TTMZ0411-93-PROYECTO MBD
CAPSTONE PROJECT

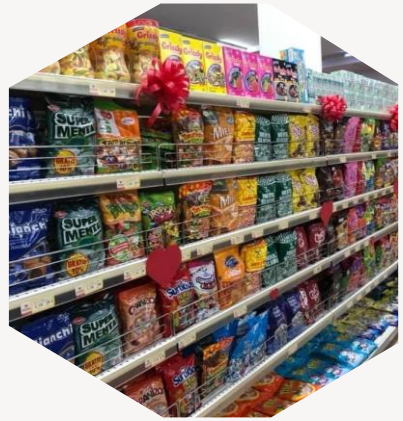
ANÁLISIS DE EFECTIVIDAD DE CAMPAÑAS EN EMPRESAS DE CONSUMO MASIVO, MEDIANTE EL USO DE MODELOS PROBABILISTICOS DE CLASIFICACIÓN BINARIA

Willy Lema M

24-08-2023

Introducción

- Finalidad de realizar un análisis de la efectividad de campañas, (marketing o promocionales)
- Empresas de consumo masivo
- Modelos probabilísticos de predicción binaria



Datos

- Repositorio Kaggle
(Registros anonimizados, demográficos, históricos y preferencias de consumo)
- Información de aceptación o rechazo de promoción
- Realizar un análisis de correlación de las variables con la variable de respuesta.
- Modelos de predicción algorítmica como son regresión logística, árbol de decisión, random forest y extra trees

Objeto de estudio

- El objeto de estudio son las respuestas que los consumidores tienen ante campañas de marketing o promocionales.

- Se pretenden comprender cuáles son los factores que más influyen.
- Identificar patrones y características de las personas que acceden.
- Identificar los mejores modelos para este tipo de predicción.

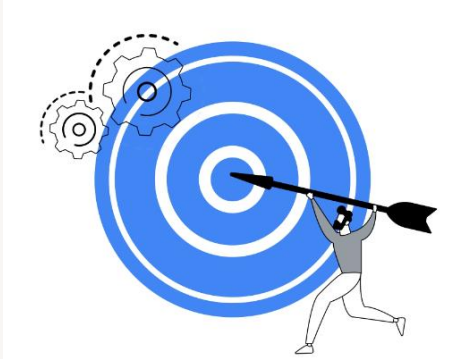
Planteamiento del problema

- En el Ecuador las empresas, casi no han aplicado metodologías de segmentación de mercado para los productos que ofrecen por falta de conocimiento o simplemente por lo que representa realizar estudios como estos.

El sector de los supermercados tampoco es la excepción, también existen empresas de menor tamaño en este sector y en otros que no ocupan este tipo de metodologías haciendo que sus operaciones no siempre sean eficientes y rentables.

El problema es que debido a la falta de identificación del mercado objetivo específico para las empresas y correcta segmentación de mercados, estas campañas no siempre podrían tener un efecto positivo en el objetivo de mejorar los ingresos o las ventas de una empresa.

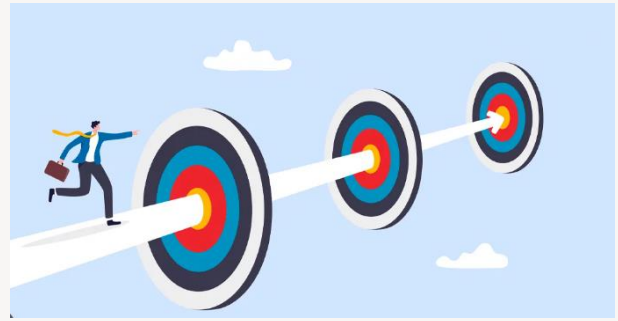
OBJETIVOS



Objetivo general

Realizar un pronóstico del éxito o rechazo de una campaña de promoción en una empresa de consumo masivo, con el fin de determinar la correlación entre variables y determinar los principales factores que influyen a la variable de respuesta. Entender la metodología y analizar los modelos de predicción usados como una herramienta que puede ser aplicable a otros proyectos similares en otras ramas productivas.

OBJETIVOS



Objetivos específicos

- Determinar a través de un análisis descriptivo los principales factores que en el caso de estudio influyen en la aceptación o rechazo por parte del cliente, definir características demográficas, socioeconómicos y específicos de los consumidores del estudio.

- Pronosticar la aceptación de los consumidores frente a la campaña realizada por la empresa en estudio, utilizando modelos de predicción clasificatoria binaria y compararlos.

- Establecer soluciones y sugerencias estratégicas que la empresa pueda llevar a cabo frente a los hallazgos del estudio de predicción de éxito o rechazo de la campaña realizada para sus clientes en el presente estudio.

- Realizar un análisis gráfico detallado para el caso de la empresa que permita encontrar patrones y rasgos de sus consumidores para los objetivos comerciales de la misma.

Selección de la base de datos

La base de datos original contiene 2240 registros clasificados en 22 campos o variables.

Variable	Descripción de campo	Tipo de Variable	Clasificación
Id	ID único de cada cliente.	Numérica / int64	Independiente
Año_Nacimiento	Edad del cliente.	Numérica / int64	Independiente
Niv_Educación	Nivel de educación del cliente.	Object	Independiente
Estado_Civil	Estado civil del cliente.	Object	Independiente
Ingresos	Ingresos familiares anuales del cliente.	Numérica / float 64	Independiente
N_Niños	Número de niños pequeños en el hogar del cliente.	Numérica / int64	Independiente
N_Adolescentes	Número de adolescentes en el hogar del cliente.	Numérica / int64	Independiente
Fecha_Cliente	Fecha de alta del cliente en la empresa.	Object	Independiente
Ult_Compra	Número de días desde la última compra.	Numérica / int64	Independiente

Selección de la base de datos

La base de datos original contiene 2240 registros clasificados en 22 campos o variables.

C_Vinos	La cantidad gastada en productos vitivinícolas en los últimos 2 años.	Numérica / int64	Independiente
C_Frutas	La cantidad gastada en productos de frutas en los últimos 2 años.	Numérica / int64	Independiente
C_Carnes	La cantidad gastada en productos cárnicos en los últimos 2 años.	Numérica / int64	Independiente
C_ProdsMar	La cantidad gastada en productos pesqueros en los últimos 2 años.	Numérica / int64	Independiente
C_Dulces	Cantidad gastada en productos dulces en los últimos 2 años.	Numérica / int64	Independiente
C_PremiumProds	La cantidad gastada en productos de oro en los últimos 2 años.	Numérica / int64	Independiente
N_CompPromos	Número de compras realizadas con descuento.	Numérica / int64	Independiente
N_CompWeb	Número de compras realizadas a través de la web de la empresa.	Numérica / int64	Independiente
N_CompCatalogo	Número de compras realizadas por catálogo (compra de productos para enviar por correo).	Numérica / int64	Independiente
N_CompTiendas	Número de compras realizadas directamente en tiendas.	Numérica / int64	Independiente
N_VisitasWebMes	Número de visitas al sitio web de la empresa en el último mes.	Numérica / int64	Independiente
Reclamo	1 si el cliente se quejó en los últimos 2 años.	Numérica / int64	Independiente
Respuesta	1 si el cliente aceptó la oferta en la última campaña, 0 en caso contrario.	Numérica / int64	Dependiente

Preprocesamiento y Limpieza

Detección datos inconsistentes

```
dataset.isna().sum()
```

Id	0
Año_Nacimiento	0
Niv_Educación	0
Estado_Civil	0
Ingresos	24
N_Niños	0
N_Adolescentes	0
Fecha_Cliente	0
Ult_Compra	0
C_Vinos	0
C_Frutas	0
C_Carnes	0
C_ProdsMar	0
C_Dulces	0
C_PremiumProds	0
N_CompPromos	0
N_CompWeb	0
N_CompCatalogo	0
N_CompTiendas	0
N_VisitasWebMes	0
Reclamo	0
Respuesta	0

dtype: int64

```
dataset.isna().sum()
```

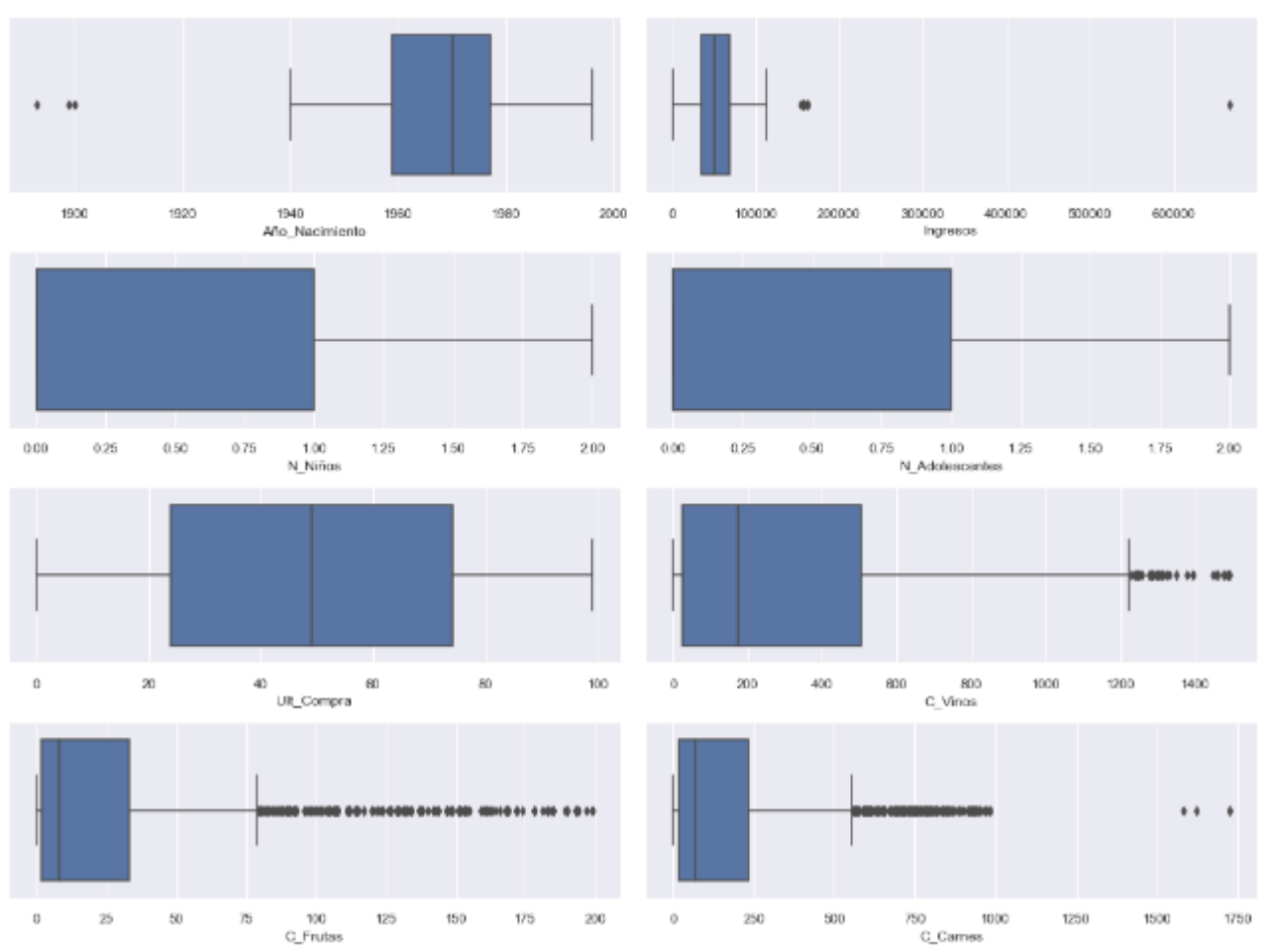
Id	0
Año_Nacimiento	0
Niv_Educación	0
Estado_Civil	0
Ingresos	0
N_Niños	0
N_Adolescentes	0
Fecha_Cliente	0
Ult_Compra	0
C_Vinos	0
C_Frutas	0
C_Carnes	0
C_ProdsMar	0
C_Dulces	0
C_PremiumProds	0
N_CompPromos	0
N_CompWeb	0
N_CompCatalogo	0
N_CompTiendas	0
N_VisitasWebMes	0
Reclamo	0
Respuesta	0

dtype: int64

Sin datos inconsistentes

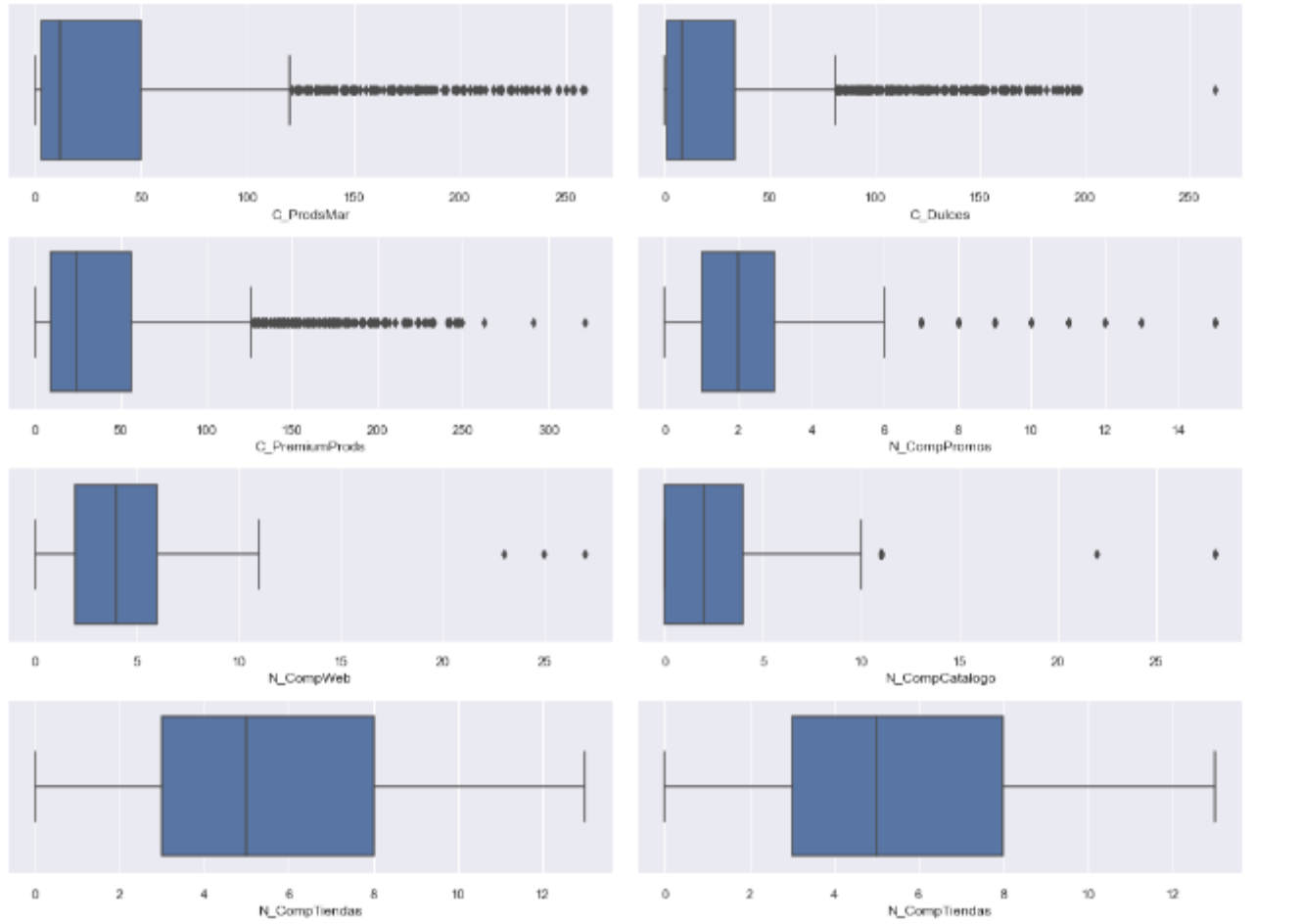
Preprocesamiento y Limpieza

Análisis de Outliers



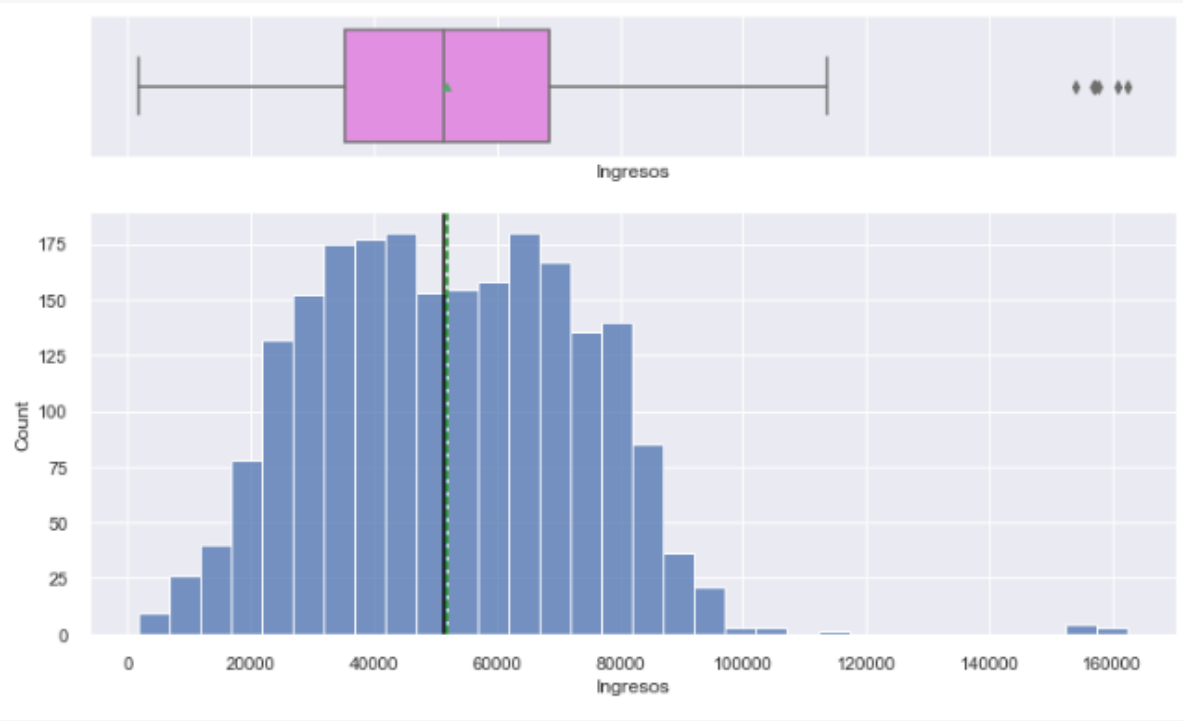
Preprocesamiento y Limpieza

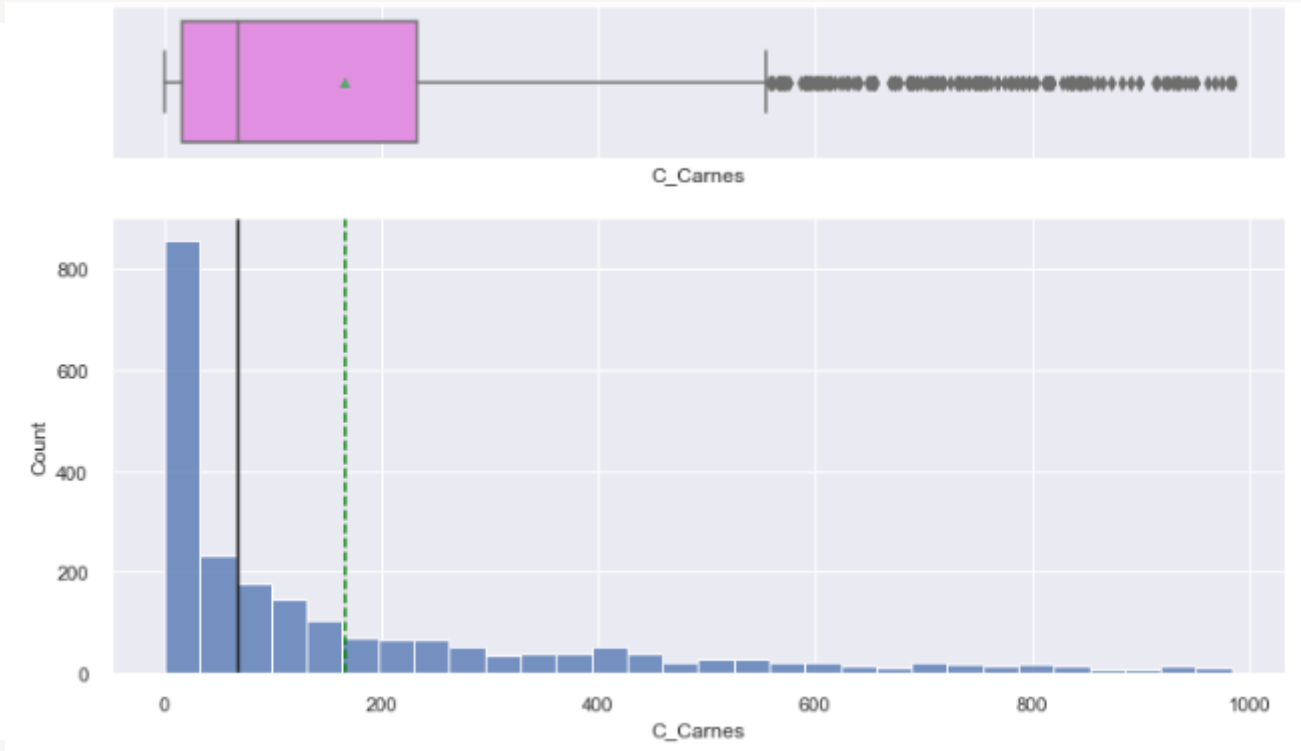
Análisis de Outliers



Análisis Exploratorio de Datos

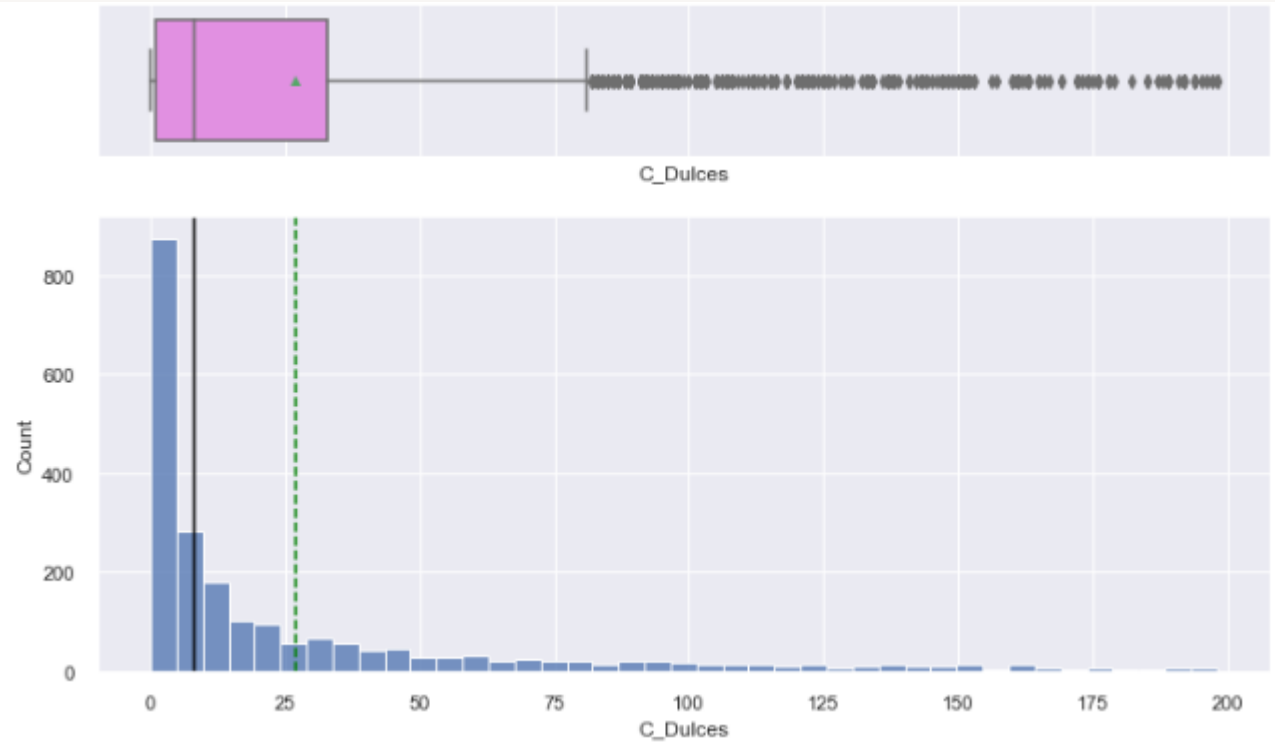
Análisis Univariable

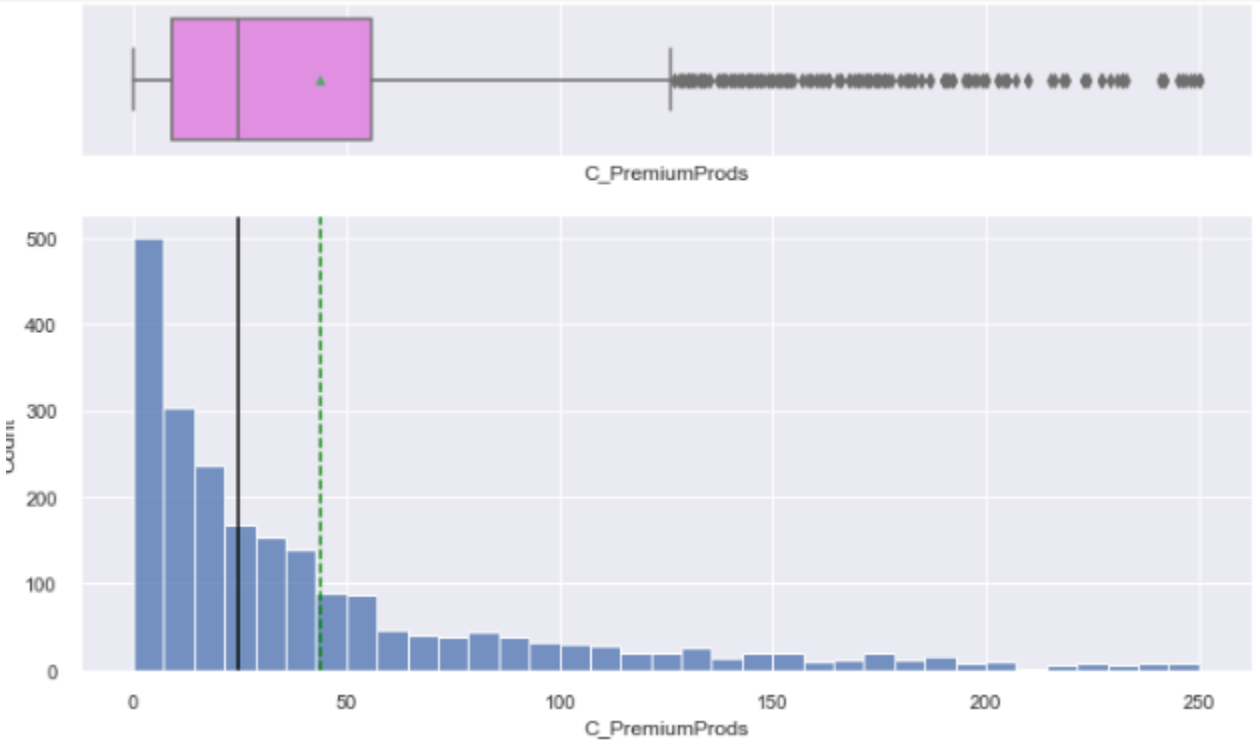




Análisis Exploratorio de Datos

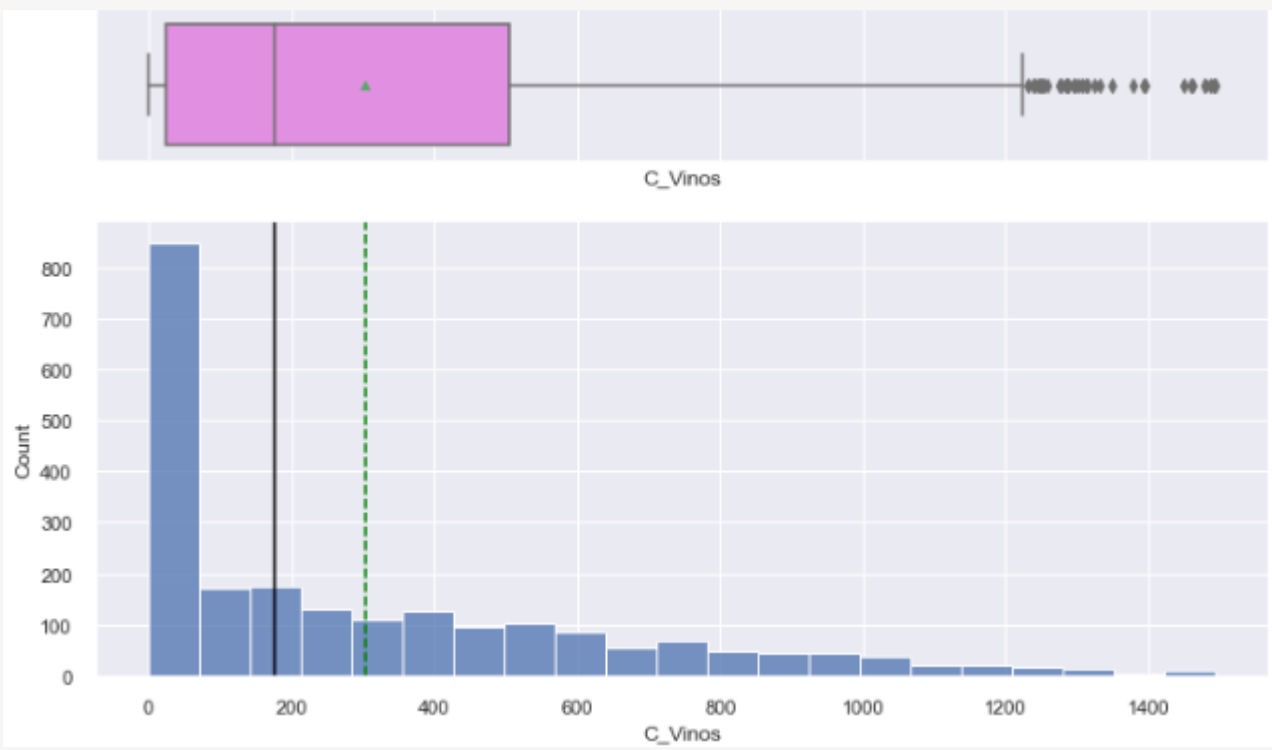
Análisis Univariable

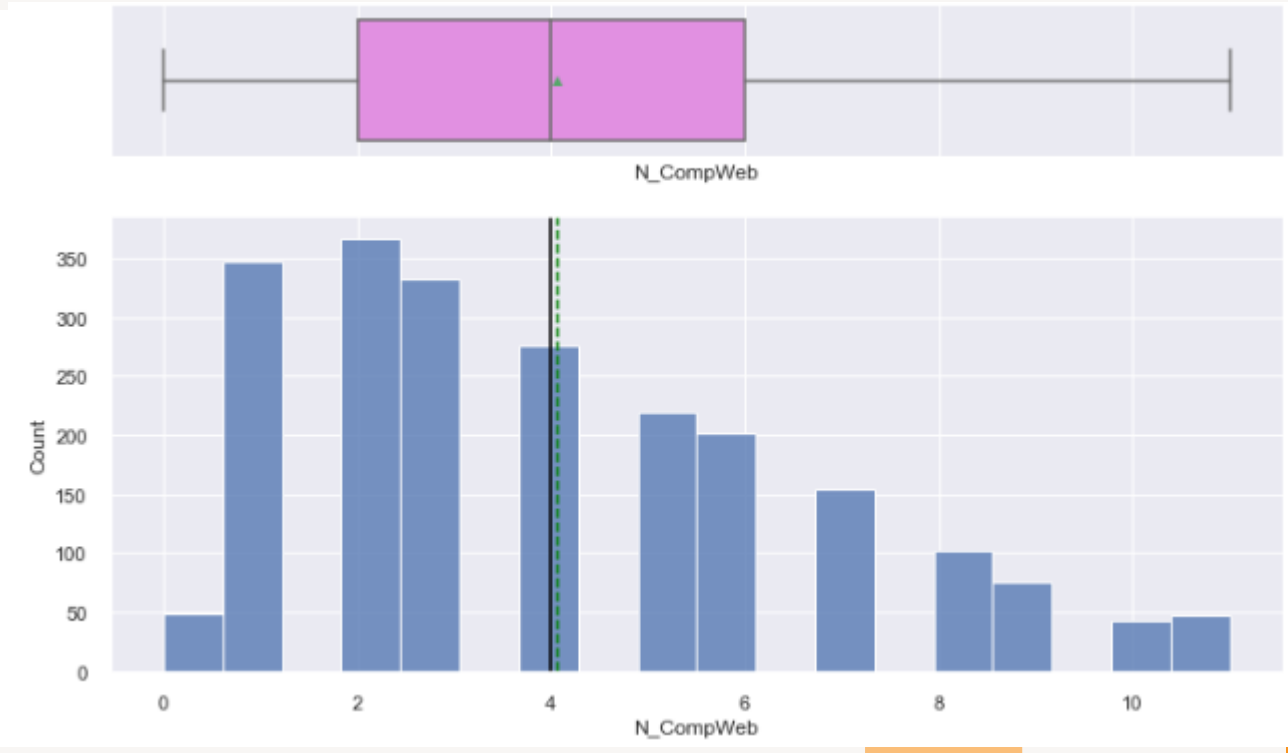




Análisis Exploratorio de Datos

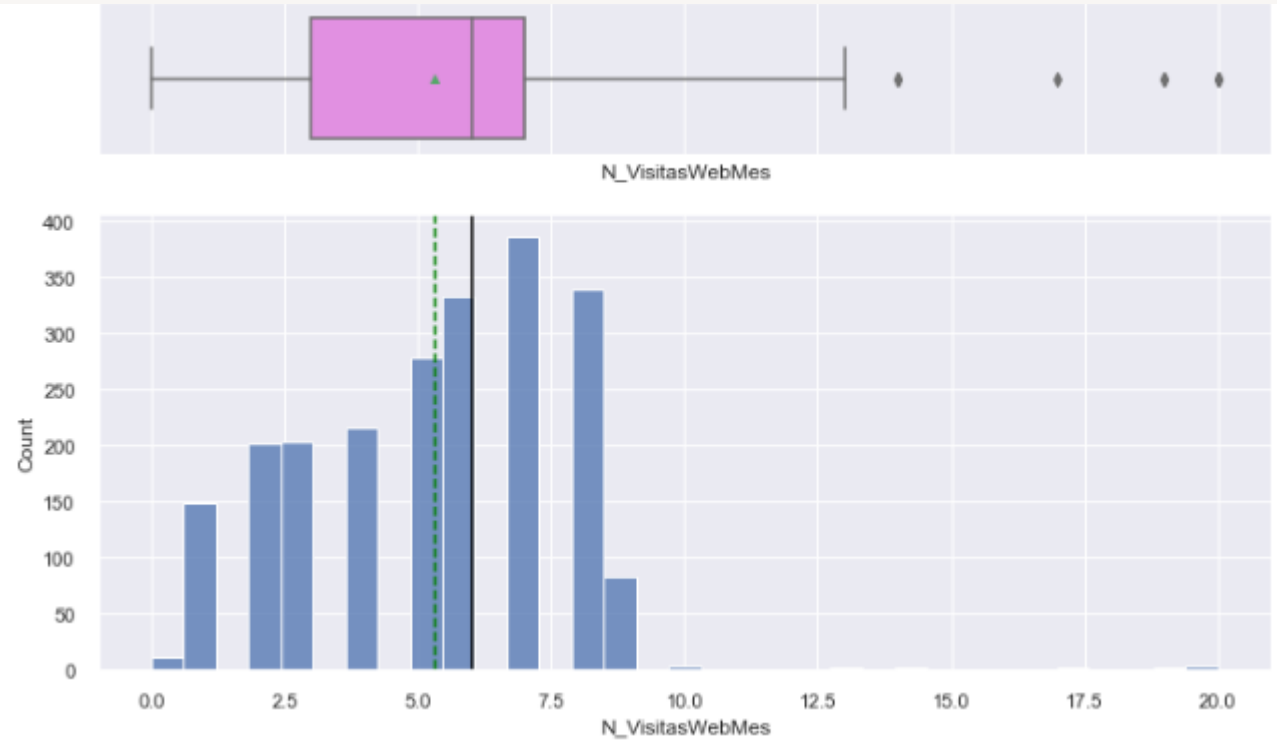
Análisis Univariable

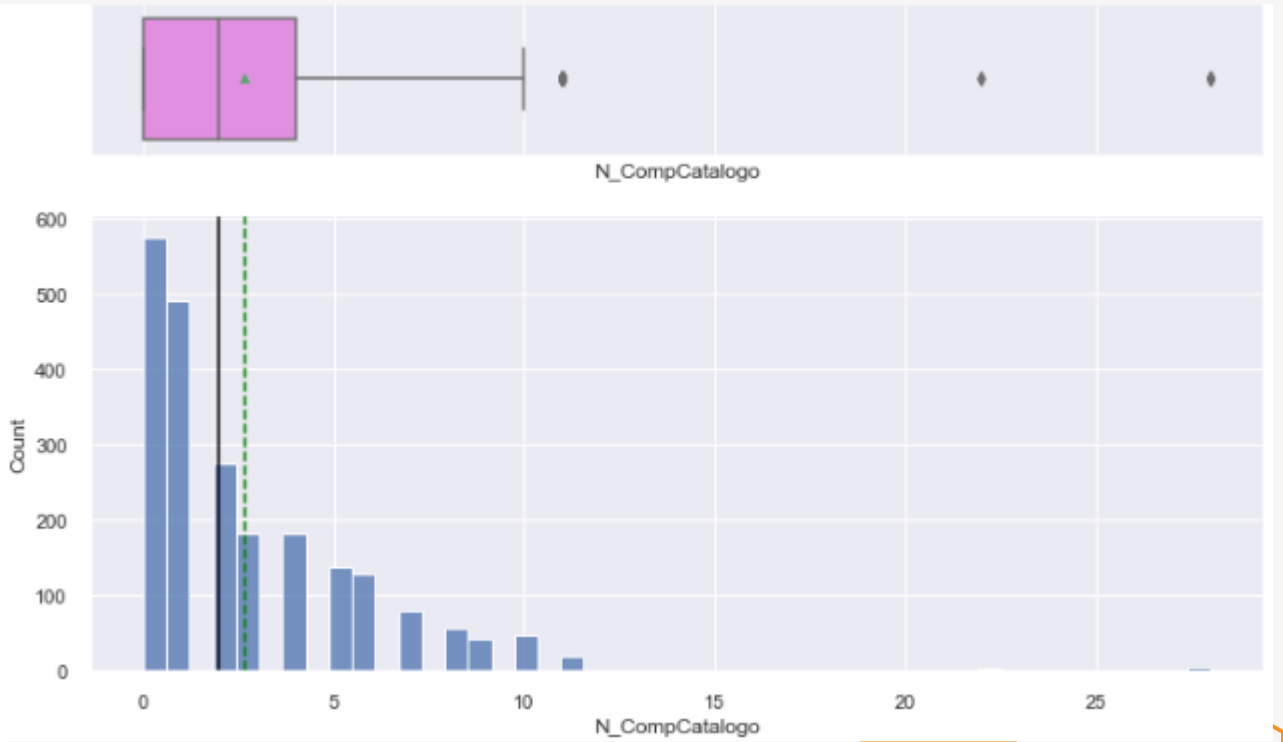




Análisis Exploratorio de Datos

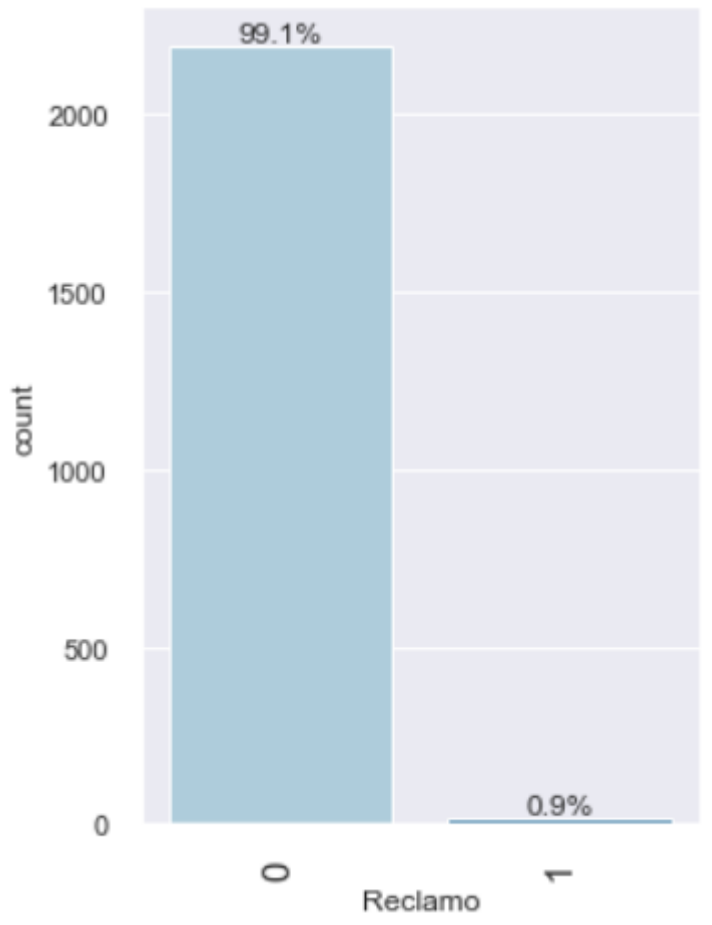
Análisis Univariable

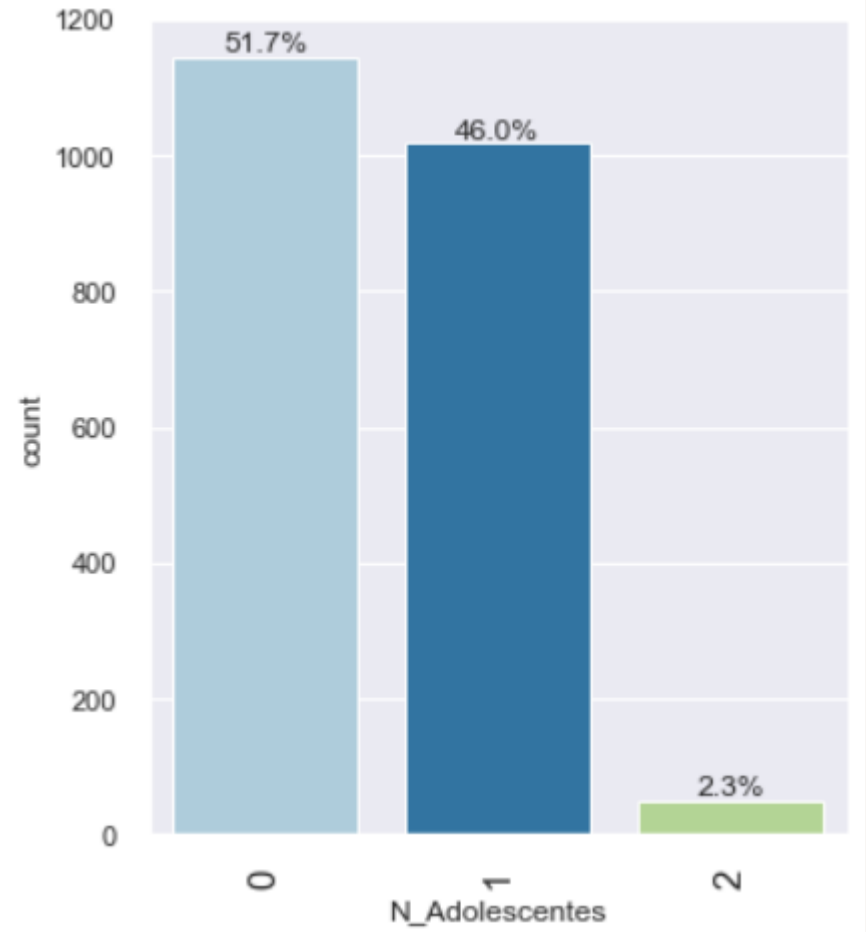




Análisis Exploratorio de Datos

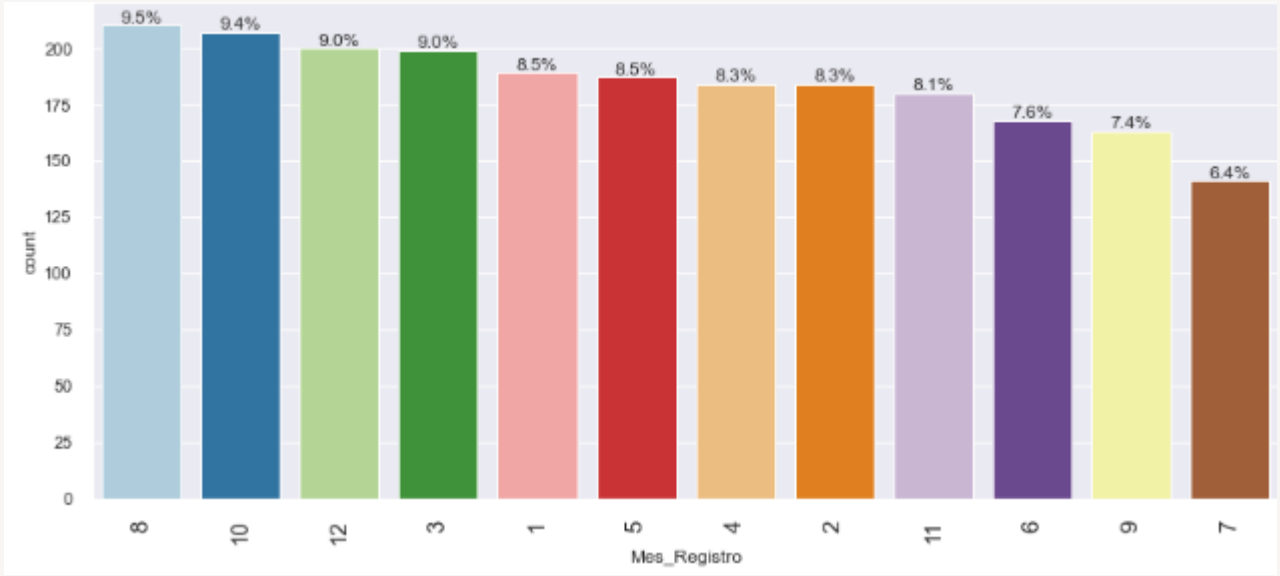
Análisis Univariable

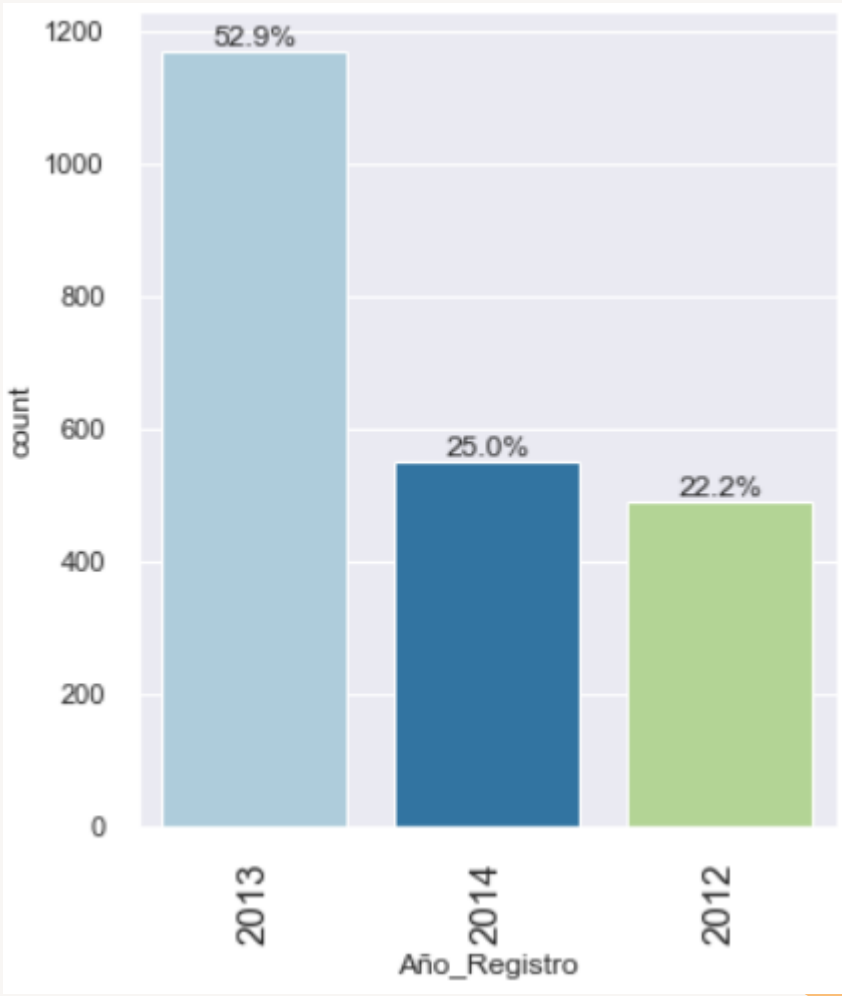




Análisis Exploratorio de Datos

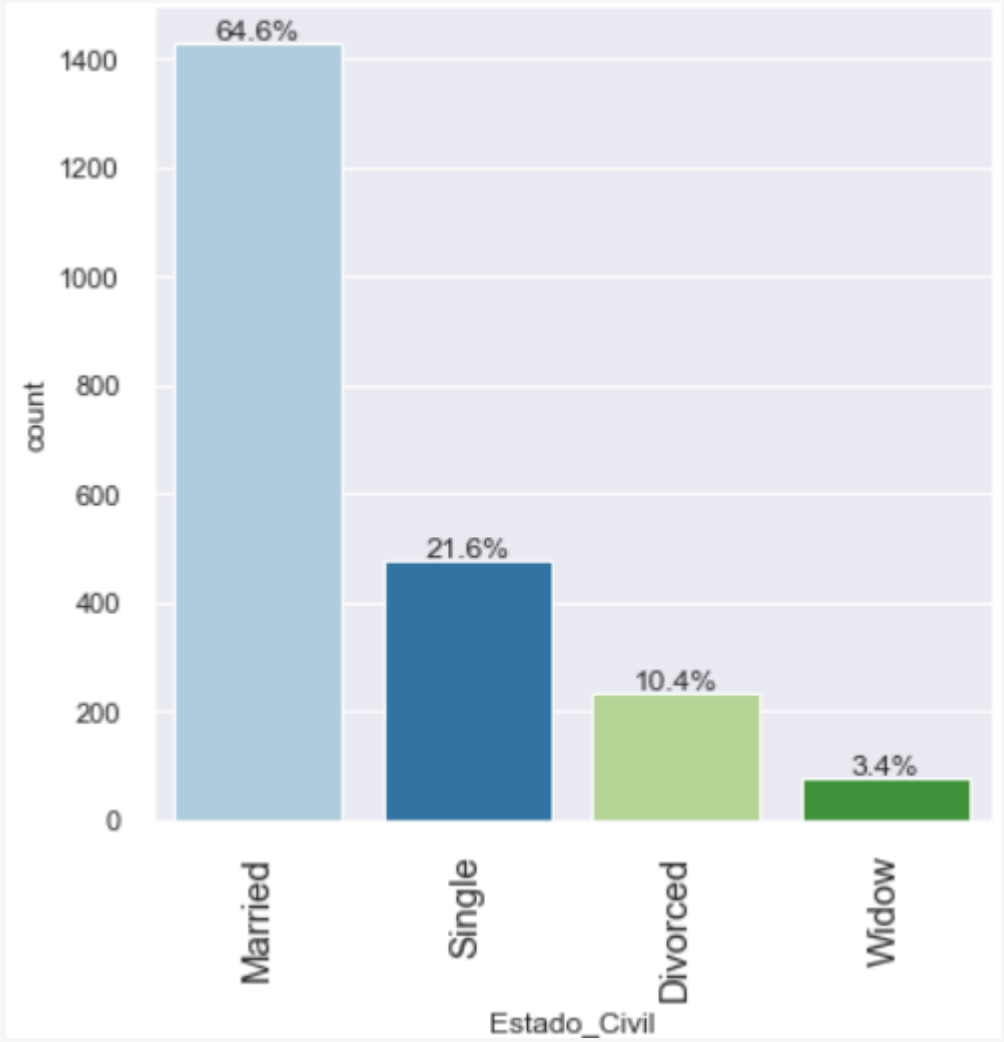
Análisis Univariable

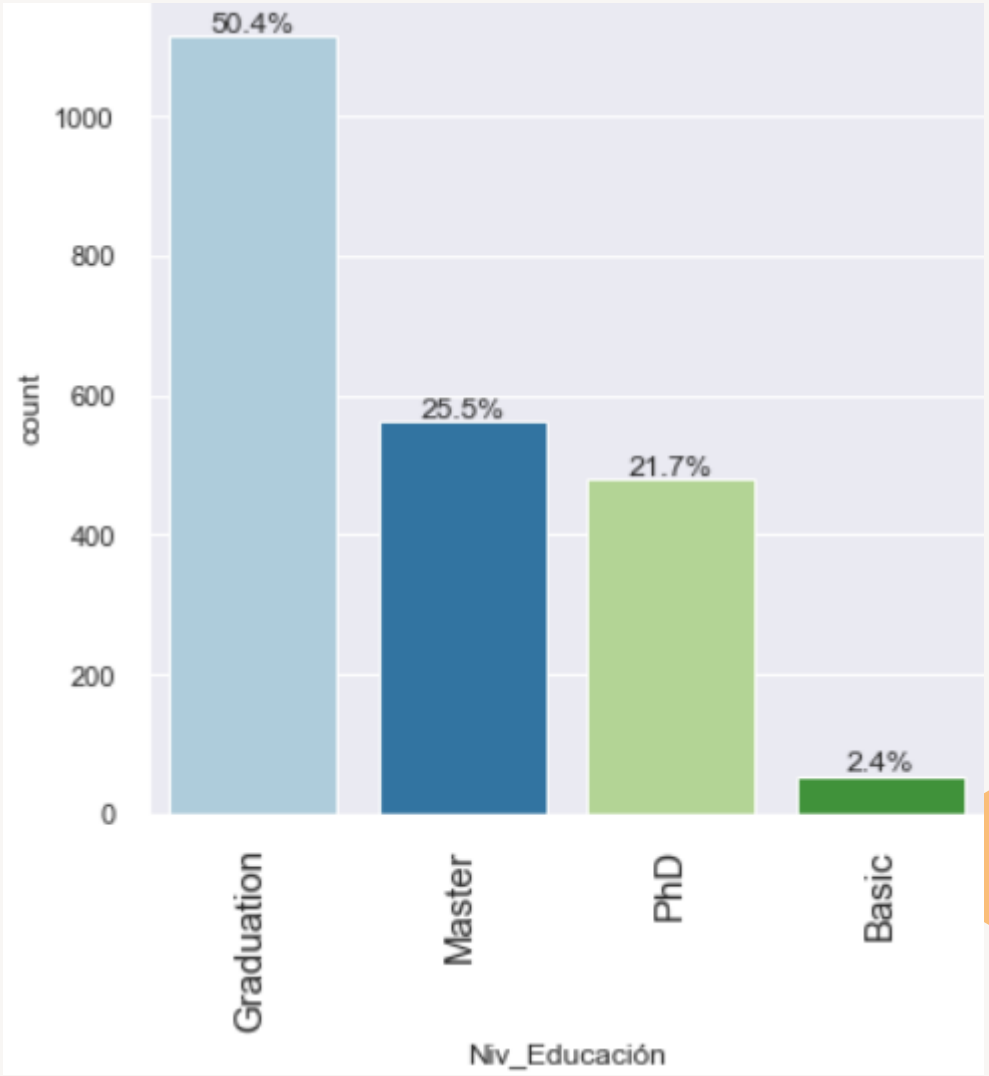




Análisis Exploratorio de Datos

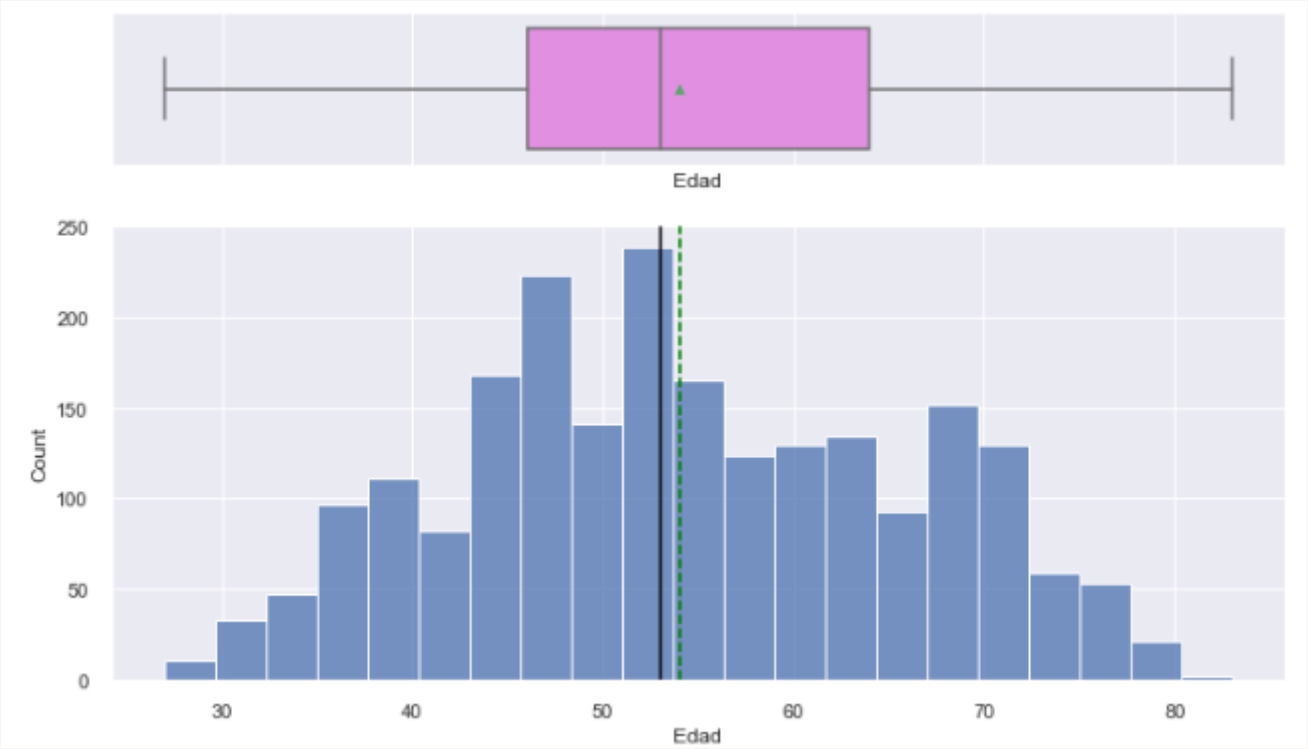
Análisis Univariable

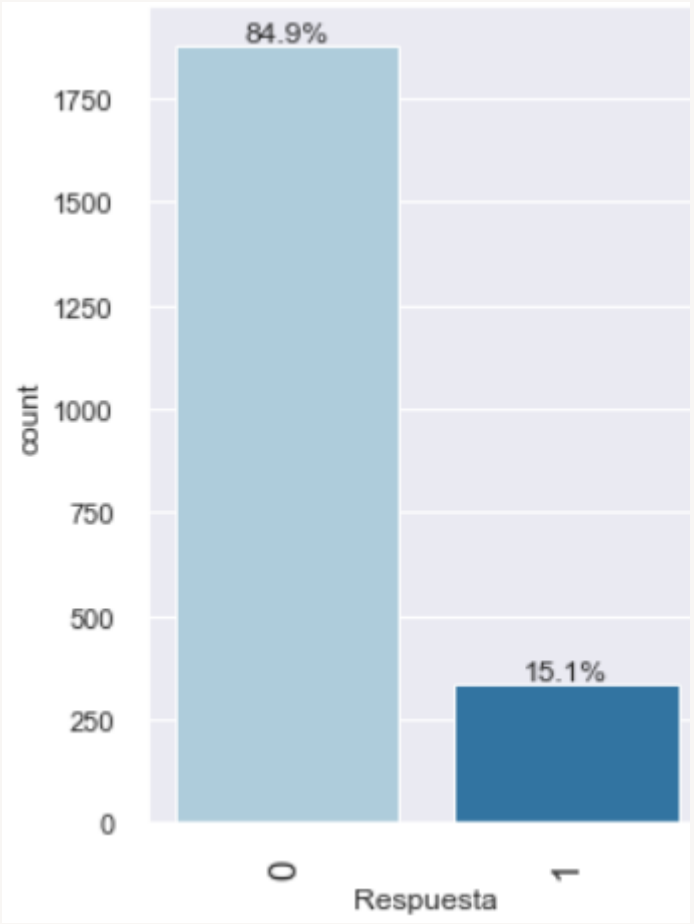




Análisis Exploratorio de Datos

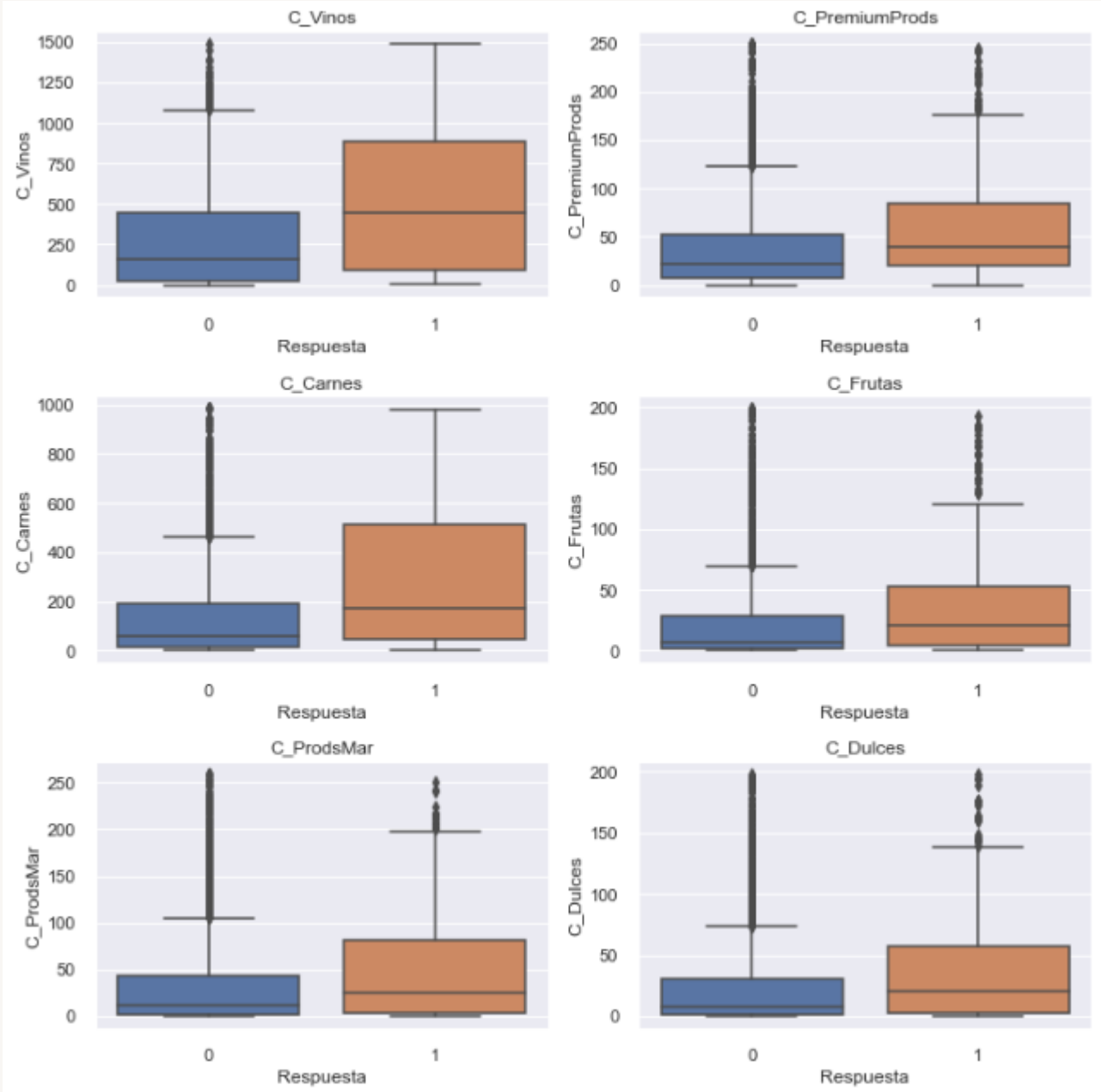
Análisis Univariable



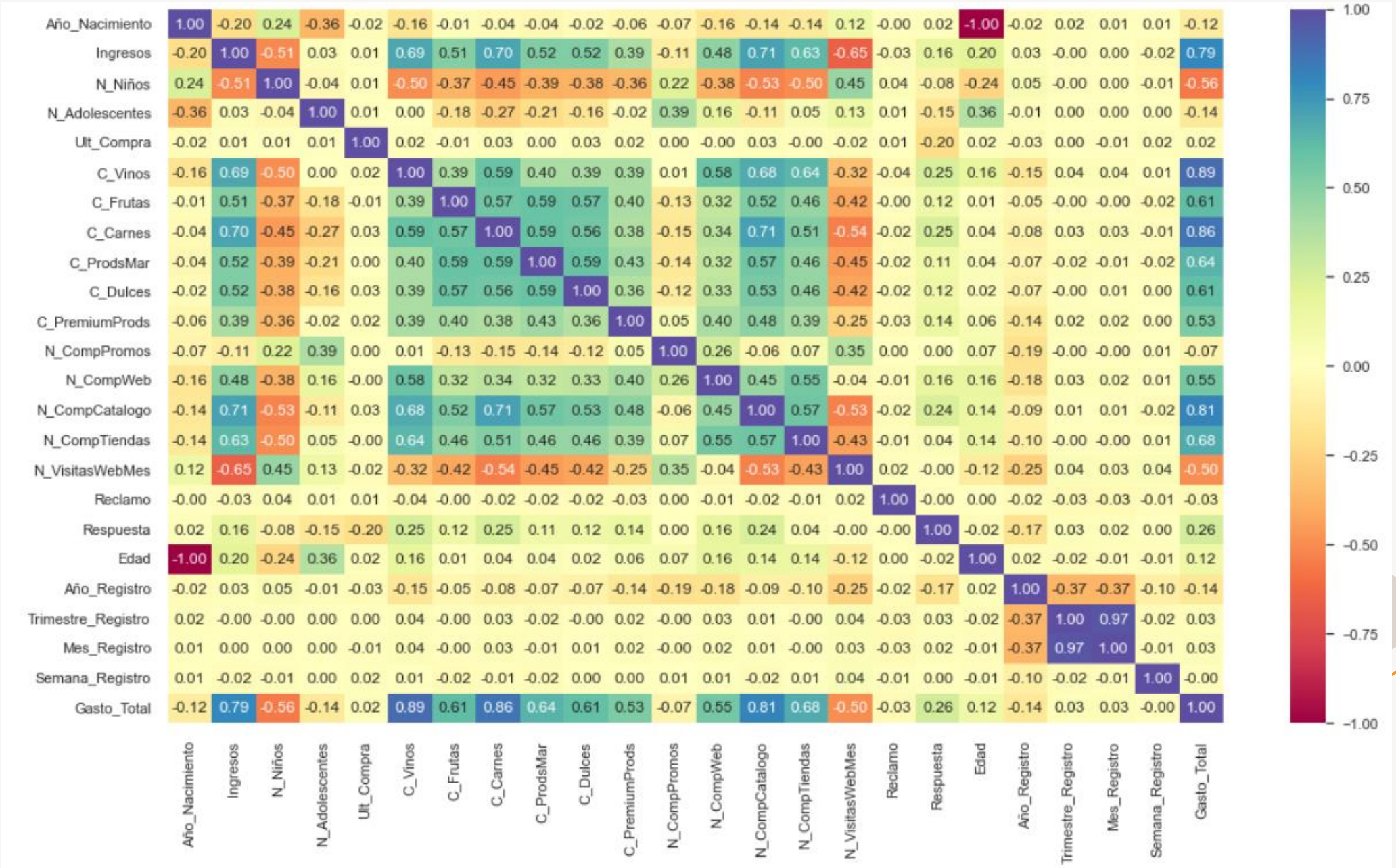


Análisis Exploratorio de Datos

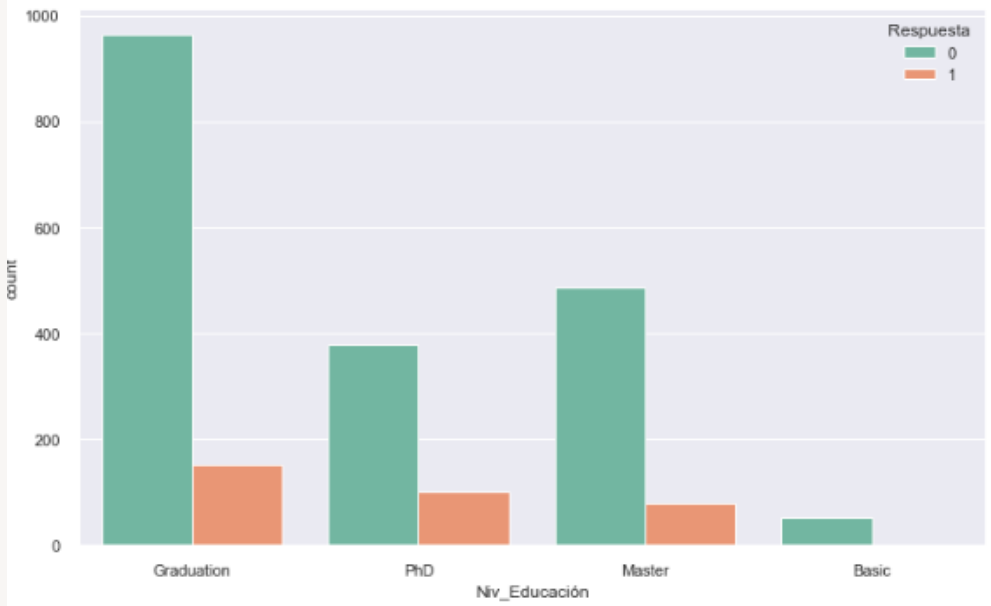
Análisis Bi Variable

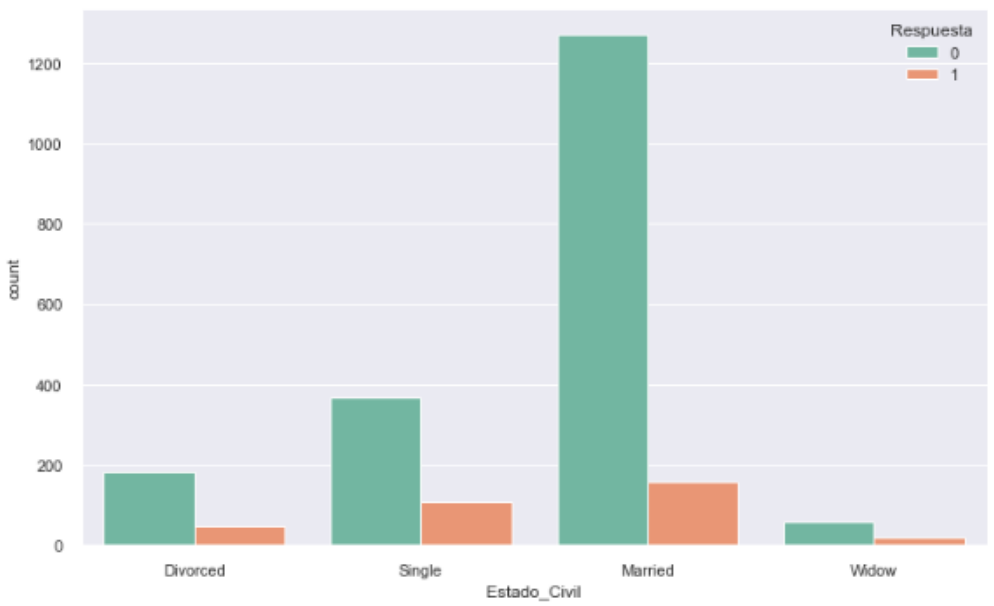


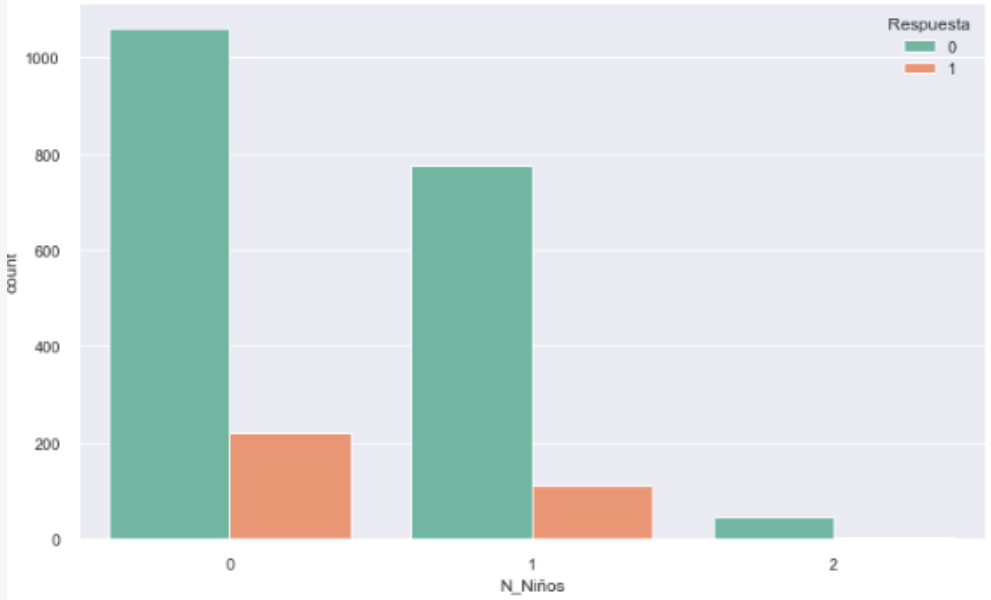
Análisis Exploratorio de Datos

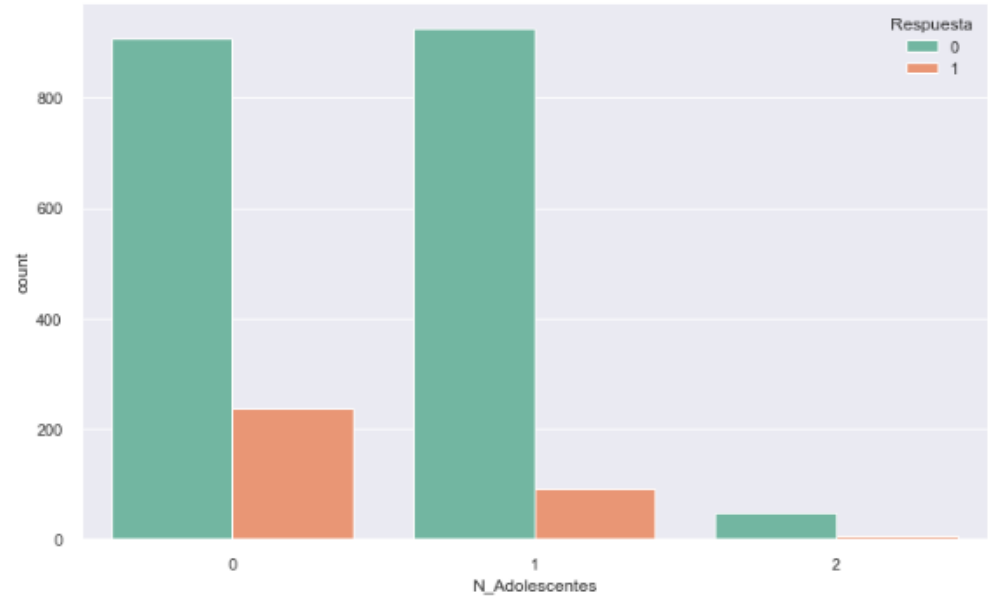


Análisis Exploratorio de Datos

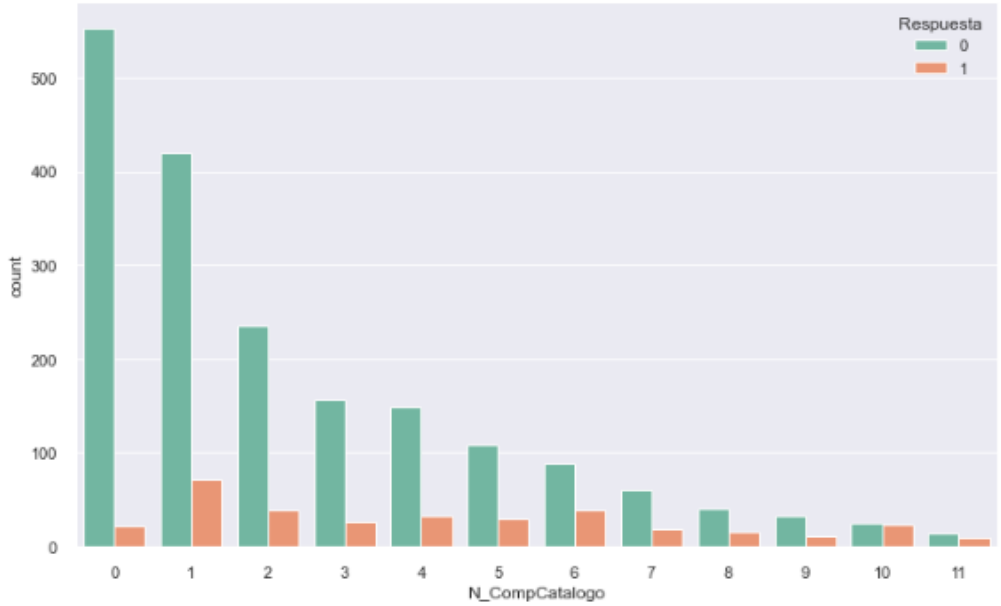


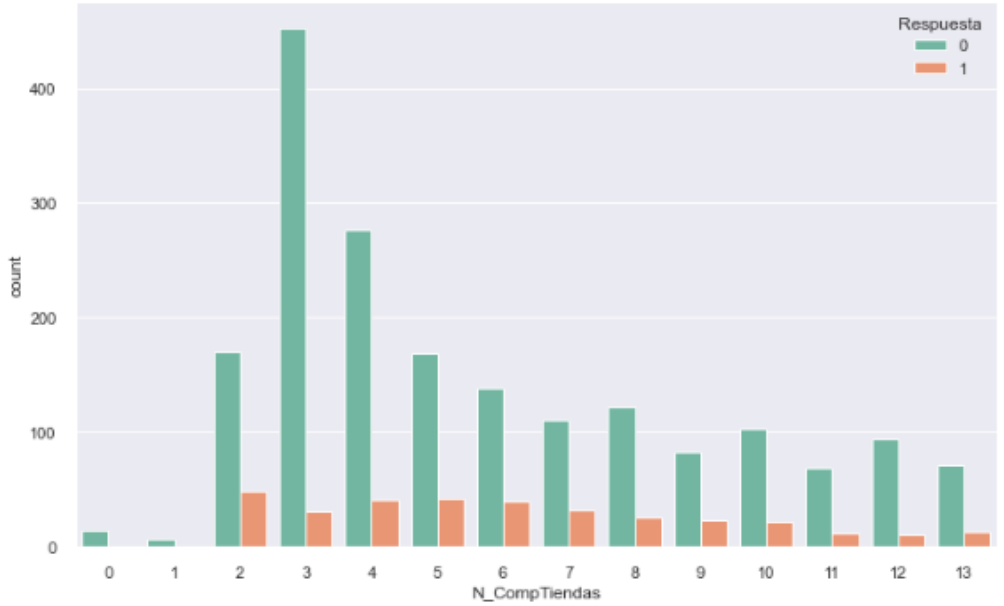


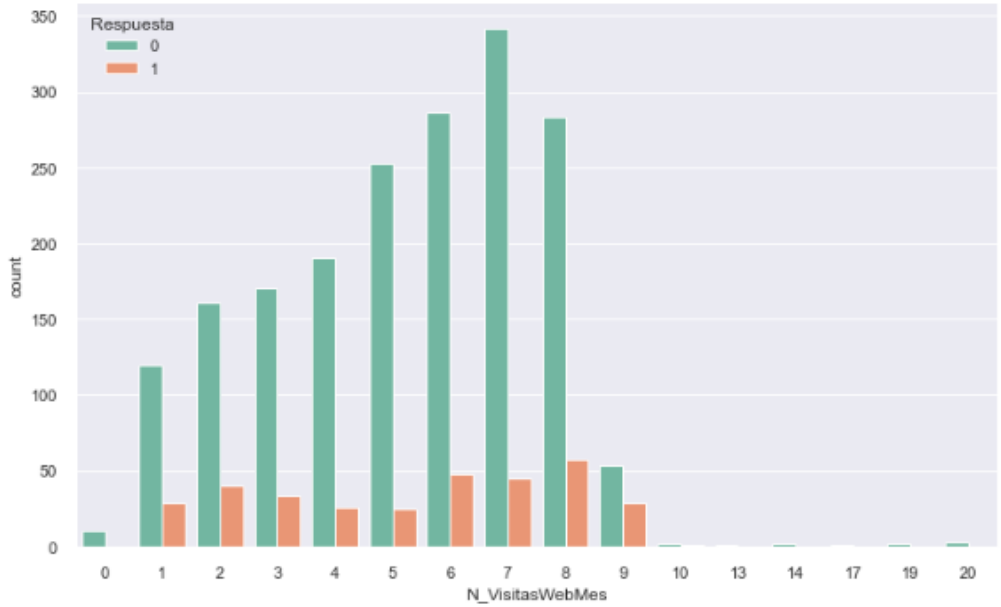


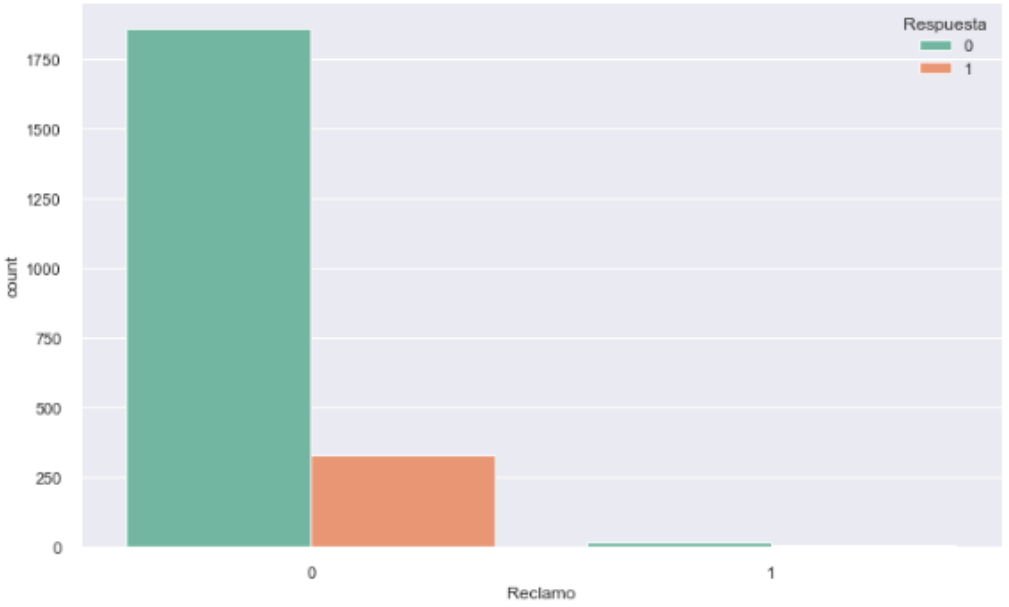


Análisis Exploratorio de Datos









Resultados

Matriz de Interpretación de matrices de confusión

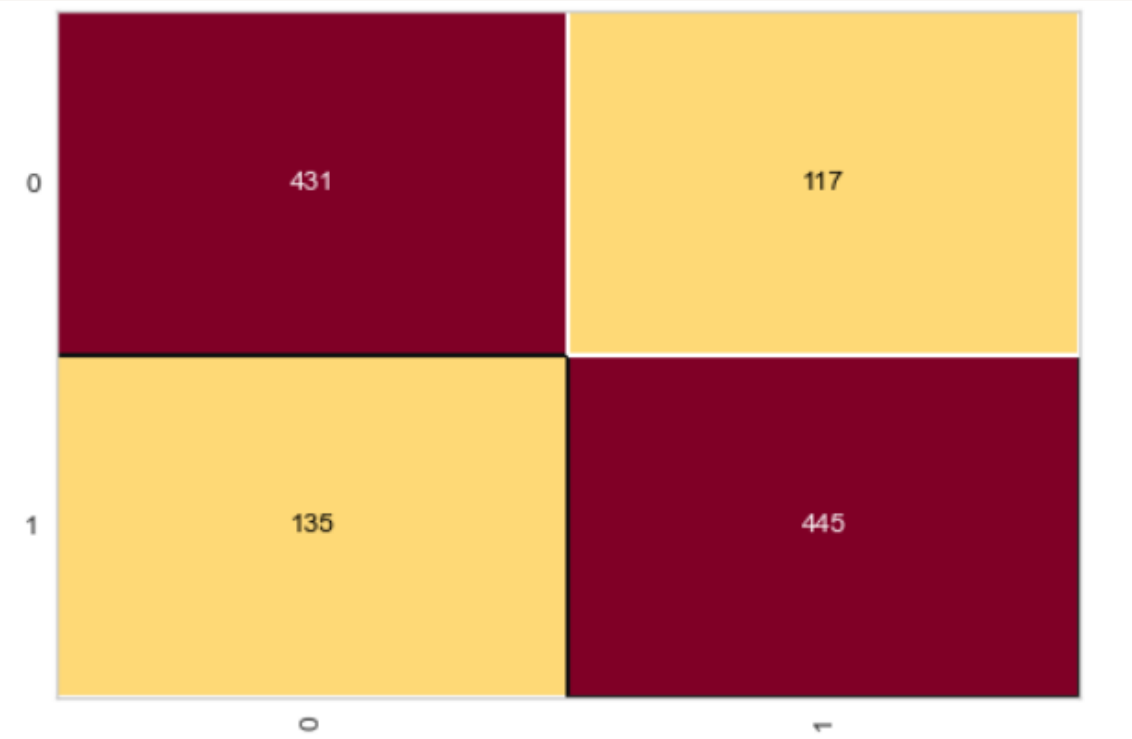
Matriz de confusión		Estimado por el modelo			
		Negativo (N)	*Positivo (P)*		
Real	*Negativo*	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	**Precisión** ***("precision")*** Porcentaje predicciones positivas correctas:	d/(b+d)
		Sensibilidad, exhaustividad ***("Recall")*** Porcentaje casos positivos detectados	**Especifidad** ***(Specifity)*** Porcentaje casos negativos detectados	**Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)*	
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)	

Fuente: Telefónica Tech / Elaboración propia

Resultados

Regresión Logística

Matriz de confusión Modelo de regresión Logística

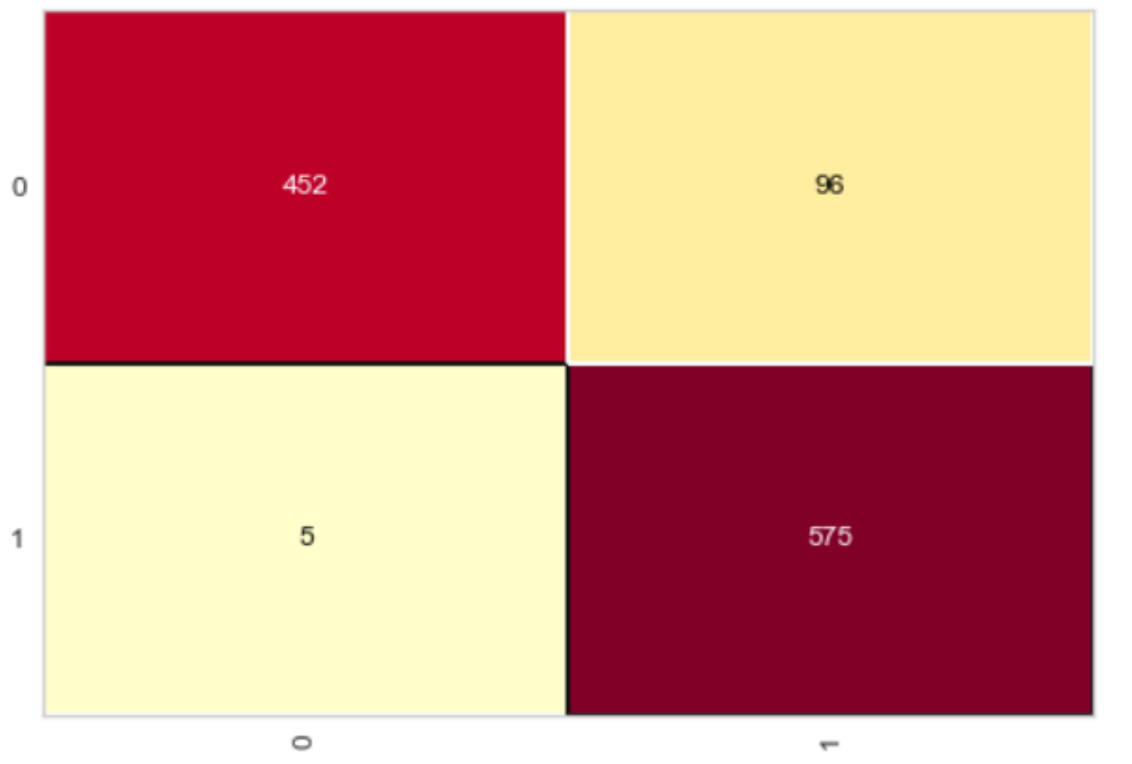


	precision	recall	f1-score	support
0	0.76	0.79	0.77	548
1	0.79	0.77	0.78	580
accuracy			0.78	1128
macro avg	0.78	0.78	0.78	1128
weighted avg	0.78	0.78	0.78	1128

Resultados

Árbol de Decisión

Matriz de confusión Modelo de árbol de decisión



	precision	recall	f1-score	support
0	0.99	0.82	0.90	548
1	0.86	0.99	0.92	580
accuracy			0.91	1128
macro avg	0.92	0.91	0.91	1128
weighted avg	0.92	0.91	0.91	1128

Resultados

Árbol de decisión

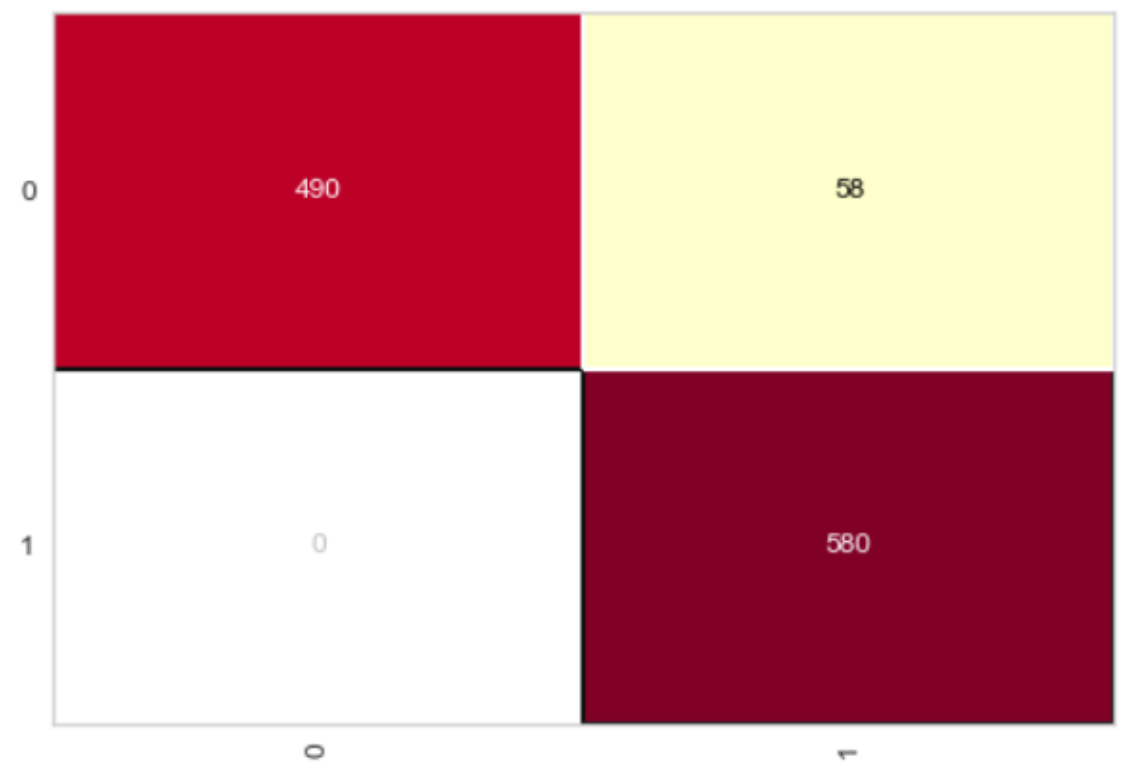
Variables de más incidencia del Modelo de árbol de decisión

Gasto_Total	0.122121
Ult_Compra	0.110246
Ingresos	0.085185
N_CompCatalogo	0.081668
N_VisitasWebMes	0.074686
Año_Registro	0.056051
C_Carnes	0.046741
C_PremiumProds	0.046604
Mes_Registro	0.046265
C_ProdsMar	0.044952
C_Vinos	0.042751
N_CompTiendas	0.040069
Estado_Civil	0.036827
C_Frutas	0.033482
N_CompPromos	0.029870
Año_Nacimiento	0.027995
Edad	0.019933
C_Dulces	0.019093
N_CompWeb	0.008894
N_Adolescentes	0.007320
Niv_Educación	0.007162
Trimestre_Registro	0.004203
N_Niños	0.004065
Semana_Registro	0.003814
Reclamo	0.000000

Resultados

Random Forest

Matriz de confusión Modelo de random forest



	precision	recall	f1-score	support
0	1.00	0.89	0.94	548
1	0.91	1.00	0.95	580
accuracy			0.95	1128
macro avg	0.95	0.95	0.95	1128
weighted avg	0.95	0.95	0.95	1128

Resultados

Random Forest

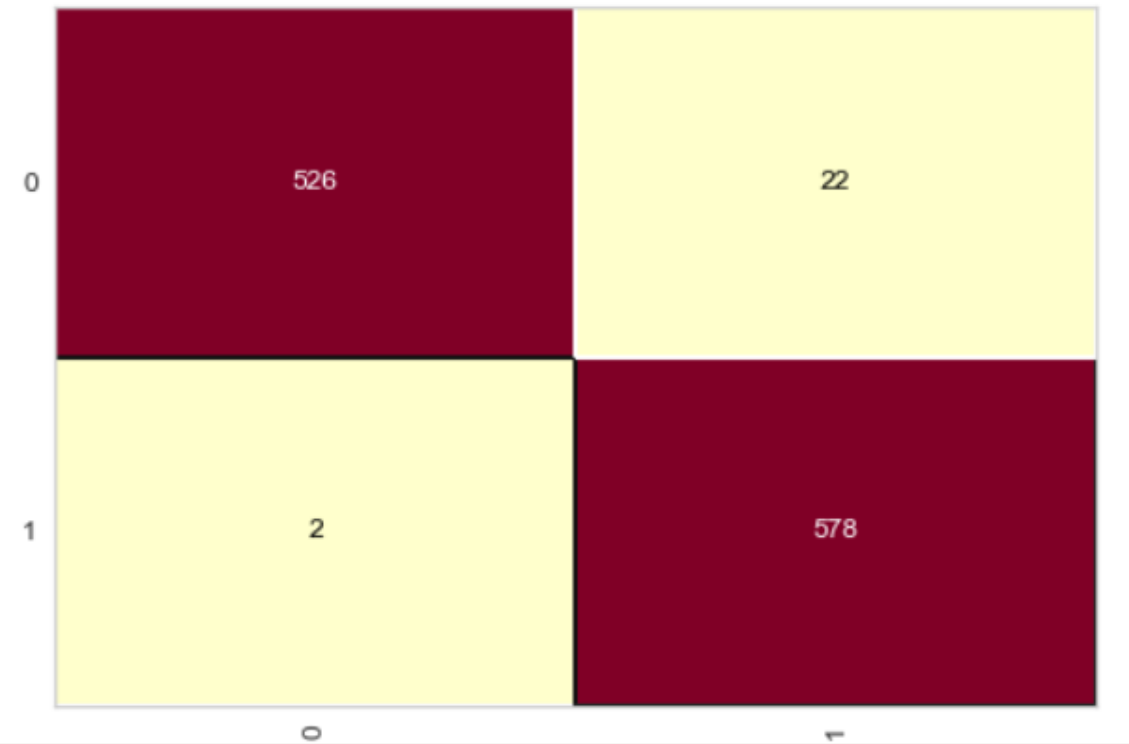
Variables de más incidencia del Modelo de random forest

Ult_Compra	0.121871
Gasto_Total	0.077065
Ingresos	0.073850
N_CompCatalogo	0.070520
C_Carnes	0.064339
C_Vinos	0.064317
C_PremiumProds	0.055420
N_CompTiendas	0.048866
N_VisitasWebMes	0.042897
Año_Registro	0.039514
C_ProdsMar	0.037021
Edad	0.035631
C_Dulces	0.034915
Año_Nacimiento	0.032873
N_CompWeb	0.032366
C_Frutas	0.032242
Mes_Registro	0.025127
N_CompPromos	0.024508
Niv_Educación	0.018804
Semana_Registro	0.017885
Estado_Civil	0.016927
N_Adolescentes	0.015441
Trimestre_Registro	0.009621
N_Niños	0.007688
Reclamo	0.000291

Resultados

Extra trees

Matriz de confusión Modelo de extra trees



	precision	recall	f1-score	support
0	1.00	0.96	0.98	548
1	0.96	1.00	0.98	580
accuracy			0.98	1128
macro avg	0.98	0.98	0.98	1128
weighted avg	0.98	0.98	0.98	1128

Resultados

Extra trees

Variables de más incidencia del Modelo de extra trees

Ult_Compra	0.092289
N_CompCatalogo	0.069126
Año_Registro	0.060676
Gasto_Total	0.059969
C_Carnes	0.052329
C_Vinos	0.050040
N_CompTiendas	0.048012
N_CompWeb	0.042402
N_VisitasWebMes	0.041493
Ingresos	0.040188
C_PremiumProds	0.039998
Niv_Educación	0.036891
N_CompPromos	0.036278
Estado_Civil	0.033640
N_Adolescentes	0.032826
Semana_Registro	0.032735
Edad	0.032355
C_Dulces	0.032351
C_Frutas	0.031334
Año_Nacimiento	0.030785
C_ProdsMar	0.030296
Mes_Registro	0.028957
Trimestre_Registro	0.025141
N_Niños	0.018890
Reclamo	0.000997

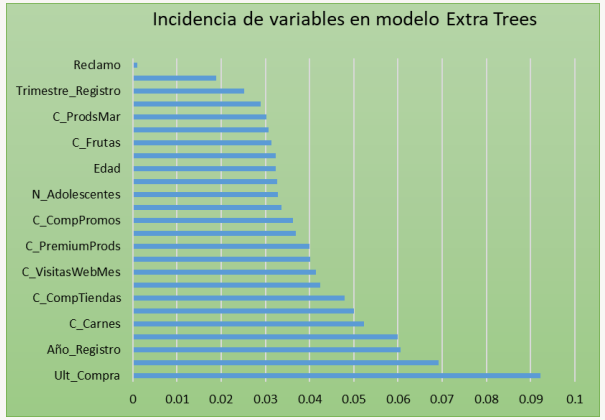
Resultados

Resumen

Resumen de resultados por modelo de predicción

	Decision Tree	Random Forest	Extra Trees	Logistic Regression
Model	Decision Tree	Random Forest	Extra Trees	Logistic Regression
Scaling	Normal Data	Normal Data	Normal Data	Normal Data
Type	Gini	Gini	Gini	-
Accuracy	0.9104	0.9485	0.9787	0.7765

Resultados

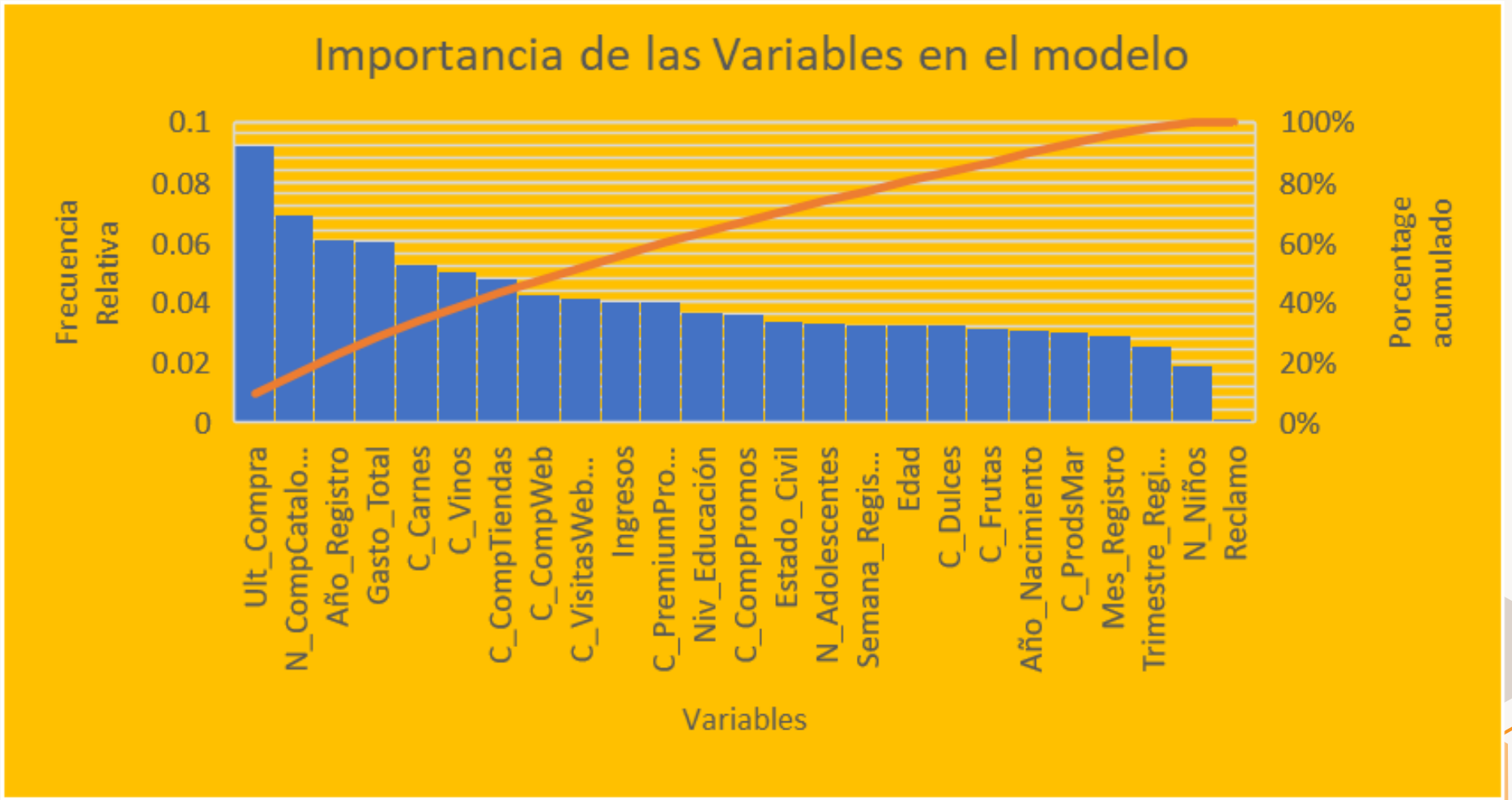


Importancia de las variables según el modelo más exacto

Variable	Importancia	% Importancia individual	% Importancia acumluada
Ult_Compra	0.092289	9%	9%
N_CompCatalogo	0.069126	7%	16%
Año_Registro	0.060676	6%	22%
Gasto_Total	0.059969	6%	28%
C_Carnes	0.052329	5%	33%
C_Vinos	0.05004	5%	38%
C_CompTiendas	0.048012	5%	43%
C_CompWeb	0.042402	4%	47%
C_VisitasWebMes	0.041493	4%	52%
Ingresos	0.040188	4%	56%
C_PremiumProds	0.039998	4%	60%
Niv_Educación	0.036891	4%	63%
C_CompPromos	0.036278	4%	67%
Estado_Civil	0.03364	3%	70%
N_Adolescentes	0.032826	3%	74%
Semana_Registro	0.032735	3%	77%
Edad	0.032355	3%	80%
C_Dulces	0.032351	3%	83%
C_Frutas	0.031334	3%	86%
Año_Nacimiento	0.030785	3%	90%
C_ProdsMar	0.030296	3%	93%
Mes_Registro	0.028957	3%	95%
Trimestre_Registro	0.025141	3%	98%
N_Niños	0.01889	2%	100%
Reclamo	0.000997	0%	100%

Resultados

Diagrama de Pareto



Variables de más incidencia del estudio

Variable	Importancia	% Importancia individual	% Importancia acumluada
Ult_Compra	0.092289	9%	9%
N_CompCatalogo	0.069126	7%	16%
Año_Registro	0.060676	6%	22%
Gasto_Total	0.059969	6%	28%
C_Carnes	0.052329	5%	33%
C_Vinos	0.05004	5%	38%
C_CompTiendas	0.048012	5%	43%
C_CompWeb	0.042402	4%	47%
C_VisitasWebMes	0.041493	4%	52%
Ingresos	0.040188	4%	56%

Propuesta de solución

IMPLICACIONES ORGANIZACIONALES

- ✓ **Entender los factores que afectan a la variable de respuesta:**
- ✓ **Predecir el éxito o la negativa de la campaña para un individuo:**
- ✓ **Clasificar a los clientes por características aportantes al modelo:**
- ✓ **Mejorar la toma de decisiones basadas en datos:**

Implicaciones sobre innovación empresarial

1. Establecer objetivos claros:

2. Grupos de clientes:

3. Recopilación y gestión de información:

4. Análisis para la predicción:

5.Configuración de campañas:

6. El aprendizaje autónomo:

7. Optimización de los canales en línea:

8. Vigilancia y medición constante:

9. Trabajo colaborativo entre equipos:

10. Educación y capacitación:

11. Respecto a la Privacidad:

12. Experimentos y pruebas A/B:

13. Mantenimiento de la tecnología:

14. Los comentarios de los clientes:

15. La adaptación continua:

Conclusiones

Personalización Mejorada: Los modelos de predicción permiten una personalización más profunda en las estrategias de marketing.

Mayor Eficiencia: La segmentación precisa y las predicciones ayudan a dirigir los recursos y esfuerzos de marketing de manera más efectiva hacia los clientes más propensos a responder positivamente

Mejor Retorno de Inversión (ROI): Al concentrarse en los clientes que tienen más probabilidades de participar en las campañas, la empresa puede lograr un ROI más alto.

Adaptación Continua: Los modelos de predicción no son estáticos. Con el tiempo, la empresa puede refinar y mejorar.

Competitividad: Ventaja competitiva al comprender mejor su base de clientes y responder rápidamente a las tendencias del mercado.

Recomendaciones

- ✓ **Inversión en Capacidades Analíticas**
- ✓ **Calidad de Datos**
- ✓ **Validación Constante**
- ✓ **Pruebas Rigurosas**
- ✓ **Colaboración entre Equipos**
- ✓ **Cumplimiento Normativo**
- ✓ **Educación Interna**
- ✓ **Flexibilidad y Adaptación**
- ✓ **Medición y Evaluación**
- ✓ **Aprendizaje Continuo**

Referencias:

Amat, J. (Octubre, 2020). Arboles de decision Python. Ciencia de datos, teoría y ejemplos prácticos en R y Python. https://cienciadedatos.net/documentos/py07_arboles_decision_python

Calva, Karen. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado. https://lajc.epn.edu.ec/index.php/LAJC/article/download/264/159/

Cárdenas, J. (Octubre, 2022). Qué es la regresión logística binaria Y Como analizarla. Networkianos. Blog de Sociología. https://networkianos.com/regresion-logistica-binaria/

Carrasco Ortega, M. (2017). Herramientas del marketing digital que permiten desarrollar presencia online, analizar la web, conocer a la audiencia y mejorar los resultados de búsqueda. Scielo(45). Obtenido de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1994-37332020000100003

Espino, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. https://openaccess.uoc.edu/bitstream/10609/59565/6/caresptimTFG0117mem%C3%

Referencias:

Huertas, A. (Junio, 2020). Algoritmos de aprendizaje automático supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas. https://repository.usta.edu.co/bitstream/handle/11634/29886/2020alexanderhuertas.pdf?sequence=1&isAllowed=y

Giraldo, L. (2018). Los desafíos del marketing en la era del big data. Sistema de Información Científica Redalyc, Red de Revistas Científicas. https://www.redalyc.org/journal/4768/476852090003/

Gonzalez, L. (Marzo, 2018. Aprendizaje Supervisado: Random forest classification. https://aprendeia.com/aprendizaje-supervisado-random-forest-classification/

Marín, J. (2019). Análisis de datos para el marketing digital emprendedor: Caso de estudio del Parque de Innovación Empresarial de Manizales. http://www.scielo.org.co/pdf/unem/v22n38/2145-4558-unem-22-38-65.pdf

Salazar, A. (2019). MPORTANCIA DE UNA INVESTIGACIÓN DE MERCADO. https://www.itson.mx/publicaciones/pacioli/documents/no71/49a.-_importancia_de_la_investigacion_de_mercado_nx.pdf

Referencias:

Marín, J. (2019). Análisis de datos para el marketing digital emprendedor: Caso de estudio del Parque de Innovación Empresarial de Manizales. http://www.scielo.org.co/pdf/unem/v22n38/2145-4558-unem-22-38-65.pdf

Salazar, A. (2019). MPORTANCIA DE UNA INVESTIGACIÓN DE MERCADO. https://www.itson.mx/publicaciones/pacioli/documents/no71/49a.-_importancia_de_la_investigacion_de_mercado_nx.pdf

Zamorano, Juan. (2018). Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea. https://docta.ucm.es/entities/publication/7f2287a4-7122-454d-803e-0a8b47786649

Zúñiga, Freddy & Poveda, Diego & Llerena, William. (2023). El BIG DATA Y SU IMPLICACIÓN EN EL MARKETING. Revista de Comunicación de la SEECI. 56. 302-321. 10.15198/seeci.2023.56.e83

Muchas Gracias