UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

Bank Marketing Campaign Analysis

Presenta:

319304899 Ascencio Díaz Enrique
319241398 Ronson García Eduardo

Profesor:

Juan Carlos González Granados

Asignatura:

Análisis Multivariado

Semestre:

2025-II

ÍNDICE

1. Introducción.

El **análisis multivariado** es una técnica estadística que permite examinar conjuntos de datos que contienen más de una variable. Su finalidad principal es identificar patrones, relaciones y asociaciones entre dichas variables. Además, busca optimizar los datos o simplificar su estructura, facilitando así su interpretación y análisis.

Este tipo de análisis resulta especialmente útil en estudios donde múltiples factores pueden influir simultáneamente en los resultados, permitiendo una visión más integral y detallada del fenómeno en estudio.

1.1. Planteamiento del problema.

En el sector bancario, uno de los grandes desafíos consiste en comprender el comportamiento de los clientes para diseñar estrategias de marketing más efectivas. Las campañas de marketing directo, como las realizadas por vía telefónica, representan una inversión significativa para las instituciones financieras, pero no siempre resultan en una tasa de conversión satisfactoria.

En particular, el presente estudio se enfoca en una base de datos proveniente de una campaña de marketing de un banco portugués, donde se registraron múltiples variables relacionadas con los clientes (como edad, nivel de educación, estado civil, situación laboral, duración de la llamada, entre otras) y la respuesta final del cliente: si aceptó o no suscribirse a un depósito a plazo fijo.

A pesar de contar con un gran volumen de información, identificar qué factores influyen realmente en la decisión del cliente representa un reto. Es aquí donde el análisis multivariado cobra relevancia, ya que permite explorar interacciones complejas entre varias variables al mismo tiempo.

1.2. Objetivo del problema.

Analizar de forma multivariada los datos obtenidos de una campaña de marketing directo de un banco portugués, con el fin de identificar los factores que más influyen en la decisión de los clientes de suscribirse a un depósito a plazo fijo, y así generar conocimiento útil para mejorar la segmentación y efectividad de futuras campañas de marketing.

1.3. Definimos Auxiliares.

El análisis multivariado es una rama de la estadística que se enfoca en el estudio y análisis simultáneo de múltiples variables interrelacionadas. Su objetivo principal es comprender las relaciones complejas entre estas variables y extraer información útil de conjuntos de datos multidimensionales. Para ello, es necesario repasar algunos conceptos fundamentales para el estudio del análisis multivariado:

Definició: Una **población** es un conjunto de personas, eventos u objetos de los cuales nos interesa estudiar alguna característica.

Definición: Una **muestra** es un subconjunto de dicha población. Usualmente, denotamos el tamaño de nuestra muestra con la letra n.

Definición: Una **variable** es una característica o atributo referente a la población. Podemos distinguir dos tipos principales de variables: **cualitativas** y **cuantitativas**.

Definición: Una **variable cualitativa** es aquella que se usa para representar cualidades o características no cuantificables de la población. A su vez, existen dos tipos:

- **Ordinales:** Representan categorizaciones o calificaciones que siguen un orden o jerarquía inherente. Ejemplo: bajo, medio, alto.
- **Nominales:** Representan nombres, etiquetas o clasificaciones sin un orden específico. Ejemplo: colores, sexo, estado civil.

Definición: Una **variable cuantitativa** es aquella que se usa para representar características cuantificables de la población. También se clasifica en dos tipos:

- **Discretas:** Aquellas que sólo toman valores dentro de un conjunto finito o numerable de valores (ejemplo: número de hijos, número de autos).
- **Continuas:** Aquellas que pueden tomar infinitos valores dentro de un intervalo de $\mathbb{R}$ (ejemplo: peso, altura, temperatura).

Definición: Una **variable aleatoria** X es una función $X : \Omega \to \mathbb{R}$ que asigna a cada resultado en el espacio muestral Ω un número real.

En otras palabras, X mapea los resultados de un experimento aleatorio a valores numéricos reales.

Definición: La **esperanza**, también conocida como media y denotada por la letra μ, para una variable aleatoria discreta X con conjunto de valores $\{x_1, x_2, \dots\}$ y probabilidades correspondientes $\{P(X = x_1), P(X = x_2), \dots\}$, se define como:

$$E[X] = \sum_i x_i \cdot P(X = x_i)$$

Para una variable aleatoria continua X con función de densidad de probabilidad $f(x)$, la esperanza se define como:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Definición: La **varianza** para una variable aleatoria, usualmente denotada por σ_X^2, se define como:

$$V[X] = E[(X - E[X])^2]$$

Una forma alternativa de calcularla está dada por:

$$V[X] = E[X^2] - (E[X])^2$$

Definición: La **desviación estándar** de una variable aleatoria X se define como:

$$\sigma_X = \sqrt{V[X]}$$

Definición: La **covarianza** entre dos variables aleatorias X, Y se define como:

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$$

Definición: El **Coeficiente de correlación de Pearson** ρ entre dos variables aleatorias X e Y se define como:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Definición: La **Matriz de datos**

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \tag{1}$$

Definición: La **media muestral** de la i-ésima variable se define como:

$$\bar{x}_i = \frac{1}{n} \sum_{r=1}^{n} x_{ri}$$

Definición: La **covarianza muestral** de la j-ésima variable se define como:

$$s_{ij} = \frac{1}{n} \sum_{r=1}^{n} (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) \tag{2}$$

Definición: La matriz $\mathbf{S}$ de dimensiones $p \times p$ con elementos dados por (2) es nombrada la **matriz de covarianza muestral**.

Definición: Se define la **varianza total** de una matriz de datos $\mathbf{X}$ como la suma de las varianzas de las variables individuales σ_{ii} donde i denota una variable de la matriz de datos. También se puede describir como la traza de la matriz de varianzas-covarianzas de $\mathbf{X}$ que se denota como Σ:

$$\text{Varianza Total} = \sum_{i=1}^{p} \sigma_{ii} = \text{tr}(\Sigma)$$

Definición: La **varianza promedio** se define en términos de la varianza total como el cociente de esta sobre el número de variables:

$$\text{Varianza Promedio} = \frac{\text{Varianza Total}}{p}$$

donde p es el número de variables.

Definición: La **varianza generalizada** se define como el determinante de la matriz de varianzas-covarianzas Σ de una matriz de datos $\mathbf{X}$:

$$\text{Varianza Generalizada} = \det(\Sigma)$$

1.4. Variables redundantes.

En esta sección estudiamos las detección de variables redundantes que puedan existir en nuestro conjunto de datos. Haciendo referencia a una variable que proporciona información que ya está contenida en otras variables del conjunto de datos. En otras palabras, es una variable que no agrega información nueva o útil al análisis y que podría se eliminada sin afectar significativamente los resultados. Algunos métodos para detectar dichas variables redundantes son:

- Matriz de correlaciones
- Análisis de componentes principales.
- Clustering.
- Análisis de Varianza Infiltrada por Tolerancia (VIF).

En tal caso del método de la matriz de correlaciones, discriminaremos como variables redundares a aquellas que estén indicadas con altas correlaciones.

1.5. Detección de valores atípicos.

Los valores atípicos son aquellas observaciones cuyos valores son muy diferentes a otras observaciones del mismo grupo de datos. Estos pueden ser ocasionados por: errores de procedimiento, acontecimientos extraordinarios, valores extremos o por causas desconocidas.

Estos datos atípicos representan un problema, ya que tienen un efecto de distorsión en el resultado del análisis de los datos. Por esta razón, es importante identificarlos para tratarlos de manera adecuada. Entre los métodos de detección de estos valores atípicos, encontramos los siguientes:

- **Método del rango intercuartílico (RIC)**
  El rango intercuartílico es una medida de variabilidad basada en la división de un conjunto de datos en cuartiles. Comprende el rango entre el primer y el tercer cuartil (los bordes de la caja en un diagrama de caja y bigotes) y se denota por:

  $$RIC = Q_3 - Q_1$$

  Bajo este método, se considerarán como valores atípicos aquellos datos que queden por debajo de $Q_1 - 1{,}5 \times RIC$ o por encima de $Q_3 + 1{,}5 \times RIC$.

- **Método Z-score**
  Los puntajes z o puntajes estándar miden la distancia entre un punto de datos y la media del conjunto en términos de desviación estándar.

  Para usar este puntaje en la detección de valores atípicos, establecemos un umbral en función del nivel de importancia o de los requisitos específicos del conjunto de datos. En general, una puntuación Z de más de 3 se considera un caso atípico. Es decir, si cumplen la condición:

  $$\frac{X - \mu_X}{\sigma_X} > 3$$

  se consideran valores atípicos.

- **Distacia de Mahalanobis**
  La distancia de Mahalanobis es un criterio que depende de los parámetros estimados de la distribución multivariada. Describe la distancia entre cada punto de datos y el centro de masa.

  La distancia de Mahalanobis se define como:

$$MSD_i = \sqrt{(x_i - \bar{x})^T S_n^{-1} (x_i - \bar{x})}$$

1.6. Componentes principales

El análisis de componentes principales busca algunas combinaciones lineales que puedan ser utilizadas para resumir los datos, perdiendo en el proceso la menor cantidad de información posible. Este intento de reducir la dimensión puede ser descrito como un resumen eficiente de los datos.

Definición: Si x es un vector aleatorio con media μ y covarianza σ, entonces la transformación de componentes principales es:

$$x \to y = \Gamma^T (x - \mu)$$

Donde Γ es ortogonal, $\Gamma^T \Sigma \Gamma = A$ es diagonal y $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$. La estricta positividad de los eigenvalores λ_i está garantizada si Σ es definida positiva. Esta representación de Σ sigue del teorema de descomposición espectral. El i-ésimo componente principal de x puede definirse como el i-ésimo elemento del vector y, es decir:

$$y_i = (\gamma_i)^T (x - \mu)$$

La combinación lineal con la varianza más grande λ_1 y la respuesta será la primera componente principal, definida por analogía sobre la expresión anterior como:

$$y_{(1)} = (X - \mathbf{1}\bar{x}^T) g_{(1)}$$

En este caso $g_{(1)}$ es el eigenvector estandarizado correspondiente al eigenvalor más grande de S (es decir $S = GLG^T$).

De forma similar el i-ésimo componente principal se define como:

$$y_{(i)} = (X - \mathbf{1}\bar{x}^T) g_{(i)}$$

La correlación entre x y el vector de componentes principales y.

Por simplicidad se asume que la media de x y por tanto de y es 0. La covarianza de x y y es:

$$E(xy) = E(xx^T \Gamma) = \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda$$

por lo tanto la covarianza entre x_i y y_i es $\gamma_{ij}\lambda_j$. Ahora x_i y y_j tienen varianzas σ_{ii} y λ_i respectivamente, así su correlación ρ_{ij} es entonces:

$$\rho_{ij} = \gamma_{ij}\lambda_j / (\sigma_{ii}\lambda_j)^{1/2} = \gamma_{ij}(\lambda_j/\sigma_{ii})^{1/2}$$

1.7. Escalamiento principal.

Definición: El escalamiento multidimensional (MDS) es una técnica estadística utilizada para analizar la similitud de un conjunto de datos, que nos va a permitir mapear (proyectar) en un espacio de menor dimensión las distancias originales entre individuos que se encuentran en un espacio de mayores dimensiones.

Esta técnica es una herramienta útil para la reducción de dimensionalidad y la visualización de datos complejos, permitiendo entender y analizar las relaciones en conjuntos de datos de manera intuitiva.

El procedimiento del MDS implica los siguientes pasos:

1. **Construcción de la matriz de distancias:** A partir de los datos originales, se calcula una matriz que contiene las distancias o disimilitudes entre todos los pares de objetos.

2. **Configuración inicial:** Se elige una configuración inicial en un espacio de dimensiones reducidas.

3. **Optimización iterativa:** Se ajustan las posiciones de los puntos en el espacio reducido para minimizar la diferencia entre las distancias en este espacio y las distancias originales. Este proceso se realiza mediante un criterio de estrés, definido como:

$$\text{Estrés} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$$

   donde d_{ij} son las distancias originales y $\hat{d}_{ij}$ son las distancias en el espacio reducido.

4. **Visualización:** Los puntos en el espacio reducido se visualizan, permitiendo una interpretación más sencilla de las relaciones entre los objetos.

1.8. Análisis de factores.

Definición: El Análisis de factores es una técnica estadística utilizada para identificar la estructura subyacente en un conjunto de variables observadas. El objetivo principal del análisis de factores es explicar las relaciones entre las variables en términos de un número menor de variables no observadas, denominadas factores.

Es una técnica esencial para la reducción de dimensionalidad y la identificación de estructuras latentes en datos complejos, facilitando la interpretación y el entendimiento de las relaciones entre variables.

El análisis de factores supone que las variables observadas pueden ser representadas como combinaciones lineales de un conjunto de factores más un término de error. Este modelo puede ser expresado matemáticamente como:

$$x = \Lambda f + \epsilon$$

donde:

- x es el vector de variables observadas.

- Λ es la matriz de cargas factoriales que indica la relación entre las variables observadas y los factores.
- f es el vector de factores.
- ϵ es el vector de términos de error específicos de cada variable.

Las cargas factoriales (Λ) representan las correlaciones entre las variables observadas y los factores. Una carga factorial alta indica que una variable está fuertemente asociada con un factor particular.

1.9. Análisis de conglomerados.

Definición: El análisis de clústeres o conglomerados es una técnica estadística utilizada para agrupar un conjunto de objetos en grupos de tal manera que los objetos en el mismo grupo sean más similares entre sí que con los objetos de otros grupos. Esta técnica es muy útil en la exploración de datos y en la identificación de patrones naturales en los datos sin una estructura predefinida.

1.9.1. Métodos Jerárquicos

Los métodos jerárquicos crean una estructura de árbol que representa cómo los objetos son agrupados en diferentes niveles de similitud. Estos métodos se dividen en:

- **Aglomerativos :** Comienzan con cada objeto en un clúster separado y fusionan los clústeres más similares en cada paso hasta que todos los objetos están en un solo clúster.
- **Divisivos:** Comienzan con todos los objetos en un solo clúster y dividen los clústeres en cada paso hasta que cada objeto está en su propio clúster.

Existen varios métodos para definir la distancia entre clústeres, tales como:

- **Enlace Simple:** La distancia entre dos clústeres es la distancia mínima entre cualquier par de puntos, uno de cada clúster.
- **Enlace Completo:** La distancia entre dos clústeres es la distancia máxima entre cualquier par de puntos, uno de cada clúster.
- **Enlace Promedio:** La distancia entre dos clústeres es el promedio de todas las distancias entre pares de puntos, uno de cada clúster.
- **Enlace Centroidal:** La distancia entre dos clústeres es la distancia entre los centroides de los clústeres.

Una herramienta clave en el clustering jerárquico es el dendrograma, que es una representación visual en forma de árbol que muestra el proceso de agrupación o división de los clústeres. En ellos las hojas representan los objetos y las ramas representan la unión de clústeres. El dendrograma permite visualizar y decidir el número óptimo de clústeres cortando el árbol en un nivel específico.

1.9.2. Métodos No Jerárquicos

Los métodos no jerárquicos, como el **k-medias**, dividen los datos en un número predefinido de clústeres. El objetivo es minimizar la variabilidad dentro de cada clúster.

Los pasos principales del algoritmo k-medias son:

1. **Inicialización:** Seleccionar k centroides iniciales.
2. **Asignación:** Asignar cada objeto al clúster cuyo centroide esté más cercano.
3. **Actualización:** Recalcular los centroides de los clústeres basados en los objetos asignados.
4. **Iteración:** Repetir los pasos de asignación y actualización hasta que los centroides ya no cambien significativamente.

Sin embargo, el método k-Medias posee desventajas significativas como:

- **Número de Clústeres:** Requiere que el número de clústeres k sea especificado de antemano.
- **Sensibilidad a la Inicialización:** La elección de los centroides iniciales puede afectar el resultado final.
- **Formas de Clústeres:** Supone que los clústeres son esféricos y de tamaño similar, lo cual puede no ser adecuado para datos con clústeres de formas arbitrarias.

1.10. Modelos de regresión.

1.10.1. Regresión lineal.

La **regresión lineal** es el modelo de regresión más simple y común. Se asume que la relación entre la variable dependiente y y las variables independientes $X_1, X_2, \ldots, X_p$ es lineal. El modelo se expresa como:

$$y = \beta_0 + \beta_1 X + \epsilon$$

donde:

- β_0 es la intersección o término constante.
- β_1, vector de coeficientes de regresión que representan la magnitud y la dirección de la relación entre cada variable independiente y la variable dependiente.
- ϵ es el término de error que captura la variabilidad en y no explicada por las variables independientes.

1.10.2. Regresión Logística

La **regresión logística** se utiliza cuando la variable dependiente es categórica, especialmente binaria (con dos categorías). El modelo estima la probabilidad de que la variable dependiente tome un valor específico (e.g., éxito o fracaso). La relación se modela utilizando la función logística:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

donde p es la probabilidad de que la variable dependiente sea 1 (por ejemplo, éxito).

1.10.3. Criterios de evaluación

Los modelos de regresión se evalúan utilizando diversos criterios y estadísticas:

- **Coeficiente de Determinación (R^2):** Indica la proporción de la varianza en la variable dependiente explicada por las variables independientes.
- **Error Cuadrático Medio (MSE):** Mide la magnitud promedio de los errores de predicción.
- **Estadístico F:** Prueba la significancia global del modelo.
- **Pruebas t:** Evalúan la significancia individual de cada coeficiente de regresión.

1.11. Árboles de decisión.

Definición: Los árboles de decisión son modelos predictivos utilizados tanto para clasificación como para regresión. Estos modelos utilizan un conjunto de reglas de decisión derivadas de los datos para predecir el valor de una variable objetivo.

Los árboles de decisión son priorizados debido a su capacidad para manejar tanto variables categóricas como continuas.

La construcción de un árbol de decisión implica los siguientes pasos:

1. **Selección de la Mejor División:** En cada nodo, seleccionar la variable y el punto de división que mejor separen los datos según una métrica específica.
2. **División del Nodo:** Crear ramas basadas en la mejor división.
3. **Repetición:** Repetir los pasos anteriores para cada nodo resultante hasta que se cumpla un criterio de detención (profundidad máxima del árbol, número mínimo de muestras en un nodo, etc.).
4. **Asignación de Valores:** Asignar valores a los nodos de hoja basándose en la mayoría de clases (para clasificación) o en la media de los valores (para regresión).

Para seleccionar la mejor división en cada nodo, se mide la impureza de un nodo, la incertidumbre de un nodo o la reducción en la variabilidad de los valores de la variable objetivo.

Para evitar el sobreajuste, se puede aplicar una técnica llamada **poda**. Puede ser previa, donde se detiene la construcción del árbol en base a la profundidad máxima o mínimo de muestras en un nodo, o posterior, donde luego de construir el árbol completo se eliminan nodos que no proporcionan poder predictivo significativo.

1.12. Análisis de discriminante lineal

Definición: El Análisis de Discriminante Lineal (LDA) es una técnica estadística utilizada para encontrar una combinación lineal de características que separe o caracterice dos o más clases de objetos o eventos.

El objetivo principal del LDA es proyectar los datos en un espacio de menor dimensión con la máxima separabilidad de clases.

LDA se basa en varias suposiciones clave:

- **Normalidad:** Las características siguen una distribución normal multivariante.
- **Covarianza Igual:** Las clases tienen la misma matriz de covarianza.
- **Linealidad:** Las relaciones entre las características son lineales.

1.12.1. Matemáticas del LDA

El proceso de LDA implica los siguientes pasos:

1. **Calcular las Medias de las Clases y la Media General**

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} x \quad y \quad \mu = \frac{1}{N} \sum_{i=1}^{k} N_i \mu_i$$

donde N_i es el número de muestras en la clase i, μ_i es la media de la clase i, N es el número total de muestras y μ es la media general.

2. **Calcular la Matriz de Varianza dentro de la clase S_W**

$$S_W = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T$$

3. **Calcular la Matriz de Varianza entre clases S_B**

$$S_B = \sum_{i=1}^{k} N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

4. **Resolver la Ecuación Generalizada de valores propios**

$$W = S_W^{-1} S_B$$

donde W es la matriz de transformación de datos. Los k vectores propios correspondientes a los mayores valores propios forman la base del espacio de características reducido.

5. **Proyección de las Muestras**

   Proyectar las muestras originales (X de dimensión M), en el nuevo espacio ($k < M$) de características utilizando:

$$Y = XV_k$$

   donde V_k es la matriz de vectores propios de W.

2. Marco Teórico

2.1. Introducción

El marketing bancario se ha convertido en una herramienta esencial para que las instituciones financieras se mantengan competitivas en un entorno globalizado, digital y cambiante. Esta disciplina busca comprender, atraer, retener y fidelizar a los clientes mediante estrategias enfocadas en satisfacer sus necesidades financieras de forma eficiente y personalizada. La importancia del marketing bancario ha aumentado conforme los clientes se vuelven más exigentes y esperan soluciones rápidas, accesibles y adaptadas a sus necesidades particulares.

El presente marco teórico tiene como objetivo abordar los antecedentes históricos y conceptuales del marketing bancario, exponer los principales conceptos y teorías aplicadas, y relacionar estos fundamentos con la problemática específica de la investigación. A través de este análisis, se busca construir una base sólida que permita comprender el contexto actual del sector bancario y su relación con las estrategias de marketing.

2.2. Antecedentes

El desarrollo del marketing en el sector bancario ha evolucionado significativamente desde la década de los setenta, cuando las instituciones financieras comenzaron a adoptar una orientación centrada en el cliente. Inicialmente, el enfoque era netamente transaccional, centrado en la venta de productos financieros, sin considerar aspectos como la satisfacción del cliente o la construcción de relaciones a largo plazo (**kotler2012**).

Con el avance de las tecnologías de la información y la creciente competencia en el mercado financiero, los bancos han debido replantear sus estrategias. En América Latina, autores como **cabrera2017** han documentado cómo la liberalización económica, la apertura de mercados y la aparición de nuevos actores como las fintechs han obligado a los bancos tradicionales a transformarse digitalmente. Esta transformación implica no solo cambios tecnológicos, sino también una nueva forma de entender al cliente y de ofrecerle valor.

En el contexto mexicano, **gomez2019** señalan que la banca ha experimentado un proceso de modernización, impulsado por la regulación gubernamental y la demanda de servicios más personalizados. La digitalización de los canales bancarios y la segmentación estratégica del mercado son elementos claves en este proceso, y constituyen ejes centrales para la formulación de estrategias de marketing más efectivas.

2.3. Conceptos clave

2.3.1. Marketing bancario

Es la aplicación de los principios y herramientas del marketing tradicional y digital en el sector financiero, con el objetivo de identificar y satisfacer las necesidades de los clientes bancarios,

generando valor tanto para el cliente como para la institución (**jobber2013**). Incluye actividades como investigación de mercado, desarrollo de productos, estrategias de comunicación y gestión de relaciones.

2.3.2. Segmentación de mercado

Proceso de dividir el mercado en grupos homogéneos de clientes que comparten características similares. En el sector bancario, la segmentación puede basarse en criterios demográficos, psicográficos, conductuales o financieros (**lamb2011**). Esto permite diseñar ofertas personalizadas que responden a las expectativas específicas de cada segmento.

2.3.3. Gestión de relaciones con el cliente (CRM)

Estrategia centrada en la administración eficiente de la relación con los clientes, apoyada generalmente en tecnologías digitales. El CRM permite personalizar ofertas, aumentar la satisfacción y fortalecer la lealtad del cliente mediante una comunicación más efectiva y la recopilación de datos significativos (**peppers2004**).

Transformación de los canales tradicionales hacia plataformas digitales que permiten a los usuarios realizar operaciones y consultas desde dispositivos electrónicos, lo cual ha revolucionado las estrategias de marketing del sector (**gonzalez2020**). La digitalización también ha permitido la automatización de procesos y la mejora de la experiencia del usuario.

3. Bases teóricas

3.0.1. Teoría del marketing relacional

Propuesta por **berry1983**, esta teoría enfatiza la importancia de establecer relaciones duraderas con los clientes, más allá de una simple transacción. En el ámbito bancario, se traduce en programas de fidelización, atención personalizada y servicios postventa que permiten construir una relación basada en la confianza mutua.

3.0.2. Modelo de las 4 P's de McCarthy

Producto, precio, plaza y promoción son elementos esenciales del marketing mix. En la banca, el *producto* incluye cuentas, créditos e inversiones; el *precio* se refiere a comisiones e intereses; la *plaza* abarca sucursales físicas y digitales; y la *promoción* incluye publicidad y campañas de captación (**kotler2016**).

3.0.3. Teoría del valor percibido

zeithaml1988 propone que la decisión del consumidor se basa en la percepción del valor recibido en relación al costo. En este sentido, los bancos deben gestionar las percepciones del cliente ofreciendo beneficios tangibles (como tasas competitivas) e intangibles (como seguridad y confianza).

3.0.4. Marketing digital y automatización

Autores contemporáneos como **chaffey2020** argumentan que el marketing en el sector financiero está siendo redefinido por la inteligencia artificial, la automatización y los análisis predictivos. Estas herramientas permiten desarrollar campañas más efectivas y personalizadas, así como anticiparse a las necesidades del cliente.

4. Cuaderno de experimentos.

4.1. Descripción del conjunto de datos

El proyecto analiza los datos de campañas de marketing previas de un banco portugués y busca predecir si el cliente suscribirá los productos de depósito a plazo fijo que ofrece el banco.

Los datos están relacionados con campañas de marketing directo de una institución bancaria portuguesa entre 2008 y 2010. Las campañas de marketing se basaron en llamadas telefónicas. A menudo, se requería más de un contacto con el mismo cliente para determinar si el producto se suscribiría o no. Por lo tanto, su conjunto de datos se basa en una clasificación binaria.

4.1.1. Información de variables.

- **age:** Edad.
- **job:** Tabajo.
- **marital:** Estatus marital.
- **education:** Nivel educativo.
- **default:** ¿Tiene crédito en mora?
- **housing:** ¿Tiene préstamo hipotecario?
- **loan:** ¿Tiene préstamo personal?
- **contact:** Tipo de contacto.
- **month:** Mes del ultimo contacto.
- **day_of_week:** Día del ultimo contacto.
- **duration:** Duración del ultimo contacto (segundos).
- **campaign:** Número de contactos realizados durante la campaña para un cliente:
- **pdays:** Número de días transcurridos desde el último contacto con el cliente de una campaña anterior
- **previous:** Número de contactos realizados antes de esta campaña y para este cliente.
- **poutcome:** Resultado de la campaña de marketing anterior.
- **emp.var.rate:** Tasa de variación del empleo -indicador trimestral.
- **cons.price.idx:** Índice de precios al consumo - indicador mensual.
- **cons.conf.idx:** Índice de confianza del consumidor - indicador mensual.
- **euribor3m:** Tasa del euríbor a 3 meses - indicador diario.
- **nr.employed:** Número de empleados - indicador trimestral
- **y:** ¿El cliente ha suscrito un depósito a plazo?

4.1.2. Clasificación de variables.

Para realizar la limpieza de nuestros datos, debemos verificar primero que nada que todo este en orden, notamos algunas inconsistencias en las columnas como ciertos espacios irrelevantes, así que los eliminamos para poder trabajar los datos de mejor manera. Lo siguiente es separar las variables de acuerdo con su clasificación, recordando que nuestro set de datos tiene variables de dos tipos, los cuales son categóricas y numéricas.

Índice	Variable	Tipo
0	age	Numérica
1	duration	Numérica
2	campaign	Numérica
3	pdays	Numérica
4	previous	Numérica
5	emp.var.rate	Numérica
6	cons.price.idx	Numérica
7	cons.conf.idx	Numérica
8	euribor3m	Numérica
9	nr.employed	Numérica
10	job	Categórica
11	marital	Categórica
12	education	Categórica
13	default	Categórica
14	housing	Categórica
15	loan	Categórica
16	contact	Categórica
17	month	Categórica
18	day_of_week	Categórica
19	poutcome	Categórica
20	y	Categórica

Cuadro 1: Lista de variables y sus tipos

4.2. Porcentaje de valores nulos.

Al revisar el conjunto de datos, algo importante salta a la vista: no hay valores nulos. Eso, de entrada, es una muy buena señal. Significa que no falta información en ninguna de las columnas, lo que nos ahorra un montón de trabajo en la etapa de limpieza. No hace falta rellenar huecos, eliminar filas incompletas ni tomar decisiones sobre cómo tratar datos ausentes.

Tener los datos completos desde el principio también ayuda a que los modelos funcionen mejor, ya que muchos algoritmos no se llevan bien con valores faltantes. Además, todo el análisis que venga después (gráficas, estadísticas, correlaciones) va a ser más confiable, porque no hay "ruido" causado por datos incompletos.

A continuación se presenta el porcentaje de valores nulos:

Índice	Variable	Porcentaje
0	age	0.0
1	duration	0.0
2	campaign	0.0
3	pdays	0.0
4	previous	0.0
5	emp.var.rate	0.0
6	cons.price.idx	0.0
7	cons.conf.idx	0.0
8	euribor3m	0.0
9	nr.employed	0.0
10	job	0.0
11	marital	0.0
12	education	0.0
13	default	0.0
14	housing	0.0
15	loan	0.0
16	contact	0.0
17	month	0.0
18	day_of_week	0.0
19	poutcome	0.0
20	y	0.0

Cuadro 2: Lista de porcentajes nulos

4.2.1. Tratamiento de la variable Pdays

Recordemos que la variable **pdays** es una variable de tipo numérico, donde el valor 999 indica que el cliente no fue contactado previamente. Este valor especial introduce un sesgo importante en el análisis estadístico y en los modelos predictivos, ya que distorsiona la distribución de la variable y no representa una medida real del tiempo, sino la ausencia de contacto.

Para abordar este problema, se puede optar por dos estrategias que permiten preservar y aprovechar la información de manera más efectiva:

1. **Imputación con la media o mediana.**
   Antes de imputar, es buena práctica revisar ambos valores:

Variable	Media	Mediana
pdays	962.48	999.00

Cuadro 3: Media y mediana sesgada.

Variable	Media	Mediana
pdays	6.00	6.01

Cuadro 4: Propuesta para imputar la varaible pdays.

Comparando los cuadros anteriores, podríamos pensar que imputar con la media o la mediana es una opción aceptable para el tratamiento de la variable. Sin embargo, realizar esta imputación cambia completamente el sentido de la variable. En consecuencia, imputar la variable **pdays** con la media o mediana no es apropiado, ya que el valor 999 no representa un número faltante real, sino que indica que el cliente no fue contactado previamente. Tratarlo como si fuera un número normal falsifica el historial del cliente y genera sesgos en el modelo.

2. **Categorización de la variable pdays.**
La decisión de categorizar la variable se basa en varias razones técnicas y analíticas que buscan preservar la integridad de los datos, evitar sesgos y mejorar la interpretación del modelo.

Índice	Variable	Tipo
0	age	Numérica
1	duration	Numérica
2	campaign	Numérica
3	pdays	Categórica
4	previous	Numérica
5	emp.var.rate	Numérica
6	cons.price.idx	Numérica
7	cons.conf.idx	Numérica
8	euribor3m	Numérica
9	nr.employed	Numérica
10	job	Categórica
11	marital	Categórica
12	education	Categórica
13	default	Categórica
14	housing	Categórica
15	loan	Categórica
16	contact	Categórica
17	month	Categórica
18	day_of_week	Categórica
19	poutcome	Categórica
20	y	Categórica

Cuadro 5: Lista de variables y sus tipos

Tomamos la decisión de cambiar el nombre de la variable **pdays** por **contacted** para refererirnos a la variable categórica que permite una lectura más clara de los resultados del modelo y de los análisis descriptivos. Categorías binarias "Yes" o "No" son fáciles de entender para los equipos de negocio o marketing, en comparación con valores numéricos.

Índice	Variable	Tipo
0	age	Numérica
1	duration	Numérica
2	campaign	Numérica
3	previous	Numérica
4	emp.var.rate	Numérica
5	cons.price.idx	Numérica
6	cons.conf.idx	Numérica
7	euribor3m	Numérica
8	nr.employed	Numérica
9	contacted	Categórica
10	job	Categórica
11	marital	Categórica
12	education	Categórica
13	default	Categórica
14	housing	Categórica
15	loan	Categórica
16	contact	Categórica
17	month	Categórica
18	day_of_week	Categórica
19	poutcome	Categórica
20	y	Categórica

Cuadro 6: Lista de variables y sus tipos

4.3. Proporciones y frecuencias de variables categóricas.

4.3.1. Proporciones

1. **Default**

Default	Proporción
no	32588
unknown	8597
yes	3

Cuadro 7: Proporción de la variable Default.

2. **Job**

Job	Proporción
admin	10422
blue_collar	9254
entrepreneur	1456
housemaid	1060
management	2924
retired	1720
self_employed	1421
services	3969
student	875
technician	6743
unemployed	1014
unknown	330

Cuadro 8: Proporción de la variable Job.

3. **Education**

Education	Proporción
basic	12513
highschool	9515
illiterate	18
professionalcourse	5243
universitydegree	12168
unknown	1731

Cuadro 9: Proporción de la variable Education.

4. **Marital**

Marital	Proporción
divorced	4612
married	24928
single	11568
unknown	80

Cuadro 10: Proporción de la variable Marital.

5. **Month**

Month	Proporción
apr	2632
aug	6178
dec	182
jul	7174
jun	5318
mar	546
may	13769
nov	4101
oct	718
sep	570

Cuadro 11: Proporción de la variable Month.

6. **Day of week**

Day of week	Proporción
fri	7827
mon	8514
thu	8623
tue	8090
wed	8134

Cuadro 12: Proporción de la variable Day of week.

7. **Housing**

Housing	Proporción
no	18622
yes	21576
unknown	990

Cuadro 13: Proporción de la variable Housing.

8. **Loan**

Loan	Proporción
no	33950
yes	6248
unknown	990

Cuadro 14: Proporción de la variable Loan.

9. **Contact**

Contact	Proporción
cellular	26144
telephone	15044

Cuadro 15: Proporción de la variable Contact.

10. **Poutcome**

Poutcome	Proporción
failure	4252
nonexistent	35563
success	1373

Cuadro 16: Proporción de la variable Poutcome.

11. **Y**

Y	Proporción
no	36548
yes	4640

Cuadro 17: Proporción de la variable Y.

12. **Contacted**

Contacted	Proporción
no	39673
yes	1515

Cuadro 18: Proporción de la variable Contacted.

4.3.2. Frecuencias

1. **Default**

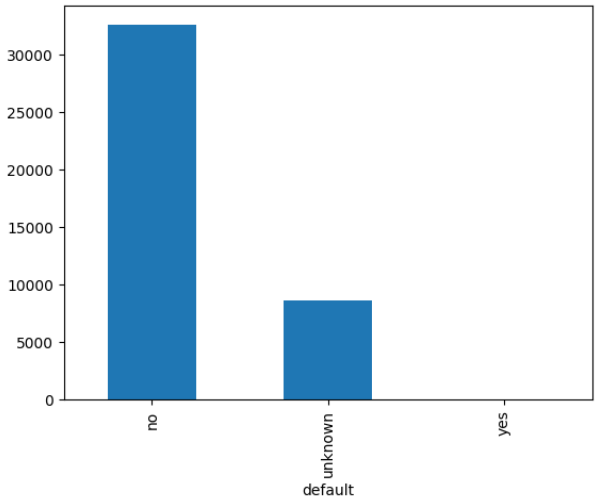

Figura 1: Frecuencias de la variable Default.

2. **Job**

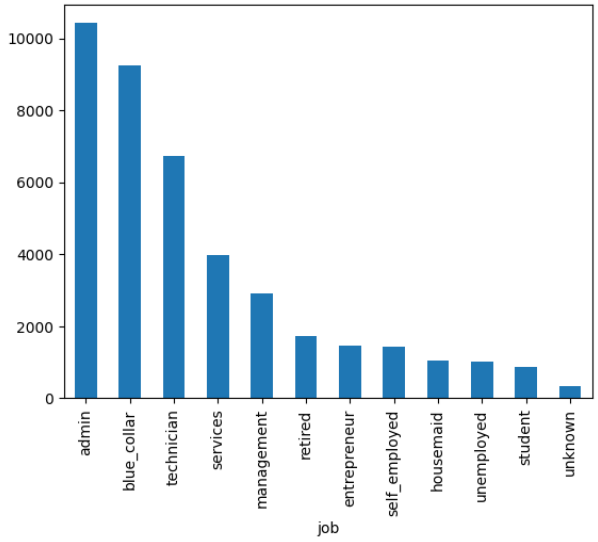

Figura 2: Frecuencias de la variable Job.

3. **Education**

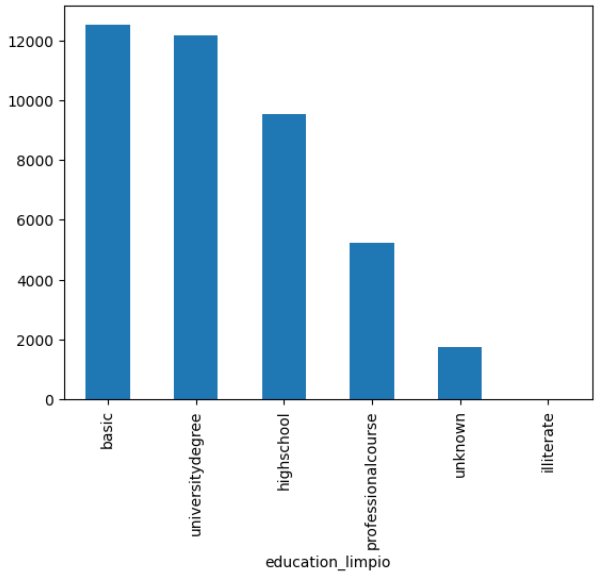

Figura 3: Frecuencias de la variable Education.

4. **Marital**

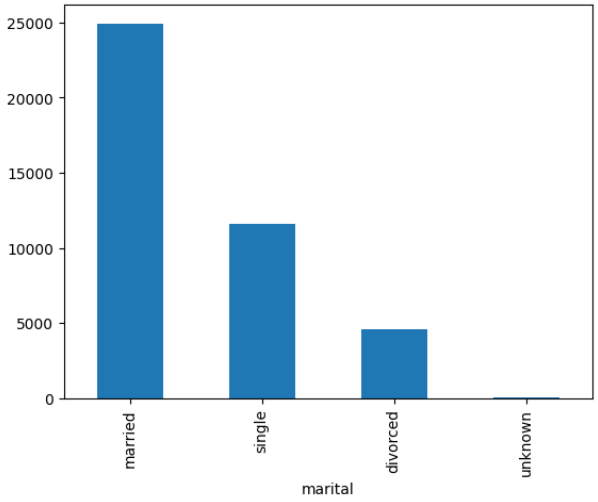

Figura 4: Frecuencias de la variable Marital.

5. **Month**

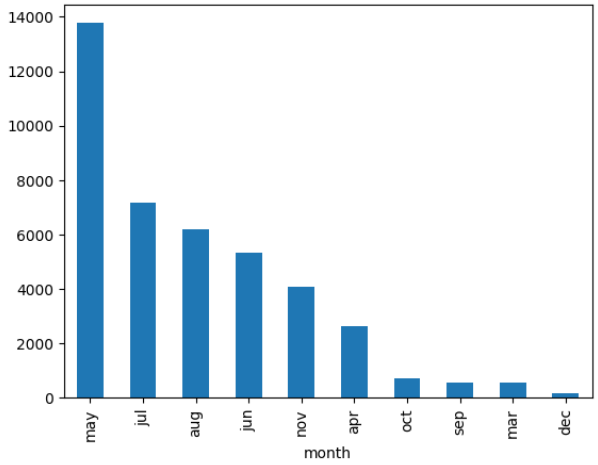

Figura 5: Frecuencias de la variable Month.

6. **Day of week**

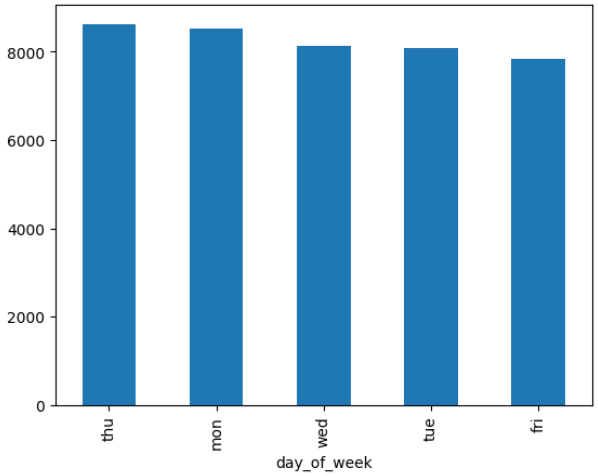

Figura 6: Frecuencias de la variable Day of week.

7. **Housing**

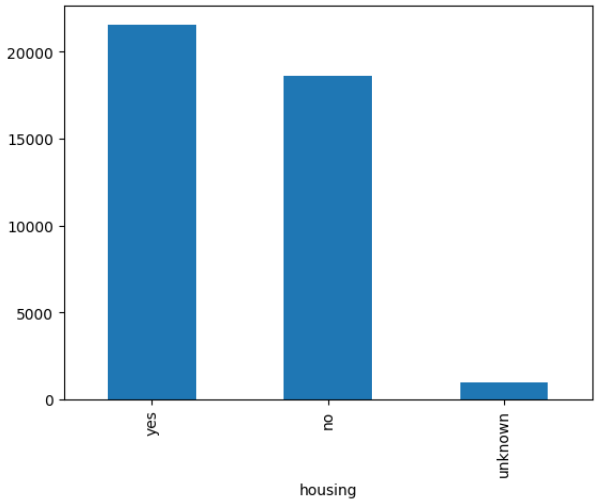

Figura 7: Frecuencias de la variable Housing.

8. **Loan**

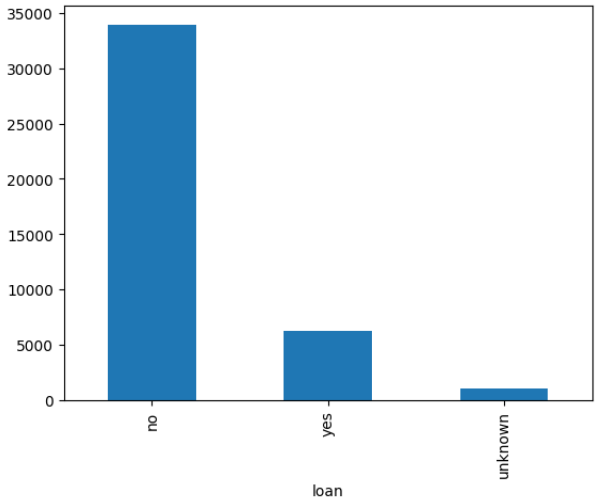

Figura 8: Frecuencias de la variable Loan.

9. **Contact**

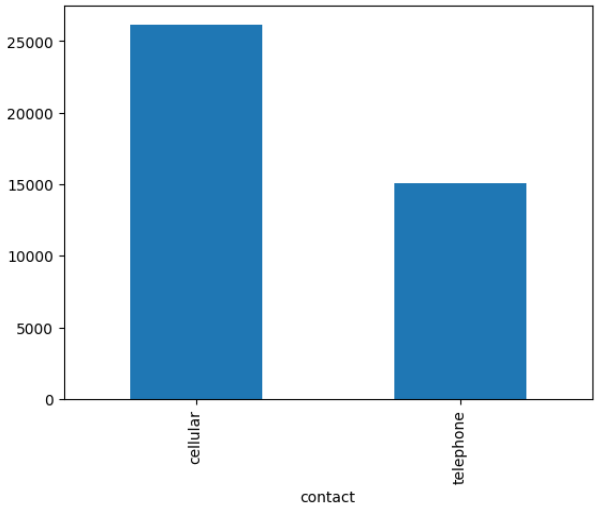

Figura 9: Frecuencias de la variable Contact.

10. **Poutcome**

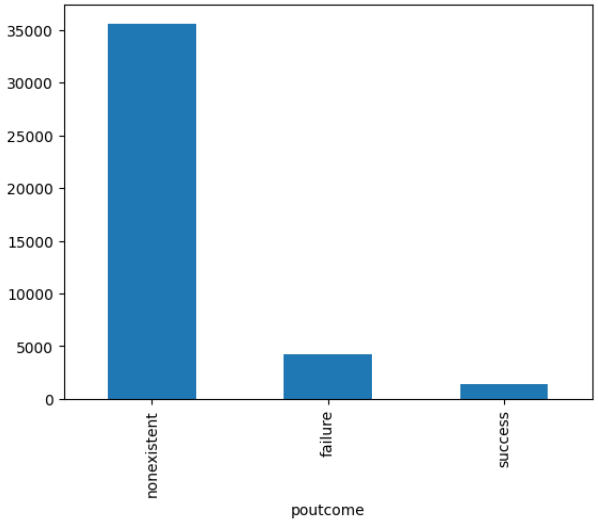

Figura 10: Frecuencias de la variable Poutcome.

11. **Y**

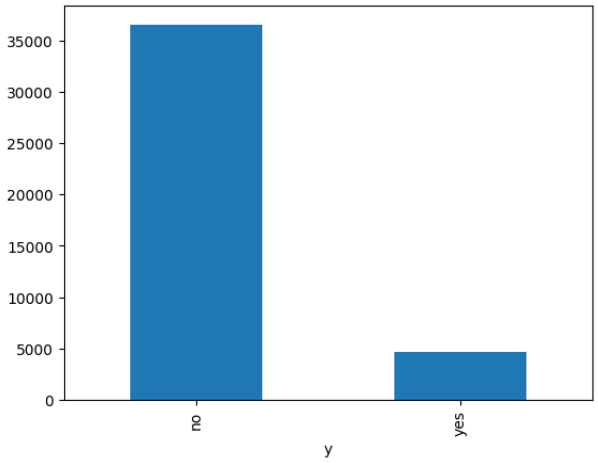

Figura 11: Frecuencias de la variable Y.

12. **Contacted**

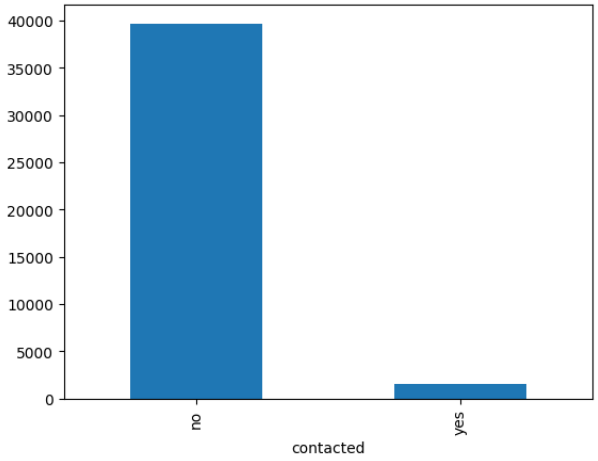

Figura 12: Frecuencias de la variable Contacted.

4.4. Estadísticas descriptivas de las variables numéricas.

Índice	Variable	Media	Mediana	Desv. Est.	Varianza
0	age	40.02	38.00	10.42	108.60
1	duration	258.29	180.00	259.28	67225.73
2	campaign	2.57	2.00	2.77	7.67
3	previous	0.17	0.00	0.49	0.24
4	emp.var.rate	0.08	1.10	1.57	2.47
5	cons.price.idx	93.58	93.75	0.58	0.34
6	cons.conf.idx	-40.50	-41.80	4.63	21.42
7	euribor3m	3.62	4.86	1.73	3.01
8	nr.employed	5167.04	5191.00	72.25	5220.28

Cuadro 19: Estadísticas descriptivas parte 1.

Índice	Variable	Mínimo	Máximo	Rango	Asimetría	Curtosis
0	age	17.00	98.00	81.00	0.78	0.79
1	duration	0.00	4918.00	4918.00	3.26	20.25
2	campaign	1.00	56.00	55.00	4.76	36.98
3	previous	0.00	7.00	7.00	3.83	20.11
4	emp.var.rate	-3.40	1.40	4.80	-0.72	-1.06
5	cons.price.idx	92.20	94.77	2.56	-0.28	-0.83
6	cons.conf.idx	-50.80	-26.90	23.90	0.30	-0.36
7	euribor3m	0.63	5.05	4.41	-0.71	-1.41
8	nr.employed	4963.60	5228.10	264.50	-1.04	-0.00

Cuadro 20: Estadísticas descriptivas parte 2.

4.4.1. Interpretación de las estadísticas descriptivas.

1. **age:** La media de edad es 40 años, lo que indica una base de clientes adultos. La mediana de 38 años sugiere que la mitad de los clientes son menores de esa edad. La diferencia entre la media y la mediana es pequeña, indicando una distribución algo simétrica, aunque con una ligera asimetría positiva (cola derecha), es decir, existen clientes de edad muy avanzada (hasta 98) que estiran la distribución. La alta desviación estándar (10.42) indica que hay bastante variedad generacional entre los clientes. Este rango etario amplio puede ser estratégico: diferentes productos financieros pueden ser dirigidos según grupos etarios.

2. **duration:** Esta variable tiene una alta media (258s) y gran dispersión. La mediana de 180s indica que más de la mitad de las llamadas duran menos de 3 minutos, pero algunos contactos mucho más largos inflan el promedio. La asimetría positiva extrema (3.26) y curtosis elevada (20.25) implican que hay muchos valores atípicos: llamadas muy largas. Estas podrían corresponder a llamadas con clientes interesados o difíciles.

3. **campaign:** El promedio de contactos es 2.57, pero hay clientes contactados hasta 56 veces. Eso indica que aunque en general se respeta un número bajo de intentos, hay sobrecontacto en algunos casos. La asimetría y curtosis muy altas refuerzan la presencia de una minoría de clientes excesivamente contactados. Esto puede ser problemático en términos de imagen de marca y eficiencia.

4. **previous:** La media es 0.17 y la mediana 0.00: la mayoría nunca había sido contactada anteriormente. Sin embargo, la asimetría alta indica que unos pocos clientes fueron contactados varias veces en campañas anteriores.

5. **emp.var.rate:** Este indicador macroeconómico refleja la salud del mercado laboral. Tiene una media cercana a 0 pero puede bajar hasta -3.4. La asimetría negativa sugiere que hay más registros en escenarios positivos, pero algunos reflejan crisis económicas.

6. **cons.price.idx:**Es un indicador relativamente estable (desviación de solo 0.58). Su estabilidad implica que no hay gran impacto en el corto plazo, pero puede ser útil para comparar con otras variables económicas.

7. **cons.conf.idx:** Todos los valores son negativos, lo cual sugiere un entorno generalizado de desconfianza (común en tiempos post-crisis). La ligera asimetría y curtosis plana indican una distribución estable.

8. **euribor3m:** Refleja las condiciones del crédito en la eurozona. Un valor alto desincentiva préstamos; uno bajo los vuelve más atractivos. La mediana más alta que la media indica

una distribución sesgada a la izquierda (la mayoría de los registros son anteriores a recortes de tasas).

9. **nr.employed:** Es un proxy del tamaño del mercado laboral activo. La distribución está centrada pero ligeramente sesgada a la izquierda. La variabilidad es baja, lo que indica un entorno relativamente estable.

4.4.2. Histogramas de las variables numéricas.

1. **Age**

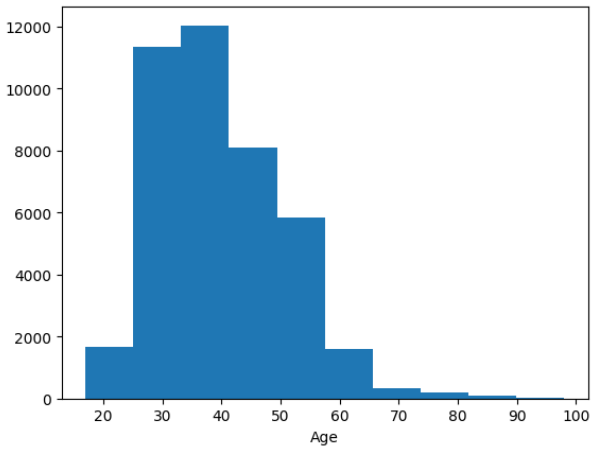

Figura 13: Histogramas de la variable Age.

2. **Duration**

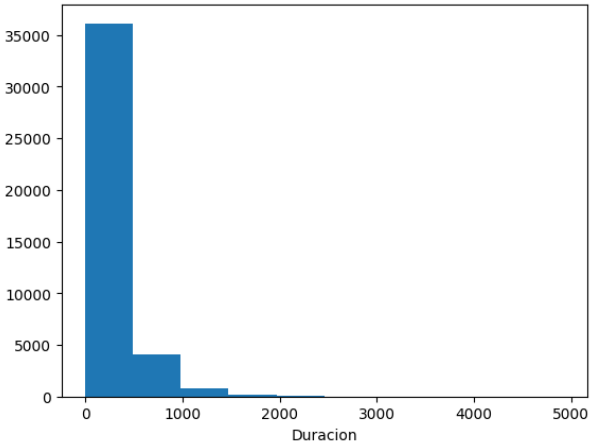

Figura 14: Histograma de la variable Duration.

3. **Campaing**

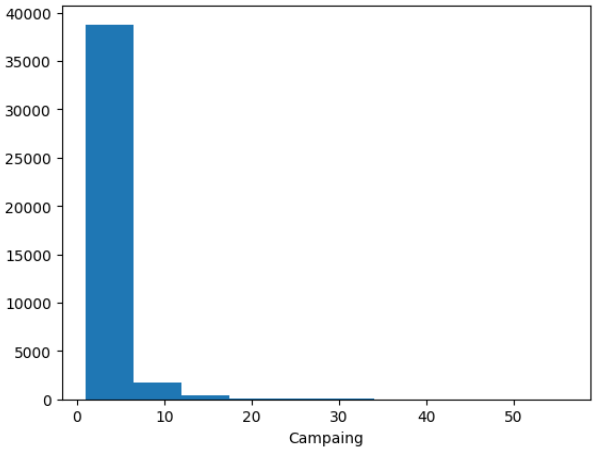

Figura 15: Histograma de la variable Campaing.

4. **emp.var.rate**

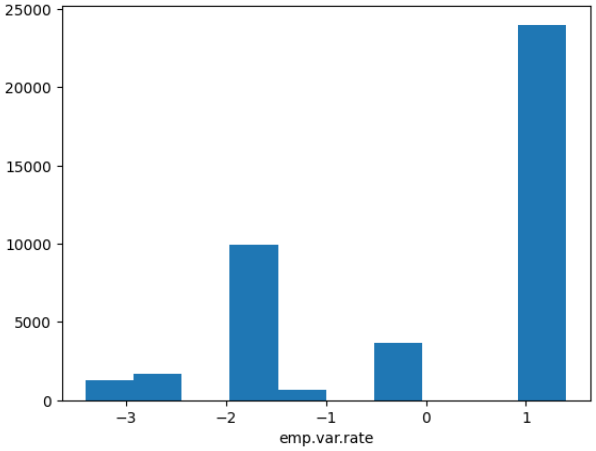

Figura 16: Histogramas de la variable emp.var.rate.

5. **cons.price.idx**

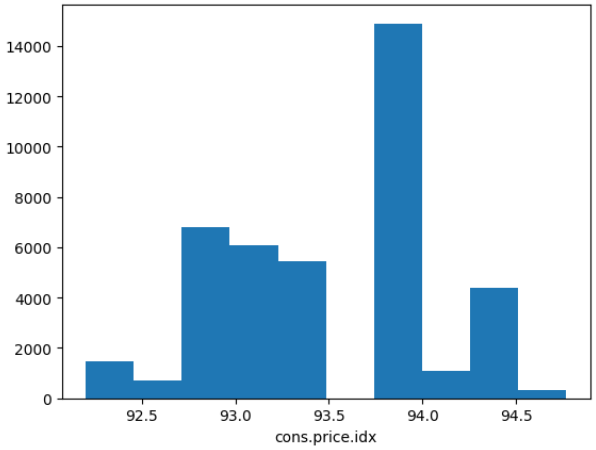

Figura 17: Histogramas de la variable cons.price.idx.

6. **cons.conf.idx**

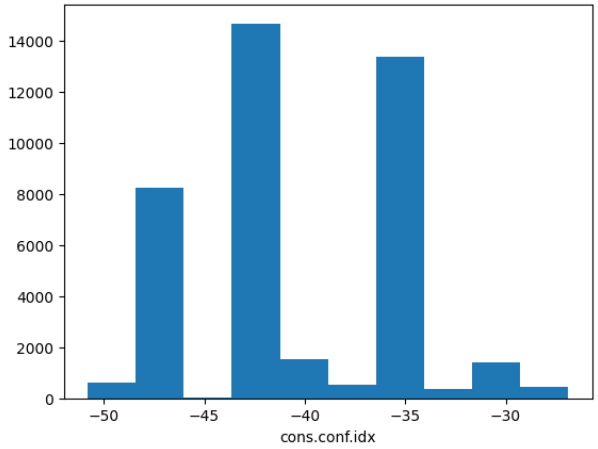

Figura 18: Histogramas de la variable cons.conf.idx.

7. **euribor3m**

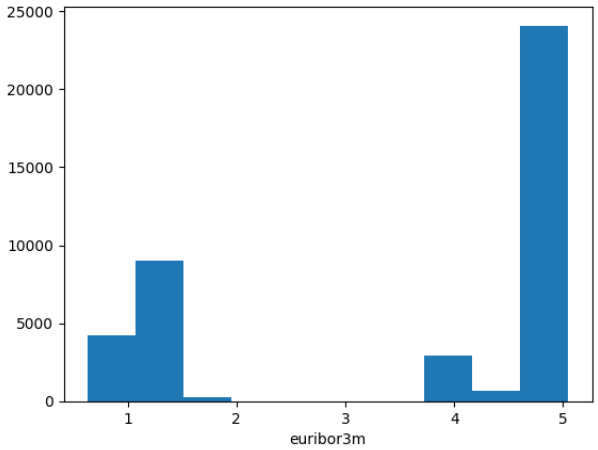

Figura 19: Histogramas de la variable euribor3m.

8. **nr.employed**

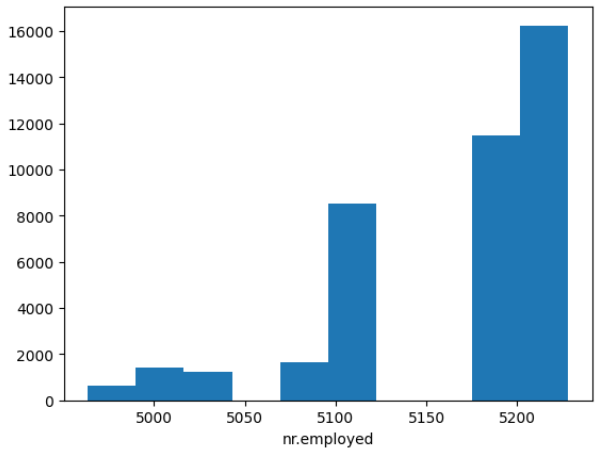

Figura 20: Histogramas de la variable nr.employed.

9. **Previous**

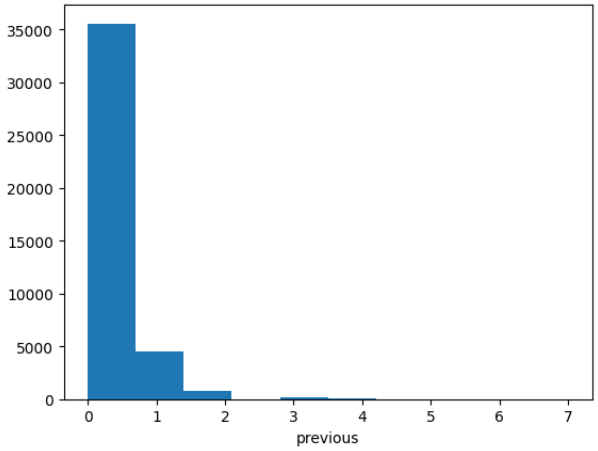

Figura 21: Histogramas de la variable Previous.

4.5. Diagramas de cajas y bigotes

1. **Age**

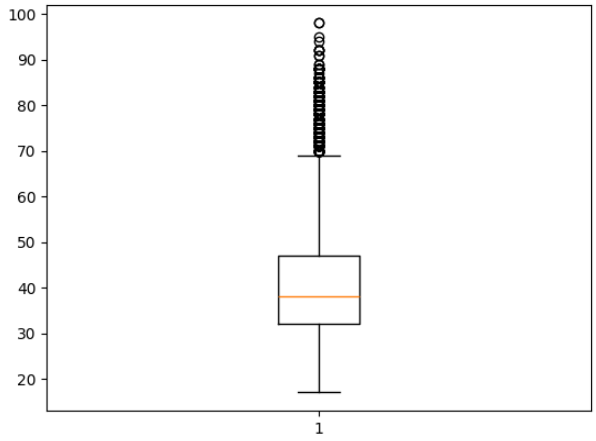

Figura 22: Boxplot de la variable Age.

2. **Duration**

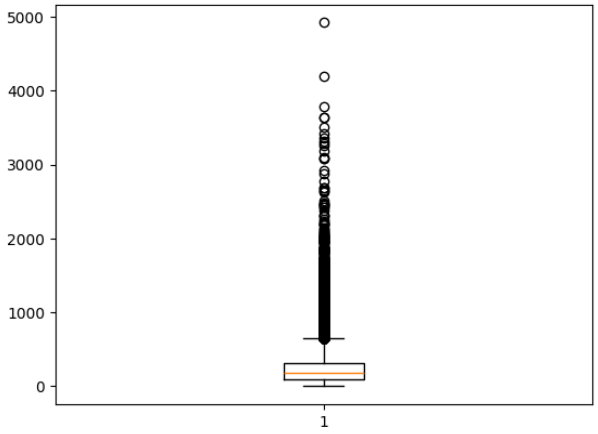

Figura 23: Boxplot de la variable Duration.

3. **Campaing**

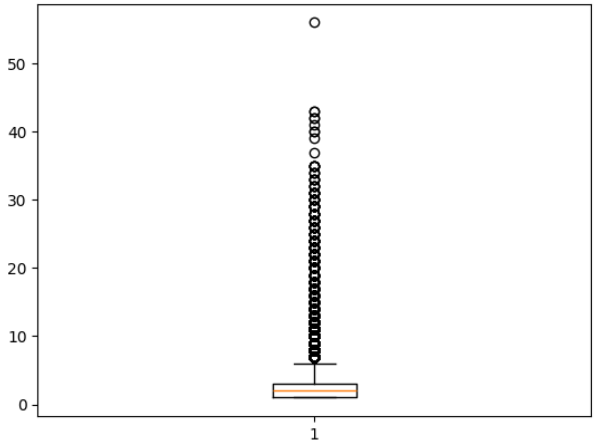

Figura 24: Boxplot de la variable Campaing.

4. **emp.var.rate**

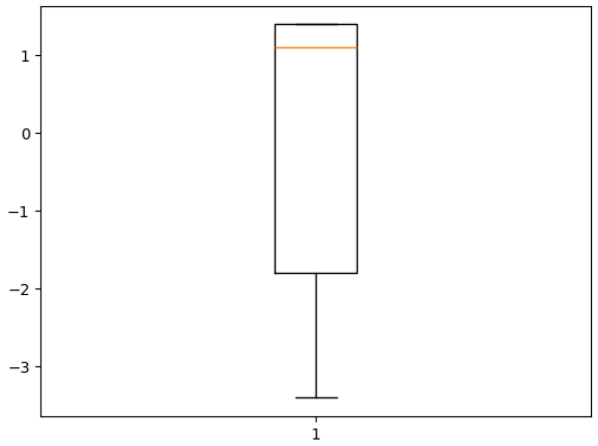

Figura 25: Boxplot de la variable emp.var.rate.

5. **cons.price.idx**

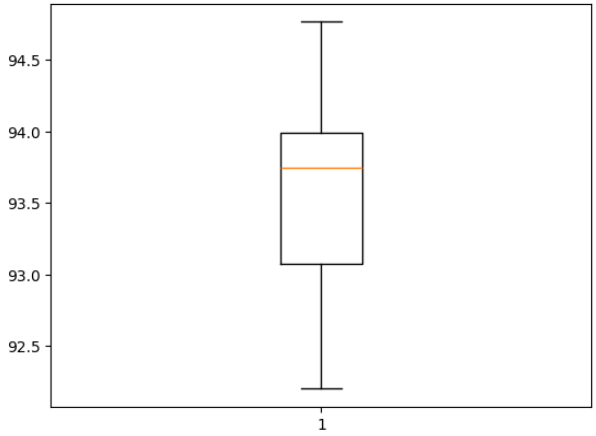

Figura 26: Boxplot de la variable cons.price.idx.

6. **cons.conf.idx**

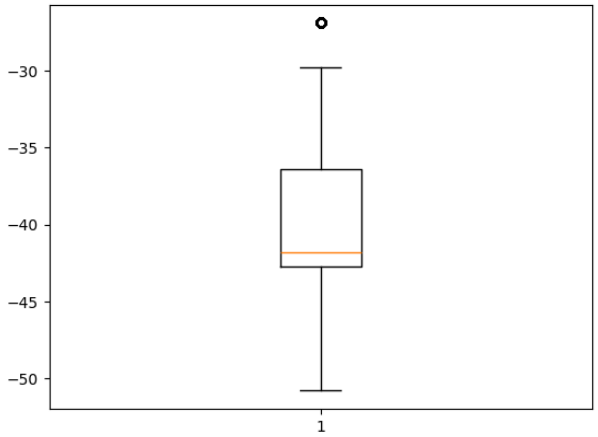

Figura 27: Boxplot de la variable cons.conf.idx.

7. **euribor3m**

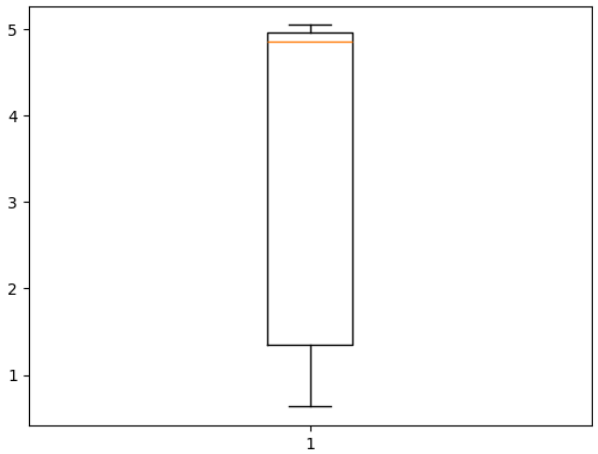

Figura 28: Boxplot de la variable euribor3m.

8. **nr.employed**

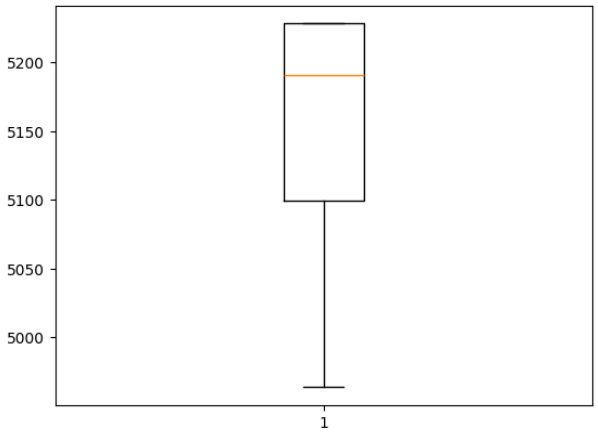

Figura 29: Boxplot de la variable nr.employed.

9. **Previous**

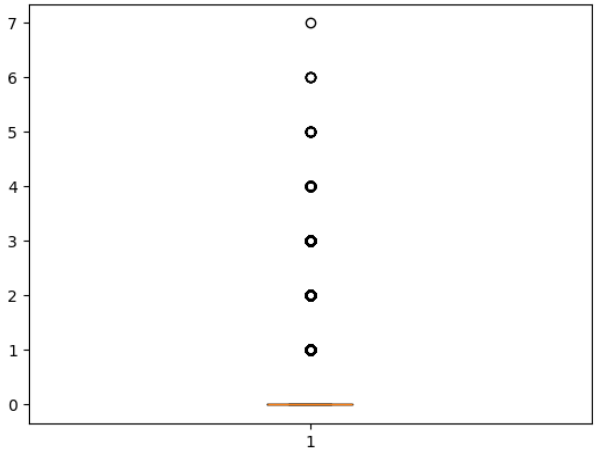

Figura 30: Boxplot de la variable Previous.

4.6. Aplicación de Z-score y RIC

Luego de visualizar los diagrama de cajas y bigotes podemos asignar que variables son aplicables a las funciones Z-score y RIC. Las variables aplicables a cada modelo son:

- **Z-score**
  - age
  - cons.price.idx
  - cons.conf.idx
- **RIC**
  - duration
  - campaign

- previous
- emp.var.rate
- euribor3m
- nr.employed

4.6.1. Diagramas de cajas y bigotes sin outliers

1. **Age**

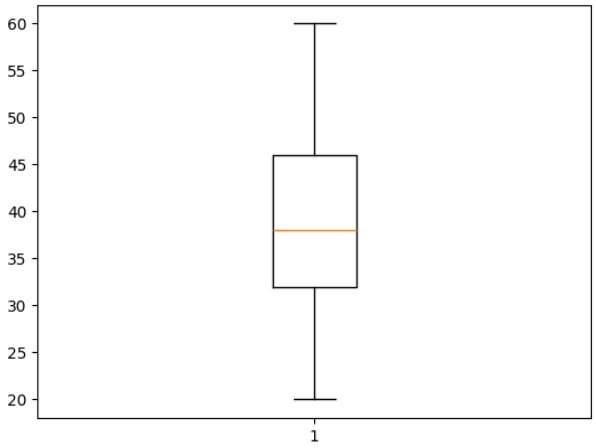

Figura 31: Boxplot sin outliers de la variable Age.

2. **Duration**

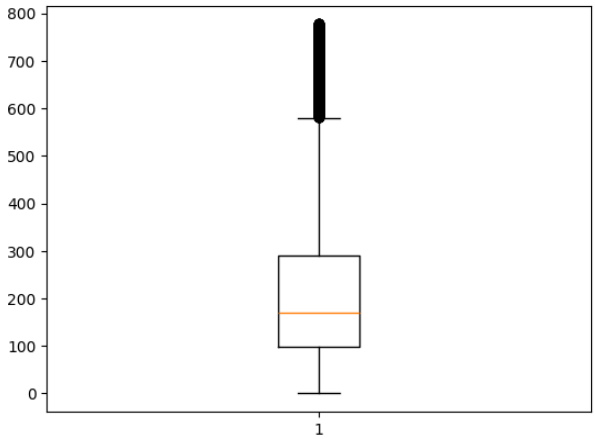

Figura 32: Boxplot sin outliers de la variable Duration.

3. **Campaing**

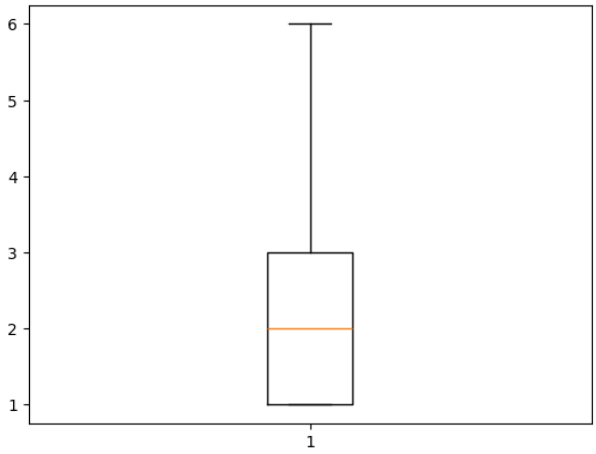

Figura 33: Boxplot sin outliers de la variable Campaing.

4. **emp.var.rate**

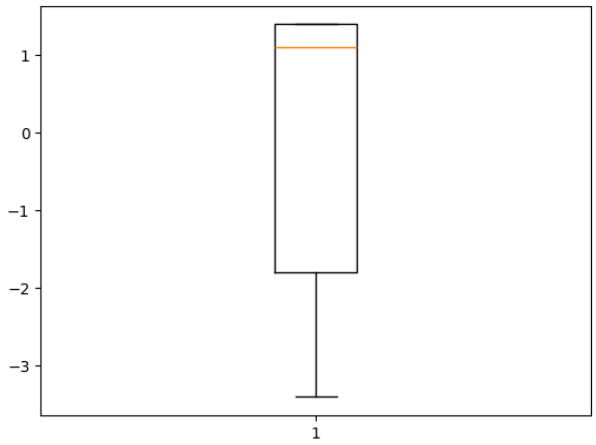

Figura 34: Boxplot sin outliers de la variable emp.var.rate.

5. **cons.price.idx**

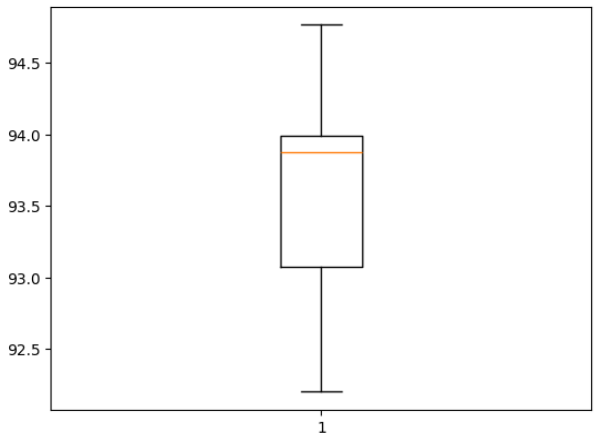

Figura 35: Boxplot sin outliers de la variable cons.price.idx.

6. **cons.conf.idx**

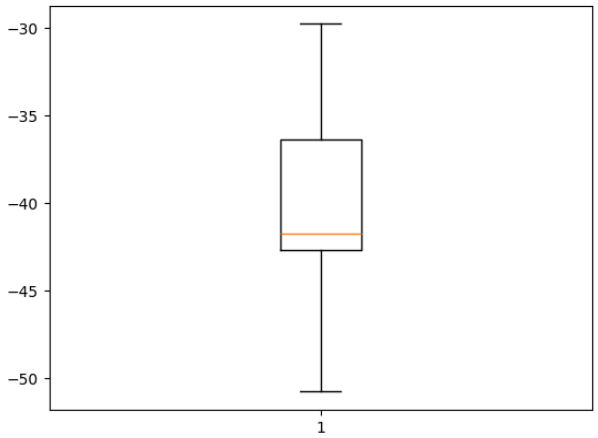

Figura 36: Boxplot sin outliers de la variable cons.conf.idx.

7. **euribor3m**

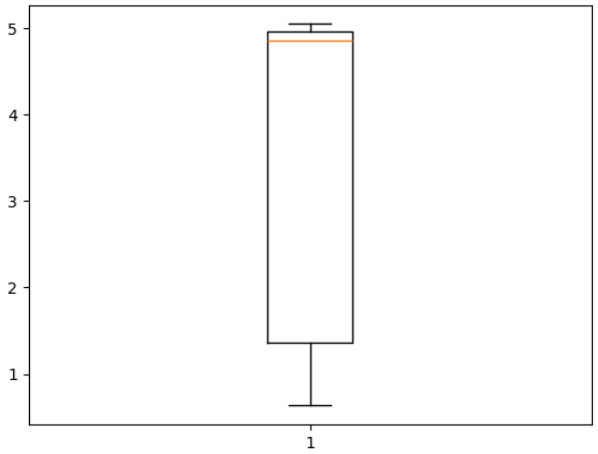

Figura 37: Boxplot sin outliers de la variable euribor3m.

8. **nr.employed**

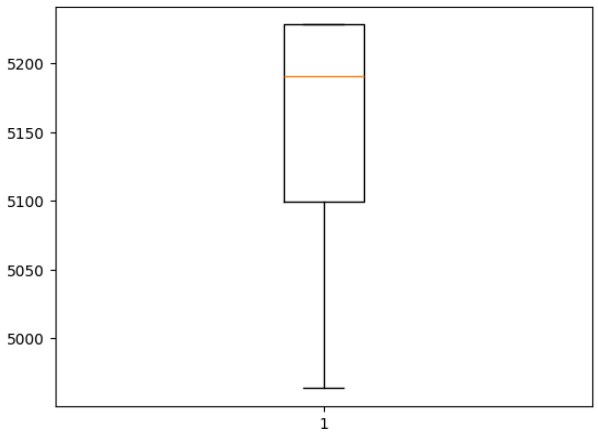

Figura 38: Boxplot sin outliers de la variable nr.employed.

9. **Previous**

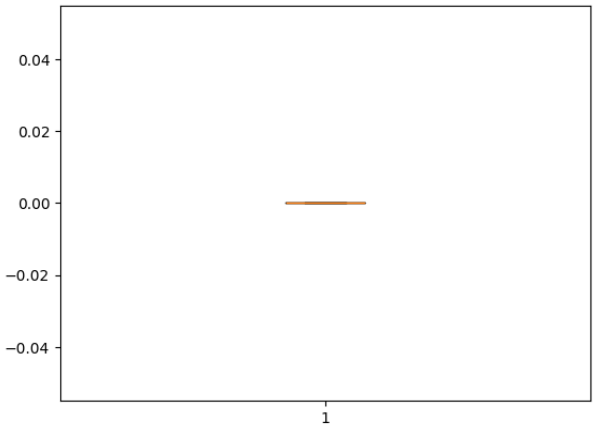

Figura 39: Boxplot sin outliers de la variable Previous.

4.6.2. Matriz de Covarianzas

	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	84.72	-2.97	0.27	-0.14	1.75	0.40	3.77	2.09	75.57
duration	-2.97	25346.32	-31.03	0.10	-6.10	0.12	-12.25	-7.54	-387.40
campaign	0.27	-31.03	4.16	-0.05	0.40	0.12	0.06	0.37	15.78
previous	-0.14	0.10	-0.05	0.11	-0.22	-0.06	-0.27	-0.25	-9.06
emp.var.rate	1.75	-6.10	0.40	-0.22	2.24	0.67	2.44	2.41	87.70
cons.price.idx	0.40	0.12	0.12	-0.06	0.67	0.30	0.55	0.70	22.95
cons.conf.idx	3.77	-12.25	0.06	-0.27	2.44	0.55	18.50	3.12	68.08
euribor3m	2.09	-7.54	0.37	-0.25	2.41	0.70	3.12	2.70	97.70
nr.employed	75.57	-387.40	15.78	-9.06	87.70	22.95	68.08	97.70	3856.28

Cuadro 21: Matriz de correlación entre variables numéricas

La matriz de covarianzas permite observar cómo varían juntas las variables considerando sus unidades originales:

- Varianza alta: Las variables duration y nr.employed muestran alta dispersión, con valores grandes en la diagonal de la matriz.
- Covarianzas altas entre variables económicas: Las variables emp.var.rate, euribor3m y nr.employed tienen covarianzas elevadas, lo cual refuerza su relación directa en contextos económicos.
- Covarianzas cercanas a cero o negativas: Algunas variables como age y campaign tienen covarianza baja con el resto, y previous presenta covarianza negativa con indicadores económicos, lo que sugiere que su comportamiento es opuesto a estos.

4.6.3. Representación gráfica por mapa de colores.

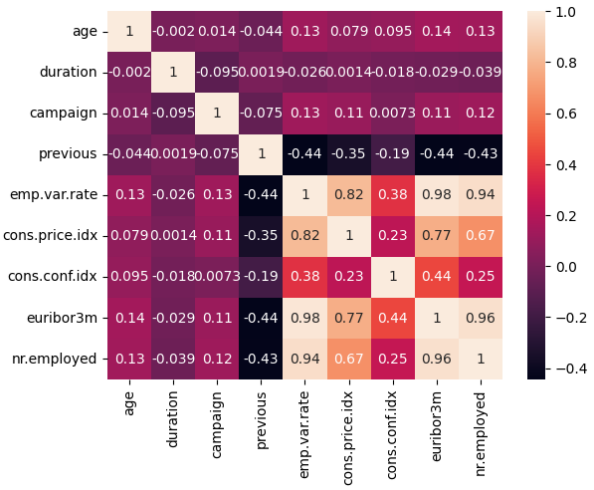

Figura 40: Heatmap de las variables numericas.

El heatmap muestra la intensidad y dirección de las relaciones lineales entre las variables analizadas:

- Fuertes correlaciones positivas: Las variables emp.var.rate, euribor3m y nr.employed están altamente correlacionadas, lo que indica que tienden a aumentar juntas en contextos económicos favorables.
- Correlaciones negativas moderadas: La variable previous presenta correlación negativa con emp.var.rate y euribor3m, lo que sugiere que un mayor número de contactos anteriores ocurre cuando las condiciones económicas no son óptimas.
- Relaciones débiles o nulas: Variables como age, campaign y duration presentan correlaciones cercanas a cero con el resto de las variables, lo que indica poca o nula relación lineal.

4.6.4. Visualización de las relaciones de la variables.

Figura 41: Pairplot de las variables numericas.

El gráfico de pares permite visualizar relaciones bivariadas y distribuciones univariadas entre las variables cuantitativas del conjunto de datos.

- Las distribuciones univariadas muestran que variables como age, duration y nr.employed tienen una distribución asimétrica con posibles valores extremos. Variables como previous y campaign presentan concentración de valores bajos, especialmente ceros.
- Se observa una relación lineal positiva entre emp.var.rate, euribor3m y nr.employed, lo cual refuerza los resultados del análisis de correlación.
- No hay evidencia de relaciones lineales entre variables como age, campaign y duration con el resto, indicando que podrían tener poca utilidad predictiva lineal.

- Existen agrupamientos en previous y campaign, lo cual puede reflejar prácticas de contacto previas o naturaleza discreta en esos datos.
- Se identifican posibles valores atípicos, especialmente en duration y nr.employed, lo cual sugiere la necesidad de un análisis de outliers.

4.6.5. Correlación de Kendall

La **correlación de Kendall**, también conocida como *tau de Kendall*, mide la concordancia entre pares de observaciones. Se basa en comparar todos los pares posibles de datos y contar cuántos están en el mismo orden (*pares concordantes*) y cuántos en orden inverso (*pares discordantes*).

La fórmula es:

$$\tau = \frac{(\text{número de pares concordantes}) - (\text{número de pares discordantes})}{\text{número total de pares}}$$

Ventajas:

- Robusta frente a valores atípicos.
- Interpretación clara: mide la probabilidad de concordancia entre pares.

Aplicando este método a nuestro set datos logramos observar lo siguiente: En general, las correlaciones son bajas, lo que sugiere relaciones débiles entre la mayoría de las variables. Sin embargo, destacan correlaciones moderadamente altas y positivas entre emp.var.rate, euribor3m y nr.employed, lo que indica que estas variables económicas están relacionadas entre sí. También se observa una correlación negativa entre previous y las variables económicas, lo cual podría reflejar patrones en el historial de contacto previo y el contexto económico.

	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.000000	-0.001458	0.004324	-0.010437	0.035119	0.032673	0.083010	0.040112	0.035517
duration	-0.001458	1.000000	-0.060037	0.034326	-0.051260	0.002488	-0.006136	-0.052531	-0.070541
campaign	0.004324	-0.060037	1.000000	-0.078623	0.129366	0.075459	-0.001085	0.106388	0.118665
previous	-0.010437	0.034326	-0.078623	1.000000	-0.389434	-0.243995	-0.100283	-0.375834	-0.396231
emp.var.rate	0.035119	-0.051260	0.129366	-0.389434	1.000000	0.525641	0.200415	0.827605	0.845130
cons.price.idx	0.032673	0.002488	0.075459	-0.243995	0.525641	1.000000	0.155028	0.352392	0.376950
cons.conf.idx	0.083010	-0.006136	-0.001085	-0.100283	0.200415	0.155028	1.000000	0.173984	0.101984
euribor3m	0.040112	-0.052531	0.106388	-0.375834	0.827605	0.352392	0.173984	1.000000	0.788706
nr.employed	0.035517	-0.070541	0.118665	-0.396231	0.845130	0.376950	0.101984	0.788706	1.000000

Cuadro 22: Matriz de correlación de Kendall entre variables

4.6.6. Correlación de Spearman (ρ)

La **correlación de Spearman** mide la relación monótona entre dos variables. Convierte los valores a rangos y luego aplica la fórmula de correlación de Pearson sobre esos rangos.

Si no hay empates, su fórmula es:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre los rangos de cada par de observaciones.

Ventajas:

- Útil para relaciones no lineales pero monótonas.
- Más rápida de calcular que Kendall.

	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.000000	-0.008947	0.017990	-0.049005	0.100111	0.074719	0.098290	0.114341	0.107867
duration	-0.008947	1.000000	-0.094299	0.043064	-0.075062	-0.001042	-0.008388	-0.080592	-0.100543
campaign	0.017990	-0.094299	1.000000	-0.084130	0.152834	0.094178	0.007202	0.136226	0.139894
previous	-0.049005	0.043064	-0.084130	1.000000	-0.433281	-0.294511	-0.144376	-0.446634	-0.427960
emp.var.rate	0.100111	-0.075062	0.152834	-0.433281	1.000000	0.658360	0.268482	0.940643	0.947337
cons.price.idx	0.074719	-0.001042	0.094178	-0.294511	0.658360	1.000000	0.287710	0.491050	0.469138
cons.conf.idx	0.098290	-0.008388	0.007202	-0.144376	0.268482	0.287710	1.000000	0.287355	0.173493
euribor3m	0.114341	-0.080592	0.136226	-0.446634	0.940643	0.491050	0.287355	1.000000	0.924368
nr.employed	0.107867	-0.100543	0.139894	-0.427960	0.947337	0.469138	0.173493	0.924368	1.000000

Cuadro 23: Matriz de correlaciones (coeficiente de correlación de Pearson)

Con los datos recabados, observamos varias correlaciones moderadas a fuertes entre las variables económicas: emp.var.rate, euribor3m y nr.employed, con valores superiores a 0.9, lo que indica que estas variables están altamente relacionadas. También se observa una correlación moderada entre cons.price.idx y estas variables económicas. Por otro lado, variables como age, duration, campaign y previous presentan correlaciones bajas o incluso negativas con el resto, lo que sugiere independencia relativa respecto a las condiciones económicas. Esto sugiere que el comportamiento del cliente en campañas puede no estar fuertemente influenciado por estos indicadores macroeconómicos.

Diferencias clave

Característica	**Kendall τ**	**Spearman ρ**
Tipo de medida	Pares concordantes/discordantes	Rangos
Interpretación	Probabilidad de concordancia	Relación monótona
Robustez	Alta	Moderada
Velocidad de cálculo	Más lenta	Más rápida
Sensibilidad	Más conservadora	Más sensible

¿Cuándo usar cada una?

- **Usar Kendall** si los datos contienen muchos empates o si se desea una medida más robusta frente al ruido.
- **Usar Spearman** si se busca detectar una relación monótona sin necesidad de linealidad, con una interpretación más sencilla y cálculo más eficiente.

4.7. Analisis de regresión

Una vez analizadas las variables por separado y en conjunto, se obtuvo un análisis de regresión en nuestro conjunto de datos para poder encontrar patrones y tendencias.

Variable	Coeficiente
age	0.022785
duration	0.012926
campaign	0.035724
previous	0.201791
emp.var.rate	0.972692
cons.price.idx	0.832399
cons.conf.idx	0.701048
euribor3m	0.985268
nr.employed	0.974137

Cuadro 24: Primera iteración del análisis de regresión

Se decidió eliminar la variable nr.employed debido a su alta correlación con euribor3m y emp.var.rate. Esta redundancia generaba multicolinealidad, lo que afectaba la estabilidad e interpretabilidad del modelo. Además, euribor3m mostró un mayor coeficiente y capacidad explicativa, mientras que emp.var.rate captura mejor la dinámica del mercado laboral. La exclusión de nr.employed mejora la parsimonia del modelo sin pérdida de poder predictivo.

Variable	Coeficiente
age	0.012114
duration	0.014280
campaign	0.031565
previous	0.054815
cons.price.idx	0.071018
cons.conf.idx	0.022706

Cuadro 25: Primera iteración del análisis de regresión

Aunque la variable cons.price.idx mostró un coeficiente positivo en el modelo, se optó por excluirla en favor de un modelo más parsimonioso. La variable representa un factor macroeconómico que, aunque relacionado, tiene una influencia marginal en comparación con variables directamente vinculadas al comportamiento del cliente. Además, su posible colinealidad con otras variables como cons.conf.idx podría introducir redundancia. Su eliminación no afecta significativamente el poder predictivo del modelo y contribuye a una interpretación más clara.

Variable	Coeficiente
age	0.009915
duration	0.009491
campaign	0.014957
previous	0.042008
cons.conf.idx	0.043808

Cuadro 26: Primera iteración del análisis de regresión

En la versión final del modelo se eliminaron variables macroeconómicas altamente correlacionadas (euribor3m, emp.var.rate, nr.employed) y posteriormente cons.price.idx, con el objetivo de evitar multicolinealidad y mejorar la interpretación. Se conservaron cinco variables que combinan

factores sociodemográficos (age), conductuales (duration, campaign, previous) y un indicador macroeconómico clave (cons.conf.idx). Esto resultó en un modelo más parsimonioso y robusto, manteniendo el poder explicativo sin redundancia estadística.

5. Conclusión

El marketing bancario ha evolucionado significativamente, adaptándose a los cambios tecnológicos y a las nuevas expectativas de los clientes. A medida que el sector financiero se digitaliza, las instituciones deben replantear sus estrategias para mantenerse competitivas. Este marco teórico ofrece una base sólida para comprender cómo las teorías y estrategias de marketing se aplican en el contexto actual del sector bancario.