UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

EVALUACIÓN DE FACTORES DETERMINANTES PARA EL INGRESO DE LOS POSTULANTES A LAS UNIVERSIDADES

PRESENTADA POR:

PABLO CESAR TAPIA CATACORA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAESTRO EN INGENIERÍA DE SISTEMAS

PUNO, PERÚ

2023

Reporte de similitud

NOMBRE DEL TRABAJO

EVALUACIÓN DE FACTORES DETERMIN ANTES PARA EL INGRESO DE LOS POST ULANTES A LAS UNIVERSIDADES

AUTOR

PABLO CESAR TAPIA CATACORA

RECUENTO DE PALABRAS

17863 Words

RECUENTO DE CARACTERES

98584 Characters

RECUENTO DE PÁGINAS

92 Pages

TAMAÑO DEL ARCHIVO

2.7MB

FECHA DE ENTREGA

Nov 29, 2023 7:32 PM GMT-5

FECHA DEL INFORME

Nov 29, 2023 7:33 PM GMT-5

6% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base c

- 4% Base de datos de Internet
- Base de datos de Crossref
- 4% Base de datos de trabajos entregados
- 1% Base de datos de publicaciones
- Base de datos de contenido publicado de Crossr

Excluir del Reporte de Similitud

- Material bibliográfico
- Coincidencia baja (menos de 10 palabras)
- Material citado
- Fuentes excluidas manualmente

Mg. Aldo H. Zanabria Galvez
INGENIERO DE SISTEMAS
CIP 84584

Ing. Estadístico e Informático
CIP. 116015

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

EVALUACIÓN DE FACTORES DETERMINANTES PARA EL INGRESO DE LOS POSTULANTES A LAS UNIVERSIDADES

PRESENTADA POR:

PABLO CESAR TAPIA CATACORA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAESTRO EN INGENIERÍA DE SISTEMAS

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

.....
D.Sc. EDWIN FREDY CALDERON VILCA

PRIMER MIEMBRO

.....
M.Sc. MAGALI GIANINA GONZALES PACO

SEGUNDO MIEMBRO

.....
M.Sc. LENIN HUAYTA FLORES

ASESOR DE TESIS

.....
Mtro. ALDO HERNAN ZANABRIA GALVEZ

Puno, 13 de octubre de 2023

ÁREA: Ciencias de la ingeniería
TEMA: Base de datos e inteligencia de negocios – Random Forest
LÍNEA: Sistemas, computación e informática

DEDICATORIA

A Fernando, Paola y a Emelina, son mis fortalezas y motivación de vida. Sin ellos no hubiese llegado donde estoy ahora.

A mi papá Claudio, mi mamá Fabiana y mi hermano Tomas, con quienes aprendí a caminar y comparto la vida desde cuanto tengo memoria, y ahora me observan y me brindan fortalezas mientras camino en la vida profesional.

A la comunidad que promueve el conocimiento abierto de la inteligencia artificial, la ciencia de datos y a los visionarios que comparten ideales para lograr el éxito personal y profesional.

AGRADECIMIENTOS

Primeramente, a Dios nuestro creador que bajo su bendición y protección divina recorrí y seguiré caminando en este mundo profesional competitivo y de permanente cambio.

A la Universidad Nacional del Altiplano mi alma mater, al Programa de Maestría en Ingeniería de Sistemas. Así mismo, a los docentes de la maestría por darme la oportunidad de culminar mi formación de maestro.

A los Jurados de esta Tesis, D.Sc. Calderon Vilca Edwin Fredy, M.Sc. Gonzales Paco Magali Gianina, M.Sc. Huayta Flores Lenin y mi asesor Mg. Aldo Hernán Zanabria Gálvez, por soportar mis conocimientos básicos en ciencia de datos y hacer más legible esta tesis.

A mis amigos de la universidad y hoy colegas en la Escuela Profesional de Ingeniería de Sistemas, en especial al M.Sc. William Eusebio Arcaya Coaquira, Dr. Edelfre Flores Velasquez, M.Sc. Marga Isabel Ingaluque Arapa, Dra. Zulema Lilian Mamani Huacani, M.Sc. Lilian Magnolia Benique Ruelas, por los ánimos y bromas, a quienes alcanzo mi agradecimiento.

A los amigos con quienes compartí la universidad y algunas experiencias profesionales de en el campo de la investigación y trabajo.

A todas las personas que se toman el tiempo para darme una recomendación, para seguir investigando en el curso de mi vida profesional.

ÍNDICE GENERAL

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

RESUMEN

Los modelos de aprendizaje supervisado basados en bosques aleatorios (Random Forest) tienen alto desempeño al momento de clasificar los factores determinantes que permiten el ingreso de un postulante a la universidad. Esta investigación es de tipo cuasi-experimental y utiliza el método de análisis cuantitativo, en consecuencia el objetivo es evaluar los factores determinantes asociados al ingreso de los postulantes a la universidad. Inicia con el preprocesamiento, comprensión de los datos de admisión y clasificarlos por áreas y procesos de admisión, esta etapa se completa con la limpieza de los datos para evitar lecturas erróneas, luego se construye el modelo de aprendizaje supervisado de bosques aleatorios cuya tarea es predecir con exactitud el ingreso o no de un postulante a la universidad, previo a ello, se establece ajustes utilizando la librería sickit-learn para separar los datos de entrenamiento y de prueba, así como para establecer los hiperparámetros optimizados para cada área y proceso de admisión, solo para garantizar su desempeño óptimo. La exactitud de los resultados depende de la pureza de los datos de entrada, esto confirma la importancia que tiene los factores determinantes asociados al área y proceso de admisión, durante este análisis exploratorio, el modelo propuesto clasifica y predice con una exactitud entre el 80% y el 91% que un postulante ingresa o no a la universidad. Finalmente, la investigación concluye que los factores determinantes: puntaje obtenido, las asignaturas con mayor ponderación, la edad y los años de haber egresado del colegio son los de mayor importancia.

Palabras clave: Árboles de decisión, aprendizaje supervisado, bosques aleatorios, ciencia de datos, proceso de admisión.

ABSTRACT

Supervised learning models based on Random Forests have high performance when it comes to classifying the determinants that allow an applicant to enter a university. This research is quasi-experimental and uses the quantitative analysis method, consequently the purpose is to evaluate the determinant factors associated with the entrance of college applicants. It begins with preprocessing, understanding the admission data, and classifying them by areas and admission processes. This stage is completed with data cleaning to avoid erroneous readings, and then building the supervised learning model of Random Forest whose task is to accurately predict the admission or not of an applicant to the university, before that, adjustments are established using the sickit-learn library to separate the training and test data, as well as to establish the optimized hyperparameters for each area and admission process, just to ensure their optimal performance. The accuracy of the results depends on the purity of the input data, this confirms the importance of the determinants associated with the area and admission process, during this exploratory analysis, the proposed model classifies and predicts with an accuracy between 80% and 91% that an applicant enters or does not enter the university. Finally, the research concludes that the determinant factors: score obtained, the subjects with the highest weighting, age and years of high school graduation are the most important.

Keywords: Admission process, data science, decision trees, supervised learning, Random Forests.

INTRODUCCIÓN

Tomar decisiones estratégicas en base a métricas y análisis visual de datos de los postulantes, garantiza que los ingresantes a la universidad son los más indicados, pero en el contexto de la universidad surgen permanentes cambios, desde los responsables en la dirección hasta los encargados de gestionar la información de admisión, también hay ausencia de profesionales que puedan interpretar la información. A raíz de ello surge la interrogante, ¿En qué medida los algoritmos de aprendizaje supervisado permiten evaluar los factores determinantes que permiten el ingreso de los postulantes a la universidad? y con el fin de satisfacer los requerimientos de la universidad, el propósito de la investigación fue utilizar algoritmos de aprendizaje supervisado comprendidos en el área de ciencias de la ingenierías, línea de investigación de sistemas, computación e informática y el tema de bosques aleatorios (Random Forest) para determinar cuáles son los factores que determinan en gran medida el ingreso o no de un postulante a la universidad. En el aprendizaje automático, existen algoritmos de clasificación y de regresión, siendo el método de la investigación en análisis cuantitativo, se utilizó los algoritmos de clasificación de bosques aleatorios porqué responden muy bien al momento de clasificar a los postulantes y ordenar en base a la media aritmética y desviación estándar sus factores asociados para lograr una vacante en la universidad.

La investigación está distribuida en secciones claramente diferenciadas. Inicia con el Capítulo I, en la cual se ilustra el sustento teórico y conceptual requerido para la investigación, por otro lado, se describe los avances actuales de la inteligencia artificial en el campo del aprendizaje supervisado y su aplicación basada en algoritmos de bosques aleatorios para dar solución a problemas de investigación de naturaleza predictiva. En el Capítulo II, se enfoca los factores predominantes que permiten el ingreso de los postulantes a las universidades, y fue el nexo motivacional de la investigación, también se establece los objetivos e hipótesis propia de la investigación. En el Capítulo III nos centramos en describir la ubicación de la investigación, así como la cantidad de la población y muestra de estudio, una descripción general de las herramientas necesarias y su aplicación en las fases de entrenamiento y predicción del algoritmo de bosques aleatorios. En el Capítulo IV, se describe los resultados alcanzados en la investigación, para una mejor comprensión se clasificó por áreas y procesos de admisión, al mismo

tiempo se muestran las figuras que ilustran visualmente lo siguiente: a) las escuelas profesionales y la distribución de sus postulantes clasificados entre ingresante y no ingresantes, b) la cuantificación de las respuestas correctas por áreas acompañado de su indicador de mínimos óptimos y de nivelación, c) los factores determinantes ordenados en base a su importancia y en forma descendiente. Se acompaña a los resultados los resultados alcanzados en la investigación y las coincidencias y mejoras significativas frente a otras investigaciones que hayan utilizado los bosques aleatorios para clasificar o predecir una situación futura. En la sección de Conclusiones, presentamos de forma resumida los resultados de la investigación después de entrenar el modelo de aprendizaje supervisado de bosques aleatorios. En la sección de las Recomendaciones incluimos algunas consideraciones a tener en cuenta en las futuras investigaciones.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1. Marco teórico

1.1.1. Arboles de decisión

Los árboles de decisión son bastante intuitivos y utilizado por todas las personas en algún momento de sus vidas, a sabiendas o no. Son distribuciones en las cuales se visualiza los posibles resultados de una serie de decisiones. Inicia con un nodo base y luego se ramifica en bloques izquierdo y derecho hasta llegar a posibles resultados (Díaz-Martínez *et al.*, 2021).

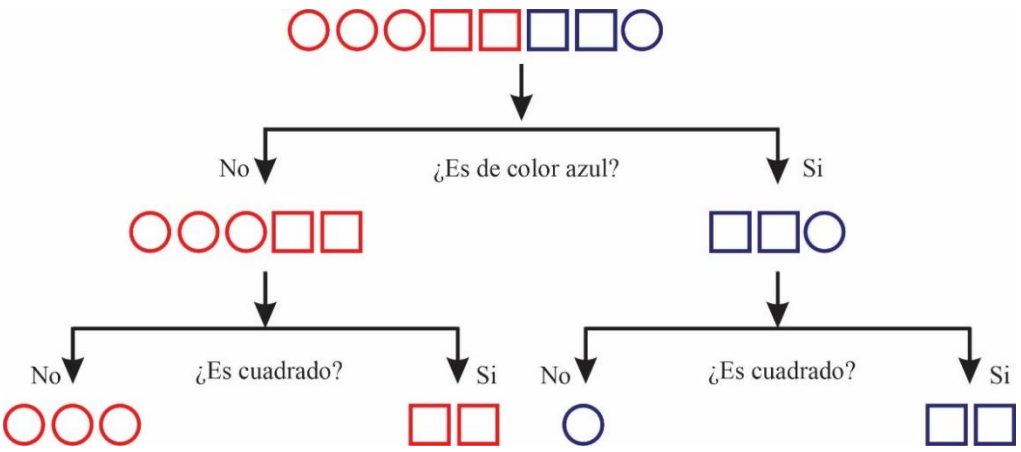


Figura 1. Ejemplo de árbol de decisión basado en formas geométricas.

En la Figura 1 se tiene 2 clases y son círculos y cuadrados, los cuales intentamos separarlos tomando en cuenta ciertas características, primero el color en particular (azul y rojo), luego si la observación es un cuadrado o no.

1.1.2. Bosques aleatorios – Random Forest

Consiste en una gran cantidad de árboles de decisión individual (simple) que operan como un conjunto. Cada árbol individual en el Random Forest, retorna como salida una predicción a la clase asociada (Yeşilkanat, 2020). La clase con la mayor cantidad de salidas se convierte en la predicción del modelo de bosque aleatorio (Schonlau & Zou, 2020).

Para obtener una alta precisión en los resultados, es importante que los árboles de decisión individuales tengan una baja correlación (Zhang *et al*., 2020). Esto hace que las predicciones en conjunto sean más precisas que cualesquiera de los árboles individuales. Esta condición hace que los árboles se protegen de sus errores individuales y evitan el sobreajuste (overfitting).

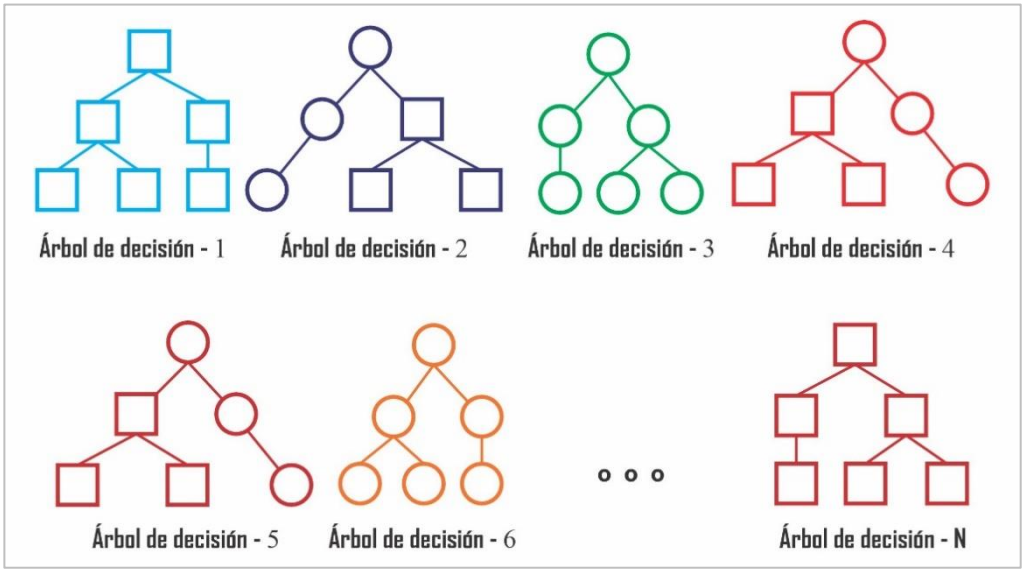


Figura 2. Bosque aleatorio conformado por N árboles de decisión.

En la Figura 2 se visualiza las diferentes distribuciones que son árboles de decisión individuales, Random Forest se basa en el aprendizaje de dicho conjunto de "n" árboles.

Importancia de los hiperparámetros de los bosques aleatorios son ajustables, permitiendo obtener mejores predicciones y los más importantes son:

- **n_estimators**: número de árboles de decisión individuales generados.

- **max_features**: manera de seleccionar la cantidad máxima de los predictores para cada árbol.

- **min_sample_leaf**: número mínimo de elementos en las ramas para una nueva división del nodo.

- **oob_score**: método que simula la validación cruzada en árboles permitiendo mejorar la precisión y el sobreajuste (overfitting).

- **boostrap**: determina los tamaños aleatorios de muestras para entrenar. Cuando se encuentra en modo ***false***, utiliza el data set completo para su entrenamiento.

- **n_jobs**: muy usado cuando el hardware utilizado dispone de múltiples núcleos en los procesadores, lo cual ayuda acortando los tiempos de entrenamiento.

1.1.3. Ventajas distintivas de Random Forest

- Reducción de riesgo de overtiffing (sobreajuste). Los árboles de decisión individuales corren el riesgo de sobreajuste, tienen una tendencia a un estrecho ajuste a todas las muestras dentro de los datos de entrenamiento. Esto se evita con una gran cantidad de árboles de decisión contenidos en el Random Forest, el clasificador no sobreajustará el modelo, debido al promedio de los árboles no correlacionados se reduce la varianza y el error de predicción.

- Brinda flexibilidad. Random Forest tiene una notable ventaja en manejar tareas de regresión y clasificación con alto grado de precisión, y muy útil en el mundo de la ciencia de datos. Aun faltando una parte de los datos, Random Forest mantiene la precisión. Esto hace que sea una herramienta eficaz para estimar datos faltantes.

- Fácil de determinar las características importantes. Facilita la evaluación de lo importante de las columnas (variables) o la contribución del modelo.

1.1.4. Campos de aplicación de Random Forest

Hasta el momento las aplicaciones de Random Forest se dieron en el campo de la industria, los casos más visuales son:

- Finanzas. Son muy utilizados para determinar los clientes con altos riesgos de crédito, para detectar fraudes, estimación de precios, ya que reduce los tiempos en las tareas de preprocesamiento y gestión de datos.

- Medicina. Permite a los profesionales de la salud abordar problemas como la clasificación de expresiones génicas, clasificación de secuencias, descubrimiento de biomarcadores, detección temprana de enfermedades. Esto implica su aplicación dentro del campo de la biología.

- Comercio electrónico. Se puede utilizar en los sistemas de recomendaciones con fines de mejorar las ventas.

- Educación. Para identificar las tendencias, predecir comportamientos, analizar encuestas, estudios de opinión y modelar la dinámica de comportamiento de los estudiantes, hacer recomendaciones personalizadas con datos de fichas socioeconómicas.

- Agricultura. En la agricultura de precisión, detectar posibles enfermedades de las plantas, optimización de uso del agua y fertilizantes, clasificar productos en base a proteínas o componentes químicos, etc.

- Regresión. Para predecir el comportamiento de compra-venta de viviendas en función a ubicación, tamaño, tiempo y otras características.

1.1.5. Ciencia de datos

Al preprocesamiento de los datos, la preparación y la interpretación de datos se le conoce como la ciencia de datos, es una tecnología en asenso que trata de comprender y encontrar patrones e información útil en medio de ingentes cantidades de datos (estructurados, no estructurados, y sin procesar), información que se puede utilizar para la toma de decisiones de manera inteligente u otros objetivos que se busca alcanzar (Lemus-Delgado & Pérez Navarro, 2020).

Su utilidad se encuentra en la capacidad de convertir datos en información valiosa y conocimientos para impulsar la toma de decisiones estratégicas de forma continua en una variada gama de campos. Las aplicaciones y utilidades de la ciencia de datos se pueden resumir en lo siguiente:

- Predicción y pronóstico. Predecir los eventos futuros o posibles resultados en sectores variados de la industria. Por ejemplo, la demanda de los productos terminados, la tendencia de los precios de las acciones, el comportamiento del clima, la propagación y contagio de las enfermedades, el tráfico (terrestre, marítimo, aéreo y espacial).

- Optimización de procesos. Optimización de las cadenas de suministros con los proveedores, procesos de manufactura y logística, mejorar la eficiencia y productividad.

- Personalización. Las tendencias de personalización de pedidos en línea de los clientes, personalización de contenidos en redes sociales y de anuncios publicitarios, recomendación de productos en base al comportamiento del cliente como puede ser la compra de un vehículo personalizado, etc.

- Detección de fraudes. Identificar transacciones financieras sospechosas, operaciones con tarjetas de créditos, seguros u otros sistemas de pagos.

- Análisis de redes sociales. Analizar datos provenientes de las redes sociales, comprender el comportamiento de los actores y grupos de las redes sociales, segmentación de la opinión pública y sus tendencias.

- Marketing y publicidad. Segmentar audiencias en base a preferencias, conocer el impacto de las campañas de publicidad, optimización de las estrategias de publicidad en línea o fuera de ella.

- Recursos humanos. Medición del impacto de difusión de convocatorias, selección estratégica de personal requerido, retención de empleados clave, medir la gestión de desempeño.

- Investigación científica. Comprender datos en los campos de la astronomía, la biología, la física, el medio ambiente, el cambio climático entre otros. Para luego analizar y modelar datos científicos complejos.

- Gobierno. En la toma de decisiones políticas clave, detección de fraudes fiscales, mejora de los servicios públicos, mejora de la infraestructura pública.

- Educación. Mejorar las estrategias de enseñanza-aprendizaje hacia los alumnos con la adaptación de contenidos, identificación de patrones en el rendimiento académico, identificación de talentos.

- Transporte y logística. Ayuda en la planificación de rutas (terrestres, aéreas, marítimas y fluviales), gestión de flotas, optimización de entregas, mejora continua de la logística.

1.1.6. Bagging

En el universo del aprendizaje, el algoritmo de Bagging es una forma organizada de lograr el aprendizaje por conjuntos, se utiliza para reducir la varianza dentro de un conjunto de datos cuya existencia posee demasiadas variaciones (ruido). En este proceso se selecciona muestras aleatorias de datos hasta convertirse en modelos, estos se entrenan de forma independiente, ya sea para regresión o clasificación. Siendo el promedio o la mayoría de dichas predicciones, una estimación de alta precisión (Breiman, 2020).

1.1.7. Boosting

El algoritmo de Boosting tiene relación directa con el Bagging, porque son métodos de aprendizaje por conjuntos, siendo su única diferencia el entrenamiento. Es decir, en el método Boosting el aprendizaje individual se entrena de forma paralela (Espinosa, 2020).

Inicia con la construcción una serie de modelos y en cada interacción, cuando un dato es clasificado de forma errónea en el modelo anterior, su peso se incrementa. Con esta redistribución en los pesos, se logra mejorar el rendimiento del algoritmo.

1.1.8. Índice GINI

Es una medida de dispersión utilizada en los árboles de decisión y bosques aleatorios. Se utiliza para evaluar la eficiencia los grupos en un nodo de un árbol de decisión, es decir, cómo se debe formar los nodos de los árboles de decisión individuales que conforman el bosque aleatorio (Algehyne *et al*., 2022).

Se calcula en base a la siguiente ecuación:

$$Gini(D) = 1 - \sum_{i=1}^{c} (p_i)^2$$

Donde:

- *Gini(D)*: índice GINI del nodo.
- *c*: número de clases o categorías predictoras en el nodo.
- p_i: proporción de muestras del nodo que pertenecen a la clase *i*.

Los valores de este índice varían entre el rango de 0 y 1, donde:

- 0: nodo completamente puro (*todas las muestras del nodo pertenecen a la misma clase*).
- 1: nodo completamente impuro (*las muestras se distribuyen de manera uniforme entre las clases*).

1.1.9. Aprendizaje por conjuntos

El aprendizaje por conjuntos se basa en la filosofía de la "sabiduría las masas" Rueda (2020), refiere a un conjunto de modelos de aprendizajes individuales trabajando en forma colectiva para lograr una mejor predicción. Este trabajo colectivo reduce el sesgo o varianza, haciendo que el modelo tenga un mejor rendimiento.

1.1.10. Overfitting

El sobreajuste (overfitting) es un escenario en la cual un algoritmo de aprendizaje automático realiza predicciones de alta precisión con los datos de entrenamiento y su precisión es muy baja con datos son nuevos o no vistos (Baba & Sevil, 2020). Esto ocurre porque el modelo aprende con mucho detalle durante el entrenamiento, confunde el ruido y la variabilidad aleatoria en lugar de encontrar patrones genuinos, limitando al modelo la capacidad de generalizar los resultados.

Cuando un modelo presenta sobreajuste, resalta algunas características como:

- Alto rendimiento con datos de entrenamiento.

- Pésimo rendimiento con datos de prueba o nuevos (validación).
- Curva de aprendizaje irregular.
- Modelos complejos con muchos parámetros y árboles de decisión muy profundos.

La ocurrencia de un sobreajuste se puede evitar eliminando árboles de decisión individuales (Schonlau & Zou, 2020). En los algoritmos de bosques aleatorios resulta muy improbable su ocurrencia porque se basa en la aleatorización durante la formación de árboles, además resulta inherente aplicar estrategias como la validación cruzada, una adecuada división de los datos de entrenamiento y prueba, un nivel adecuado de complejidad tomando en cuenta la cantidad y calidad de datos (Breiman, 2020).

1.1.11. Validación cruzada

Es una técnica muy utilizada en algoritmos de aprendizaje supervisado y estudios estadísticos para validar la predicción y rendimiento de un modelo (Seraj *et al*., 2022). Proporcionan estimación confiable cuando el modelo funcione con datos nuevos, de esta forma ayuda a evaluar con precisión la capacidad de generalización de un modelo y su independencia de los resultados.

Utiliza únicamente los datos de entrenamiento durante la construcción del modelo predictivo, luego se valida con los datos de prueba. Su precisión se soporta en la comparación de los resultados predichos con los datos reales.

1.1.12. Out of bag score (OOB)

Es una potente técnica de validación para los algoritmos de aprendizaje supervisado como Random Forest para obtener resultados de varianza muy baja (Mohandoss *et al*., 2021). Permite a los modelos evitar la pérdida de datos y al mismo tiempo poder validarlo.

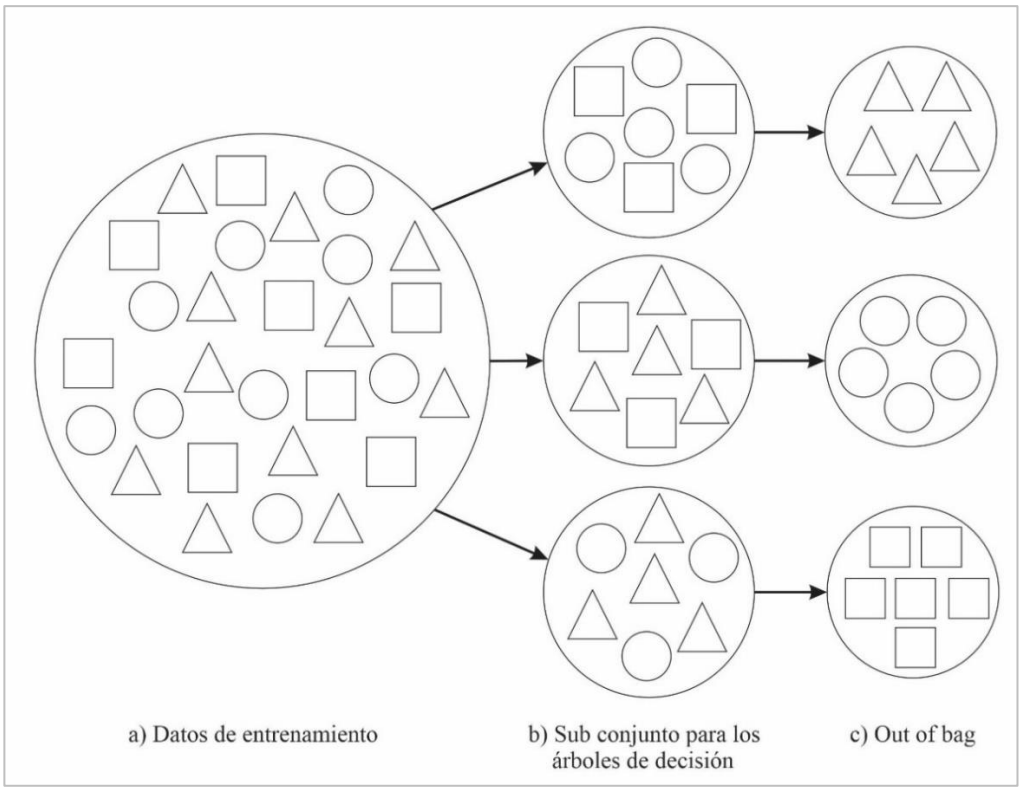


Figura 3. Ejemplo de Out of Bag - Datos fuera del subconjunto de árboles de decisión.

En la Figura 3-c se muestra los tres escenarios y ejemplos de Out of Bag presente en los bosques aleatorios: En los datos del subconjunto de círculos y cuadrados, se excluye los triángulos, esta exclusión representa los datos fuera de subconjunto (Out fo Bag). Del mismo modo en el último subconjunto formado por círculos y triángulos, se excluyen los cuadrados, y ello es otro ejemplo de datos fuera de subconjunto.

1.1.13. Exploratory Data Analysis

El análisis exploratorio de datos es considerado un método que se utiliza para manipular los orígenes de datos, analizar datos, investigar conjuntos de datos y resumir las principales características de datos y sus relaciones interesantes entre ellos utilizando métodos de visualización de datos (Tukey, 1977). Permitiendo de esta forma descubrir patrones no visibles, detectar anomalías y valores atípicos, comprobar los supuestos o hipótesis planteadas (Sahoo *et al*., 2019).

La importancia del análisis exploratorio de datos radica en la generación de resultados válidos y aplicables a conclusiones y objetivos planteados durante la

investigación. Estas características permiten crear modelos de machine learning mucho más complejos para analizar e interpretar datos (Abelairas-Etxebarria & Astorkiza, 2020).

1.1.14. Aprendizaje automático – Machine Learning

Es parte de la inteligencia artificial centrado en el desarrollo de algoritmos y modelos computacionales que permiten el aprendizaje de las máquinas mediante la experiencia y el análisis de datos (Janiesch *et al*., 2021). Consecuentemente las máquinas adquieren la capacidad de aprender y tomar decisiones o bien realizar predicciones basados en patrones y conocimientos extraídos a partir de los datos.

Elementos básicos del proceso de aprendizaje

- Datos de entrenamiento: Comprende el conjunto de datos de entrada y salida debidamente etiquetadas para el entrenamiento del modelo. Son la esencia para que el algoritmo aprenda patrones y sus relaciones.
- Elección del algoritmo: la elección se basa en el tipo de tarea que puede ser de clasificación, de regresión o bien de agrupamiento.
- Entrenamiento del modelo: es el proceso de construcción del modelo matemático en el cual un modelo se ejecuta con los datos de entrenamiento y a la vez, ajustando sus parámetros. Modelo que finalmente es capaz de realizar predicciones o tomar decisiones basadas en nuevos datos proporcionados.
- Validación y ajuste del modelo: utiliza datos de validación para evaluar el rendimiento del modelo a la vez, ajustando sus hiperparámetros. Esto con el fin de mejorar su precisión y generalización.
- Pruebas y despliegue: esta parte del proceso implica realizar predicciones con datos nuevos que no forman parte del entrenamiento. Esto es la aplicación del modelo en diferentes escenarios del mundo real.

1.1.15. Algoritmo de Machine Learning

Es un conjunto de instrucciones finitas y reglas matemáticas que permite un computador aprender y mejorar su rendimiento en tareas específicas mediante el

entrenamiento y análisis de datos (Glaser *et al*., 2020). Estos algoritmos permiten completar tareas de identificación de patrones, realizar predicciones precisas y tomar decisiones basadas en datos sin estar programados de forma exclusiva para dicha tarea.

Los algoritmos de machine learning se pueden agrupar en tres categorías muy bien diferenciadas:

- Aprendizaje supervisado. Los algoritmos de esta categoría se entrenan con un conjunto de datos etiquetados de entrada y salida correspondiente. Su tarea fundamental es que el algoritmo aprenda a mapear las entradas y salidas, en base a ello pueda realizar predicciones precisas con nuevos datos.

- Aprendizaje no supervisado. Estos algoritmos se entrenas con conjuntos de datos no etiquetados y dentro de ese mar de datos descubrir patrones o relaciones ocultas entre los datos. Incluye tareas como agrupación, reducción de dimensionalidad de los datos, autoencoders.

- Aprendizaje por refuerzo. En estos algoritmos utilizan un aprendizaje acumulativo durante el proceso de entrenamiento, este aprendizaje se da mediante la retroalimentación del entorno, ajustando progresivamente para mejorar su aprendizaje y precisión. Son utilizando en entornos como la robótica, sistemas de control y los juegos estratégicos.

Son algoritmos esenciales en campos como: visión computacional, procesamiento de lenguaje natural (NLP), detección de fraudes, la medicina, predicción mediante series de tiempo, optimización de las cadenas de suministro, recomendación de productos terminados, la conducción autónoma (Ponce, 2010).

El funcionamiento óptimo del tipo de algoritmo depende de los datos y los objetivos que se espera alcanzar.

1.1.16. Redes neuronales

Las redes neuronales son modelos computacionales simulados estructuralmente en base al funcionamiento del cerebro de los humanos (Alcaide Martínez, 2020). Están diseñados para aprender y completar tareas específicas a partir de datos reales y en

ocasiones simulados (Alamilla-Jiménez *et al*., 2022). Se compone de nodos y capas interconectadas, donde cada conexión tiene un peso que se ajusta continuamente durante el proceso de entrenamiento para lograr el resultado esperado (Rojas *et al*., 2020).

Le ventajas de utilizar las redes neuronales para completar con precisión las tareas radica en:

- Capacidad de aprendizaje. Tienen una particularidad de aprender y adaptarse para buscar soluciones a una amplia variedad de problemas de aprendizaje automático partir de los datos.
- Manejo de datos complejos. Tiene la capacidad de manejar datos no lineales y complejos, permitiendo ser efectivas en tareas como la visión computacional y procesamiento de imágenes.
- Generalización. Pueden generalizar a partir de ejemplos de entrenamiento para lograr predicciones precisas sobre datos nuevos o no vistos anteriormente, lo que permite ser muy versátil su aplicación en el mundo real.
- Paralelismo. Su entrenamiento e inferencia son paralelizables cuando se cuenta con el hardware adecuado, siendo muy eficientes en su procesamiento.

Así como existen ventajas notables, también presenta algunas desventajas como las que siguen:

- Requiere enormes cantidades de datos. Para lograr una mejor precisión de los resultados, las redes neuronales dependen de grandes cantidades de datos, lo que representa un desafío cuando se cuenta con una cantidad reducida de datos.
- Sobreajuste (overfitting). En la mayoría de los casos las reden neuronales son propensas al sobreajuste, esto implica que pueden aprender en demasía los detalles de los datos de entrenamiento y no generalizar muy bien con datos nuevos. Esto ocurre cuando no se tiene un control adecuado.

- Interpretabilidad. Cuanto mayor es el número de capas, por lo general tienden a complicarse la interpretación lo que puede desencadenar problemas en aplicaciones donde se requiere mayor explicación y transparencia.

- Selección de hiperparámetros. Cuando no se tiene la suficiente experiencia y tiempo, ajustar los hiperparámetros se puede volver complejo e impreciso.

- Dependencia de datos de entrenamiento. Para evitar los sesgos en los datos, las redes neuronales dependen mucho de la calidad y representatividad de los datos de entrenamiento, esto se puede traducir en desafíos por solucionar.

1.2. Antecedentes

El trabajo de investigación refiere a los antecedentes que ayudaron con mayor trascendencia en el desarrollo de la misma, se consideró los siguientes:

Speiser *et al*. (2019) indica que el bosque aleatorio es un método para desarrollar modelos de predicción basados en clasificación aleatoria en entornos de investigación, mientras que para (Baba & Sevil, 2020) se utiliza para predecir los rendimientos iniciales de las IPOs (del inglés Initial Public Offering) emitidas en Borsa Istanbul. La precisión con la que un bosque aleatorio predice se prueba con métodos de regresión robusta. Los resultados de la predicción muestran que los bosques aleatorios superaron con creces a otros métodos en todas las categorías de la comparación.

Para Gomes *et al.* (2017), en la investigación sobre el algoritmo de adaptive Random Forest (ARF) para la clasificación de flujos de datos en evolución, demuestra que es muy preciso y utiliza los recursos de forma razonable y Xu & Hoang (2021) agrega un método de remuestreo efectivo con operaciones adaptativas a desviaciones de conceptos y sin optimizaciones complejas para grandes conjuntos de datos. Ambos autores concluyen que los algoritmos ARFs son precisos y de gran fiabilidad al momento de mostrar los resultados, además, utilizan una cantidad factible de recursos.

Charbuty & Abdulazeez (2021) en su investigación indica que los bosques aleatorios son herramientas del aprendizaje automático basado en árboles de decisión que son altamente adaptables a grandes conjuntos de datos, para Ao *et al*. (2019) los bosques aleatorios son particularmente atractivos para el análisis de datos genómicos de alta dimensión, este proceso incluye: predicción y clasificación, selección de variables, análisis de vías, asociación genética y detección de epistasis, y aprendizaje no supervisado. De ello se deduce que los árboles de decisión constituyen la base de un bosque aleatorio.

Asadi *et al*. (2021) en su artículo propone un nuevo enfoque multiobjetivo evolutivo para predecir enfermedades del corazón combinando multi-objective particle swarm optimization (MOPSO). Este enfoque lleva generar árboles de decisión diversificados y precisos, con esto se busca mejorar la precisión de la predicción del modelo. La eficacia del método propuesto se investiga comparando su rendimiento en seis conjuntos de datos

cardíacos con clasificadores individuales y de conjunto. Concluyendo que el método MOPSO tenga mejor rendimiento que los bosques aleatorios.

Para Biau *et al*. (2019), una red neuronal multicapa se puede construir a partir de un conjunto de árboles aleatorios de regresión, asignándoles pesos de conexión. Para Nieto *et al*. (2019) los predictores utilizan el conocimiento previo de los árboles de regresión, tienen menos parámetros por ajustar en comparación con las redes neuronales estándar y menos restricciones en los límites de decisión que los árboles. Después de probar resultados de consistencia con datos reales o sintéticos, los autores concluyen que se obtiene excelentes resultados en una gran variedad de problemas de predicción.

Dogan & Birant (2021) indica que la minería de datos juega un papel importante en varias actividades humanas porque extrae los patrones (o conocimientos) útiles desconocidos. Ahora Gupta & Chandra (2020) agrega sus campos de aplicación indicando que debido a sus capacidades, la minería de datos se convierte en una tarea esencial en una gran cantidad de dominios de aplicación, como la banca, el comercio minorista, la medicina, los seguros, la bioinformática, etc. Ambas investigaciones concluyen lo importante que es rol de la minería de datos en la obtención de conocimientos y soluciones reales a problemas reales.

Para Lemay *et al*. (2021) el enfoque de minería de datos se ha implementado con éxito en la educación superior y emerge como un área interesante en la investigación educativa de minería de datos. Está centrada en la identificación y extracción de conocimiento nuevo y potencialmente valioso a partir de los datos. A partir de ello se deriva conclusiones sobre el éxito académico de los estudiantes. La investigación utiliza varias técnicas de modelado predictivo de los modelos K-Nearest Neighbor Lubis *et al*. (2020), Naive Bayes, Decision Tree y Logistic Regression Model Ahmadini, (2022) para predecir el rendimiento de los estudiantes, ya sea excelente o no excelente (Yaacob *et al*., 2019).

El análisis de componentes principales (PCA) para Reddy *et al*. (2020) es un algoritmo muy utilizado en la clasificación de datos, inicia con la estandarización, sus posibles visualizaciones de los resultados, y finalmente la detección de valores atípicos (Gewers *et al*., 2021). El principio que subyace a este algoritmo es la técnica de reducción de dimensionalidad de grandes volúmenes de datos presentes en el mundo real

(Basavegowda & Dagnew, 2020). Los investigadores concluyen que PCA se basa en la reducción de predictores cuantitativos para responder de manera simple a los objetivos buscados.

Para Julianto *et al*. (2021) en su investigación utiliza clasificación de opiniones mediante el análisis de sentimientos mediante un enfoque de minería de textos, clasifica reseñas positivas o negativas proporcionadas por los usuarios sobre el producto Samsung 850 Evo SSD en NewEgg Store, el estudio compara los resultados de 2 modelos, el valor de precisión del modelo de máquina de vectores de soporte, cuyo resultado es 0,87 o 87%, mientras que el valor de precisión del modelo de árbol de decisión es de 0,82 o 82%. Demostrando que los mejores modelos son aquellos basados en máquina de vectores de soporte.

El campo de la medicina no es ajeno para los algoritmos de aprendizaje supervisado, Li *et al*. (2011) utiliza un modelo basado en support vector machines (SVM) para clasificar electrocorticogramas, que son fuentes de señal provenientes de campos de interfaces cerebro-computador, en la cual la extracción de características es crucial para aumentar la tasa de precisión de la clasificación. La tasa de precisión de clasificación alcanzada por el modelo es del 83%. Una vez más las máquinas de vectores de soporte demuestran su capacidad predictiva.

Para Yu (2021) la educación en línea es un campo donde se puede predecir el rendimiento académico en línea de los estudiantes, analiza el árbol de decisión del algoritmo de clasificación única y el bosque aleatorio del algoritmo de aprendizaje conjunto. El modelo se evalúa mediante análisis empírico y se compara la precisión de las pruebas de varios algoritmos diferentes. Se encuentra que la precisión de predicción del algoritmo bosques aleatorios es superior al 90%, lo que demuestra que se pueden optimizar alternativas de enseñanza-aprendizaje para mejorar las actividades de aprendizaje en línea.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

Cuando la información almacenada no es utilizada, se convierte en desechos de la informática. La universidad no cuenta con un marco legal estandarizado para el almacenamiento de la información de todos los procesos de admisión, no hay procedimientos para depurar información, tampoco hay evidencias de información limpia que pueda generar resultados plausibles para que las autoridades universitarias puedan tomar decisiones acertadas. Ello ocurre casi con toda la información que dispone y almacena la universidad.

Gestionar la información de los postulantes a la Universidad Nacional del Altiplano es tarea de los miembros encargados de la Dirección de Admisión, ya sea con fines puramente académicos o para la toma de decisiones, estas acciones deben ser coordinadas con las autoridades de turno. No hay evidencias visibles de la gestión adecuada de la información y tampoco se conoce las razones.

A fin de brindar una mejor formación profesional a los nuevos ingresantes, es necesario conocer el dominio de sus destrezas cognitivas, los factores que permitieron lograr una vacante y aptitudes de los ingresantes a las diferentes escuelas profesionales. Dicha información debería conocer las autoridades de las facultades y escuelas profesionales, a fin de se pueda gestar programas que permitan de nivelación de los ingresantes. Es un vacío que existe en la actualidad.

Otra de las ocurrencias son las vacantes perdidas, ocurre cuando algunos ingresantes terminan renunciando a la escuela profesional al cual postularon e ingresaron. Por citar

un ejemplo, en un escenario donde un ingresante al área de ingeniería tiene mejor desempeño en carreras profesionales del área de sociales o viceversa, como consecuencia de ello, probablemente el postulante tome la decisión de abandonar o renunciar a la vacante lograda y tenga que postular nuevamente. Esta acción es perjudicial para otro postulante que pudo haber logrado su ingreso la universidad.

2.2. Enunciados del problema

Estos vacíos existentes en la gestión de la información de los procesos de admisión en la Universidad Nacional del Altiplano, descritos en el ítem anterior, nos lleva a plantearnos la siguiente interrogante.

¿Los algoritmos de aprendizaje supervisado de bosques aleatorios permiten evaluar con precisión los principales factores que determinan el ingreso de los postulantes a la Universidad Nacional del Altiplano?

2.3. Justificación

Gran parte de la tarea del aprendizaje automático es la clasificación, esta investigación buscó conocer los factores determinantes para el ingreso de los postulantes a la universidad. Esta capacidad de conocer dichos factores con precisión resulta esencial para conocer a los postulantes que logran una vacante. Al final utilizando un algoritmo de clasificación se buscó predecir si un postulante con ciertas características en particular ingresaría o no a la universidad.

En la actualidad la ciencia de datos nos proporciona una variedad de algoritmos de clasificación como el clasificador Naive Bayes, máquina de vectores de soporte, regresión logística y árboles de decisión. Estos algoritmos ayudan a organizar e interpretar los datos que finalmente se convierten en conocimientos.

Es muy importante conocer las características y factores que permitieron el ingreso de los postulantes a la universidad, los algoritmos de aprendizaje juegan un rol importante al permitir conocer el grado de influencia de cada factor del postulante asociado a su ingreso, ello permite a los encargados de la Dirección de Admisión y a las autoridades de la universidad, adoptar acciones estratégicas para identificar y seleccionar a los mejores estudiantes de la región y del país.

La investigación evalúa el funcionamiento de los árboles de decisión individuales, después de varias combinaciones de todos ellos se construye un modelo de bosques aleatorios capaz de evaluar y determinar el grado de influencia de cada factor asociado al postulante. Con el modelo validado y entrenado se clasifica los factores comunes que mejor influyen en el ingreso o no de los postulantes a la Universidad Nacional del Altiplano.

La investigación forma parte de la ciencia de datos y proporciona a las futuras investigaciones un modelo validado de algoritmo de clasificación basado en bosques aleatorios para clasificar los factores que más influyen al momento de postular y lograr cubrir una vacante en la Universidad Nacional del Altiplano. El mismo que puede ser replicado para otras universidades del ámbito nacional e internacional.

2.4. Objetivos

2.4.1. Objetivo general

Evaluar los principales factores que determinan el ingreso de los postulantes a la universidad utilizando algoritmos de aprendizaje supervisado de bosques aleatorios.

2.4.2. Objetivos específicos

- Organizar los datos de los postulantes e ingresantes a la Universidad Nacional del Altiplano de los diferentes procesos de admisión.
- Analizar las salidas de los diferentes escenarios generados por los modelos del algoritmo de bosques aleatorios.
- Identificar los factores más determinantes en el ingreso de los postulantes a la Universidad Nacional del Altiplano.

2.5. Hipótesis

2.5.1. Hipótesis general

Los algoritmos de aprendizaje supervisado de bosques aleatorios permiten evaluar con alta precisión los principales factores que permiten el ingreso de los postulantes a la Universidad Nacional del Altiplano.

2.5.2. Hipótesis específicas

- Cuando se dispone de los datos organizados de los postulantes e ingresantes a la Universidad Nacional del Altiplano, la fiabilidad de los datos a procesar es alto.
- Los resultados de los árboles de decisión individuales facilitan el análisis exhaustivo sobre los factores influyentes.
- Los resultados del modelo de bosques aleatorios permiten clasificar con eficacia factores determinantes en el ingreso de los postulantes a la Universidad.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. Lugar de estudio

El estudio se realizó en la Universidad Nacional del Altiplano que se encuentra en el departamento, provincia y distrito de Puno, situado al extremo sur este del Perú, en la meseta altiplánica del Collao, entre los 13°00'00" y 17°17'30" de latitud sur y los 71°06'57" y 68°48'46" de longitud oeste del meridiano de Greenwich, se encuentra a una altitud de 3808 msnm.

3.2. Población

La población de esta investigación son los postulantes a la Universidad Nacional del Altiplano provienen de las diferentes provincias del departamento de Puno, y un 3.6% provienen de otros departamentos según los datos de admisión 2022-II. Todos ellos son egresados de colegios privados, estatales y combinación de ambos.

En base al párrafo precedente, para la investigación se consideró los postulantes a los procesos de admisión del Examen CEPREUNA 2022-II con 5104 postulantes, el Examen General 2022-II con 9193 postulantes, haciendo un total de 14297 postulantes que conformaron la población del estudio.

3.3. Muestra

Para Bowater & Denise (2013) La estadística dentro de la investigación y la ciencia tiene diversas funciones, y sostiene que está el de cuantificar las ocurrencias y su frecuencia.

Por tanto, para la investigación se considera el muestreo no probabilístico decisivo por las siguientes razones:

a) Los bosques aleatorios son un tipo de aprendizaje supervisado.

b) Un postulante tiene la misma posibilidad de ser escogido al momento de entrenar el modelo.

Los bosques aleatorios mejoran su precisión cuando son entrenados con una mayor cantidad de datos y mayor número de variables determinantes, por esta razón la muestra de la investigación es la totalidad de nuestra población, es decir, 14297 postulantes entre ingresantes y no ingresantes.

3.4. Método de investigación

Siendo el problema a resolver la evaluación de los factores que determinan el ingreso de los postulantes a la universidad, la investigación requiere herramientas de software y datos numéricos válidos de los procesos de admisión CEPREUNA y Examen General del semestre 2022-II. De existir datos categóricos, estas se normalizan con el fin de analizar mediante la estadística y luego compararlos.

En base al párrafo precedente y por ser datos numéricos, el tipo de investigación es cuasi-experimental, el método de investigación asociado es el análisis cuantitativo con un nivel de profundidad exploratorio, porque en todo momento se buscó determinar, interpretar y jerarquizar los factores más influyentes que permiten el ingreso de un postulante a la universidad. Por naturaleza estas tareas a realizar consisten en recopilar datos de los postulantes, comprender los datos, descubrir patrones comunes y sus relaciones para determinar su ingreso o no a la Universidad Nacional del Altiplano.

La investigación requirió completar los siguientes procesos:

a) Recopilación de datos de los postulantes de los procesos de admisión 2022-II.

b) Recopilación de datos de las instituciones educativas secundarias de las diferentes gestiones, niveles y modalidades de todo el Perú.

c) Lectura de datos de los resultados de los exámenes de admisión 2022-II de los postulantes a la Universidad.

d) Combinación y tabulación de los datos de los postulantes, instituciones educativas y resultados de admisión.

e) Clasificación de los postulantes por áreas: Biomédicas, Ingenierías y Sociales.

f) Preprocesamiento y filtrado de los datos de los postulantes asociado con los factores determinantes asociados a cada postulante.

g) Separación y clasificación de los datos de los postulantes en grupos de datos para entrenamiento (train) y prueba (test) respectivamente.

h) Selección y creación del modelo de aprendizaje automático de bosques aleatorios (Random Forest).

i) Establecimiento de los mejores hiperparámetros para el entrenamiento mediante validación cruzada.

j) Entrenamiento del modelo de aprendizaje automático y predicción con los datos de prueba. Predicciones que se obtuvieron en base a la matriz de confusión y reporte de clasificación.

Matriz de confusión basado en la librería Scikit Learn

		Predicción	
		Negativo	Positivo
Escenario Real	Negativo	Verdaderos Negativos	Falso Positivo
	Positivo	Falsos Negativos	Verdaderos Positivos

Reporte de clasificación

	Precision	**Recall**	**F1-Score**	**Support**
0	0.xx	0.xx	0.xx	X1
1	0.xx	0.xx	0.xx	X2
accuracy			0.xx	X1 + X2
macro avg	0.xx	0.xx	0.xx	X1 + X2
weighted avg	0.xx	0.xx	0.xx	X1 + X2

k) Cálculo de las principales métricas de clasificación a partir de la matriz de confusión.

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

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

$$F1 - Score = 2 * \left(\frac{recall * precision}{recall + precision}\right)$$

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

l) Extracción e interpretación de los factores determinantes ordenados de mayor a menor predominancia en función a los datos de la matriz de confusión y reporte de clasificación.

m) Tabulación y gráfica de los resultados obtenidos.

n) Redacción de las conclusiones y contraste de los resultados alcanzados.

3.5. Descripción detallada de métodos por objetivos específicos

Para lograr nuestro primer objetivo específico: *Organizar los datos de los postulantes e ingresantes a la Universidad Nacional del Altiplano de los diferentes procesos de admisión*. se procedió como se describe a continuación.

a) Recopilar los datos de inscripción a los procesos de admisión del Centro de Estudios Pre Universitario (CEPREUNA) 2022-II y examen general 2022-II.

b) Recopilar los datos de calificación de las pruebas (número de documento nacional de identificación, tipo de prueba, aula, respuestas).

c) Sistematizar los datos de admisión, los datos de las instituciones educativas secundarias y los datos de calificación de las pruebas.

d) Filtrar las columnas con datos incompletos (null, NaN y otros).

e) Normalizar los datos categóricos de los postulantes e ingresantes a la universidad nacional del altiplano.

f) Selección de las columnas determinantes según se detalla en la Tabla 1

Tabla 1

Columnas predictoras determinantes del ingreso o no a la Universidad Nacional del Altiplano.

Predictor	Descripción
SEXO	Sexo del postulante 1: Masculino 2: Femenino
ESTAD_CIVIL	Estado civil del postulante 1: Soltero 2: Conviviente 3: Casado 4: Divorciado 5: Viudo
INGRESO	Variable objetivo: Ingreso o No
PUNTAJE_TOTAL	Puntaje total obtenido
NIV_MOD	Código de nivel/modalidad 1 - F0: Secundaria 2 - D0: Básica Alternativa 3 - D2: Básica Alternativa Avanzado 4 - G0: Secundaria de Adultos
TIPSSEXO	Código de género 1: Varones 2: Mujeres 3: Mixto
GESTION	Código de gestión del servicio educativo 1: Pública de gestión directa 2: Pública de gestión privada 3: Privado
AREA_CENSO	Código del área geográfica

Predictor	Descripción
	1: Urbana 2: Rural
MATEMATICA I	Preguntas contestadas correctamente en la asignatura de matemática I
MATEMATICA II	Preguntas contestadas correctamente en la asignatura de matemática II
FISICA	Preguntas contestadas correctamente en la asignatura de física
QUIMICA	Preguntas contestadas correctamente en la asignatura de química
BIOLOGIA	Preguntas contestadas correctamente en la asignatura de biología
PSICOLOGIA Y FILOSOFIA	Preguntas contestadas correctamente en la asignatura de psicología y filosofía
GEOGRAFIA	Preguntas contestadas correctamente en la asignatura de geografía
HISTORIA	Preguntas contestadas correctamente en la asignatura de historia
EDUCACION CIVICA	Preguntas contestadas correctamente en la asignatura de educación cívica
ECONOMIA	Preguntas contestadas correctamente en la asignatura de economía
COMUNICACIÓN	Preguntas contestadas correctamente en la asignatura de comunicación
LITERATURA	Preguntas contestadas correctamente en la asignatura de literatura
RAZONAMIENTO MATEMATICO	Preguntas contestadas correctamente en la asignatura de razonamiento matemático
RAZONAMIENTO VERBAL	Preguntas contestadas correctamente en la asignatura de razonamiento verbal
ACTITUDINAL	Preguntas contestadas correctamente en el área aptitudinal
EDAD	Edad del postulante
ANNIOS_EGRESO	Años transcurridos desde que el postulante egresó del colegio

g) agrupar los datos por modalidad (CEPREUNA y Examen General) y áreas (biomédicas, ingenierías, sociales).

Para lograr nuestro segundo objetivo específico: *Analizar las salidas de los diferentes escenarios generados por los modelos del algoritmo de bosques aleatorios*. Se completó los siguientes procedimientos.

a) Establecer los hiper-parámetros para la construcción del modelo de bosques aleatorios.

   - **n-estimators**: establecer el número de árboles de decisión a 150.
   - **max_features**: máximo número de factores determinantes o características (*variables o columnas: 5, 7, 9*)
   - **max_depth**: número de divisiones o niveles para cada árbol de decisión (*None, 3, 10, 20*)
   - **criterion**: criterio de creación de los árboles de decisión (*gini, entropy*)

b) Evaluar la importancia de los factores determinantes en base a diferentes escenarios y la combinación de los hiper-parámetros basados en:

   - Out-of-bag score
   - Versión paralelizada
   - Validación cruzada

c) Selección de los hiper-parámetros de mejor desempeño para la construcción de los bosques aleatorios.

Para lograr nuestro tercer objetivo específico: *Identificar los factores más determinantes en el ingreso de los postulantes a la Universidad Nacional del Altiplano*. Se completó os siguientes procesos.

a) Establecer los hiper-parámetros de mejor desempeño para la construcción del modelo de bosques aleatorios con los siguientes datos.

   - **criterion** : gini

- **max_depth** : None,
- **max_features** : 7
- **n_estimators** : 150

b) Evaluación las predicciones del modelo en base a la matriz de confusión y la precisión con los datos de prueba.

c) Clasificación de los predictores en base a la mayor probabilidad.

d) Determinación de la importancia de los factores predictores para el ingreso a la universidad considerando la importancia de la media aritmética y la desviación estándar.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. Resultados

Para una mejor comprensión de la investigación iniciaremos con los resultados de organizar los datos, luego el análisis de las salidas producidas por el modelo de aprendizaje supervisado de bosques aleatorios.

En el último apartado de esta sección y con la finalidad de vislumbrar con detalle los resultados del modelo y determinar los factores con mayor influencia en el ingreso de los postulantes a la universidad, procesamos los datos por áreas (biomédicas, ingenierías, sociales) y procesos de admisión (CEPREUNA, Examen General), considerando hiperparámetros optimizados para cada área y procesos de admisión.

1. Resultados para el primer objetivo

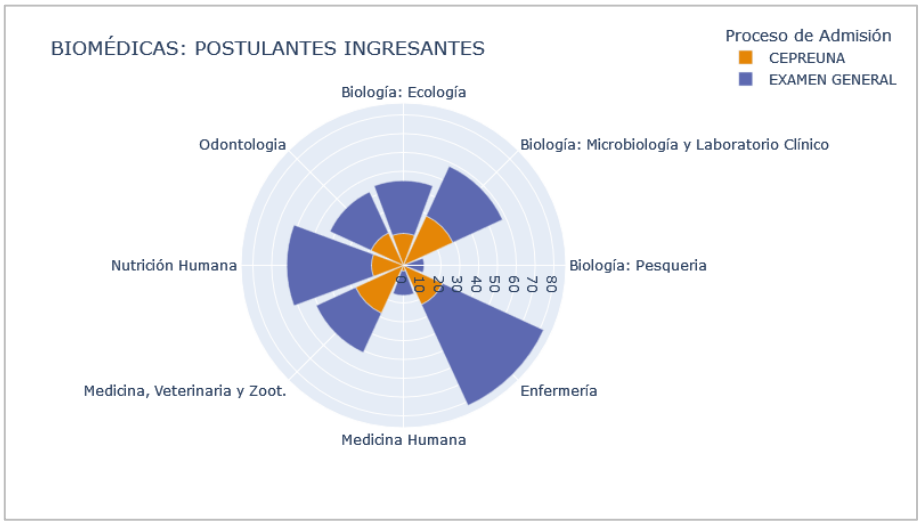


Figura 4. Postulantes ingresantes del área de Biomédicas.

La Figura 4 muestra los datos organizados de los ingresantes a las escuelas profesionales del área de Biomédicas en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor número de ingresantes son: Enfermería, Nutrición Humana, Biología: Microbiología y Laboratorio Clínico.

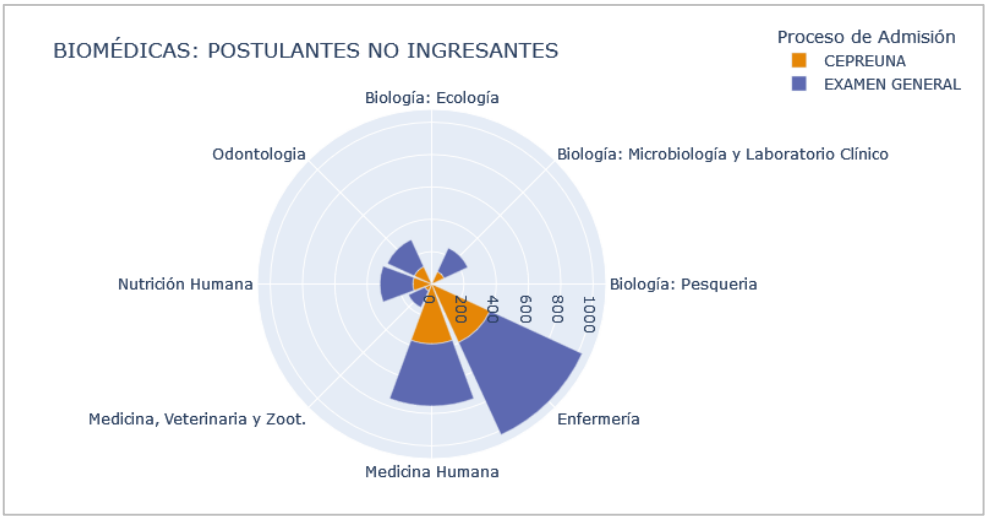


Figura 5. Postulantes no ingresantes del área de Biomédicas.

La Figura 5 muestra los datos organizados de los postulantes que no lograron su ingreso a las escuelas profesionales del área de Biomédicas en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor demanda son: Enfermería y Medicina Humana.

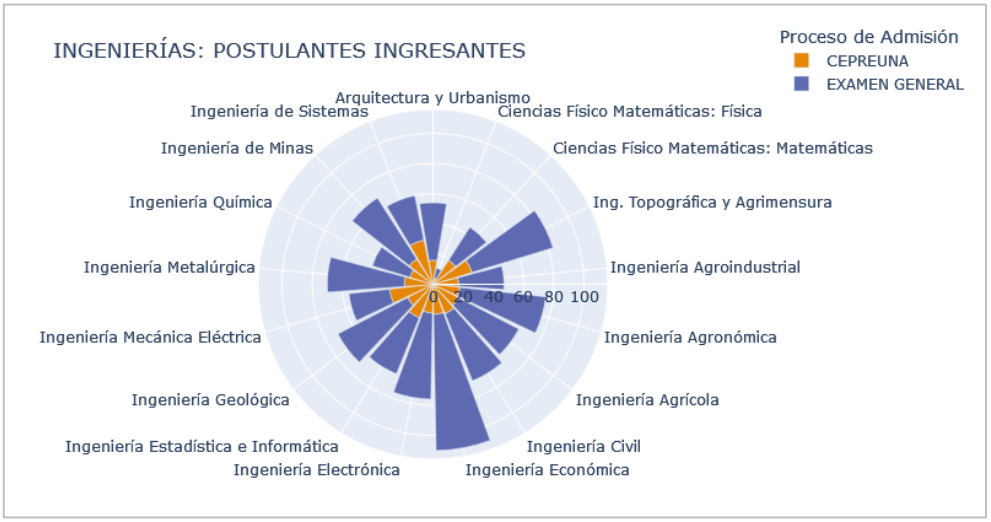


Figura 6. Postulantes ingresantes del área de Ingenierías.

La Figura 6 muestra los datos organizados de los ingresantes a las escuelas profesionales del área de Ingenierías en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor número de ingresantes son: Ingeniería Económica, Ingeniería Topográfica y Agrimensura, Ingeniería Agronómica y algunos más.

Figura 7. Postulantes no ingresantes del área de Ingenierías.

La Figura 7 muestra los datos organizados de los postulantes que no lograron su ingreso a las escuelas profesionales del área de Ingenierías en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor demanda son: Ingeniería Civil, Ingeniería de Sistemas e Ingeniería Económica.

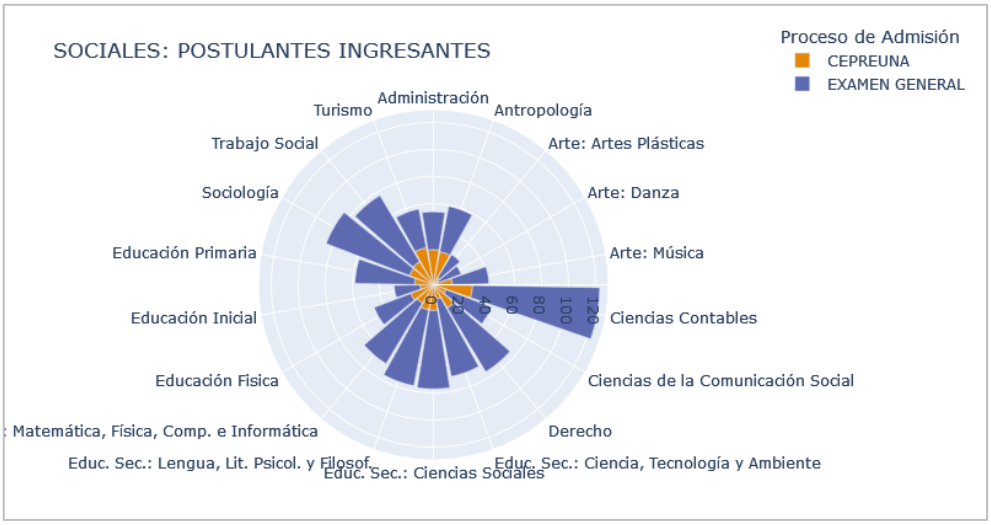


Figura 8. Postulantes ingresantes del área de Sociales.

La Figura 8 muestra los datos organizados de los ingresantes a las escuelas profesionales del área de Sociales en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor número de ingresantes son: Ciencias Contables, Sociología y Trabajo Social.

Figura 9. Postulantes no ingresantes del área de Sociales.

La Figura 9 muestra los datos organizados de los postulantes que no lograron su ingreso a las escuelas profesionales del área de Sociales en los procesos de admisión Examen General y CEPREUNA correspondiente al semestre 2022-II. En ello se observa que las escuelas profesionales con mayor demanda son: Derecho, Ciencias Contables y Administración.

Prever el impacto prescriptivo con datos organizados de los postulantes sobre el ingreso a la universidad es un paso fundamental en el contexto de la ciencia de datos, así lo demuestran el análisis visual de las Figuras 4, 5, 6, 7, 8 y 9 donde se visualiza gráficamente y con precisión alta el comportamiento inicial de los datos de los postulantes ingresantes y no ingresantes a la universidad.

2. Resultados para el segundo objetivo

En esta sección se analiza las salidas obtenidas por el modelo de aprendizaje supervisado de bosques aleatorios para las diferentes áreas y procesos de admisión.

Área biomédicas: proceso de admisión CEPREUNA

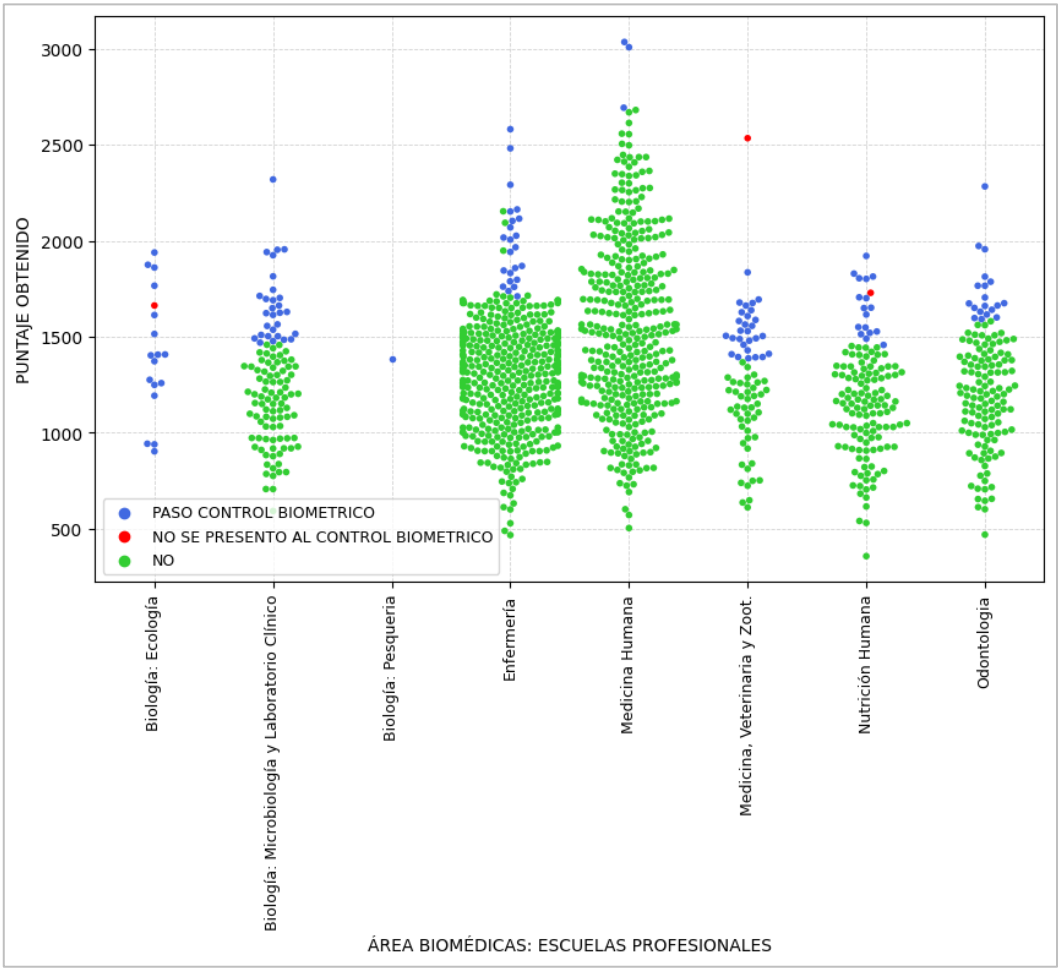


Figura 10. Postulantes del proceso de admisión CEPREUNA 2022-II, área de biomédicas. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 10 se visualiza que las escuelas profesionales con mayor demanda son: Enfermería y Medicina Humana y en las mismas, los postulantes obtuvieron los mayores puntajes. Por otro lado, la escuela profesional de Biología en sus especialidades de Ecología y Pesquería tienen muy poca demanda, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante.

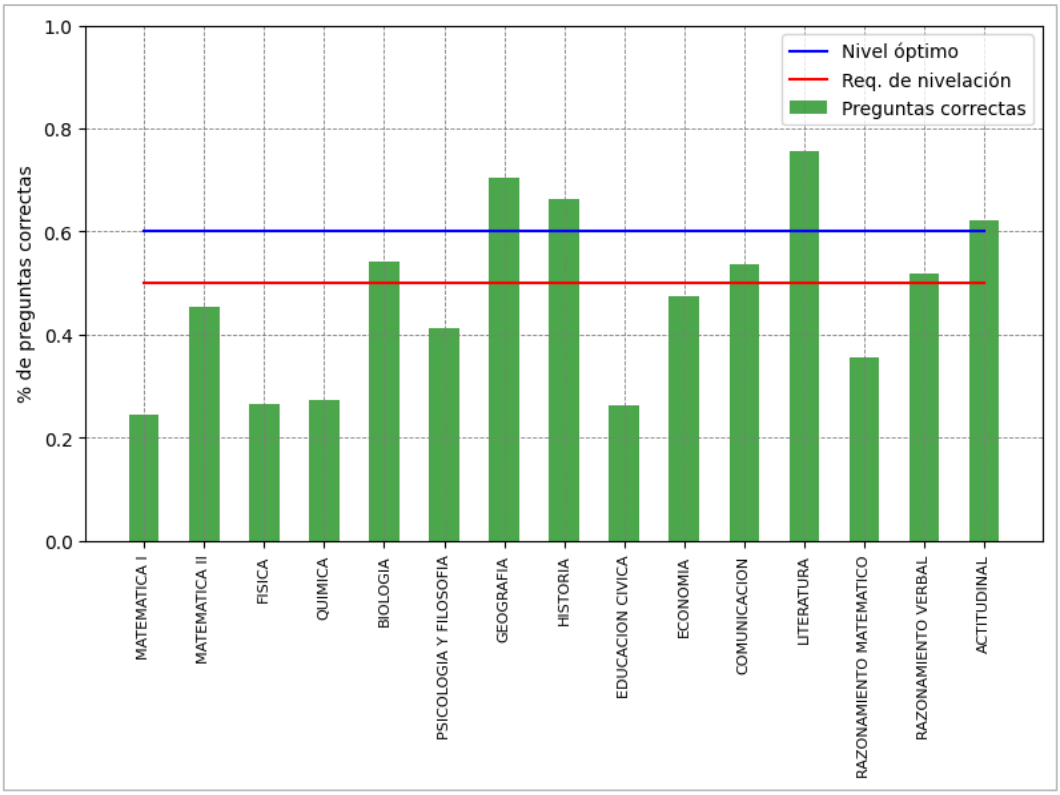


Figura 11. Porcentaje de preguntas contestadas correctamente por los ingresantes CEPREUNA 2022-II área biomédicas. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 11 resume el desempeño de los ingresantes del proceso de admisión CEPREUNA 2022-II área biomédicas, en ello se puede apreciar que los ingresantes tienen un buen desempeño en las asignaturas de geografía, historia, literatura y actitudinal. Un desempeño regular en las asignaturas de biología, comunicación y razonamiento verbal. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 11 corresponde a los resultados de todos los ingresantes del área de biomédicas de la modalidad CEPREUNA, será muy distinto al momento de mostrar la gráfica por escuelas profesionales.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión CEPREUNA 2022-II, área de Biomédicas, obteniéndose los siguientes resultados.

Escenario Real	Predicción: No ingresa	Predicción: Si Ingresa
No Ingresa	283	1
Si Ingresa	31	4

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 89.97 %

Tabla 2

Predicciones del área de Biomédicas CEPRE-UNA 2022-II

	Función objetivo	precision	recall	f1-score	support
0	(No ingresó)	0.90	1.00	0.95	284
1	(Ingresó)	0.80	0.11	0.20	35
Total de datos de prueba (test)					**319**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del CEPREUNA 2022-II área de biomédicas a la universidad.

1. Precisión = 80%, nos indica que el modelo puede equivocarse un 20% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{4}{31+4}$ = 11%, significa que el 11% del total de los postulantes lograron su ingreso, por lo tanto, el modelo puede identificar a 1 de cada 9 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.80\ x\ 0.11}{0.80+0.11}\right)$ = 0.2, es un indicador que tanto la precisión y la exhaustividad tienen la misma importancia.

4. Exactitud (Accuracy) = $\frac{4+283}{4+283+1+31}$ = 89.96%, el modelo tiene un acierto del 90% del total de predicciones realizadas.

Área biomédicas: proceso de admisión Examen General

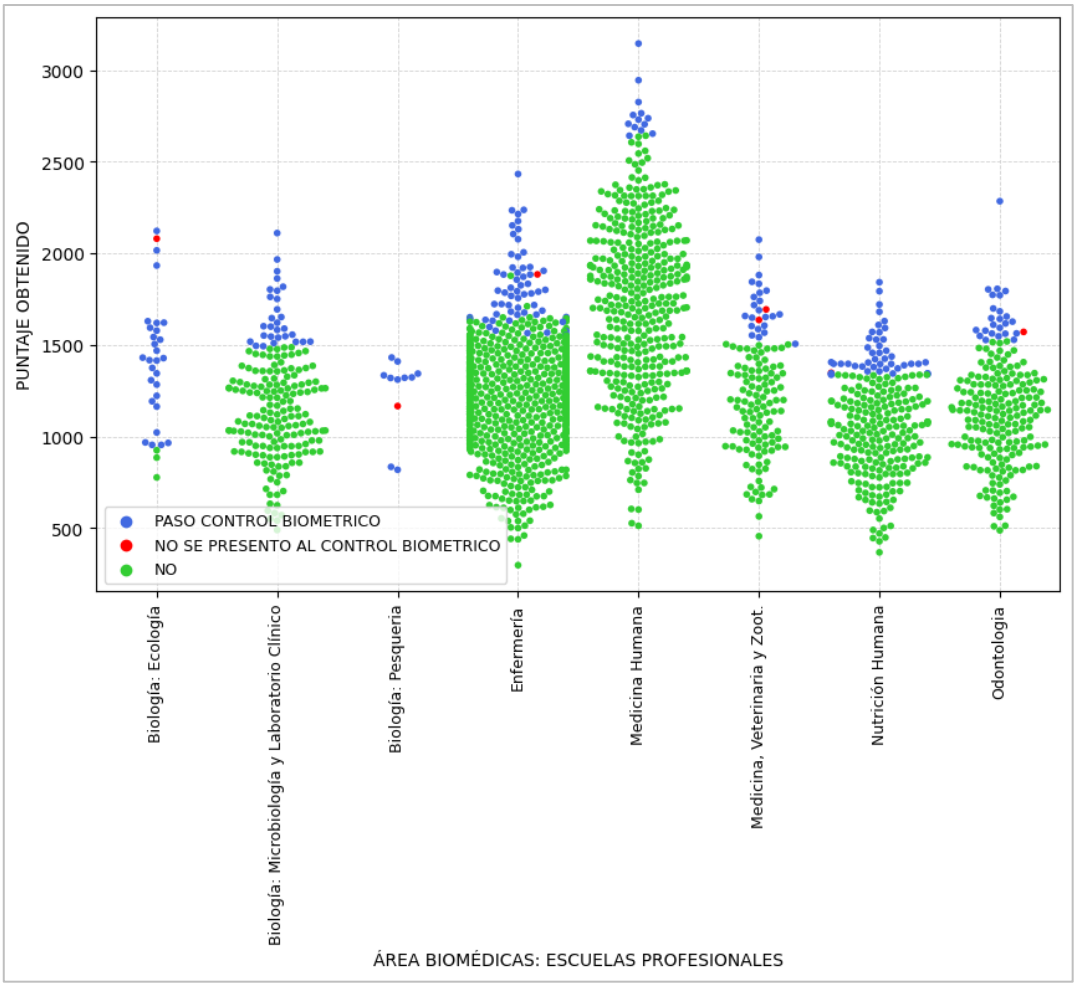


Figura 12. Postulantes del proceso de admisión Examen General 2022-II, área de biomédicas. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 12 se visualiza que las escuelas profesionales con mayor demanda son: Enfermería y Medicina Humana y en las mismas, los postulantes obtuvieron los mayores puntajes. Por otro lado, la escuela profesional de Biología en su especialidad de Pesquería tiene muy poca demanda, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante. Siendo los ingresantes a la escuela profesional de Medicina Humana los que obtuvieron altos puntajes.

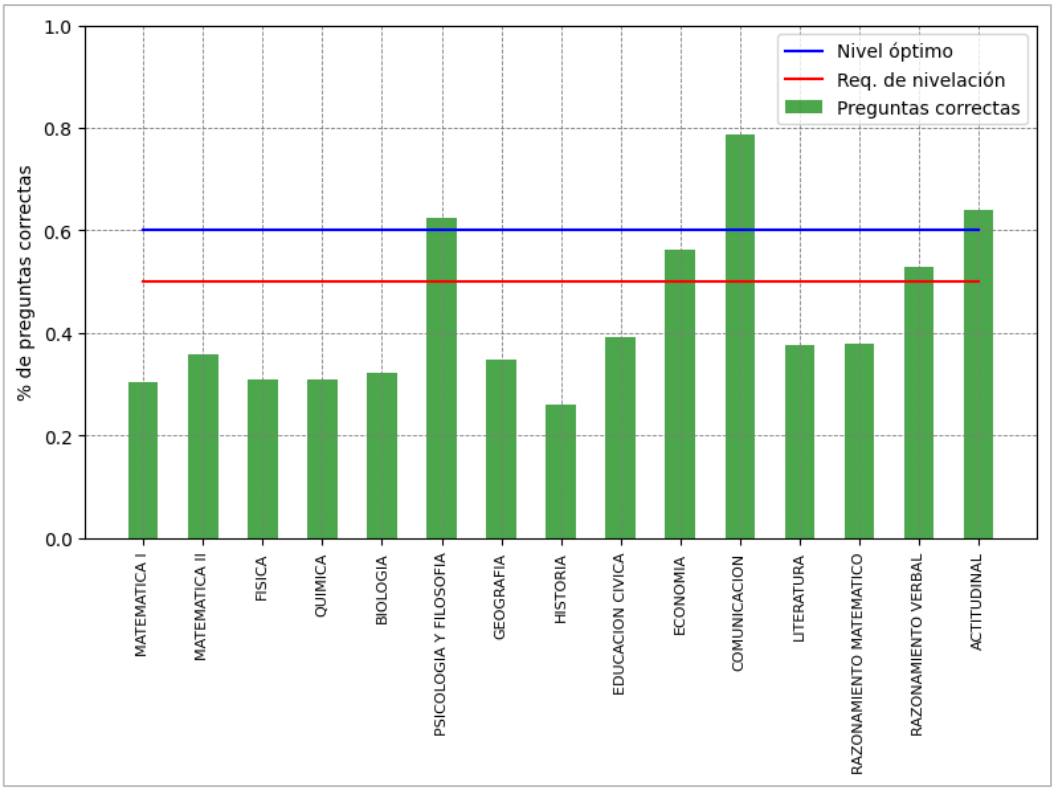


Figura 13. Porcentaje de preguntas contestadas correctamente por los ingresantes del Examen General 2022-II área biomédicas. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 13 resume el desempeño de los ingresantes del proceso de admisión Examen General 2022-II área biomédicas, en ello se puede apreciar que los ingresantes tienen un buen desempeño en las asignaturas de comunicación, actitudinal, psicología y filosofía. Un desempeño regular en las asignaturas de economía y razonamiento verbal. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 13 corresponde a los resultados de todos los ingresantes del área de biomédicas de la modalidad Examen General, será muy distinto al momento de mostrar la gráfica por escuelas profesionales como en la figura 12.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión Examen General 2022-II, área de biomédicas, obteniéndose los siguientes resultados.

Escenario Real	Predicción: No ingresa	Predicción: Si Ingresa
No Ingresa	412	4
Si Ingresa	50	11

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 88.68 %

Tabla 3

Predicciones del área de Biomédicas Examen General 2022-II

	Función objetivo	precision	recall	f1-score	support
0	(No ingresó)	0.89	0.99	0.94	416
1	(Ingresó)	0.73	0.18	0.29	61
Total de datos de prueba (test)					**477**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del Examen General 2022-II área de biomédicas a la universidad.

1. Precisión = 73%, nos indica que el modelo puede equivocarse un 27% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{11}{50+11}$ = 0.18, significa que el 18% del total de los postulantes lograron su ingreso, por tanto, el modelo puede identificar a 1 de cada 6 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.73\ x\ 0.18}{0.73+0.18}\right)$ = 0.29, es un indicador que tanto la precisión y la exhaustividad tienen la misma importancia.

4. Exactitud (Accuracy) = $\frac{11+412}{11+412+4+50}$ = 88.68%, el modelo tiene un acierto del 90% del total de predicciones realizadas.

Área ingenierías: proceso de admisión CEPREUNA

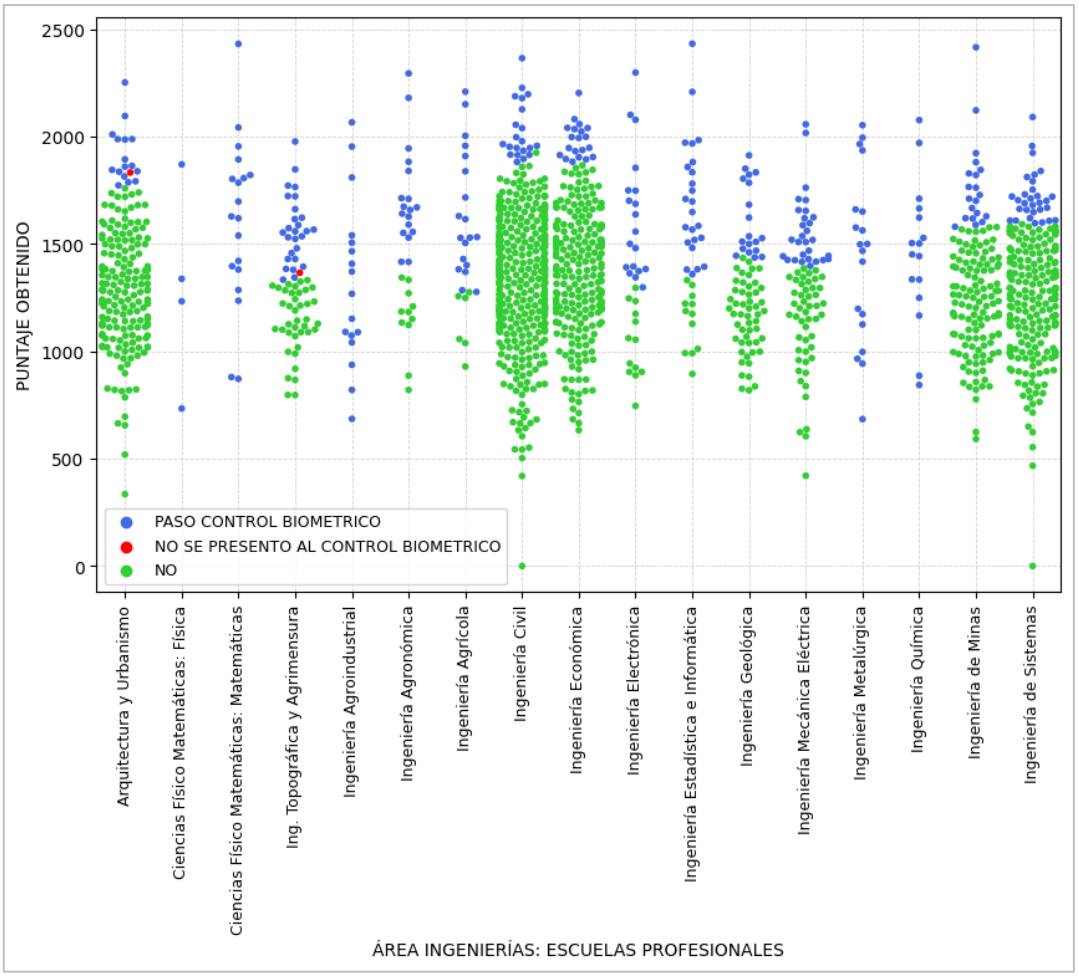


Figura 14. Postulantes del proceso de admisión CEPREUNA 2022-II, área de ingenierías. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 14 se visualiza que las escuelas profesionales con mayor demanda son: Ingeniería Civil, Ingeniería Económica e Ingeniería de Sistemas y en las escuelas profesionales de Ciencias Físico-Matemáticas: Matemáticas, Ingeniería Estadística e Informática e Ingeniería de Minas, los postulantes obtuvieron los mayores puntajes. Por otro lado, Ciencias Físico-Matemáticas: Física tiene muy poca demanda, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante. Siendo los ingresantes a la escuela profesional de Ingeniería Civil los que obtuvieron altos puntajes.

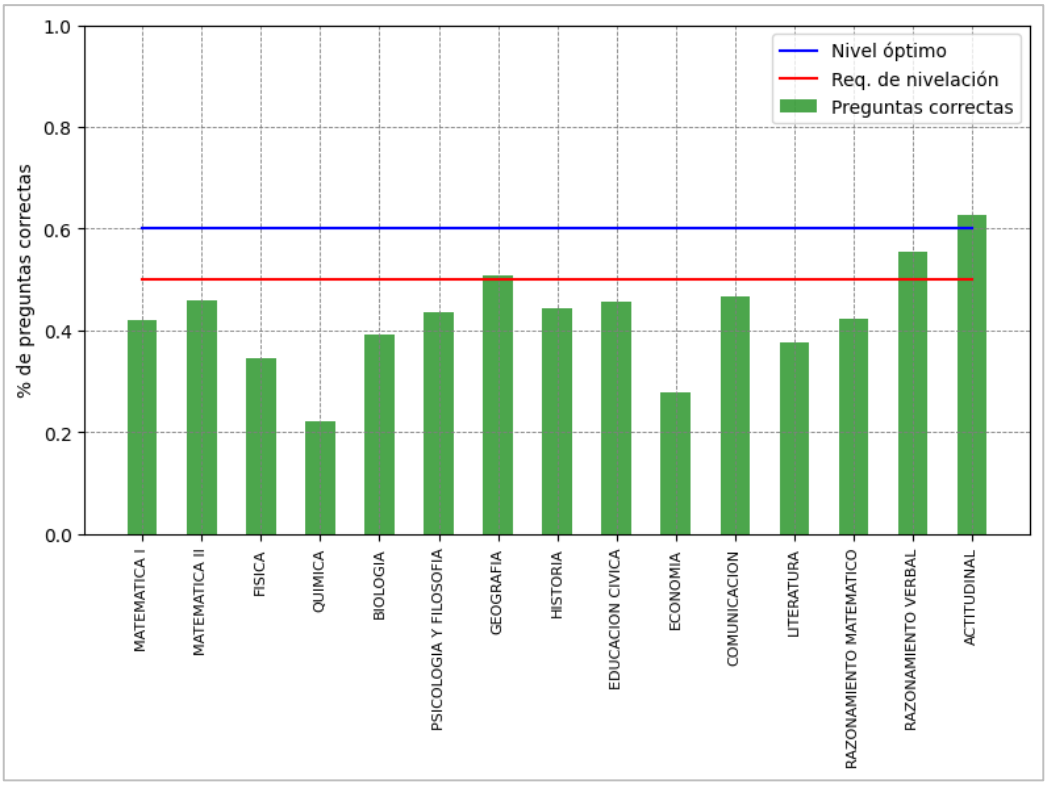


Figura 15. Porcentaje de preguntas contestadas correctamente por los ingresantes CEPREUNA 2022-II área ingenierías. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 15 resume el desempeño de los ingresantes del proceso de admisión CEPREUNA 2022-II área ingenierías, en ello se puede apreciar que los ingresantes tienen un buen desempeño en la asignatura de actitudinal. Un desempeño regular en las asignaturas de geografía y razonamiento verbal. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 15 corresponde a los resultados de todos los ingresantes del área de ingenierías de la modalidad CEPREUNA, será muy distinto al momento de mostrar la gráfica por escuelas profesionales muy similar a los resultados de la figura 14.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión CEPREUNA 2022-II, área de ingenierías, obteniéndose los siguientes resultados.

		Predicción	
		No ingresa	Si Ingresa
Escenario Real	No Ingresa	284	8
	Si Ingresa	60	40

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 82.65 %

Tabla 4

Predicciones del área de Ingenierías CEPRE-UNA 2022-II

Función objetivo	precision	recall	F1-score	support
0 (No ingresó)	0.83	0.97	0.89	292
1 (Ingresó)	0.83	0.40	0.54	100
Total de datos de prueba (test)				**392**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del CEPREUNA 2022-II área de ingenierías a la universidad.

1. Precisión = 83%, nos indica que el modelo puede equivocarse un 17% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{40}{40+60}$ = 40%, significa que el 40% del total de los postulantes lograron su ingreso, por lo tanto, el modelo puede identificar a 1 de cada 3 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.83\ x\ 0.40}{0.83\ +\ 0.40}\right)$ = 0.54, es un indicador que tanto la precisión y la exhaustividad tienen una relativa dependencia de los resultados del modelo.

4. Exactitud (Accuracy) = $\frac{40\ +\ 284}{40\ +\ 60\ +\ 8\ +\ 284}$ = 82.65%, el modelo tiene un acierto del 83% del total de predicciones realizadas.

Área ingenierías: proceso de admisión Examen General

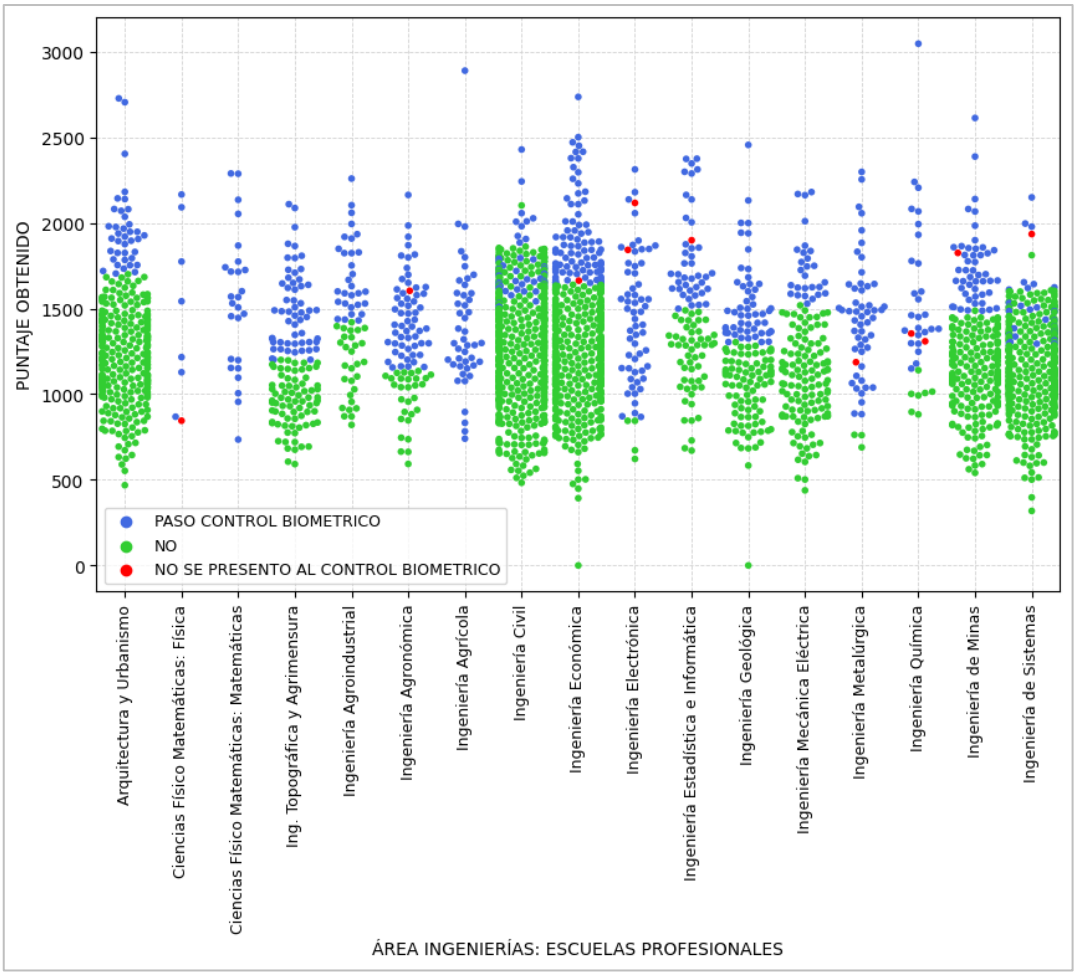


Figura 16. Postulantes del proceso de admisión Examen General 2022-II, área de ingenierías. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 16 se visualiza que las escuelas profesionales con mayor demanda son: Ingeniería Civil, Ingeniería Económica e Ingeniería de Sistemas y en las escuelas profesionales de Ingeniería Química, Ingeniería Agrícola e Ingeniería Económica, los postulantes obtuvieron los mayores puntajes. Por otro lado, Ciencias Físico-Matemáticas: Física tiene muy poca demanda, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante. Siendo los ingresantes a la escuela profesional de Ingeniería Económica los que obtuvieron altos puntajes.

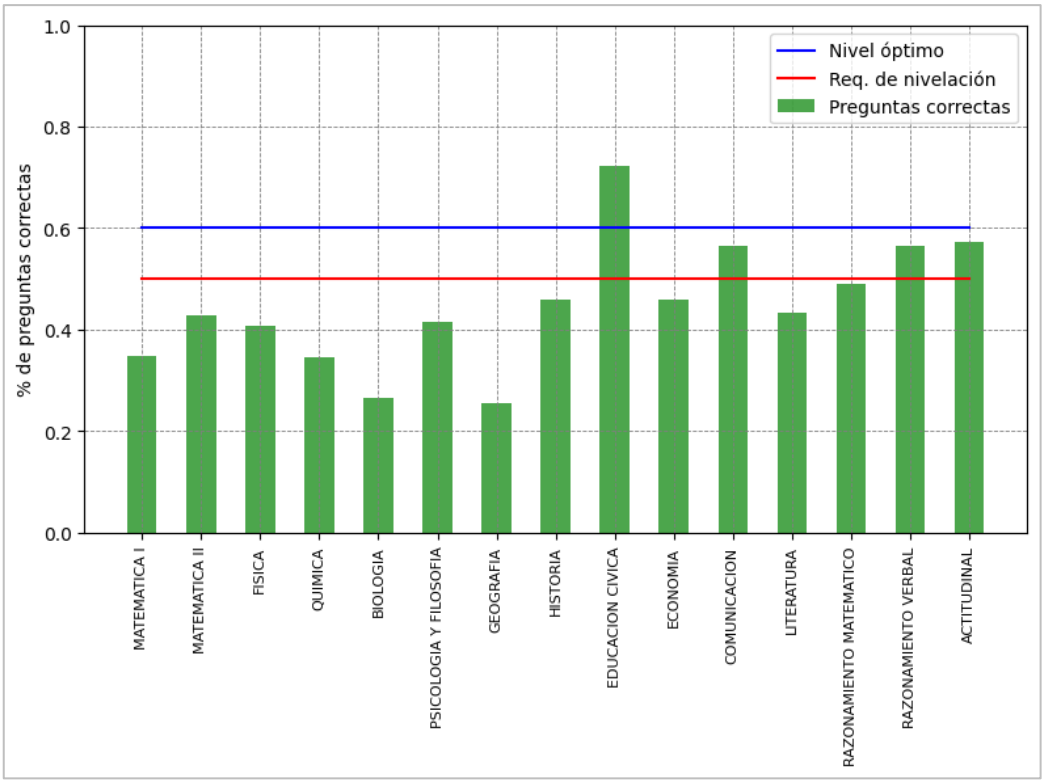


Figura 17. Porcentaje de preguntas contestadas correctamente por los ingresantes del Examen General 2022-II área ingenierías. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 17 resume el desempeño de los ingresantes del proceso de admisión Examen General 2022-II área ingenierías, en ello se puede apreciar que los ingresantes tienen un buen desempeño en la asignatura de educación cívica. Un desempeño regular en las asignaturas de actitudinal, comunicación y razonamiento verbal. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 17 corresponde a los resultados de todos los ingresantes del área de ingenierías de la modalidad Examen General, será muy distinto al momento de mostrar la gráfica individual por escuelas profesionales muy similar a los resultados de la figura 16.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión Examen General 2022-II, área de ingenierías, obteniéndose los siguientes resultados.

Escenario Real	Predicción: No ingresa	Predicción: Si Ingresa
No Ingresa	498	29
Si Ingresa	79	102

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 84.74 %

Tabla 5

Predicciones del área de Ingenierías Examen General 2022-II

Función objetivo	**precision**	**recall**	**f1-score**	**support**
0 (No ingresó)	0.86	0.94	0.90	527
1 (Ingresó)	0.78	0.56	0.65	181
Total de datos de prueba (test)				**708**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del Examen General 2022-II área de ingenierías a la universidad.

1. Precisión = 78%, nos indica que el modelo puede equivocarse un 22% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{79}{79+102}$ = 0.43, significa que el 43% del total de los postulantes lograron su ingreso, por lo tanto, el modelo puede identificar a 1 de cada 3 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.78\ x\ 0.56}{0.78\ +\ 0.56}\right)$ = 0.65, es un indicador que tanto la precisión y la exhaustividad tienen una dependencia de los resultados del modelo.

4. Exactitud (Accuracy) = $\frac{102+\ 498}{498+29+79+102}$ = 0.8474, el modelo tiene un acierto del 84.74% del total de predicciones realizadas.

Área sociales: proceso de admisión CEPREUNA

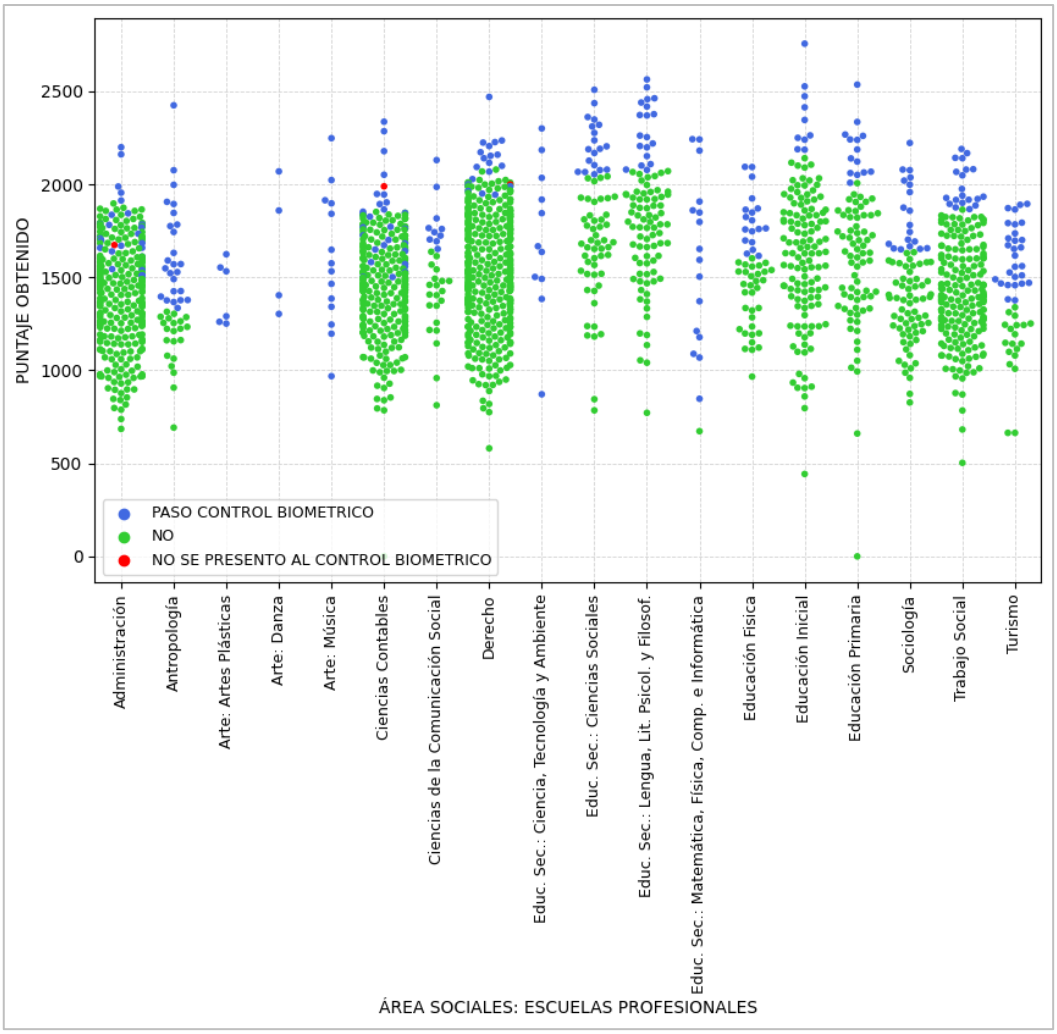


Figura 18. Postulantes del proceso de admisión CEPREUNA 2022-II, área de sociales. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 18 se visualiza que las escuelas profesionales con mayor demanda son: Derecho, Administración y Ciencias Contables. En las escuelas profesionales de Educación Inicial, Educación Primaria y Educación Secundaria en su especialidad de Lengua Literatura, Psicología y Filosofía, los postulantes obtuvieron los mayores puntajes. Por otro lado, Las escuelas profesionales de Arte en sus especialidades Danza y Artes Plásticas tienen muy poca demanda de postulantes, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante.

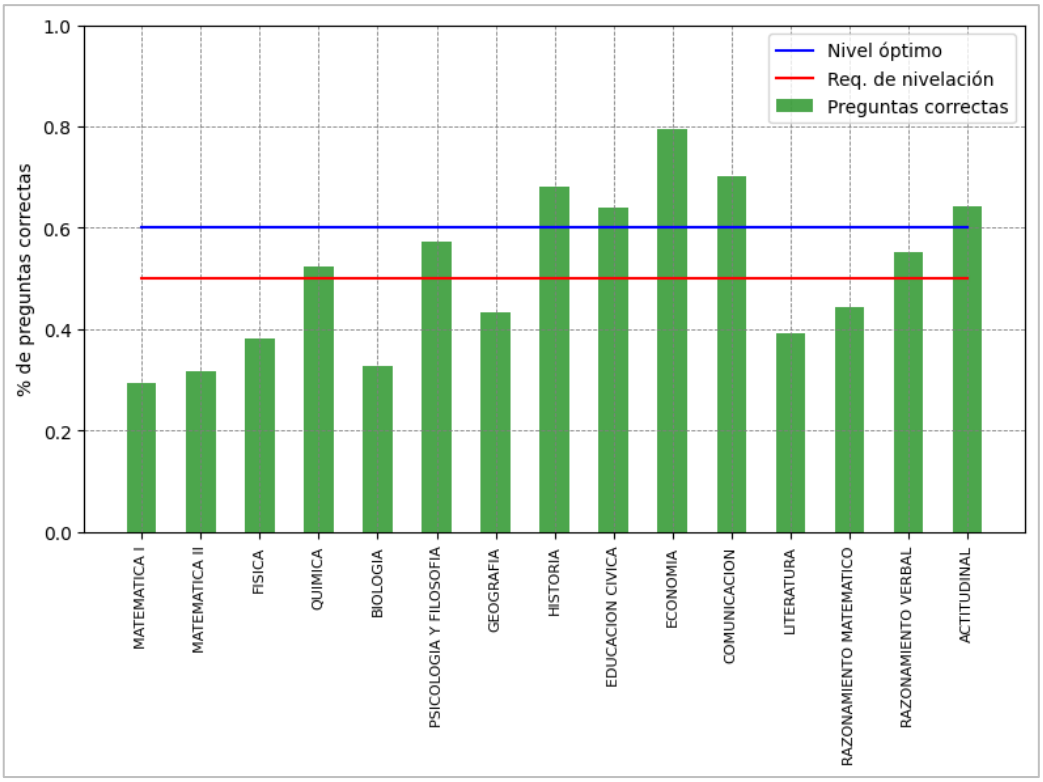


Figura 19. Porcentaje de preguntas contestadas correctamente por los ingresantes CEPREUNA 2022-II área sociales. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 19 resume el desempeño de los ingresantes del proceso de admisión CEPREUNA 2022-II área sociales, en ello se puede apreciar que los ingresantes tienen un buen desempeño en las asignaturas de economía, comunicación, historia, educación cívica y actitudinal. Regular en las asignaturas de química, psicología y filosofía, razonamiento verbal. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 19 corresponde a los resultados de todos los ingresantes del área de sociales de la modalidad CEPREUNA, será muy distinto al momento de mostrar la gráfica individual por escuelas profesionales, muy similar a los resultados de la figura 18.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión CEPREUNA 2022-II, área de sociales, obteniéndose los siguientes resultados.

Escenario Real	Predicción: No ingresa	Predicción: Si Ingresa
No Ingresa	390	3
Si Ingresa	41	23

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 90.37 %

Tabla 6

Predicciones del área de Sociales CEPRE-UNA 2022-II

Función objetivo	**precision**	**recall**	**f1-score**	**support**
0 (No ingresó)	0.90	0.99	0.95	393
1 (Ingresó)	0.88	0.36	0.51	64
Total de datos de prueba (test)				**457**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del CEPREUNA 2022-II área de sociales a la universidad.

1. Precisión = 88%, nos indica que el modelo puede equivocarse un 12% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{23}{23+41}$ = 0.36, significa que el 36% del total de los postulantes lograron su ingreso, por lo tanto, el modelo puede identificar a 1 de cada 3 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.88\ x\ 0.36}{0.88+0.36}\right)$ = 0.52, es un indicador que tanto la precisión y la exhaustividad dependen relativamente de los resultados del modelo.

4. Exactitud (Accuracy) = $\frac{23+390}{23+390+3+41}$ = 90.37%, el modelo tiene un acierto del 90.37% del total de predicciones realizadas.

Área sociales: proceso de admisión Examen General

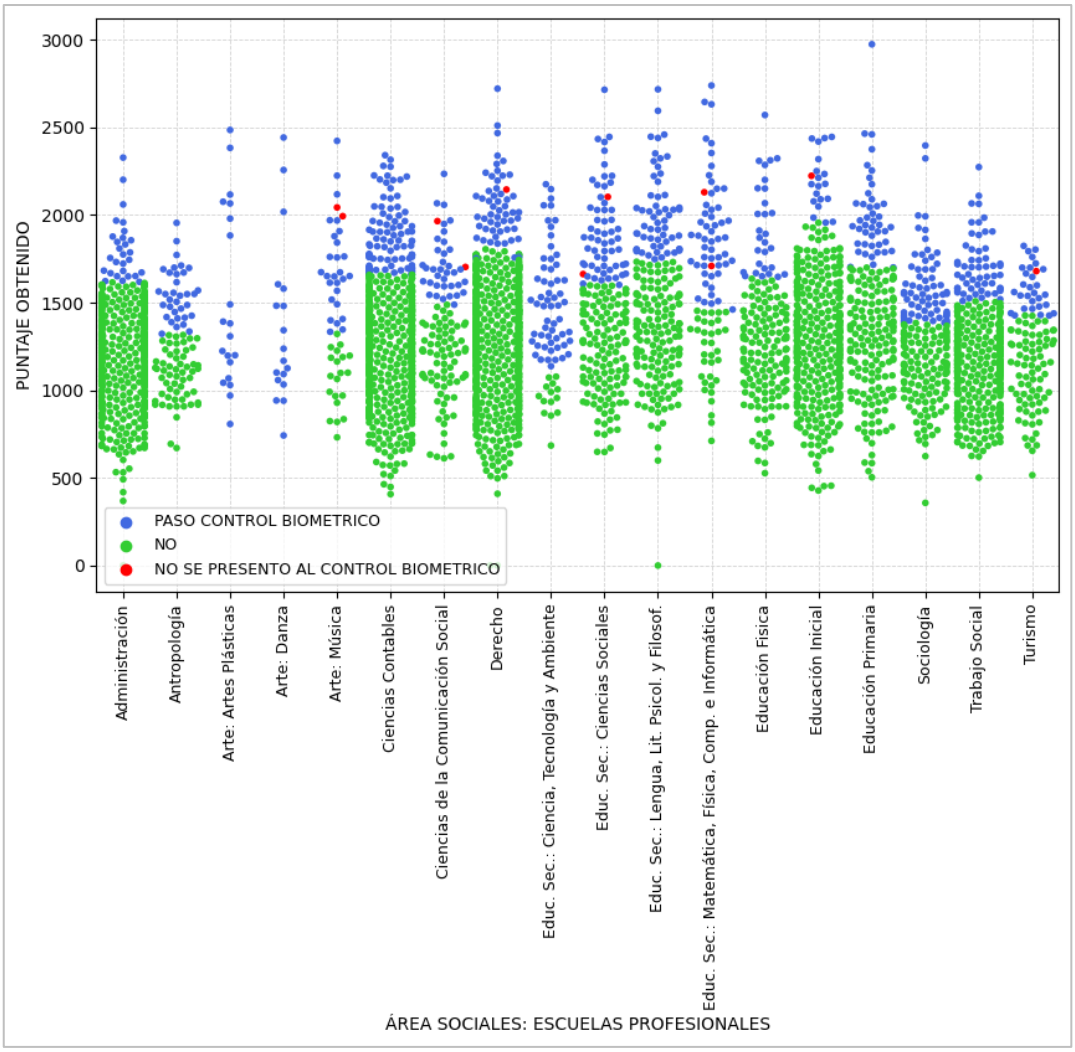


Figura 20. Postulantes del proceso de admisión Examen General 2022-II, área de sociales. a) ingresantes de color azul, b) no ingresantes de color verde y c) postulantes aptos que no lograron presentarse al control biométrico.

En la figura 20 se visualiza que las escuelas profesionales con mayor demanda son: Derecho, Administración, Ciencias Contables y Educación Inicial. En las escuelas profesionales de Educación Primaria, Derecho y Educación Secundaria en su especialidad de Matemática, Física, Computación e Informática, los postulantes obtuvieron los mayores puntajes. Por otro lado, Las escuelas profesionales de Arte en sus especialidades Danza y Artes Plásticas tienen muy poca demanda de postulantes, razón por la cual no se visualiza postulantes que no hayan alcanzado una vacante.

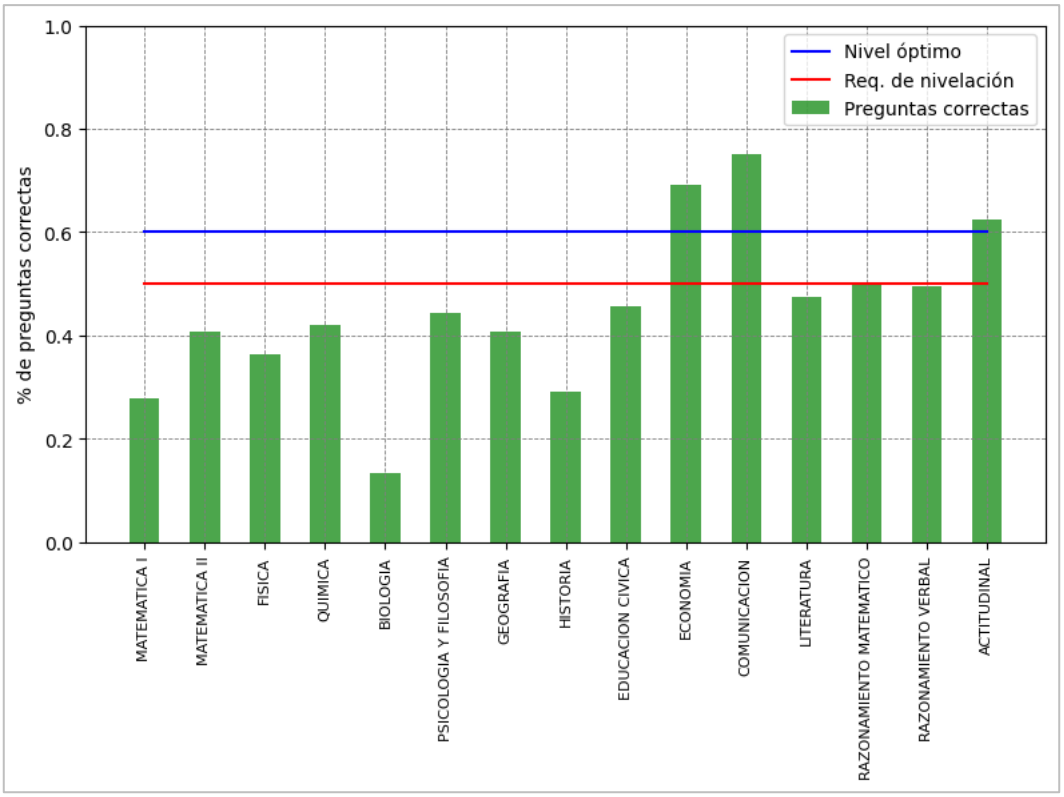


Figura 21. Porcentaje de preguntas contestadas correctamente por los ingresantes del Examen General 2022-II área sociales. a) la línea azul representa el mínimo óptimo, b) la línea roja representa el nivel mínimo de aprobación.

La figura 21 resume el desempeño de los ingresantes del proceso de admisión Examen General 2022-II área sociales, en ello se puede apreciar que los ingresantes tienen un buen desempeño en la asignatura de economía, comunicación y actitudinal. Un desempeño regular en la asignatura de razonamiento matemático. Bajo desempeño en las demás asignaturas.

También se debe considerar que los resultados de la figura 21 corresponde a los resultados de todos los ingresantes del área de sociales de la modalidad Examen General, será muy distinto al momento de mostrar la gráfica individual por escuelas profesionales, muy similar a los resultados de la figura 20.

Matriz de confusión

La siguiente matriz de confusión permite evaluar el rendimiento del modelo de clasificación propuesto para el proceso de admisión Examen General 2022-II, área de sociales, obteniéndose los siguientes resultados.

Escenario Real	Predicción: No ingresa	Predicción: Si Ingresa
No Ingresa	746	24
Si Ingresa	85	110

El porcentaje de aciertos (accuracy) del modelo de bosques aleatorios luego de optimizar los hiperparámetros es: 88.70 %

Tabla 7

Predicciones del área de Sociales Examen General 2022-II

Función objetivo	**precision**	**recall**	**f1-score**	**support**
0 (No ingresó)	0.90	0.97	0.93	770
1 (Ingresó)	0.82	0.56	0.67	195
Total de datos de prueba (test)				**965**

Interpretación de las métricas obtenidas por los bosques aleatorios para la clasificación y determinación de los factores que permiten el ingreso de los postulantes del Examen General 2022-II área de sociales a la universidad.

1. Precisión = 82%, nos indica que el modelo puede equivocarse un 18% de las veces cuando prediga que un postulante ingresará a la universidad.

2. Exhaustividad (Recall) = $\frac{110}{110+85}$ = 0.56, significa que el 56% del total de los postulantes lograron su ingreso, por lo tanto, el modelo puede identificar a 1 de cada 2 postulantes que pueden lograr su ingreso a la universidad.

3. F1-score = $2\ x\left(\frac{0.82\ x\ 0.56}{0.82\ +\ 0.56}\right)$ = 0.67, es un indicador que tanto la precisión y la exhaustividad dependen en gran medida de las predicciones del modelo.

4. Exactitud (Accuracy) = $\frac{746\ +\ 110}{746\ +\ 371\ +\ 85\ +\ 24}$ = 88.70%, el modelo tiene un acierto del 88.70% del total de predicciones realizadas.

3. Resultados para el tercer objetivo

En esta sección identificamos los factores más determinantes en el ingreso de un postulante a la universidad, para una mayor versatilidad en su comprensión lo separamos por áreas y procesos de admisión.

Área biomédicas: proceso de admisión CEPREUNA

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Biomédicas, proceso de admisión CEPREUNA.

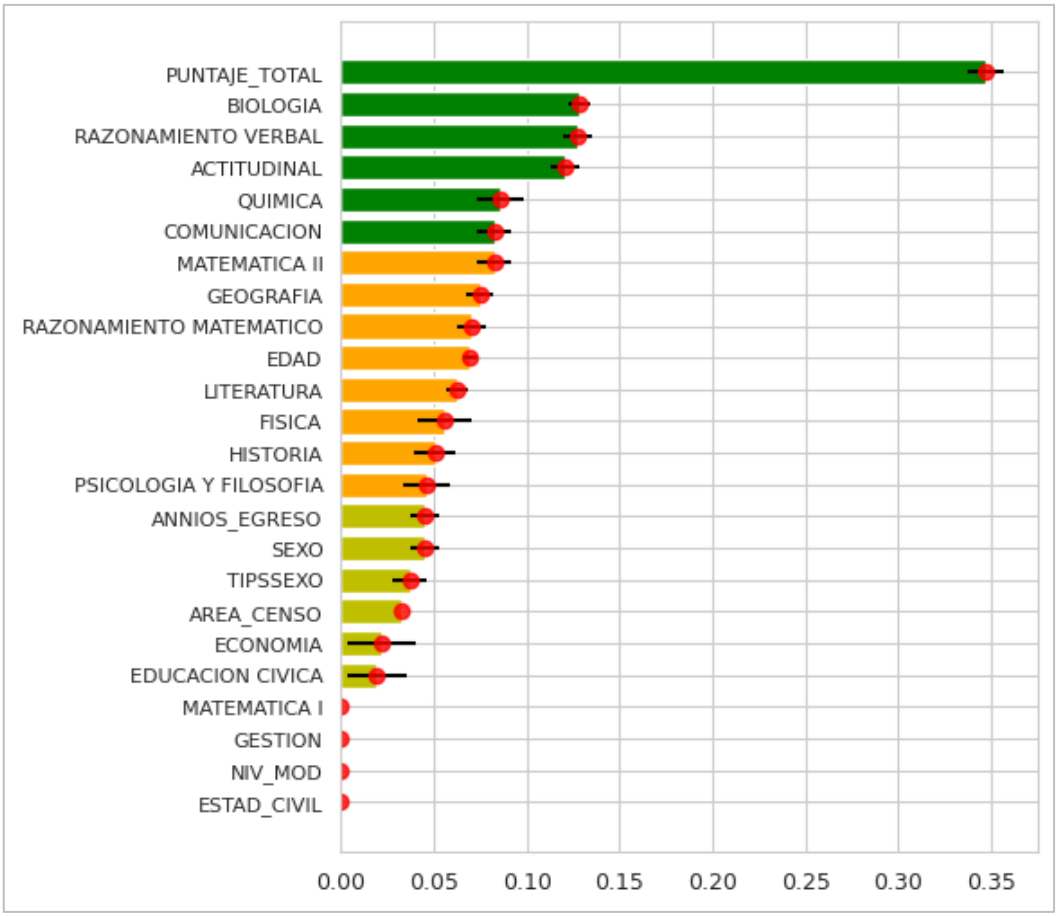


Figura 22. Importancia de los factores de ingreso a la universidad de los postulantes del CEPREUNA 2022-II área de biomédicas.

De la figura 22, se deduce que los factores predictores: puntaje total obtenido, las asignaturas de biología, razonamiento verbal, actitudinal, química y comunicación, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de matemática II, geografía, razonamiento matemático, literatura, física, historia, psicología y filosofía, así mismo la edad del postulante son factores de influencia media. Los demás

factores tienen una influencia baja. Consecuentemente el estado civil, la gestión institucional (estatal o privado), la modalidad de la institución secundaria y la asignatura de matemática I no tienen influencia alguna en el ingreso a la universidad.

Área biomédicas: proceso de admisión Examen General

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Biomédicas, proceso de admisión Examen General.

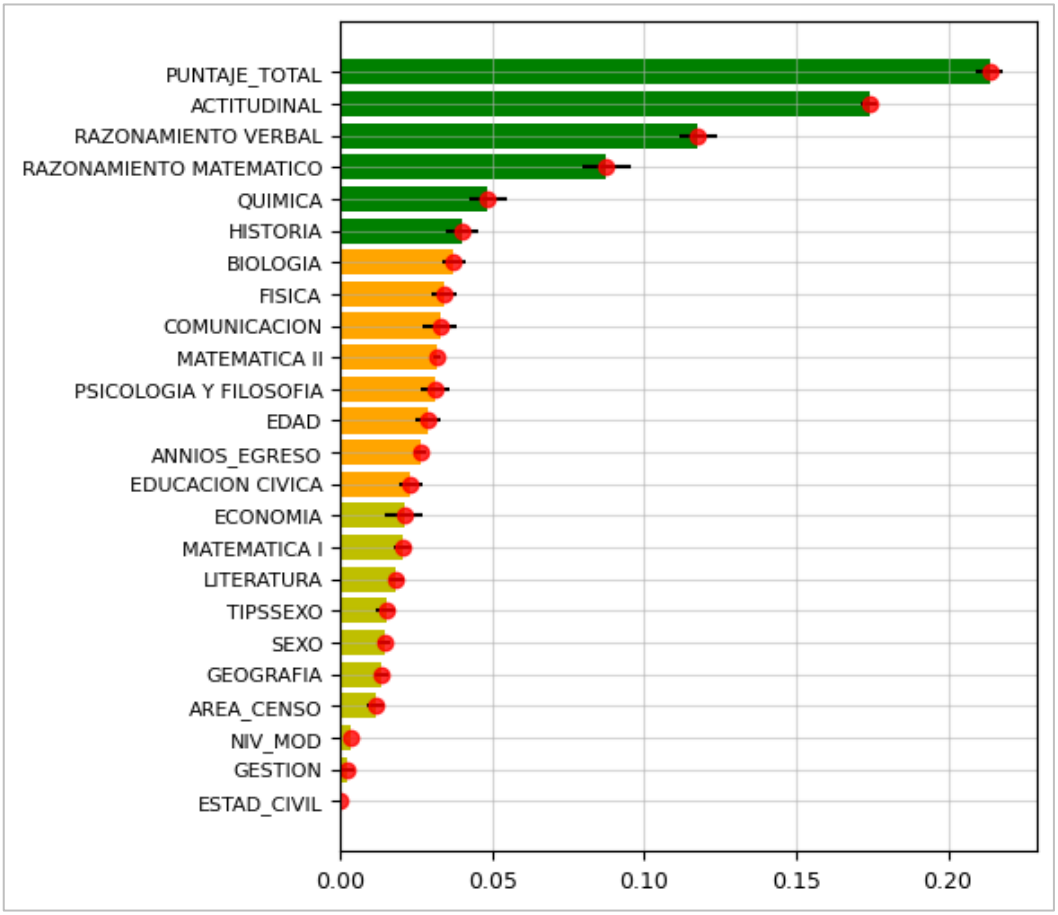


Figura 23. Importancia de los factores de ingreso a la universidad de los postulantes del Examen General 2022-II área de biomédicas.

De la figura 23, se deduce que los factores: puntaje total obtenido, las asignaturas de: actitudinal, razonamiento verbal, razonamiento matemático, química e historia, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de: biología, física, comunicación, matemática II, psicología y filosofía, educación cívica, así mismo la edad del postulante y los años de egresado son factores de influencia media.

Los demás factores tienen una influencia baja. Consecuentemente el estado civil, la gestión institucional (estatal o privado), la modalidad de la institución secundaria no tiene influencia alguna o es muy mínima en el ingreso a la universidad.

Área ingenierías: proceso de admisión CEPREUNA

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Ingenierías, proceso de admisión CEPREUNA.

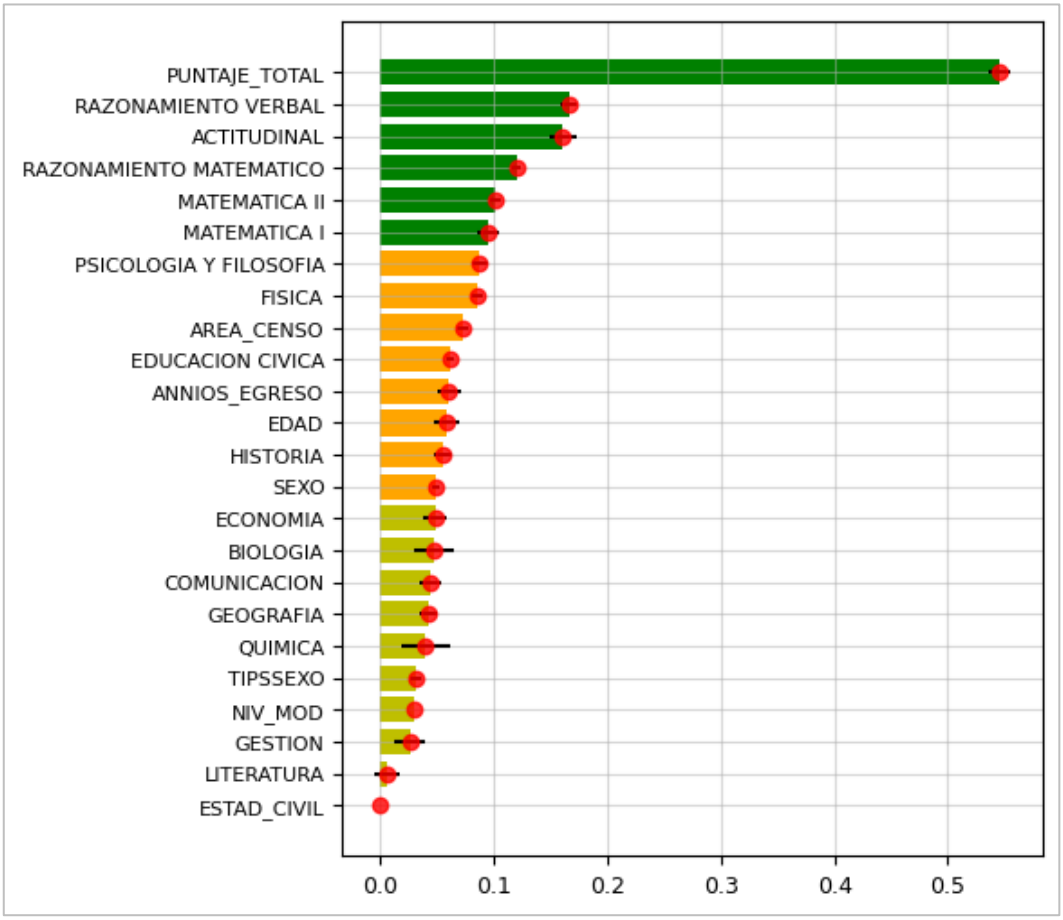


Figura 24. Importancia de los factores de ingreso a la universidad de los postulantes del CEPREUNA 2022-II área de ingenierías

De la figura 24, se deduce que los factores: puntaje total obtenido, las asignaturas de: razonamiento verbal, aptitudinal, razonamiento matemático, matemática II y matemática I, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de: psicología y filosofía, física, educación cívica e historia, así mismo el área geográfica (rural y urbano), la edad del postulante, los años de egresado y sexo del postulante son factores de influencia media. Los demás factores tienen una influencia baja o nula.

Consecuentemente el estado civil, la asignatura de literatura no tiene influencia alguna o es muy mínima en el ingreso a la universidad.

Área ingenierías: proceso de admisión Examen General

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Ingenierías, proceso de admisión Examen General.

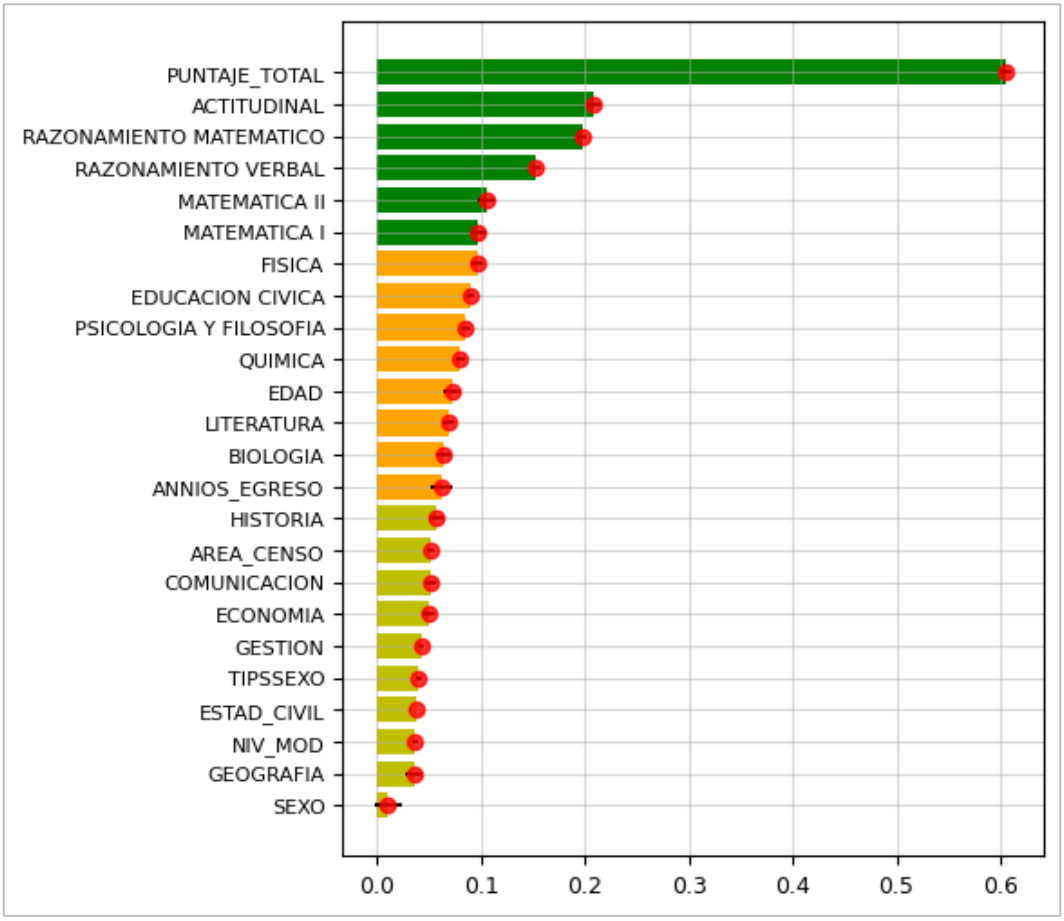


Figura 25. Importancia de los factores de ingreso a la universidad de los postulantes del Examen General 2022-II área de ingenierías.

De la figura 25, se deduce que los factores: puntaje total obtenido, las asignaturas de: actitudinal, razonamiento matemático, razonamiento verbal, matemática II y matemática I, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de: física, educación cívica, psicología y filosofía, química, literatura y biología, así mismo la edad del postulante y los años de egresado son factores de influencia media.

Los demás factores tienen una influencia baja. Consecuentemente el sexo no tiene influencia alguna o es muy mínima en el ingreso a la universidad.

Área sociales: proceso de admisión CEPREUNA

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Sociales, proceso de admisión CEPREUNA.

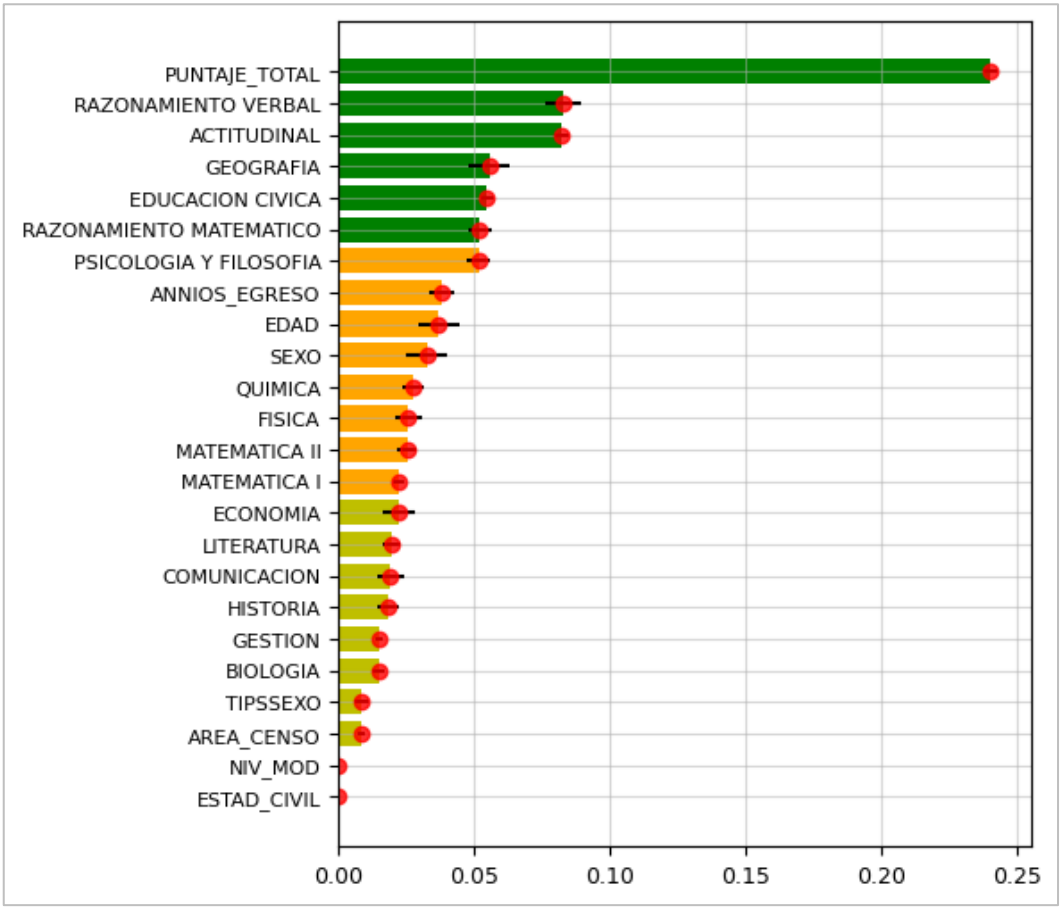


Figura 26. Importancia de los factores de ingreso a la universidad de los postulantes del CEPREUNA 2022-II área de sociales.

De la figura 26, se deduce que los factores: puntaje total obtenido, las asignaturas de: razonamiento verbal, aptitudinal, geografía, educación cívica y razonamiento matemático, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de: psicología y filosofía, química, física, matemática II y matemática I, así mismo los años de egresado, la edad del postulante y el sexo son factores de influencia media. Los demás factores tienen una influencia baja. Consecuentemente el estado civil

y modalidad de la institución secundaria no tiene influencia alguna o es muy mínima en el ingreso a la universidad.

Área sociales: proceso de admisión Examen General

Determinación de la importancia de los factores predictores de ingreso a la universidad del área de Sociales, proceso de admisión Examen General.

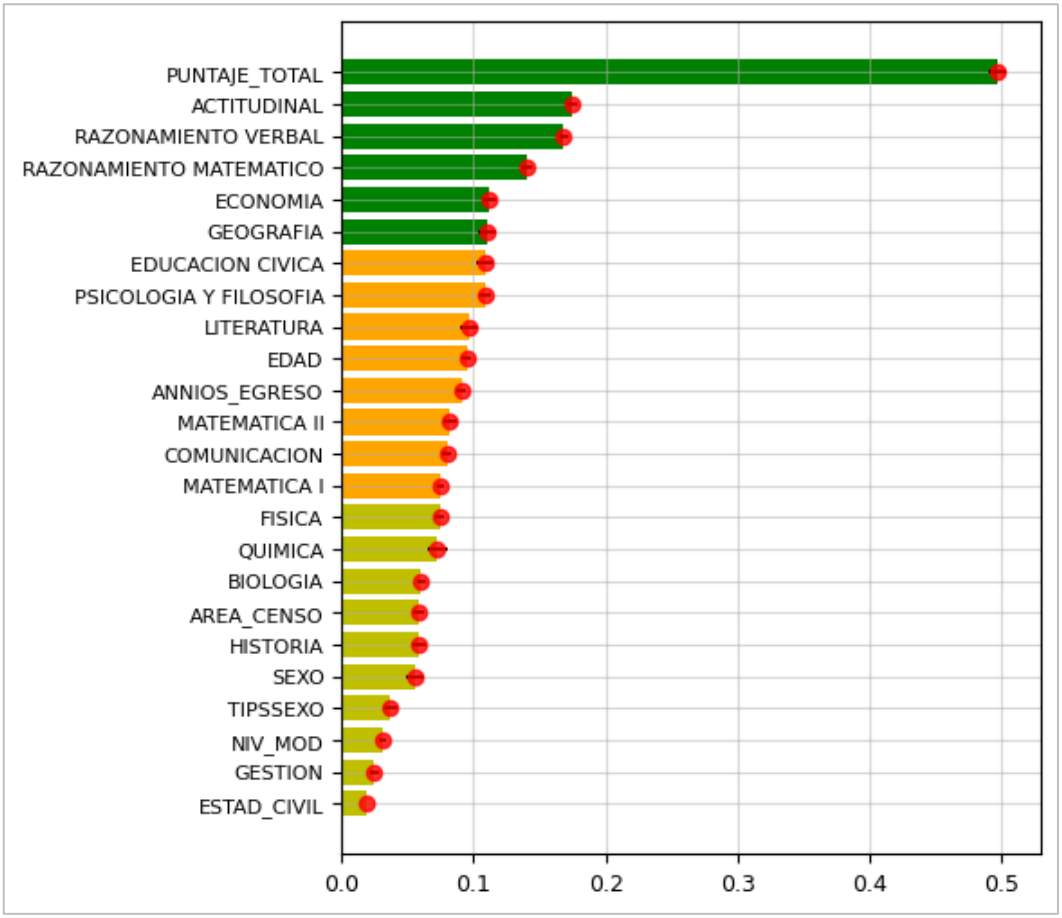


Figura 27. Importancia de los factores de ingreso a la universidad de los postulantes del Examen General 2022-II área de sociales.

De la figura 27, se deduce que los factores: puntaje total obtenido, las asignaturas de: actitudinal, razonamiento verbal, razonamiento matemático, economía y geografía, son factores que tienen mayor influencia en el ingreso a la universidad. Las asignaturas de: educación cívica, literatura, matemática II, comunicación y matemática I, así mismo la edad del postulante y los años de egresado son factores de influencia media. Los demás factores tienen una influencia baja. Consecuentemente el estado civil, la gestión

institucional (estatal o privado) y modalidad de la institución secundaria tiene influencia mínima en el ingreso a la universidad.

4.2. Discusiones

Antes de entrenar el modelo de bosques aleatorios, se consideró los hiperparámetros siguientes:

- **criterion**: entropy, gini.
- **max_depth**: 10, 20, None.
- **max_features**: 5, 7, 9.
- **n_estimators**: 150

Estos valores de los hiperparámetros son ajustables, pero hay que considerar que a mayor profundidad se torna inestable y confuso la predicción del árbol de decisión. El mejor criterio es entropy pero también en algunos casos gini, dependerá de los datos de entrada. El número de predictores (max_features) para esta investigación trabaja mejor con números impares.

En el área de Biomédicas y tomando en cuenta las figuras 5 y 8 respectivamente, donde se ilustra el porcentaje de respuestas correctas proporcionadas por los ingresantes en las 15 asignaturas para el área de Biomédicas, por otro lado, las asignaturas con mayor ponderación según el reglamento de admisión son actitudinal, biología, razonamiento verbal, razonamiento matemático, química y física respectivamente. Con excepción de la asignatura de actitudinal, los ingresantes demostraron un rendimiento inferior al mínimo nivel óptimo (60%), razonamiento verbal, biología de los ingresantes de CEPREUNA sobrepasan el 50% del rendimiento mínimo. Con una ponderación intermedia continúan las asignaturas de literatura, educación cívica, historia, psicología y filosofía. Las asignaturas de literatura e historia tuvieron respuestas correctas los ingresantes del proceso CEPREUNA, en la asignatura de psicología y filosofía tuvieron mejor rendimiento los ingresantes del Examen General.

Este escenario es un indicador que los ingresantes a las escuelas profesionales de Biomédicas requieren de un periodo de nivelación en las asignaturas con mayor ponderación y cuyo rendimiento sea inferior al 50%, por tanto, las asignaturas con mayor

ponderación para el área de Ingenierías son factores determinantes para que el postulante logre su ingreso a la universidad.

Para el área de Ingenierías y tomando en cuenta las figuras 11 y 14 respectivamente, donde se ilustra el porcentaje de respuestas correctas proporcionadas por los ingresantes en las 15 asignaturas para el área de Ingenierías, por otro lado, las asignaturas con mayor ponderación según el reglamento de admisión son actitudinal, razonamiento verbal, razonamiento matemático, matemática II, matemática I y física respectivamente. Con excepción de la asignatura de actitudinal, los ingresantes demostraron un rendimiento inferior al mínimo nivel óptimo (60%), en razonamiento verbal sobrepasan el 50% del rendimiento mínimo. En este grupo surge una particularidad, el rendimiento en la asignatura de educación cívica esta por encima del 60% pero su ponderación es baja.

Este escenario es un indicador que los ingresantes a las escuelas profesionales de Ingenierías requieren de un periodo de nivelación en las asignaturas con mayor ponderación y cuyo rendimiento sea inferior al 50%, por tanto, dichas asignaturas para el área de Ingenierías son factores determinantes para que el postulante logre su ingreso a la universidad.

Para el área de sociales y tomando en cuenta las figuras 17 y 20 respectivamente, donde se ilustra el porcentaje de respuestas correctas proporcionadas por los ingresantes en las 15 asignaturas para el, por otro lado, las asignaturas con mayor ponderación según el reglamento de admisión son actitudinal, razonamiento verbal, razonamiento matemático, historia, Psicología y filosofía, literatura, comunicación, geografía y economía. En las asignaturas de economía, comunicación y actitudinal los ingresantes demostraron un rendimiento superior al mínimo nivel óptimo (60%), en historia y educación cívica los ingresantes del proceso CEPREUNA demostraron rendimiento superior al 60%, en las asignaturas de psicología y filosofía, razonamiento verbal y química los ingresantes del proceso CEPREUNA demostraron un rendimiento superior al 50%.

Este escenario es un indicador que los ingresantes a las escuelas profesionales de Sociales requieren de un periodo de nivelación en las asignaturas con mayor ponderación y cuyo rendimiento sea inferior al 50%, por tanto, las asignaturas con mayor ponderación para el

área de Sociales son factores determinantes para que el postulante logre su ingreso a la universidad.

Según Reddy *et al*., (2020) y Gewers *et al*., (2021) el algoritmo más utilizado en la clasificación de datos es el análisis de componentes principales (PCA) para detectar los valores atípicos, se basa en la técnica de reducción de dimensiones para responder a los objetivos buscados, coincide con esta investigación ya que se trata de clasificar a los ingresantes a la universidad en base a factores predictores. El análisis de componentes principales reduce dimensiones y esta investigación clasifica y ordena en base a una combinación de parámetros determinados y ajustes cíclicos.

Lemay *et al*. (2021) en su investigación sobre minería de datos para predecir el rendimiento de los estudiantes entre excelente y no excelente, aplica varias técnicas de modelado predictivo, las que resaltan con mejores predicciones de rendimiento son K-Nearest Neighbor, Naive Bayes, Decision Tree y Logistic Regression Model. En la investigación realiza predicciones en base a resultados de bosques aleatorios, alcanzando precisiones entre 80% y 90% al momento de predecir si un postulante ingresa o no a la universidad. Considerando ambos modelos predictivos y los predictores empleados para responder a los objetivos buscados, hay que considerar los datos de entrada del modelo.

Dogan y Birant (2021) centra su investigación en la extracción de patrones (conocimientos) útiles desconocidos utilizando técnicas de minería de datos, mientras Gupta y Chandra (2020) amplia los campos de aplicación de la minería de datos hacia la banca, el comercio minorista, la medicina, los seguros, la bioinformática, etc. Todo se orienta a la extracción de conocimientos y plantear soluciones reales a problemas reales. La presente investigación también busca proporcionar conocimientos a las autoridades y a la par, plantear soluciones para que los ingresantes a la universidad sean los más indicados.

Asadi *et al*. (2021) utiliza árboles de decisión para predecir enfermedades del corazón, propone un modelo predictivo denominado multi-objective particle swarm optimization (MOPSO), este modelo compara el rendimiento de seis conjuntos de datos cardíacos con clasificadores individuales y de conjunto, mejorando su rendimiento frente a los bosques aleatorios. En toda investigación se debe comprender el contexto de aplicación y la pureza

de los datos, para datos cardiacos la pureza de los datos el alta y por lo tanto las predicciones serán altas, en cambio, para predecir el ingreso a la universidad, la pureza de los datos es relativamente alta porque hay datos propios del postulante y datos de origen exógeno. Por lo tanto, la pureza de los datos determina la precisión de la predicción de un modelo.

Cuando se afirma que los árboles de decisión constituyen la base fundamental de los bosques aleatorios, la investigación coincide con las conclusiones planteadas por Charbuty y Abdulazeez (2021) al indicar que los bosques aleatorios son herramientas del aprendizaje automático altamente adaptables a grandes conjuntos de datos, esta investigación consideró 14,297 postulantes entre ingresantes y no ingresantes, esta cantidad puede ampliarse a mucho más, y el modelo de bosques aleatorios puede mejorar su precisión para determinar cuándo un postulante puede lograr su ingreso o no a la universidad.

Todas las investigaciones donde utilizan algoritmos de aprendizaje supervisado coinciden en su afirmación que los bosques aleatorios (Random Forest) evitan el sobreajuste (overfitting) del modelo al momento de generalizar la predicción de los resultados. Esta investigación no es ajena a ello, porque predice y generaliza adecuadamente.

CONCLUSIONES

Después de considerar que el aprendizaje automático supervisado ayuda a las universidades a resolver el problema de selección de sus ingresantes, por otro lado, tabular y evaluar los resultados obtenidos por el modelo aprendizaje supervisado de bosques aleatorios y compararlos por procesos (CEPREUNA, Examen General) y áreas (biomédicas, ingenierías y sociales), llegamos a concluir los siguientes:

1. En todas las áreas y procesos de admisión, el puntaje final es el factor determinante que se posiciona con mayor importancia y se obtiene como resultado de la acumulación de respuestas correctas en las asignaturas con mayor ponderación, por lo tanto, estas asignaturas se convierten en factores determinantes con mayor influencia en el ingreso de los postulantes a la universidad, estas últimas son: actitudinal, razonamiento verbal, razonamiento matemático. Adicionalmente existen factores con mayor influencia por área y son: para el área de biomédicas las asignaturas de biología, química, comunicación e historia, para el área de ingenierías las asignaturas de matemática II y matemática I, para el área de sociales las asignaturas de geografía, educación cívica y economía. Los factores con regular influencia en el ingreso de los postulantes a la universidad son asignaturas, algunos propios del postulante y del colegio de procedencia. En todas las áreas y procesos influye la edad, los años de haber egresado, respecto a las asignaturas prevalece la asignatura de psicología y filosofía, educación cívica y física. Con todos estos factores determinantes, el modelo de aprendizaje supervisado de bosques aleatorios propuesto clasifica y predice con una exactitud entre el 80% y el 91% que un postulante ingresa a la universidad.

2. Para obtener mejores resultados, es muy importante la pureza de los datos de entrada de un modelo, por lo tanto, la recopilación de los datos se debe realizar utilizando adecuados sistemas de inscripción y los mismos sean proporcionados por los postulantes. Esta etapa y antes de ser etiquetados se debe acompañar de un buen preprocesamiento y limpieza de datos, luego fijar hiperparámetros optimizados que mejoran significativamente el entrenamiento y predicción del modelo propuesto. No hacerlo implicaría la omisión de datos que ayuden a mejorar la precisión de los resultados.

3. Las salidas obtenidas por el modelo de aprendizaje supervisado de bosques aleatorios dependen en gran medida de los hiperparámetros propuestos (criterion, max_depth, max_features y n_estimators), la eficiencia de estos valores se logra comparando los resultados basados en out-of-bag score, paralelizado de ciclos y validación cruzada. Estos hiperparámetros se debe ajustar para cada área y proceso de admisión, debido a que el comportamiento de los factores determinantes muestra diferentes resultados para cada área y proceso de admisión.

4. En el área de biomédicas, los ingresantes del CEPREUNA-2022 II demuestran acumular la mayor cantidad de respuestas correctas en comparación a los postulantes del Examen General 2022-II. En el área de ingenierías la cantidad de respuestas correctas que acumulan los ingresantes es casi uniforme para ambos procesos y mientras en el área de sociales ocurre lo contrario al área de biomédicas, los ingresantes del Examen General acumulan la mayor cantidad de respuestas correctas frente a los ingresantes del CEPREUNA-2022 II. Estas diferencias obedecen al nivel de preparación que tuvieron los ingresantes en función al área y proceso de admisión a que postulan.

RECOMENDACIONES

1. En la investigación no se consideró el factor tiempo de preparación de los postulantes previo al examen de admisión, sea en academias preuniversitarias, de forma particular, auto preparación o bien sin ninguna preparación. En las futuras investigaciones considerar este factor para evaluar el nivel de influencia en el ingreso de los postulantes a la universidad.

2. Para lograr un buen contraste y con los mismos datos, es posible utilizar otros algoritmos de clasificación como Maquina de Vectores de Soporte (SVM del inglés Support Vector Machine), K-Vecinos más cercanos (KNN), también es posible utilizar algoritmos de regresión logística. Previo a la elección del algoritmo de clasificación, se sugiere considerar los siguientes criterios: a) comprender los datos, si son lineales o no, la cantidad de predictores y si están adecuadamente etiquetados b) experimentar y conocer otros algoritmos de clasificación comparando su precisión del modelo, tiempo de ejecución y sencillez a la hora de implementar, c) considerar la posibilidad de escalar según la cantidad de datos.

3. Considere la posibilidad de aplicar el modelo de aprendizaje supervisado de bosques aleatorios para clasificar los sentimientos de los egresados de las instituciones educativas secundarias en múltiples categorías, específicamente de las promociones 2020 y 2021 cuando el país y el mundo entero estuvo en pandemia. Ello para comprender la percepción de los postulantes y su deseo de estudiar la escuela profesional elegida, considerando siempre el posible sesgo que puede ocasionar algunas variables exógenas.

4. A las futuras investigaciones, se recomienda utilizar algoritmos de aprendizaje supervisado de bosques aleatorios para categorizar y jerarquizar la importancia de los factores determinantes, porqué estos algoritmos tienen un buen desempeño al momento de generalizar las predicciones con datos nuevos, además, evitan el sobreajuste (overfitting) en comparación con predicciones realizadas con árboles de decisión. Para lograr mejores resultados, considerar también el entrenamiento del modelo con hiperparámetros optimizados.

BIBLIOGRAFÍA

Abelairas-Etxebarria, P., & Astorkiza, I. (2020). From exploratory data analysis to exploratory spatial data analysis. *Mathematics and Statistics*, *8*(2). https://doi.org/10.13189/ms.2020.080202

Ahmadini, A. A. H. (2022). A novel technique for parameter estimation in intuitionistic fuzzy logistic regression model. *Ain Shams Engineering Journal*, *13*(1). https://doi.org/10.1016/j.asej.2021.06.004

Alamilla-Jiménez, E., Bolivar-Cime, A., & Nájera, E. (2022). REDES NEURONALES Y SU APLICACIÓN EN LA CLASIFICACIÓN DE PATRONES. *Revista de La Facultad de Ciencias*, *11*(1). https://doi.org/10.15446/rev.fac.cienc.v11n1.99173

Alcaide Martínez, A. (2020). Redes neuronales convolucionales aplicadas a la indentificación y medición automatizadas. *Universidad Carlos III de Madrid*, 99. recuperado de: https://e-archivo.uc3m.es/bitstream/handle/10016/32814/TFG_Asier_Alcaide_Martinez.pdf

Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., & Alzahrani, A. K. (2022). Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data and Cognitive Computing*, *6*(1). https://doi.org/10.3390/bdcc6010013

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear Random Forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering*, *174*(August 2018), 776–789. https://doi.org/10.1016/j.petrol.2018.11.067

Asadi, S., Roshan, S. E., & Kattan, M. W. (2021). Random Forest swarm optimization-based for heart diseases diagnosis. *Journal of Biomedical Informatics*, *115*. https://doi.org/10.1016/j.jbi.2021.103690

Baba, B., & Sevil, G. (2020). Predicting IPO initial returns using Random Forest. *Borsa Istanbul Review*, *20*(1). https://doi.org/10.1016/j.bir.2019.08.001

Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. *CAAI Transactions on Intelligence Technology*, *5*(1). https://doi.org/10.1049/trit.2019.0028

Biau, G., Scornet, E., & Welbl, J. (2019). Neural Random Forests. *Sankhya A*, *81*(2). https://doi.org/10.1007/s13171-018-0133-y

Bowater, R. J., & Denise, G. H. (2013). *ESTADÍSTICA Y CIENCIA. investigación cuantitativa en diversas disciplinas*. Fontamara.

Breiman, L. (2020). Bagging predictors. *Kluwer Academic Publishers*, *8*(3), 1–26. https://doi.org/10.3390/risks8030083

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01). https://doi.org/10.38094/jastt20165

Díaz-Martínez, M. A., Ahumada-Cervantes, M. D. los Á., & Melo-Morín, J. P. (2021). Árboles de Decisión como Metodología para Determinar el Rendimiento Académico en Educación Superior. *Revista Lasallista de Investigación*, *18*(2), 94–104. https://doi.org/10.22507/rli.v18n2a8

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. In *Expert Systems with Applications* (Vol. 166). https://doi.org/10.1016/j.eswa.2020.114060

Espinosa Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, *21*(3), 1–16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Gewers, F. L., Ferreira, G. R., De Arruda, H. F., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. D. F. (2021). Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys*, *54*(4). https://doi.org/10.1145/3447755

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine learning for neural decoding. *ENeuro*, *7*(4). https://doi.org/10.1523/ENEURO.0506-19.2020

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., & Abdessalem, T. (2017). Adaptive Random Forests for evolving data stream classification. *Machine Learning*, *106*(9–10). https://doi.org/10.1007/s10994-017-5642-8

Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology (Singapore)*, *12*(4). https://doi.org/10.1007/s41870-020-00427-7

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3). https://doi.org/10.1007/s12525-021-00475-2

Julianto, M. F., Malau, Y., Hidayat, W. F., Nugroho, W., & Indriyani, F. (2021). COMPARATION OF DECISION TREE MODEL AND SUPPORT VERCTOR MACHINE IN SENTIMENT ANALYSIS OF REVIEW DATASET SAMSUNG SSD 850 EVO AT NEW EGG SHOP. *Jurnal Riset Informatika*, *3*(4). https://doi.org/10.34288/jri.v3i4.278

Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100016

Lemus-Delgado, D., & Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. *Colombia Internacional*, *102*, 41–62. https://doi.org/10.7440/colombiaint102.2020.03

Li, L., Xiong, D., & Wu, X. (2011). Classification of imaginary movements in ECoG. *5th International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2011*. https://doi.org/10.1109/icbbe.2011.5780688

Lubis, A. R., Lubis, M., & Al-Khowarizmi. (2020). Optimization of distance formula in k-nearest neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1). https://doi.org/10.11591/eei.v9i1.1464

Mohandoss, D. P., Shi, Y., & Suo, K. (2021). Outlier Prediction Using Random Forest Classifier. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 0027–0033.

https://doi.org/10.1109/CCWC51732.2021.9376077

Nieto, Y., Gacia-Diaz, V., Montenegro, C., Gonzalez, C. C., & Gonzalez Crespo, R. (2019). Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions. *IEEE Access*, *7*. https://doi.org/10.1109/ACCESS.2019.2919343

Ponce Cruz, P. (2010). *Inteligencia artificial con aplicacion a la ingeniería*.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.2980942

Rojas, B. S. C., Rodríguez, C. U. C., Osorio, D. J. E., & Bello, Y. T. G. (2020). Redes neuronales artificiales y estado del arte aplicado en la ciberseguridad. *Revista Matices Tecnológicos*, *12*.

Rueda, A. M. (2020). Estudio exploratorio sobre “la sabiduría de las masas” Atenuación y cortesía verbal en las reseñas gastronómicas de TripAdvisor. *Les Enjeux Du Numérique En Sciences Sociales et Humaines*, *2004*. https://doi.org/10.17184/eac.9782813003867

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12). https://doi.org/10.35940/ijitee.L3591.1081219

Schonlau, M., & Zou, R. Y. (2020). The Random Forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Seraj, A., Mohammadi-Khanaposhtani, M., Daneshfar, R., Naseri, M., Esmaeili, M., Baghban, A., Habibzadeh, S., & Eslamian, S. (2022). Cross-validation. In *Handbook of HydroInformatics: Volume I: Classic Soft-Computing Techniques*. https://doi.org/10.1016/B978-0-12-821285-1.00021-X

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of Random Forest variable selection methods for classification prediction modeling. In *Expert Systems with Applications* (Vol. 134). https://doi.org/10.1016/j.eswa.2019.05.028

Tukey, J. W. (1977). *Exploratory data analysis Vol. 2*.

Xu, W., & Hoang, V. T. (2021). MapReduce-Based Improved Random Forest Model for Massive Educational Data Processing and Classification. *Mobile Networks and Applications*, *26*(1). https://doi.org/10.1007/s11036-020-01699-w

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(3). https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592

Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using Random Forest machine learning algorithm. *Chaos, Solitons & Fractals*, *140*, 110210. https://doi.org/10.1016/j.chaos.2020.110210

Yu, J. (2021). Academic Performance Prediction Method of Online Education using Random Forest Algorithm and Artificial Intelligence Methods. *International Journal of Emerging Technologies in Learning*, *16*(5). https://doi.org/10.3991/ijet.v16i05.20297

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2020). Random Forest Prediction Intervals. *The American Statistician*, *74*(4), 392–406. https://doi.org/10.1080/00031305.2019.1585288

ANEXOS

Anexo 1. Código fuente del núcleo del modelo y sus hiperparámetros.

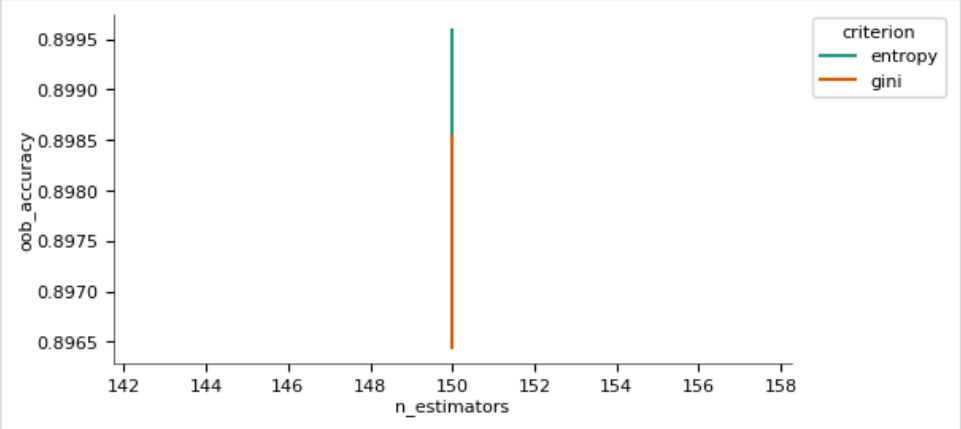
```
# Ajuste del modelo y optimización de hiperparámetros
X_train, X_test, y_train, y_test = train_test_split(
    postulantes_GRAL_ING.drop(columns = 'INGRESO'),
    postulantes_GRAL_ING['INGRESO'],
    random_state = 145
    )

# Hiperparámetros evaluados para su optimización
params_grid = ParameterGrid(
    {'n_estimators': [150],
     'max_features': [5, 7, 9],
     'max_depth'   : [None, 3, 10, 20],
     'criterion'   : ['gini', 'entropy']})

# Grid Search basado en out-of-bag score
resultados = {'params': [], 'oob_accuracy': []}
for params in params_grid:
    modelo = RandomForestClassifier(
        oob_score    = True,
        n_jobs       = -1,
        random_state = 123,
        ** params
    )
    modelo.fit(X_train, y_train)
    resultados['params'].append(params)
    resultados['oob_accuracy'].append(modelo.oob_score_)

# Grid Search en su versión paralelizada
def eval_oob_error(X, y, modelo, params, verbose=True):
        modelo.set_params(
        oob_score    = True,
        n_jobs       = -1,
        random_state = 123,
        ** params
    )
    modelo.fit(X, y)
    return {'params': params, 'oob_accuracy': modelo.oob_score_}

n_jobs     = multiprocessing.cpu_count() - 1
pool       = multiprocessing.Pool(processes=n_jobs)
resultados = pool.starmap(
    eval_oob_error,
    [(X_train, y_train, RandomForestClassifier(), params) for params
in param_grid]
)
```

```
# Grid Search basada en validación cruzada
validacionCruzada = GridSearchCV(
    estimator  = RandomForestClassifier(random_state = 123),
    param_grid = param_grid,
    scoring    = 'accuracy',
    n_jobs     = multiprocessing.cpu_count() - 1,
       cv                =   RepeatedKFold(n_splits=5,   n_repeats=3,
random_state=123),
    refit      = True,
    verbose    = 0,
    return_train_score = True
)
# with tf.device('/device:GPU:0'):
validacionCruzada.fit(X = X_train, y = y_train)
resultados = pd.DataFrame(validacionCruzada.cv_results_)
```

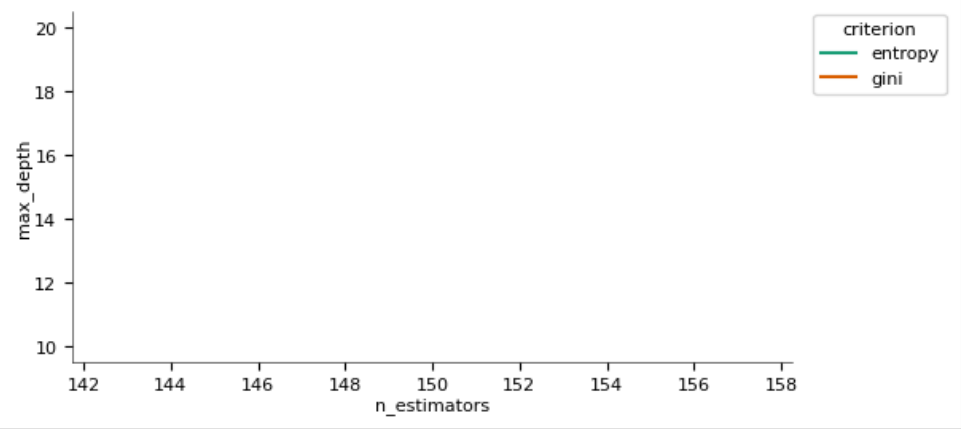
Anexo 2. Resultados de evaluación de hiperparámetros basados en out-of-bag score y paralelizada.

	oob_accuracy	criterion	max_depth	max_features	n_estimators
12	0.899582	entropy	NaN	5	150
2	0.898536	gini	NaN	9	150
21	0.898536	entropy	20.0	5	150
1	0.898536	gini	NaN	7	150
11	0.898536	gini	20.0	9	150
10	0.898536	gini	20.0	7	150
9	0.897490	gini	20.0	5	150
0	0.897490	gini	NaN	5	150
8	0.896444	gini	10.0	9	150
7	0.896444	gini	10.0	7	150

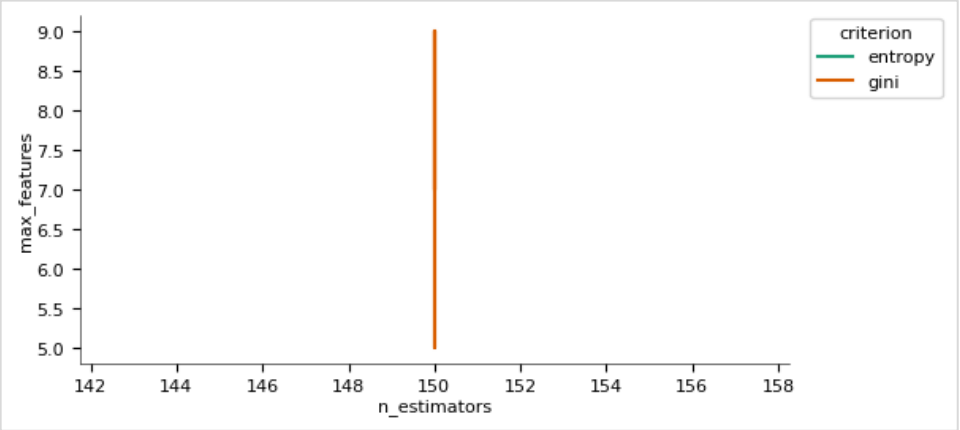
Precisión alcanzada por el modelo de bosques aleatorios en base al criterio es de 89.9%



Profundidad máxima optimizada para lograr un mejor desempeño del modelo en base al criterio es vacío (None)



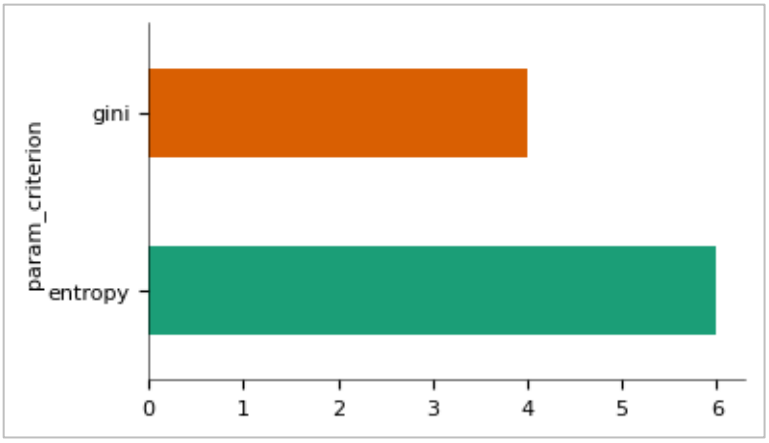
Bloques de factores predictores para el modelo en base al criterio se establece a 9, es decir se agrupa de 9 en 9 para la construcción de los árboles de decisión.



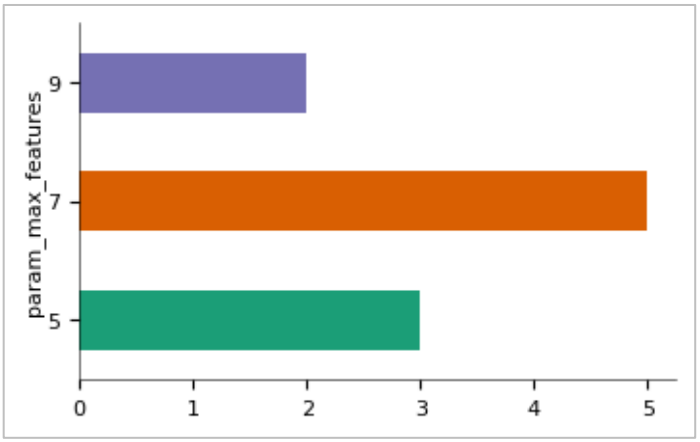
Anexo 3. Resultados de evaluación de hiperparámetros basados en validación cruzada.

	criterion	max_depth	max_features	n_estimators	mean_test_score
12	entropy	None	5	150	0.899231
21	entropy	20	5	150	0.899231
14	entropy	None	9	150	0.898537
23	entropy	20	9	150	0.898537
19	entropy	10	7	150	0.898188
10	gini	20	7	150	0.898184
1	gini	None	7	150	0.898184
13	entropy	None	7	150	0.897838
7	gini	10	7	150	0.897835
0	gini	None	5	150	0.897493

Mejor criterio para la construcción de los árboles es "*entropy*"



Bloques de factores predictores optimizados para el modelo en base al criterio se establece a 5, es decir se agrupa de 5 en 5 para la construcción de los árboles de decisión.



Anexo 4. Código fuente y datos tabulados de la importancia de los factores predictores de ingreso a la universidad.

```
importancia = permutation_importance(
    estimator    = modelo_final,
    X            = X_train,
    y            = y_train,
    n_repeats    = 5,
    scoring      = 'neg_root_mean_squared_error',
    n_jobs       = multiprocessing.cpu_count() - 1,
    random_state = 123
    )

df_importancia = pd.DataFrame(
    { k: importancia[k] for k in ['importances_mean', 'importances_std'] } )
df_importancia['predictor'] = X_train.columns
print(df_importancia.sort_values('importances_mean',
ascending=False))
```

	predictor	importances_mean	importances_std
2	PUNTAJE_TOTAL	0.126787	0.005654
20	RAZONAMIENTO VERBAL	0.058045	0.005765
21	ACTITUDINAL	0.040748	0.003810
11	BIOLOGIA	0.036335	0.010406
16	ECONOMIA	0.022910	0.003891
12	PSICOLOGIA Y FILOSOFIA	0.018930	0.002171
8	MATEMATICA II	0.018881	0.004727
19	RAZONAMIENTO MATEMATICO	0.018353	0.001837
17	COMUNICACION	0.014234	0.002365
13	GEOGRAFIA	0.013638	0.002400
9	FISICA	0.010640	0.001484
10	QUIMICA	0.010618	0.003088
22	EDAD	0.010598	0.004062
0	SEXO	0.008811	0.001234
14	HISTORIA	0.006332	0.001969
18	LITERATURA	0.006310	0.003379
15	EDUCACION CIVICA	0.003831	0.001257
4	TIPSSEXO	0.003831	0.001257
7	MATEMATICA I	0.003819	0.002370
23	ANNIOS_EGRESO	0.003819	0.002370
1	ESTAD_CIVIL	0.000000	0.000000
6	AREA_CENSO	0.000000	0.000000
3	NIV_MOD	0.000000	0.000000
5	GESTION	-0.001307	0.001601

Anexo 5. Código fuente y datos tabulados de la importancia de los factores predictores de ingreso a la universidad en base a la pureza de los nodos.

```
importancia_predictores = pd.DataFrame(
    {'predictor': X_train.columns,
     'importancia': modelo_final.feature_importances_ }
     )

print("IMPORTANCIA DE LOS FACTORES DE INGRESO A LA UNIVERSIDAD")
print(importancia_predictores.sort_values('importancia',
ascending=False))
```

	predictor	importancia
2	PUNTAJE_TOTAL	0.261205
21	ACTITUDINAL	0.062907
20	RAZONAMIENTO VERBAL	0.058947
19	RAZONAMIENTO MATEMATICO	0.055126
22	EDAD	0.051337
11	BIOLOGIA	0.050938
8	MATEMATICA II	0.044169
23	ANNIOS_EGRESO	0.041837
10	QUIMICA	0.041772
12	PSICOLOGIA Y FILOSOFIA	0.037962
9	FISICA	0.037035
16	ECONOMIA	0.036821
17	COMUNICACION	0.030577
7	MATEMATICA I	0.029937
14	HISTORIA	0.025952
13	GEOGRAFIA	0.025799
15	EDUCACION CIVICA	0.023202
18	LITERATURA	0.022032
0	SEXO	0.017775
5	GESTION	0.015554
4	TIPSSEXO	0.014533
6	AREA_CENSO	0.010682
3	NIV_MOD	0.003822
1	ESTAD_CIVIL	0.000080

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

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo PABLO CESAR TAPIA CATACORA,
identificado con DNI 40270043 en mi condición de egresado de:

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

MAESTRÍA EN INGENIERÍA DE SISTEMAS,

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

"EVALUACIÓN DE FACTORES DETERMINANTES PARA EL INGRESO DE LOS POSTULANTES A LAS UNIVERSIDADES"

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 30 de noviembre del 2023

FIRMA (obligatoria) Huella

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 PABLO CESAR TAPIA CATACORA,
identificado con DNI 40270043 en mi condición de egresado de:

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

MAESTRÍA EN INGENIERÍA DE SISTEMAS,
informo que he elaborado el/la ☑ **Tesis** o ☐ **Trabajo de Investigación** denominada:

"EVALUACIÓN DE FACTORES DETERMINANTES PARA EL INGRESO DE LOS POSTULANTES A LAS UNIVERSIDADES"

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 30 de noviembre del 2023

FIRMA (obligatoria)

Huella