ANÁLISIS DE LA INFLUENCIA DE FACTORES SOCIOECONÓMICOS EN LAS PRUEBAS SABER PRO

JUAN SEBASTIAN RODRIGUEZ CARREÑO - 20231020107

SEBASTIAN RICARDO ACEVEDO BALDOVINO - 20222020095

CARLOS ANDRES PARDO ANGEL - 20231020041

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

PROBABILIDAD Y ESTADÍSTICA

GRUPO 020-84

ALBERTO ACOSTA LOPEZ

2025

Introducción

El siguiente trabajo busca encontrar la relación entre los factores socioeconómicos y el desempeño académico de los estudiantes colombianos en las pruebas Saber Pro 2023. Se busca identificar y cuantificar la influencia de diversas variables socioeconómicas en los resultados obtenidos por los estudiantes en el examen de este año, para ambas pruebas y comparar con los resultados del año pasado. Comprender estas dinámicas puede llevar a una conciencia mayor de enfoque para las políticas públicas, así como un mejor entendimiento de la población y que es lo que más la afecta.

Para abordar esta problemática, se emplearán metodologías de machine learning, específicamente técnicas de regresión lineal, regresión logística y árboles de decisión. La regresión lineal permitirá modelar la relación entre variables socioeconómicas continuas y el puntaje global en la prueba, mientras que la regresión logística se utilizará para analizar la probabilidad de alcanzar ciertos umbrales de desempeño o la elección de programas académicos específicos en función de estos factores. Finalmente, los árboles de decisión ofrecerán una perspectiva no lineal y permitirán identificar interacciones complejas entre las variables predictoras, así como segmentar a la población estudiantil en función de sus características socioeconómicas y su rendimiento académico. El análisis resultante proporcionará información sobre los determinantes socioeconómicos del éxito en las pruebas Saber Pro 2023, y una vez que se tenga esta información, se podrá comparar con los resultados del año 2022 para saber si el modelo que se planteó es viable.

El grupo de muestra para este trabajo va a pertenecer a los resultados de todas las universidades de Colombia con un programa de ingeniería, y cuyos estudiantes hayan realizado el examen Saber Pro en el periodo académico 2023-3.

Objetivos

Objetivo General:

Analizar la relación entre los factores socioeconómicos y el desempeño académico de los estudiantes colombianos de la facultad de ingeniería de seis universidades específicas en las pruebas Saber Pro 2023, y comparar los hallazgos con los resultados del año 2022 para evaluar la viabilidad del modelo planteado.

Objetivos Específicos:

- Identificar y cuantificar la influencia de variables socioeconómicas específicas (a definir según la disponibilidad de datos) en el puntaje global obtenido por los estudiantes de ingeniería en las pruebas Saber Pro 2023, utilizando modelos de regresión lineal.
- Analizar la probabilidad de los estudiantes de ingeniería de alcanzar umbrales de desempeño académico predefinidos en las pruebas Saber Pro 2023, en función de sus factores socioeconómicos, mediante la aplicación de modelos de regresión logística.
- Segmentar a la población estudiantil de ingeniería de las universidades seleccionadas según sus características socioeconómicas y su rendimiento en las pruebas Saber Pro 2023, utilizando los resultados de los modelos de machine learning implementados.
- Comparar los resultados obtenidos para el año 2023 con los resultados del año 2022 para las mismas universidades y facultad de ingeniería, con el fin de identificar cambios en la relación entre los factores socioeconómicos y el desempeño académico, y evaluar la consistencia y viabilidad del modelo propuesto a lo largo del tiempo.

Índice

Marco Conceptual

Para mayor entendimiento del proceso realizado para el método de predicción es fundamental entender términos relevantes del tema, el procesamiento de los datos, la metodología y la interpretación cruda de los datos para generar una conclusión asertiva. “Los modelos predictivos utilizan datos históricos y técnicas de machine learning para identificar patrones y predecir resultados futuros” (Reuters Institute, 2023; Mckinsey & Company, 2022), cabe aclarar que en este documento de investigación se harán uso de estas técnicas de machine learning, junto con modelos de regresión y árboles de decisión para hallar patrones y generalizar una predicción.

La tecnología se encuentra en avance continuo y se ve evidenciado en las nuevas formas de investigación y divulgación científica Según el Ministerio de Telecomunicaciones y de la sociedad de la información (2025) “ El machine learning (ML) se basa en el desarrollo de algoritmos que permiten a las computadoras aprender de los datos sin ser programadas explícitamente . Estos algoritmos identifican patrones y hacen predicciones o toman decisiones basadas en los datos que han aprendido“. El machine learning será una de las herramientas más útiles durante el procedimiento pues hará todo el procesamiento en la búsqueda de patrones.

“La regresión lineal múltiple es un modelo estadístico que examina la relación entre una

variable dependiente y dos o más variables independientes. Este modelo permite identificar cómo diferentes factores influyen en un resultado específico.”(Martinez Alba et al., 2025).

Este tipo de modelo es esencial al querer buscar una relación específica entre los factores socioeconómicos de los partícipes a los resultados presentados en la prueba de Estado, siendo este último la variable independiente.

Eso junto a los árboles de decisión que se definen como:

> "Un árbol de decisión es una regresión a través de la clasificación para minería de datos y otras aplicaciones, representado con una estructura invertida en forma de árbol, donde la raíz en la parte superior es la entrada y las hojas en la parte inferior son los resultados o decisiones" (Panda & Daya Sagar, 2022)

Y finalmente también junto a la regresión logística definida como:

> "La regresión logística es un tipo de modelo lineal generalizado, que pertenece a una familia de modelos en los que se relajan las principales suposiciones lineales. La regresión logística es una herramienta excelente para modelar relaciones con resultados que no se miden en una escala continua" (Starbuck, 2023).

Estos modelos más una trata estadística de los datos harán posible la formulación de un método de predicción funcional en torno al contexto deseado y explicado durante el documento.

El Icfes Saber Pro es la prueba de Estado mencionada anteriormente, en la página oficial de la prueba se define según ICFES(2025) como: "El examen de Estado de la Calidad de la Educación Superior, Saber Pro, es un instrumento de evaluación estandarizada para la medición externa de la calidad de la educación superior que evalúa las competencias de los estudiantes que están próximos a culminar los distintos programas profesionales universitarios."

Metodología

La metodología empleada en este estudio se estructuró en cinco etapas clave: recolección de datos, preprocesamiento, selección de variables, modelado estadístico y validación de resultados. A continuación, se detalla cada una de estas etapas.

1. Recolección de Datos:

Se utilizaron datos públicos de las pruebas Saber Pro 2023, proporcionados por el Instituto Colombiano para la Evaluación de la Educación (ICFES), correspondientes a estudiantes de la facultad de ingeniería de todas las instituciones educativas que presentaron la prueba.

Las variables socioeconómicas consideradas incluyeron: pago matrícula anterior semestre (de donde se paga y cuánto), la educación de sus padres, el propio trabajo, el estrato, si tiene conexión a internet, computador, lavadora, horno microondas, servicio de tv, vehiculo (auto o moto), consola de videojuegos, y las horas que trabaja en la semana.

2. Presentación del algoritmo utilizado, para procesamiento de datos:

I. Importación de Bibliotecas:

```
import pandas as pd
import numpy as np
import os
from openpyxl import Workbook
from openpyxl.utils.dataframe import dataframe_to_rows
from openpyxl.styles import Font, PatternFill, Border, Side
```

- **Propósito:** Cargar herramientas para:

- Manipulación de datos (pandas, numpy).
- Generación de Excel con formato (openpyxl).
- Manejo de rutas de archivos (os).

II. Función Principal (procesar_a_excel)

A. Lectura de Datos

```python
def procesar_a_excel(input_file, output_excel):
    try:
        # Leer el archivo con el delimitador correcto (;) o '¬' dependiendo si es 20232 0 20231 respectivamente

        #df = pd.read_csv(input_file, sep=';', encoding='utf-8')
        df = pd.read_csv(input_file, sep='¬', encoding='utf-8', engine='python')
```

- **Detalle:**
  - **Lee archivos CSV con delimitador ¬ o ; (evita conflictos con texto).**
  - **`engine='python'` asegura compatibilidad.**

B. Limpieza de Datos:

```python
        # Limpieza básica de datos
        df = df.apply(lambda x: x.str.strip() if x.dtype == 'object' else x)
        df.replace(['', 'NA', 'N/A', 'NaN'], np.nan, inplace=True)
```

- **Acciones:**
  - Elimina espacios en blanco.
  - Estandariza valores faltantes como NaN.

C. Creación de Excel

```
# Crear archivo Excel con openpyxl para mejor control
wb = Workbook()
ws = wb.active
ws.title = "Datos ICFES"
```

- **Salida:**
    - Crea un libro de Excel con hoja "Datos ICFES".

D. Configuración de Estilos para Excel:

```
# Estilos para los encabezados
header_font = Font(bold=True, color="FFFFFF")
header_fill = PatternFill(start_color="2c3e50", end_color="2c3e50", fill_type="solid")
thin_border = Border(left=Side(style='thin'),
                     right=Side(style='thin'),
                     top=Side(style='thin'),
                     bottom=Side(style='thin'))
```

- **Propósito:** Define el formato visual del archivo Excel resultante
- **Detalles:**
    - header_font: Texto en negrita y color blanco para encabezados
    - header_fill: Fondo azul oscuro para la fila de encabezados
    - thin_border: Borde delgado alrededor de todas las celdas

E. Escritura de Datos en Excel

```
# Escribir los datos
for r_idx, row in enumerate(dataframe_to_rows(df, index=False, header=True), 1):
    for c_idx, value in enumerate(row, 1):
        cell = ws.cell(row=r_idx, column=c_idx, value=value)
        if r_idx == 1:  # Encabezados
            cell.font = header_font
            cell.fill = header_fill
            cell.border = thin_border
```

- **Funcionamiento:**
    - Convierte el DataFrame fila por fila al formato Excel

- Aplica los estilos solo a la primera fila (encabezados)
- enumerate(..., 1) comienza la numeración desde 1 (formato Excel)

F. Ajuste Automático de Columnas

```python
# Ajustar el ancho de las columnas
for column in ws.columns:
    max_length = 0
    column_letter = column[0].column_letter
    for cell in column:
        try:
            if len(str(cell.value)) > max_length:
                max_length = len(str(cell.value))
        except:
            pass
    adjusted_width = (max_length + 2) * 1.2
    ws.column_dimensions[column_letter].width = adjusted_width
```

- **Lógica:**
  - Calcula el ancho necesario basado en el contenido más largo de cada columna
  - (max_length + 2) * 1.2 proporciona un espacio adicional del 20%

G. Funcionalidades Adicionales

```python
# Congelar encabezados
ws.freeze_panes = 'A2'

# Guardar el archivo
os.makedirs(os.path.dirname(output_excel), exist_ok=True)
wb.save(output_excel)

print("\n" + "="*50)
print("PROCESAMIENTO COMPLETADO EXITOSAMENTE")
print(f"Archivo Excel generado en: {output_excel}")
print("="*50 + "\n")
```

- **Mejoras de Usabilidad:**
  - Encabezados visibles al desplazarse
  - Creación automática de directorios
  - Guardado seguro del archivo

III. Manejo de Errores:

```
except Exception as e:
    print("\n" + "="*50)
    print("ERROR EN EL PROCESAMIENTO")
    print(f"Error: {str(e)}")
    print("="*50 + "\n")
    raise
```

- **Control de Calidad:**
    - Bloque try-except para capturar cualquier error
    - Mensajes claros de error con formato visual
    - Re-lanza la excepción para depuración

IV. Función Principal (main):

```
def main():
    input_file = 'tu_data.txt'
    output_excel = '/output/icfes_data.xlsx'

    procesar_a_excel(input_file, output_excel)

if __name__ == "__main__":
    main()
```

- **Buenas Prácticas:**
    - Punto único de entrada al programa
    - Rutas configurables en un solo lugar
    - Ejecución condicional (solo si es script principal)

3. Filtrado de datos

Una vez se limpio la base de datos, se procedió a filtrar la información pertinente con la que se trabajara, se realizó únicamente la comparacion de el campo de grupo de referencia filtrando por carreras afines con el área de ingeniería, el script con el que se desarrolló dicho filtrado es el siguiente:

main.py > ...

```python
import pandas as pd

file_path = 'icfes_data.xlsx'
data = pd.read_excel(file_path)

grupo = "INGENIERÍA"

dataFiltered = data[
    data['gruporeferencia'] == grupo
]

dataFiltered.to_excel('icfesFiltered.xlsx', index=False)

print(dataFiltered)
```

Se puede evidenciar el uso de la librería pandas (en conjunto con la librería openPyxl) para la función read_excel, y se filtraron los datos con el conjunto de caracteres “INGENIERÍA” dentro de la columna pertinente en la base de datos, una vez que este proceso se realizó, se guardaron los datos en otro excel llamado “icfesFiltered.xlsx” haciendo uso de la función to_excel (con un index falso para no generar cabeceras adicionales a los datos), en total se contaría con 18899 datos que servirán como nuestro muestreo en el análisis.

Análisis de la influencia de factores socioeconómicos en las pruebas saber pro

Resultados

Conclusiones

Se

Bibliografía

- Garcia M. (2024). *Técnicas de Machine Learning para la predicción del rendimiento académico en las pruebas Saber Pro en Colombia.* https://repository.unad.edu.co/bitstream/handle/10596/62951/mgarcia.pdf?sequence=1&isAllowed=y
- Paterson M. (2024). *Gender, socioeconomic status, and numeracy test scores.* https://www.sciencedirect.com/science/article/pii/S0167268124003652?via%3Dihub
- Reuters Institute. (2023). *Resumen ejecutivo y hallazgos clave del informe de 2023. Reuters Institute for the Study of Journalism.* https://reutersinstitute.politics.ox.ac.uk/es/digital-news-report/2023/dnr-resumen-ejecutivo
- McKinsey & Company. (2022). *El estado de la IA en 2022 y el balance de media década*. https://www.mckinsey.com/featured-insights/destacados/el-estado-de-la-ia-en-2022-y-el-balance-de-media-decada/es
- Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2025). *Política Pública para la Transformación Digital del Ecuador 2025-2030.*
- Martínez Alba, M. G., García Munguía, C. A., Meráz Jiménez, A. de J., Mata Zamores, S., Olvera González, J. E., & García Munguía, A. M. (2025). *La gestión del conocimiento y su relación con el empoderamiento de las mujeres universitarias. RIDE Revista Iberoamericana Para La Investigación Y El Desarrollo Educativo*, 15(30).
- Panda, R. M., & Daya Sagar, B. S. (2022). Decision tree [Árbol de decisión]. En B. S. Daya Sagar, Q. Cheng, J. McKinley, & F. Agterberg (Eds.), *Encyclopedia of Mathematical Geosciences (pp. 1–7). Springer*.

https://doi.org/10.1007/978-3-030-26050-7_81-2

- Starbuck, C. (2023). *Logistic regression. En The Fundamentals of People Analytics (pp. 223–238). Springer*. https://doi.org/10.1007/978-3-031-28674-2_12SpringerLink

- ICFES. (2025). *Acerca del examen Saber Pro*. https://www.icfes.gov.co/evaluaciones-icfes/acerca-del-examen-saber-pro/