Benemérita Universidad Autónoma de Puebla

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

INGENIERIA EN CIENCIA DE DATOS

Segmentación de Clientes Bancarios y Detección de Clientes Potenciales

MATERIA: INTRODUCCION A LA CIENCIA DE DATOS

PROFESOR: JAIME ALEJANDRO ROMERO SIERRA

ALUMNO: JAIME ALEXIS HERNANDEZ FRANCO

29 de Noviembre del 2025

1. **INTRODUCCIÓN**

El propósito de este proyecto es analizar una base de datos bancaria real y aplicar técnicas de aprendizaje no supervisado, principalmente K-Means, para segmentar clientes con patrones de comportamiento similares y detectar aquellos con mayor probabilidad de aceptar una oferta de marketing. El proyecto cubre todo el ciclo de ciencia de datos: limpieza, análisis exploratorio, preprocesamiento, clustering, visualización y generación de insights accionables para áreas de negocio.

Justificación

Simplicidad: K-means es un algoritmo fácil de implementar y entender, lo que lo hace accesible para los usuarios con diferentes niveles de experiencia en análisis de datos.

Eficiencia: Es rápido y escalable, capaz de manejar grandes conjuntos de datos de manera eficiente y paralelizada.

Flexibilidad: Se puede trabajar con diferentes tipos de datos y medidas de distancia, y se puede combinar con otros métodos o técnicas.

Los bancos suelen realizar campañas masivas, pero una gran parte de los clientes contactados nunca acepta una oferta. Esto provoca:

- Pérdida de tiempo del personal.
- Aumento de costos operativos.
- Saturación y molestia del cliente.

Por ello, es necesario un método que permita dirigir esfuerzos solo hacia clientes que realmente tienen potencial.

2. **DESCRIPCIÓN DE LA BASE DE DATOS**

La base de datos contiene información demográfica, financiera y de comportamiento del cliente, incluyendo:

• Edad

• Saldo bancario

• Estado civil y ocupación

• Nivel educativo

• Historial de campañas previas

• Tipo de comunicación usada

• Resultado de la campaña (suscrito o no)

Contiene variables numéricas y categóricas, por lo que se requiere un proceso cuidadoso de preparación.

La base de datos contiene **7567** filas y **12** columnas.

3. **LIMPIEZA Y TRANSFORMACIÓN**

Se realizó una limpieza completa:

• Conversión de “saldo” a valores numéricos.

• Relleno de valores faltantes usando la mediana.

• Reducción de categorías muy poco frecuentes.

• Codificación One-Hot Encoding para variables categóricas.

• Estandarización de variables numéricas mediante StandardScaler para igualar la escala.

Columnas y su descripción

Columna	Tipo	Descripción
Edad	Numérica	Edad del cliente en años. Ayuda a segmentar perfiles demográficos.
saldo	Numérica	Dinero disponible en la cuenta bancaria. Puede influir en la capacidad o interés del cliente para adquirir productos.
campaign	Numérica	Número de contactos realizados durante la campaña actual. Mide insistencia de la campaña.
contactos_previos	Numérica	Número de contactos realizados en campañas previas. Indica historial de marketing con el cliente.
trabajo	Categórica	Ocupación del cliente (administrativo, técnico, servicios, etc.). Describe nivel socioeconómico y perfil laboral.
estado_civil	Categórica	Estado civil del cliente (soltero, casado, divorciado). Aporta información demográfica relevante.
educacion	Categórica	Nivel educativo del cliente (primaria, secundaria, universidad). Influye en decisiones financieras.
credito_vivienda	Categórica (Sí/No)	Indica si el cliente tiene un crédito hipotecario. Útil para identificar cargas financieras.
prestamo	Categórica (Sí/No)	Indica si el cliente posee un préstamo personal. Ayuda a evaluar riesgo y comportamiento financiero.
contacto	Categórica	Método de contacto (celular, teléfono). Determina el canal de comunicación utilizado.
poutcome	Categórica	Resultado de campañas previas (exitoso, fallido, inexistente). Indicador clave del comportamiento histórico.
suscrito	Categórica (yes/no)	Indica si el cliente aceptó la campaña actual. Es la variable objetivo original.
suscrito_bin	Numérica (0/1)	Versión binaria de *suscrito*, usada para cálculos y métricas.
cluster	Numérica	Grupo asignado por K-Means tras el clustering. Representa el segmento al que pertenece el cliente.
pca1	Numérica	Primer componente principal generado por PCA para visualizar grupos.
pca2	Numérica	Segundo componente principal del PCA. Permite graficar datos en 2D.
dist_center	Numérica	Distancia del cliente al centroide de su cluster. Valores menores = cliente más representativo del cluster.

4. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

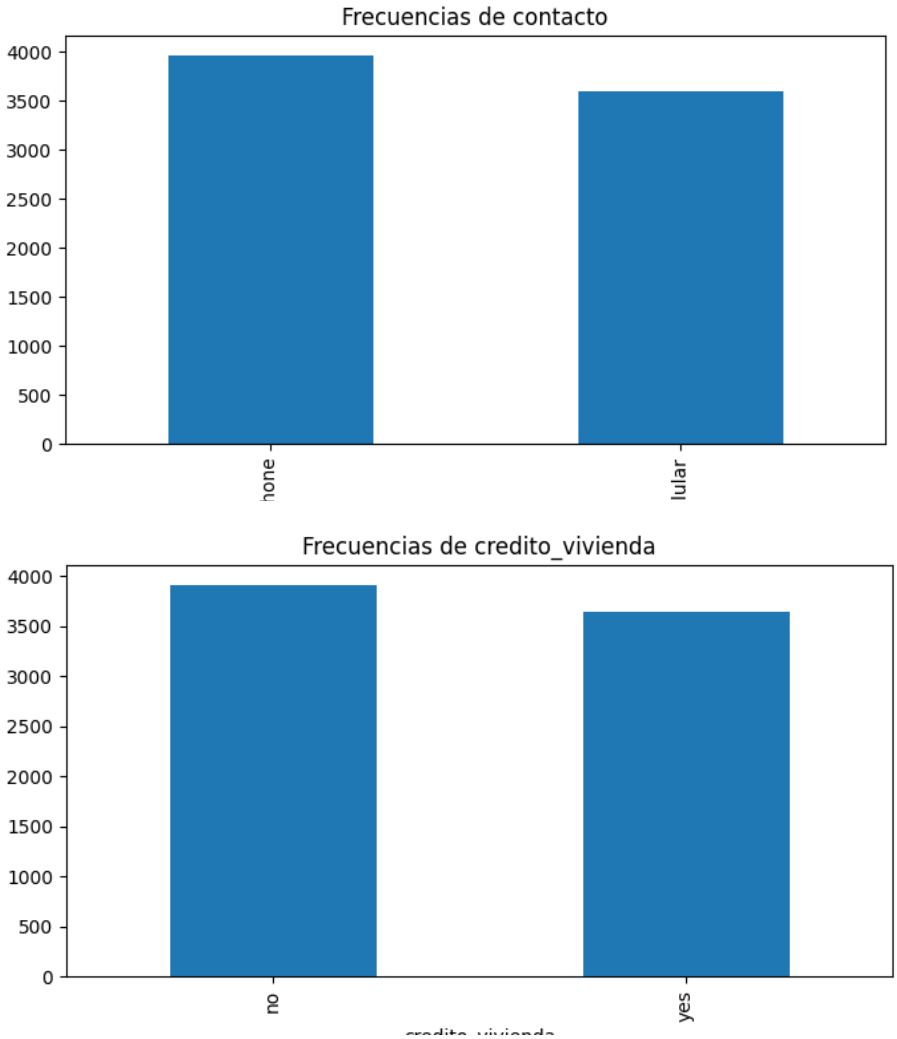
Se generaron visualizaciones y estadísticas clave:

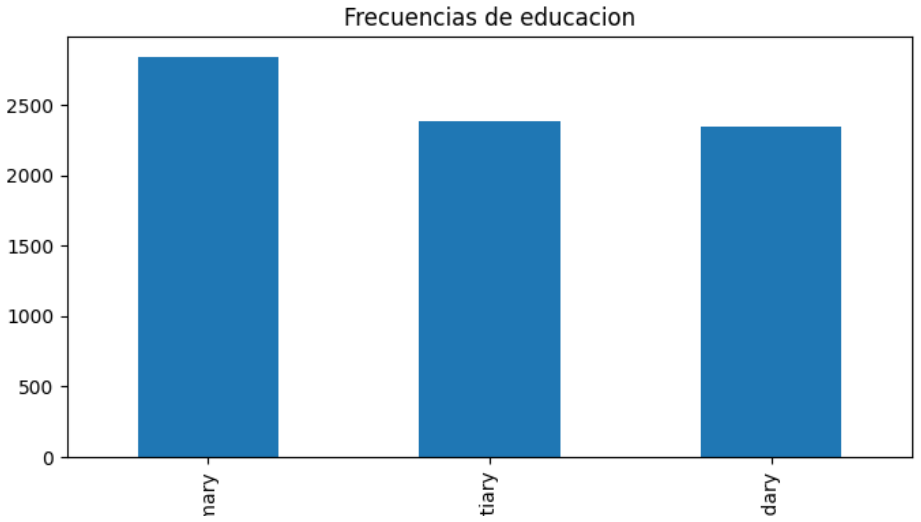
• Histogramas y boxplots para valores numéricos.

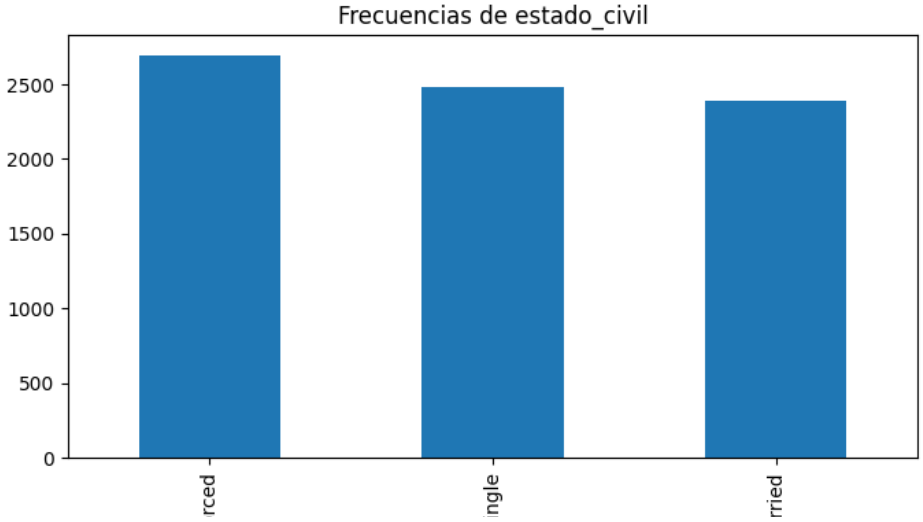
• Gráficos de barras para variables categóricas.

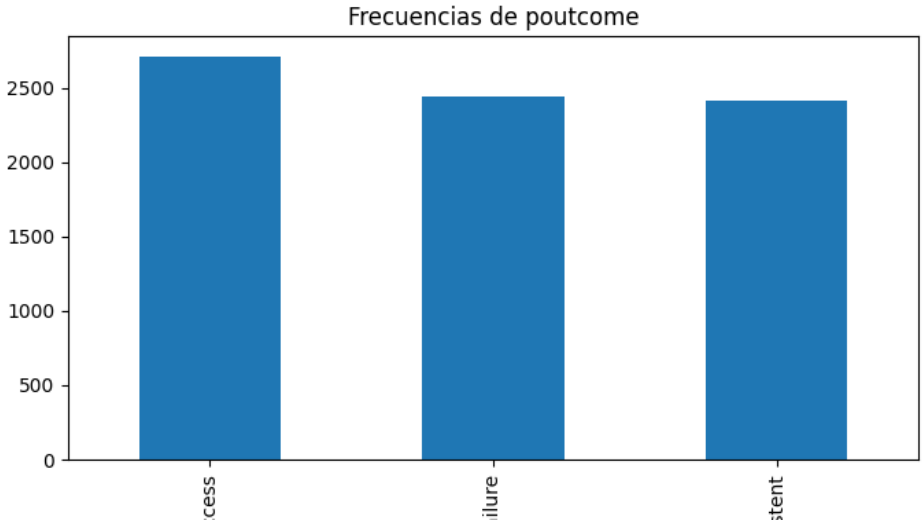
• Matriz de correlación para entender relaciones entre variables.

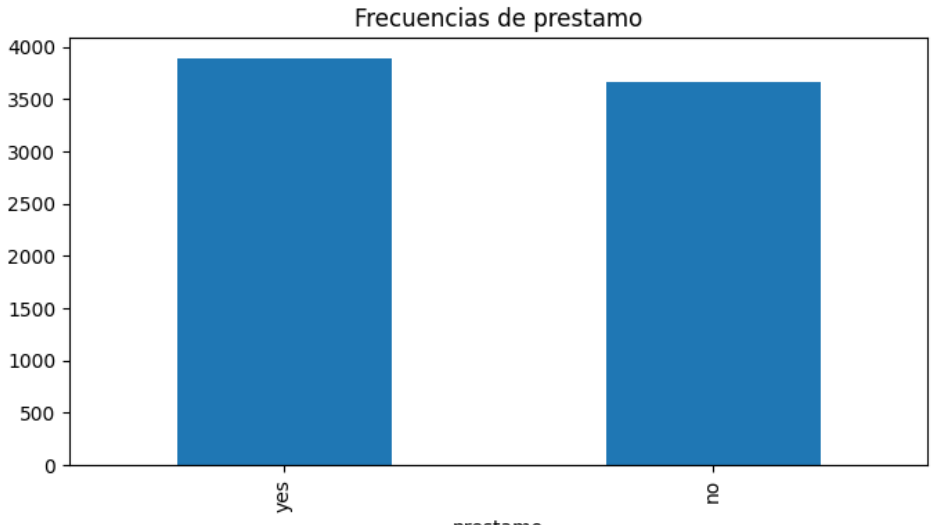
Este análisis permitió conocer la distribución y variabilidad de los datos antes del modelado.

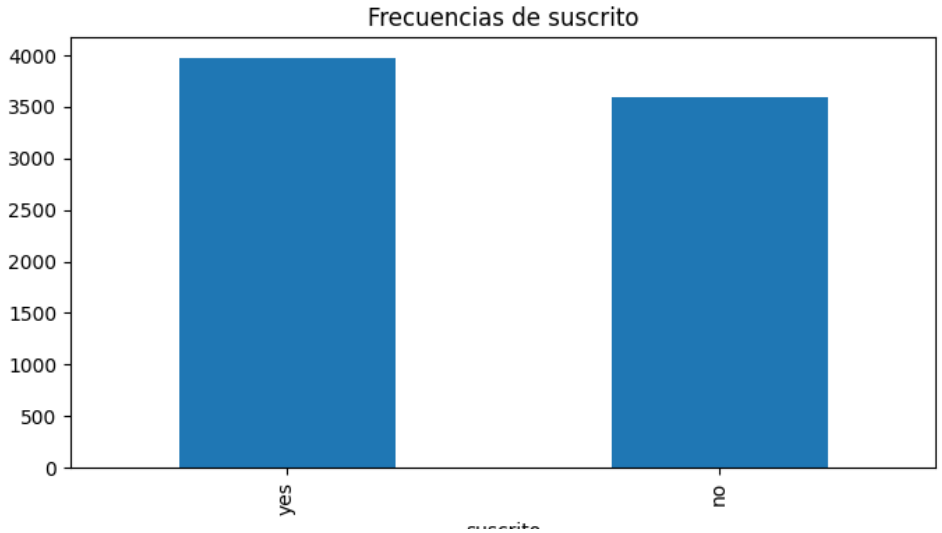


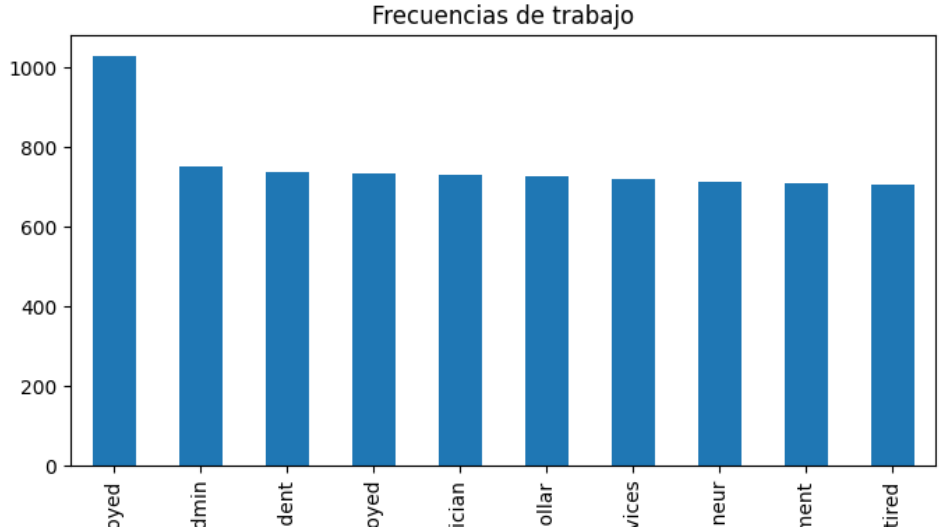

Frecuencias de educacion
2500
2000
1500
1000
500
0
nary
tiary
dary

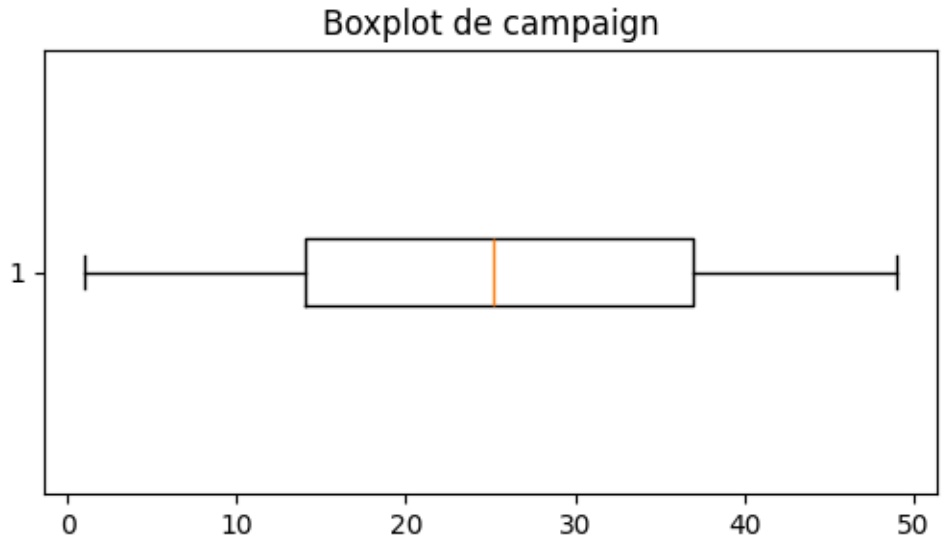

Frecuencias de estado_civil
2500
2000
1500
1000
500
0
rced
ngle
rried

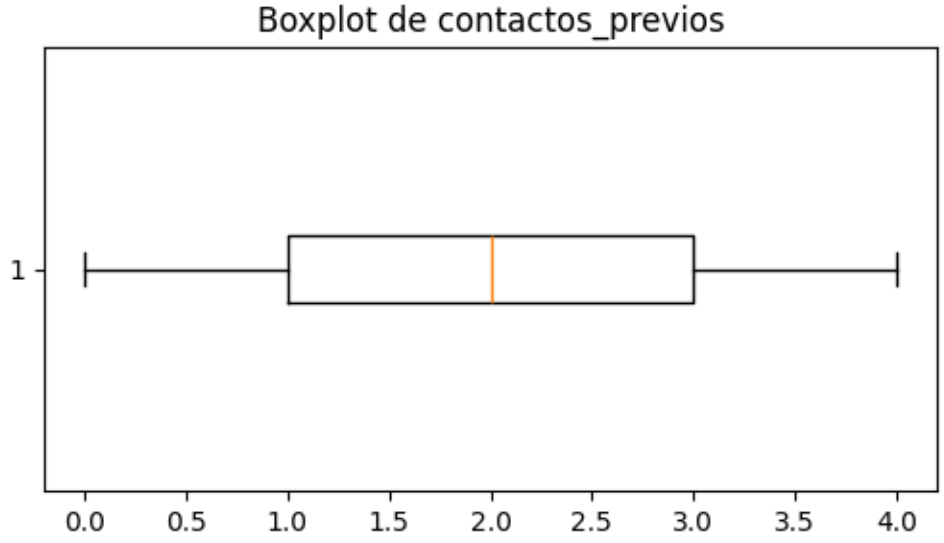

Frecuencias de poutcome
2500
2000
1500
1000
500
0
cess
ilure
stent

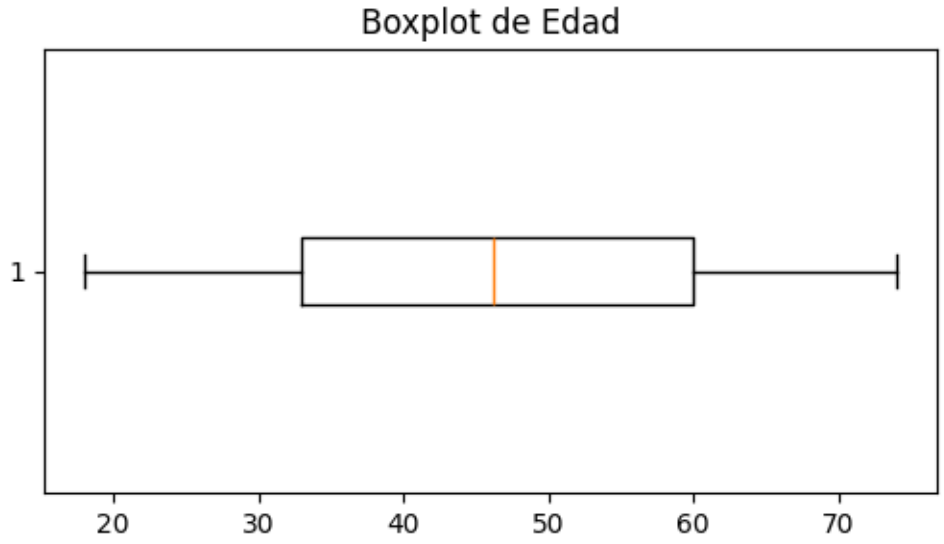

Frecuencias de prestamo
4000
3500
3000
2500
2000
1500
1000
500
0
yes
no

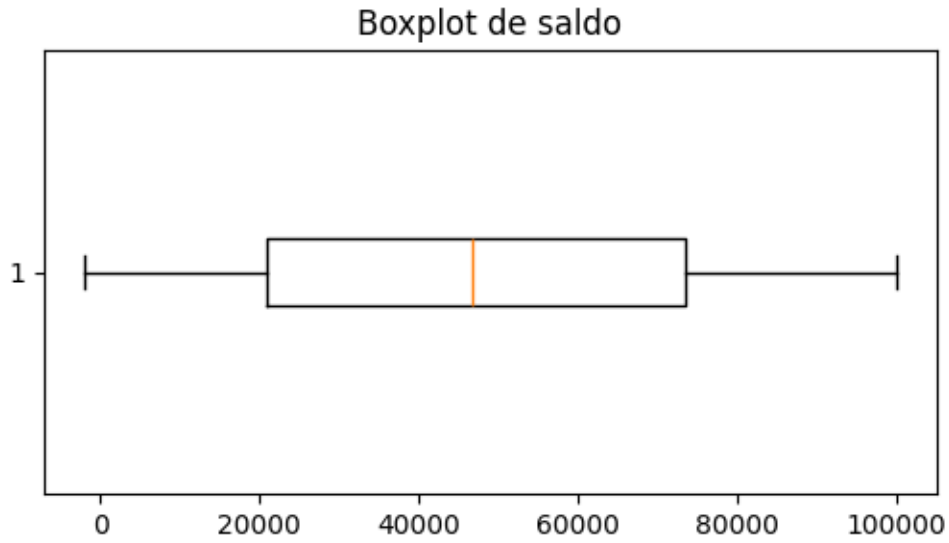

Frecuencias de suscrito
4000
3500
3000
2500
2000
1500
1000
500
0
yes
no

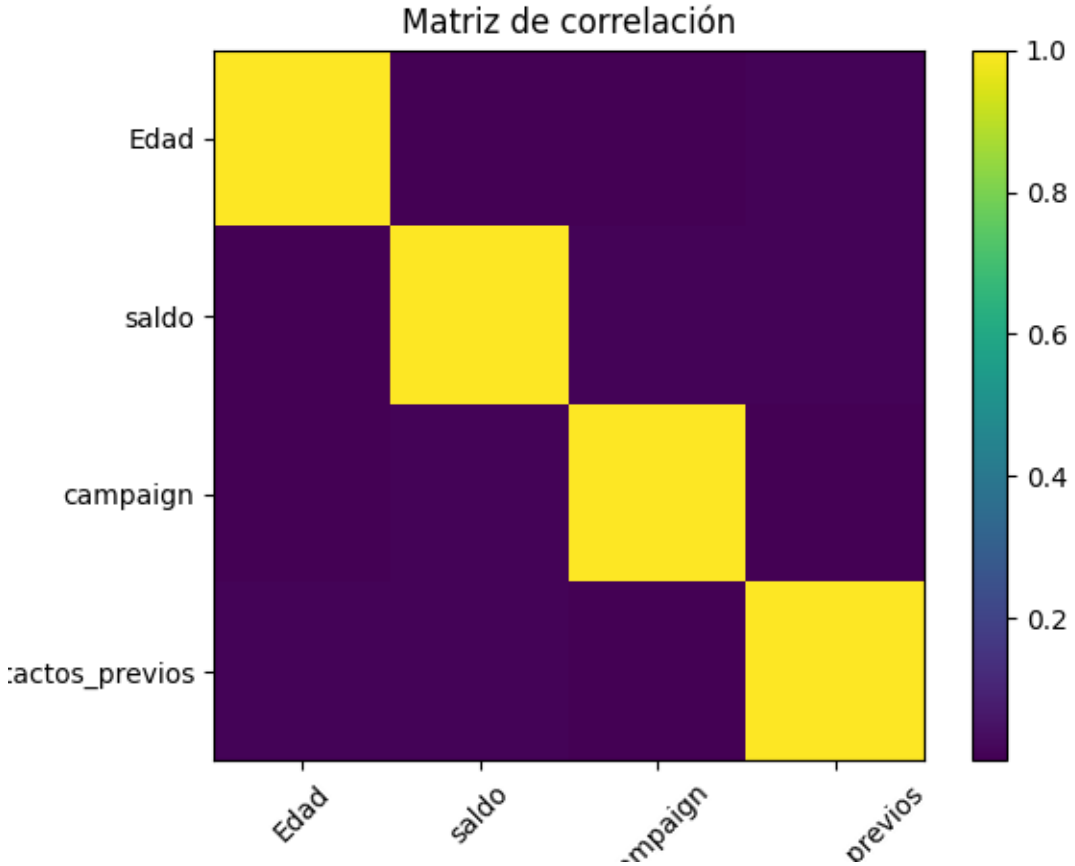

Frecuencias de trabajo
1000
800
600
400
200
0

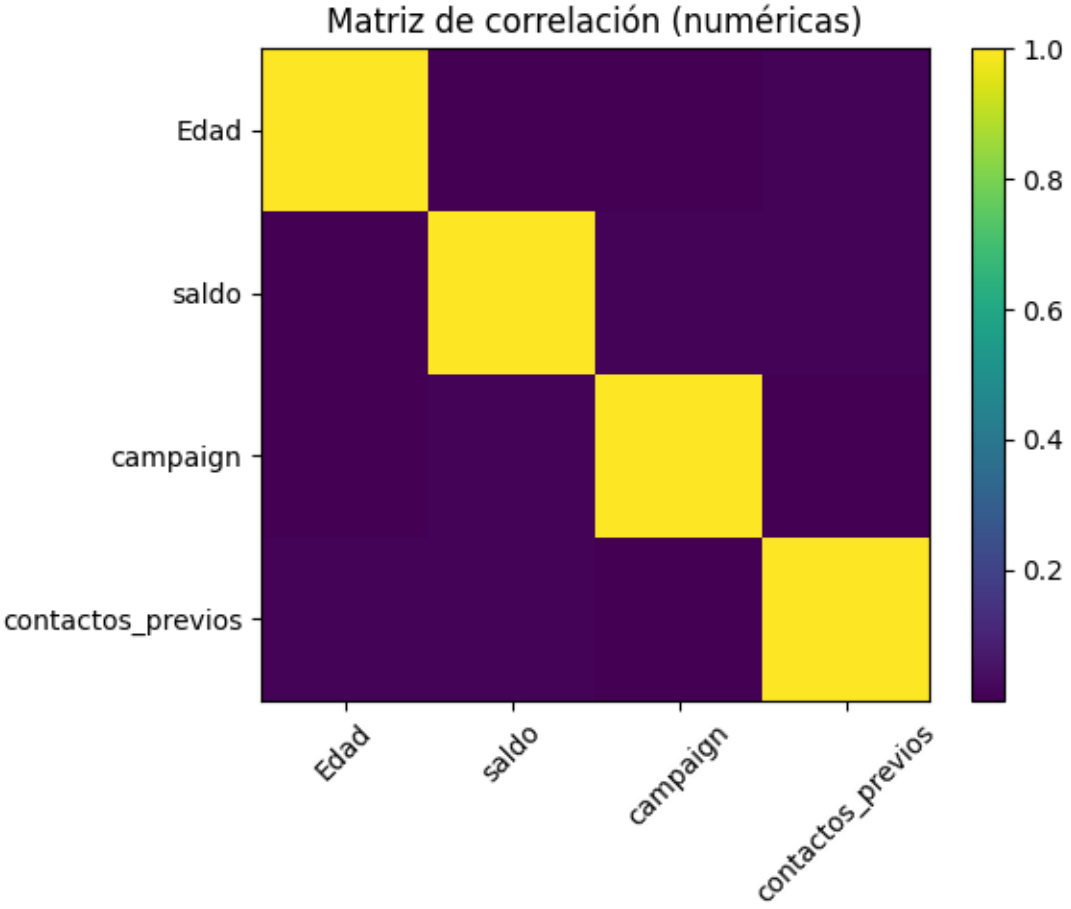

Boxplot de campaign
1
0
10
20
30
40
50

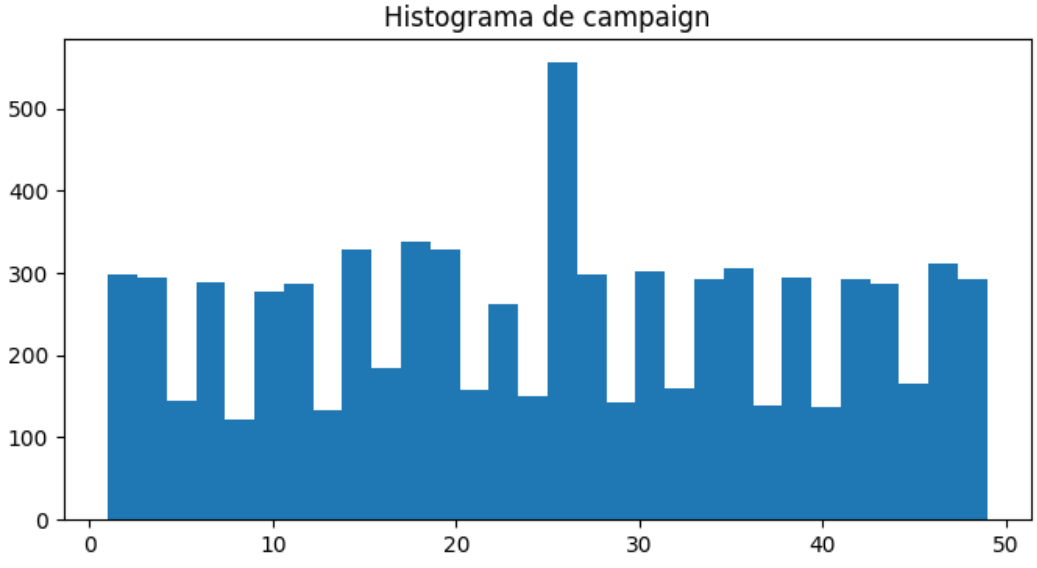

Boxplot de contactos_previos
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

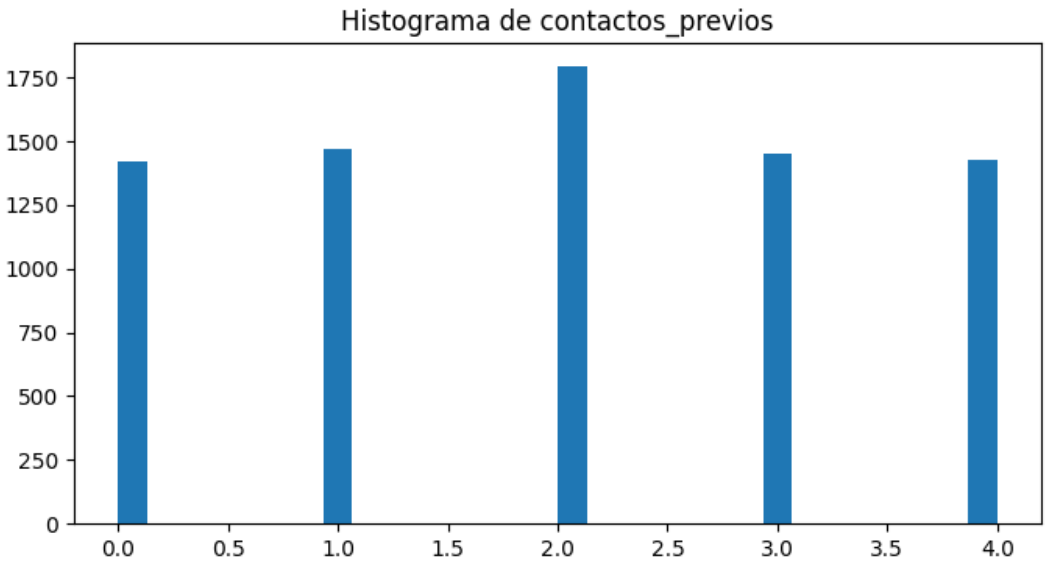

Boxplot de Edad
1
20
30
40
50
60
70

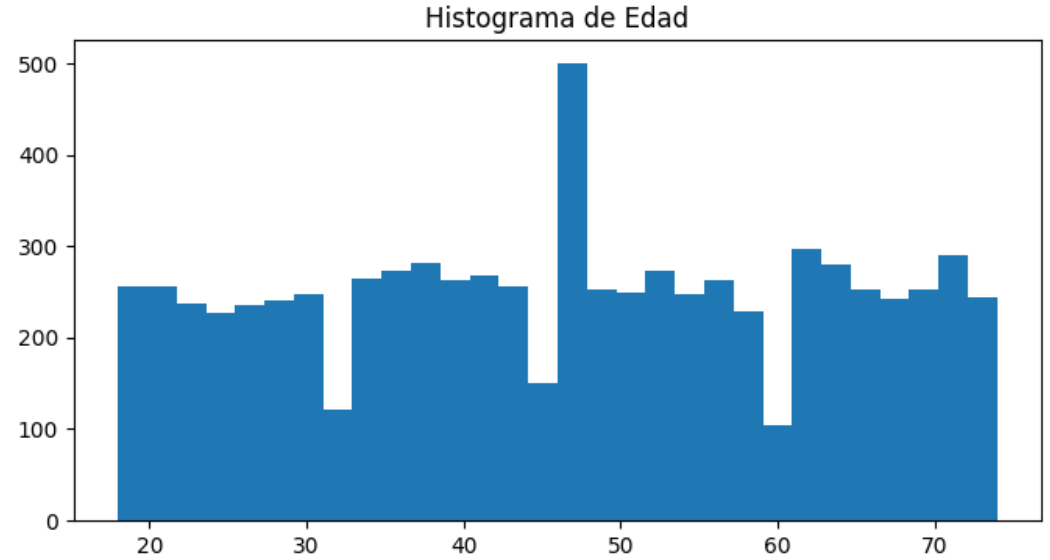

Boxplot de saldo
1
0
20000
40000
60000
80000
100000

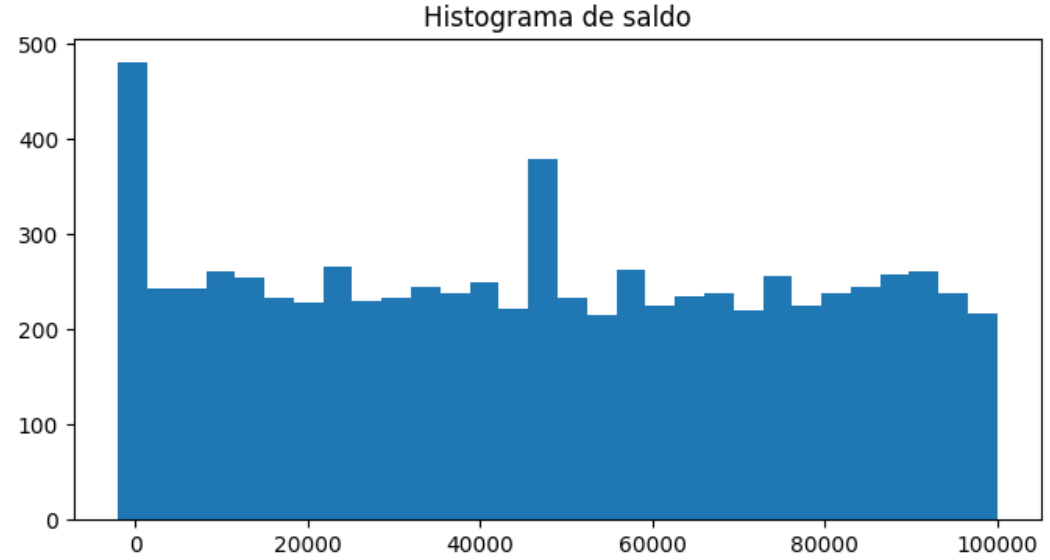

Matriz de correlación
Edad
saldo
campaign
:actos_previos
Edad
saldo
previos
1.0
0.8
0.6
0.4
0.2

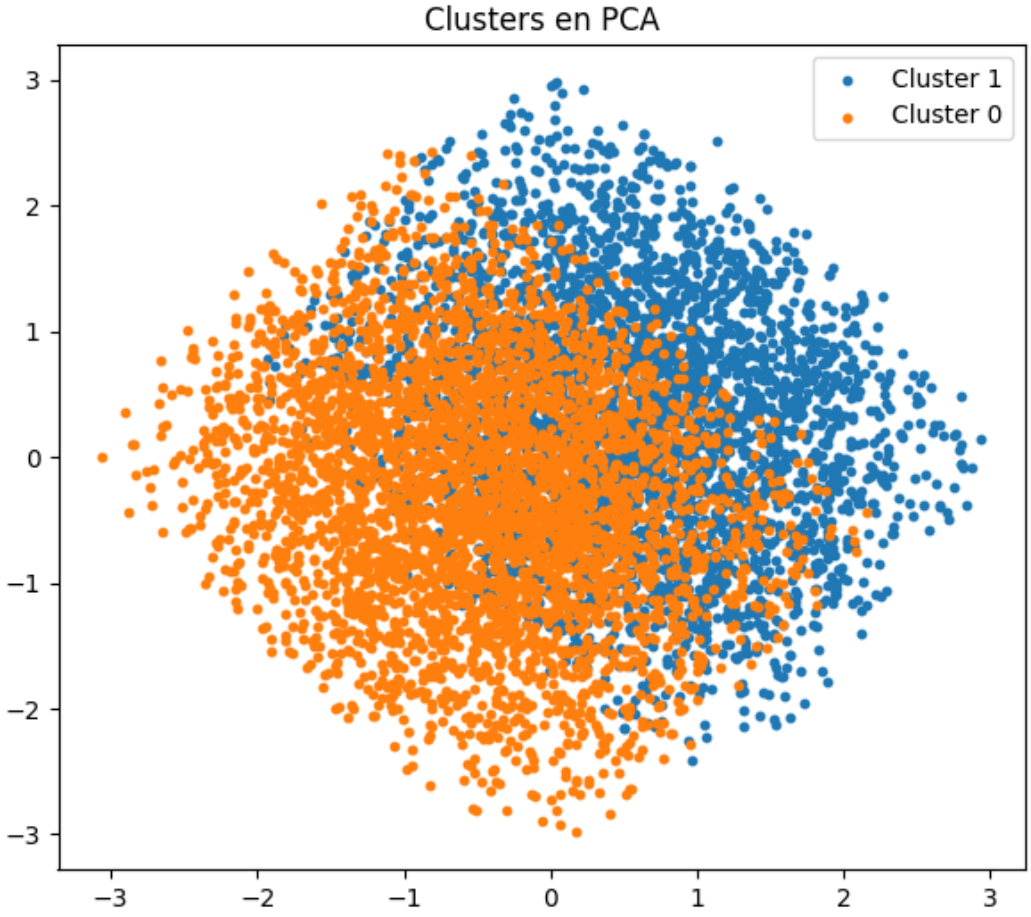

Matriz de correlación (numéricas)
Edad
saldo
campaign
contactos_previos
1.0
0.8
0.6
0.4
0.2
Edad
saldo
campaign
contactos_previos


Histograma de campaign
500
400
300
200
100
0
0
10
20
30
40
50


Histograma de contactos_previos
1750
1500
1250
1000
750
500
250
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Histograma de Edad
500
400
300
200
100
0
20
30
40
50
60
70


Histograma de saldo
500
400
300
200
100
0
0
20000
40000
60000
80000
100000


Clusters en PCA
Cluster 1
Cluster 0
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

5. **SELECCIÓN DEL NÚMERO ÓPTIMO DE CLUSTERS**

Se evaluó K desde 2 a 6 utilizando el silhouette score.

El mejor K fue el que obtuvo el mayor silhouette, garantizando clusters con buena cohesión interna y separación entre sí.

6. **MODELADO CON K-MEANS**

Una vez elegido el mejor K:

• Se entrenó el modelo final con todas las variables procesadas.

• Se asignó un cluster a cada cliente.

• Se generaron perfiles por cluster sobre edad, saldo, campañas previas y tasas de suscripción.

7. **VISUALIZACIÓN CON PCA**

Se aplicó PCA para reducir la dimensionalidad y visualizar los clusters en un plano de dos componentes principales.

Esto permite ver la estructura de los grupos y validar visualmente la separación entre segmentos.

8. **IDENTIFICACIÓN DE CLIENTES POTENCIALES**

Procedimiento:

• Se calculó la tasa global de suscripción.

Se encontraron los clusters con tasa superior al promedio.

- Se seleccionaron los clientes NO suscritos que pertenecen a esos clusters.
- Se midió su distancia al centroide para elegir los Top 50 mejores prospectos.

9. **RESULTADOS PRINCIPALES**

- Segmentos bien definidos de clientes.
- Clusters con alta probabilidad de conversión.
- Lista priorizada de clientes potenciales para futuras campañas.
- Visualizaciones que facilitan la presentación del análisis.

10. **IMPACTO DE NEGOCIO**

El modelo permite:

- Reducir costos dirigiendo campañas a clientes con mayor probabilidad de convertir.
- Mejorar la eficiencia y segmentación de estrategias de marketing.
- Basar decisiones comerciales en análisis estadístico real.

11. **USOS Y BENEFICIOS DEL DASHBOARD**

El dashboard sirve como herramienta profesional para:

- Explorar el EDA.
- Revisar justificación del número óptimo de clusters.
- Visualizar el PCA y entender separación de grupos.
- Consultar perfiles de clusters y diseñar estrategias.
- Ver los clientes potenciales ordenados por relevancia.

Es una herramienta clave para equipos ejecutivos y de marketing.

12. **CONCLUSIONES**

El proyecto logró segmentar de manera efectiva a los clientes del banco, permitiendo identificar grupos con características, comportamientos y probabilidades de suscripción diferentes. Esto demuestra que los datos contienen patrones reales que pueden aprovecharse para estrategias comerciales. El proceso de limpieza y preprocesamiento fue fundamental para garantizar resultados confiables.

Las transformaciones aplicadas (conversión numérica, estandarización y codificación One-Hot) aseguraron que todas las variables aportaran información equilibrada al modelo K-Means.

1. El análisis exploratorio (EDA) nos permitió comprender la base, revelando distribuciones, valores extremos y tendencias relevantes, lo que facilitó la interpretación posterior de los clusters y la toma de decisiones basada en evidencia.

2. La selección del número óptimo de clusters mediante silhouette score proporcionó una justificación estadística sólida del valor de K utilizado. Esto aumentó la coherencia y calidad del modelo final.

3. El algoritmo K-Means formó clusters interpretables y útiles, con diferencias claras en edad, saldo, campañas previas y, especialmente, en tasas de suscripción. Esto permitió identificar segmentos con mayor potencial de conversión.

4. La visualización con PCA confirmó la separación entre clusters, demostrando que el modelo no solo funcionó matemáticamente, sino también de manera intuitiva y visual, facilitando la comunicación de resultados.

5. Se identificaron clientes potenciales basados en evidencia, seleccionando aquellos no suscritos pero con características similares a los que sí suelen aceptar ofertas. La priorización de los Top 50 clientes ofrece un recurso directo y aplicable para campañas futuras.

6. El proyecto aporta un valor real al negocio, ya que permite reducir costos, mejorar la focalización de estrategias y aumentar la efectividad de las campañas, transformando datos crudos en decisiones concretas.

7. El dashboard creado funciona como herramienta ejecutiva, facilitando la exploración del análisis, la revisión del modelo y la visualización de los clientes objetivo, permitiendo presentaciones profesionales y decisiones rápidas.

8. En conjunto, el proyecto demuestra un flujo completo de ciencia de datos, desde la limpieza hasta la aplicación práctica. Su metodología es escalable y adaptable para cualquier institución financiera que busque fortalecer sus campañas mediante segmentación inteligente.