

CLASIFICACIÓN DE CLIENTES POR POTENCIALES INGRESOS





01 OBJETIVO

¿Que impulsa este trabajo?
¿Que se busca determinar?

02 EDA

¿Qué nos dicen los datos?

03 MODELOS

¿Qué modelos se
desarrollaron?

04 CONCLUSIONES

¿Qué resultados se
obtienen con los modelos?

01. OBJETIVO

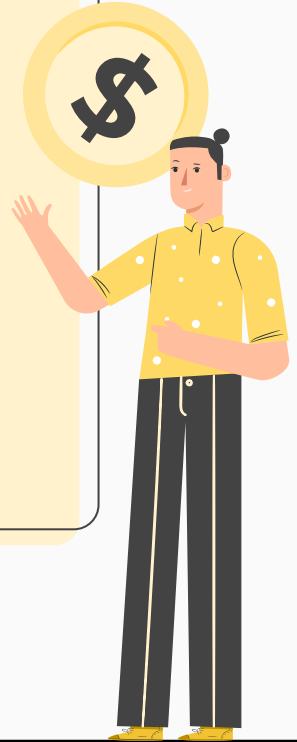
¿Que impulsa este trabajo?
¿Que se busca determinar?



INTRODUCCIÓN

Este trabajo busca crear modelos de clasificación de posibles clientes de empresas relacionadas al mundo financiero (bancos, fintechs y otros) con el fin de poder ofrecerles productos en base a su nivel de ingresos estimado.

Al clasificar a los clientes, además, se pueden implementar estrategias de marketing segmentadas a cada grupo, en las cuales se les ofrecen productos a medida de sus ingresos.



OBJETIVO DEL TRABAJO

En este proyecto se desarrolla un sistema de **clasificación de ingresos personales** basado en modelos de *machine learning*, cuyo objetivo es predecir si un individuo gana **más o menos de \ \$50K anuales** utilizando datos demográficos y laborales extraídos del censo.

El sistema está diseñado para analizar registros individuales y, mediante un algoritmo de **clasificación supervisada**, identificar patrones socioeconómicos relevantes que puedan ser útiles en estudios de mercado, políticas públicas, o decisiones comerciales.

OBJETIVO DEL TRABAJO

Los objetivos específicos son:

- Desarrollar un modelo capaz de **predecir el nivel de ingresos** de una persona a partir de datos censales.
- Aplicar técnicas de **preprocesamiento, entrenamiento y evaluación** de modelos de clasificación binaria.
- Diseñar un pipeline reproducible que permita aplicar el modelo en nuevos conjuntos de datos.

El proyecto incluye dentro de su alcance:

- Exploración de datos (EDA) y visualización de variables clave.
- Entrenamiento y validación de modelos de clasificación (por ejemplo: Regresión Logística, Árboles de Decisión, Random Forest).
- Evaluación de métricas como *accuracy*, *precision*, *recall*, y *f1-score*.
- Identificación de las variables más relevantes mediante técnicas de importancia de características.

02. EDA

Exploratory Data Analysis



DATASET: ADULT CENSUS INCOME

Se utiliza el *dataset* **Adult Census Income**^{**}, extraído del Censo de EE. UU. de 1994 y preparado por Ronny Kohavi y Barry Becker para tareas de minería de datos.

Características principales del *dataset*:

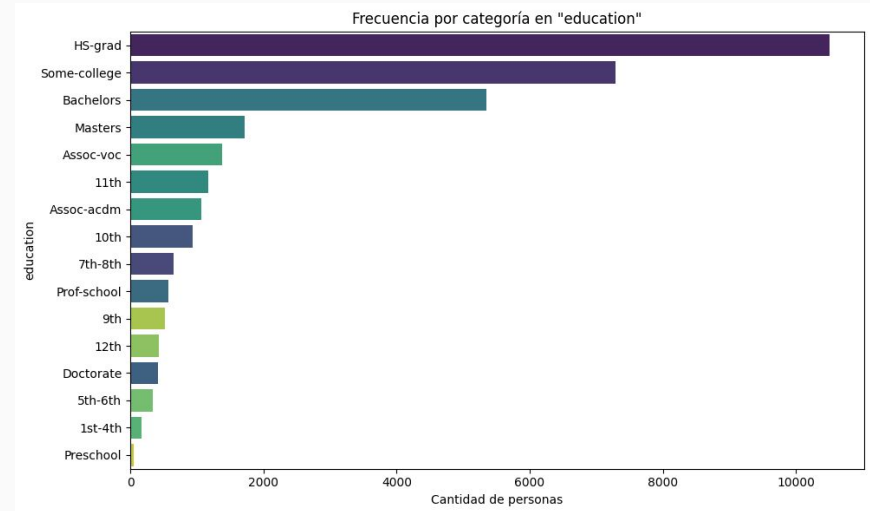
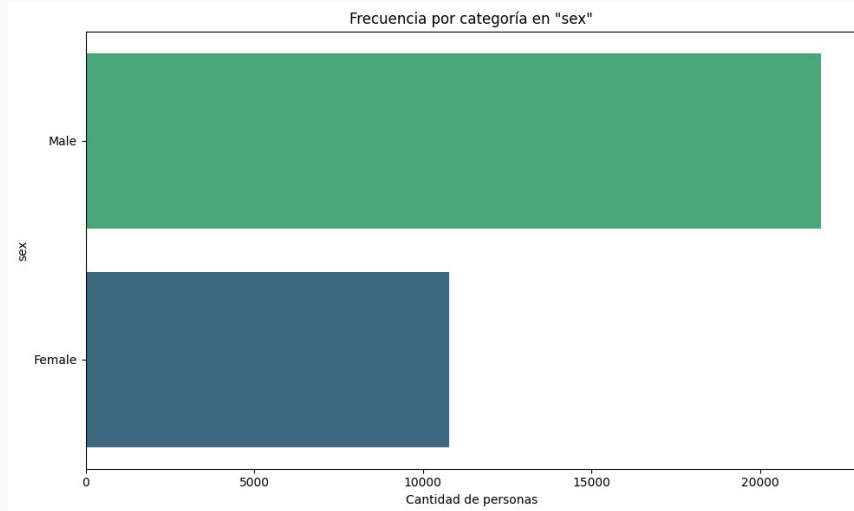
- **Número de instancias:** 32561 registros.
- **Atributos:** 14 variables incluyendo edad, educación, ocupación, horas trabajadas por semana, entre otros.
- **Variable objetivo:** *income*, que indica si el individuo gana `>50K` o `<=50K` al año.
- **Tipo de datos:** Mixto (categóricos y numéricos).
- **Fuente original:** Base de datos de la Oficina del Censo de los Estados Unidos.

ATRIBUTOS DEL DATASET

- **Age:** Variable numérica que representa la edad del individuo.
- **Workclass:** Variable categórica que determina el tipo de empleo (privado, gobierno, autónomo, etc.).
- **Fnlwgt:** Variable numérica que representa el peso muestral (indica cuántas personas representa esta muestra).
- **Education:** Variable categórica que determina el nivel educativo (HS-grad, Bachelors, etc.).
- **Education-num:** Variable numérica que representa el nivel educativo en formato numérico.
- **Marital-status:** Variable categórica que determina el estado civil del individuo.
- **Occupation:** Variable categórica que determina la ocupación laboral.
- **Relationship:** Variable categórica que determina la relación familiar (esposo/a, hijo/a, etc.).
- **Race:** Variable categórica que determina la raza declarada.
- **Sex:** Variable categórica que determina el Género (Male/Female).
- **Capital-gain:** Variable numérica que representa ganancias de capital obtenidas.
- **Capital-loss:** Variable numérica que representa pérdidas de capital registradas.
- **Hours-per-week:** Variable numérica que determina la cantidad de horas trabajadas por semana.
- **Native-country:** Variable categórica que expresa el país de origen.
- **Income:** Variable categórica y objetivo que representa el ingreso del individuo: `>50K` o `<=50K` (clase a predecir).

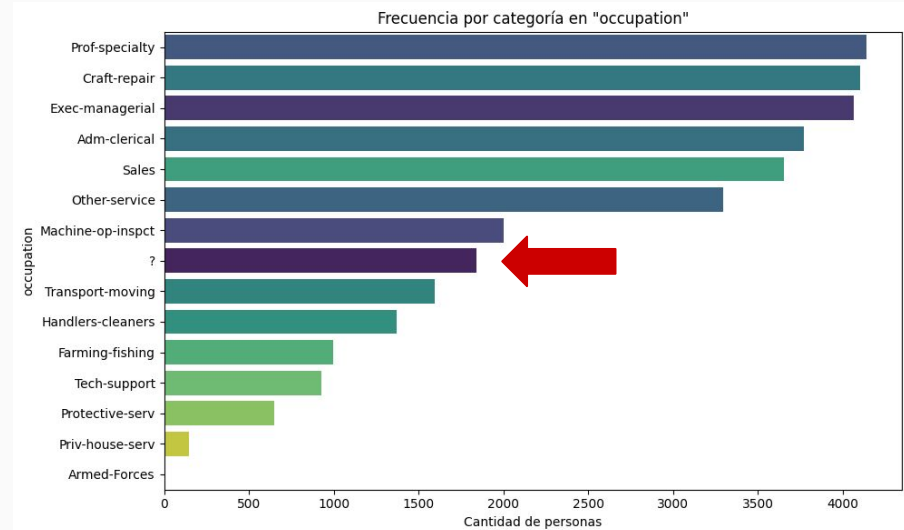
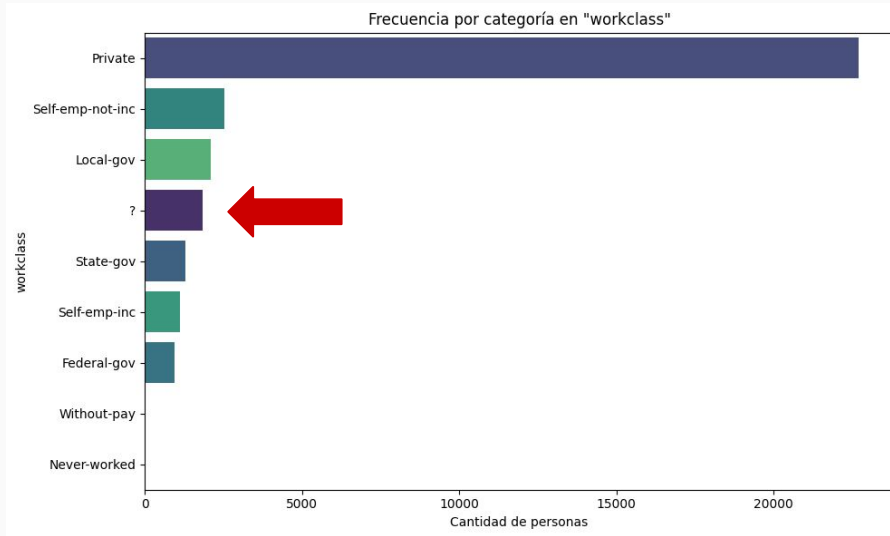
ATRIBUTOS DEL DATASET

Algunos ejemplos de la atributos categóricos en el dataset



ATRIBUTOS DEL DATASET

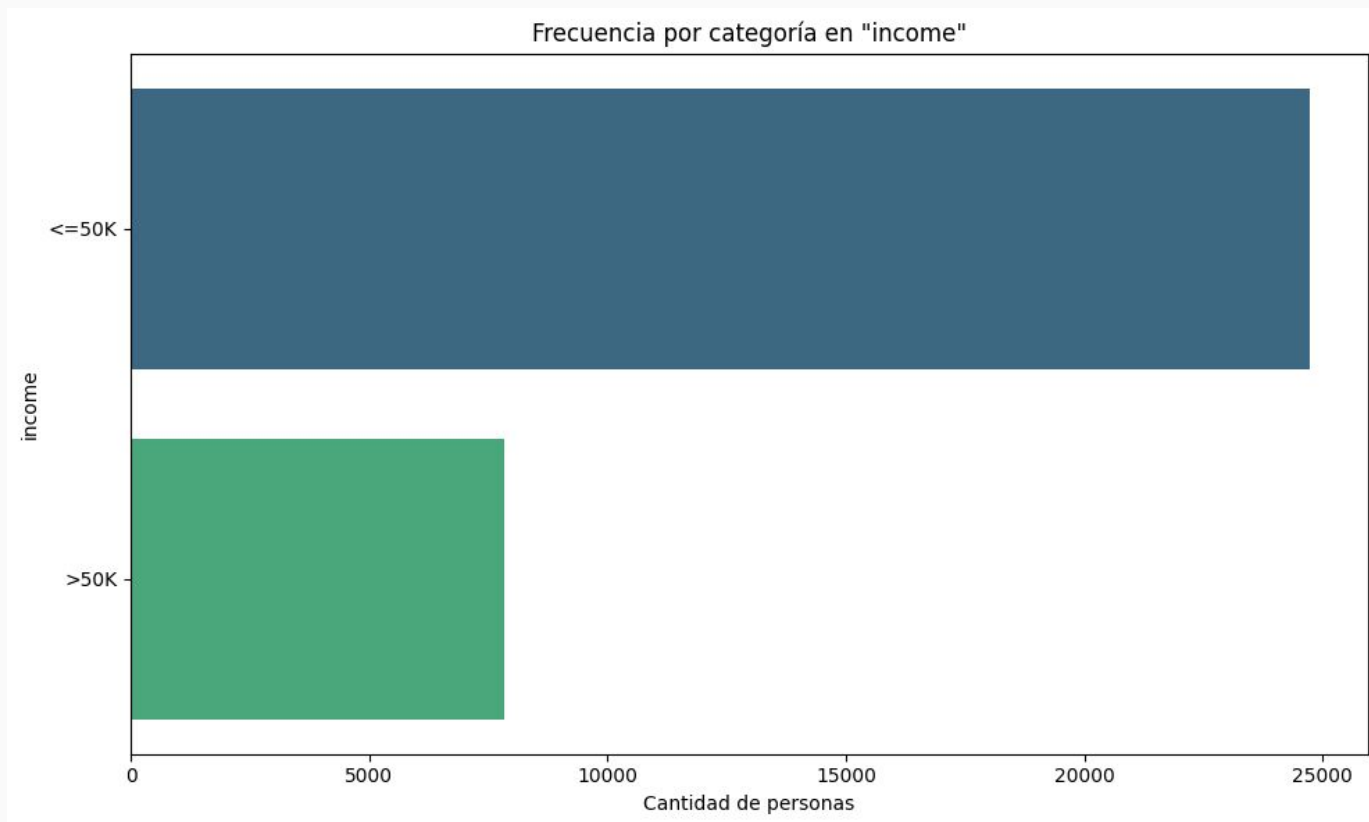
Existen datos **sin categoría definida** en los atributos *workclass* y *occupation*



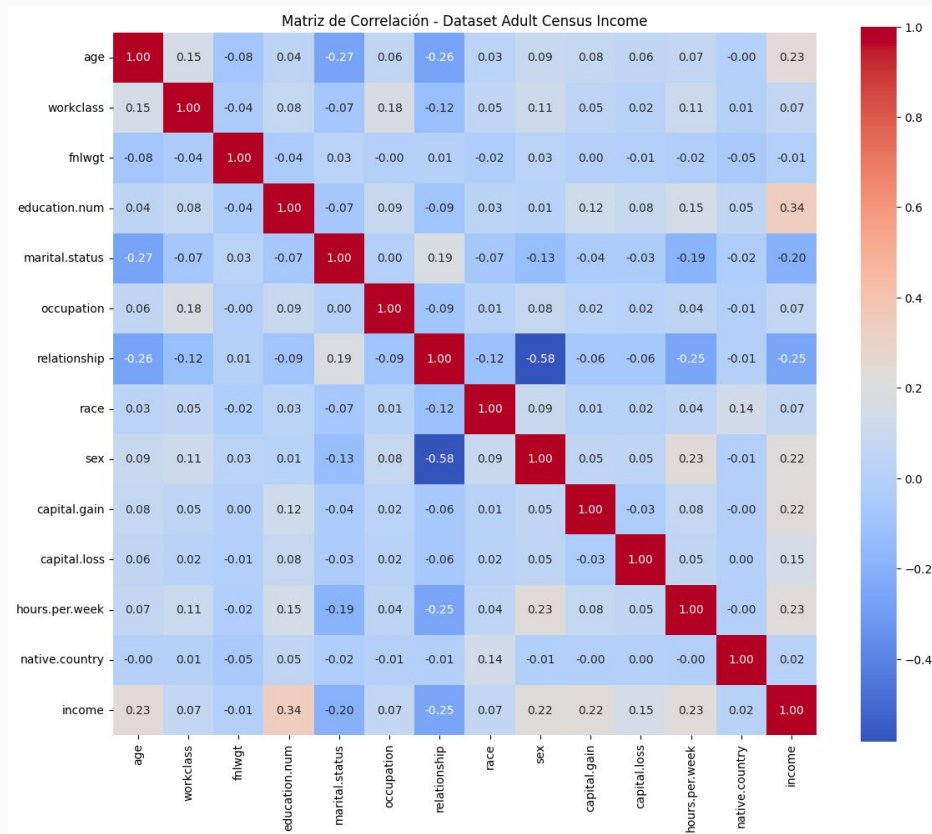
ATRIBUTOS DEL DATASET


El dataset se encuentra desbalanceado.


Esto será una complicación para entrenar el modelo para predecir personas con ingresos >50K





CORRELACIÓN ENTRE ATRIBUTOS




 Baja correlación general: La mayoría de los atributos del dataset presentan correlaciones muy bajas (≤ 0.10), por lo que no son significativas.

 Mayor correlación positiva: Entre education e income (0.34).

 Mayor correlación negativa: Entre sex y relationship (-0.58).

 Atributos más conectados: age, marital.status, sex, hours.per.week e income muestran más relaciones con otras variables.

 education.num tiene baja correlación con casi todos los atributos, excepto con income.

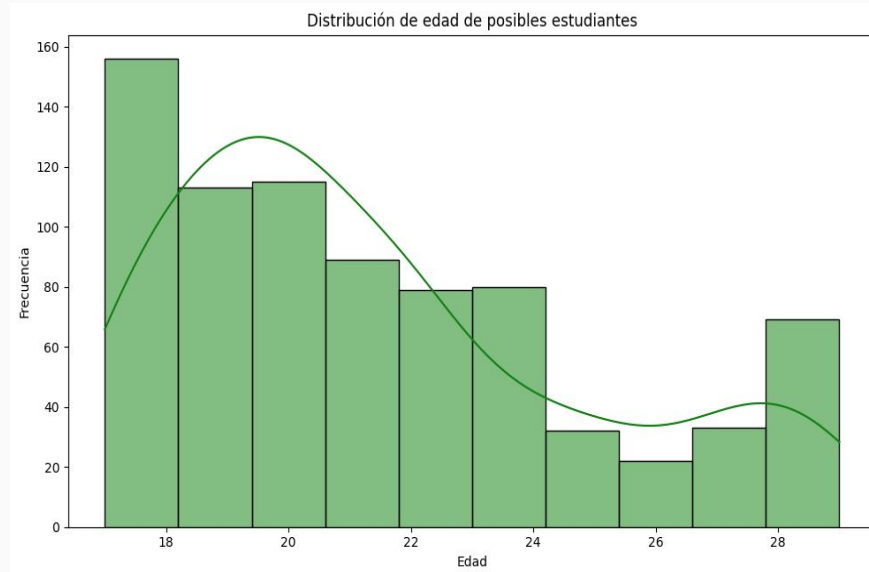
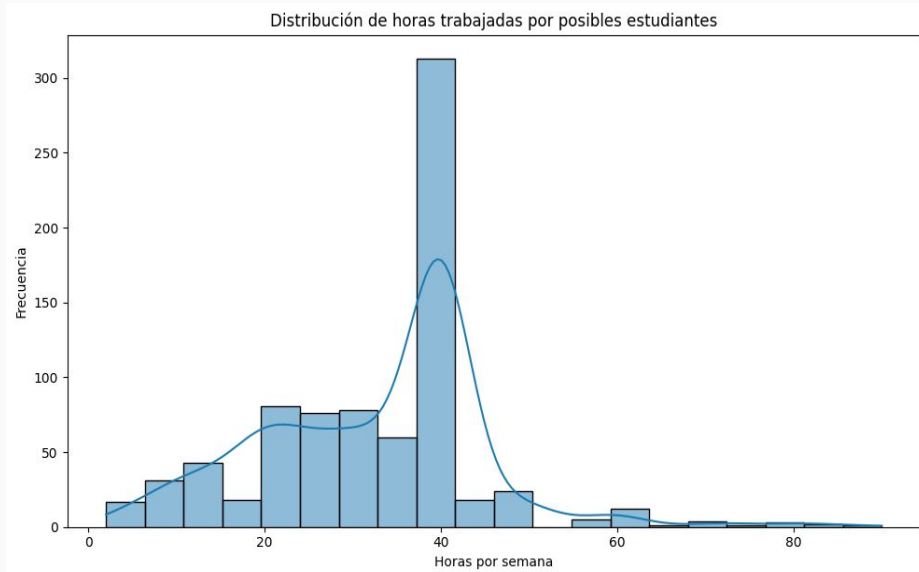
HIPÓTESIS



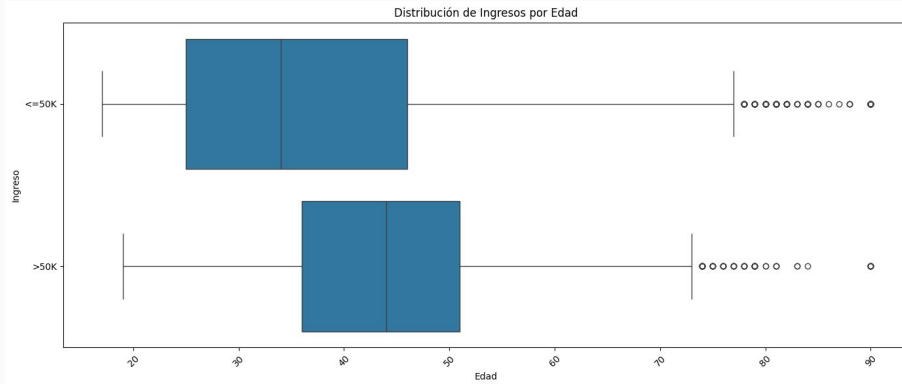
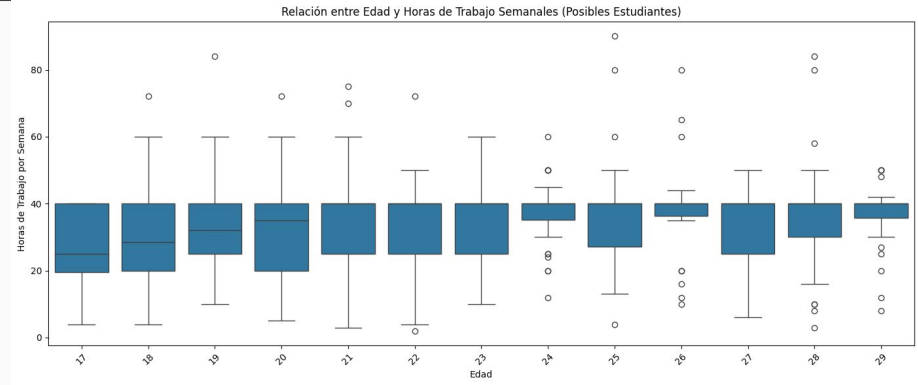
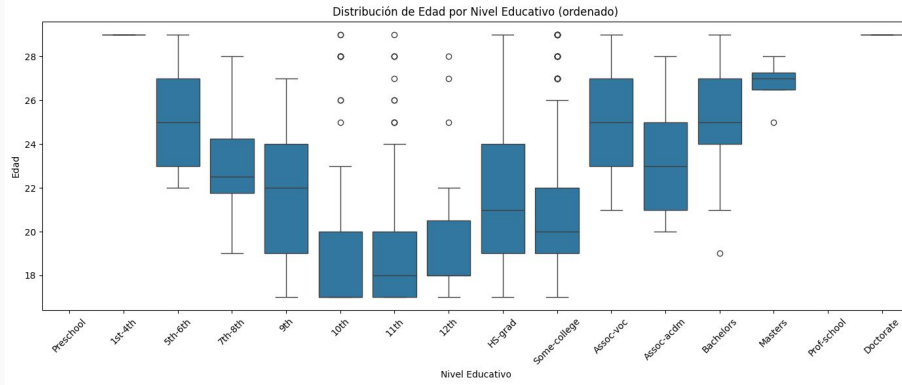
Hipótesis sobre estudiantes: Personas jóvenes, sin ocupación ni clase laboral definida (?), que trabajan horas semanales y no están casadas, podrían ser estudiantes activos no representados explícitamente.



Nuevo subconjunto: Este grupo se almacena como **posibles estudiantes** para análisis posterior.



ANÁLISIS DE POSIBLES ESTUDIANTES



Nuevas categorías creadas para occupation y workclass iguales a “?”:

student: Personas de 17 a 22 años, con educación Some-college, ingresos ≤50K y entre 20–40 horas semanales trabajadas.

informal: Personas con más de 25 horas semanales trabajadas, excluyendo estudiantes (Some-college), sin restricción de edad o nivel educativo.

unemployed: Personas con menos de 25 horas semanales trabajadas, también excluyendo estudiantes, sin restricción de edad o educación.

03. MODELOS

Modelos Seleccionados



MODELOS SELECCIONADOS

Con el fin de predecir si una persona obtiene ingresos mayores a 50K anuales, se evaluarán los siguientes modelos de clasificación supervisada:

- **Regresión Logística:** Modelo base para clasificación binaria, útil como referencia inicial.
- **Árboles de Decisión (Decision Tree Classifier):** Permite interpretar decisiones basadas en reglas y condiciones.
- **Random Forest:** Ensamble de árboles de decisión que mejora la generalización y reduce el overfitting.
- **Gradient Boosting (XGBoost o GradientBoostingClassifier):** Modelo basado en boosting, que corrige los errores de modelos anteriores.
- **Support Vector Machines (SVM):** Eficiente en espacios de alta dimensión, especialmente con kernels no lineales.
- **K-Nearest Neighbors (KNN):** Modelo basado en la similitud de instancias vecinas, útil para entender la estructura local.

MÉTRICAS A EVALUAR

Cada modelo será evaluado utilizando las siguientes métricas de rendimiento:

- **Accuracy (Exactitud):** Proporción de predicciones correctas sobre el total de instancias.
- **Precision (Precisión):** Proporción de verdaderos positivos sobre el total de predicciones positivas realizadas.
- **Recall (Sensibilidad o Tasa de Verdaderos Positivos):** Proporción de verdaderos positivos sobre el total de instancias realmente positivas.
- **F1-Score:** Media armónica entre precisión y recall, útil ante clases desbalanceadas.
- **Matriz de Confusión:** Visualización de los aciertos y errores del modelo en cada clase.

04. CONCLUSIONES

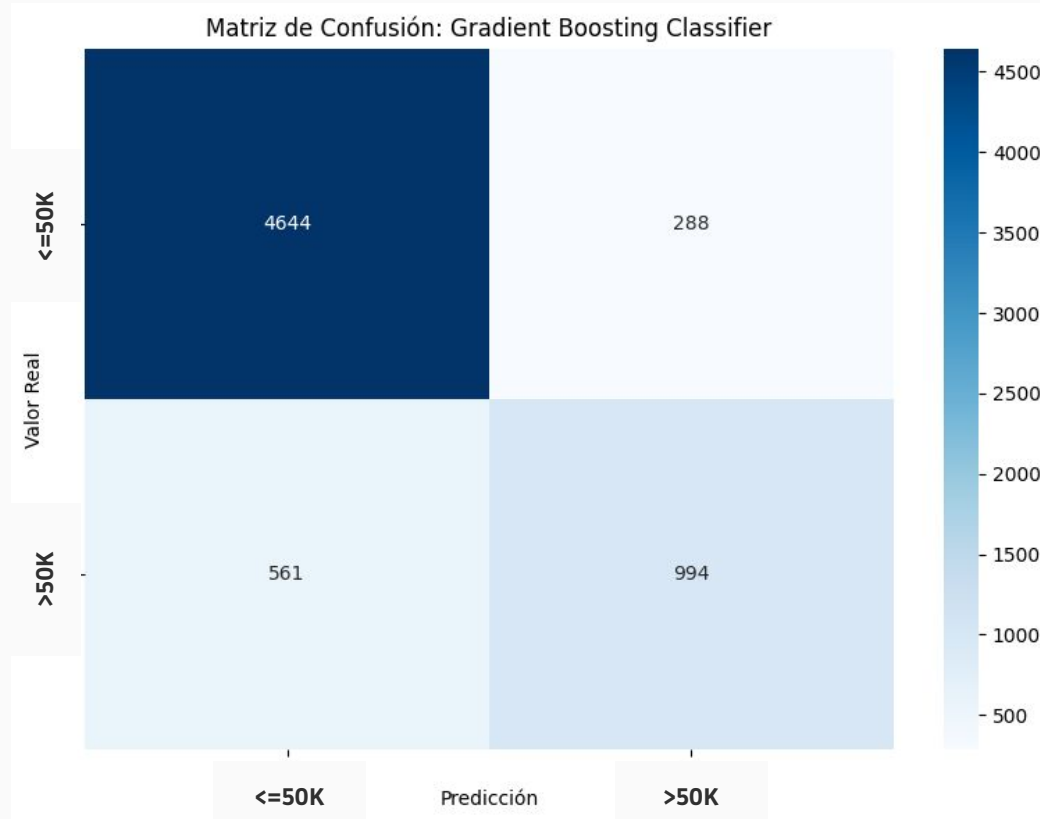
¿Cual es el mejor modelo?











RESULTADO DE LOS MODELOS

Modelo	Clase	Precision	Recall	F1-Score
Logistic Regression	<=50K	0.8818	0.927	0.9038
Logistic Regression	>50K	0.7235	0.6058	0.6594
Decision Tree	<=50K	0.8796	0.8783	0.879
Decision Tree	>50K	0.6159	0.6186	0.6173
Random Forest	<=50K	0.8859	0.926	0.9055
Random Forest	>50K	0.726	0.6219	0.6699
SVM	<=50K	0.8769	0.9376	0.9062
SVM	>50K	0.7463	0.5826	0.6544
KNN	<=50K	0.8779	0.9065	0.892
KNN	>50K	0.6693	0.6	0.6328
XGBoost	<=50K	0.8922	0.9416	0.9162
XGBoost	>50K	0.7754	0.6392	0.7007

RESULTADO DE LOS MODELOS



CONCLUSIONES

-  Todos los modelos predicen muy bien la clase $\leq 50K$, con **f1-scores superiores a 0.88**.
-  **SVM** logró el **mejor recall** para $\leq 50K$ (0.9376), pero
-  **XGBoost** fue el **modelo más equilibrado**, con el mayor f1-score para $\leq 50K$ (0.9162).
-  **La clase $>50K$ fue más difícil de predecir** para todos los modelos, con f1-scores significativamente menores.
-  **XGBoost también lideró para la clase $>50K$** , con mejores valores de precision (0.7754), recall (0.6392) y f1-score (0.7007).
-  **Decision Tree y KNN** mostraron el **rendimiento más bajo** para esta clase minoritaria.
-  El desbalance de clases afectó la calidad de predicción, especialmente para $>50K$.
-  Se sugiere implementar técnicas de balanceo (oversampling, undersampling, pesos ajustados) para mejorar el rendimiento en futuras versiones.