UNIVERSIDAD TECNOLÓGICA NACIONAL

FACULTAD REGIONAL ROSARIO

Cátedra: Minería de datos

Electiva de Ingeniería en Sistemas de Información

2023

Trabajo práctico integrador - ETAPA 1

Profesores/as: Scime, Silvia - Di Carlo, Martina

Turno: Tarde

Comisión N°: 5E05

Grupo N°: 1

Integrantes:

- 47771 – Fernández Cariaga, Ezequiel
- 47854 – Pinciroli, Franco Iván
- 48618 – Sola, Martín Ricardo
- 44768 - Yodice, Martín Adrián

Índice

Fase de análisis del problema

Definición de los objetivos

Nos presentamos ante un trabajo en el cual fuimos contactados por el gerente general de la empresa “AllHome”, quien nos solicitó la colaboración para el diseño de campañas de publicidad por correo electrónico de su nueva marca propia. Teniendo un listado de posibles clientes, es preciso saber quiénes de ellos pudiesen estar interesados en comprar bicicletas para enviarle publicidad por e-mail.

Nuestro objetivo es encontrar un modelo que pueda predecir con gran porcentaje de acierto quiénes de esos destinatarios podrían ser potenciales clientes para enviarles la publicidad.

Para ello tendremos que crear un modelo que se ajuste lo más posible a la realidad a partir de los datos de clientes ya existentes en la empresa.

Pondremos a prueba tres técnicas predictivas como son los árboles de decisiones, el vecino más próximo (KNN) y el análisis discriminante (LDA). Luego de entrenarlos, a cada uno se le evaluará su rendimiento con una porción de los datos para determinar cómo responde. Es allí cuándo podremos compararlos para determinar el modelo de predicción que nos resulte más preciso.

Es importante tener en cuenta que, de acuerdo con lo que se nos comunica, no es tan importante que el modelo se enfoque en predecir quienes no están interesados, sino que se asegure de no dejar pasar a los potenciales compradores por no enviarles publicidad.

Fase de preprocesamiento de los datos

Dentro de esta fase vamos a realizar las siguientes actividades:

- Análisis exploratorio de los datos (univariante y multivariante).
- Proceso de limpieza de datos.
- Especificación de las vistas minables.

El primer paso es eliminar las columnas que creemos son irrelevantes para el análisis de los datos, y que estorban a la hora de crear el modelo de predicción.

Análisis Univariante

En esta etapa vamos a analizar las variables de manera independiente observando los parámetros principales de cada una.

Análisis de parámetros principales

Como primera parte del análisis univariante analizamos los primeros parámetros importantes como promedio, desvío estándar, valor máximo, valor mínimo, percentiles, cantidad de valores, entre otros. Ellos se detallan en la tabla de la siguiente página.

Analizando la tabla, existen puntos importantes a destacar:

- Como primer punto, la cantidad de clientes en la lista es de 6400, de los cuales 10 se desconoce la información sobre sus ingresos anuales. Se debe decidir qué hacer con estos valores nulos, para así mejorar la calidad del conjunto de datos.
- Otro resultado que destaca es la existencia de al menos un ingreso anual máximo de 170.000. Se debe tener precaución con él, para así determinar si se trata de un valor aislado o si se repite en más de una ocasión, clasificándose como outlier o no, respectivamente.
- Además, se observa que los rangos de las variables varían significativamente, por lo que antes de pasar a la modelización es necesario estandarizarlas.

	Estado Civil	Género	Ingreso Anual	TotalHijos	Educación	Ocupación	Propietario	Cant. Automóviles	Distancia	Región	Edad	Compró Bicicleta
Cantidad	6400	6400	6390	6400	6400	6400	6400	6400	6400	6400	6400	6400
Valores únicos	3	2	-	-	5	5	-	-	5	4	-	-
Valor más frecuente	C	M	-	-	Licenciatura	Profesional	-	-	'0-1 Km.'	Norte	-	-
Frecuencia	3504	3223	-	-	1800	1946	-	-	2166	3310	-	-
Media	-	-	57532.08	1.90	-	-	0.68	1.55	-	-	51.20	0.39
Desviación estándar	-	-	32331.97	1.63	-	-	0.47	1.15	-	-	11.52	0.49
Mínimo	-	-	10000	0	-	-	0	0	-	-	32	0
Primer cuartil	-	-	30000	0	-	-	0	1	-	-	42	0
Mediana	-	-	60000	2	-	-	1	2	-	-	49	0
Tercer cuartil	-	-	70000	3	-	-	1	2	-	-	59	1
Máximo	-	-	170000	5	-	-	1	4	-	-	102	1

Varianza y Desvío estándar

A continuación, se detallan las varianzas y el desvío estándar con redondeo a 4 decimales:

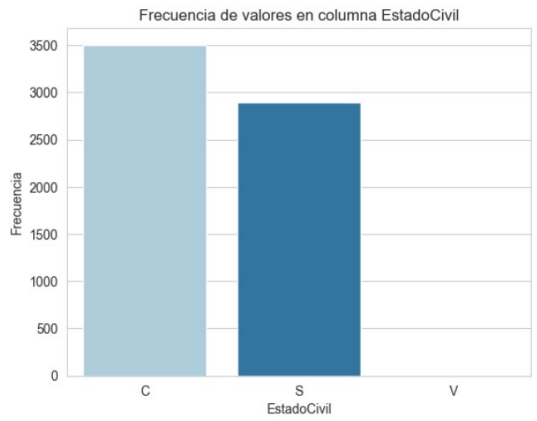
```
----------Varianza----------
IngresoAnual       1045356225.2806
TotalHijos                2.6601
Propietario               0.2189
CantAutomoviles           1.3157
Edad                    132.6574
ComproBicicleta           0.2389

----------Desvío estándar----------
IngresoAnual       32331.9691
TotalHijos             1.6310
Propietario            0.4678
CantAutomoviles        1.1471
Edad                  11.5177
ComproBicicleta        0.4888
```

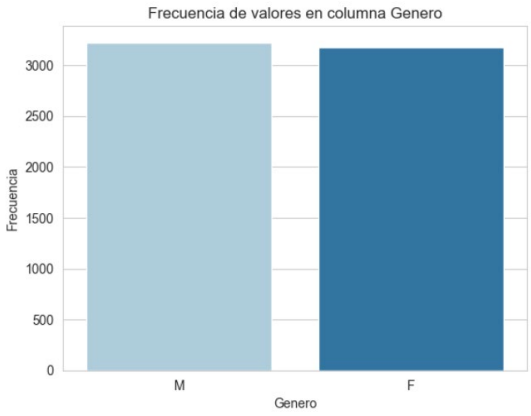
Analizando los resultados y basándose además en las medidas analizadas anteriormente, se puede concluir que los valores de los ingresos anuales se encuentran muy dispersos, y que su rango es muy amplio. Como se ve en el punto anterior, estos varían desde 10.000 hasta 170.000.

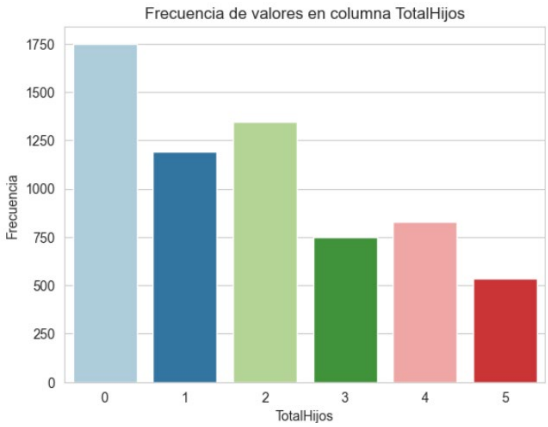
Por otro lado, refiriéndose a las demás columnas, no se encuentran datos relevantes para realizar un análisis en profundidad.

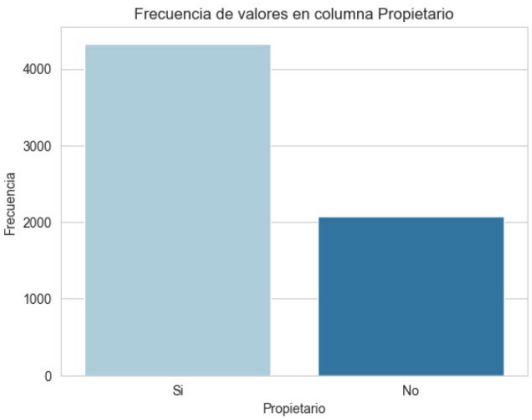
Frecuencia absoluta y relativa

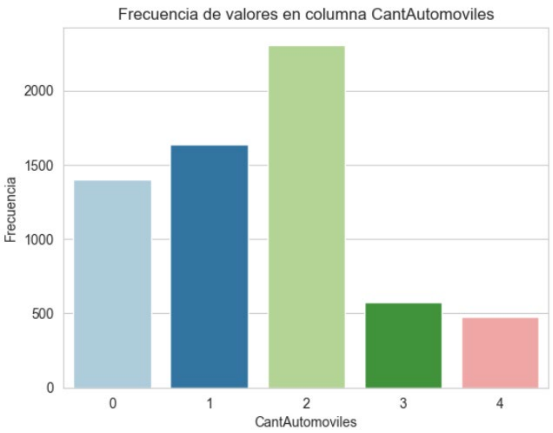
A continuación, se detallan los gráficos de las frecuencias absolutas y relativas de cada variable, es decir, la cantidad de veces que aparece cada valor en la columna y su proporción en porcentaje.

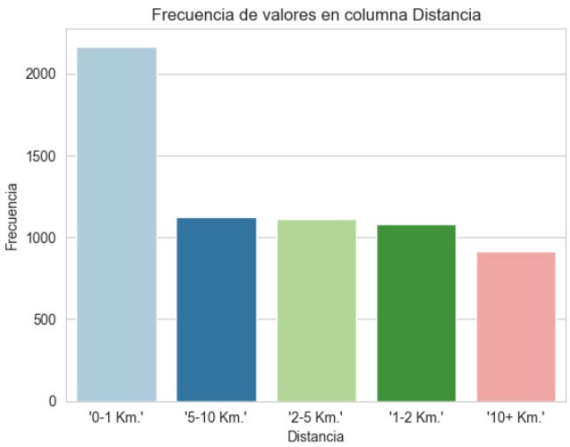

Frecuencia de valores en columna EstadoCivil
Frecuencia
3500
3000
2500
2000
1500
1000
500
0
C
S
V
EstadoCivil

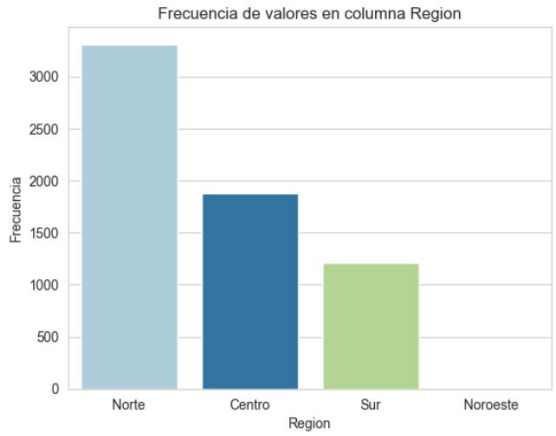

Frecuencia de valores en columna Genero
Frecuencia
3000
2500
2000
1500
1000
500
0
M
F
Genero

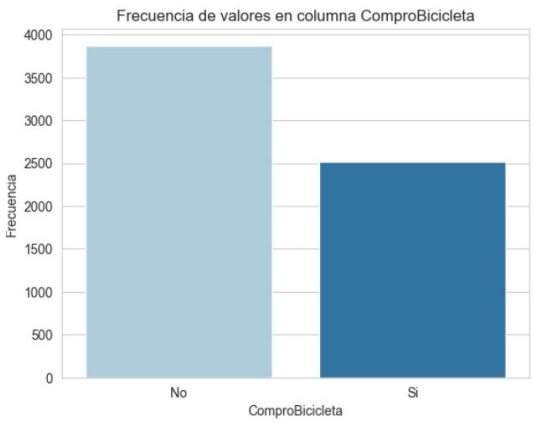

Frecuencia de valores en columna TotalHijos
Frecuencia
1750
1500
1250
1000
750
500
250
0
0
1
2
3
4
5
TotalHijos

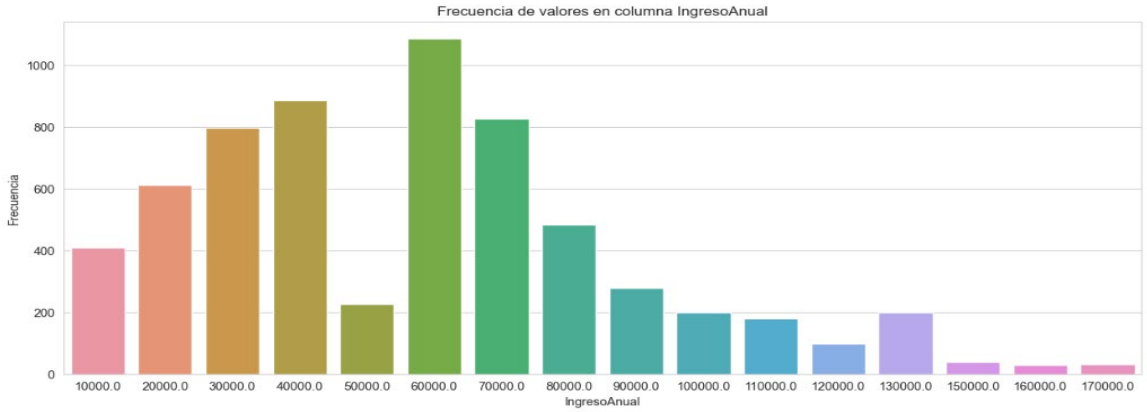

Frecuencia de valores en columna Propietario
Frecuencia
4000
3000
2000
1000
0
Si
No
Propietario

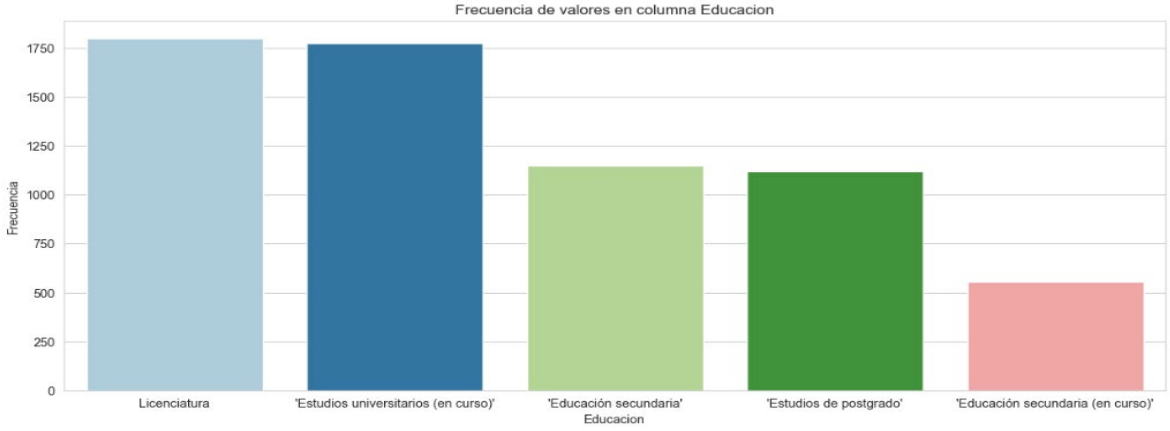

Frecuencia de valores en columna CantAutomoviles
Frecuencia
2000
1500
1000
500
0
0
1
2
3
4
CantAutomoviles

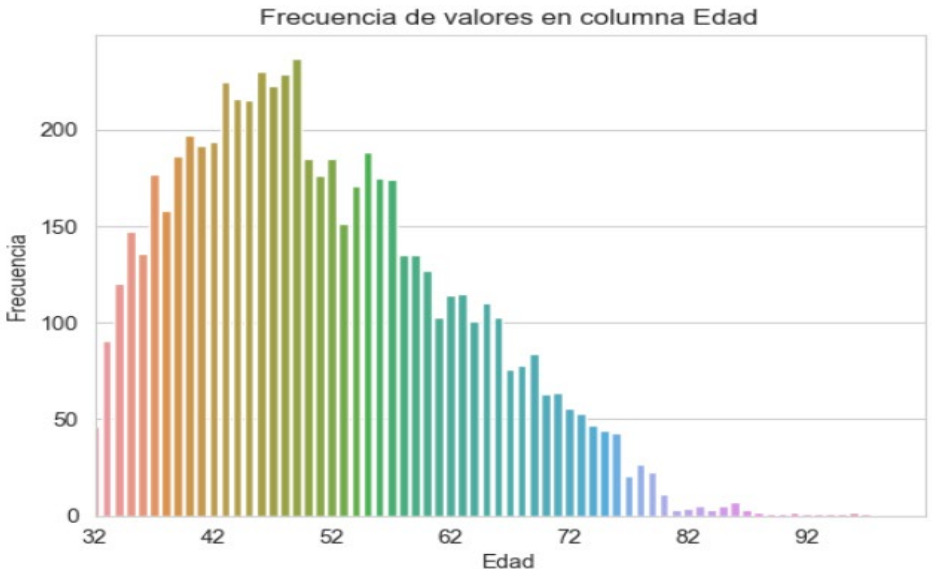

Frecuencia de valores en columna Distancia
Frecuencia
2000
1500
1000
500
0
'0-1 Km.'
'5-10 Km.'
'2-5 Km.'
'1-2 Km.'
'10+ Km.'
Distancia


Frecuencia de valores en columna Region
Frecuencia
3000
2500
2000
1500
1000
500
0
Norte
Centro
Sur
Noroeste
Region


Frecuencia de valores en columna ComproBicicleta
Frecuencia
4000
3500
3000
2500
2000
1500
1000
500
0
No
Si
ComproBicicleta


Frecuencia de valores en columna IngresoAnual
Frecuencia
0
200
400
600
800
1000
10000.0
20000.0
30000.0
40000.0
50000.0
60000.0
70000.0
80000.0
90000.0
100000.0
110000.0
120000.0
130000.0
150000.0
160000.0
170000.0
IngresoAnual


Frecuencia de valores en columna Educacion
Frecuencia
0
250
500
750
1000
1250
1500
1750
Licenciatura
'Estudios universitarios (en curso)'
'Educación secundaria'
'Estudios de postgrado'
'Educación secundaria (en curso)'
Educacion


Frecuencia de valores en columna Edad
Frecuencia
0
50
100
150
200
32
42
52
62
72
82
92
Edad

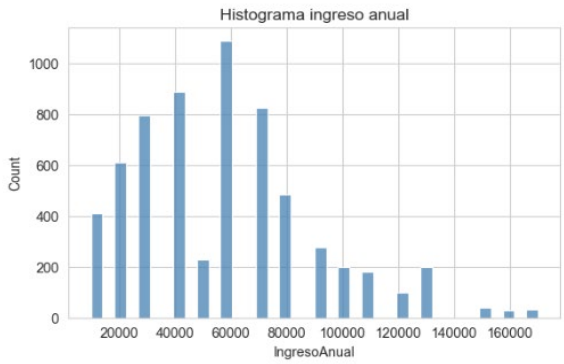
En cuanto a la variable ingreso anual, si se analiza su gráfico es simple detectar que existe más de un valor de $170.000 y que, además, existen otros valores cercanos como $160.000 o $150.000 que se encuentran lejos del rango de la media y sobre los cuales se debe decidir si tenerlos presentes para el análisis, considerando en qué medida afectan al promedio.

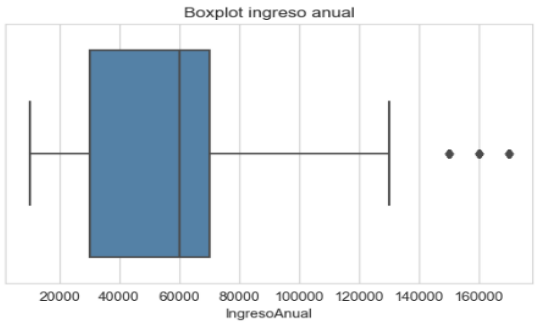
Análisis de histograma y boxplots

Ingreso anual:

La mayor parte de los clientes tienen ingresos anuales entre $30.000 y $70.000, por lo que la variable está sesgada a la izquierda.

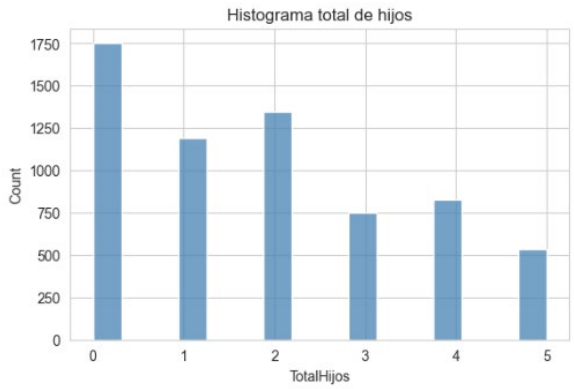
En el gráfico de boxplot del ingreso anual se observan tres valores que están fuera del intervalo, al ser más grandes que el bigote del máximo (tercer cuartil + 1.5 RIQ). Si bien son valores atípicos, puede tratarse sólo de personas con un ingreso anual considerablemente mayor a la media, pero no de forma excesiva. Es importantes tenerlos en cuenta para verificar si afectan en gran proporción al promedio y si deben ser eliminados o no.

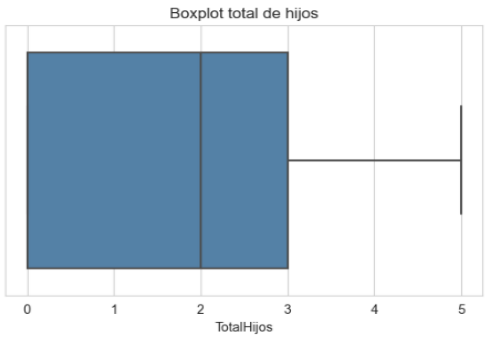




Total de hijos:

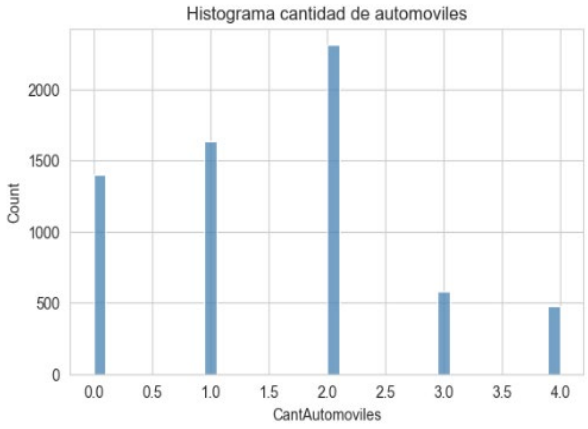
Observando el gráfico, se puede notar que gran parte de los clientes no tiene hijos (alrededor del 25%), información sumamente importante, a sabiendas de que la marca ofrece bicicletas para niños, evitando enviarles publicidad de este tipo. Todos los valores están dentro de los intervalos del boxplot, es decir que todos los clientes tienen entre 0 y 5 hijos.

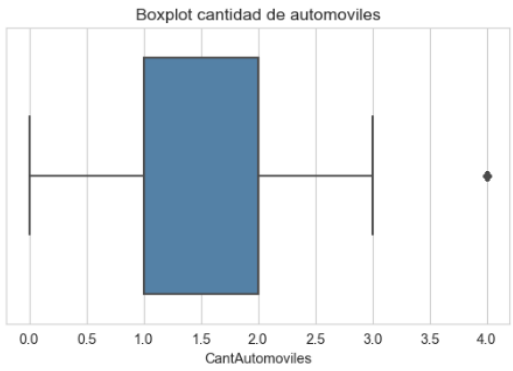




Cantidad de automóviles:

Tanto el boxplot como el histograma muestran que la media de la cantidad de automóviles es igual a 2. Como se observó anteriormente en el análisis de los parámetros principales, se puede confirmar que la mediana coincide con el tercer cuartil y es por eso por lo que no se dibuja una línea dentro del rectángulo. Se ve también que existe un valor anómalo en 4.



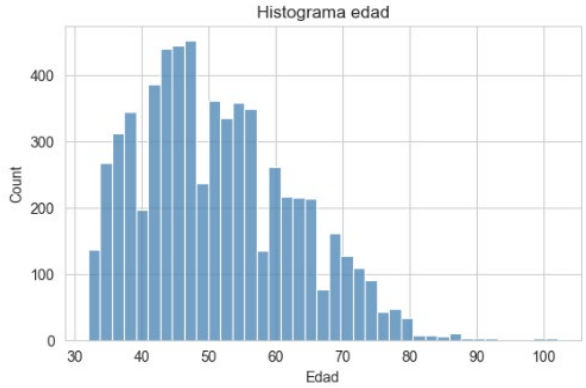


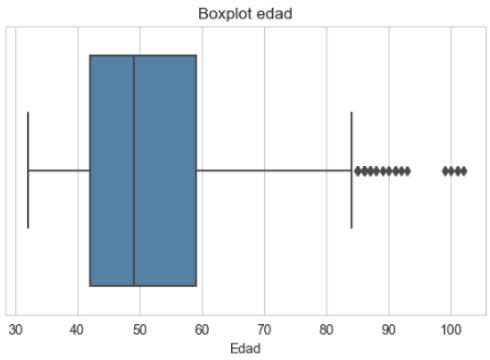
Edad:

Analizando el histograma de la edad es fácil notar que la variable tiene una distribución normal sesgada a la izquierda, y que posee pocos valores entre 80 y 100 años que se alejan de la media.

Ahora, apoyándose en el diagrama de boxplot de esta variable se observa que existen algunos valores que están fuera del máximo que está calculado mediante Q3 + 1.5 RIQ. Esto puede indicar que estamos frente a valores anómalos o “outliers”, y que quizás debamos modificar o retirar.

En la etapa final del análisis exploratorio se deberá tomar una decisión acerca de qué hacer con estos valores.



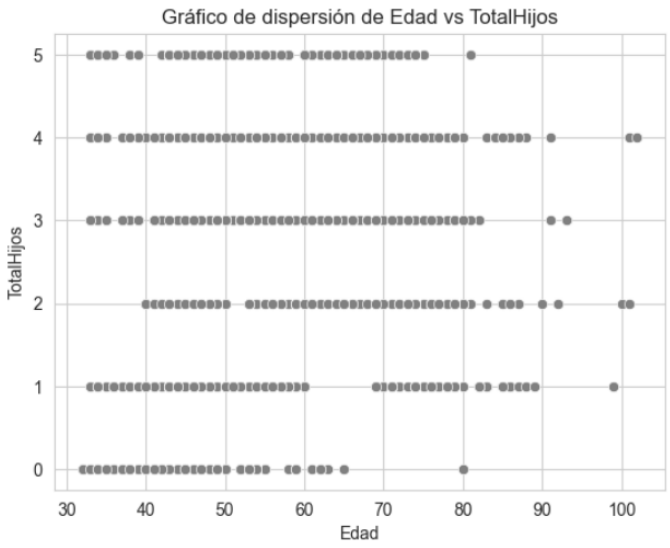


Análisis de diagramas de dispersión (Scatterplot)

Edad - Total de hijos

De este gráfico se destaca que los clientes que no tienen hijos se encuentran por debajo de los 65 años, a pesar de que exista uno solo con 80 años.

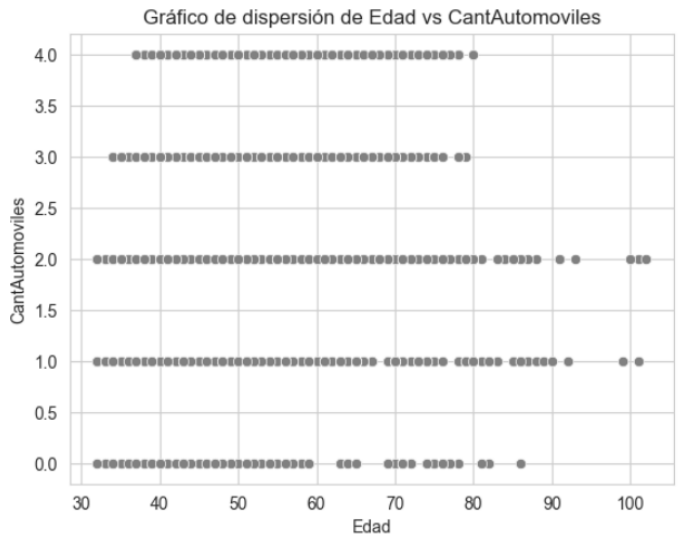
En un histograma anterior se había visto que 0 hijos era la mayor frecuencia de la variable total de hijos dentro de los clientes, por lo que es importante considerar este grupo. Los demás tienen una dispersión bastante similar, con el detalle de que cuando el total de hijos es igual a 2, no se tienen clientes menores de 40 años.



Edad - Cantidad de automóviles

No se aprecia una relación significativa entre la edad y la cantidad de automóviles, sólo una pequeña concentración en las edades a medida que se incrementa el número de vehículos.

Los únicos detalles para destacar son que a partir de 3 y 4 automóviles no se encuentran clientes con edades mayores a 80 años, y que los pocos clientes registrados que están en torno a los 100 años tienen al menos un automóvil.

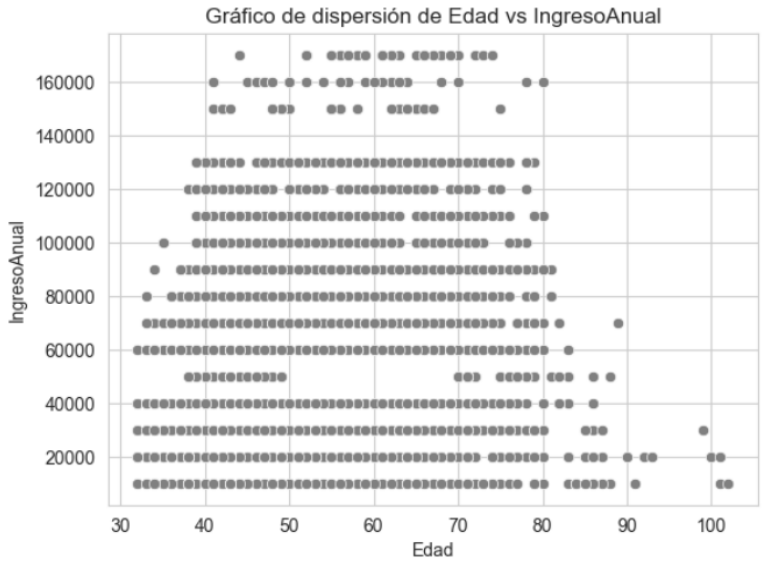


Edad - Ingreso anual

En el gráfico se aprecia que en los niveles más bajos de ingresos anuales existen clientes de todos los rangos de edad.

A medida que se incrementan los ingresos, la edad mínima aumenta gradualmente hasta estar por encima de los 40 años, mientras que la edad máxima de los clientes en esos rangos también se ve reducida.

Se puede observar que los clientes con los mayores ingresos anuales (más de $140.000) se encuentran todos entre 40 y 80 años.



Análisis Multivariante

En esta etapa de la fase de preprocesamiento de datos, se analizará cómo se relacionan las distintas variables entre sí y qué combinación de variables puede servir para clasificarlas mejor.

Análisis de matriz S

	IngresoAnual	TotalHijos	CantAutomoviles	Edad
IngresoAnual	1.00	0.22	0.47	0.15
TotalHijos	0.22	1.00	0.27	0.50
CantAutomoviles	0.47	0.27	1.00	0.17
Edad	0.15	0.50	0.17	1.00

Interpretando la matriz S y se puede notar que las variables que están más relacionadas entre sí son las variables de ingreso anual con la cantidad de automóviles, así como también el total de hijos con la edad. Ambas se relacionan directamente, es decir, que a mayor ingreso aumenta la cantidad de automóviles, y a menor edad menos cantidad de hijos.

Esto sirve para prestar atención a los scatterplots (diagramas de dispersión) y a los boxplots estratificados de esos pares de variables y poder determinar de qué manera están relacionadas.

Análisis de matriz R

La construcción de una Matriz R no es adecuada en este caso, ya que las unidades de las variables son diferentes y la covarianza no está normalizada.

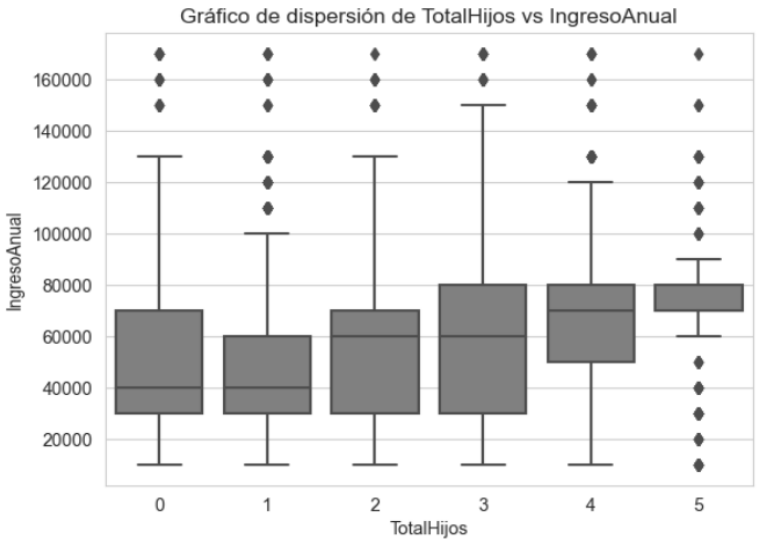
Análisis de Boxplots estratificados

Total de hijos - Ingreso anual

Se aprecia que la mediana es igual para los que tienen uno o ningún hijo, aumenta para los que tienen entre 2 y 3 y vuelve a incrementar para aquellos que tienen entre 4 y 5.

Se presenta un rango considerablemente menor de ingresos anuales para los que tienen 5 hijos. Para todas las cantidades de hijos existen valores de ingresos anuales por encima de la media, superiores a $160.000.

Como conclusión se entiende que la cantidad de hijos no depende del ingreso anual del cliente.



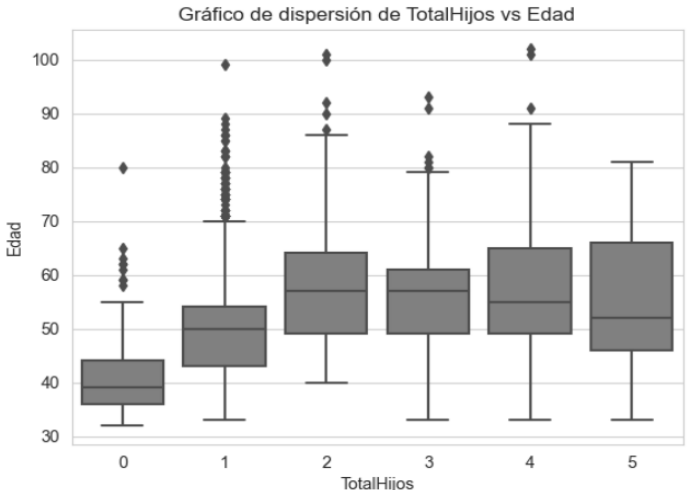
Total de hijos - Edad

Siguiendo las edades que cubre la caja del boxplot, se percibe que van desplazándose hacia arriba desde 0 hasta 2 hijos.

Las medianas de edad también aumentan hasta los 2 hijos y de allí en adelante se mantienen estables, incluso disminuyendo un poco en el caso de 5 hijos.

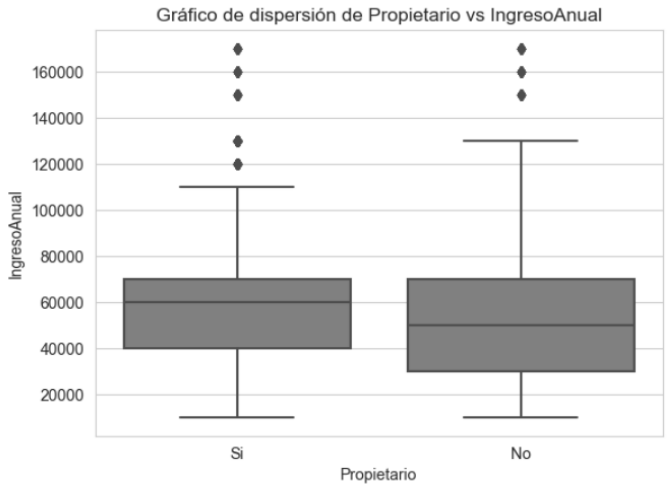
No se ven valores atípicos para aquellos clientes con 5 hijos, pero en el valor de se trata de un sólo hijo, existe una gran cantidad de valores anómalos.

Quienes no tienen hijos presentan en su mayoría edades más jóvenes que los demás.



Propietario - Ingreso anual

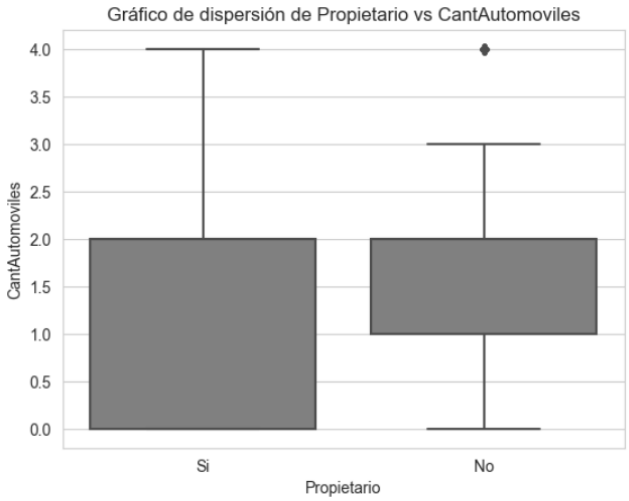
Se aprecia que para quienes son propietarios hay una mediana más alta con asimetría negativa, además de que la caja es más estrecha hacia arriba, por lo que los ingresos anuales menores de los propietarios son más altos.



Propietario - Cantidad de automóviles

El 75% de los valores de los propietarios se encuentran entre 0 y 2, mientras que los no propietarios en su mayoría tienen 1 o 2 automóviles.

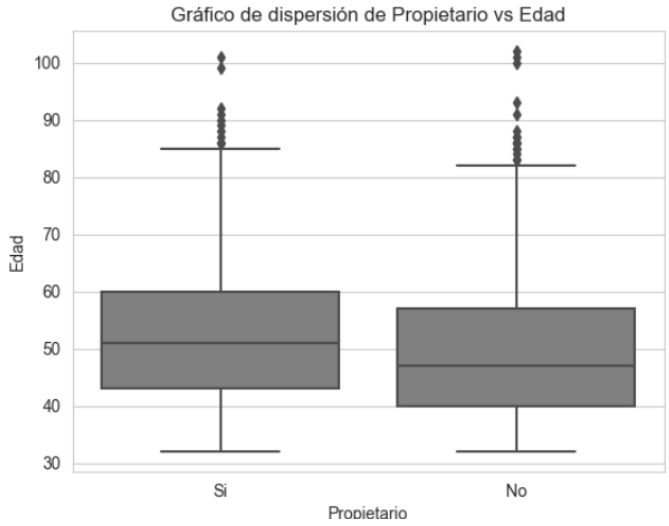
Luego, el intervalo de los propietarios llega hasta el valor máximo de automóviles (4), mientras que en los no propietarios hay unos pocos valores atípicos en dicho valor 4.



Propietario - Edad

No se aprecian diferencias significativas en la edad entre aquellos clientes que son propietarios y los que no. La mediana de los propietarios es un poco mayor (más de 50 años) que los que no lo son (menos de 50 años), y es prácticamente simétrica.

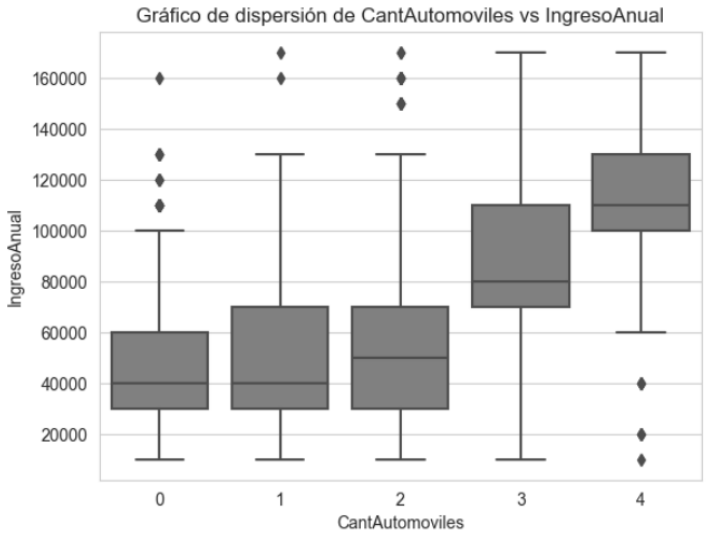
Los cuartiles de los propietarios están entre 45 y 60 años, mientras que los no propietarios entre 40 y más de 55 años aproximadamente.



Cantidad de automóviles - Ingreso anual

Aquí la mediana de ingreso anual lógicamente es proporcional a la cantidad de automóviles que poseen. Aumenta a partir de los 2 automóviles, y a partir de 3 crece considerablemente.

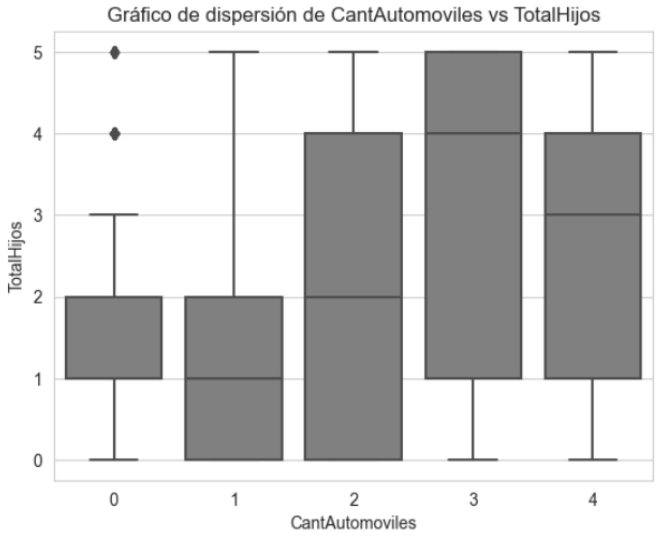
Existen varios valores atípicos, sin embargo, los más llamativos se encuentran en los 4 automóviles al encontrar varios clientes con un ingreso anual considerablemente bajo (menor a $40.000) que poseen 4 automóviles. Esto resulta ilógico si se tiene en cuenta que quienes tienen mayor cantidad de autos deberían tener un ingreso anual alto. Por ello, debe considerarse cada caso cuando se revisen los “outliers”.



Cantidad de automóviles - Total de hijos

Se puede advertir que la mediana aumenta a medida que también aumenta la cantidad de automóviles. Esto quiere decir que se mantiene, en la mayoría de los casos, una relación directa, en la que a mayor cantidad de automóviles se cuenta con un mayor número de hijos.

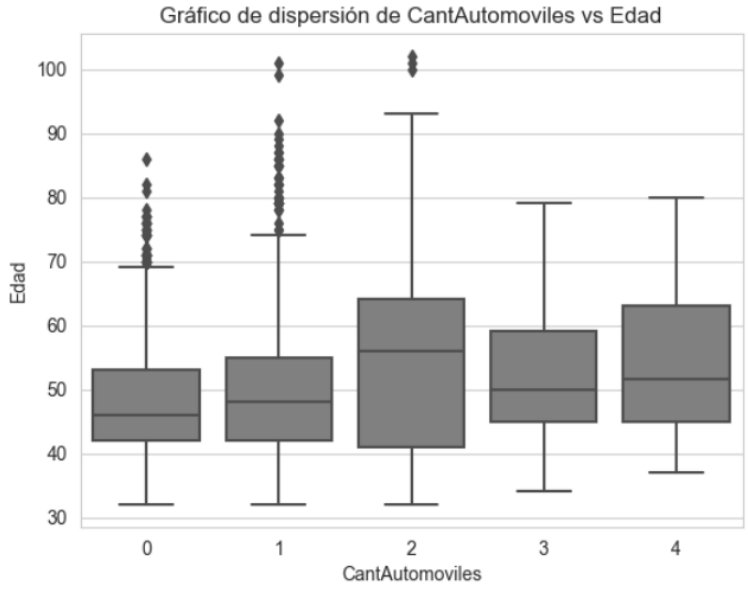
Además, es perceptible que para quienes cuentan con dos automóviles, se mantiene una distribución casi uniforme y el rango es tan amplio que abarca casi todo el dominio de la cantidad de hijos. Sin embargo, cuando se cuenta con un sólo automóvil, se aprecia que el valor del tercer cuartil (75%) es de 2 hijos y su caja es pequeña, por lo que es estrecho el rango, a pesar de existir 2 outliers.



Cantidad de automóviles - Edad

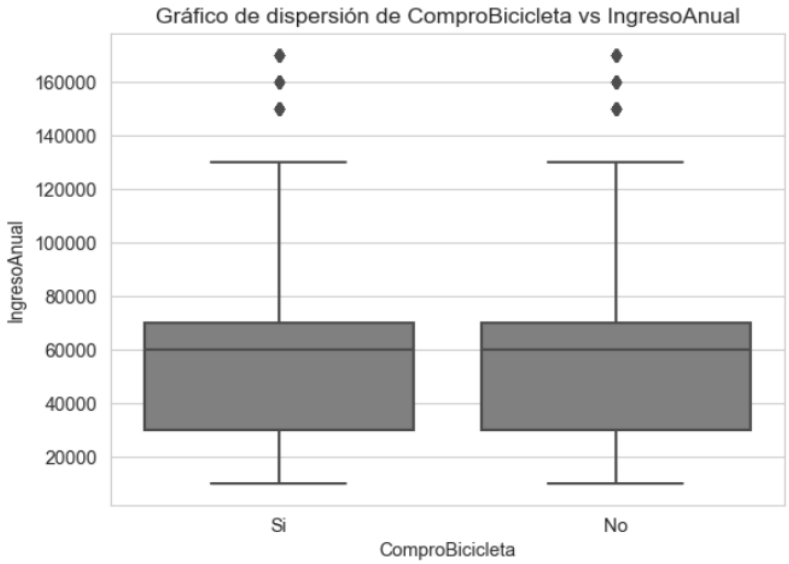
Lo primero que es evidente notar, es que para cada valor de la cantidad de automóviles la mediana de la edad varía poco (45 a 55 años). Se puede inferir que para las personas que cuentan con 0,1 o 2 automóviles normalmente tienen una edad en un rango de 45 a 60 años, sin evidentemente obviar los múltiples outliers existentes en estos tres valores.

Cuando se trata de 1 sólo automóvil, analizando su caja se ve que es prácticamente uniforme, pero también que es el valor para el cual se encuentran más valores anómalos. Para los valores de 3 a 4 automóviles no existen outliers.



Compró bicicleta - Ingreso anual

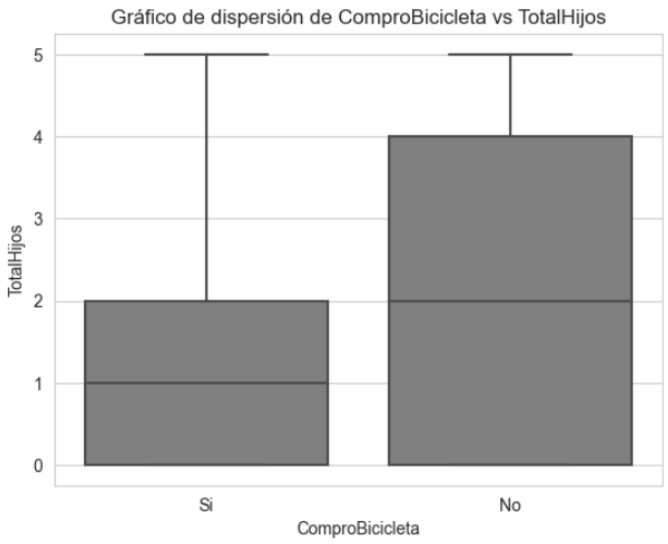
En este caso se avista que ambos boxplots son exactamente iguales para ambos valores de la variable compró bicicleta. Por este motivo, y también como se demostró anteriormente en la matriz S, queda claro que ambas variables no están relacionadas entre sí y que, por ello, el comprar o no una bicicleta no depende del ingreso anual.



Compró bicicleta - Total de hijos

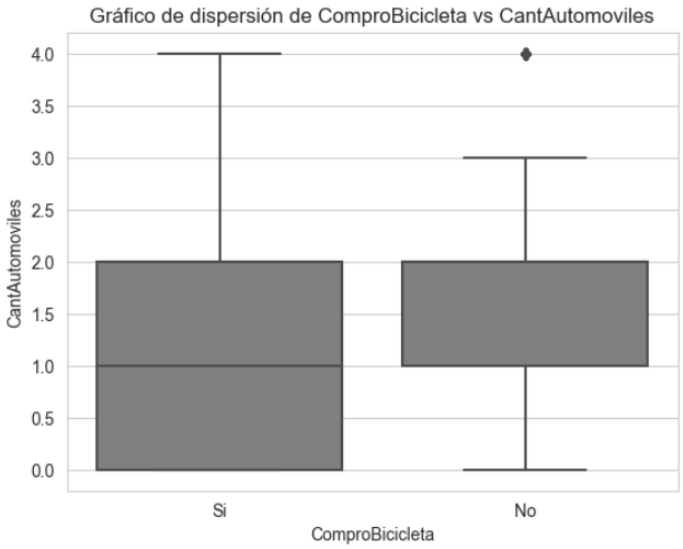
Lo primero que se nota es que la media disminuye en el caso de los clientes que sí compraron bicicleta. Es decir, el 50% de los que compraron bicicleta tienen 1 hijo, mientras que el 50% de los clientes que no compraron bicicleta tienen 2 hijos.

Mirando el tamaño de las cajas se determina que es más amplio el rango de aquellos clientes que no compraron bicicleta que los que sí lo hicieron. Como un análisis predictivo se puede suponer que aquellos clientes que tengan 4 o 5 hijos es más probable es que no compren una bicicleta.



Compró bicicleta - Cantidad de automóviles

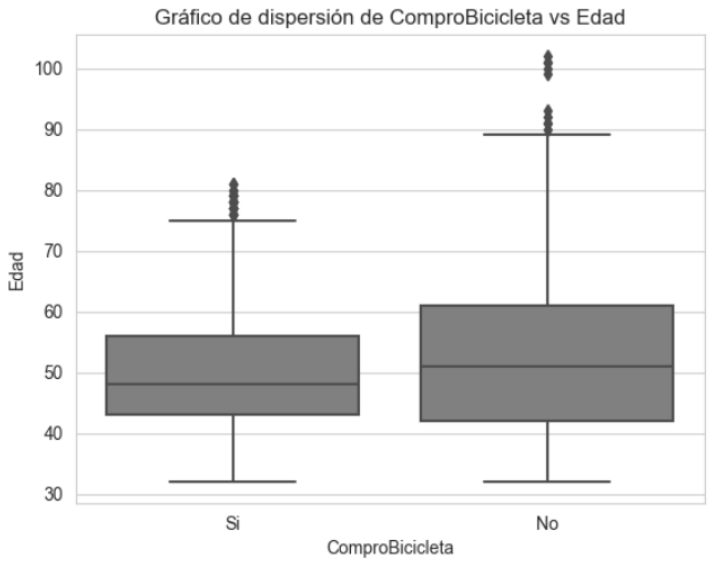
Lo importante a analizar en este boxplot estratificado es que el rango de los clientes que necesitan o compran una bicicleta son aquellos que poseen entre 0 a 2 automóviles. Este dato puede ser muy importante a la hora de decidir a quién enviarle un mail con las promociones, aunque no es definitorio.



Compró bicicleta - Edad

Se percibe que ambas cajas son muy similares, por lo que se entiende que ambas variables no están estrechamente relacionadas. La caja del boxplot cuando compraron bicicletas es más corta, probablemente debido a que los datos están más concentrados y su rango intercuartílico es más pequeño. En ambos casos existe la presencia de outliers.

Se concluye que la decisión de comprar o no una bicicleta no depende de la edad del cliente.

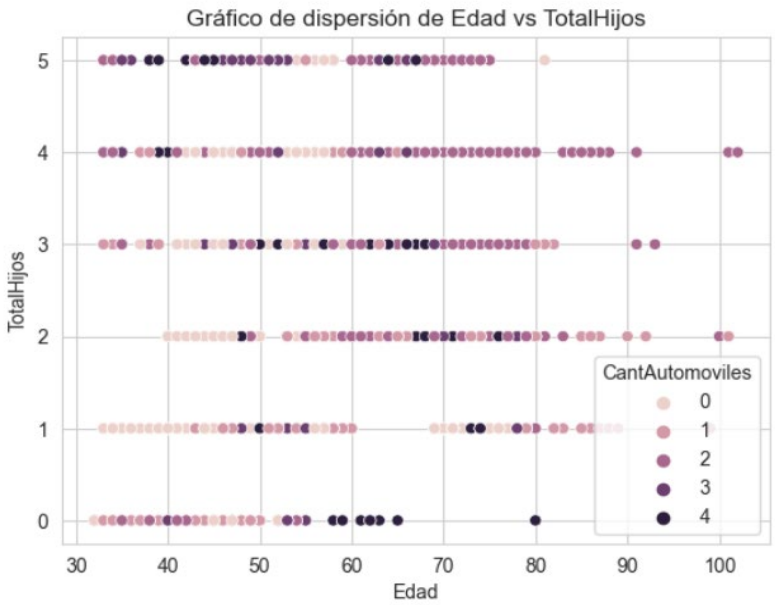


Análisis de diagramas de dispersión estratificados

Edad - Total de hijos

- **Cantidad de automóviles**

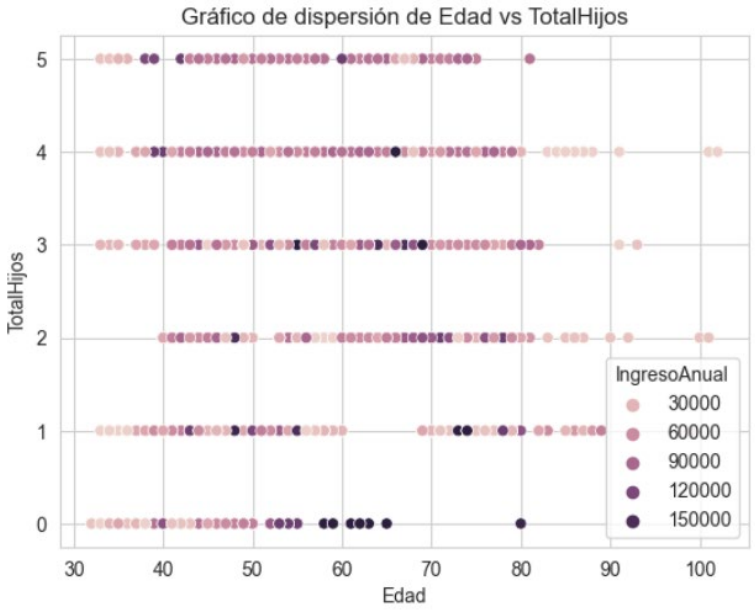
Analizando el gráfico vemos que los clientes sin hijos mayores a 55 años tienen 4 automóviles. A partir de un hijo se encuentra repartido, aunque se puede notar que a medida que tienen mayor cantidad de hijos aparecen clientes con un mayor número de automóviles.



- **Ingreso anual**

Gráfico bastante parejo, en cada nivel de hijos se encuentran todo tipo de nivel de ingresos. Se pueden ver varios clientes con el rango más alto de ingresos que no tienen hijos.

A partir de los 80 años los clientes tienen un ingreso anual menor, independientemente de la cantidad de hijos.

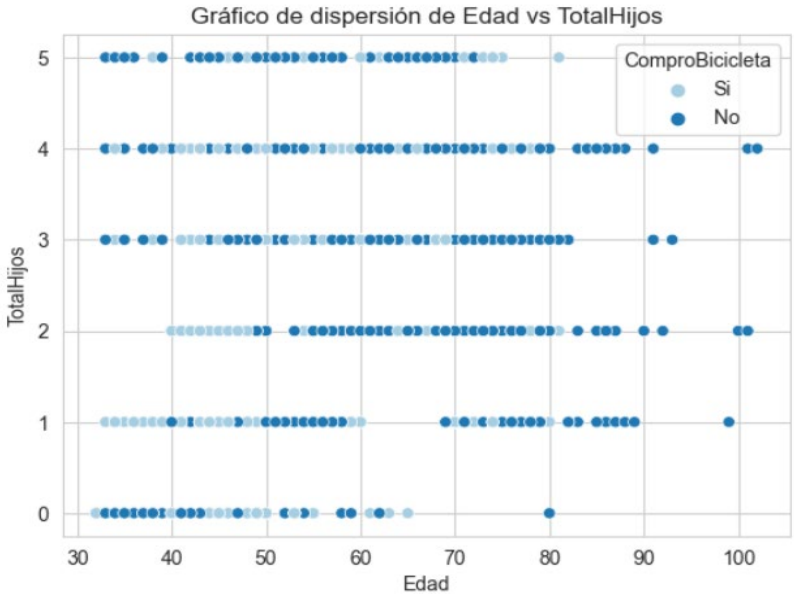


- **Compró bicicleta**

Quienes tienen 5 hijos compraron menos bicicletas comparado con el resto, seguidos muy de cerca por los clientes sin hijos.

En general se pueden ver dispersiones bastante similares, y se aprecia una cierta tendencia a que gran parte de los compradores de bicicletas son clientes menores de 50 años. Esto se evidencia aún más en los casos de 1 o 2 hijos donde prácticamente todos los que tienen entre 30 y 50 años compraron bicicletas en la empresa.

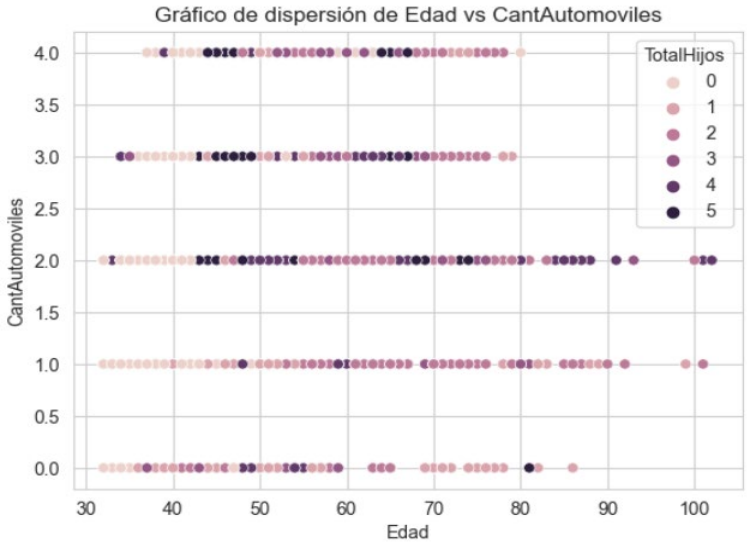
Podrían ser datos muy útiles a la hora de decidir a quienes enviarles la publicidad.



Edad - Cantidad de automóviles

- **Total de hijos**

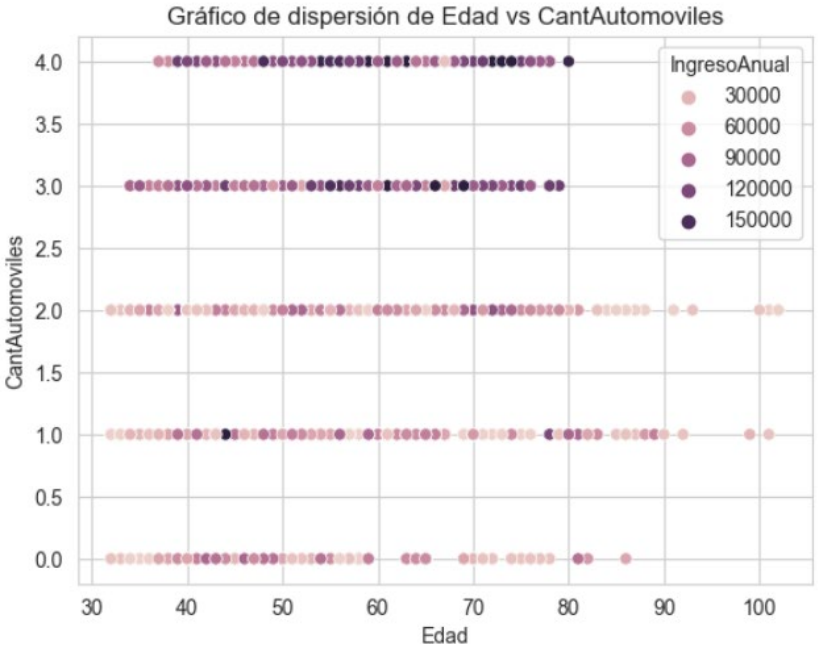
Para destacar se ven clientes con 5 hijos que no tienen automóviles. Además, se puede notar que hay una cantidad considerable de clientes con más de 1 automóvil y que no tienen hijos.



- **Ingreso anual**

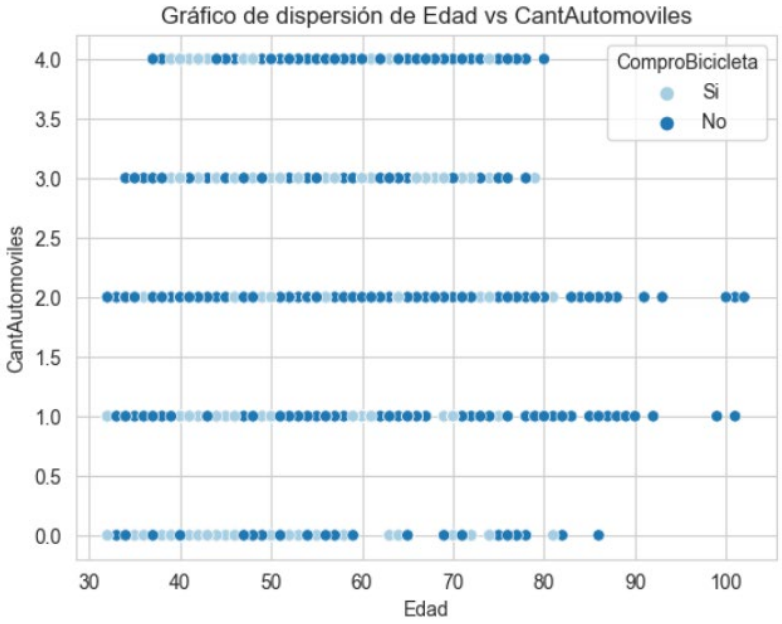
En el gráfico no se detectan demasiadas cosas a destacar, ya que es lógico pensar que quienes tienen mayor cantidad de automóviles veamos mayores ingresos.

Hay algunas excepciones a analizar detalladamente, en especial los que tienen el nivel más bajo de ingreso anual, pero tienen 3 o 4 automóviles registrados.



- **Compró bicicleta**

Si bien los clientes que no tienen automóviles son los que más compran bicicletas, en los otros valores de cantidad de automóviles no se ve una tendencia clara, ya que en proporción los clientes que menos compran bicicletas son los que tienen 2 vehículos, incluso menos que los que tienen 4.

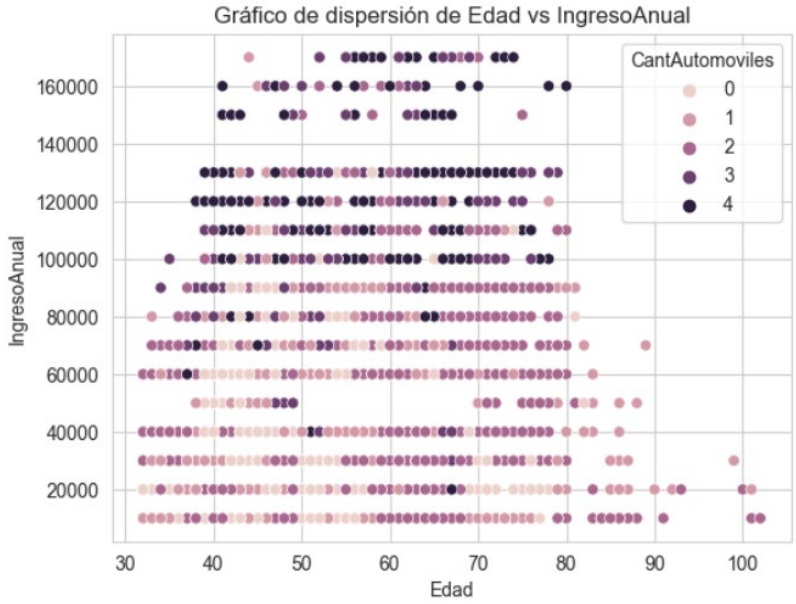


Edad - Ingreso anual

- **Cantidad de automóviles**

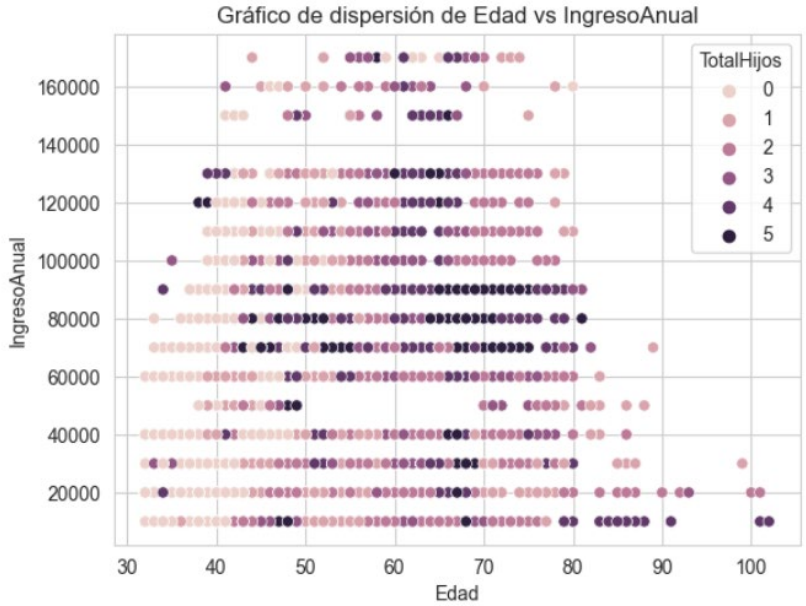
Se ve que a partir de un ingreso anual de $100.000 aumenta la cantidad de automóviles, encontrando a la mayoría de los clientes que poseen 4.

Las excepciones más notorias de algunos clientes son aquellos que tienen un ingreso menor a los $40.000.



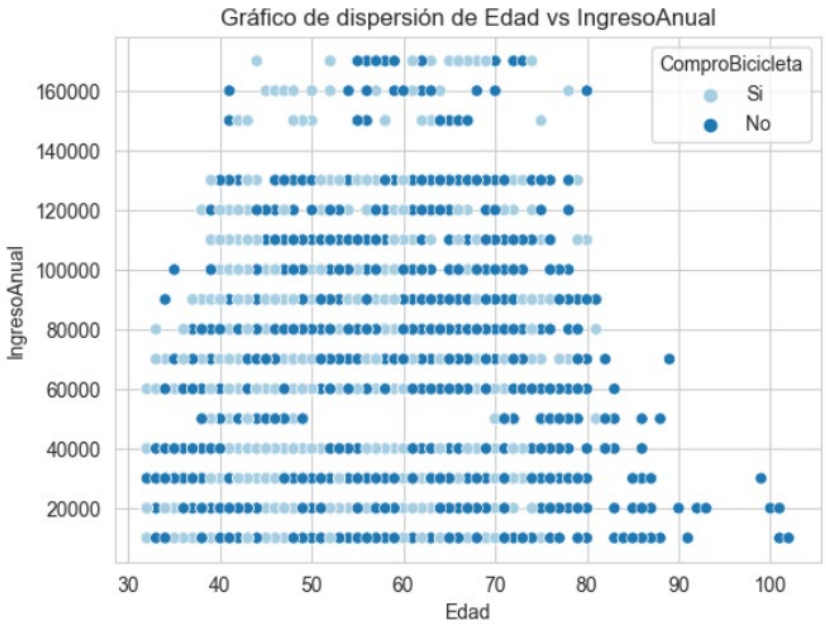
- **Total hijos**

Muestra una considerable concentración de clientes que tienen 5 hijos en el rango entre los $70.000 y los $90.000 de ingresos anuales, y en especial, en edades de entre 65 y 80 años. En los niveles más altos de ingresos anuales no hay un número tan elevado de hijos, en general.



- **Compró bicicleta**

No se nota una clara tendencia o patrón a destacar. En general la edad o el nivel de ingresos no afecta considerablemente a la compra o no de bicicletas, aunque es interesante destacar que dentro del nivel más alto de ingresos anuales hay una gran proporción de bicicletas compradas.



Calidad de los datos

Análisis de valores anómalos (outliers)

Outliers en ingresos anuales:

En los análisis hechos anteriormente se detectó la existencia de outliers en los ingresos anuales. Luego de realizar el análisis univariante y multivariante, se debe tomar una decisión sobre qué hacer con estos registros, si se decide eliminarlos o dejarlos.

Se determinó que el factor más importante se trata de la media o promedio, por lo que para tomar una decisión sobre qué hacer con estos valores anómalos, hay que tener en cuenta cómo y en qué medida afectará al promedio. Es por ello que lo primero que se realizó fue encontrar la media actual, cuyo valor es de $57.532,08.

Como siguiente paso se realiza un análisis de frecuencias, y se grafican el histograma y el boxplot de la variable para encontrar cuántos registros pertenecían a los outliers. Con esta información se obtuvieron 100 registros que presentaban valores anómalos de ingresos anuales. Posteriormente, se descartan de forma temporal los 100 registros, para así calcular nuevamente la media con los registros restantes (6290).

Utilizando las frecuencias absolutas se calculó el nuevo valor que adopta el promedio, que es de $55.915,74.

Con este nuevo valor, se realizó una resta entre ambos para ver cuál es la diferencia entre el valor de la media con outliers presente y sin estos, lo que nos dio una diferencia de $1.616,34.

Analizando tanto este número como así también los demás parámetros de los ingresos anuales, se sabe que su rango varía entre $10.000 y $120.000, por lo que la diferencia de $1.616,34 no es para nada significante.

Es por este motivo por el que se puede concluir que los valores anómalos no modifican significativamente a la media, y, sin embargo, eliminarlos puede llevar a perder datos importantes sobre otras variables.

Análisis de valores nulos

Dentro de nuestro conjunto de datos se contempló que existen 10 valores ausentes en los ingresos anuales. Para mejorar la calidad de nuestro conjunto de datos debemos solucionar este problema de los datos faltantes.

Sabemos que existen 3 formas de solucionarlo.

1) Eliminar los registros que presenten esos datos faltantes.
2) Imputar los datos faltantes con una constante, ya sea:
    a) Media
    b) Mediana
    c) Moda
3) Utilizar el modelo de regresión

La opción de eliminar los registros es una buena idea, ya que sólo son 10 los registros que presentan valores nulos dentro de los 6400 totales, por lo que no debería afectar significativamente el conjunto de datos. Sin embargo, no se considera quizás el método óptimo.

Pensar en utilizar el modelo de regresión para sólo predecir 10 valores es un proceso ineficiente, ya que se trata de un método que no es fácil de implementar y lleva más tiempo que las otras opciones.

Por lo tanto, la conclusión es que la opción óptima para solucionar el problema de los datos faltantes es la de imputar dichos valores con una constante. Para ello, debemos decidir qué constante vamos a utilizar.

Como la variable es del tipo numérica se descarta la moda, al utilizarse esta para variables no numéricas o cualitativas. Esto hace que se deba elegir entre la mediana y la

media. Informándose sobre cuál sería la mejor decisión y se sabe que para casos donde la variable hace referencia a ingresos o sueldos, no es conveniente utilizar la media.

Por todo lo detallado anteriormente, se decidió completar los 10 valores faltantes de los ingresos anuales con la mediana.

Sin embargo, realizando un análisis de mayor profundidad, es perfectible que la opción óptima sería agrupar las medianas a través de alguna otra variable y así intentar acercarnos lo más posible al valor real que podría tener ese cliente. Es por ello por lo que se agrupó a los clientes por ocupación y se calculó la mediana del ingreso anual de cada uno de los posibles valores (profesional, obrero especializado, gestión, administrativo y obrero), asignándole a cada cliente el ingreso anual que corresponde a la mediana de su ocupación.

Fase de modelado

En esta fase se realizan varios tipos de modelado, utilizando diferentes técnicas y modelos para predecir correctamente a quien conviene enviarle un mail con las promociones de la empresa. Finalmente se elige el modelo más confiable basándose en su matriz de confusión.

Árbol de decisión

Un árbol de decisión es una técnica utilizada en el modelado de minería de datos y el aprendizaje automático. Se trata de un modelo predictivo que utiliza una estructura de árbol para tomar decisiones basadas en múltiples atributos o características de un conjunto de datos.

En un árbol de decisión, cada nodo interno del árbol representa una característica o atributo del conjunto de datos, y las ramas que salen de ese nodo representan las posibles respuestas o valores de esa característica. A medida que se desciende por el árbol, se toman decisiones basadas en los valores de las características hasta llegar a las hojas, que representan las predicciones o clasificaciones finales.

El proceso de construcción de un árbol de decisión implica dividir recursivamente el conjunto de datos en subconjuntos más pequeños con base en los atributos más relevantes. El objetivo es maximizar la homogeneidad dentro de cada subconjunto y maximizar la heterogeneidad entre los subconjuntos.

Existen diferentes parámetros que se pueden modificar para obtener así diferentes resultados y poder lograr el árbol que mejor prediga y evitar los errores más comunes como el sobreajuste o sobre entrenamiento. A continuación, se definen los 3 parámetros más importantes que se irán modificando:

- **Profundidad:** Se refiere a cuándo el árbol va a dejar de separar los datos. Es el número de hijos máximo.
- **Criterio:** Medida utilizada para evaluar la calidad de una división o partición en el árbol
- **Proporción de entrenamiento**: Es la proporción que se utiliza para separar los datos en conjunto de entrenamiento y conjunto de predicción.

Profundidad: 6 - Criterio: Information gain - Proporción: 70/30

accuracy: 70.26%

	true Si	true No	class precision
pred. Si	430	244	63.80%
pred. No	327	919	73.76%
class recall	56.80%	79.02%	

En el gráfico se muestra la matriz de confusión del árbol. En ella se observa la precisión general (70,26%) y luego la precisión para cada valor de nuestra variable a predecir.

En este caso, de 757 registros que comprarían bicicleta, el árbol predijo que solo 430 lo harían. Lo cual nos deja una efectividad del 56,80%. Para los clientes que no comprarían, existe una eficacia de 79,02%. Es muy importante recordar que para el negocio es preferible enviar innecesariamente un correo a una persona que no resulte comprador, y no perder un potencial cliente por no enviarle la publicidad. Por lo tanto, el valor que más interesa aumentar es el porcentaje de aciertos de la columna “Si”.

Profundidad: 8 - Criterio: Gain ratio - Proporción: 80/20

accuracy: 65.94%

	true Si	true No	class precision
pred. Si	262	193	57.58%
pred. No	243	582	70.55%
class recall	51.88%	75.10%	

Se puede apreciar que claramente esta configuración del árbol es incluso peor que la anterior.

Profundidad: 10 - Criterio: Information gain - Proporción: 70/30

accuracy: 76.35%

	true Si	true No	class precision
pred. Si	549	246	69.06%
pred. No	208	917	81.51%
class recall	72.52%	78.85%	

En este caso se observa que la precisión es mayor y más pareja entre ambos posibles resultados de la variable a predecir, lo que ayuda a saber que el modelo no está sobre ajustado.

Profundidad: 10 - Criterio: Information gain - Proporción: 80/20

accuracy: 78.44%

	true 1	true 0	class precision
pred. 1	378	149	71.73%
pred. 0	127	626	83.13%
class recall	74.85%	80.77%	

Finalmente se decidió utilizar estos valores de los parámetros, al tratarse del que más confiabilidad otorga. Sin embargo, aún se deben tener en cuenta otros modelos como KNN y LDA para verificar cual es el mejor.

KNN (*K* vecinos más cercanos)

La técnica de KNN, o también llamado como método de *K* vecinos más próximos, es un algoritmo de machine learning de aprendizaje supervisado. Es un clasificador robusto y versátil que a menudo se usa como un punto de referencia para clasificadores más complejos como las redes neuronales artificiales y vectores de soporte.

A pesar de su simplicidad, KNN puede superar a los clasificadores más potentes y se usa en una variedad de aplicaciones tales como pronósticos económicos, compresión de datos y genética.

A diferencia de los árboles de decisión es un método retardado, no crea un modelo general. Para clasificar un nuevo caso, buscan entre los ejemplos anteriores almacenados el más parecido y resuelven como en esa ocasión.

Se trabaja con un parámetro *k* que determina la cantidad de vecinos cercanos con los cuales se comparará la nueva observación. Es importante seleccionar un valor de *k* acorde a los datos para tener una mayor precisión en la predicción.

Consiste en clasificar un nuevo caso en función de su distancia con los casos vecinos, al momento del análisis los *k* datos más cercanos al valor que se desea predecir serán la solución.

Hay distintas formas para calcular las distancias:

- **Euclídea:** Toma la longitud de la recta que une dos puntos en el espacio, aunque tiene el problema de depender de la unidad de medida de las variables al no estar normalizada.
- **Euclídea ponderada:** Se toma una matriz diagonal para estandarizar variables que puede tener las desviaciones estándares en la diagonal principal.
- **Mahalanobis:** Recomendada para variables correlacionadas, utiliza matriz de covarianzas.
- **Manhattan:** Recorre un camino en forma de "zigzag" en lugar de hacerlo en diagonal.
- **Chebyshev:** Calcula la discrepancia más grande en alguna de las dimensiones.

En variables categóricas la distancia euclídea entre dos elementos será 0 si ambos elementos coinciden en el valor del atributo o 1 en caso de que tengan distinto valor

Analizando los posibles K

El primer paso para modelar con esta técnica es averiguar cuántos grupos conviene realizar para lograr la mayor confiabilidad posible. Para ello se dividen los datos entre entrenamiento y validación, que son aquellos con los cuales se verifica que tan bueno es el modelo.

En una primera instancia, se programó el método en lenguaje Python para facilitar la decisión de qué valor de k era el óptimo. Esto arrojó la siguiente tabla:

Número de K	Accuracy de KNN train	F1 de KNN train	Accuracy de KNN test	F1 de KNN test
1	0.76	0.75	0.69	0.68
2	0.77	0.74	0.71	0.67
3	0.80	0.78	0.71	0.69

4	0.80	0.77	0.72	0.69
5	0.81	0.79	0.72	0.70
6	0.81	0.79	0.72	0.70
7	0.81	0.79	0.72	0.70
8	0.81	0.79	0.72	0.70
9	0.81	0.80	0.73	0.71
10	0.81	0.79	0.72	0.70

Como se observa, la precisión empieza aumentando desde ‘K=1’ hasta ‘K=9’ pero luego en ‘K=10’ disminuye. Por lo tanto, el valor de ‘K’ que proporcionará el modelo más confiable y preciso es el de ‘K=9’. El próximo paso sería aplicar el modelo de KNN con ‘K=9’ y observar la matriz de confusión para determinar qué tan confiable y preciso es el modelo.

Sin embargo, existe un problema con estos valores de k. Todos ellos fueron generados sólo teniendo en cuenta las variables numéricas. Es por ello por lo que a continuación se implementó el modelo KNN mediante la herramienta “RapidMiner”, que proporciona una mayor simplicidad a la hora de analizar cualquier tipo de variable, sea esta numérica o no.

El único inconveniente aquí presente es que se debe establecer manualmente el valor de k, hasta llegar a uno que se crea que es el mejor. Luego de varios intentos, se determinó que dicho valor es k = 4, lo que generó la siguiente matriz de confusión:

accuracy: 76.02%

	true 1	true 0	class precision
pred. 1	364	166	68.68%
pred. 0	141	609	81.20%
class recall	72.08%	78.58%	

La precisión general de este modelo es de alrededor de un 76%, teniendo la predicción deseada por el negocio la menor tasa de acierto, con un 72% aproximadamente. Es este el motivo por el cual, hasta el momento, el modelo de árboles resulta más preciso y confiable.

LDA (Análisis discriminante de datos)

El análisis discriminante de datos es una técnica estadística utilizada en el campo del aprendizaje automático y la estadística multivariante. Su objetivo principal es encontrar una combinación lineal de características (variables independientes) que permita diferenciar o discriminar entre diferentes clases o categorías en un conjunto de datos.

El LDA busca una proyección de las variables que maximice la distancia entre las medias de las clases y minimice la variabilidad dentro de cada clase. A través de esta proyección, se puede reducir la dimensionalidad del conjunto de datos y, al mismo tiempo, mantener la mayor cantidad de información relevante para la discriminación entre clases.

El proceso de LDA implica calcular las matrices de dispersión entre clases y dentro de las clases, y luego encontrar los vectores propios (auto vectores) asociados con los valores propios más grandes de la matriz resultante. Estos auto vectores representan la dirección óptima para la proyección de las variables.

El LDA se utiliza comúnmente en problemas de clasificación, donde se desea predecir la pertenencia a una o varias clases en función de un conjunto de características. También se puede utilizar como una técnica de reducción de dimensionalidad para visualizar o comprimir datos en problemas con múltiples variables.

Análisis de resultados obtenidos

Una vez realizado el análisis correspondiente, se obtuvo la siguiente matriz de confusión:

Resultados de clasificación[a]

			Pertenencia a grupos pronosticada		
		ComproBicicleta	No	Si	Total
Original	Recuento	No	2352	1524	3876
		Si	947	1577	2524
	%	No	60.7	39.3	100.0
		Si	37.5	62.5	100.0

a. 61.4% de casos agrupados originales clasificados correctamente.

Se puede observar una precisión del modelo de aproximadamente un 61%, siendo la predicción más buscada (quiere bicicleta) de un 63%. Si bien la tasa de acierto en la predicción de los que quieren bicicletas y los que no son muy similares, se está frente a una muy baja proporción de ellos. Por este motivo, el modelo aquí presente no es el óptimo a utilizar para evaluar los potenciales clientes.

Fase de evaluación

Conclusiones

Después de realizar un exhaustivo análisis comparativo entre tres modelos diferentes: KNN, LDA y Árboles de Decisión, se ha llegado a la conclusión de que el modelo de Árboles de Decisión supera a los otros dos en términos de precisión y rendimiento.

Al evaluar los resultados obtenidos, se observa que tanto el modelo KNN como el modelo LDA mostraron una precisión general inferior en comparación con el modelo de Árboles de Decisión. Esto implica que, en general, el modelo de Árboles de Decisión es más efectivo para realizar predicciones y clasificar correctamente los potenciales clientes.

Además, el modelo de Árboles de Decisión tiene la capacidad de manejar tanto variables numéricas como categóricas, lo cual es una ventaja significativa en comparación con el modelo KNN, que funciona mejor con variables numéricas. Esto proporciona una mayor flexibilidad y adaptabilidad al modelo de Árboles de Decisión, lo que puede explicar su mayor rendimiento.

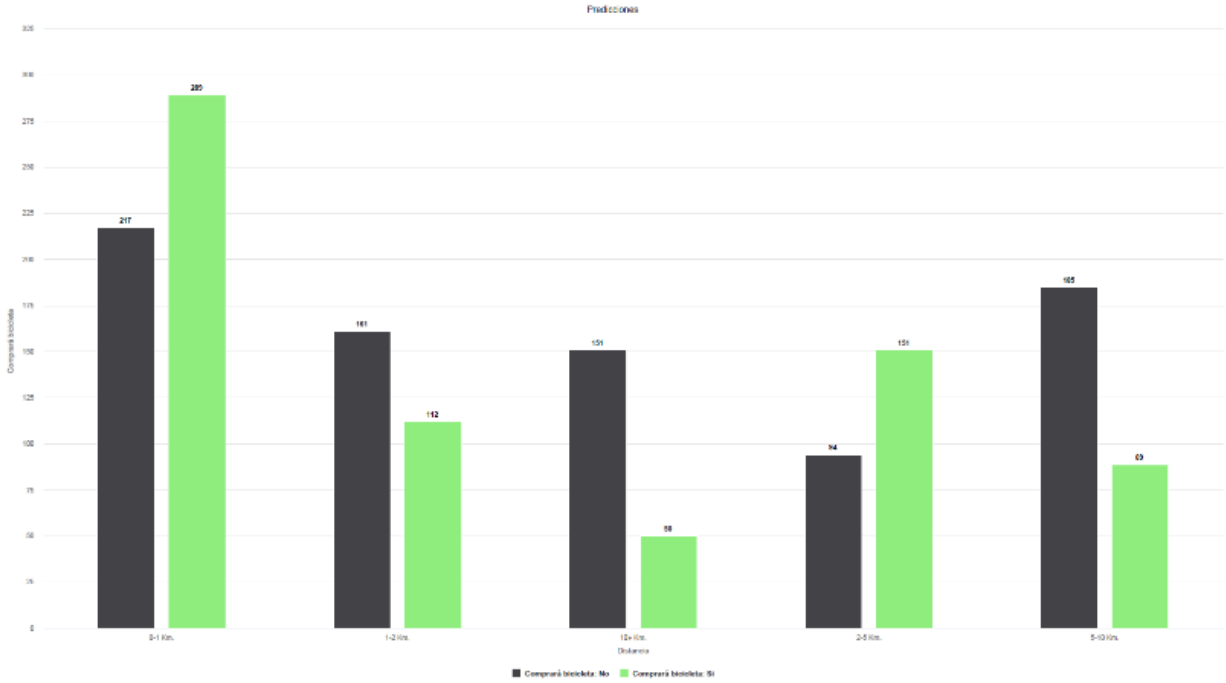
Otro aspecto importante para considerar es el costo computacional de cada modelo. Si bien el modelo de Árboles de Decisión puede requerir más recursos computacionales para construir y evaluar el árbol, una vez que se ha construido, su proceso de predicción es rápido y eficiente. Por otro lado, el modelo KNN y el modelo LDA pueden tener un mayor costo computacional durante la fase de predicción.

En conclusión, los resultados obtenidos respaldan firmemente la elección del modelo de Árboles de Decisión como el mejor en términos de precisión y rendimiento. Su capacidad para capturar relaciones complejas, su flexibilidad para manejar diferentes tipos de variables y su eficiencia computacional lo convierten en una opción destacada para la tarea de elección de potenciales clientes en este estudio.

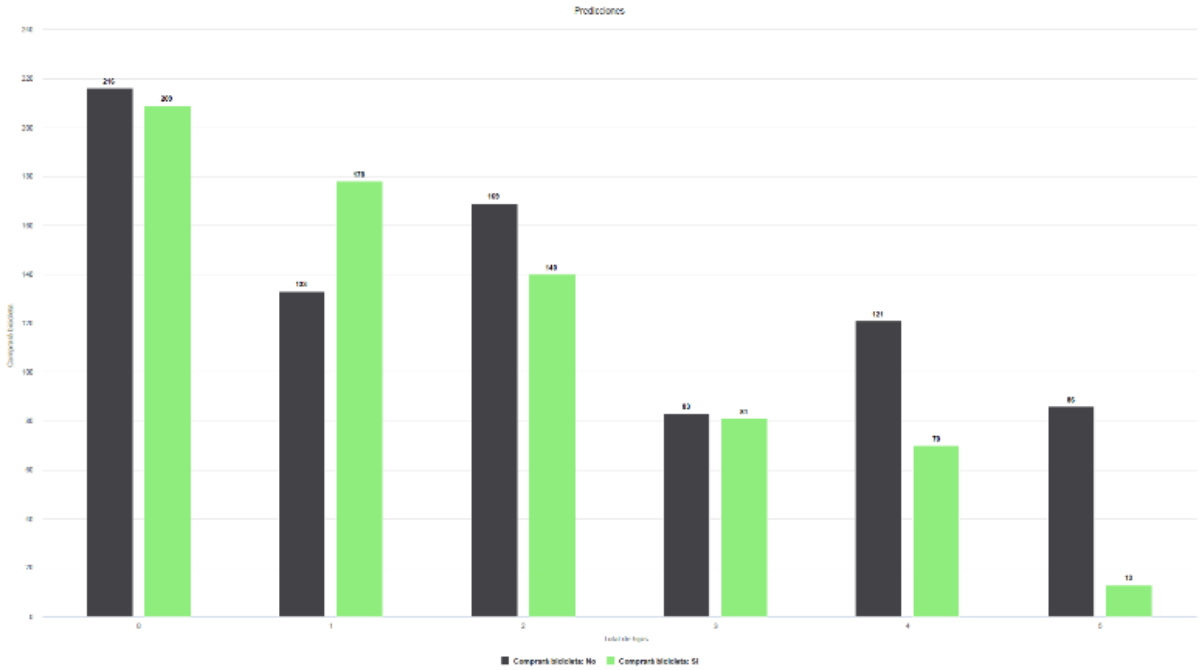
Resultados

A continuación, se detallan mediante gráficos los resultados obtenidos utilizando el método de predicción de árboles de decisión. Se cuenta con un total de 1500 clientes, de los cuales 808 no comprarán bicicleta, mientras que 692 si lo harán.

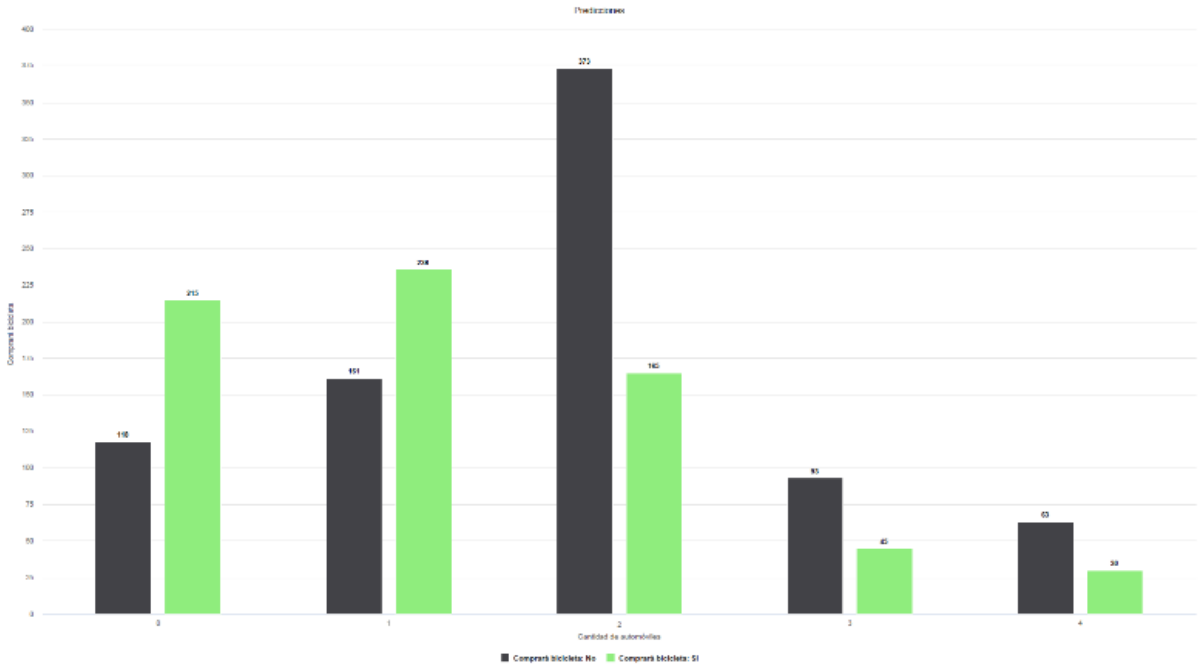
Distancia



Total de hijos



Cantidad de automóviles



Ingreso anual