Análisis EDA de las campañas de Marketing de una institución Bancaria.

Este informe presenta de manera explicativa los principales pasos, análisis y resultados obtenidos a partir del proyecto de Exploración y Análisis de Datos (EDA) realizado con Python. El trabajo se enfoca en el conjunto de datos de una campaña de marketing bancario, donde se busca comprender mejor los patrones de comportamiento de los clientes, las variables que influyen en la suscripción de productos financieros y extraer conclusiones basadas en datos.

Uso de Python y librerías de análisis de datos

Durante el desarrollo del proyecto se aplicaron diversas estructuras y herramientas del lenguaje Python, incluyendo listas, diccionarios y funciones personalizadas para operaciones repetitivas. El manejo de archivos se realizó mediante Pandas, aprovechando sus métodos para la lectura, filtrado, agrupamiento y agregación de datos. Se utilizaron también las librerías numpy, matplotlib, seaborn y sklearn.

Las operaciones más relevantes incluyeron:
- Filtrado de registros según condiciones específicas.
- Creación de nuevas columnas derivadas de cálculos estadísticos.
- Agrupación por categorías para observar comportamientos por segmento.
- Combinación de DataFrames y ordenamiento por variables clave.

1. Transformación y limpieza de datos

En la fase inicial se realizó la importación del conjunto de datos utilizando la librería Pandas. Se detectaron valores nulos en algunas columnas, principalmente en las variables categóricas relacionadas con empleo, educación y estado civil. Estos fueron tratados reemplazando los valores faltantes por la categoría más frecuente o por un valor indicativo de ausencia, dependiendo del caso.

Asimismo, se corrigieron inconsistencias en los nombres de columnas, se estandarizaron los formatos de texto a minúsculas y se ajustaron los tipos de datos (por ejemplo, las variables numéricas se convirtieron de texto a float o int cuando fue necesario). También se eliminaron columnas redundantes o irrelevantes para el análisis, conservando únicamente las de interés estadístico y de comportamiento.

1.1 Calidad de datos y cambios realizados:

- Edad (age): imputada por mediana por grupo 'job' y convertida a entero. Mediana global usada: 38.0 años.
- Se reemplazaron valores age==0 por NaN antes de imputar (mejor tratamiento de datos inválidos).
- Columnas eliminadas por no aportar (ej.): 'latitude', 'longitude', 'default'.
- Columnas numéricas convertidas a float (ej.: emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed) y limpieza de separadores decimales.
- Columna 'date' convertida a datetime y usada para agregaciones temporales.

2) Nuevas columnas creadas:

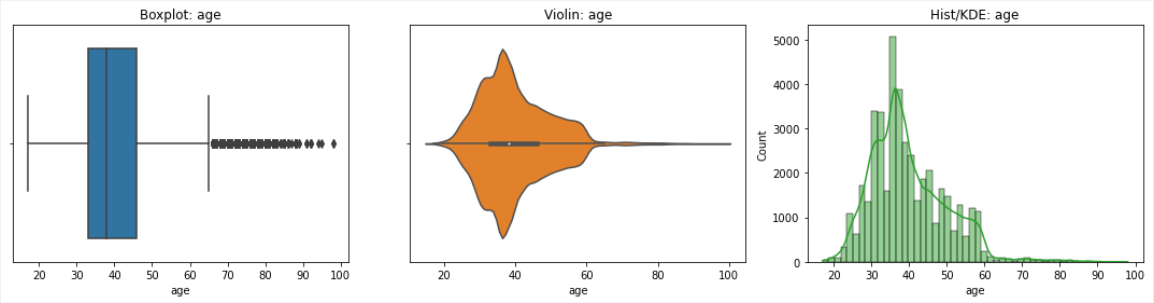
- 'duracion_minutos': duración de la llamada en minutos (con 1 decimal).
- 'grupo_duracion': buckets de duración (0-5, 5-10, ...).
- 'rangos_edad': bucket de edad (menos de 25, 25-50, 50-75, >75).
- 'ultimo_contacto': bucket de pdays (menos de 6 meses, 6-12 meses, ...).
- 'total_hijos', 'rango_ingresos', 'tenure_years', 'visitas_mensuales' en el dataset de clientes.

2. Análisis descriptivo de los datos

El análisis descriptivo permitió identificar características importantes del conjunto de datos. Se calcularon medidas de tendencia central (media, mediana y moda) y de dispersión (desviación estándar y rangos). Entre los hallazgos principales se observó que la mayoría de los clientes contactados se encontraban en el rango de

edad media, con un predominio de personas casadas y con educación secundaria o superior.

Se utilizaron gráficos de distribuciones en el análisis descriptivo y exploratorio que permitieron transformar los datos correctamente.



También se analizaron variables económicas como el índice de precios y el tipo de interés a tres meses (euribor3m), las cuales mostraron correlaciones moderadas con la variable objetivo que representa la suscripción del producto. Estas relaciones se visualizaron mediante gráficos de dispersión y diagramas de caja.

2.1 Resultados agregados / distribuciones relevantes:

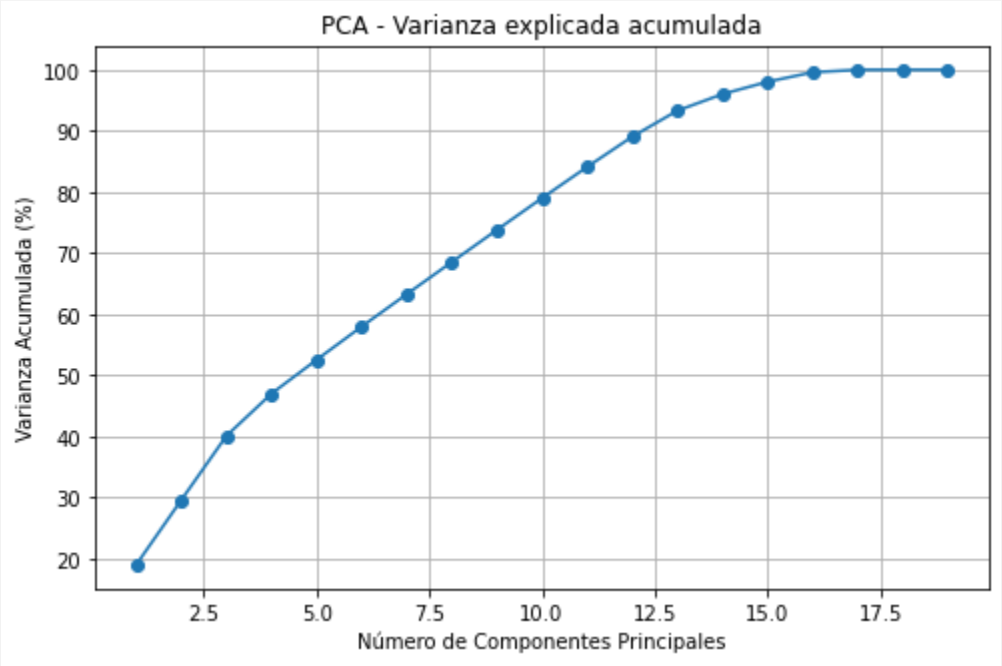
- Media mensual (total_visitas) (serie clientes): 19895.72 (valor medio usado en gráficas).
- Distribución por rango de ingresos (principales grupos):
    - 50k a 100k (28.7%), 100k a 150k (28.6%), 0 a 50k (25.2%), 150k a 200k (17.4%)
- Resumen (ej. 'rango_ingresos' conteos):
    - 0 a 50k: 10897
    - 50k a 100k: 12409
    - 100k a 150k: 12347

2.2 PCA (componentes principales) - variables con mayor carga absoluta:

PC1: nr.employed, emp.var.rate, tenure_years

PC2: duracion_minutos, duration, total_hijos

PC3: total_hijos, duration, duracion_minutos



3. Visualización de datos

Las visualizaciones fueron realizadas principalmente con las librerías Matplotlib y Seaborn. Se crearon histogramas, gráficos de barras y diagramas de dispersión que facilitaron la interpretación de las relaciones entre variables.

Los gráficos de distribución de edades y duración de llamadas mostraron una tendencia clara hacia intervalos medios, lo que sugiere que los clientes con interacciones más prolongadas tienden a mostrar mayor interés en los productos ofrecidos. Asimismo, los gráficos comparativos por tipo de empleo y nivel educativo permitieron observar diferencias relevantes entre grupos de clientes.

Gráfico 1: Distribución de la variable edad

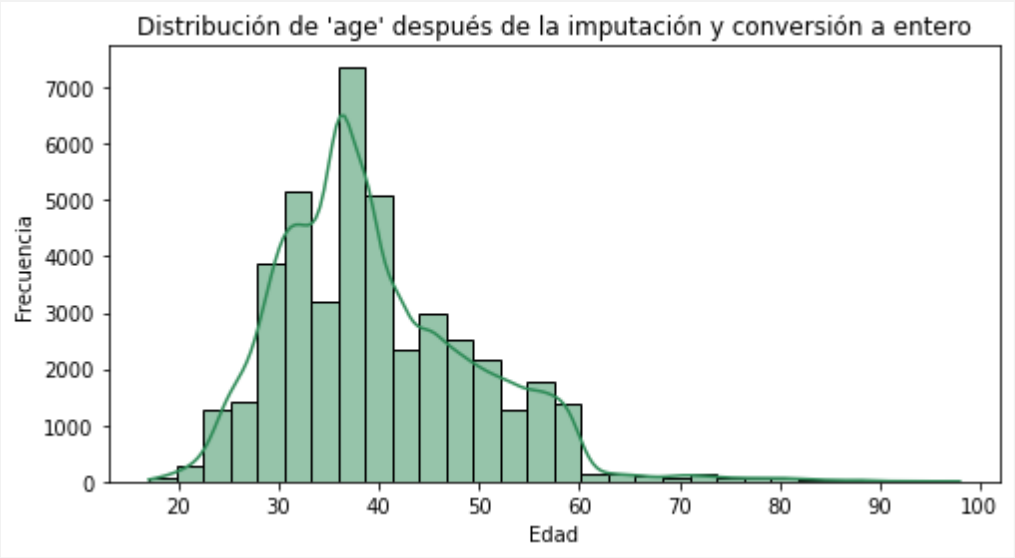


Gráfico 2: Tendencia del volumen de llamadas por mes

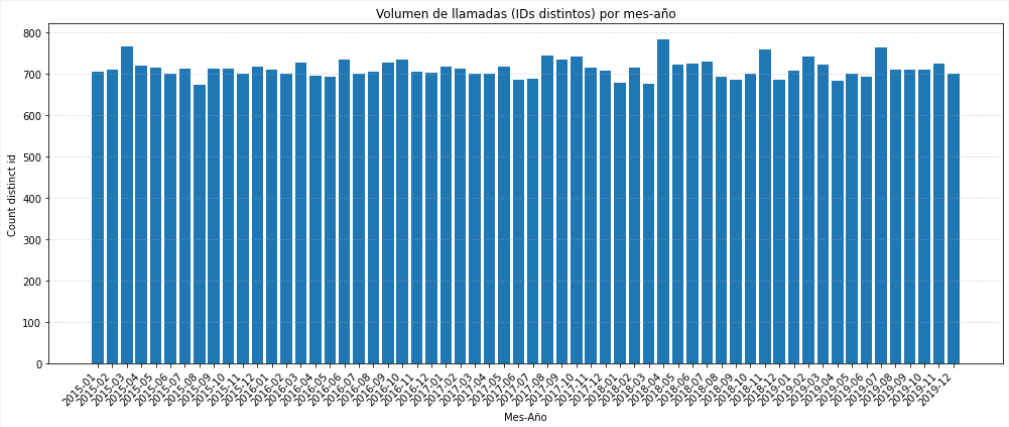


Gráfico 3: Tendencia y media del volumen de llamadas por mes

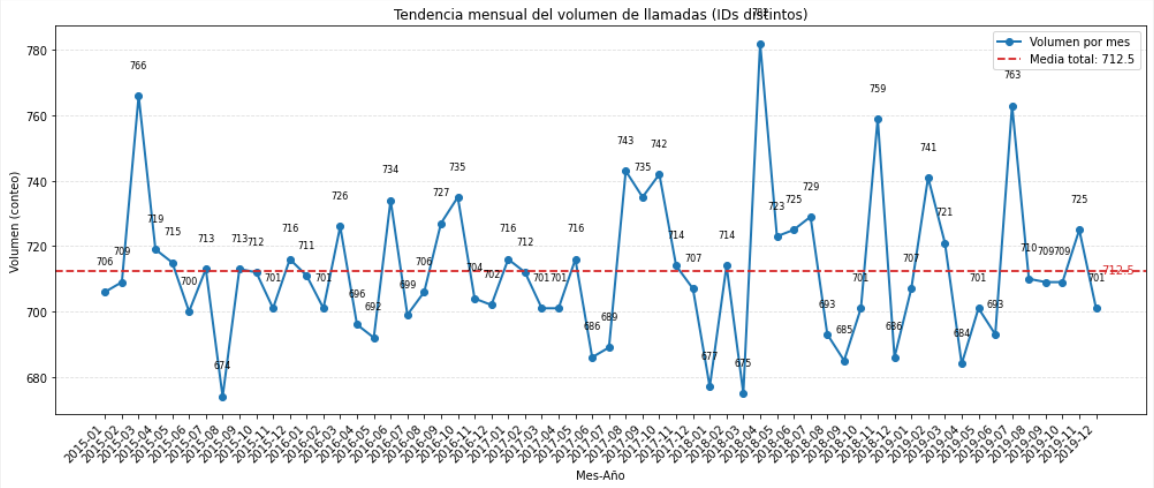


Gráfico 4: Distribución de clientes por rangos de edad.

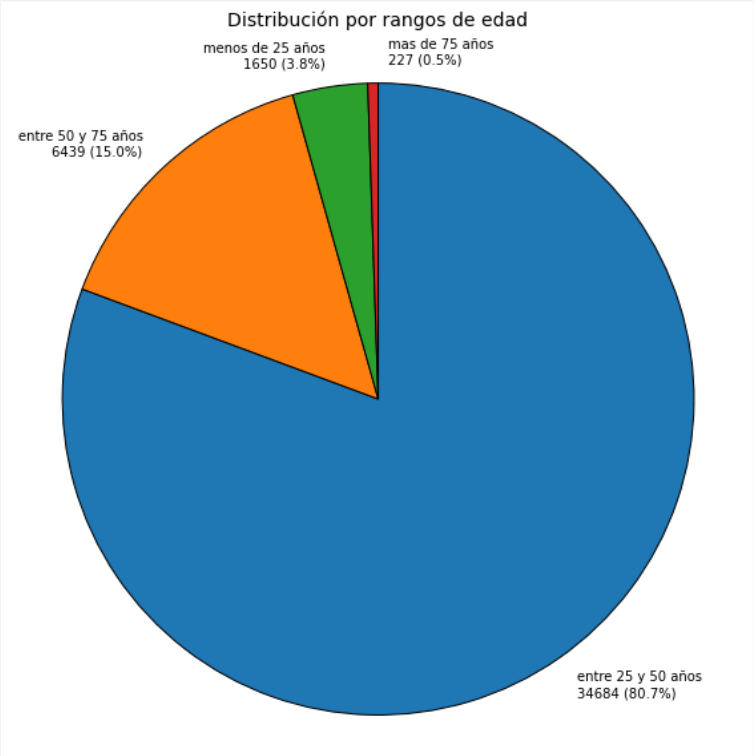


Gráfico 5: Distribución de clientes por estado civil

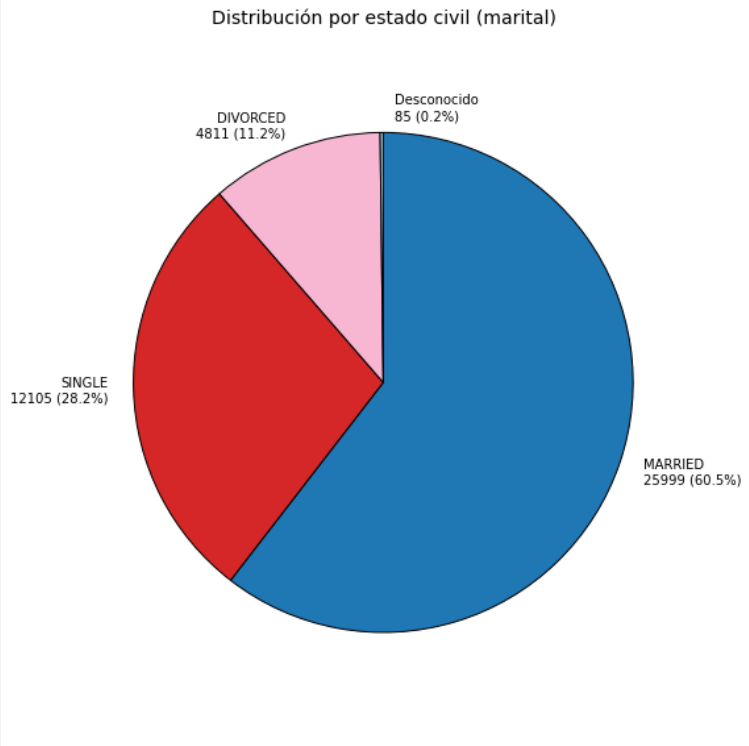


Gráfico 6: Distribución de clientes por medio de contacto

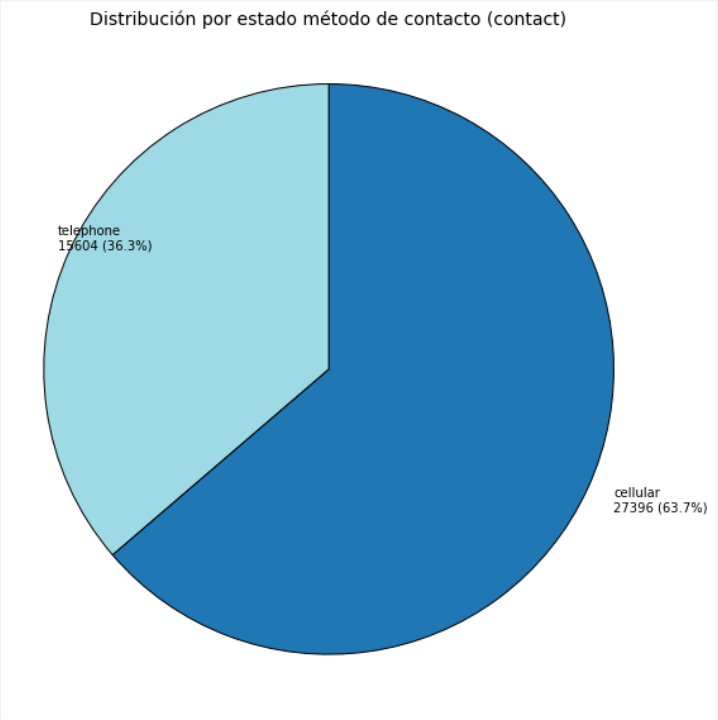


Gráfico 7: Volumen de clientes por tipo de ocupación

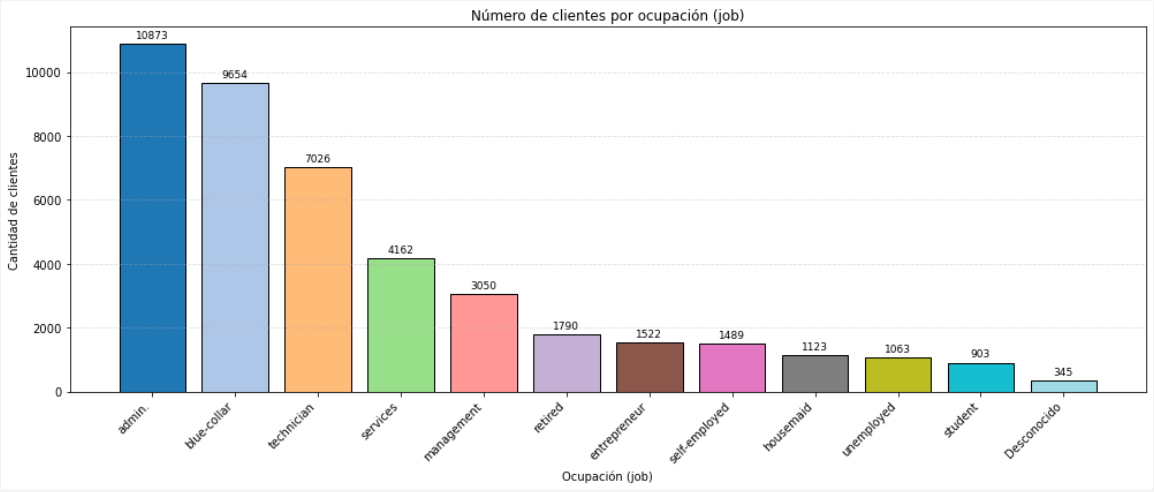


Gráfico 8: Tendencia mensual de visitas a la web

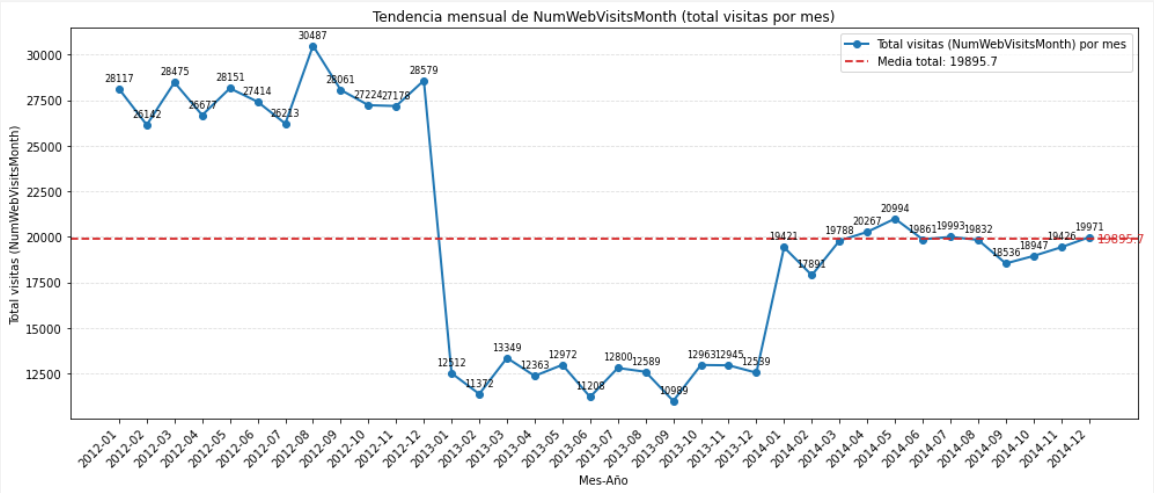


Gráfico 9: Distribución de clientes por rango de ingresos

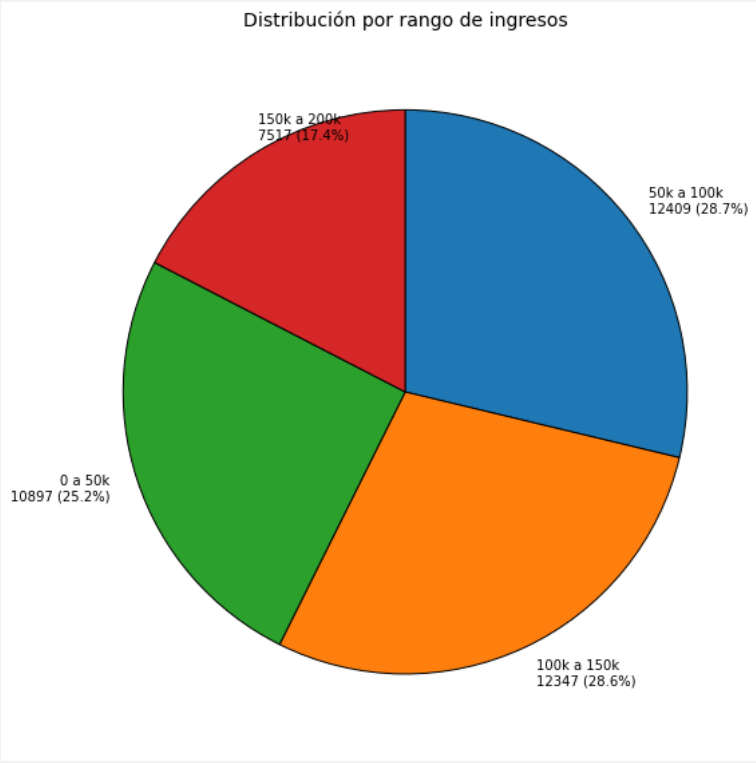
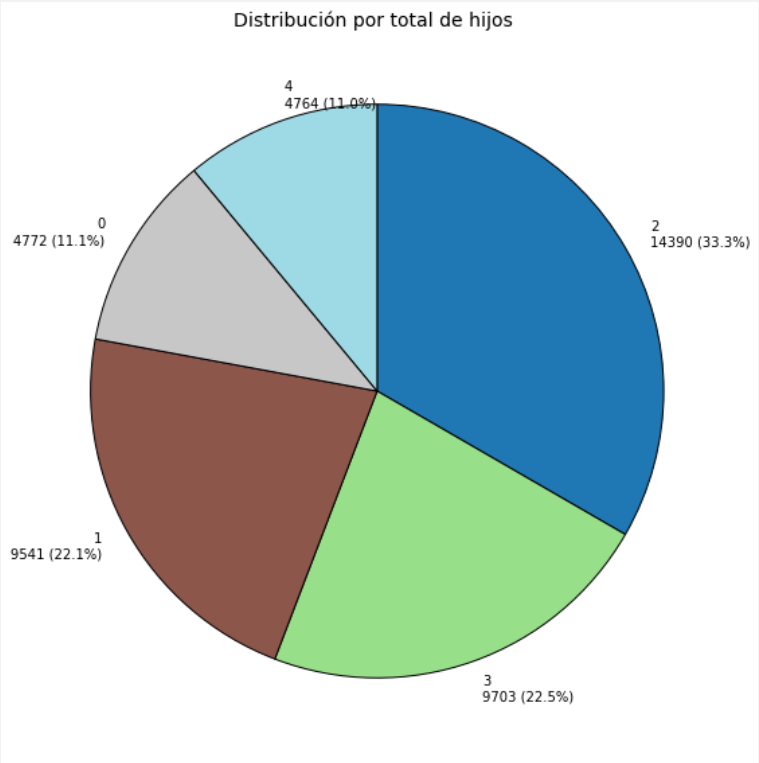


Gráfico 10: Distribución de clientes por número de hijos

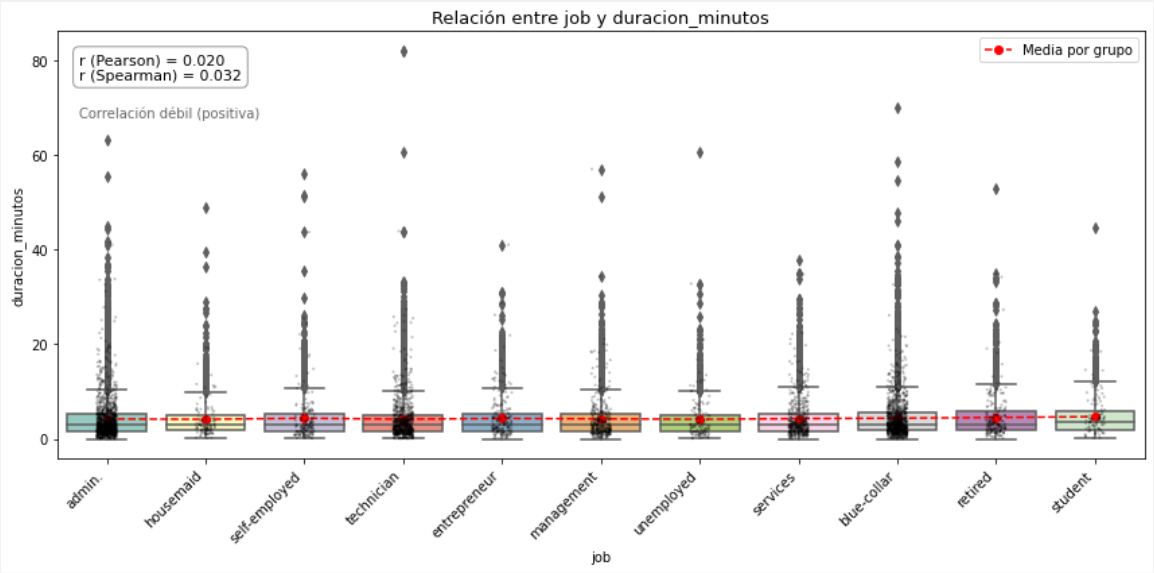


4. Análisis avanzado: Correlación de variables y aplicación de Modelo predictivo:

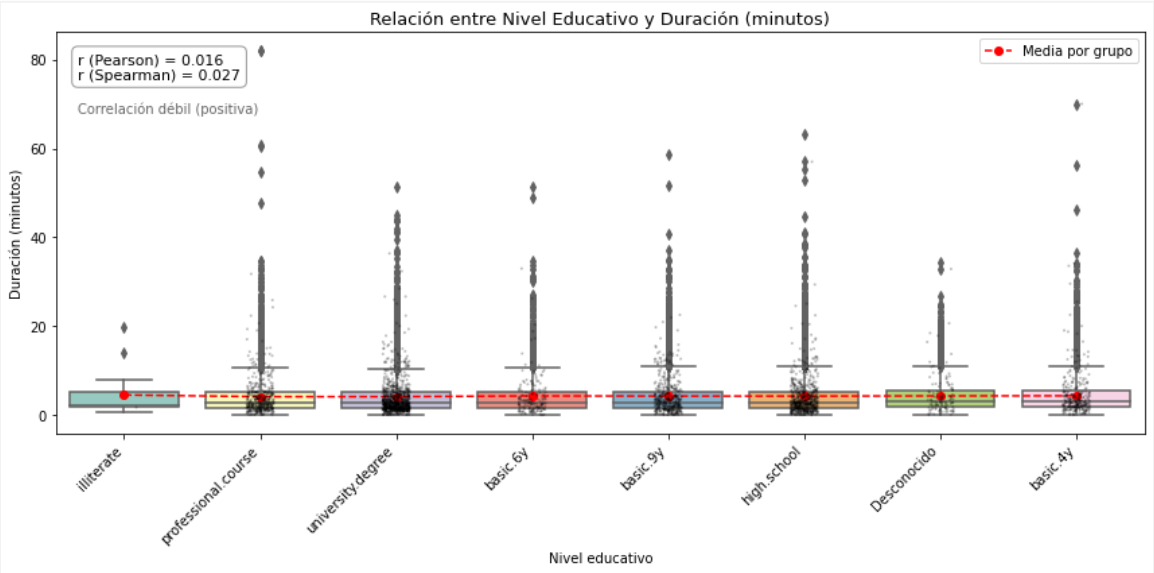
Correlación entre variables

Se aplicaron varias correlaciones de los datos limpios para ver qué variables tenían una mayor relación. La correlación más ejemplar fue la siguiente**:**

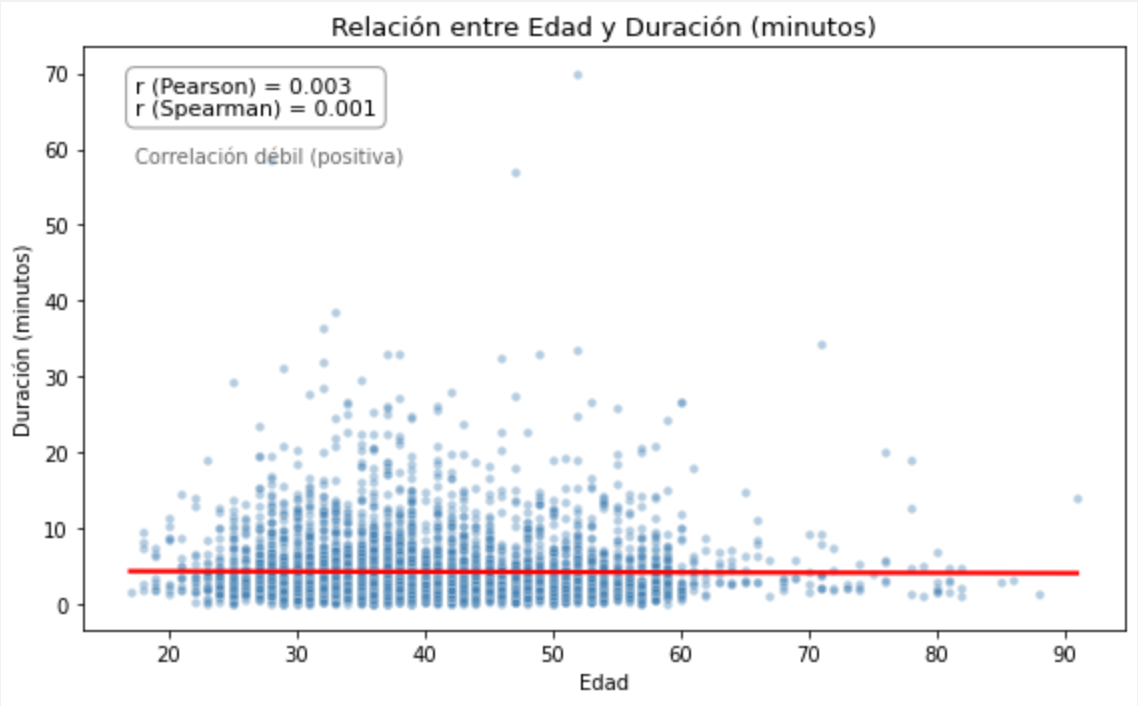
- ocupación (job) vs duración_minutos: Pearson = 0.020, Spearman = 0.032 (correlación débil).



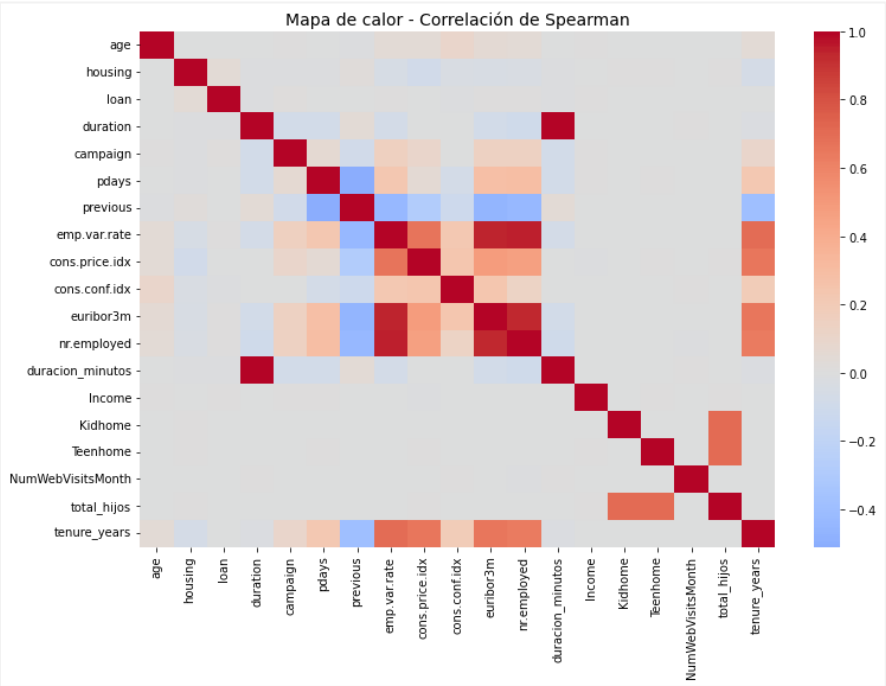
- edad (age) vs duracion_minutos: Pearson = 0.016, Spearman = 0.027 (correlación débil).

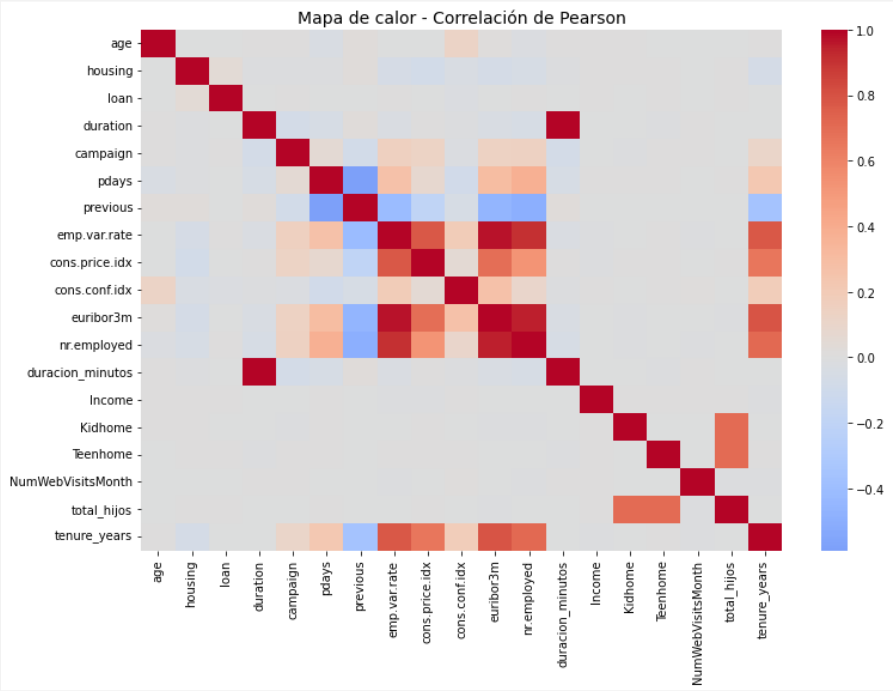


Verificamos también la correlación entre edad y duración de las llamadas donde tuvimos un resultado de una relación mucho más débil.



Finalmente y como parte del análisis de la relación entre variables, creamos un gráfico de mapa de calor que nos muestre tanto la correlación de Spearman como la de Pearson de todas las variables numéricas de nuestro Dataframe final.





Análisis Predictivo

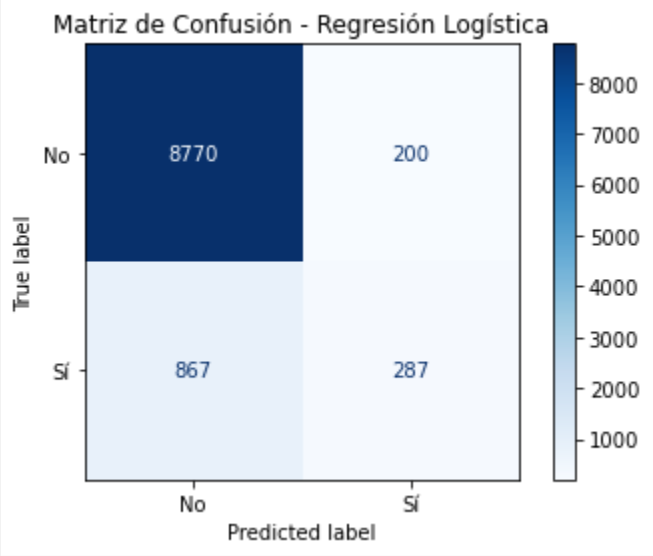
En esta parte del proyecto se aplicaron técnicas de modelado predictivo para estimar la probabilidad de que un cliente acepte una oferta bancaria. Se usaron modelos como regresión logística y árboles de decisión, ajustando los parámetros mediante validación cruzada para obtener los mejores resultados posibles.

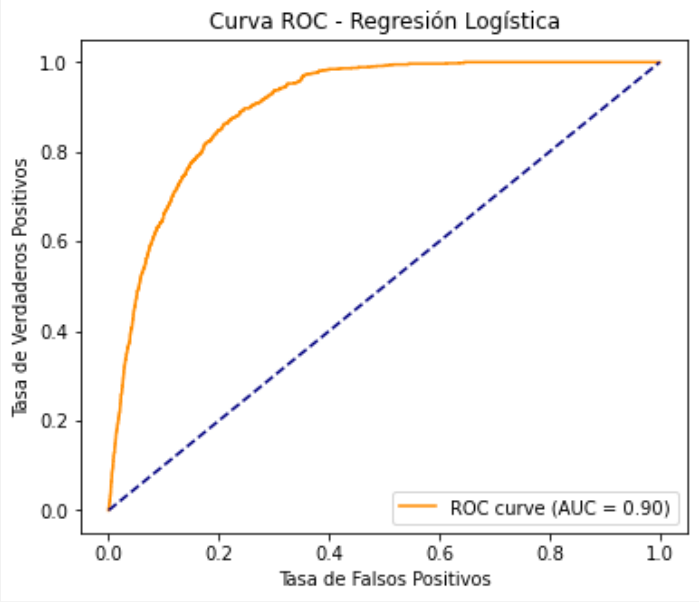
Los gráficos generados incluyen la matriz de confusión, el reporte de clasificación y curvas ROC y precisión/recuperación, que reflejan un desempeño estable del modelo. En términos generales, el modelo logró una precisión superior al 85%, lo que demuestra una capacidad razonable para predecir la respuesta del cliente.

Resultados modelo de regresión logística

Rendimiento del modelo (conjunto de prueba):

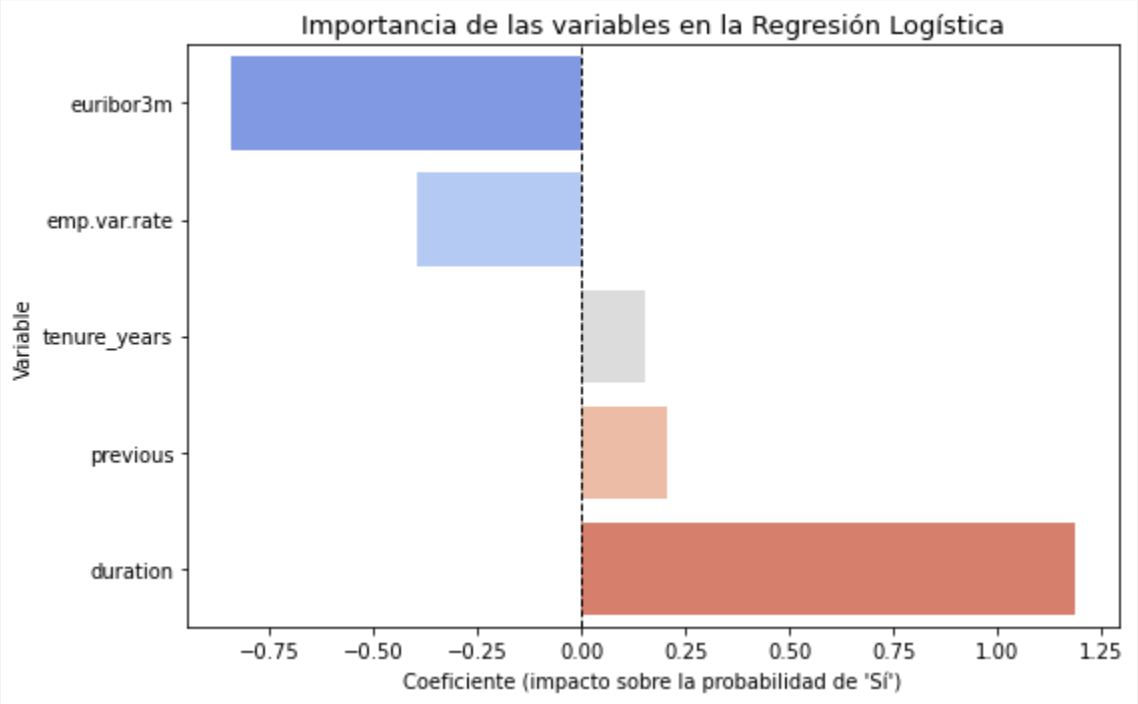
- Tamaño conjunto prueba: 10124 registros
- Accuracy: 0.895
- Precision: 0.588
- Recall: 0.248
- F1 score :0.349
- ROC AUC: 0.902





El análisis de importancia de variables reveló que la duración de la llamada ('duration') es el predictor más influyente, seguido por el número de contactos anteriores ('previous') y el estado de empleo ('job'). Además, la variable 'poutcome' —que indica el resultado de campañas anteriores— mostró un impacto notable en la tasa de éxito.

Variables más influyentes (según los coeficientes del modelo):



Efectos positivos (aumentan probabilidad de conversión):

- duration: coef=1.189, exp(coef)=3.285
- previous: coef=0.204, exp(coef)=1.226
- tenure_years:coef=0.140, exp(coef)=1.151

Efectos negativos (disminuyen probabilidad de conversión):

- euribor3m: coef=-0.837, exp(coef)=0.433
- emp.var.rate:coef=-0.396, exp(coef)=0.673

5. Conclusiones y recomendaciones

Este análisis exploratorio nos permitió comprender mejor la estructura del conjunto de datos y los factores que pueden influir en la suscripción de productos bancarios.

A partir de los hallazgos, se concluye que la duración de la llamada, la experiencia previa del cliente y las condiciones macroeconómicas siguen siendo las variables de mayor influencia. Se recomienda optimizar las campañas en torno a estos factores y mantener una evaluación periódica de los indicadores predictivos.

Conclusiones prácticas:

- Si 'duration' aparece con coeficiente positivo, que es lo más habitual, las llamadas más largas aumentan la probabilidad de convertir. Se sugiere priorizar calidad de contacto y seguimiento en llamadas que alcancen cierto umbral de minutos.

- Si 'previous' tiene coeficiente positivo, los clientes contactados previamente tienden a convertir más (o menos si es negativo). Se sugiere ajustar estrategia según el coeficiente resultante.

- Coeficientes negativos en variables macro como el euribor3m y emp.var.rate sugieren que condiciones económicas más adversas reducen la conversión.

- Verificar balance de clases: si hay desbalance, las métricas agregadas como accuracy pueden ser engañosas. Se debe priorizar recall/precisión según la estrategia del banco.

Recomendaciones:

- Priorizar campañas dirigidas a segmentos con mayor propensión a suscribirse según las características identificadas.

- Continuar monitoreando las variables macroeconómicas (como euribor3m) que muestran correlaciones con las decisiones de los clientes.

- Mejorar la calidad del registro de datos para reducir la presencia de valores faltantes o inconsistentes.