

# Análisis Exploratorio de Datos (EDA)

## 1. Información General del Dataset

- **Número de filas:** 40,841
- **Número de columnas:** 18
- **Sin valores nulos ni duplicados**
- **Variable objetivo:** response / response\_binary

## 2. Variables Numéricas

### Estadísticas principales:

Variable	Mín	Máx	Media	Desv. Est.
Age	18	95	~40.8	~10.5
Balance	-6847	10443	~1073	~1712
Duration	0.1	81.9	~4.31	~4.31
Campaign	1	58	~2.77	~3.10

## 3. Visualizaciones Destacadas

### Respuesta por Educación

Clientes con educación terciaria tienen mayor tasa de respuesta.

### Respuesta por Trabajo

Estudiantes y jubilados muestran mayor probabilidad de responder.

### Duración de llamada

Duraciones más largas están claramente asociadas con una mayor tasa de respuesta.

### Boxplots comparativos

Visualización de age, balance, duration, campaign vs response\_binary mostró:

- duration destaca como variable predictiva fuerte.
- balance y campaign muestran más dispersión sin una separación tan clara.

## 4. Conclusiones

- El dataset está limpio y bien estructurado.
  - duration, education, y job son variables clave en la predicción de respuesta.
  - El modelo de regresión logística mostró buena **precisión general (89.7%)**, pero bajo **recall para clase positiva (31.7%)**.
- 

## 5. Recomendaciones

- Considerar balancear el dataset para mejorar el recall (SMOTE, penalización de clase).
  - Incluir visualizaciones interactivas (Dash, Streamlit) para presentaciones de negocio.
  - Usar modelos más complejos (Random Forest, XGBoost) si se desea priorizar performance.
- 

**Autor:** Análisis automatizado con ChatGPT + Python

**Fecha:** 2025