Sabana Centro Cómo Vamos

# PROYECTO CORRELACIÓN Y PREDICCIÓN

Juan Esteban Ocampo, Juan Pablo Corral, Felipe Orjuela, Sebastián Serrano, Daniel Hurtado

Sabana Centro Cómo Vamos


# Introducción

Para garantizar un análisis sólido, se realizó un proceso estructurado de búsqueda y selección de bases de datos. Esto incluyó la identificación de fuentes confiables, la evaluación de la calidad y relevancia de los datos y la integración con bases de datos externas. Posteriormente, se analizaron las correlaciones entre estas fuentes para validar su compatibilidad y extraer información significativa.

# Cronograma: Fases

| Recopilación y Preparación de Datos | Análisis de Correlación | Desarrollo del Modelo Predictivo |
|---|---|---|
| • Se recolectan y limpian datos de calidad de vida en educación, salud y seguridad.<br>• Se seleccionan variables relevantes para el análisis de correlaciones. | • Se identifican relaciones entre variables mediante análisis estadísticos y machine learning<br>• Se visualizan los resultados con gráficos y mapas de calor | • Se selecciona y entrena un modelo de predicción basado en los datos analizados<br>• Se evalúa la precisión del modelo con métricas como RMSE y $R^2$ |

# Fase 1

## PROCESO

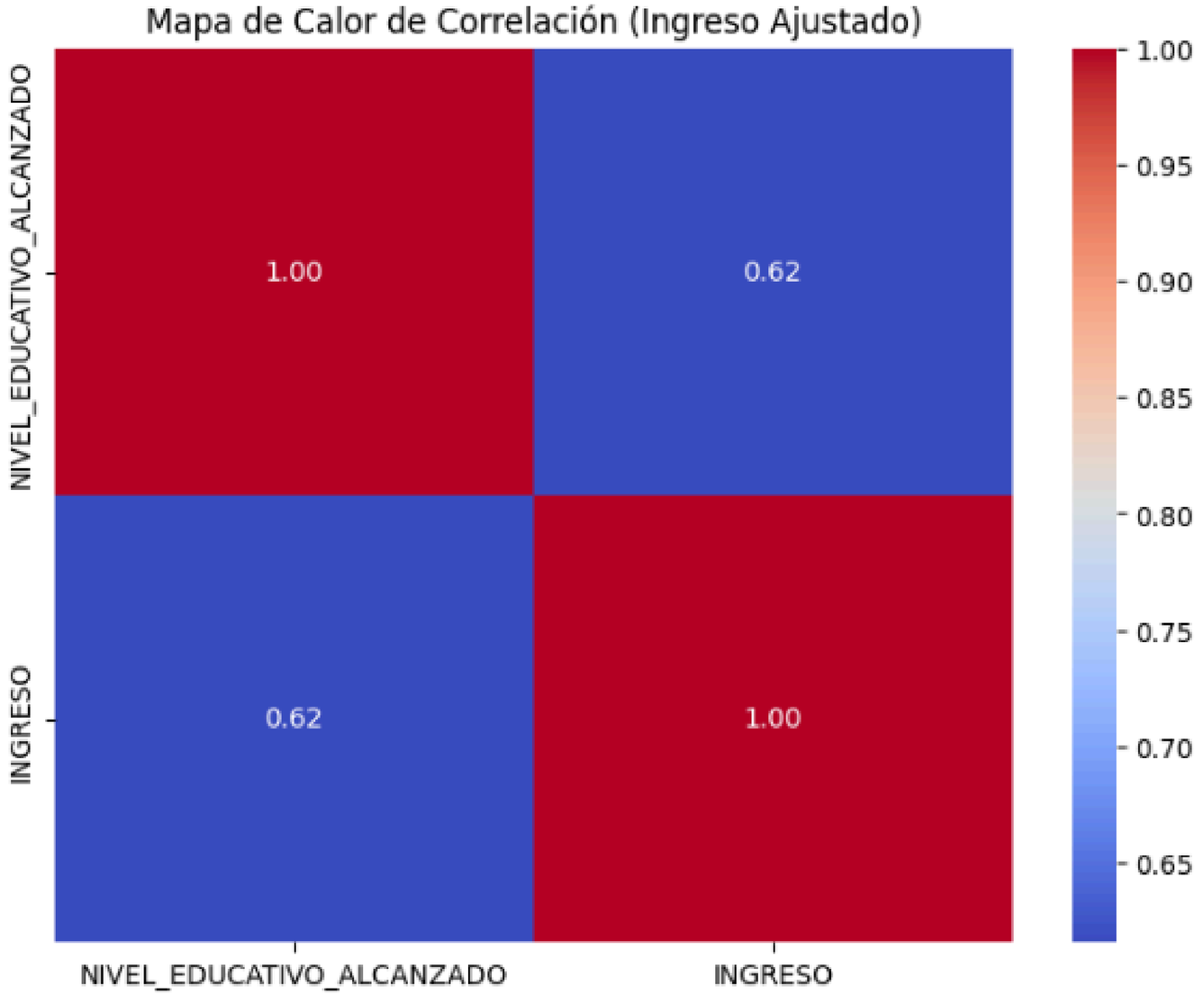
- Se agrupan varias bases de datos de áreas diversas, aparte de sabana centro, con criterio de división por municipio
- Se realiza un estudio rápido de correlación entre las bases para seleccionar bases afines con conceptos entrelazados
- Se seleccionan bases relacionadas con nivel académico e ingreso neto
- Se descartan variables irrelevantes para el estudio y se realiza filtrado de municipios relevantes (Sabana centro)
- Se realiza limpieza final de datos vacios y datos sin sentido

# Datos utilizados

```
<class 'pandas.core.frame.DataFrame'>
Index: 17723 entries, 108 to 245365
Data columns (total 10 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   ID_VIVIENDA                  17723 non-null  float64
 1   ID_HOGAR                     17723 non-null  int64
 2   ID_PERSONA                   17723 non-null  int64
 3   SECUENCIA_HOGAR              17723 non-null  int64
 4   ORDEN_PERSONA                17723 non-null  int64
 5   DEPARTAMENTO                 17723 non-null  int64
 6   MUNICIPIO                    17723 non-null  int64
 7   NIVEL_EDUCATIVO_ALCANZADO    17723 non-null  float64
 8   NIVEL_EDUCATIVO              15 non-null     float64
 9   INGRESO                      17723 non-null  float64
dtypes: float64(4), int64(6)
memory usage: 1.5 MB
None
```

Datos filtradros y limpiados considerados relevantes para el estudio y generación de modelo

# Mapa de calor




Nivel de correlación entre nivel educativo e ingreso en municipios de sabana centro

# Fase 2

## ANÁLISIS ESTADÍSTICO

Se aplicarán técnicas, como el coeficiente de Pearson para variables numéricas y chi-cuadrado para categóricas.

## MACHINE LEARNING

Se estudiarán algoritmos y modelos, como regresión linea, regresión polinómica, árboles de decisión y redes neuronales, para detectar patrones y relaciones complejas.

## VISUALIZACIÓN

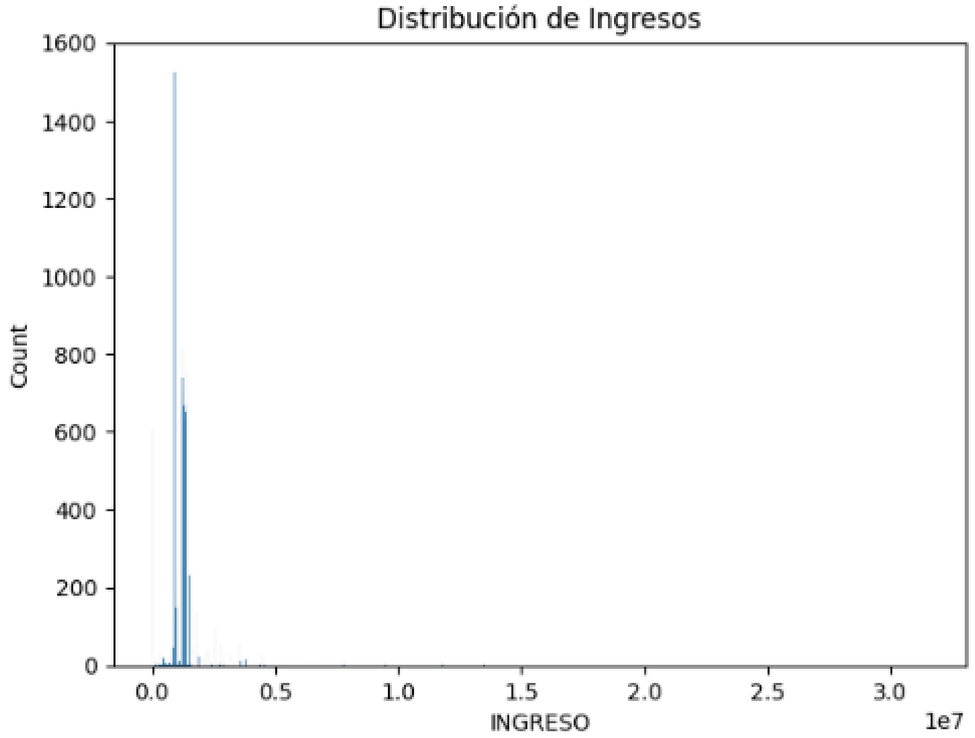
Las correlaciones se visualizarán con gráficos y mapas de calor para facilitar la comprensión de las relaciones entre las variables

# Exploration Data Analysis (EDA)

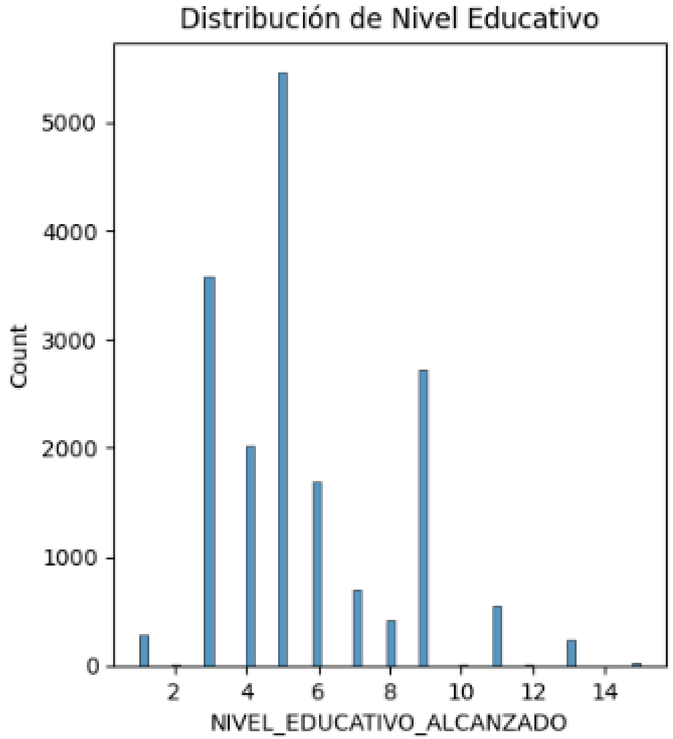
## EDA, GRÁFICAS Y ESTUDIOS REALIZADOS

- Histograma de Distribución de Ingresos
- Histograma de Distribución de Nivel Educativo Alcanzado
- Diagrama de Caja de Ingresos
- Diagrama de Dispersión: Ingreso vs Nivel Educativo Alcanzado

# FASE 2

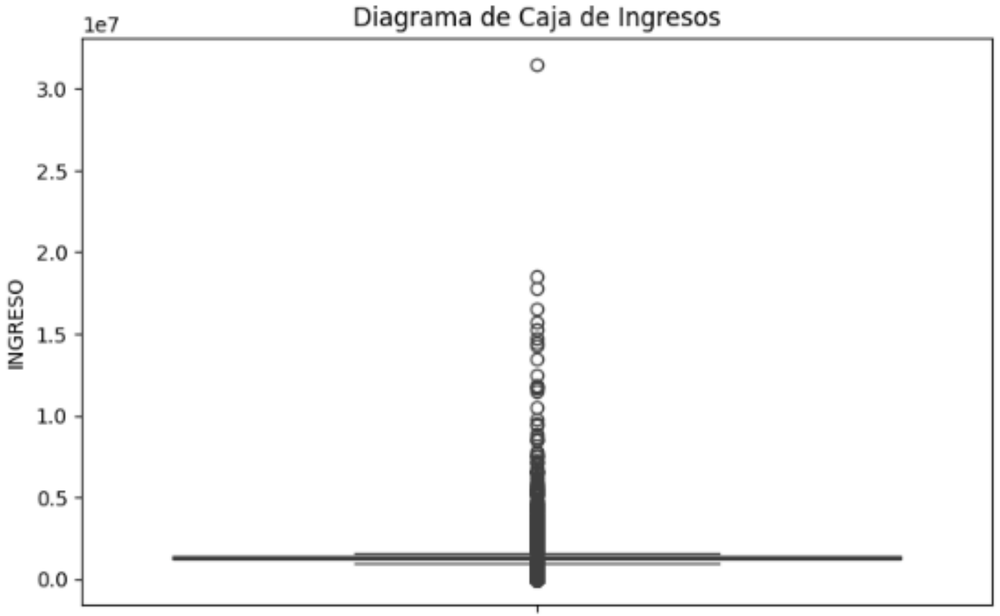



la mayoría de los ingresos se concentran entre 0.0 y 1.5, con una frecuencia máxima alrededor de 1400 individuos en el rango más bajo. Esto indica una concentración de ingresos en los niveles más bajos.

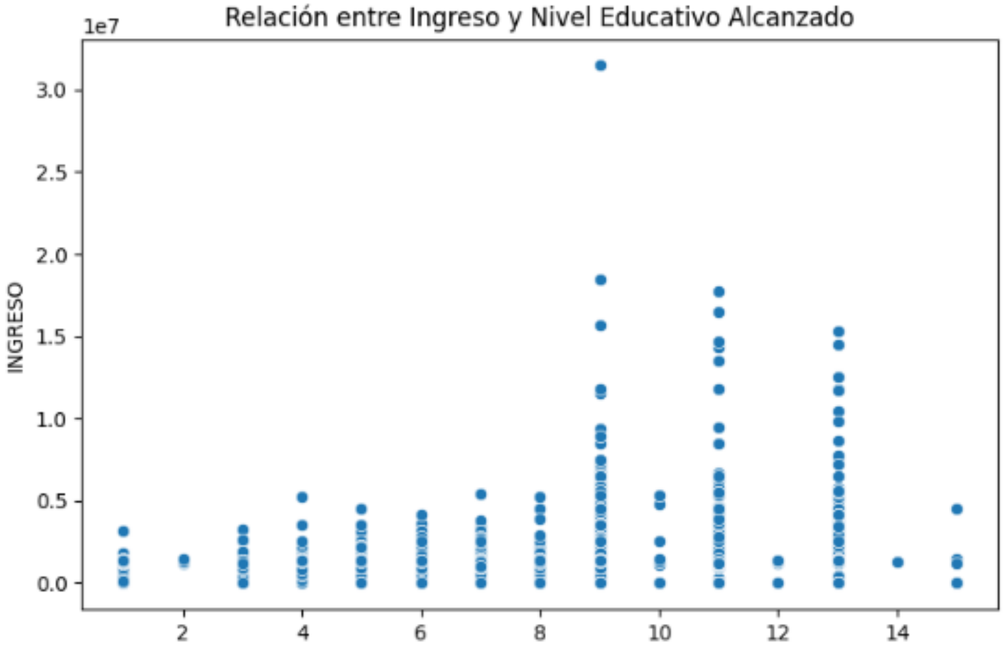



Revela que el nivel educativo más común está entre 2 y 5 (preescolar y bachillerato), con una frecuencia de alrededor de 5000 individuos. Esto sugiere que la mayoría de la población tiene un nivel educativo básico o intermedio.

# Ingresos

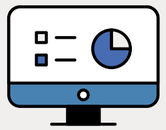



Indica que la mediana de los ingresos está alrededor de 0.3, con algunos valores atípicos que superan 1.0. Esto muestra una distribución sesgada hacia ingresos más bajos.



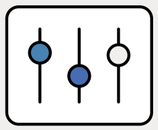

Muestra que a medida que aumenta el nivel educativo, los ingresos tienden a ser más altos, aunque la relación no es perfectamente lineal. Los ingresos más altos se observan en niveles educativos entre 8 y 12.

# Fase 3



## SELECCIÓN DEL MODELO

Se seleccionará un modelo predictivo, como series temporales o regresión, según el análisis de correlación.



## ENTRENAMIENTO DEL MODELO

Los datos recolectados se usarán para entrenar y optimizar el modelo

## VALIDACIÓN DEL MODELO

La precisión y confiabilidad del modelo se evaluarán con métricas como RMSE y $R^2$.

Sabana Centro Cómo Vamos

# Modelos utilizados

- **Regresión Lineal:**

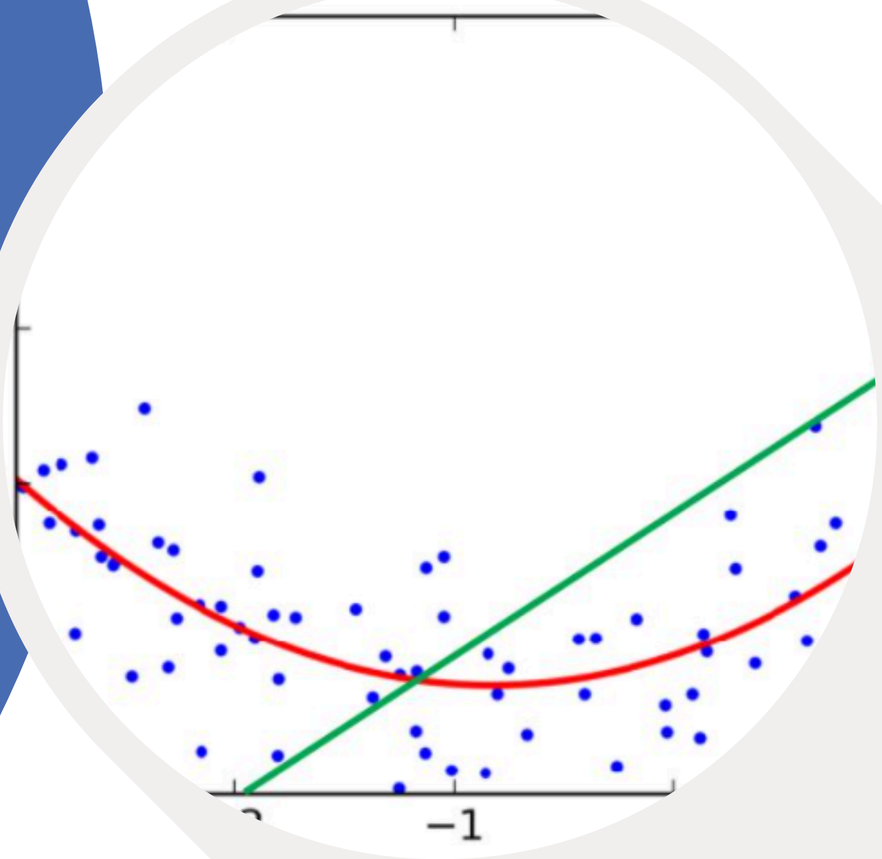
Modelo que estima una relación lineal entre el nivel educativo alcanzado y el ingreso, permitiendo predecir valores continuos.

- **Regresión Polinomica:**

**E**s una extensión de la regresión lineal que permite modelar relaciones no lineales entre variables.
La diferencia con la regresión lineal, que asume una relación directa, la regresión polinómica incorpora términos elevados a distintas potencias, lo que permite capturar patrones más complejos.

- **Regresión Logística**

El modelo de regresión logística se usa para clasificar datos en dos o más categorías. A diferencia de la regresión lineal, que predice valores numéricos, la regresión logística estima probabilidades y asigna una clase según un umbral (por ejemplo, 0 o 1)




Estos modelos nos permiten tanto predecir valores específicos de ingresos como clasificar tendencias en los datos, facilitando la toma de decisiones estratégicas.

# Generación de modelo de regresion lineal

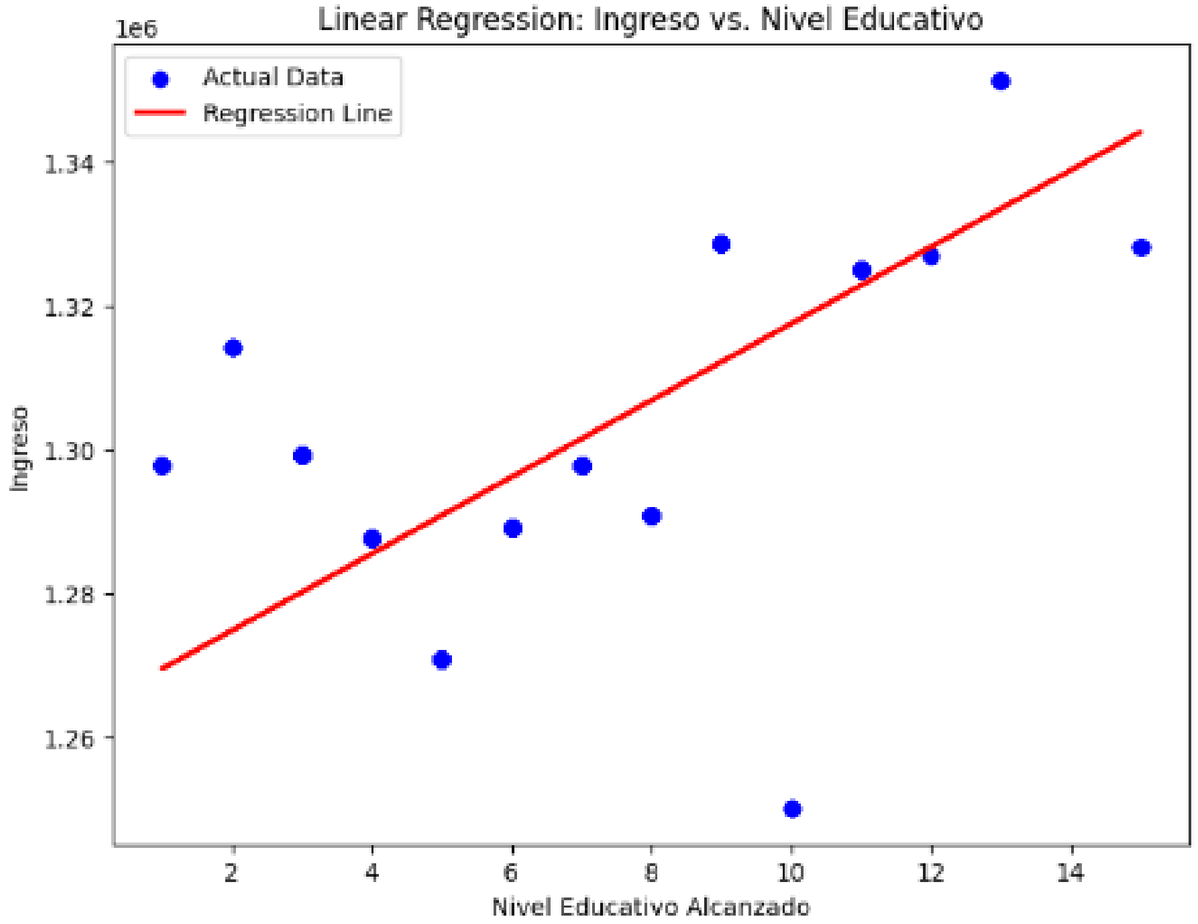
- Preparación de Datos: Se selecciona el nivel educativo como variable independiente (X) y el ingreso como variable dependiente (y).

- División de Datos: Los datos se dividen en conjuntos de entrenamiento (70%) y prueba (30%).

- Evaluación del Modelo: Se calcula el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación ($R^2$) para evaluar el rendimiento del modelo.

- Entrada: Nivel educativo

- Salida: Predicciones de ingreso y métricas (MSE, $R^2$).




# Modelo de regresion lineal

- El modelo tiene un MSE de 272,327,665.83 (bastante alto) y un $R^2$ de 0.37, lo que indica que el nivel educativo explica el 37% de la variabilidad en el ingreso.
- La gráfica muestra una tendencia positiva: a mayor nivel educativo, mayor ingreso.

# Generación de modelo de regresion polinomica

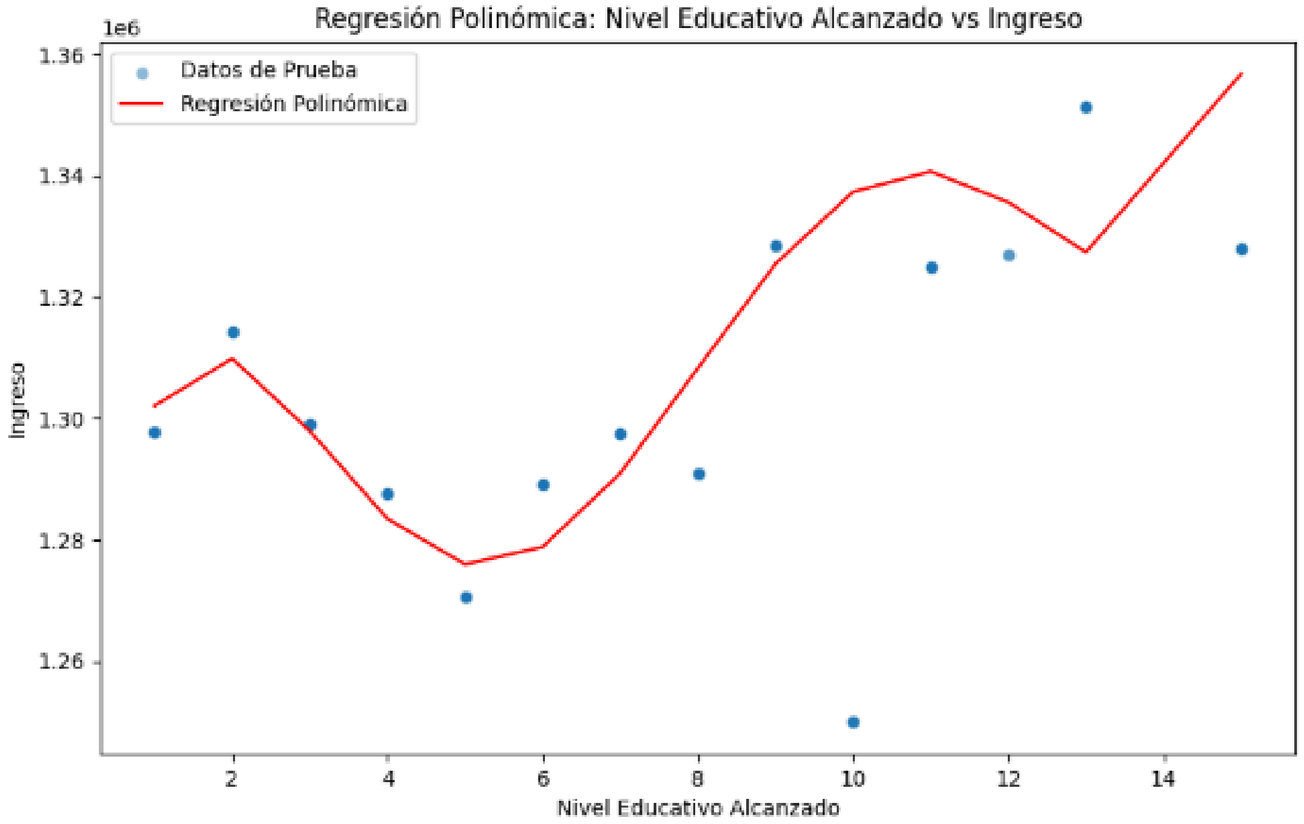
- Preparación de Datos: Se selecciona el nivel educativo como variable independiente (X) y el ingreso como variable dependiente (y).

- División de Datos: Los datos se dividen en conjuntos de entrenamiento (75%) y prueba (25%).

- Evaluación del Modelo: Se calcula el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación ($R^2$) para evaluar el rendimiento del modelo.

- Entrada: Nivel educativo y datos generales de ubicación

- Salida: Predicciones de ingreso y métricas (MSE, $R^2$).




# Modelo de regresión polinomica

- El modelo tiene un MSE de 59,368,643.40 (menor que el modelo lineal, lo que indica un mejor ajuste) y un $R^2$ de 0.86, lo que sugiere que el 86% de la variabilidad en el ingreso puede explicarse por el nivel educativo cuando se utiliza un modelo polinómico de grado 5.

- La gráfica muestra una curva de regresión polinómica que se ajusta mejor a los datos que la línea recta del modelo lineal, capturando relaciones no lineales entre el nivel educativo y el ingreso.

# Generación de modelo de regresion logistica

- En este caso, seleccionamos NIVEL_EDUCATIVO_ALCANZADO y MUNICIPIOcomo la variable independiente (X) e INGRESO (definida como 1 para ingresos por encima de la mediana y 0 para ingresos por debajo) como la variable dependiente (y).

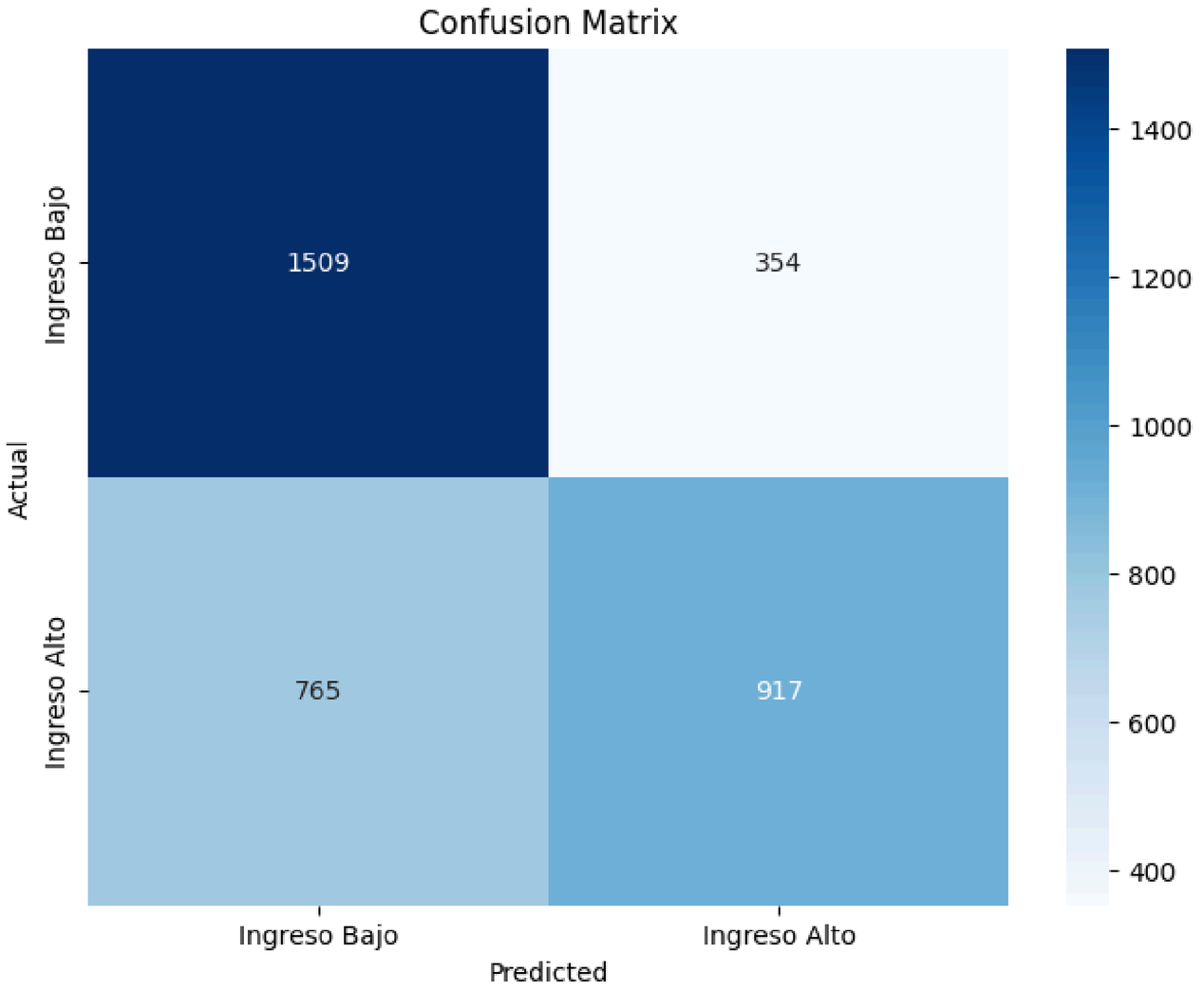
- Usamos la función train_test_split para dividir el conjunto de datos en conjuntos de entrenamiento y prueba. Aquí se utiliza un 80% para entrenamiento y un 20% para prueba, con una semilla aleatoria para reproducibilidad.

- Creamos una instancia del modelo LogisticRegression y lo entrenamos utilizando los datos de entrenamiento.

- Realizamos predicciones utilizando el conjunto de prueba (X_test) y evaluamos el rendimiento del modelo utilizando métricas como la matriz de confusión y el informe de clasificación.

- Salida: Las predicciones del modelo se refieren a INGRESO, donde se clasifica si un individuo tiene un ingreso alto (1) o bajo (0) basado en el modelo entrenado.

# Modelo de regresión logistica




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.66 | 0.81 | 0.73 | 1863 |
| 1 | 0.72 | 0.55 | 0.62 | 1682 |
| accuracy | | | 0.68 | 3545 |
| macro avg | 0.69 | 0.68 | 0.68 | 3545 |
| weighted avg | 0.69 | 0.68 | 0.68 | 3545 |

Sabana Centro Cómo Vamos

# El por qué de nuestras desiciones

Las decisiones tomadas en este proyecto se basan en un análisis exhaustivo de datos y en la aplicación de modelos predictivos confiables. Seleccionamos las estrategias óptimas considerando:

- Correlaciones clave que nos permitieron entender mejor los factores sobre la calidad de vida
- Hemos elegido un modelo de regresión polinómica porque nos permite capturar la relación no lineal entre el nivel educativo alcanzado y el ingreso.
- A diferencia de la regresión lineal, que asume una relación lineal, la regresión polinómica puede ajustarse a curvas, lo que nos permite modelar relaciones más complejas presentes en nuestros datos.

- Resultados del modelo predictivo

- Impacto y viabilidad

# Bibliografia


https://aprendeia.com/algoritmo-regresion-polinomial-machine-learning-practica-con-python/

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://platzi.com/tutoriales/1766-regresion-python/11159-de-donde-viene-el-algoritmo-de-regresion-lineal/

https://microdatos.dane.gov.co/index.php/catalog/743/get-microdata