TRABAJO DE GRADO

Opción seminario-diplomado

MODELO DE PREDICCIÓN DE ABANDONO DE CLIENTES BANCO A PARTIR DE DATOS, UTILIZANDO ESTRATEGIAS DE MACHINE LEARNING.

Corporación universitaria Remington.

Facultad de ingeniería industrial

Programa académico: Ingeniería industrial.
Ingeniería de Sistemas

Estudiantes:

Luis Guillermo Chalacan L

Edinson Joaquín Muñoz Gutiérrez

Jeisson Portilla Rios.

Tutor: Juan Carlos Briñez de León.

Opción de trabajo de grado seminario-diplomado

Noviembre 2024.

Dedicatoria.

Antes que nada, dar gracias a DIOS por habernos dado la oportunidad de estudiar esta bonita carrera y poder formarnos como profesionales, a nuestras esposas e hijos que han estado en esta batalla académica y por quienes somos hoy.

Queremos destacar la contribución de trabajo en grupo lo cual fue de vital importancia para el desarrollo de este trabajo de opción de grado.

Dedicamos este nuevo logro obtenido en primer lugar a Dios quien es nuestro motor para seguir adelante día a día. De igual manera agradecerles a mis compañeros de trabajo y de manera lo dedicamos a nuestras familias, padres y hermanos que nos han brindado su apoyo incondicional al afrontar cada uno de los retos que la carrera así exigía. La culminación de esta etapa es el fruto del gran esfuerzo, dedicación, perseverancia y disciplina que tuvimos a lo largo de los cinco años transcurridos, en los que se vivieron momentos inolvidables colmados de grandes satisfacciones, culminando con la satisfacción de convertirnos en Ingeniero de sistemas – Ingenieros Industrial

Agradecimientos

Antes que nada, la gloria y la honra sea para DIOS, a la UNIVERSIDAD, A NUESTROS COMPAÑEROS, AL PROFESOR JUAN CARLOS POR SU DEDICACION Y COMPROMISO HACIA NOSOTROS, MERO TALENTO, y a todas aquellas personas que nos sirvieron para sumar conocimiento en este proceso de formación.

Quisiera ante todo agradecerle antes de nada, a la honra de Dios que nos permitió a la universidad a nuestros Profesores y compañeros de trabajos, a muestro profesor juan Carlos por su apoyo colaborativo en este proyecto sus esfuerzos y dedicaciones fueron fundamental para que hoy estemos aquí y que ha sido para enriquecer este nuevo proyecto de formación.

Tabla de contenido

INTRODUCIÓN.

La consecución de ese objetivo ha traído consigo un cambio en el tradicional proceso de compras en el que el comprador buscaba al proveedor y luego compraba, al proceso según el que el proveedor busca al cliente y luego lo mantiene. Esto se refiere al concepto de captación y fidelización del cliente. El abandono de clientes se ha convertido en un tema de gran interés para las empresas, por sus altos costos, que son difíciles de calcular. La salida de los clientes representa una pérdida segura de los ingresos futuros que derivarían de dichos clientes.

El problema no solo reside en que los clientes dejen de comprar un producto o pagar un servicio, sino en la incertidumbre con la que las empresas conviven día a día, ya que el saber cuáles clientes van a permanecer y cuáles no implicaría mejoras en los procesos administrativos y los esfuerzos enfocados en retener clientes. Esto causa que los departamentos de mercadotecnia de las empresas se orienten hacia lo referente al abandono del cliente a fin de lograr disminuirlo. Mediante esta información, orientarán los esfuerzos en retener a aquellos clientes que realizan más compras.

El objetivo principal de este trabajo es desarrollar un modelo de abandono de clientes que permita predecir el comportamiento eventual del mismo, con el objeto de mejorar los niveles de fidelización del cliente mediante el desarrollo de estrategias diferenciadas por segmentos de clientes que lleven a disminuir la tasa de abandono.

Resumen.

La fidelización de los clientes es un concepto fundamental para todas las empresas al momento de conocer la relación comercial que presentan con sus usuarios. Por ende, los bancos actualmente prestan mayor atención a su servicio posventa y buscan conocer las mejores estrategias que permitan obtener una alta tasa de retención de clientes. El presente trabajo de grado aborda el análisis de la fidelización y comportamiento de los clientes en el sector bancario, las alternativas de solución ante el problema de la falta de estrategias de fidelización y baja retención de clientes. Para ello, se ha realizado un análisis de enfoque a través de datos recopilados con revisión de fuentes de investigaciones académicas previas, relacionados a la fidelización de clientes, administración y retención de clientes, entrevistas internas realizadas a colaboradores de múltiples bancos, con el fin de obtener una base sólida para brindar una alternativa de solución adecuada según las necesidades de la empresa. Es gracias a ello, que se comprueba la importancia de la implementación del área de fidelización con personal interno de la compañía, realizando una convocatoria interna para formar el equipo que estará a cargo de generar datos con la ayuda de algoritmos y poder predecir por qué los clientes se están retirando de las entidades bancarias y así poder mejorar su retención.

Palabras clave. Fidelización, experiencia del cliente, clientes leales, retención de clientes

1. Marco conceptual y contextual

2. 1.1 Contexto:
3. 1.1.1 Sistemas de recomendación.

En las organizaciones, los procesos de transformación digital han permitido la consolidación de datos que registran las decisiones tomadas en su momento, y las variables asociadas a ellas. Hace 20 años la posibilidad de perder un cliente en la banca era baja. No era de vital importancia estudiar la tasa de abandono dado que no había, por parte de los clientes, una inclinación a cambiar de entidad. Hoy los clientes son cada vez más exigentes en cuanto a calidad percibida, precios, oferta de servicios e imagen de marca. Son libres de elegir y aquellos que no tienen una fuerte vinculación pueden cambiar de entidad fácilmente, casi con un solo clic. Un banco que sea capaz de predecir el abandono de sus clientes más valiosos podrá segmentarlos de tal manera que aquellos que tengan muchas posibilidades de abandonar estarán localizados. Se les podrá prestar un servicio más adecuado para que reconsideren su posible decisión de abandono. El objetivo principal de este trabajo es desarrollar un modelo predictivo de probabilidad de abandono. Mediante técnicas de aprendizaje automático de machine learning, basándose en datos históricos, el modelo buscará patrones que puedan identificar posibles fugas. El propósito es predecir el abandono y segmentar a los clientes de manera que se puedan priorizar y distribuir los recursos definiendo las acciones comerciales de retención oportunas.

El problema no solo reside en que los clientes dejen de comprar un producto o pagar un servicio, sino en la incertidumbre con la que las empresas conviven día a día, ya que el saber cuáles clientes van a permanecer y cuáles no implicaría mejoras en los procesos administrativos y los esfuerzos enfocados en retener clientes. Esto causa que los departamentos de mercadotecnia de las empresas se orienten hacia lo referente al abandono del cliente a fin de lograr disminuirlo. Mediante esta información, orientarán los esfuerzos en retener a aquellos clientes que realizan más compras.

Una tasa cercana a un 55% de abandono de clientes al término del primer año, implica pérdidas significativas para la compañía y altos costos de mantención, llegando incluso al término de relaciones comerciales con sus principales socios estratégicos. Con estos antecedentes, se hace cada vez más necesario estudiar el comportamiento de fuga de los clientes dada la alta competitividad del negocio y las nuevas regulaciones del mercado, obligando a actuar velozmente para lograr atraer a nuevos asegurados y sobre todo mantener a los actuales.

1.1.2 Algoritmos de Machine learning en sistemas de recomendación.

En el proceso de recopilación de los datos, el nivel de calidad y el grado de la anotación semántica son variables; y según desarrolla la actividad, los clientes tienen o no continuidad en su comportamiento. Como respuesta a esta realidad, se detecta la necesidad de proveer a los clientes de mecanismos de ayuda que permitan reconocer, encontrar, seguir y disfrutar. En la actualidad, todavía resulta complicado para los usuarios encontrar aquellas informaciones que buscan rápidamente entre los datos disponibles, dado que es complicado distinguir el grano de la paja. Los llamados sistemas de recomendación utilizan algoritmos que rastrean enormes bases de datos para presentar sugerencias a los usuarios. Estos algoritmos están basados usualmente en inteligencia artificial, más concretamente en técnicas de aprendizaje automático (machine learning).

1.2 Pregunta problema:

¿Qué factores son más influyentes en la decisión de un cliente de abandonar el banco haciendo uso del machine learning a través de datos computacionales?

1.4 Hipótesis:

El análisis computacional de los datos sobre el abandono de clientes a bancos, acompañado de algoritmo de aprendizaje automatizado y algoritmos de dataframe, permitirá implementar un sistema de estrategias que permita garantizar la conformidad y continuidad a los clientes en el banco, con miras a la personalización y mejoramiento del servicio.

Objetivos.

Avaluar ,analizar y revisar todos estos procesos de desarrollo de los clientes y el desarrollo de un sistema de software , como un sistema de gestión de calidad y gestión de riesgo con los que cuente la empresa , los cuales buscan un modelo de desarrollo que sea capas de tener tareas especificas ,como clasificar ,predecir o detectar ,o resolver un problema o mejorar un proceso de la empresa , en donde evalúa diferentes modelos de ML (como modelos supervisados y no supervisados) y supervisar la selección de unos o varios modelos específicos basados en su característica y el problema a resolver en donde recolecte ,limpie y procesar un conjunto de datos adecuado que sea representativo del problema a resolver , en donde aplique técnicas de manejo de datos faltantes ,normalizando y selección de carasteristicas.este desarrollo del sistema se clasifica basado en machine learming para una detección temprana que los bancos tengan que utilizar un algoritmo de aprendizaje supervisado para ayudar a los clientes a tener un mejer finanzas y tasas justas que se ajusten a su comodidad y precisión de las evaluaciones financieras.

3 Desarrollo e implementación del aprendizaje.

```
#Para cargar los datos
import pandas as pd
from google.colab import files
uploaded = files. Upload()
for filename in uploaded.keys():
   Conjunto_Datos = pd.read_csv(filename,sep=',')
Conjunto_Datos.head()
```

Elegir archivos Churn.csv

- **Churn.csv**(text/csv) - 709705 bytes, last modified: 31/10/2024 - 100% done

Saving Churn.csv to Churn (2).csv

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2.0	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1.0	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8.0	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1.0	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2.0	125510.82	1	1	1	79084.10	0

Figura 1

Esta imagen muestra el estrato de un archivo de datos de clientes , probablemente para un análisis de churn o abandono del cliente en una institución financiera en donde muestra algunos puntos clave de los datos , 1.columnas , customenrid , identificador único del cliente ; (CrediScore) puntaje crediticio del cliente ,geografía , país de residencia del

cliente ,genero , genero del cliente ,edad del cliente ,tiempo de permanencia en la institución ,(en años) saldo de la cuenta del cliente , numero de productos adquiridos por el cliente . si el cliente tiene tarjeta de crédito (para así o para no) si el cliente es miembro activo (1 para si o para no) salario estimado del cliente; (salido) si el cliente a abandonado (1 para si o para no), se espera que se obtenga una salida que muestre las primeras cinco filas de este Dataframe en donde nos dará la idea de cómo está estructurado los datos.

```
[72] #Información de la estructura de datos
     Conjunto_Datos.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           9091 non-null   float64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(3), int64(8), object(3)
memory usage: 1.1+ MB
```

Figura 2

Esta figura de salida ha proporcionado es que el resultado ejecutado en el conjunto de datos es un dataframe lo cual nos brinda un resumen conciso sobre la estructura de los datos:

El índice total de la entrada (Rangelndex 10000 entries 0 a 9999 esto indica que este dataframe tiene un total de 10,000 filas numeradas desde 0 a 9999

Un total de 14 columnas en este dataframe en donde cada columna demuestra los detalles.

Nombre de las columnas

Non-Null Count el número de valores no nulos en cada columna

Dtype el tipo de datos de cada columna

Detalles de las columnas RowNumber, Customerid, Surname, CrediScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsactiveMember, EstimatedSalary, Exited: las cuales todas estas variables tienen 10,000 entradas de tipo float64 unas de numero entero y otra en decimal, la mayoría de las columnas tienen 10,000 valores no nulos lo que significa que no faltan datos en esas columnas.

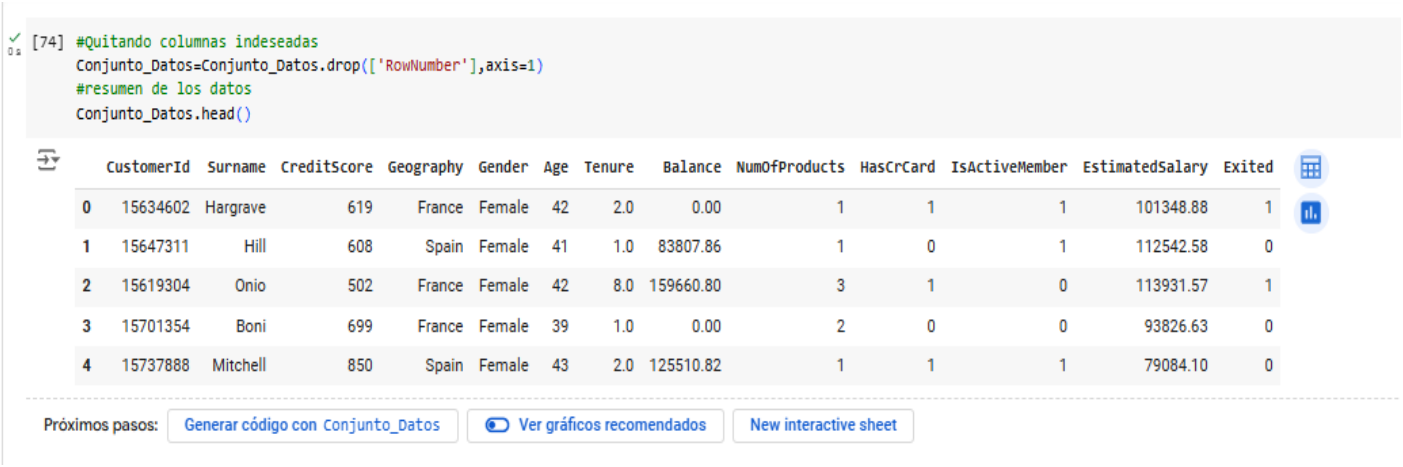
```
#Análisis de los datos
Conjunto_Datos.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	9091.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	4.997690	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.894723	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	2.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Figura 3

Los resultados obtenidos en la figura 3 son estadísticas descriptivas generadas por un conjunto de datos que proporcionan un resumen numérico de las columnas numéricas en el dataFrame, __ Count 10,000 (se espera ya que es un total de filas), Mean 5000,5 lo que indica que es un índice que varía desde 1 hasta 10,000 Atd 2886.9, lo que muestra muestra variabilidad en el índice. Customerld indica que todos los IDs estan presente y también representa el promedio de los identificadores de clientes ,CrediScore surgiere el puntaje de crédito promedio de los clientes que estén dentro de un rango considerado bueno, lo que indica que haya una variabilidad moderada en los puntajes de crédito , los valores van desde 350 hasta 850 que es un rango común para los puntajes de créditos, Age lo que indica que la edad promedio de los clientes es de aproximadamente de 39 años lo que indica una diversidad de edades en el conjunto, las edades oscilan entre 18 y 92 años, lo que surgiere que el conjunto de datos abarca una amplia gama de edades Tenure: indica que hay 909 valores nulos lo que es significativo y debe ser tratado ,también surgiere que los clientes has estado con la empresa durante el promedio de 5 años lo que depuesta que algunos clientes ha estado más tiempo que otros a un rango de 0na 10 años , Balance lo que indica que el saldo promedio de los clientes es alto lo que surgiere que haya una gran variabilidad en los saldos que van desde 0 hasta 250,989,09 lo que muestra que hay clientes con saldos muy bajos y otros con saldos muy altos, NumOfProducts: indica que el promedio ,los clientes tienen alrededor de 1,5 productos con la empresa lo que surgiere que la mayoría de los clientes tienen de 1 a 2 productos , con algunos pocos que tienen más ,HasCrCardindica que aproximadamente el 71% de loa clientes tiene una tarjeta de crédito, lo cual muestra una variabilidad normal en la tenencia de tarjetas todos los valores son de 0 a 1 (no tienen o tienen tarjeta) ,IsaactiveMember: lo que surgiere aproximadamente la mitad de los clientes son miembros activos ,lo que refleja la variabilidad en el estatus de membresía y que todos los valores son de 0 a 1EstimatedSalary:indica que el salario está estimado en un

promedio de los clientes es de bastante alto e indica variabilidad significativa en los ingresos estimados de 11,580 hasta 199,992,48 mostrando una amplia gama de salarios, Exited indica que aproximadamente el 20% de los clientes has abandonado el servicio lo que refleja la variabilidad en la tasa de abandonos ,todos los valores son de 0 a 1 representado en el estado de los clientes si han abandonado o no , esto indica que la mayoría de los clientes jóvenes están en la categoría de edad media alrededor de los 39 años.



```
[74] #Quitando columnas indeseadas
     Conjunto_Datos=Conjunto_Datos.drop(['RowNumber'],axis=1)
     #resumen de los datos
     Conjunto_Datos.head()
```

	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	15634602	Hargrave	619	France	Female	42	2.0	0.00	1	1	1	101348.88	1
1	15647311	Hill	608	Spain	Female	41	1.0	83807.86	1	0	1	112542.58	0
2	15619304	Onio	502	France	Female	42	8.0	159660.80	3	1	0	113931.57	1
3	15701354	Boni	699	France	Female	39	1.0	0.00	2	0	0	93826.63	0
4	15737888	Mitchell	850	Spain	Female	43	2.0	125510.82	1	1	1	79084.10	0

Próximos pasos: Generar código con Conjunto_Datos | Ver gráficos recomendados | New interactive sheet

Figura 4

Esta figura muestra que después de haber eliminado la columna del Dataframe en el conjunto de datos en donde tiene un conjunto de datos más limpio y enfocado en la variable que son relevantes para el análisis después de haber eliminado la columna indeseada aparecen los registros en la siguientes Dataframe como nos muestra la imagen una vez cargado los datos

```
[75] # Realizando tabla de frecuencia según el numero de rangos
     numero_de_rangos = 5 #el número de rangos es aproximadamente la raiz cuadrada del número de muestras

     # Calcular la tabla de frecuencia absoluta con bins
     Tabla_frecuencia = pd.cut(Conjunto_Datos['Age'], bins=numero_de_rangos, include_lowest=True, right=False).value_counts().reset_index()
     Tabla_frecuencia.columns = ['Intervalo', 'Frecuencia Absoluta']

     # Calcular la frecuencia acumulada
     Tabla_frecuencia['Frecuencia Acumulada'] = Tabla_frecuencia['Frecuencia Absoluta'].cumsum()

     # Mostrar la tabla de frecuencia
     print("Tabla de Frecuencia:")
     print(Tabla_frecuencia)
```

```
Tabla de Frecuencia:
        Intervalo  Frecuencia Absoluta  Frecuencia Acumulada
0    [32.8, 47.6)                 5500                  5500
1    [18.0, 32.8)                 2790                  8290
2    [47.6, 62.4)                 1351                  9641
3    [62.4, 77.2)                  335                  9976
4  [77.2, 92.074)                   24                 10000
```

Figura 5

Esta tabla nos muestra la frecuencia obtenida donde divide la variable Age en 5 intervalos, calculando para cubrir todo el rango de edades en el conjunto de datos, en donde la edad promedio esta desde los 18 a 77 años de edad intervalo 32 tiene una mayor frecuencia absoluta con 5,500 clientes lo que representa la mayoría del conjunto de datos, frecuencia acumulada después de este intervalo es de 5,500 lo que surgiere que más de la mitad de los clientes tiene edad en este rango.

Intervalo 18 es el intervalo más poblado con 2,790 clientes al sumar la frecuencia de los dos primeros intervalos la frecuencia acumulada es de 8,290 lo que significa que aproximadamente el 82.9% del cliente tiene edades, menores de 47 años

Intervalos 47 disminuye a 1,351 clientes lo que indica que hay un número menor cantidad de personas estén en el rango de edad

Intervalo 62 solo las 335 clientes estan en este rango lo que demuestra una caída significativa en la frecuencia

Intervalo 77 este intervalo con menor frecuencia con solo 24 clientes la frecuencia acumulada al final es de 10,000 que es el total de la muestra

En general la mayoría de los clientes tienen edad comprometida entre los 32 y 47 años lo que indica que la base de datos esta predominantemente compuesta por clientes de mediana edad.

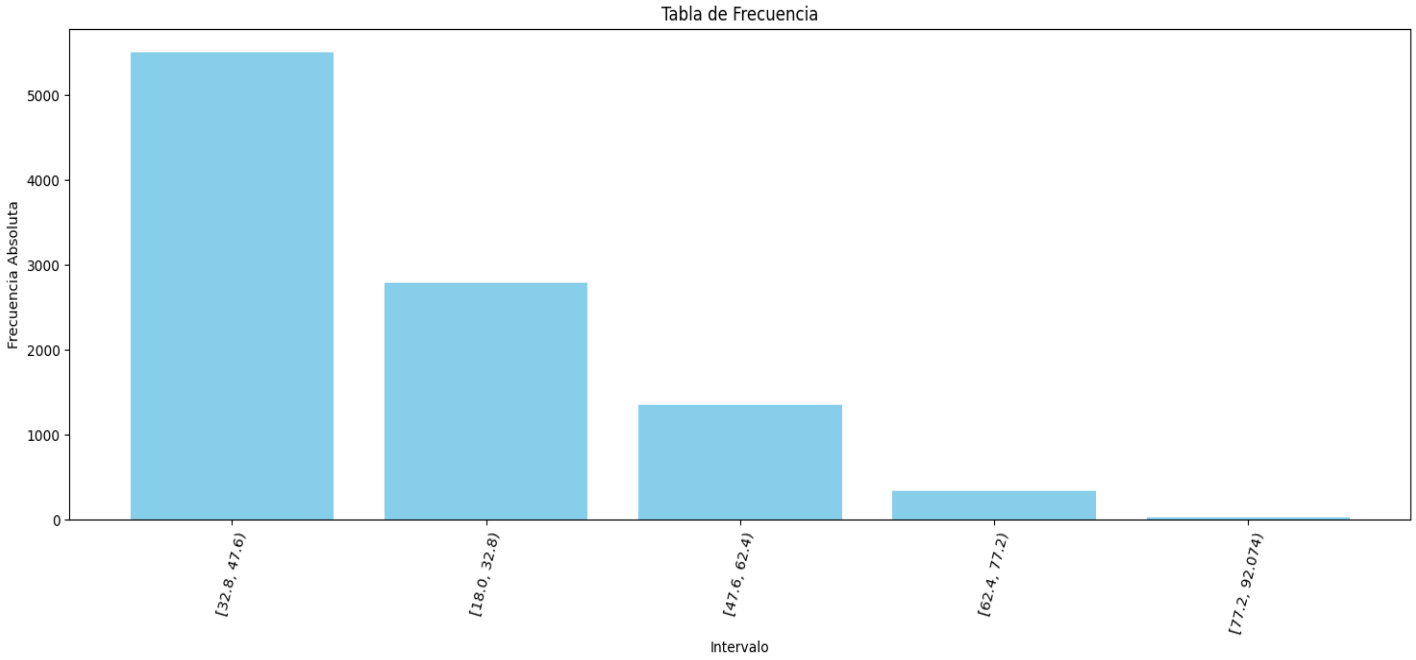


Figura 6

Esta imagen de barra que representa la tabla de frecuencia en la variable edad muestra una distribución de los datos en intervalos específicos, veremos el análisis detallado de los resultados obtenidos a partir de esta visualización

El eje x representa los intervalos en los que se ha dividido la variable como el rango de edades ,cada barra representa un intervalo y su altura indica la cantidad de datos que caen dentro del rango la cantidad de intervalos son 5 lo cual el promedio máximo de edad esta entre los 32 a 47 años y la mínima esta entre los 77 a 92 años , este eje muestra la frecuencia absoluta , es decir el número de observaciones en cada intervalo esta barra vemos que los intervalos más alto indica que los intervalos con más datos , lo que refleja los grupos mas representados en el conjunto de datos, si la distribución es uniforme las barras tendrá una altura similar ,si hay intervalos claramente domínate esto se reflejara en una barra significativamente más alta ,si la barra disminuye gradualmente hacia un lado ,la distribución puede estar segada si la mayoría de los datos se encuentra en uno o dos lados intervalo se observa una concentración clara en cierto rangos. Si vemos que la barra esta alta se encuentra en el intervalo de 30 a 40 años significa que la mayor parte de los datos de edad se concentra en ese rango.

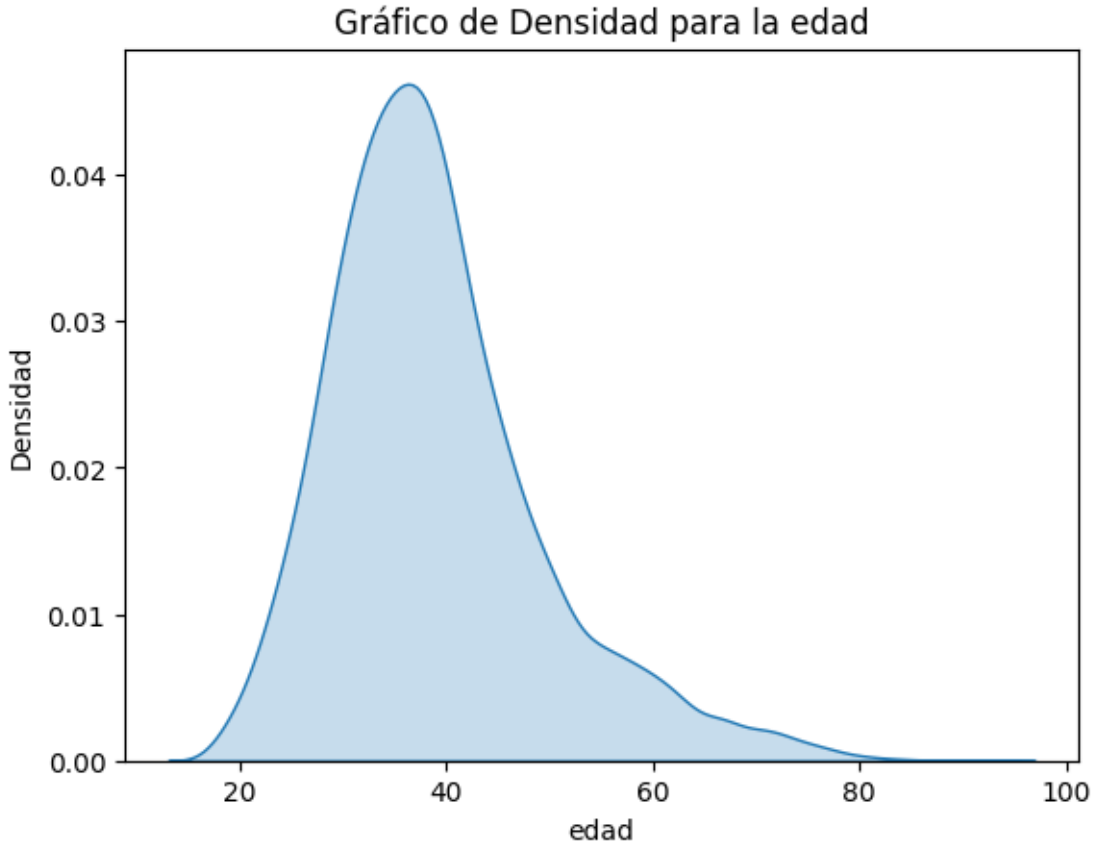


Figura 7

el grafico densidad para esta variable Age muestras una distribución de edades de los clientes en conjunto de datos ,donde este tipo de datos es útil para observar la forma de la distribución y obtener información sobre la concentración y dispersión de las edades ,podemos observar que si la curva es simétrica vemos que la probabilidad de edad esta dentro de los 18 a 80 años ,si esta curva tiene mas de un pico la distribución es multimodal lo que surgiere la presencia de varios grupos de edades predominantes en los datos ,la altura del pico muestra la concentración de los datos .un pico alto indica que muchas observaciones tienen edades similares. Sesgo – si esta cola de la distribución se extiende mas hacia la derecha o izquierda la distribución esta sesgada un sesgo a la derecha (positivo) implica una mayor cantidad de edades más jóvenes que un sesgo a la izquierda (negativa indica una mayor cantidad de edades mayores , se puede observar en el eje x si el pico mas alto se encuentra entre los 30 y 40 años esto aplica que la mayor parte de los clientes tiene esa edad , estos aplica que los clientes se encuentra en una etapa productiva de la vida.

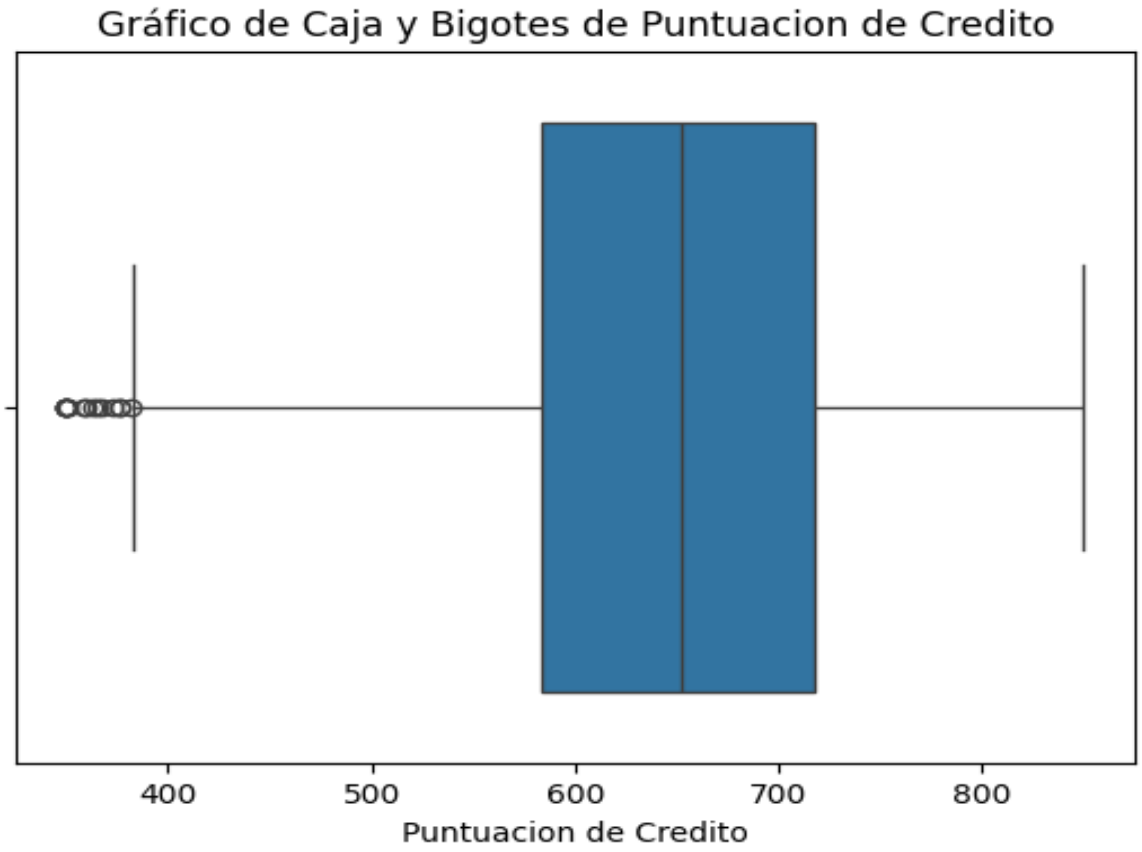


Figura 8

Esta imagen de grafica de la caja de bigote de la variable CrediScore muestra una distribución de la puntuación de créditos de los clientes en donde este grafico es útil para identificar la mediana ,los cuartiles los rangos y los valores atípicos de la puntuación de créditos en el conjunto de datos lo cual muestra un valor de crédito entre la puntuación esta entre 580 a 730 puntos , en donde cada línea representa una mediana puntuación de créditos y que las parte inferíos y superior de la caja representa un 50% de la puntuación de crédito donde se encuentra dentro de este rango , el rango Q1 y Q3 representa la dispersión de la mayoría de los datos , si la mediana esta entre 580 es que común en muchos análisis de créditos esto surgiere que la mayoría de los clientes tiene una puntuación alrededor de ese valor, este análisis ayuda a la empresa a identificar los grupos de clientes con diferentes niveles de solvencia y ajustar sus estrategias de créditos y riesgo en consecuencia.

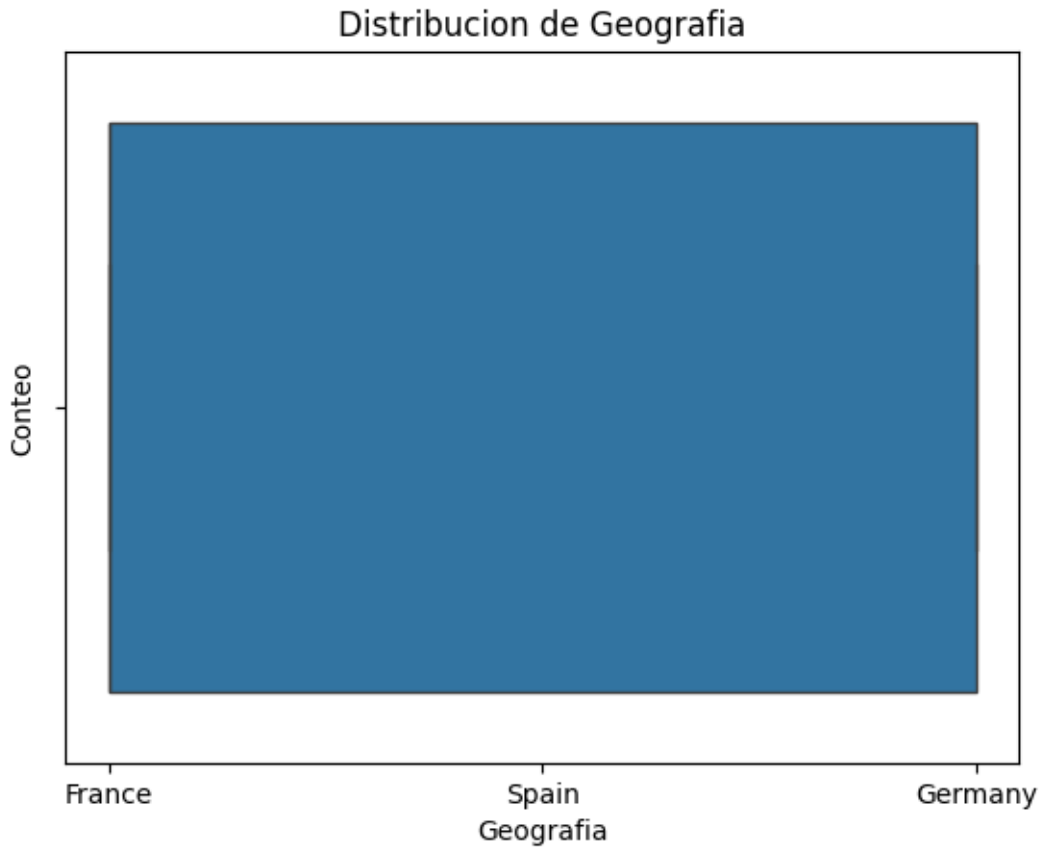


Figura 9

Esta imagen muestra que el grafico mostrara un número de clientes por regiones en cada región cuantos abran, si una región tiene un número significativo mayor de clientes esto puede indicar que la empresa tiene una mayor presencia o popularidad en esta área , conocer la distribución de clientes por regiones puede ayudar a que la empresa ajustes sus estrategias , en donde si una región tiene pocos clientes se podrían explorar las razones y considerar estrategias para aumentar la presencia en esa zona, de esta forma la mejor forma de visualizar los datos geográficos es mediante graficas de barras que de muestres la cantidad de regiones en cada categoría en donde proporcionara una visión clara de cuantos clientes provienen de cada región o país.

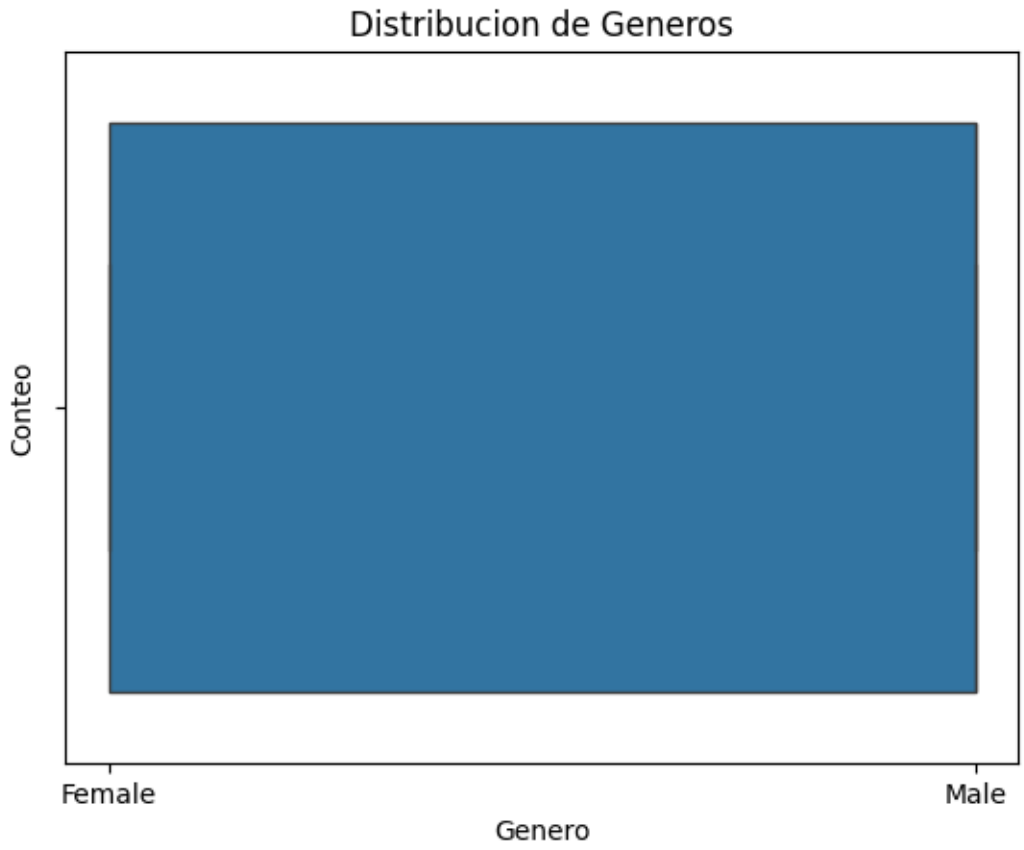


Figura 10

Esta grafica muestras que la distribución de esta caja y bigote muestra el análisis grafico de conteo , lo cual indica los registros que hay de cada genero de cuantos hombre y mujeres hay en el conjunto de los datos , esto permite que si hay un equilibrio en la base de datos o si un género esta sobre representado, esto equivale que el conteo muestra un número similar de hombre y mujeres sonde se puede inferir que la muestra esta equilibrada, o si vemos que un género es más predominante ,puede ser un indicativo de un sesgo en la muestra o refleja la composición del público objetivo.

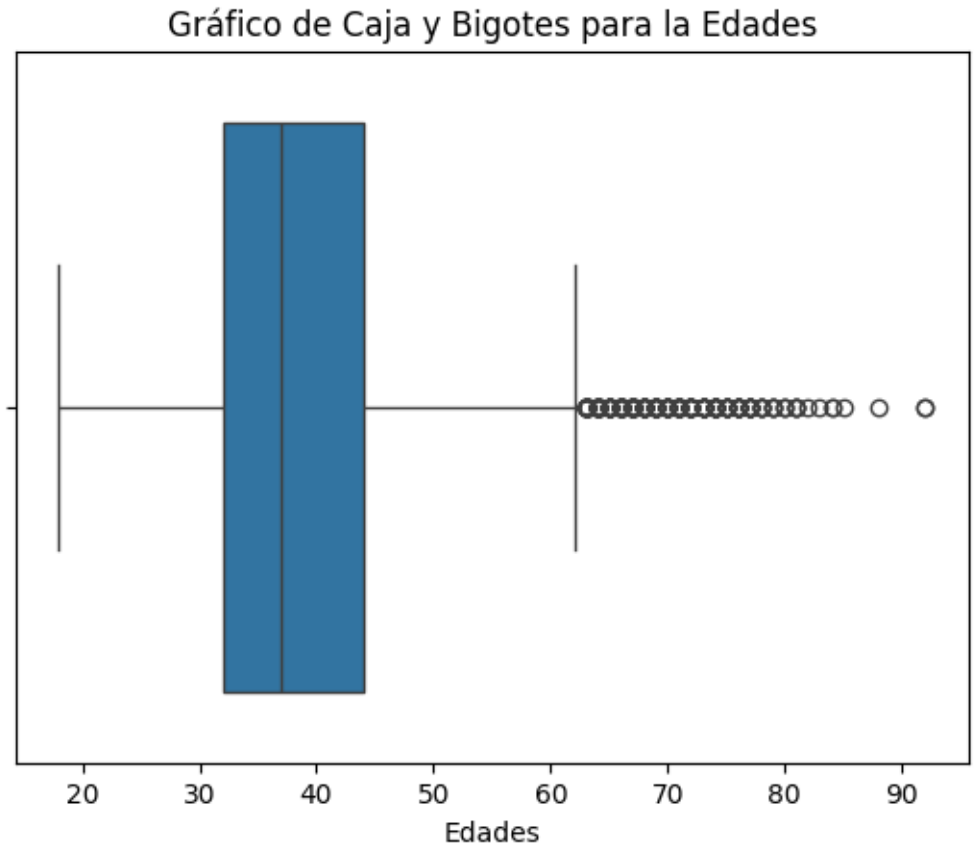


Figura 11

Esta imagen proporciona una representación visual de la distribución de las edades de los clientes en el conjunto de datos, donde muestras la línea central de la caja de bigote que la edades usan entre los 32 a 44 años de edad esto equivale que el 50% de los clientes es más joven el otro 50% es más viejo, si la mayoría de los clientes tiene edades dentro de un rango definido ,(entre 32 a 44 años) esto podría indicar que la base de los clientes está compuesta principalmente por adultos por edad de trabajo, la presencia de los jóvenes o mayores podría señalar un pequeño número de clientes que no se encuentra dentro del rango típico de edades, este análisis de las edades de los clientes proporciona una visión valiosa sobre la composición demográfica de la base de clientes lo que es útil para la toma de decisiones en ventas.

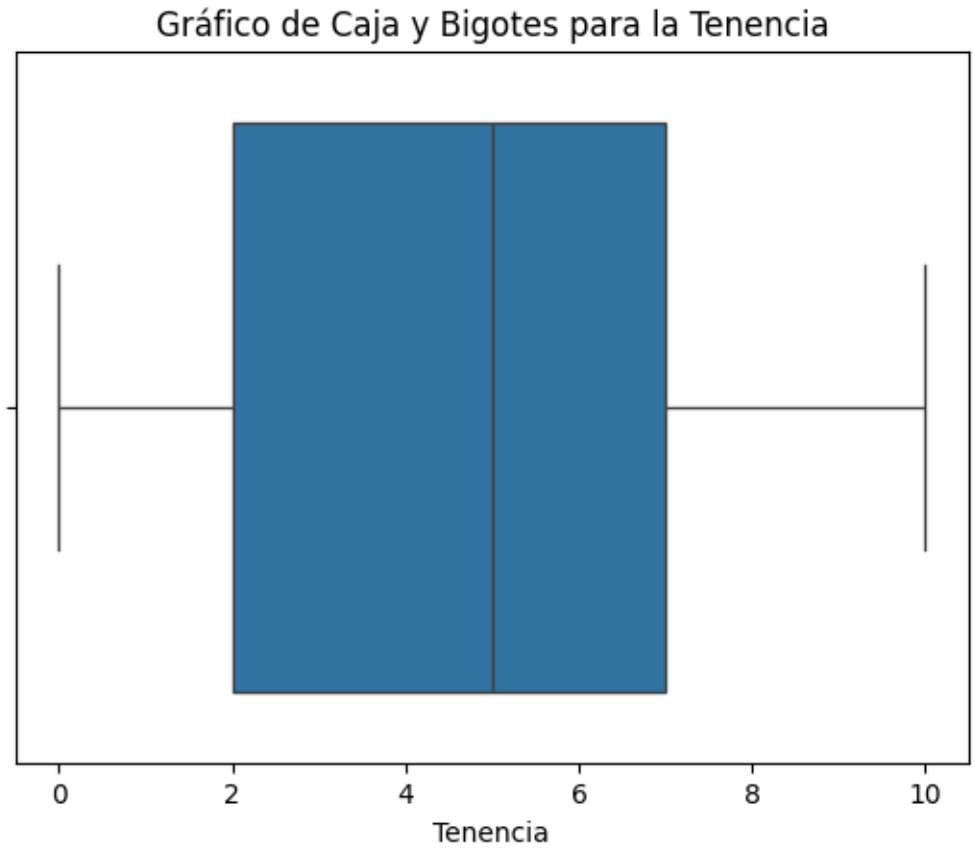


Figura 12

Esta imagen nos muestras que la caja y bigote para la variable Tenure (tenencia) muestras una distribución de la cantidad de años que los clientes han permanecido en la empresa lo cual muestra los años que han permanecido esta entre el rango de la línea central que representa la mediana variable es de 2 a 7 años mostrando la tenencia que tiene el cliente en promedio

Los bordes de la caja indican el primer cuartil Q1 y el tercer cuartil Q3 el 50% de los clientes tiene un tiempo de tenencia que se encuentra dentro del rango

Una caja más estrecha significa que la mayoría de los clientes tienen una tenencia similar, mientras que la caja mas amplia indica una mayor variabilidad.

La longitud de los bigotes puede indicar la presencia de clientes con tenencias muy altas o muy bajas en comparación con el resto, si la mediana se encuentra centrada en la caja, la tenencia está distribuida de manera más simétrica .si esta un extremo hay un sesgo en la distribución.

Otro extremo muestras si la gráfica muestra una concentración de valores bajos de tenencia, amplia que hay muchos clientes nuevos, si los valores son altos puede haber mas clientes con una relación a largo plazo con la empresa. Esto representa que la distribución sugiere que la tenencia promedio es representativa mientras que una distribución sesgada podría

indicar un grupo de clientes predominante, (como una mayoría de clientes nuevos o muy leales).

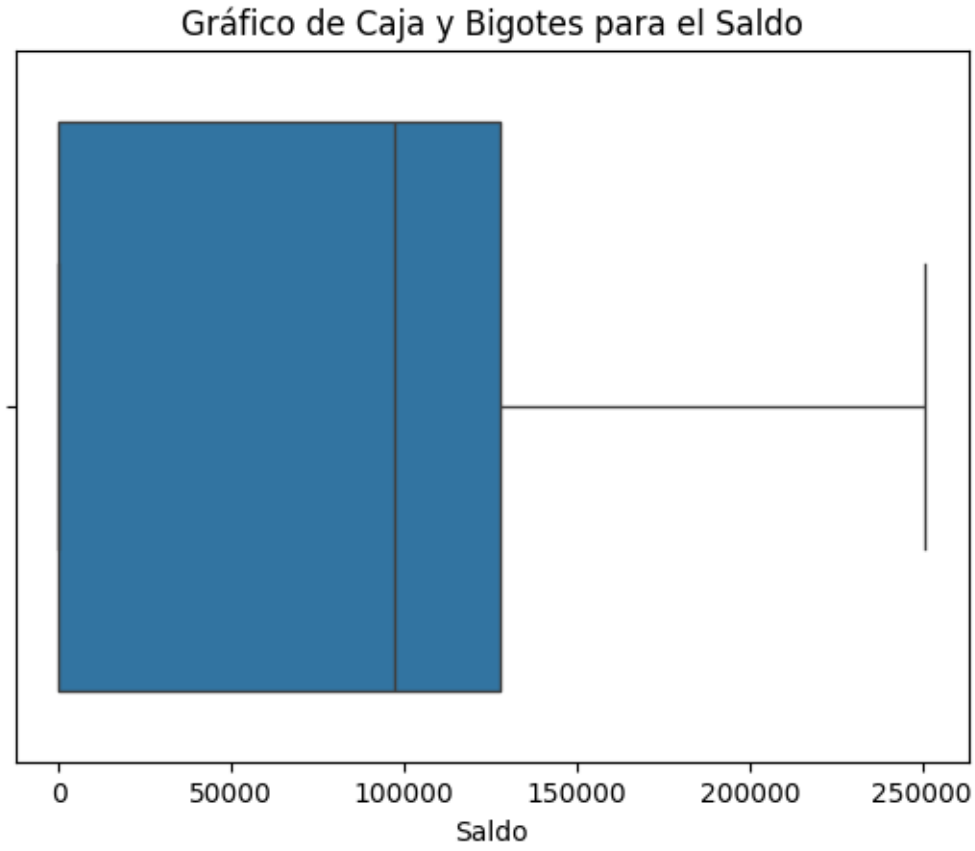


Figura 13

Esta grafica de la caja de bigote representa una visual de la distribución del saldo de los clientes en el conjunto de datos veremos el resultado detallado

En donde la línea central de la caja representa la mediana del saldo de los clientes es decir el valor en el que el 50 % de los clientes tiene un saldo menor y el otro 50% tiene un saldo mayor el cual esta reflejado entre los 50000 dólares a 140000 dólares en donde los bigotes se extienden desde la caja y muestra el tango de datos que nos e consideran atípicos los valores que caen fuera de estos bigotes son posibles , los bigotes que represente los valores atípicos en el caso del saldo pueden ser clientes con saldo significativamente más altos o bajos que el promedio , si esta medida está cerca del centro de la caja la distribución es relativamente simétrica es la mediana se desplaza hacia uno de los lados la distribución es sesgada , un gráfico asimétrico hacia la derecha indica que hay un mayor cantidad de clientes con saldos bajos muestras que la asimétrica hacia la izquierda mostrario lo contrario.

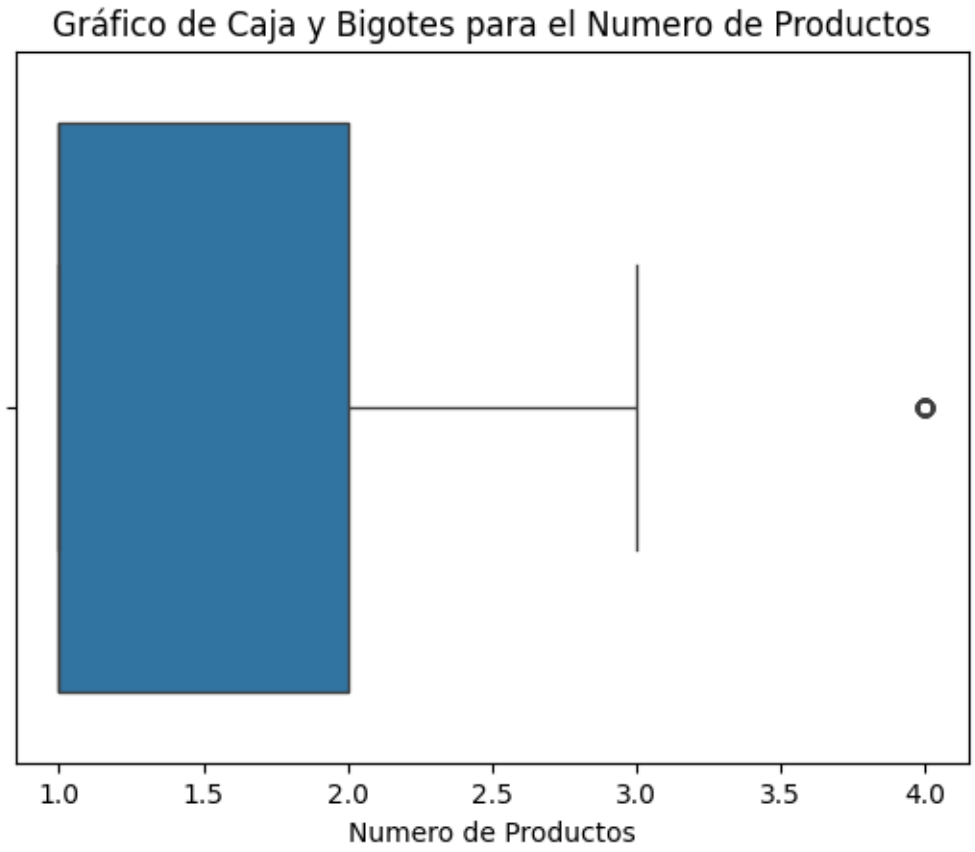


Figura 14

Este grafico de la caja de bigote muestra como están distribuidos los datos sobre la cantidad de productos que los clientes tienen esta variable es numérica y discreta ya que representa un recuento de los productos.

Tiene una línea central en la caja indica la mediana del número de productos que los clientes tienen esto muestra un valor central de la distribución dividiendo al conjunto de datos en dos mitades, el primer cuartil Q1 y el tercer cuartil Q3 delimitan los bordes de la caja encerrando el 50% de los datos centrales. Si la caja estrecha significa que la mayoría de los clientes tiene un numero de productos similar, una caja más amplia indica una mayor variabilidad en la cantidad de productos, los bigotes se extiendes desde la caja hasta los datos que no se consideran atípicos el rango de estos bigotes proporciona una idea del rango general de la cantidad de productos. Los puntos fuera de los bigotes se consideran outlier esto podría representar clientes que no tienen una cantidad inusualmente baja o alta de productos. Es probable que la mayoría de los datos se concentren en un pequeño de productos ya que los clientes típicos suelen manejar de uno a dos productos.

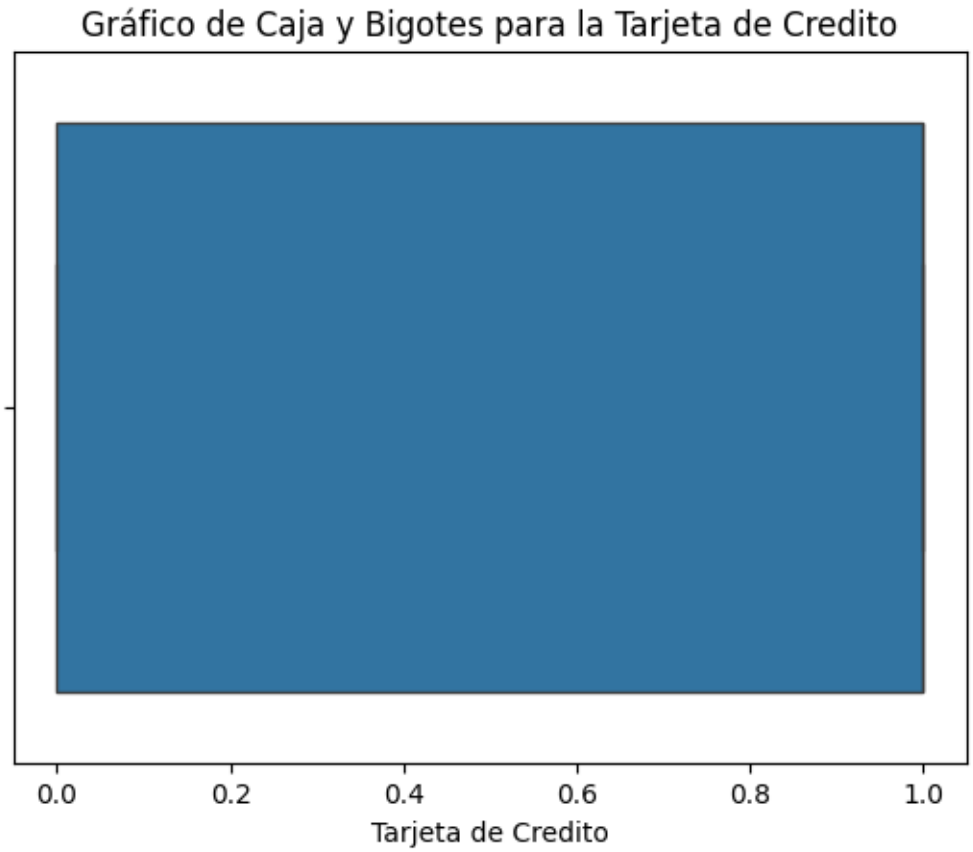


Figura 15

Esta figura nos muestra la variable tarjeta de crédito donde muestra la distribución de los datos sobre si el cliente tiene una tarjeta de crédito asociada.

Este grafico representa que eje x con dos categorías:

o clientes que no tienen tarjeta de crédito

1 cliente que si tiene tarjeta de crédito

Esta datos que la variables continuas en lugar de eso se podría observar cómo los valores de la variable están distribuidos entre dos categorías , esto mostrara si los datos están concentrados en un punto especifico , dependiente de esta grafica podemos ver si hay una diferencia notable en la cantidad de clientes que tienen tarjetas de crédito en comparación con los que no tienen esto indica que cualquier punto fuera de la caja de bigote representa un valor atípico aunque en variables binarias esto es menos común. , es probable que este grafico este dividido en dos puntos sin cajas extensas.

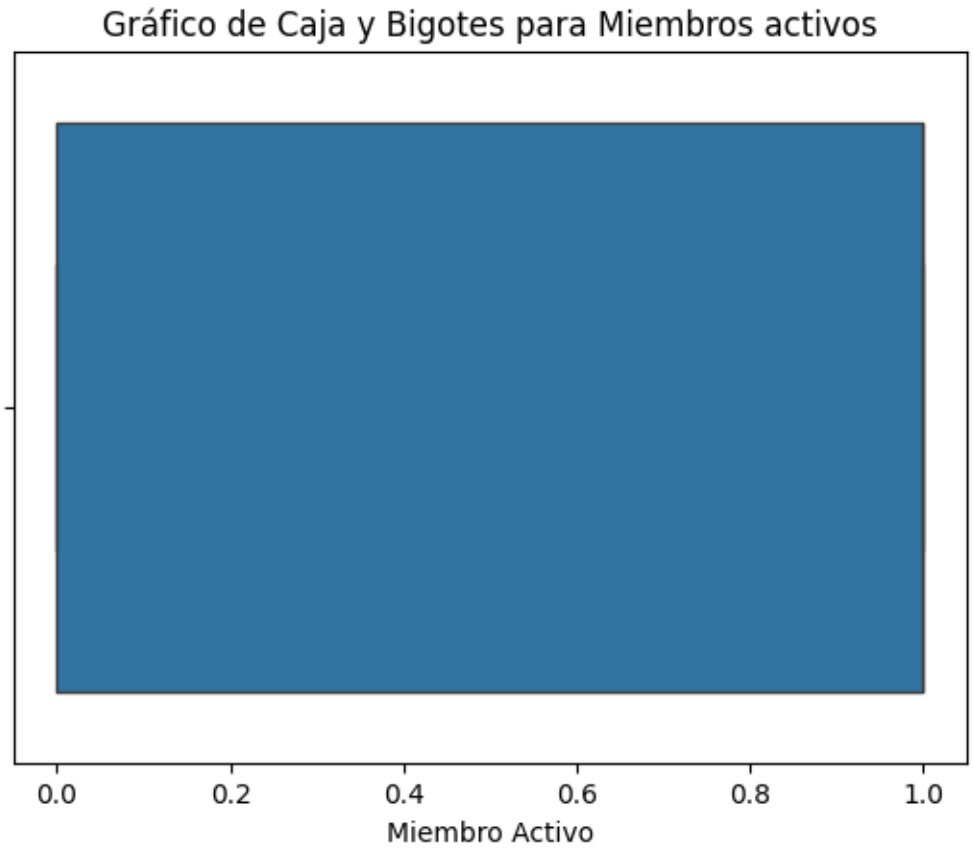


Figura 16

Esta figura muestra el análisis de esta variable de miembros activos los cuales es el conjunto de datos es importante que esta variable binaria en donde esta representa una categoría discreta en los valores continuos en donde nos muestra los miembros que están activos y mostrara a los clientes que no están activos en donde esta será útil para entender la distribución de los miembros en ambas categorías.

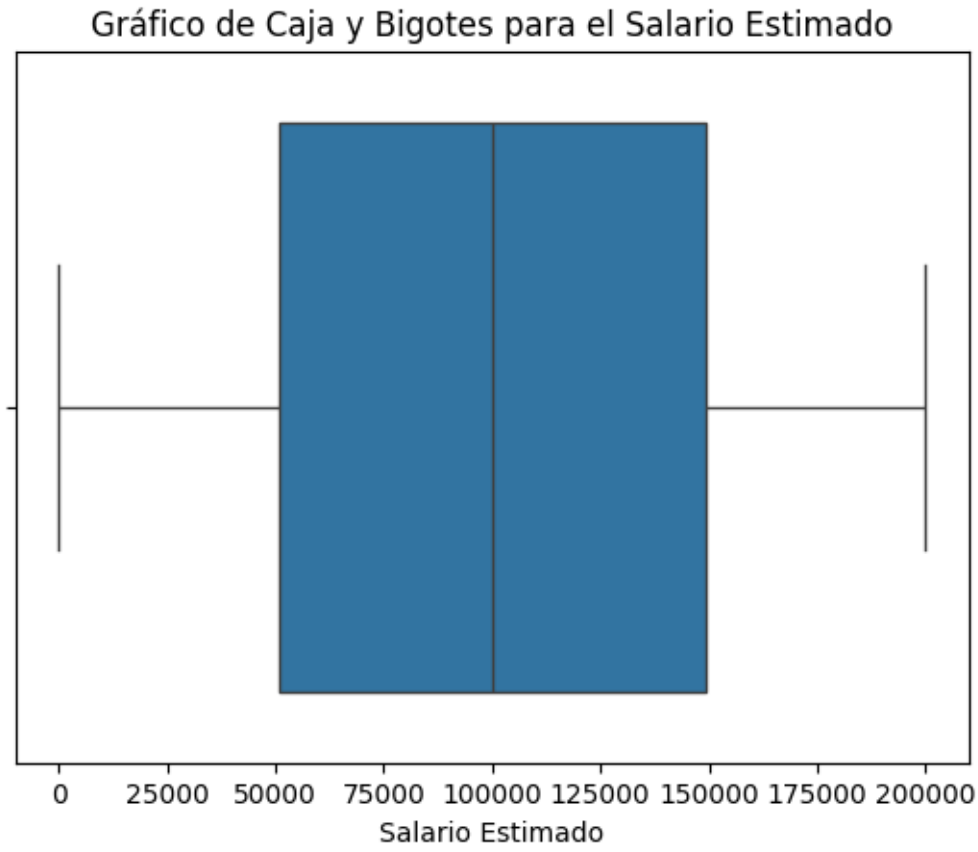


Figura 17

Esta grafica de la caja de Bigote para el salario estimado nos muestra una distribución estimada del salario de los clientes en el conjunto de datos en donde algunos puntos clave de interpretación en el rango intercuartílico que es la distancia entre el primer cuartil Q1 y el tercer cuartil Q3 en donde el 50 % de los datos se encuentra dentro de esta caja donde muestra el central el valor de los salarios estimados que son entre 50000 dólares a 150000 dólares estimados.

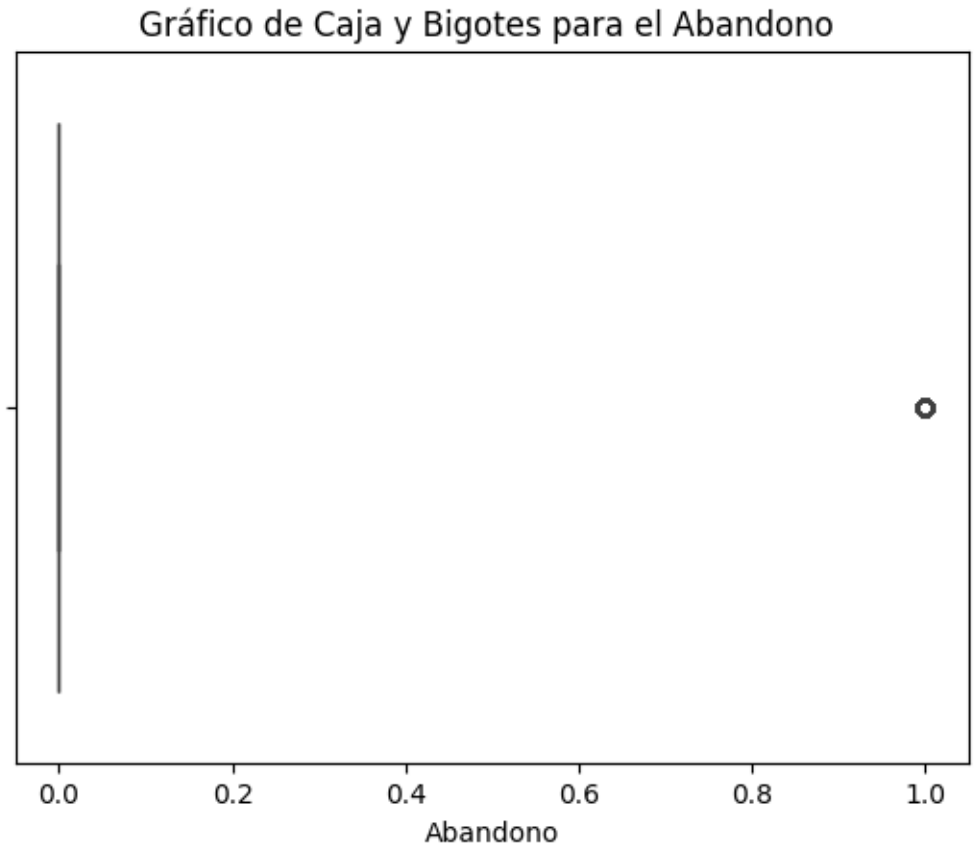


Figura 18

Esta grafica binaria de (0 a 1) indica que si un cliente abandono lo cual nos indica el resultado si los resultados obtenidos mediante el crédito y el balance del salario que es estimado por el cliente fueron abandonados, o si no vemos que esta grafica muestra un valor significativo sin valor alguno en donde muestra dos cajas que representan 0 a 1 pero sin ningún valor numérico que permita ver la dispersión, cuartiles o mediana

```
#Análisis de medidas estadísticas
from scipy.stats import describe, skew, kurtosis, scoreatpercentile

stats = describe(Conjunto_Datos['Age'])

asimetria = skew(Conjunto_Datos['Age'])
print('La asimetría está en: ',asimetria)

curt = kurtosis(Conjunto_Datos['Age'])
print('la curtosis es: ',curt)


# Calcular percentiles
percentiles = [25, 50, 75]
valores_percentiles = [scoreatpercentile(Conjunto_Datos['Age'], p) for p in
percentiles]
print('Los percentiles 25, 50 y 75% son: ',valores_percentiles)
```

Figura 29

Esta imagen de análisis estadístico de la variable edad utilizando la librería scipy.stats proporciona información clave sobre la distribución y las características de los datos, analizaremos la medida estadística calculadas.

Esta función describe la proporcionalidad de un resumen de la estadística básica que incluye el número de observaciones en total de entrada en el conjunto de datos , promedio de edades , desviación estándar ,mide la dispersión de las edades a la media ,como la mínima y máximas rango de edades , esta asimetría mide la distribución del valor positivo la cual indica que la cola derecha es más larga o está más dispersa que la izquierda sesgo positivo ,el valor negativo indica que la cola izquierda es más larga o está más dispersa que la derecha sesgo negativo ,esta distribución es aproximadamente simétrica, los percentiles indican el valor bajo el cual se encuentra un porcentaje de los datos

Porcentaje 25% de los datos esta por debajo de este valor (primer cuartil)

Porcentaje 50% de los datos estan por debajo (mediana)

Porcentaje 75 75% de los datos está por debajo (tercer cuartil)

Si los valores son 30(25%) ,40(50%), y 50(75%) significa que es el 25% de los clientes tiene 30 años o menos la mitad tiene 40 años o menos y el 75% tiene 50 años o menos, estos valores de asimetría y curtosis ayudan a entender la forma de la distribución de las edades esto puede tener implicaciones en la segmentación de clientes y estrategias.

Los percentiles permiten identificar grupos específicos dentro de la población y puede guiar en la personalización de servicios y productos, este análisis nos ofrece una comprensión profunda sobre esta variable edad la cual es crucial para la toma de decisiones basadas en los datos demográficos de los clientes.

Modelo de predicción de abandono de clientes banco

Los clientes de Beta Bank se están yendo, cada mes, poco a poco. Los banqueros descubrieron que es más barato salvar a los clientes existentes que atraer nuevos. Necesitamos predecir si un cliente dejará el banco pronto. Crearemos un modelo de clasificación.

Hemos tomado un datase en el cual encontramos 14 columnas con más de diez mil observaciones, donde identificamos que EXITED como nuestro variable objetivo.

Características

. *RowNumber*: índice de cadena de datos
. *CustomerId:* identificador de cliente único
. *Surname:* apellido
. *CreditScore:* valor de crédito
. *Geography:* país de residencia
. *Gender:* sexo
. *Age:* edad
. *Tenure:* período durante el cual ha madurado el depósito a plazo fijo de un cliente (años)
. *Balance:* saldo de la cuenta
. *NumOfProducts:* número de productos bancarios utilizados por el cliente
. *HasCrCard:* el cliente tiene una tarjeta de crédito
. *IsActiveMember:* actividad del cliente
. *EstimatedSalary:* salario estimado
. Exicted: el cliente se ha ido

```
#Cargando datos
import pandas as pd
from google.colab import files
uploaded = files.upload()
for filename in uploaded.keys():
    Datos_Loan = pd.read_csv(filename,sep=',')

Datos_Loan.head(7)
```

Elegir archivos Churn.csv
• Churn.csv(text/csv) - 709705 bytes, last modified: 29/10/2024 - 100% done
Saving Churn.csv to Churn (1).csv

1 to 7 of 7 entries Filter

index	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2.0	0.0	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1.0	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8.0	159660.8	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1.0	0.0	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2.0	125510.82	1	1	1	79084.1	0
5	6	15574012	Chu	645	Spain	Male	44	8.0	113755.78	2	1	0	149756.71	1
6	7	15592531	Bartlett	822	France	Male	50	7.0	0.0	2	1	1	10062.8	0

Show 25 per page

Se explica en una Tabla de datos:

Género	País	Edad	Salario en dólares	Salida o existente
Femenino	France	42	10134888	Salió
Femenino	Spain	41	11254258	existente
Femenino	France	42	11393157	Salió
Femenino	France	39	9382663	existente
Femenino	Spain	43	790841	existente
Masculino	Spain	44	14975671	salió
Masculino	France	50	100628	existente

En la siguiente se corre para analizar los datos para mirar que información nos arroja, en el cual nos muestra un índice de rango de 10000 con unas entradas de 0 a 9999.

También podemos ver que tiene 13 columnas de entrada y 1 de salida.

Analizando los datos

```
Datos_Loan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           9091 non-null   float64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(3), int64(8), object(3)
memory usage: 1.1+ MB
```

En este procedimiento es para hacer una descripción para que nos muestre una estadística o resumen de los valores mínimos y máximos del Datase.

```
[8] Datos_Loan.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	9091.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	4.997690	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.894723	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	2.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Preparación de los datos

En este procedimiento vamos a descartar una característica de nuestro conjunto de datos que no nos será de utilidad, (Suname).

```
#Quitando columnas indeseadas
Datos_Loan=Datos_Loan.drop(columns=['Surname'],axis=1)
Datos_Loan.head()
```

1 to 5 of 5 entries Filter

index	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	15634602	619	France	Female	42	2.0	0.0	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1.0	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8.0	159660.8	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1.0	0.0	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2.0	125510.82	1	1	1	79084.1	0

Corremos los datos para que corra las filas que se encuentra errados o mal escritos para que no vaya a generar algún problema, en el cual nos quedamos con 11 características y nuestro variable objetivo.

```
#Elimina filas que tengan datos nulos
Datos_Loan=Datos_Loan.dropna()
Datos_Loan.head()
```

	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	15634602	619	France	Female	42	2.0	0.00	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1.0	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8.0	159660.80	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1.0	0.00	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2.0	125510.82	1	1	1	79084.10	0

Próximos pasos: Generar código con Datos_Loan | Ver gráficos recomendados | New interactive sheet

Para poder trabajar con algoritmos de inteligencia artificial debemos de Cambiar las variables a numéricas, para actualizar a Geography donde nos arroja que (France, Spain y Germany) y nuestra Segunda variable para actualizar Gender (Female y Male).

```
[12] #Verificación de las opciones de la variable
     print('Analizando pais de residencia')
     Datos_Loan['Geography'].unique()

Analizando pais de residencia
array(['France', 'Spain', 'Germany'], dtype=object)
```

```
#Verificación de las opciones de la variable
print('Analizando el género')
Datos_Loan['Gender'].unique()

Analizando el género
array(['Female', 'Male'], dtype=object)
```

Al correr vemos que el conjunto de datos en texto pasa hacer numérico como lo vemos en esta Imagen.

```
#Mapeando todas la variables categóricas a numéricas
Reemplazo_1={'France':1,'Spain':2,'Germany':3}
Datos_Loan['Geography']=Datos_Loan['Geography'].map(Reemplazo_1)

Reemplazo_2={'Female':1,'Male':2}
Datos_Loan['Gender']=Datos_Loan['Gender'].map(Reemplazo_2)

Datos_Loan.head()
```

	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	15634602	619	1	1	42	2.0	0.00	1	1	1	101348.88	1
1	15647311	608	2	1	41	1.0	83807.86	1	0	1	112542.58	0
2	15619304	502	1	1	42	8.0	159660.80	3	1	0	113931.57	1
3	15701354	699	1	1	39	1.0	0.00	2	0	0	93826.63	0
4	15737888	850	2	1	43	2.0	125510.82	1	1	1	79084.10	0

Creando modelos de IA para toma de decisiones

```
# Evaluando casos mediante todos los clasificadores
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,precision_score

Modelo_0 = KNeighborsClassifier(5)
Modelo_0.fit(X_train, Y_train)
Y_pred_0 =Modelo_0.predict (X_test)
print("Accuracy KNN",accuracy_score(Y_test, Y_pred_0))

Modelo_1 = GaussianNB()
Modelo_1.fit(X_train, Y_train)
Y_pred =Modelo_1.predict (X_test)
print("Accuracy Bayes",accuracy_score(Y_test, Y_pred))

Modelo_2 = LinearDiscriminantAnalysis()
Modelo_2.fit(X_train, Y_train)
Y_pred_2 =Modelo_2.predict (X_test)
print("Accuracy LDA",accuracy_score(Y_test, Y_pred_2))
```

```
print("Accuracy QDA",accuracy_score(Y_test, Y_pred_3))

Modelo_4 = DecisionTreeClassifier()
Modelo_4.fit(X_train, Y_train)
Y_pred_4 =Modelo_4.predict (X_test)
print("Accuracy Tree",accuracy_score(Y_test, Y_pred_4))

Modelo_5 = SVC()
Modelo_5.fit(X_train, Y_train)
Y_pred_5 =Modelo_5.predict (X_test)
print("Accuracy SVM",accuracy_score(Y_test, Y_pred_5))
```

```
Accuracy KNN 0.810989010989011
Accuracy Bayes 0.8307692307692308
Accuracy LDA 0.8065934065934066
Accuracy QDA 0.8417582417582418
Accuracy Tree 0.80659340659340 66
Accuracy SVM 0.8395604395604396
```

=========Un análisis de desempeño mas completo (mayor esfuerzo)==========

```
[19] import matplotlib.pyplot as plt
     from sklearn.metrics import confusion_matrix, roc_curve, auc, classification_report
     import seaborn as sns

      # Calcular métricas
     accuracy = Modelo_4.score(X_test, Y_test)
     cm = confusion_matrix(Y_test, Y_pred_0)
     report = classification_report(Y_test, Y_pred_0)
```

Con este modelo de predicción podemos observar y analizar que los clientes estuvieron existentes o hicieron abandono, teniendo los siguientes resultados.

Accuracy KNN 0.810989010989011 - la primera IA tomo una decisión con unos resultados y una seguridad del 81%

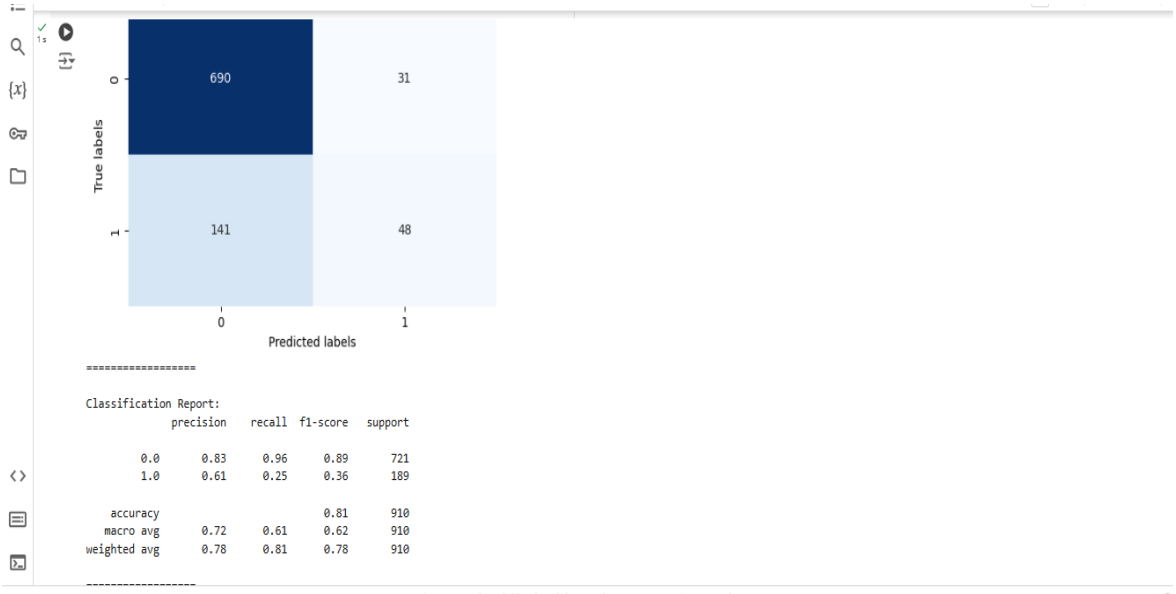
Accuracy Bayes 0.8307692307692308 – la segunda IA tomo una decisión con unos resultados y una seguridad del 83%

Accuracy LDA 0.8065934065934066 – la tercera IA tomo una decisión con unos resultados y una seguridad del 80%

Accuracy QDA 0.8417582417582418 – la cuarta IA tomo una decisión con unos resultados y una seguridad del 84%

Accuracy Tree 0.8065934065934066 – la quinta IA tomo una decisión con unos resultados y una seguridad del 80%

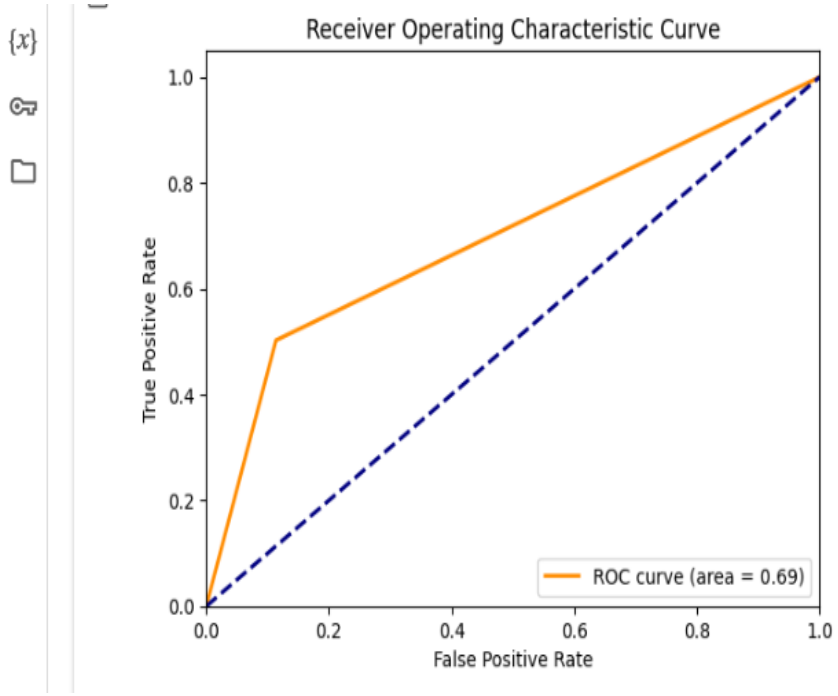
Accuracy SVM 0.8395604395604396 – la sesta IA tomo una decisión con unos resultados y una seguridad del 83%



Podemos evidenciar en esta gráfica y sumando las dos filas tenemos como resultado 721 clientes donde le decimos al modelo que intentara adivinar desde cliente 0, que 690 adivino bien, es decir que esos clientes si estaban abandonando el banco, y 31 veces se equivocó.

En la segunda fila decimos que clientes de tipo 1 tenemos y sumando nos da un resultado de 189 clientes, y que desde 1, dice el modelo que adivino bien, 141 clientes existentes en el banco y 48 veces se equivocó.

En donde decimos que el 83% no se equivoca.



El Grafico nos dice que solo aprendió muy poco con un resultado de ROC 0.69

Probando los datos:

Para hacer pruebas para nuevos clientes hemos hecho 12 variables.

count	9.091000e+03	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000	9091.000000
mean	1.569105e+07	650.736553	1.751732	1.547135	38.949181	4.997690	76522.740015	1.530195	0.704983	0.515565	100181.214924	0.203938
std	7.161419e+04	96.410471	0.831361	0.497801	10.555581	2.894723	62329.528576	0.581003	0.456076	0.499785	57624.755647	0.402946
min	1.556570e+07	350.000000	1.000000	1.000000	18.000000	0.000000	0.000000	1.000000	0.000000	0.000000	11.580000	0.000000
25%	1.562899e+07	584.000000	1.000000	1.000000	32.000000	2.000000	0.000000	1.000000	0.000000	0.000000	51227.745000	0.000000
50%	1.569106e+07	652.000000	1.000000	2.000000	37.000000	5.000000	97318.250000	1.000000	1.000000	1.000000	100240.200000	0.000000
75%	1.575285e+07	717.000000	3.000000	2.000000	44.000000	7.000000	127561.890000	2.000000	1.000000	1.000000	149567.210000	0.000000
max	1.581566e+07	850.000000	3.000000	2.000000	92.000000	10.000000	250898.090000	4.000000	1.000000	1.000000	199992.480000	1.000000

```
[21] #Probando el modelo entrenado sobre un nuevo sujeto
     Target=np.zeros((1,11))
     Target[0,0]=float(input('Ingrese identificador de cliente unico: '))
     Target[0,1]=float(input('Ingrese valor de credito: '))
     Target[0,2]=float(input('Ingrese pais de recidencia: '))
     Target[0,3]=float(input('Ingrese el genero, 1 para femenino y 2 para masculino: '))
     Target[0,4]=float(input('Ingrese la edad: '))
     Target[0,5]=float(input('Ingrese deposito a plazo fijo de un cliente años: '))
     Target[0,6]=float(input('Ingreso saldo de la cuenta, entre 0 y 26000: '))
     Target[0,7]=float(input('Cantidad de productos bancarios utilizados por el cliente, entre 0 y 4: '))
     Target[0,8]=float(input('Ingrese tarjeta de credito, 1 si y 0 no: '))
     Target[0,9]=float(input('Activida del cliente, 1 si y 0 no: '))
     Target[0,10]=float(input('Ingreso salarial, entre 0 y 20000: '))

     Target = scaler.transform(Target) #Normalizar los datos
```

```
print(" ")

if Prediction_1==0:
  print("Según Bayes, Cliente que se esta yendo")
else:
  print("Según Bayes, Cliente existente")

print(" ")

if Prediction_2==0:
  print("Según LDA, Cliente que se esta yendo")
else:
  print("Según LDA, Cliente existente")

print(" ")

if Prediction_3==0:
  print("Según QDA, Cliente que se esta yendo")
else:
  print("Según QDA, Cliente existente")

print(" ")

if Prediction_4==0:
  print("Según Tree, Cliente que se esta yendo")
else:
  print("Según tree, Cliente existente")

print(" ")

if Prediction_5==0:
  print("Según SVM, Cliente que se esta yendo")
else:
```

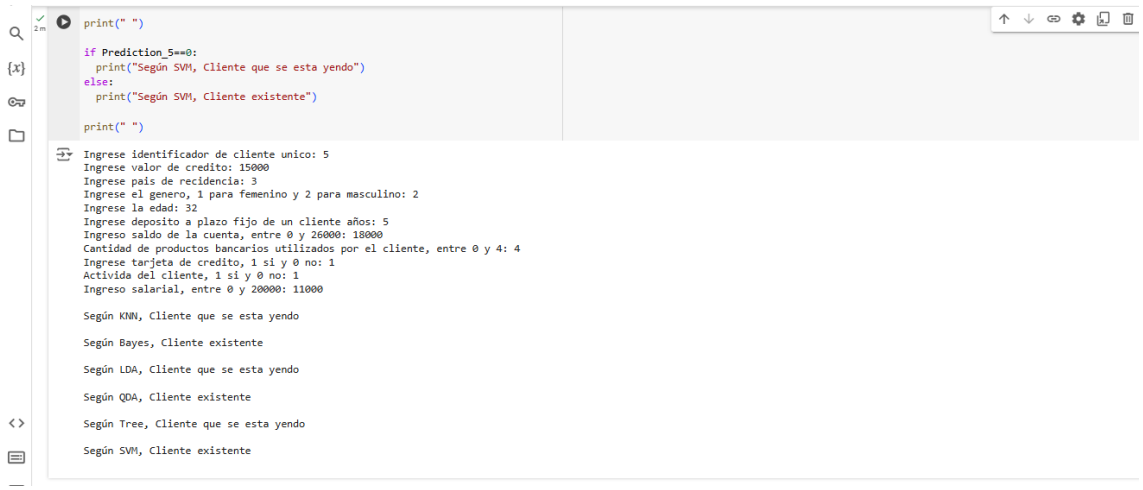
```
print(" ")

Ingrese identificador de cliente unico: 1
Ingrese valor de credito: 1000
Ingrese pais de recidencia: 2
Ingrese el genero, 1 para femenino y 2 para masculino: 2
Ingrese la edad: 32
Ingrese deposito a plazo fijo de un cliente años: 3
Ingreso saldo de la cuenta, entre 0 y 26000: 16000
Cantidad de productos bancarios utilizados por el cliente, entre 0 y 4: 2
Ingrese tarjeta de credito, 1 si y 0 no: 1
Activida del cliente, 1 si y 0 no: 1
Ingreso salarial, entre 0 y 20000: 12000

Según KNN, Cliente que se esta yendo

Según Bayes, Cliente existente

Según LDA, Cliente existente

Según QDA, Cliente existente

Según Tree, Cliente que se esta yendo

Según SVM, Cliente existente
```

Hacemos hecho una consulta a la IA donde entrevistamos una mujer como nueva cliente en el banco en donde tiene una edad de 32 años y nos arrojó unos resultados donde esta persona se queda en el banco.

```
print(" ")

if Prediction_5==0:
  print("Según SVM, Cliente que se esta yendo")
else:
  print("Según SVM, Cliente existente")

print(" ")
```

```
Ingrese identificador de cliente unico: 2
Ingrese valor de credito: 20000
Ingrese pais de recidencia: 1
Ingrese el genero, 1 para femenino y 2 para masculino: 1
Ingrese la edad: 38
Ingrese deposito a plazo fijo de un cliente años: 1
Ingreso saldo de la cuenta, entre 0 y 26000: 8000
Cantidad de productos bancarios utilizados por el cliente, entre 0 y 4: 1
Ingrese tarjeta de credito, 1 si y 0 no: 2
Activida del cliente, 1 si y 0 no: 3
Ingreso salarial, entre 0 y 20000: 13000

Según KNN, Cliente que se esta yendo

Según Bayes, Cliente existente

Según LDA, Cliente que se esta yendo

Según QDA, Cliente existente

Según Tree, Cliente que se esta yendo

Según SVM, Cliente existente
```

Le hemos hecho la consulta a la IA donde hemos entrevistado una mujer como nueva cliente en el banco en donde tiene una edad de 38 años y nos arrojó unos resultados del 50% se queda o se ira del banco.



```
print(" ")

if Prediction_5==0:
  print("Según SVM, Cliente que se esta yendo")
else:
  print("Según SVM, Cliente existente")

print(" ")
```

```
Ingrese identificador de cliente unico: 5
Ingrese valor de credito: 15000
Ingrese pais de recidencia: 3
Ingrese el genero, 1 para femenino y 2 para masculino: 2
Ingrese la edad: 32
Ingrese deposito a plazo fijo de un cliente años: 5
Ingreso saldo de la cuenta, entre 0 y 26000: 18000
Cantidad de productos bancarios utilizados por el cliente, entre 0 y 4: 4
Ingrese tarjeta de credito, 1 si y 0 no: 1
Activida del cliente, 1 si y 0 no: 1
Ingreso salarial, entre 0 y 20000: 11000

Según KNN, Cliente que se esta yendo

Según Bayes, Cliente existente

Según LDA, Cliente que se esta yendo

Según QDA, Cliente existente

Según Tree, Cliente que se esta yendo

Según SVM, Cliente existente
```

Le hemos hecho la consulta a la IA donde hemos entrevistado un hombre como nuevo cliente en el banco en donde tiene una edad de 32 años y nos arrojó unos resultados del 50% se queda o se puede ir del banco.

BIBLIOGRAFIA:

kaggle kernels output valentinmaz/modelo-de-predicci-n-de-abandono-de-clientes-banco -p /path/to/dest

Cárdenas Herrera, P. M. (2022). El abandono de clientes en banca.

NK Pari Apaza - 2024 - repositorio.upt.edu.pe. Marketing digital y su relación con la captación de clientes del Centro Comercial La Multiplaza Ollaraya, La Victoria, Lima, 2024. upt.edu.pe

Beyreuther, C., & Abalde, R. Minería de datos en CRM: modelos de predicción de abandono de clientes bancarios.

A Lerma Micó - dspace.umh.es. Sistemas de recomendación en R y aplicación a datos reales. umh.es

V Osorio Urrea - 2024 - repository.eafit.edu.co. Analizando patrones de éxito en YouTube: un sistema de recomendación para creadores de contenidos educativos. eafit.edu.co

N San Emeterio Rodríguez - 2024 - digibuo.uniovi.es. Aplicación de compra y venta de productos basada en un sistema de recomendación. uniovi.es

Decisiontreeclassifier (no date) scikit. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html (Accessed: 04 November 2024).

Gaussiannb (no date) scikit. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.naive_bayes.GaussianNB.html (Accessed: 04 November 2024).

Lineardiscriminantanalysis (no date) scikit. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html (Accessed: 04 November 2024).

Precision score (no date) scikit. Available at: https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.precision_score.html (Accessed: 04 November 2024).

Quadraticdiscriminantanalysis (no date) scikit. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html (Accessed: 04 November 2024).

Salunke, D. (2024) *SVC (support vector classifier), LinkedIn*. Available at: https://www.linkedin.com/pulse/svc-support-vector-classifier-dishant-salunke (Accessed: 04 November 2024).

Sklearn.neighbors. kneighborsclassifier¶ (no date) *sklearn.neighbors.KNeighborsClassifier - documentación de scikit-learn - 0.24.1*. Available at: https://qu4nt.github.io/sklearn-doc-es/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (Accessed: 04 November 2024).