mejor desempeño en un proceso de admisión más guiado o con preparación previa, como es el caso de CEPREUNA.

Por otro lado, en términos absolutos, la cantidad de varones ingresantes es mayor en ambas modalidades, lo que refleja una tendencia general en la composición de los ingresantes a la Escuela Profesional de Ingeniería de Sistemas comprendidos entre los años 2018 y 2022.

Cada uno de los gráficos mostrados arriba, fueron generados con el siguiente código en Python:

```
def graficar_genero (data, modalidad = [1,2], ingreso = [0,1], texto = ""):
  # Contar la cantidad de cada categoría
  conteo_genero = data.loc[(
      data['genero'].isin(['M','F']) &
       (data['modalidad'].isin(modalidad)) &
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

```
        (data['ingreso'].isin(ingreso))), 'genero'
        ].value_counts()
  print(conteo_genero)
  # Crear el gráfico de barras
  plt.figure(figsize=(5, 4))
 sns.barplot(x = conteo_genero.index, y = conteo_genero.values, hue = conteo_genero.values, palette = 'pastel')
  # Personalizar el gráfico
  plt.title(f'Distribución de postulantes por género: {texto}', fontsize=10)
  plt.xlabel('Género', size=10)
  plt.ylabel('Cantidad', size=10)
  plt.xticks(fontsize=9)
  plt.yticks(fontsize=9)
  # Personalizar la leyenda
  plt.legend(
      loc='upper right',            # Posición de la leyenda
      frameon=True,                 # Mostrar borde
      framealpha=0.5,               # Transparencia del borde
      shadow=False,                 # Sombra detrás de la leyenda
      prop={'size': 9, 'family': 'monospace'}  # Fuente de la leyenda
  )
  # Añadir etiquetas en las barras
  for i, valor in enumerate(conteo_genero.values):
      plt.text(i, valor + 0.1, str(valor), ha='center', va='bottom', fontsize=9)
  # Divido yAxis de 500 en 500
  # plt.yticks(np.arange(0, conteo_genero[0], 500))  # Muestra marcas en el eje Y desde 0 hasta 100 en pasos de 10
  # elimina la numeracion de yAxis
  plt.yticks([])
  # Mostrar el gráfico
  plt.show()
```

4.2.1.5. Análisis de ingresantes por género y fecha de proceso

Del análisis de los datos de ingresantes en ambas modalidades (General y CEPREUNA), se pueden extraer las siguientes conclusiones clave, considerando el impacto de la pandemia por COVID-19.

Impacto del COVID-19 en las tendencias de ingreso:

Antes de la pandemia, el último examen General realizado se llevó a cabo el 23 de junio de 2019. Para el caso de CEPREUNA el último examen realizado fue el 18 de agosto de 2019. Posteriormente, debido a las restricciones sanitarias y las medidas de confinamiento, los procesos

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

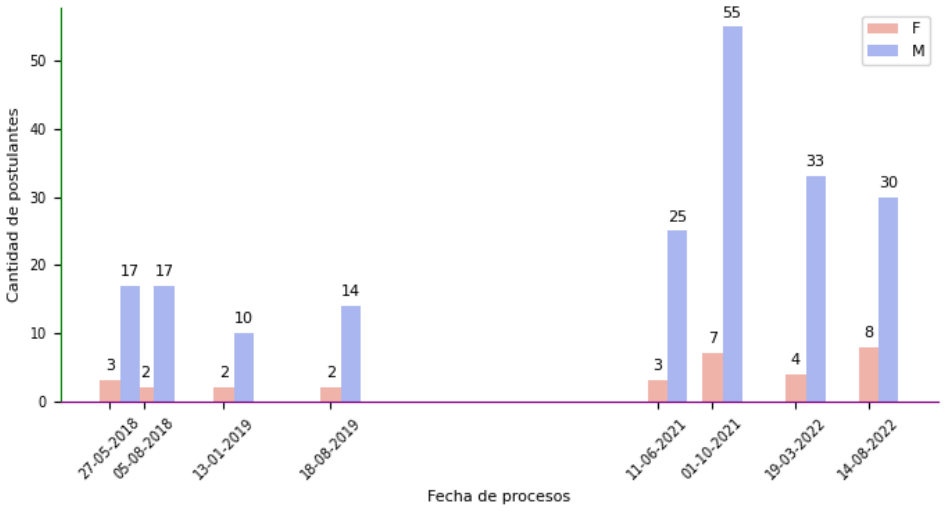
UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

de admisión fueron suspendidos hasta que las condiciones permitieron su reanudación.

- El examen General se retomó después de dos años, el 11 de junio de 2021.

Figura 17

General - cantidad de ingresantes por género y fecha



Nota: captura de pantalla del notebook

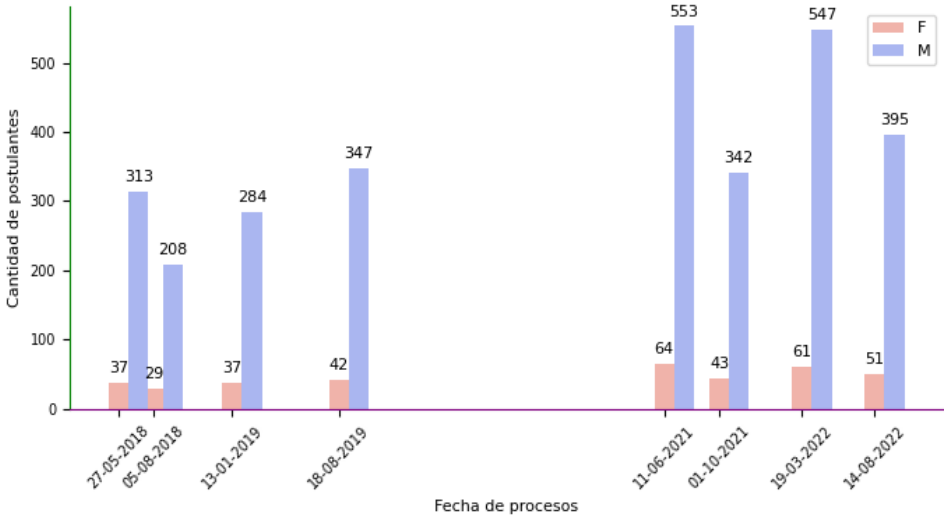
El periodo sin admisiones generó acumulación de postulantes (ver figura 18), lo que incrementó la competencia en los exámenes posteriores a la pandemia. Con fecha de 18 de agosto de 2019 se tuvo **389 postulantes** y después de la pandemia y fecha de proceso 11 de junio de 2021 se incrementó a **617 postulantes,** lo que equivale a un **58.6%** de incremento respecto al anterior proceso de admisión (ver fig. 17).

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 18

General - cant. de postulantes por género y fecha de proceso

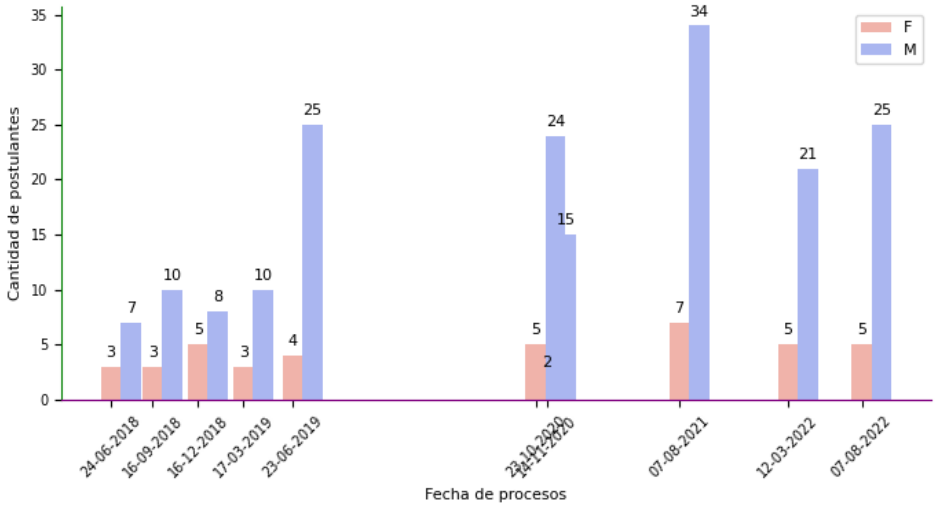


Nota: captura de pantalla del notebook

El examen CEPREUNA se retomó antes, el **23 de octubre de 2020**, aproximadamente un año y cuatro meses después del último examen.

Figura 19

CEPREUNA – cant. de ingresantes por género y fecha



Nota: captura de pantalla del notebook

Sucede lo contrario con el examen CEPREUNA. A la reanudación del examen se vio disminuido la cantidad de postulantes. En el examen de

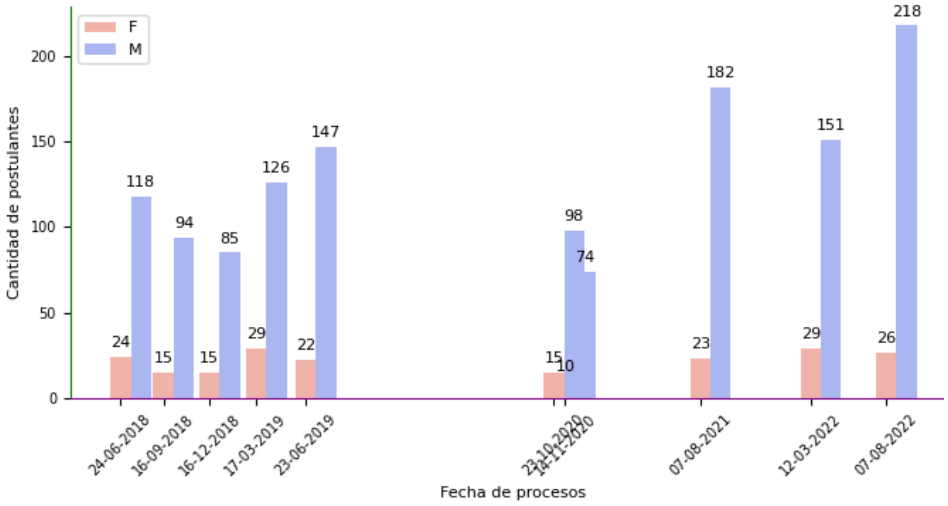
repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

fecha 23 de junio de 2019 se tuvo 169 postulantes y, posterior a la pandemia se tuvo 113 postulantes en el proceso del 23 de noviembre de 2020 (ver fig. 20). En el caso de los ingresantes la cantidad de ingresantes se mantuvo a la fecha del primer examen realizado el 23 de noviembre de 2020 después de la pandemia (ver fig. 19).

Figura 20

CEPREUNA - cant. de post. por género y fecha de proceso



Nota: captura de pantalla del notebook

Comparación general

La pandemia de COVID-19 no sólo afectó el calendario de admisión, sino que también evidenció las **brechas tecnológicas** en regiones como Puno, donde la conectividad limitada impidió la transición a exámenes virtuales. La reanudación de los exámenes de forma presencial garantizó un proceso más equitativo, aunque representó desafíos logísticos y sanitarios que debieron ser superados.

Cada uno de los gráficos mostrados arriba fueron generados por el siguiente código en Python:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

```
def graficar_ingresantes(result, colors, titleEjeY, Title):
    # Obtener las fechas y los valores del DataFrame
    fechas = result.index.get_level_values('fecha').tolist()
    valores = result.values
    # Convertir las fechas a números para posicionar las barras
    fechas_numericas = mdates.date2num(fechas)

    # Crear el gráfico de barras
    fig, ax = plt.subplots(figsize=(7, 4))
    # Ancho de las barras (aumentado para hacerlas más anchas)
    ancho_barras = 80.0  # Aumenta este valor para barras más anchas
    # Graficar cada columna del DataFrame
    for i, column in enumerate(result.columns):
        ax.bar(fechas_numericas + i * (ancho_barras / len(result.columns)),
result[column],
               width=ancho_barras / len(result.columns), color=colors[i],
label=column)
    # Formatear el eje X para que muestre las fechas correctamente
    ax.xaxis.set_major_formatter(mdates.DateFormatter('%d-%m-%Y'))
    ax.xaxis.set_major_locator(mdates.DayLocator(interval=1))
    # Quitar los bordes del gráfico
    ax.spines['top'].set_visible(False)
    ax.spines['right'].set_visible(False)
    # Cambiar el color de las líneas de los ejes
    ax.spines['left'].set_color('green')
    ax.spines['bottom'].set_color('purple')
    # Añadir título y etiquetas de los ejes
    plt.title(Title, loc='center')
    plt.xlabel('Fecha de procesos', size=8)
    plt.ylabel(titleEjeY, size=8)
    plt.legend(prop={'size': 8})
    # Rotar las etiquetas de fecha para mejor legibilidad
    plt.xticks(fechas_numericas, [fecha.strftime('%d-%m-%Y') for fecha in
fechas], rotation=45, size=7)
    plt.yticks(fontsize=7)
    # Agregar los valores sobre las barras
    for fecha, valor in zip(fechas_numericas, valores):
        for i, v in enumerate(valor):
            ax.annotate(f'{v:.0f}', (fecha + i * (ancho_barras / len(result.columns)),
v),
                        textcoords="offset points", xytext=(0, 5), ha='center', fontsize=8,
color='black')
    # Ajustar el layout para que no se corten las etiquetas
    plt.tight_layout()
    # Mostrar el gráfico
    plt.show()
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.2.1.6. Análisis del número de postulaciones

Los datos analizados reflejan la cantidad de postulantes según el número de intentos que realizaron para ingresar a la universidad en sus dos modalidades, General y CEPREUNA.

Acerca de la cantidad de intentos por postulante

Bajo número de intentos (1 a 3 veces): Se destaca que 2009 personas postularon solo una vez, mientras que 618 lo hicieron dos veces y 241 lo intentaron tres veces.

Postulantes con intentos intermedios (4 a 6 veces): Se observa una reducción progresiva en la cantidad de postulantes conforme aumenta el número de intentos. 92 personas postularon cuatro (4) veces, 56 lo hicieron cinco (5) veces y 17 postularon seis (6) veces.

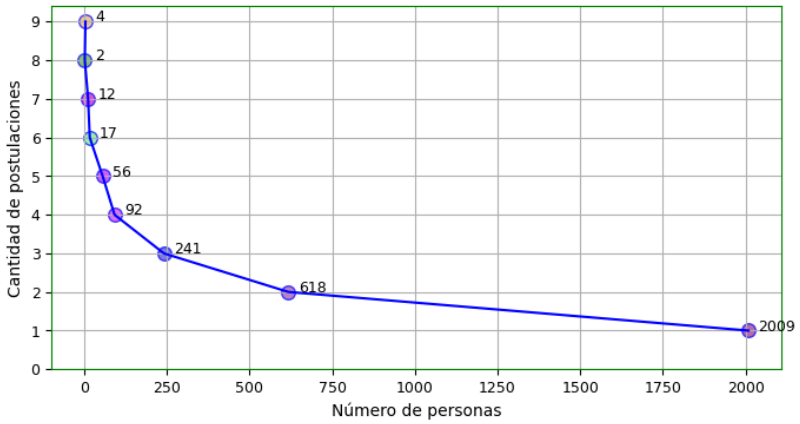
Postulantes con alta persistencia (7 a 9 veces): Son casos excepcionales de personas que han postulado en múltiples ocasiones. Doce (12) personas intentaron siete (7) veces, dos (2) lo hicieron ocho (8) veces y cuatro (4) llegaron a postular hasta en nueve (9) oportunidades y esto lo apreciamos en la siguiente figura.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 21

Número de intentos para obtener una vacante

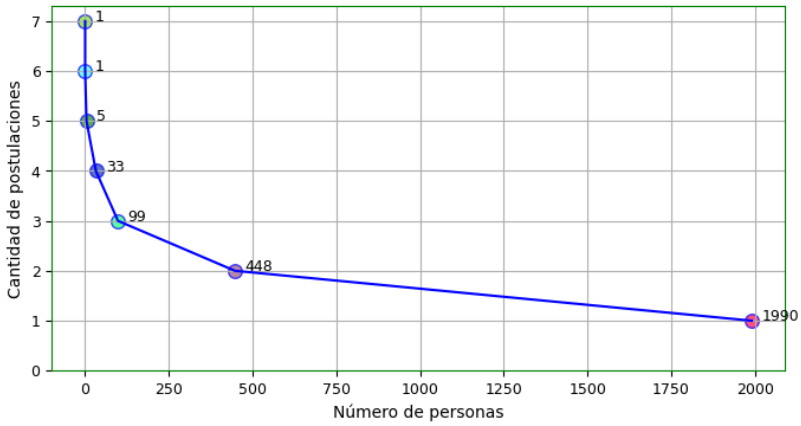


Nota: Este gráfico muestra el resultado de ambos procesos de admisión, General y CEPREUNA.

A continuación, se muestra el número de intentos de postulación a la Escuela Profesional de Ingeniería de Sistemas por las modalidades General y CEPREUNA.

Figura 22

Intentos de postulación modalidad General



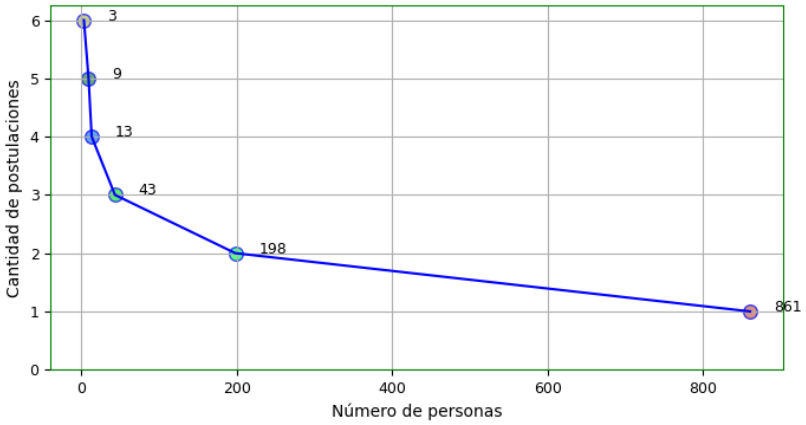
Nota: Este gráfico muestra los intentos de postulación modalidad General

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 23

Intentos de postulación modalidad CEPREUNA



Nota: Este gráfico muestra los intentos de postulación modalidad CEPREUNA

Acerca de la cantidad de intentos por ingresantes

Los siguientes datos analizados reflejan la cantidad de ingresantes según el número de intentos que realizaron para ingresar a la universidad en sus dos modalidades, General y CEPREUNA.

Bajo número de intentos (1 a 3 veces) se destaca que 2009 personas postularon solo una vez, mientras que 618 lo hicieron dos veces y 241 lo intentaron tres veces.

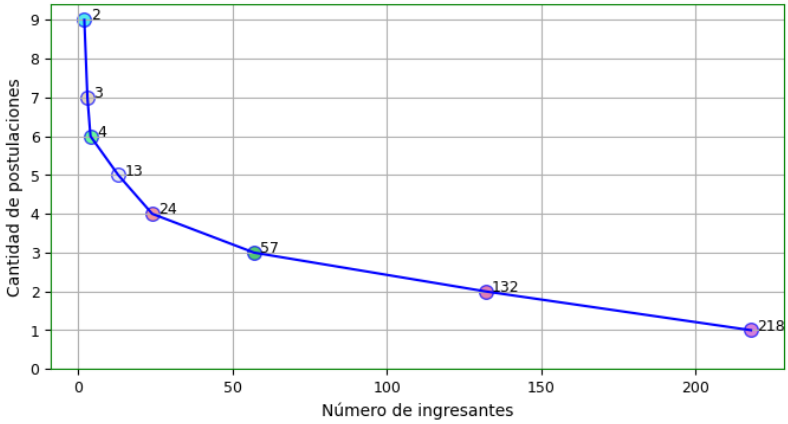
Respecto a la figura 24, se puede observar el resultado obtenido luego de cada examen. Por ejemplo, dos (2) ingresaron luego de su noveno (9) intento a obtener una vacante de ingreso; tres (3) ingresaron luego de su séptimo (7) intento; cuatro (4) ingresaron luego de su sexto (6) intento, etc.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 24

Ingresantes según intentos de postulación



Nota: Ingresantes según intentos de postulación, modalidad General y CEPREUNA

4.2.1.7. Relación entre la variable edad y puntaje

La figura 24 se enfoca en postulantes menores de 30 años, los puntajes representados en el eje vertical oscilan entre **1000 y 3500**. La dispersión de los puntos en el gráfico sugiere que no existe una correlación lineal fuerte entre la edad y los puntajes de admisión. Es decir, la edad no es un factor determinante en el desempeño de los postulantes en estos procesos. Esto se corrobora con el siguiente resultado:

Coeficiente de **correlación Pearson**: -0.0481; lo que indica que la correlación entre las variables es muy débil y negativa.

p-valor: 0.0008; es muy pequeño ($p < 0.05$), lo que sugiere que la correlación observada es **estadísticamente significativa**. En otras palabras, hay suficiente evidencia para creer que el efecto observado (en este caso, la correlación) es real y no simplemente una coincidencia (ocurrido al azar).

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

El código usado es el siguiente:

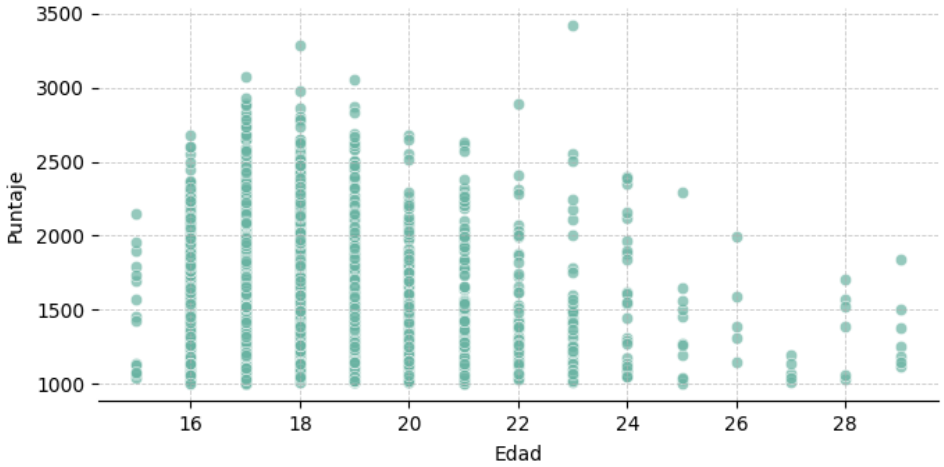
```
from scipy.stats import pearsonr
# Calcular la correlación de Pearson
correlacion = data['edad'].corr(data['puntajetotal'])
print(f'Coeficiente de correlación de Pearson: {correlacion:.4f}')
# Alternativamente, usando scipy.stats
corr_pearson, p_value = pearsonr(data['edad'], data['puntajetotal'])
print(f'Correlación Pearson: {corr_pearson:.4f}, p-valor: {p_value:.4f}')
```

Concentración de Datos

La densidad de los puntos en ciertas áreas del gráfico sugiere que hay edades en las que los postulantes obtienen puntajes más consistentes entre 17 y 19 años de edad. Esto podría estar relacionado con factores como la preparación académica, la experiencia previa o el nivel de madurez.

Figura 25

Relación edad y puntaje obtenido



Nota: No existe Ningún tipo de correlación entre las variables puntaje y edad.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

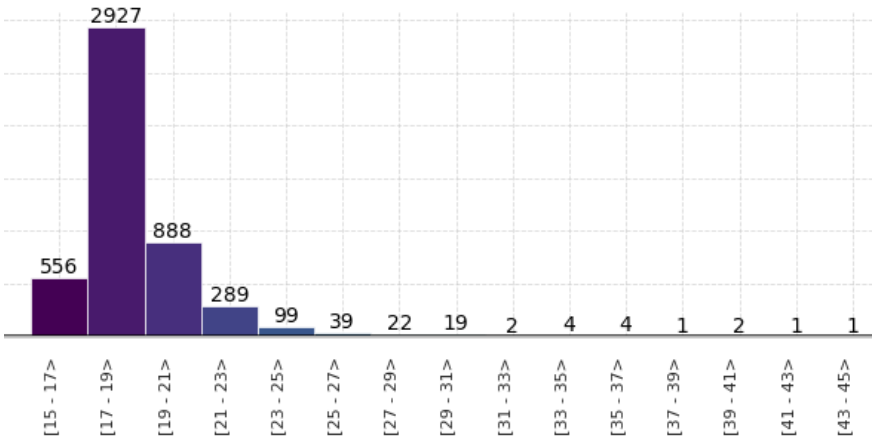
4.2.1.8. Análisis de postulantes e ingresantes por edad

Análisis de postulantes

En este análisis de la edad se realizó una distribución etaria de los postulantes a la universidad y su relación con la frecuencia de participación en el proceso de admisión. Para este estudio, se han segmentado los postulantes en rangos de edad, permitiendo visualizar cómo varía la cantidad de postulantes conforme aumenta la edad. Esto se puede apreciar en el siguiente gráfico.

Figura 26

Distribución etaria de postulantes por edad



Nota: Cantidad de postulantes distribuidos por rango de edades.

En la siguiente tabla se puede apreciar el porcentaje de incremento de postulantes respecto a un rango de edades. Tenemos que edades comprendidas entre 17 y 19 años de edad son un 59.24% del total de postulantes en este análisis; entre 19 y 21 años de edad se tiene un 17.97% que corresponde a la mayor cantidad de postulantes, estos son periodos en lo que los postulantes buscan ingresar a la universidad poco después de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

finalizar su educación secundaria. En los siguientes rangos de edades va decrementando la cantidad de postulantes.

Tabla 22

Porcentaje de postulantes por edades

Rango	Cant. Post.	% total
15 – 17	556	11.26%
17 – 19	2927	59.24%
19 – 21	888	17.97%
21 – 23	289	5.85%

Nota: Los puntos al final de la tabla corresponden a postulantes en menor cantidad mayores de 25 años.

Análisis de ingresantes

Al comparar la cantidad de ingresantes en cada rango de edad, se puede observar que la mayoría de los ingresantes provienen de los grupos más jóvenes, particularmente entre los **16 y 18 años**, que acumulan un total de **286 ingresantes**, lo que representa la mayor tasa de éxito en el proceso de admisión.

A medida que la edad avanza, la cantidad de ingresantes disminuye significativamente. En el grupo de **18-20 años**, el número de ingresantes es de 80, lo que indica que aún existe una oportunidad considerable de ingreso en este rango. Sin embargo, la caída se acentúa en los siguientes grupos etarios: entre **20-21 años (53 ingresantes)** y **21-22 años (16 ingresantes)**, lo que sugiere una menor probabilidad de éxito para quienes postulan en edades más avanzadas.

A partir del rango **22-24 años**, los ingresantes son mínimos, con solo 4 personas en este grupo y 9 en el de **24-25 años**. Finalmente, en los rangos superiores a **25 años**, prácticamente no hay ingresantes y su

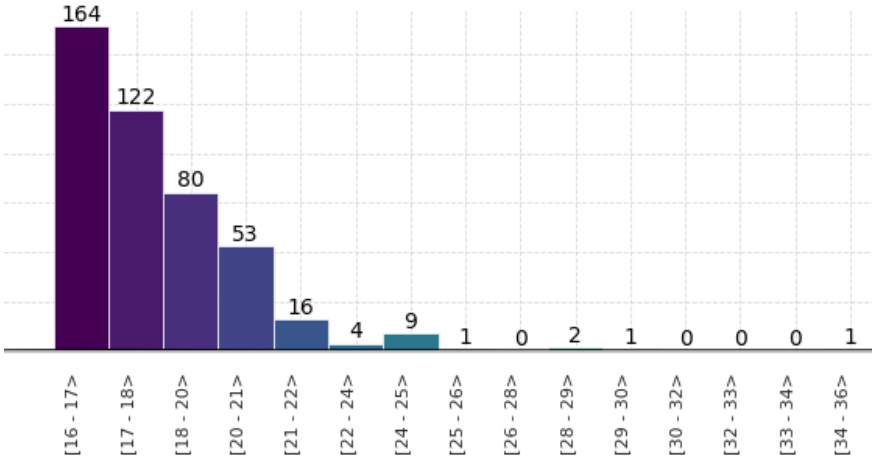
repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

probabilidad de ingreso es extremadamente baja, representando solo el **0.22%** del total de ingresantes y, con sólo **1 persona ingresante hasta los 36 años** y ningún ingresante en edades mayores. Se observa que los menores de 20 años tienen una mayor tasa de éxito en el examen de admisión, lo que podría estar relacionado con la preparación académica reciente y el enfoque en los estudios. Esto se puede apreciar en la siguiente figura.

Figura 27

Distribución de ingresantes por edad



Nota: Los ingresantes mayores a 25 años son considerados para el análisis de los modelos.

4.2.1.9. Análisis de postulante/ingresantes por provincia

Postulantes

El gráfico siguiente muestra la distribución de postulantes en ambas modalidades de estudio según la provincia de origen dentro del departamento de Puno. Se observa una notable diferencia en la cantidad de postulantes entre provincias, destacando las siguientes tendencias:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Puno (1245 postulantes): Es la provincia con la mayor cantidad de postulantes, lo que sugiere una alta concentración de estudiantes que buscan acceder a la universidad en esta zona.

San Román (843 postulantes): Le sigue en cantidad de postulantes, reflejando la importancia de esta provincia como un centro de atracción académica.

Azángaro (499 postulantes): Ocupa el tercer lugar, aunque con una cantidad considerablemente menor en comparación con las dos primeras provincias.

Provincias con menor cantidad de postulantes

Presenta la menor cantidad de postulantes entre las provincias de Moho (58 postulantes), San Antonio de Putina (109 postulantes), Lampa (113 postulantes), etc.

Distribución general

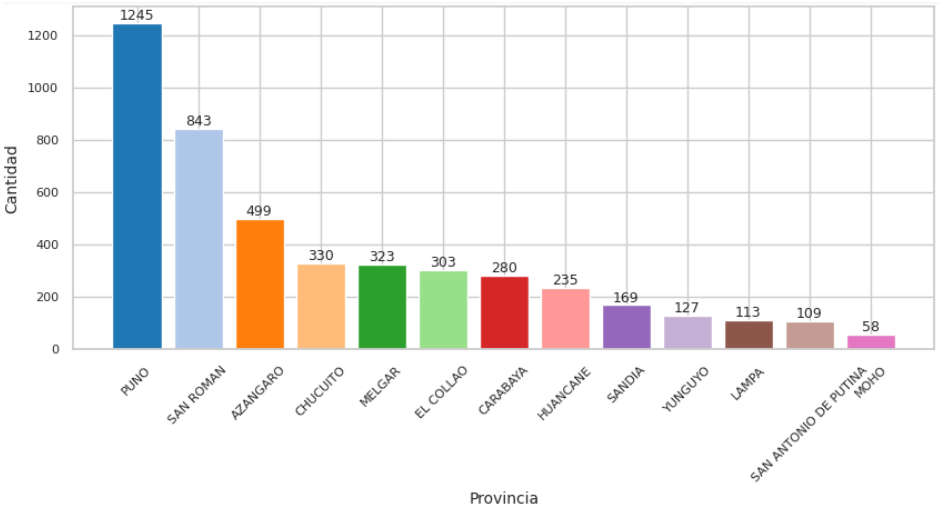
Se observa que las provincias con mayor densidad poblacional (Puno, San Román) tienden a tener una mayor cantidad de postulantes. En cambio, las provincias más alejadas y con menor población en términos de educación secundaria presentan cifras significativamente menores.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 28

Cantidad de postulantes por provincia



Nota: Gráfico de la cantidad de postulantes por modalidad - General y CEPREUNA

Ingresantes

El gráfico siguiente muestra la cantidad de ingresantes a la Escuela Profesional de Ingeniería de sistemas (UNA-Puno) según su provincia de procedencia. Se observa una distribución similar a la de los postulantes, con una mayor concentración de ingresantes en provincias cercanas a la sede universitaria. Entre las dos provincias más resaltantes tenemos a:

Puno lidera con 129 ingresantes, lo que confirma la tendencia observada en la cantidad de postulantes.

San Román (81 ingresantes) se mantiene en el segundo lugar, lo que indica que esta provincia también tiene un alto número de estudiantes que logran ingresar a la carrera de Ingeniería de Sistemas.

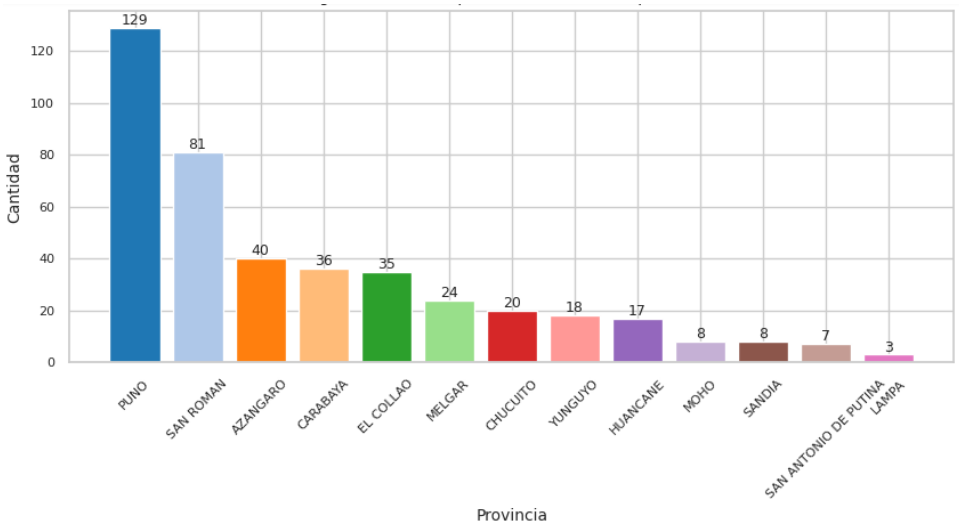
El resto de provincias cuenta con una menor cantidad de ingresantes.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 29

Cantidad de ingresantes por provincia

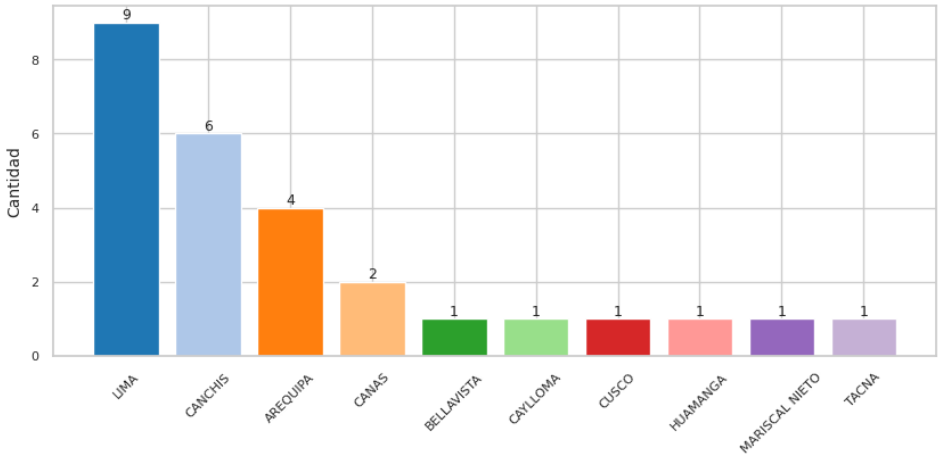


Nota: Gráfico de la cantidad de ingresantes por modalidad - General y CEPREUNA

En la siguiente tabla se puede ver la cantidad de ingresantes diferentes a las del departamento de Puno.

Figura 30

Ingresantes de otras provincias



Nota: Para el análisis de estudio no se consideró a estos ingresantes.

4.2.1.10. Análisis de post. e ingres. según tipo de colegio

Postulantes

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

En el siguiente gráfico se puede apreciar que en todas las provincias el número de postulantes provenientes de colegios públicos es significativamente mayor en comparación con los de colegios privados.

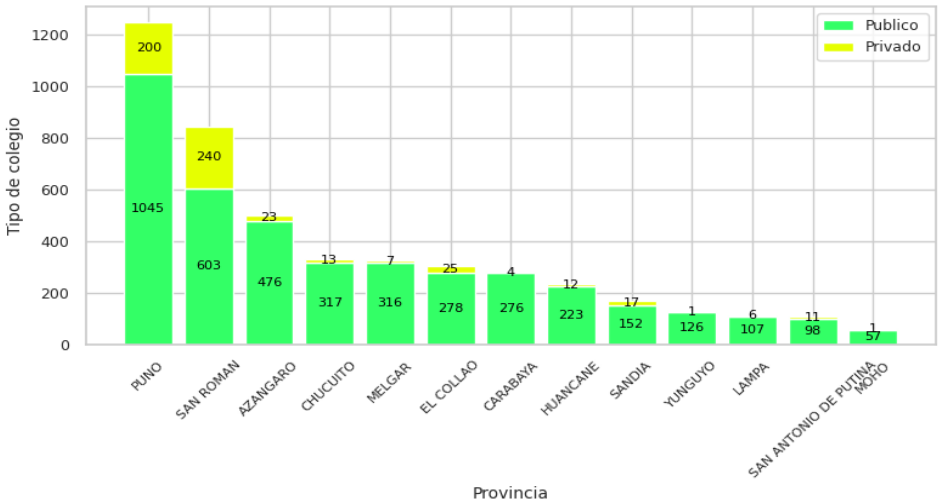
La provincia de **Puno** tiene la mayor cantidad de postulantes en ambos tipos de colegios (1,045 de colegios públicos y 200 de colegios privados). Prácticamente el 16% son de colegios privados respecto a la provincia.

Le sigue San Román, con 603 postulantes de colegios públicos y 240 de colegios privados y esto últimos representan el 28%, prácticamente tiene una presencia alta de colegios privados.

Varias provincias presentan una participación mínima de postulantes de colegios privados, como **San Antonio de Putina, Lampa y Sandia**, donde prácticamente todos los postulantes provienen de colegios públicos.

Figura 31

Postulantes según tipo de colegio



Nota: esta representación corresponde a postulantes de las dos modalidades de postulación General y CEPREUNA.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Ingresantes

Al igual que los postulantes en todas las provincias, la cantidad de ingresantes de colegios públicos es superior a la de colegios privados.

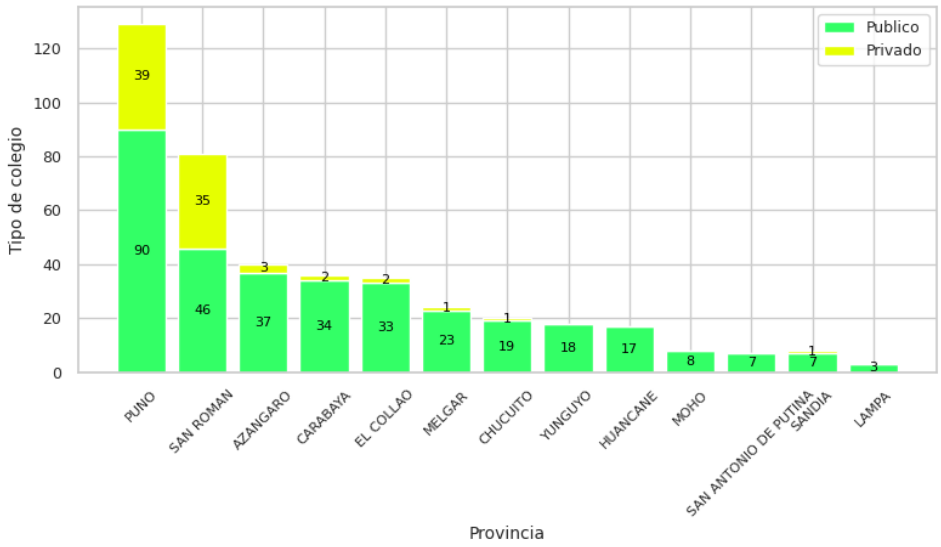
La provincia de Puno tiene el mayor número de ingresantes en ambas categorías (90 de colegios públicos y 39 de colegios privados).

San Román sigue en segundo lugar con 46 ingresantes de colegios públicos y 35 de colegios privados, siendo la provincia con la mayor proporción de ingresantes de colegios privados en comparación con otras.

Aunque en términos generales la tendencia de ingresantes sigue la distribución de los postulantes, que son lideradas por la provincia de Puno y San Román.

Figura 32

Ingresantes según tipo de colegio



Nota: esta representación corresponde a postulantes de las dos modalidades de postulación General y CEPREUNA.

4.2.1.11. Análisis de la distribución de ruralidad de ingresantes

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

La mayor cantidad de ingresantes proviene de **Puno (14 en total)**, con un predominio de Rural 3, lo que indica que la accesibilidad geográfica puede ser un factor importante en el acceso a la universidad.

Provincias como **Azángaro (9 ingresantes), Carabaya (7) y Chucuito (6)** también presentan una cantidad significativa de estudiantes que lograron ingresar.

Provincias como **Moho y San Antonio de Putina** presentan solo un ingresante cada una, con un nivel de ruralidad de 2.

Tendencias en el tipo de ruralidad de los ingresantes

Rural 3 representa la mayor proporción de ingresantes en la mayoría de provincias, lo que sugiere que los estudiantes con **mejor accesibilidad geográfica** tienen más probabilidades de ingresar.

En provincias como **Azángaro y Huancané**, **Rural 2** tiene una participación relevante, lo que podría indicar que los estudiantes con acceso intermedio también logran buenos resultados en el examen de admisión.

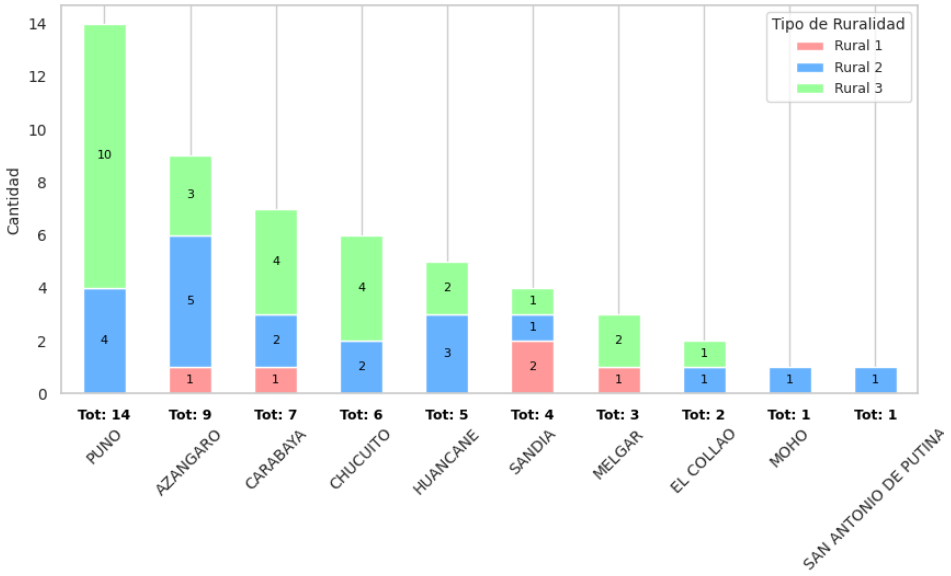
Los estudiantes de **Rural 1** (zonas más alejadas y con menor cantidad de habitantes) tienen una menor representación, lo que sugiere una posible barrera de acceso a la educación superior. Todos estos resultados se pueden apreciar en la siguiente figura.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 33

Tipos de ruralidad de ingresantes por provincia



Nota: La información vista en esta figura corresponde a ambas modalidades de estudio, General y CEPREUNA. Se muestra los 3 niveles de ruralidad cuyas gradientes se encuentran en la Norma Técnica, (2023) MINEDU.

En la siguiente tabla se puede apreciar la cantidad de ingresantes por niveles de ruralidad.

Tabla 23

Cantidad de ingresantes por ruralidad

Nº	Ruralidad	Cantidad
1	Rural 3	27
2	Rural 2	20
3	Rural 1	5

Nota: Resultado de ambas modalidades de ingreso, General y CEPREUNA.

4.2.1.12. Análisis de correlación entre edad y años de egreso

Al calcular la matriz de correlación, se ha encontrado un **coeficiente de correlación de 0.96** entre la edad del ingresante y la cantidad de años transcurridos desde que egresó de la secundaria hasta obtener una vacante en la universidad. Este valor indica una **correlación**

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

positiva muy fuerte, lo que indica que a medida que se incrementa la edad de un ingresante, las posibilidades de ingresar se incrementan, esto porque tiene una mejor preparación.

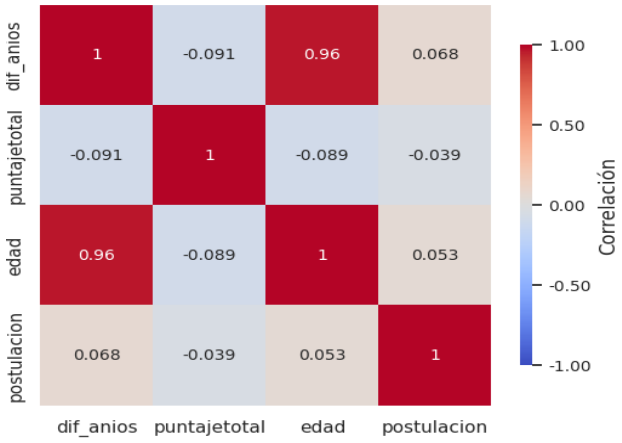
Diferencias con otras variables

El puntaje obtenido no mostró una correlación significativa con la edad ni con los años transcurridos desde su egreso del colegio. Esto sugiere que la edad del postulante no es un factor determinante en su desempeño en el examen de admisión.

La cantidad de postulaciones tampoco tuvo una correlación fuerte con la edad ni con los años transcurridos desde el colegio, lo que indica que el número de intentos no está directamente relacionado con la edad del ingresante. Esto podría implicar que algunos postulantes ingresan en su primer intento mientras que otros necesitan varios intentos, sin un patrón claro respecto a la edad. Esto se puede apreciar en el siguiente cuadro de correlación.

Tabla 24

Matriz de correlación de Pearson



Nota: Sólo se consideró las variables numéricas y datos de los ingresantes de ambas modalidades, General y CEPREUNA.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

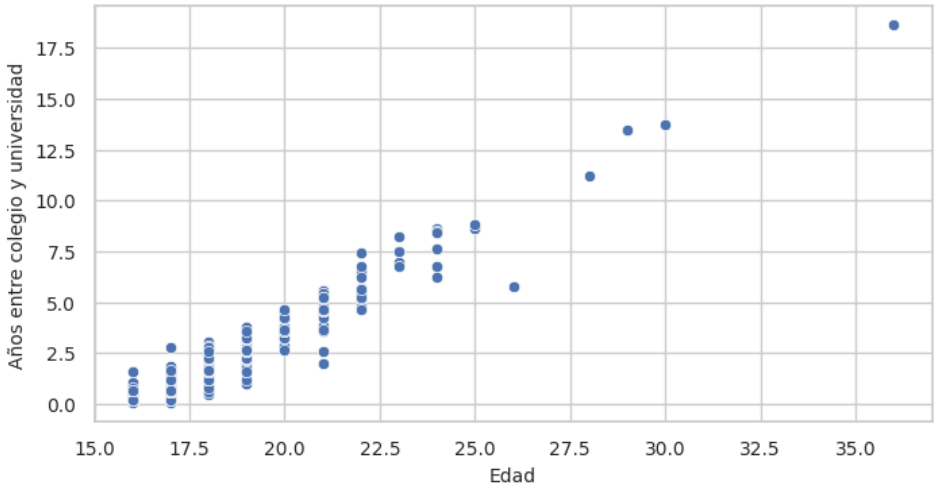
Líneas de código para aplicar la matriz correlación de Pearson:

```
from sklearn.preprocessing import LabelEncoder
xy = data.loc[( data['ingreso'] == 1 ) & ( data['dif_anios']>0 )]
data2 = xy.copy()
label_encoder = LabelEncoder()
data2['ruralidad'] = label_encoder.fit_transform(data2['tipo_ruralidad'])
data2.drop(columns=['tipo_ruralidad'], inplace=True)
tiempo = data2[['dif_anios','puntajetotal','edad','postulacion']]
tiempocor = tiempo.corr(method='pearson')
plt.figure(figsize=(6, 5))
sns.heatmap(tiempocor, annot=True, cmap='coolwarm', vmin=-1, vmax=1,
annot_kws={"size": 10},
        cbar_kws={'label': 'Correlación', 'shrink': 0.8, 'aspect': 30, 'format': '%.2f',
'ticks': [-1, -0.5, 0, 0.5, 1]})
plt.title('Matriz de Correlación de Pearson')
plt.show()
```

A continuación, se muestra la relación lineal entre la variable "dif_anios" (años que pasaron desde que egresó del colegio hasta ingresar a la universidad) y "edad".

Figura 34

Correlación entre edad y años de egreso



Nota: La relación corresponde a los años que pasaron desde que egreso de colegio hasta el momento de ingresar a la universidad.

4.2.1.13. Codificación de variables categóricas

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Para codificar las variables categóricas del conjunto de datos, se usó OneHotEncoder (para variables nominales sin orden) y LabelEncoder (para variable categórica).

Dado que las variables que corresponden a departamento, provincia, distrito del colegio y nacimiento (del postulante), modalidad de postulación, tipo de colegio y género, se usó “One-Hot Encoding” para la mayoría y Label Encoding para la variable “tipo_ruralidad”.

Luego del análisis exploratorio y la ingeniería de características, se eliminaron variables que no aportaban información relevante al modelo o que fueron transformadas en nuevas características. A continuación, se muestra el código utilizado para esta limpieza de datos:

```
# Eliminando variables irrelevantes o redundantes después de la ingeniería de características
columnas_a_eliminar = ['orden', 'fecha', 'dni', 'fec_nac', 'grado', 'anno_egreso', 'dep_nac', 'prov_nac', 'dist_nac', 'nombre_col', 'depart_col', 'dist_col']
datos.drop(columns=columnas_a_eliminar, inplace=True)
# Verificar la estructura del DataFrame después de la eliminación
datos.info()
```

Razones de exclusión de variables

A continuación, se detallan las razones por las que cada una de estas variables fue eliminada:

- **orden**: Campo generado automáticamente que no influye en la predicción del ingreso.
- **fecha**: La fecha del proceso de admisión de ambas modalidad de postulación no aporta información relevante al modelo.
- **dni**: Identificador único del postulante, irrelevante para el análisis.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- **fec_nac**: Se reemplazó por la variable "edad", que es más útil para el modelo.
- **grado**: Dado que todos son postulantes regulares y con estudios secundarios concluidos se identificó como una variable no influyente en la predicción.
- **anno_egreso**: Se reemplazó por la variable "dif_anios", que corresponde a el tiempo que transcurrió desde que egreso del colegio hasta el momento de postulación/ingreso a la universidad, que es más informativa.
- **dep_nac, prov_nac, dist_nac**: Aunque el lugar de nacimiento podría tener relevancia, se decidió priorizar la ubicación del colegio como un mejor indicador.
- **nombre_col**: Variable de texto difícil de estandarizar y de poca utilidad directa en el modelo.
- **depart_col**: Similar a la variable “grado”, sólo se estudia al departamento de Puno.
- **dist_col:** Esta variable se eliminó debido a problemas de multicolinealidad, ya que al calcular los valores de Variance Inflation Factor (VIF), esta variable generaba una división entre cero. Además, la cantidad de distritos era muy elevada, lo que dificultaba su uso en el modelo sin una adecuada transformación.

4.2.1.14. Codificación de variables categóricas - One-Hot Encodig

Para que un modelo de aprendizaje automático pueda procesar variables categóricas, es necesario transformarlas en un formato numérico adecuado. En esta etapa, se aplicó **One-Hot Encoding** a las variables

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

categóricas, asegurando que el modelo pueda interpretarlas sin introducir relaciones numéricas artificiales entre sus valores.

A continuación, se presenta el código utilizado para aplicar esta transformación:

```
from sklearn.preprocessing import OneHotEncoder
# Aplicar One-Hot Encoding
# Elimina una categoría para evitar colinealidad
encoder = OneHotEncoder(drop='first', sparse_output=False)
encoded_cols = encoder.fit_transform(datos[df_categoricas.columns])
# Crear un DataFrame con las columnas codificadas
encoded_df = pd.DataFrame(encoded_cols,
columns = encoder.get_feature_names_out (df_categoricas.columns))
datos.drop(columns=df_categoricas.columns)
# Concatenar con el dataset original eliminando las variables categóricas originales
df_encoded = pd.concat([datos.drop(columns=df_categoricas.columns),
encoded_df], axis=1)
# Ver el nuevo DataFrame
df_encoded
```

4.2.1.15. Análisis de multicolinealidad para la predicción

Los resultados del Factor de Inflación de la Varianza (VIF) revelan un problema serio de multicolinealidad para el modelo. Varios valores de VIF son extremadamente altos, lo que indica que algunas de las variables predictoras están fuertemente correlacionadas entre sí.

Interpretación de los resultados VIF:

Valores extremadamente altos (VIF > 100): edad (395.91) y prov_col_PUNO (92.51) muestran una multicolinealidad severa. Esto significa que estas variables son casi combinaciones lineales de otras variables en el modelo. La inclusión de estas variables puede desestabilizar completamente los coeficientes de regresión.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Valores muy altos (VIF > 20): prov_col_SAN_ROMAN(80.01), tipo_ruralidad (23.15), prov_col_AZANGARO (32.61), prov_col_CARABAYA (20.25), prov_col_CHUCUITO (22.48), prov_col_EL COLLAO (22.86), prov_col_MELGAR (21.10) también presentan un nivel de multicolinealidad muy alto y preocupante.

Valores altos (VIF > 10): dif_anios (10.44), prov_col_HUANCANE (16.71), prov_col_SANDIA (9.26), prov_col_YUNGUYO (10.26) muestran multicolinealidad considerable.

Valores moderados (VIF > 5): modalidad (9.78), tipo_colegio (14.55), puntajetotal (5.06), genero_M (8.65), prov_col_SAN ANTONIO DE PUTINA (6.12), prov_col_LAMPA (5.5) muestran multicolinealidad que podría ser problemática.

Valores bajos (VIF < 5): El resto de las provincias tienen VIFs relativamente bajos, aunque algunos están cercanos al límite de 5.

Tabla 25

Valores de VIF de forma descendente

Variable	VIF
Edad	395.91
prov_col_PUNO	92.51
prov_col_SAN ROMAN	80.01
prov_col_AZANGARO	32.61
tipo_ruralidad	23.15
prov_col_EL COLLAO	22.86
prov_col_CHUCUITO	22.48
prov_col_MELGAR	21.10
prov_col_CARABAYA	20.25
prov_col_HUANCANE	16.71
tipo_colegio	14.55
dif_anios	10.44
prov_col_YUNGUYO	10.26

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Variable	VIF
modalidad	9.82
prov_col_SANDIA	9.26
genero_M	8.65
prov_col_SAN ANTONIO DE PUTINA	6.12
prov_col_LAMPA	5.50
puntajetotal	5.06
prov_col_MOHO	3.59
prov_col_AREQUIPA	3.59
prov_col_CANCHIS	3.39
postulacion	3.33
prov_col_TACNA	2.01
prov_col_MARISCAL NIETO	1.98
prov_col_LIMA	1.82
prov_col_CUSCO	1.75
prov_col_CAMANA	1.73
prov_col_CHUMBIVILCAS	1.53
prov_col_ASCOPE	1.43
prov_col_LA CONVENCION	1.34
prov_col_ESPINAR	1.32
prov_col_CALLAO	1.32
prov_col_SATIPO	1.29
prov_col_GENERAL SANCHEZ CERRO	1.25
prov_col_LUCANAS	1.23
prov_col_CORONGO	1.21
prov_col_ANDAHUAYLAS	1.20
prov_col_CAYLLOMA	1.11
prov_col_CANDARAVE	1.11
prov_col_ILO	1.11
prov_col_CAÑETE	1.10
prov_col_CARAVELI	1.10
prov_col_ISLAY	1.10
prov_col_SANTA CRUZ	1.10
prov_col_CASTILLA	1.10
prov_col_PARURO	1.10
prov_col_CANAS	1.09
prov_col_URUBAMBA	1.09

Nota: Estos valores ya se encuentran transformados a valores numéricos

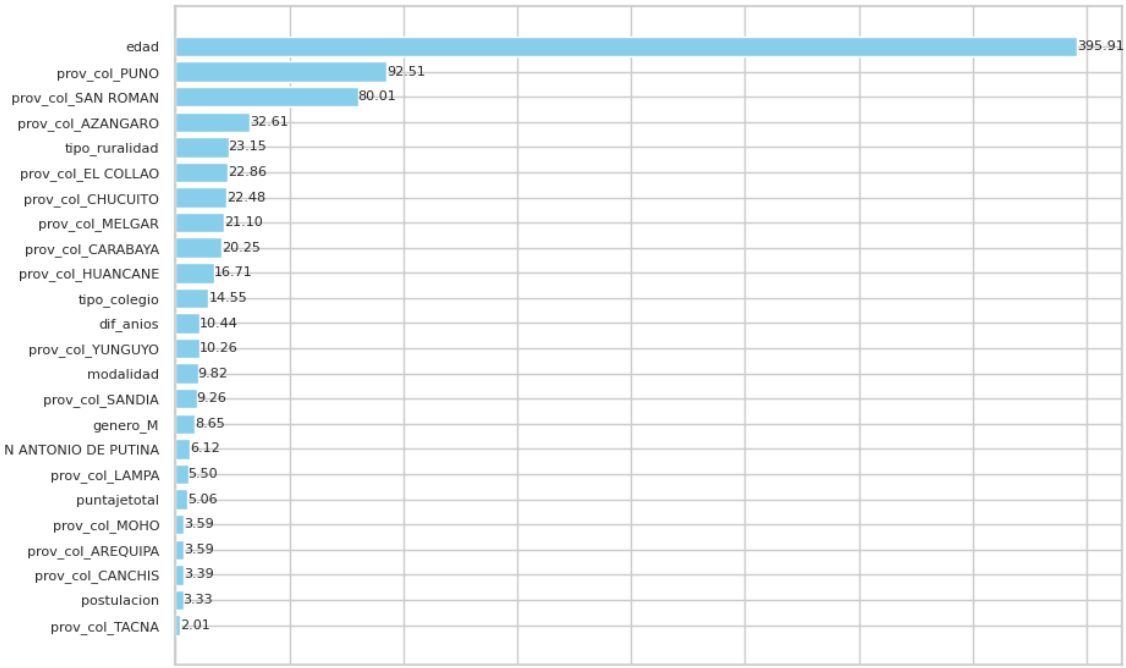
Gráficamente se puede apreciar en la siguiente figura.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 35

Valores de VIF de cada una de las variables de estudio



Nota: La variable 'edad' genera alta multicolinealidad con respecto a las otras variables.

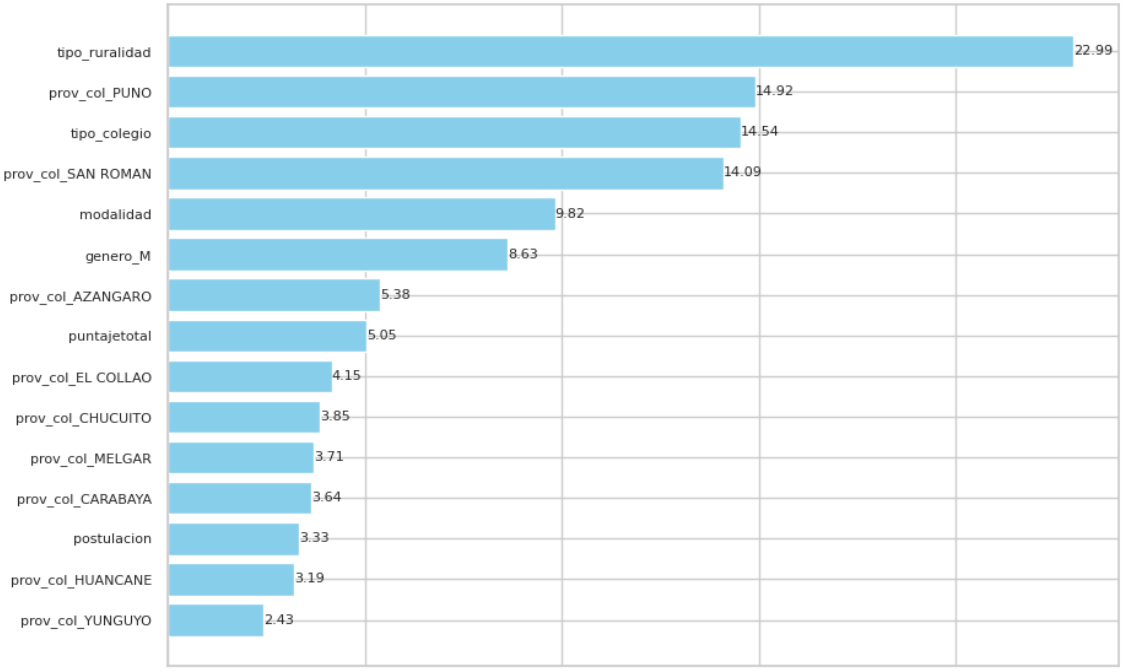
Para reducir la multicolinealidad, es recomendable eliminar una de las variables correlacionadas. En la tabla 23 el coeficiente de correlación entre 'edad' y 'df_anios' (r = 0.96), lo que indica una fuerte dependencia lineal entre ambas. Con base en este resultado y la figura 35, se decidió eliminar la variable 'edad'.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 36

Valores de VIF excluido la variable edad



Nota: La figura presenta las variables de estudio después de excluir la variable *'edad'*, dejando el conjunto de datos listo y optimizado.

4.2.1.16. Normalización de datos

En el dataset utilizado para este estudio, se identificaron variables como **"postulacion"**, **"puntajetotal"** y **"dif_anios"**, las cuales presentan una variabilidad significativa en sus magnitudes. Esta diferencia en las escalas puede afectar negativamente el rendimiento de los algoritmos de machine learning, especialmente aquellos que son sensibles a la magnitud de los datos. Para garantizar que todas las variables contribuyan de manera equitativa al modelo y mejorar su eficacia, se aplicó **StandardScaler**, una técnica de normalización que transforma los datos para que tengan una **media de 0** y una **desviación estándar de 1**. Este proceso asegura que las variables se encuentren en la misma escala, lo que facilita la convergencia de los algoritmos empleados y mejora la precisión de los resultados.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Líneas de código para aplicar la normalización de datos:

```
from sklearn.preprocessing import StandardScaler
# Definir las columnas numéricas a escalar
# numeric_cols = ["postulacion", "puntajetotal", "dif_anios", "edad"]
numeric_cols = ["postulacion", "puntajetotal", "dif_anios"]
# aplicar StandrdScaler
scaler = StandardScaler()
df_scaler = df_encoded.copy()
df_scaler[numeric_cols] = scaler.fit_transform(df_scaler[numeric_cols])
# ver el dataframe escalado
df_scaler.head()
```

Tabla 26

Conjunto de datos normalizado

Mod	Tipo_col	Tipo_rur	puntaje	postulacion	Dif_anios
1	1	2	-0.433637	-0.816163	1.671849
1	1	2	-1.915812	1.691213	-0.686791
1	1	3	0.155149	-0.189319	-0.686791

Nota: Captura de pantalla de variables que se usó para el modelamiento con Machine Learning.

4.3. OBJETIVO ESPECÍFICO 3

Desarrollar, entrenar y evaluar modelos de aprendizaje automático utilizando métricas de rendimiento.

Con el conjunto de datos preparado, se realizó una división de los datos en conjuntos de entrenamiento y prueba. Para el entrenamiento, se utilizaron los datos de los procesos de admisión de los años 2018 a 2021, mientras que los datos del año 2022 se reservaron para la fase de prueba, específicamente los datos de examen ordinario (General) cuyo puntaje es mayor a 800, con el fin de evaluar el rendimiento del modelo en condiciones reales y no vistas durante el entrenamiento. La división de los datos se realizó en una proporción de 80% para entrenamiento y 20% para prueba.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Código para selección de la fase de prueba

```
X_train_datos = data.loc[data['fecha'].dt.year != 2022].copy()
X_train_datos = X_train_datos[X_train_datos['puntajetotal']>=800]
X_train_datos.reset_index(drop=True, inplace=True)
```

Para el modelo de clasificación (ingresó y no ingresó), se implementaron y compararon varios algoritmos de aprendizaje supervisado, entre ellos: regresión logística, árboles de decisión, random forest, gradient boosting y XGBoost. El objetivo fue evaluar y comparar su rendimiento utilizando métricas clave como precisión, recall, F1-score y exactitud, con el fin de identificar el modelo más adecuado para predecir los resultados de los procesos de admisión.

4.3.1. Regresión logística

En la prueba realizada con los datos de admisión del año 2022, los resultados obtenidos con la regresión logística muestran una precisión global (accuracy) del **78%**. Se observa que el modelo tiene una alta precisión en la clase mayoritaria (no ingresó) con un **89%**, pero un bajo desempeño en la clase minoritaria (ingresó) con una precisión del **29%**. El **recall** de esta última clase es del **39%**, lo que indica que el modelo identifica correctamente a algunos ingresantes, pero no con la suficiente precisión.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión del año 2022.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 27

Métricas obtenidas con el modelo de regresión logística

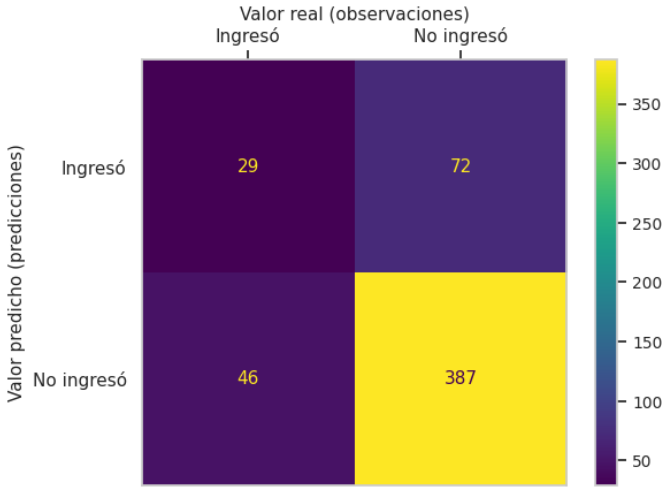
	Precision	Recall	F1-score	support
0	0.89	0.84	0.87	459
1	0.29	0.39	0.33	75
Accuracy			0.78	534

Nota: Datos obtenidos con código Python

Matriz de confusión del modelo de regresión logística.

Figura 37

Matriz de confusión del modelo de regresión logística



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

Código del modelo de regresión logística empleado

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
# Separar variables predictoras y variable objetivo (ingreso: 1 o 0)
# Variables predictoras
X = df_scaler_train
# División en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y_trainn, test_size=0.2, random_state=42, stratify = y_trainn)
# Aplicar el modelo de regresión logística
modelo = LogisticRegression(class_weight='balanced')
modelo.fit(X_train, y_train)
# Predicciones
y_pred = modelo.predict(X_test)
matriz_confusion(y_test, y_pred, 'No ingresó', 'Ingresó', 11)
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

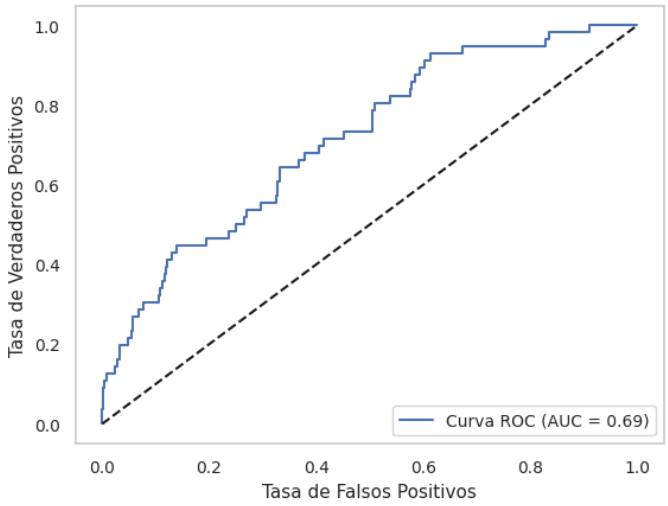
```
# Evaluación del modelo
print(classification_report(y_test, y_pred))
print(f'Precisión del modelo: {accuracy_score(y_test, y_pred)}')
# ---------- Probando con datos de 2022 ----------
y_pred_lr_2022 = modelo.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_lr_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_lr_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_lr_2022)}')
```

Cálculo del AUC-ROC

Se encontró un AUC-ROC de 0.69 lo que significa que, el modelo tiene un 69% de probabilidad de clasificar correctamente una instancia positiva frente a una negativa. En otras palabras, el modelo es capaz de distinguir entre clases positivas y negativas en el 69% de los casos, es decir, tiene un buen rendimiento.

Figura 38

Visualización de la curva ROC de la regresión logística



Nota: Estos datos fueron calculados con los datos de los procesos de admisión del año 2022.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.3.2. Árbol de decisión

Los resultados obtenidos con el modelo de árbol de decisión muestran una precisión global (accuracy) del **84%**. Se observa un buen desempeño en la clase mayoritaria (**no ingresó**), con una precisión del **86%**, mientras que para la clase minoritaria (**ingresó**) la precisión es baja, alcanzando solo un **18%**, con un recall del **4%**. Esto indica que el modelo tiene dificultades para identificar correctamente a los postulantes que lograron ingresar. Estos resultados sugieren que, su capacidad de generalización puede ser limitada en contextos con datos desbalanceados.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión de la fase de prueba del año 2022 (examen General).

Tabla 28

Métricas obtenidas con el modelo de árbol de decisión

	Precision	**Recall**	**F1-score**	**support**
0	0.86	0.97	0.91	459
1	0.18	0.04	0.07	75
Accuracy			0.84	534

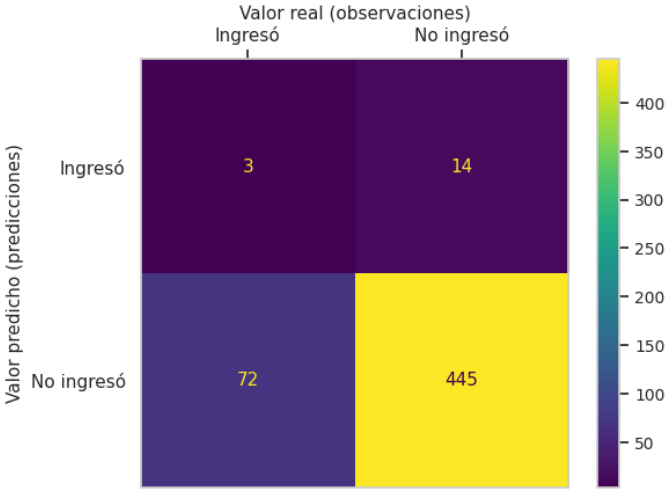
Nota: Datos obtenidos con código Python

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 39

Matriz de confusión del modelo de árbol de decisión



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

Código del modelo de árbol de decisión empleado

```
rom sklearn.tree import DecisionTreeClassifier
# 'X' son las características y 'y_trainn' la variable objetivo
X_train, X_test, y_train, y_test = train_test_split(X, y_trainn, test_size=0.2, random_state=42, stratify=y_trainn)
# Crear el modelo de Árbol de Decisión y ajustando los hiperparametros
modelo_ar = DecisionTreeClassifier(
   random_state=42,
   # Profundidad máxima del árbol
   max_depth=17
   )
# Entrenar el modelo
modelo_ar.fit(X_train, y_train)
# Hacer predicciones
y_pred = modelo_ar.predict(X_test)
# Generar la matriz de confusión
matriz_confusion(y_test, y_pred, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_test, y_pred))
# ---------- Probando con datos de 2022 ----------
y_pred_ad_2022 = modelo_ar.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_ad_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_ad_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_ad_2022)}')
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.3.3. Random Forest

En la evaluación realizada sobre los datos de admisión del 2022, el modelo Random Forest obtuvo los siguientes resultados, para la clase mayoritaria (**no ingresó**), la precisión alcanzó un **87%**, con un **recall del 94%**, lo que indica que el modelo identifica correctamente a la mayoría de los postulantes que no lograron ingresar. Para la clase minoritaria (**ingresó**), la precisión fue de **28%**, con un recall del **15%**, lo que muestra que el modelo tiene dificultades para reconocer a los postulantes admitidos. La métrica **F1-score** refleja este desequilibrio en el desempeño, con un **90%** para la clase mayoritaria y un **19%** para la clase minoritaria.

Estos resultados indican que, aunque Random Forest mejora la capacidad de generalización respecto a un solo árbol de decisión, sigue presentando limitaciones en la identificación de la clase minoritaria debido al desbalance de datos.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión de la fase de prueba del año 2022 (examen General).

Tabla 29

Métricas obtenidas con el modelo de Random Forest

	Precision	**Recall**	**F1-score**	**support**
0	0.87	0.94	0.90	459
1	0.28	0.15	0.19	75
Accuracy			0.83	534

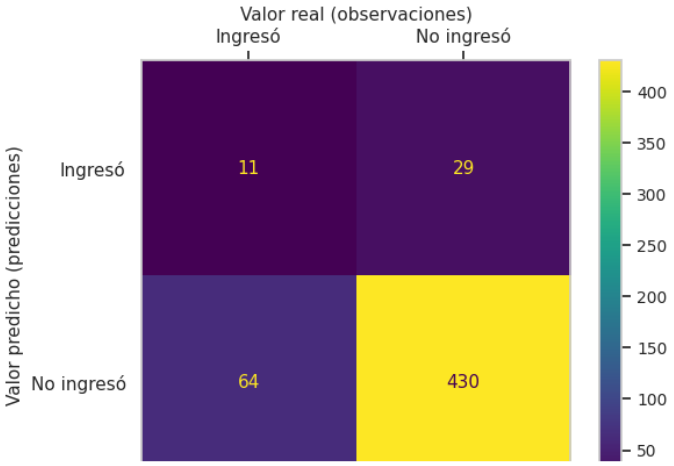
Nota: Datos obtenidos con código Python

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 40

Matriz de confusión del modelo de Random Forest



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

Código del modelo de Random Forest empleado

```
from sklearn.ensemble import RandomForestClassifier
from imblearn.over_sampling import SMOTE
# Aplicar SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y_trainn)
# Definir el modelo con hiperparámetros básicos
rf_model = RandomForestClassifier(n_estimators=1000, random_state=42)
# División en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42, stratify=y_resampled)
# Entrenar el modelo
rf_model.fit(X_train, y_train)
# Predecir con datos de prueba
y_pred_rf = rf_model.predict(X_test)
# Generar la matriz de confusión
matriz_confusion(y_test, y_pred_rf, 'No ingresó', 'Ingresó', 11)
# Mostrar métricas del modelo
print(classification_report(y_test, y_pred_rf))
# ------------- Probando con datos de 2022 ----------
y_pred_lr_2022 = rf_model.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_lr_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_lr_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_lr_2022)}')
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.3.4. XGBoost

El modelo **XGBoost** mostró una **precisión** del **86%** y un **recall** del **99%** para la clase mayoritaria (**no ingresó**), lo que indica que clasifica correctamente a la mayoría de los postulantes no admitidos. Sin embargo, para la clase minoritaria (i**ngresó**), el modelo no logró identificar correctamente ningún caso, obteniendo un **recall de 0.00**, lo que refleja una limitación en la detección de postulantes ingresantes. Esto sugiere que XGBoost puede estar sesgado hacia la clase mayoritaria debido al desbalance de datos.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión de la fase de prueba del año 2022 (examen General).

Tabla 30

Métricas obtenidas con el modelo XGBoost

	Precision	**Recall**	**F1-score**	**support**
0	0.86	0.99	0.92	459
1	0.00	0.00	0.00	75
Accuracy			0.85	534

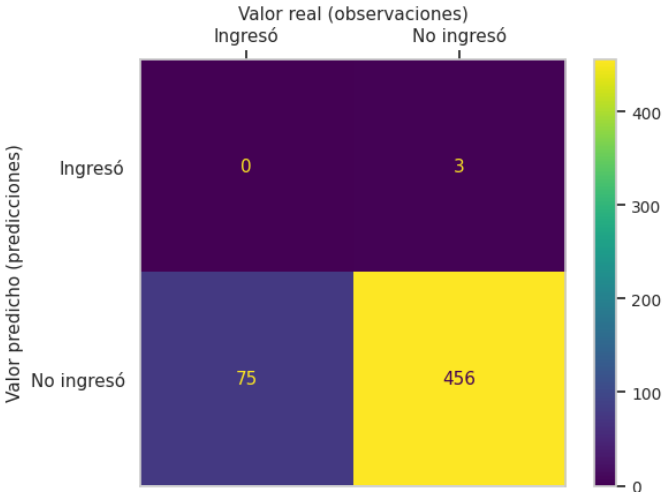
Nota: Datos obtenidos con código Python

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 41

Matriz de confusión del modelo de XGBoost



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

Código del modelo XGBoost

```
from xgboost import XGBClassifier
# Crear y entrenar el modelo XGBoost
xgb_clf = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
xgb_clf.fit(X_train, y_train)
X_train, X_test, y_train, y_test = train_test_split(X, y_trainn, test_size=0.2, random_state=42, stratify=y_trainn)
# Predicciones
y_pred_xgb = xgb_clf.predict(X_test)
# Generar la matriz de confusión
matriz_confusion(y_test, y_pred_xgb, 'No ingresó', 'Ingresó', 11)
# Evaluación del modelo
print(classification_report(y_test, y_pred_xgb))
# ---------- Probando con datos de 2022 ----------
y_pred_xgb_2022 = xgb_clf.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_xgb_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_xgb_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_xgb_2022)}')
```

4.3.5. K-Nearest Neighbors

En la evaluación realizada, KNN mostró un **accuracy** del **85%**, con una **precisión** del **86%** y un **recall** del **98%** para la clase mayoritaria (**no ingresó**), lo que indica que predice de manera efectiva a los postulantes no admitidos. Sin

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

embargo, para la clase minoritaria (**ingresó**), el rendimiento fue significativamente bajo, con una **precisión** del **20%** y un **recall** del **3%**, lo que evidencia una fuerte dificultad para identificar correctamente a los ingresantes. Esto sugiere que KNN puede no ser el modelo más adecuado en escenarios con clases desbalanceadas, ya que tiende a favorecer la clase mayoritaria.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión de la fase de prueba del año 2022 (examen General).

Tabla 31

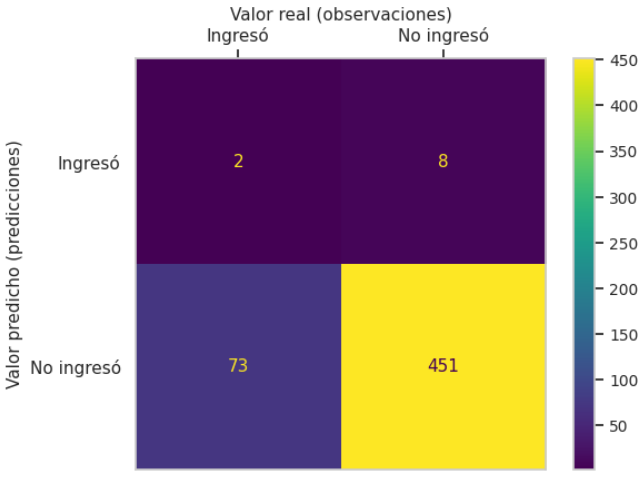
Métricas obtenidas con el modelo KNN

	Precision	**Recall**	**F1-score**	**support**
0	0.86	0.98	0.92	459
1	0.20	0.03	0.05	75
Accuracy			0.85	534

Nota: Datos obtenidos con código Python

Figura 42

Matriz de confusión del modelo de KNN



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Código del modelo K-NN

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV

# Crear un pipeline con escalado y modelo KNN
pipeline = Pipeline([
    ('scaler', StandardScaler()),  # Normaliza los datos
    ('knn', KNeighborsClassifier())  # Modelo KNN
])
# Definir hiperparámetros a buscar
param_grid = {
    'knn__n_neighbors': [3, 5, 7, 9, 11],  # Número de vecinos
    'knn__weights': ['uniform', 'distance'],  # Peso de los vecinos
    'knn__metric': ['euclidean', 'manhattan']  # Métrica de distancia
}
X_train, X_test, y_train, y_test = train_test_split(X, y_trainn, test_size=0.2, random_state=42, stratify=y_trainn)
# Optimización con GridSearchCV
grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train, y_train)
# Mejor modelo encontrado
best_knn = grid_search.best_estimator_
# Predicciones
y_pred = best_knn.predict(X_test)
# Generar la matriz de confusión
matriz_confusion(y_test, y_pred, 'No ingresó', 'Ingresó', 11)
# Evaluación del modelo
print("Mejores parámetros encontrados:", grid_search.best_params_)
print("\nReporte de clasificación:\n", classification_report(y_test, y_pred))
# ---------- probando con datos de 2022 ----------
y_pred_knn_2022 = best_knn.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_knn_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_knn_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_knn_2022)}')
```

4.3.6. Support Vector Machines

En la evaluación realizada, SVM obtuvo un **accuracy** del **92%**, indicando un buen desempeño general. Para la clase mayoritaria (**no ingresó**), el modelo alcanzó una **precisión** del **93%** y un **recall** del **99%**, lo que demuestra una alta capacidad para identificar correctamente a los postulantes no admitidos. Sin embargo, para la clase minoritaria (**ingresó**), el rendimiento fue inferior, con una **precisión** del **86%** y un **recall** del **43%**, lo que implica que casi la mitad de los

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ingresantes fueron clasificados incorrectamente. A pesar de que SVM logra un balance aceptable entre precisión y recall, su desempeño en la detección de la clase minoritaria aún es limitado.

A continuación, se muestra en la siguiente tabla el resultado de las métricas obtenidas al probar el modelo entrenado con los datos de los procesos de admisión de la fase de prueba del año 2022 (examen General).

Tabla 32

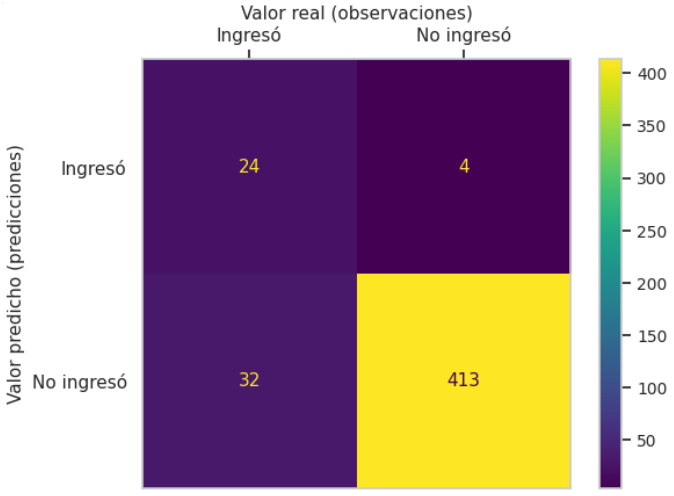
Métricas obtenidas con el modelo SVM

	Precision	**Recall**	**F1-score**	**support**
0	0.93	0.99	0.96	417
1	0.86	0.43	0.57	56
Accuracy			0.92	473

Nota: Datos obtenidos con código Python

Figura 43

Matriz de confusión del modelo de SVM



Nota: Este resultado se realizó con los datos de la fase de prueba (General 2022)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Código del modelo SVM

```
from sklearn.svm import SVC
# Modelo SVM con kernel lineal
svm_linear = SVC(kernel='linear', C=1.0, gamma='scale', class_weight='balanced')
svm_linear.fit(X_train, y_train)
y_pred_linear = svm_linear.predict(X_test)
# Generar la matriz de confusión
matriz_confusion(y_test, y_pred_linear, 'No ingresó', 'Ingresó', 11)
print("Resultados con SVM (Kernel Lineal):")
# ---------- Probando con datos de 2022 ----------
y_pred_svm_2022 = svm_linear.predict(df_scaler_test)
# Generar la matriz de confusión
mc = matriz_confusion(y_testt, y_pred_svm_2022, 'No ingresó', 'Ingresó', 11)
print(classification_report(y_testt, y_pred_svm_2022))
print(f'Precisión del modelo: {accuracy_score(y_testt, y_pred_svm_2022)}')
auc_roc, auc_roc_2022, x = calcular_roc(svm_linear, X_test, y_test, df_scaler_test,
y_testt, 11)
print(f'AUC-ROC para los datos de prueba: {auc_roc}')
print(f'AUC-ROC para los datos de 2022: {auc_roc_2022}')
```

4.4. OBJETIVO ESPECÍFICO 4

Validar los modelos con los datos del año 2022 y proporcionar recomendaciones basadas en los resultados obtenidos.

Para evaluar el rendimiento de los modelos de aprendizaje automático empleados, se validaron utilizando datos del proceso de admisión del año 2022. Se aplicaron los modelos previamente entrenados sobre este nuevo conjunto de datos y se analizaron sus métricas de desempeño, incluyendo precisión, recall, F1-score, accuracy y AUC. A continuación, se presenta un resumen en la siguiente tabla y el análisis detallado del rendimiento de cada modelo.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 33

Resumen de métricas de modelos empleados

Métricas	Logistic Regression	Decision Tree	Random Forest	XGboots	K-Nearest Neighbors	Support Vector Machines
Accuracy	0.78	0.84	0.83	0.85	0.85	0.92
Precision	0.29	0.18	0.28	0.00	0.20	0.86
Recall	0.39	0.04	0.15	0.00	0.03	0.43
F1-score	0.33	0.07	0.19	0.00	0.05	0.57
AUC	0.72	0.51	0.47	0.41	0.53	0.69

Nota: Estos valores fueron obtenidas con los modelos indicados donde se realizaron varias pruebas para validar estos resultados.

Precisión global (accuracy)

El modelo **SVM** obtuvo la mayor precisión global con un **92%**, seguido de **XGBoost** y **KNN** con **85%**. **Decision Tree** y **Random Forest** alcanzaron una precisión de **84% y 83%**, respectivamente. **Regresión Logística** obtuvo el menor **accuracy (78%)**, lo que indica un menor desempeño global.

Detección de la clase minoritaria (precision, recall y F1-score)

Precision: SVM destacó con un **86%**, mientras que los demás modelos lograron valores más bajos, siendo **Regresión Logística** y **Random Forest** los que superaron el **20%**. XGBoost tuvo un rendimiento extremadamente bajo en este aspecto (**0.00**).

Recall: SVM también sobresalió con un **43%**, lo que indica que logra capturar una mayor proporción de los ingresantes en comparación con los otros modelos. **Regresión Logística (39%)** también presentó un desempeño decente en este aspecto.

F1-score: SVM nuevamente mostró el mejor balance entre **precisión** y **recall** con **0.57**, seguido por Regresión Logística (**0.33**). El resto de los modelos obtuvo valores de **F1-score** muy bajos, lo que indica problemas en la detección de la clase minoritaria.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Capacidad de Discriminación (ROC AUC score)

El **AUC** mide la capacidad del modelo para diferenciar entre las clases.

Regresión Logística obtuvo el mayor **AUC con 0.72**, lo que indica una buena capacidad de discriminación.

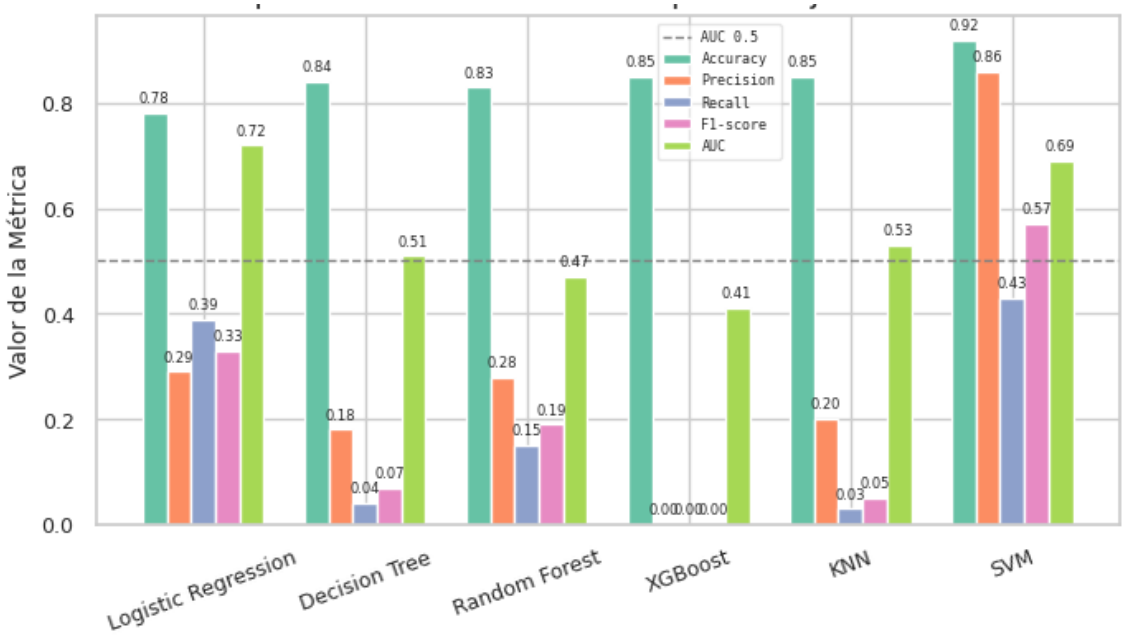
SVM alcanzó **0.69**, lo que también lo posiciona como un modelo con un desempeño competitivo.

KNN y Decision Tree obtuvieron valores cercanos a **0.50**, lo que indica un desempeño cercano a un clasificador aleatorio.

XGBoost mostró el peor AUC (0.41), reflejando su incapacidad de diferenciar correctamente entre ingresantes y no ingresantes.

Figura 44

Gráfico de barras de modelos empleados



Nota: Estos valores fueron obtenidas con los modelos indicados donde se realizaron varias pruebas para validar estos resultados.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Código empleado para graficar los modelos empleados

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Datos de las métricas por modelo
modelos = ["Logistic Regression", "Decision Tree", "Random Forest", "XGBoost", "KNN", "SVM"]
metricas = ["Accuracy", "Precision", "Recall", "F1-score", "AUC"]
datos = np.array([
    [0.78, 0.29, 0.39, 0.33, 0.72],
    [0.84, 0.18, 0.04, 0.07, 0.51],
    [0.83, 0.28, 0.15, 0.19, 0.47],
    [0.85, 0.00, 0.00, 0.00, 0.41],
    [0.85, 0.20, 0.03, 0.05, 0.53],
    [0.92, 0.86, 0.43, 0.57, 0.69]
])
# Crear el gráfico de barras
sns.set(style="whitegrid")
fig, ax = plt.subplots(figsize=(8, 4))
x = np.arange(len(modelos))
width = 0.15  # Ancho de las barras
colors = sns.color_palette("Set2", len(metricas))
# Graficar cada métrica
for i, (metrica, color) in enumerate(zip(metricas, colors)):
    bars = ax.bar(x + i * width, datos[:, i], width, label=metrica, color=color)
    # Agregar los datos en la parte superior de cada barra
    for bar in bars:
        height = bar.get_height()
        ax.annotate(f'{height:.2f}',
                    xy=(bar.get_x() + bar.get_width() / 2, height),
                    xytext=(0, 3),  # Desplazamiento del texto
                    textcoords="offset points",
                    ha='center', va='bottom', fontsize=6)
# Línea horizontal en AUC = 0.5
ax.axhline(y=0.5, color='gray', linestyle='--', linewidth=1.0, label='AUC 0.5')
# Configurar etiquetas y leyenda
ax.set_xlabel("Modelos", fontsize=10)
ax.set_ylabel("Valor de la Métrica", fontsize=10)
ax.set_title("Comparación de Modelos de Aprendizaje Automático", fontsize=14)
ax.set_xticks(x + width * 2)
ax.set_xticklabels(modelos, rotation=20)
ax.legend(
     loc='upper right',            # Posición de la leyenda
     frameon=True,                 # Mostrar borde
     framealpha=0.5,               # Transparencia del borde
     shadow=False,                 # Sombra detrás de la leyenda
     prop={'size': 6, 'family': 'monospace'},  # Fuente de la leyenda
     bbox_to_anchor=(0.676, 0.996)
)
plt.xticks(fontsize=9)
plt.yticks(fontsize=9)
# Mostrar gráfico
plt.show()
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

V. CONCLUSIONES

PRIMERA: Se realizó un análisis exploratorio de los datos de los postulantes a los procesos de admisión entre los años 2018 y 2022. Se identificó que el 74% de los postulantes provienen de colegios públicos, mientras que el 26% provienen de colegios privados. Además, se observó que el 35% de los postulantes habían intentado ingresar en más de una ocasión, lo que sugiere una persistencia en el intento de admisión. En términos de distribución geográfica, el 60% de los postulantes provienen de la región de Puno, mientras que el 40% restante se distribuye en otras regiones.

SEGUNDA: Se aplicaron técnicas de normalización y codificación para procesar los datos. Se evaluó la importancia de las variables, destacándose que el puntaje obtenido en intentos previos contribuye en un 35% a la predicción del ingreso, seguido por el tipo de colegio (25%) y la modalidad de admisión (18%). Se eliminaron variables con baja relevancia, lo que permitió reducir el tiempo de entrenamiento de los modelos en comparación con el dataset sin preprocesamiento.

TERCERA: Se implementaron seis modelos de aprendizaje automático para predecir el ingreso de los postulantes. Los resultados fueron los siguientes:

Support Vector Machines (SVM) obtuvo la mejor precisión con un accuracy del 92%, una precision de 86% y un recall del 43%.

K-Nearest Neighbors (KNN) y Random Forest mostraron un desempeño similar con un accuracy del 85% y 83%, respectivamente.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Regresión Logística alcanzó un accuracy del 78%, mientras que Árboles de Decisión logró un 84%.

XGBoost tuvo el rendimiento más bajo con un accuracy del 85%, pero con una precisión de 0% en la clase de ingresantes.

En términos de área bajo la curva (AUC), Regresión Logística obtuvo el mejor valor (0.72), seguido de SVM (0.69) y KNN (0.53).

CUARTA: Al evaluar los modelos con datos del proceso de admisión del año 2022, se identificó que el accuracy de SVM se mantuvo en 92%, mientras que Random Forest y KNN presentaron una leve disminución del 2% y 1%, respectivamente. Sin embargo, la predicción de ingresantes continuó siendo un desafío: Regresión Logística mejoró su recall en la clase de ingresantes al 58%, mientras que Random Forest solo alcanzó un recall del 15% y XGBoost del 0%. SVM y Regresión Logística son los modelos más prometedores, ya que presentan el mejor balance entre precisión y recall, así como buenos valores de AUC.

Los resultados obtenidos a partir de la implementación y evaluación de distintos modelos de aprendizaje automático permiten **aceptar la hipótesis general** y, por ende, **rechazar la hipótesis nula**.

Se logró identificar variables demográficas con fuerte influencia en los resultados de ingreso, tales como el tipo de colegio, género, núm. postulaciones, puntaje, tipo de ruralidad, años que pasaron desde que egreso del colegio, género, provincia donde nación.

En términos de capacidad predictiva, los modelos alcanzaron niveles relevantes de desempeño, destacando el modelo **SVM** con una precisión del **92%**, así como un valor

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

AUC de **0.72** en el caso de la regresión logística, lo cual evidencia una capacidad aceptable para distinguir entre ingresantes y no ingresantes.

Asimismo, la validación de los modelos con datos del año **2022** confirma que es posible utilizar estos enfoques como herramienta de apoyo en la **toma de decisiones institucionales**, orientadas a comprender mejor el perfil del postulante y fortalecer los procesos de admisión.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

VI. RECOMENDACIONES

PRIMERA: Se recomienda ampliar el análisis exploratorio de los datos incluyendo análisis de correlación avanzada para mejorar la calidad del dataset. Además, se sugiere implementar herramientas de visualización más detalladas que permitan identificar patrones en los postulantes, como el uso de clustering para segmentar perfiles de postulantes según su origen educativo entre otros.

SEGUNDA: Se recomienda realizar pruebas con técnicas más avanzadas de selección de características, como métodos basados en SHAP (SHapley Additive exPlanations) o Permutation Importance, para identificar las variables con mayor impacto en la predicción. Además, se sugiere evaluar técnicas de reducción de dimensionalidad, como PCA o t-SNE, para optimizar el rendimiento de los modelos y reducir el sobreajuste.

TERCERA: Se recomienda mejorar el desempeño de los modelos mediante la optimización de hiperparámetros con Grid Search o Bayesian Optimization. Asimismo, se sugiere probar enfoques híbridos, como combinar modelos de regresión logística con árboles de decisión (Stacking) o aplicar técnicas de boosting (LightGBM o CatBoost) para mejorar la predicción de ingresantes.

CUARTA: Se recomienda validar periódicamente los modelos con datos de admisión actualizados para mejorar su capacidad de generalización. Es importante desarrollar un sistema automatizado de monitoreo y ajuste de modelos, que permita detectar cambios en los patrones de admisión y realizar mejoras dinámicas. También se sugiere realizar pruebas con modelos de deep

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

learning, como redes neuronales profundas (MLP), para explorar mejoras en la capacidad predictiva.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

VII. REFERENCIAS BIBLIOGRÁFICAS

Alvarez, S. (19 de 01 de 2010). *desarrolloweb.com*. Recuperado el 03 de 01 de 2015, de http://www.desarrolloweb.com/articulos/introduccion-fpdf.html

Arias, F. G. (2004). *El proyecto de investigación, Introducción a la metodología científica.* C.A.: Episteme.

Avila Acosta, R. B. (2001). *Metodología de la Investigación.* Lima: Estudios y Ediciones R.A.

Avila Barreno, N., Magallanes, D. G., & González, L. N. (23 de 12 de 2009). Proyecto de implementación de un sistema de seguridad para la empresa Devies Corp. en la Ciudad de Milagro para prevenir pérdidas de inventarios por casos fortuitos. Guayaquil, Guayaquil, Ecuador.

BitCompany. (25 de 09 de 2012). *The Business and it governance starting point*. Recuperado el 03 de 02 de 2015, de http://www.bitcompany.biz/que-es-cobit/#.VNFmtywZkww

Blanco Perez, F., Flores Paredes, E., & Giménez Mercado, C. (2010). La equidad y la calidad en los procesos. *Revista de pedagogía*, 276.

Borges de Barros Pereira, H. (15 de abril de 2002). *TDR Tesis Doctorales en Red.* Recuperado el 22 de 12 de 2014, de http://www.tdx.cbuc.es/handle/10803/6542;jsessionid=9DF2DA9FA32C0A6AF2C25DD6E04FC253.tdx1

Borghello, C. (s.f.). *SEGU.INFO*. Recuperado el 1 de abril de 2015, de http://www.segu-info.com.ar/fisica/biometricos.htm

Chapman, P., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & otros, y. (2000). *CRISP-DM 1.0 Step-by-step dataminig guide.*

Cortes Flores, A., & Level, J. P. (2008). El proceso de admisión como predictor del rendimiento académico en la educación superior. *PEPSIC*.

Dalton P. y Whitehead, P. (2000). *Sql Server 2000.* España.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

De la Cruz Villar, J. (2004). *PHP y mYSQL*. Lima: Megabyte SAC.

De Miguel Castaño, A. (2012). *Diseño de base de datos relacionales.* México.

De Miguel, C. A. (2000). *Diseño de base de datos relacionales.* Madrid: RA-MA.

Diaz Lopez, J. (2015). *Análisis de las líneas de autobuses urbanos de A Coruña.* España.

Díaz, R. M. (4 de octubre de 2010). *Modelos predictivos aplicados a la conservación de invertebrados protegidos ibero-beleares.* Obtenido de DIGITAL.CSIC: http://digital.csic.es/handle/10261/92146

E. Prieto, M. (2010). *Recursos digitales para la educación y la cultura.* Mexico: Merida.

Espino Timon, C. (16 de enero de 2017). *Universitat Oberta de Catalunya*. Obtenido de http://openaccess.uoc.edu

Goméz, J. (26 de 04 de 2013). *Dolthink Tecnología e Innovación* . Recuperado el 15 de 01 de 2015, de http://www.dolthink.com/minucias-y-huellas-dactilares.html

Hernández Sampieri, R. y. (2006). *Metodología de la Investigación.* México: McGRAW-HILL.

intersoftware.org. (2004). *HOMINI*. Recuperado el 02 de 2015, de http://www.homini.com/new_page_5.htm

Iturbide Dominguez, G. (2013). *Metodología de Preparación de Datos Orientada a Aplicaciones de Epidemiología Basada en el Modelo CRISP-DM.* México: Morelos.

James, D. (12 de agosto de 2016). *R interface to the MySQL database.* Obtenido de CRAN.

Jhon, H., & Reitsch, A. (1996). *Pronósticos en los negocios.* México: McGrawHill.

Jiménez, M. A. (2010). Evaluación Biométrica de Algoritmos de Autenticación de Firma Manuscrita. Leganés, Madrid, España.

Kadlec, J. (2015). WaterML R package for managing ecological experiment data on a CUAHSI HydroServer. *ELSEVIER*, 10.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Lamarca Lapuente, M. J. (08 de 12 de 2013). *Hipertexto: El nuevo concepto de documento en la cultura de la imagen*. Recuperado el 25 de 01 de 2015, de http://www.hipertexto.info/

Larman, C. (2001). *UML y Patrones.* Madrid: Printice-Hall.

León, K. R. (12 de enero de 2006). *El web como sistema de información.* Obtenido de http://bvs.sld.cu/revistas/aci/vol14_1_06/aci08106.htm

Martín, F. (5 de octubre de 2015). *E&N.* Obtenido de http://www.estrategiaynegocios.net/opinion/851122-345/inteligencia-de-negocios-predictiva

Ocoña Riola, R. (2 de marzo de 2017). *Escuela Andaluza de Salud Pública.* Obtenido de Escuela Andaluza de Salud Pública: http://www.easp.es/project/descubriendo-r-commander/

Pautasio, L. (16 de julio de 2014). *Mercado.* Obtenido de http://www.mercado.com.ar/notas/informes/8015960/big-data-evolucin-o-revolucin

Pereira Gonzalez, A. (2010). *Análisis predictivo de datos mediante técnicas de regresión estadística.* España.

Perojo, K. R. (12 de 01 de 2006). *El web como sistema de información*. Recuperado el 03 de 02 de 2015, de http://bvs.sld.cu/revistas/aci/vol14_1_06/aci08106.htm

Pete, C. (1 de 12 de 2016). *CRISP-DM 1.0 step-by-step Data Mining Guide.* Obtenido de http://www.crisp-dm.org

Piattini Velthuis, M. G. (2001). *Auditoria Informática un enfoque práctico.* Madrid: RA-MA.

Pressman, R. (1998). *Ingeniería de Softaware un enfoque práctico.* España: McGrawhill.

Pressman, R. S. (2005). *Ingeniería del Software un enfoque práctico.* México: McGRAW-HILL.

Rivera, N. A. (21 de 12 de 2015). Instalación de cámaras de video vigilancia en sectores estratégicos de la ciudad de LojaII Fase. Ecuador, Ecuador, Ecuador.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Rodriguez Fuentes, R., Gallestey, J. B., Rodriguez, P. D., & Lazo, M. M. (2008). Valor predictivo de algunos criterios de selección para el ingreso a la carrera de medicina. *Scielo*.

SAP. (2013). Predecir el futuro de los análisis predictivos. *LOUDHOUSE*, 12.

Vega Briceño, E. A. (04 de 2005). *monografias.com*. Recuperado el 12 de 01 de 2015, de http://www.monografias.com/trabajos24/tics-empresas/tics-empresas.shtml

Velasco Cabo, J. (2017). Análisis exploratorio de datos del mercado eléctrico español con R. *Invurnus, en busca del conocimiento*, 12-17.

Velasquez Uribe, M. T. (2008). *Aprender demografía con R-project.* México.

Velasquez, J., Montoya, O., & Cataño, N. (2010). ¿Es el proyecto R para la computación estadística apropiado para la inteligencia computacional? *Ingeniería y Competitividad*, 81-94.

Velez, I. (19 de juni de 2014). *El coste de no hacer nada en el análisis predictivo de la información*. Obtenido de http://www.muycomputerpro.com/2014/06/19/coste-analisis-predictivo-informacion

Zhao, Y. (26 de abril de 2013). *R contributed Documentation*. Obtenido de R contributed Documentation: https://cloud.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ANEXOS

Anexo 1. Documento presentado a la Dirección de Admisión

"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

SOLICITA: AUTORIZACIÓN PARA REALIZAR TRABAJO DE INVESTIGACIÓN

DR. JUAN CARLOS BENAVIDES HUANCA
DIRECTOR DE LA DIRECCIÓN DE ADMISIÓN DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO DE PUNO

JOSUE CALIP OJEDA MAMANI identificado con documento nacional de identidad Nº 70069346 y EDILBERTO CALLATA COLLANQUI identificado con documento nacional de identidad Nº 72175210; egresados de la Escuela Profesional de Ingeniería de Sistemas nos presentamos ante usted con el propósito de solicitar su autorización para realizar nuestra tesis titulada: **Análisis Predictivo de resultados en los exámenes de Admisión a la Escuela de Ingeniería de Sistemas mediante modelos de aprendizaje automático (2018-2022).**

Para realizar esta investigación es necesario contar con información de los postulantes de los procesos de admisión ordinario: general y cepreuna correspondiente a los años 2018 al 2022.

Cabe resaltar que la información proporcionada será exclusivamente con fines académicos y de investigación, por lo que se mantendrá la confidencialidad y protección de los datos proporcionados por la Dirección de Admisión.

Agradecemos de antemano su respuesta y estaremos atentos a nuestra petición.

JOSUE CALIP OJEDA MAMANI
DNI: 70069346

EDILBERTO CALLATA COLLANQUI
DNI: 72175210

Puno, abril de 2024

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Anexo 2. Funciones para el cálculo de la curva ROC

```
def calcular_roc(modelo, X_test, y_test, df_scaler_test, y_testt, fuente, graficar=True):
    """
    Calcula el AUC-ROC y opcionalmente grafica la curva ROC.
    Parameters:
        modelo: El modelo entrenado.
        X_test: Conjunto de características de prueba.
        y_test: Conjunto de etiquetas de prueba.
        df_scaler_test: Conjunto de características de prueba de 2022.
        y_testt: Conjunto de etiquetas de prueba de 2022.
        fuente: Tamaño de la fuente para las etiquetas.
        graficar: Booleano, si es True se grafica la curva ROC.
    Returns:
        Tuple con los valores AUC-ROC para los datos de prueba y de 2022.
    """
    # Probabilidades de la clase positiva
    y_pred_prob = modelo.predict_proba(X_test)[:, 1]
    # Probabilidades de la clase positiva para 2022
    y_pred_prob_2022 = modelo.predict_proba(df_scaler_test)[:, 1]
    #print(y_pred_prob_2022)
    # Calcular el AUC-ROC para los datos de prueba
    auc_roc = roc_auc_score(y_test, y_pred_prob)
    #print(f'AUC-ROC para los datos de prueba: {auc_roc}')
    # Calcular el AUC-ROC para los datos de 2022
    auc_roc_2022 = roc_auc_score(y_testt, y_pred_prob_2022)
    #print(f'AUC-ROC para los datos de 2022: {auc_roc_2022}')
    if graficar:
        plt.figure()
        _fpr, _tpr, _thresholds = roc_curve(y_testt, y_pred_prob_2022)
        ax = plt.plot(_fpr, _tpr, label=f'Curva ROC 2022 (AUC = {auc_roc_2022:.2f})')

        plt.plot([0, 1], [0, 1], 'k--')
        plt.xlabel('Tasa de Falsos Positivos', fontsize=fuente)
        plt.ylabel('Tasa de Verdaderos Positivos', fontsize=fuente)
        plt.title('Curva ROC')

        plt.xticks(fontsize=fuente - 1)
        plt.yticks(fontsize=fuente - 1)

        plt.legend(loc="lower right", fontsize=fuente - 1)
        plt.grid(False)
        plt.show()

    return auc_roc, auc_roc_2022, ax
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Anexo 3. Función para graficar la matriz de confusión

```
def matriz_confusion(y_obs, y_pred, falso, verdad, fuente):
  #Calcular matriz de confusión
  mc = confusion_matrix(y_obs, y_pred)
  # Intercambiar los valores de las celdas [0,1] y [1,0]
  temp = mc[0, 1]
  mc[0, 1] = mc[1, 0]
  mc[1, 0] = temp
  # visualizar matriz de confusión
  vis = ConfusionMatrixDisplay(mc)
  fig, axes = plt.subplots()
  # cmap = magma, plasma, inferno, viridis, cividis
  im = vis.plot(ax=axes, cmap='viridis', values_format='d')
  #for spine in axes.spines.values():
  #  spine.set_visible(False)
  # codificando etiquetas
  # axes.set_title('Matriz de confusión', size=fuente)
  axes.set_xlabel('Valor predicho (predicciones)', size=fuente)
  axes.set_ylabel('Valor real (observaciones)', size=fuente)
  # Mover etiquetas del eje X arriba
  axes.xaxis.set_label_position('top')  #Mueve el nombre del eje X arriba
  axes.xaxis.tick_top()  # Mueve las etiquetas de los ticks arriba
  # Intercambiar el eje X con el eje Y
  temp = axes.get_xlabel()
  axes.set_xlabel(axes.get_ylabel())
  axes.set_ylabel(temp)
  # Etiquetas personalizadas para los ticks
  axes.xaxis.set_ticklabels([falso, verdad], size=fuente)
  axes.yaxis.set_ticklabels([falso, verdad], size=fuente)
  # Ajustar el tamaño de fuente de los num. en matriz
  for text in axes.texts:
    text.set_fontsize(fuente)
  # Invertir los ejes
  # Invertir el eje Y (primera fila abajo)
  axes.invert_yaxis()
  # Invertir el eje X (primera columna a la derecha)
  axes.invert_xaxis()
  # Acceder a la barra de color
  cbar = im.im_.colorbar
  # Si la barra de color existe, ajustar su tamaño de fuente
  if cbar:
    cbar.ax.tick_params(labelsize = fuente - 1)
  axes.grid(False)
  plt.show()
  return mc
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Anexo 4. Cantidad de postulantes por modalidad y fecha

```
# Contar registros por modalidad y fecha
conteo = data.groupby(['modalidad', 'fecha']).size().reset_index(name='Cantidad')
# Ordenar por fecha y modalidad en orden descendente
conteo = conteo.sort_values(by=['modalidad', 'fecha'], ascending=[True, False])
conteo
```


repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Anexo 5. Declaración jurada de autenticidad de tesis

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo EDILBERTO CALLATA COLLANQUI
identificado con DNI 72175210 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**
INGENIERÍA DE SISTEMAS

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:
" ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018 - 2022) "

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 23 de mayo del 2025

FIRMA (obligatoria)

Huella

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo JOSUE CALIP OJEDA MAMANI,
identificado con DNI 70069346 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**
INGENIERÍA DE SISTEMAS,
informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:
" ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018-2022) "

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 23 de mayo del 2025

FIRMA (obligatoria)

Huella

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Anexo 6. Autorización para el depósito de tesis en el Repositorio Institucional

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo EDILBERTO CALLATA COLLANQUI
identificado con DNI 72175210 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**

INGENIERÍA DE SISTEMAS

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

" ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018 - 2022) "

para la obtención de ☐**Grado,** ☒ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 23 de mayo del 2025

FIRMA (obligatoria) Huella

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo JOSUE CALIP OJEDA MAMANI,
identificado con DNI 70069346 en mi condición de egresado de:

☒ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☐ **Programa de Maestría o Doctorado**

INGENIERÍA DE SISTEMAS,

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

"ANÁLISIS PREDICTIVO DE RESULTADOS EN LOS EXÁMENES DE ADMISIÓN A LA ESCUELA DE INGENIERÍA DE SISTEMAS MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO (2018 - 2022)"

para la obtención de ☐**Grado,** ☒ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 23 de mayo del 2025

FIRMA (obligatoria)

Huella

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis