



CASO DE ESTUDIO COMPAÑÍA E-CORP

Osiris Contreras
Maritza zapata
Juan Jose Molina
David Toro



CONTENIDO

1. Presentación del caso
2. Análisis de los datos
3. Limpieza y transformación de los datos
4. Preparación de los datos
5. Selección de variables
6. Aplicación y comparación de técnicas de modelado
7. Evaluación del mejor modelo
8. Conclusiones



Presentación del caso

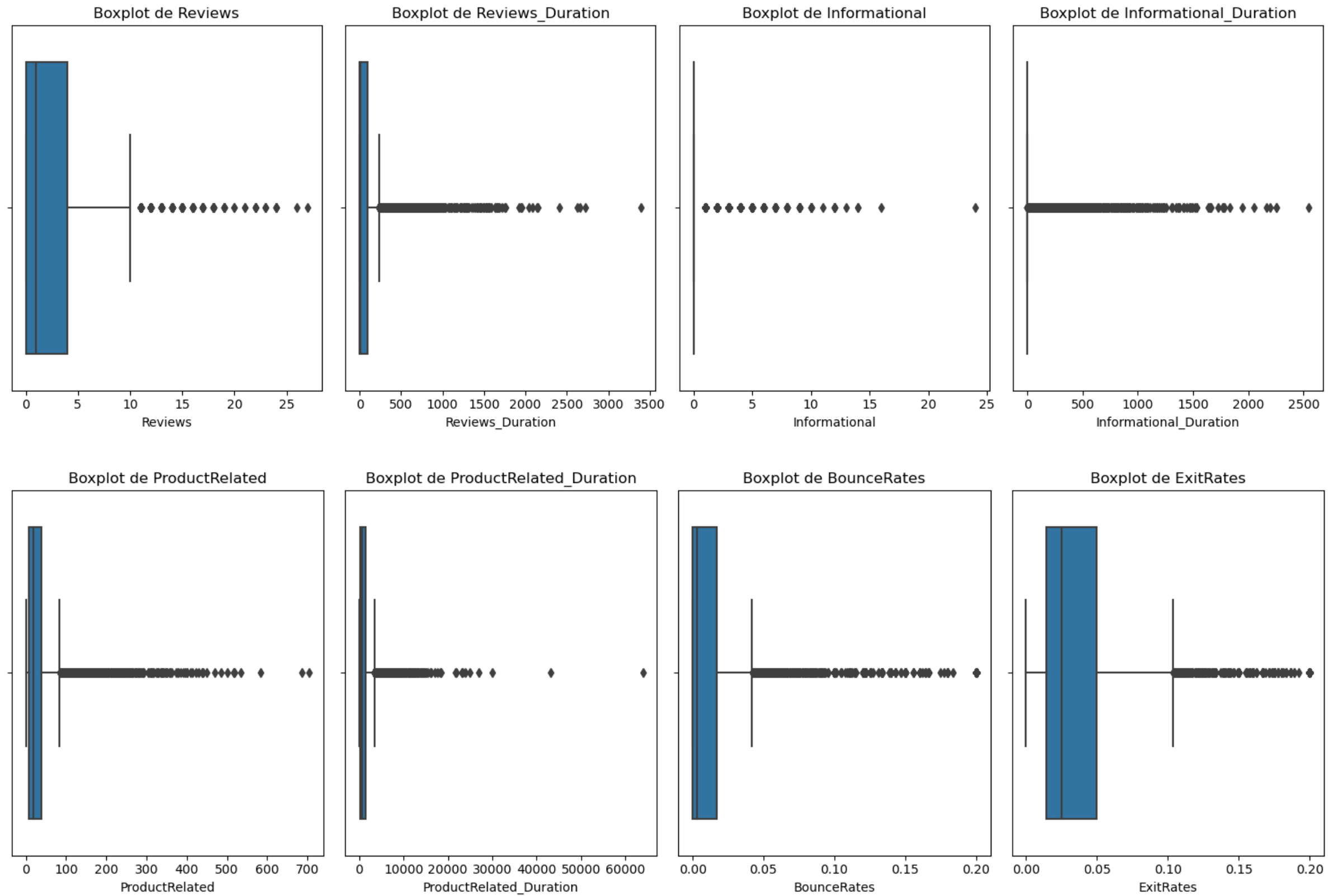


E-Corp, líder en productos de lujo, busca expansión digital.

- Problema: Ventas digitales bajas.
- Sospecha: Ineficacia en marketing digital y alcance de audiencia.
- Solución: Contratación de consultores y uso de Machine Learning para:
 - Identificar clientes potenciales.
 - Optimizar inversión publicitaria.
 - Mejorar relevancia de campañas.
 - Incrementar impacto en decisiones de compra.

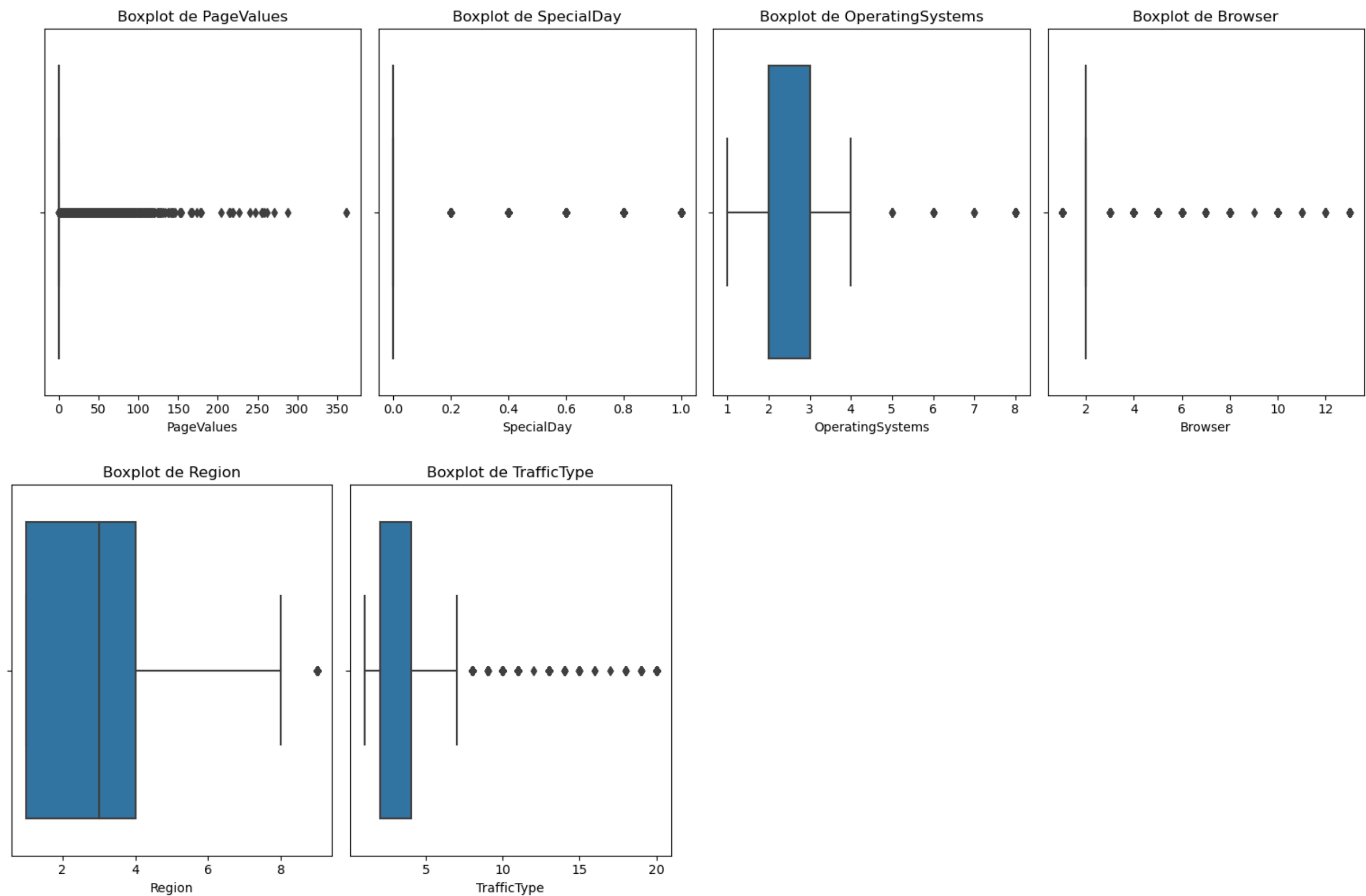


Analisis de los Datos



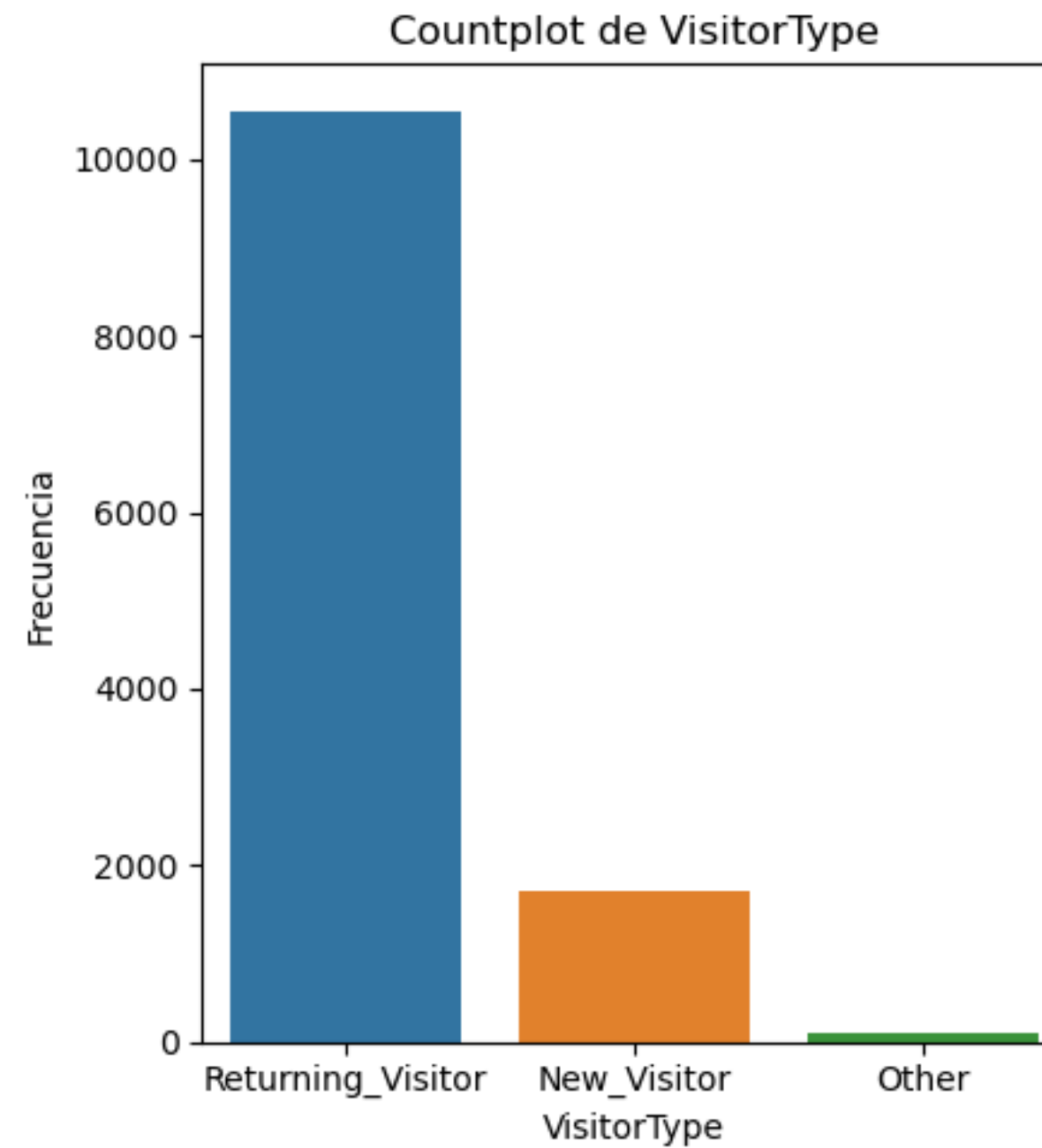
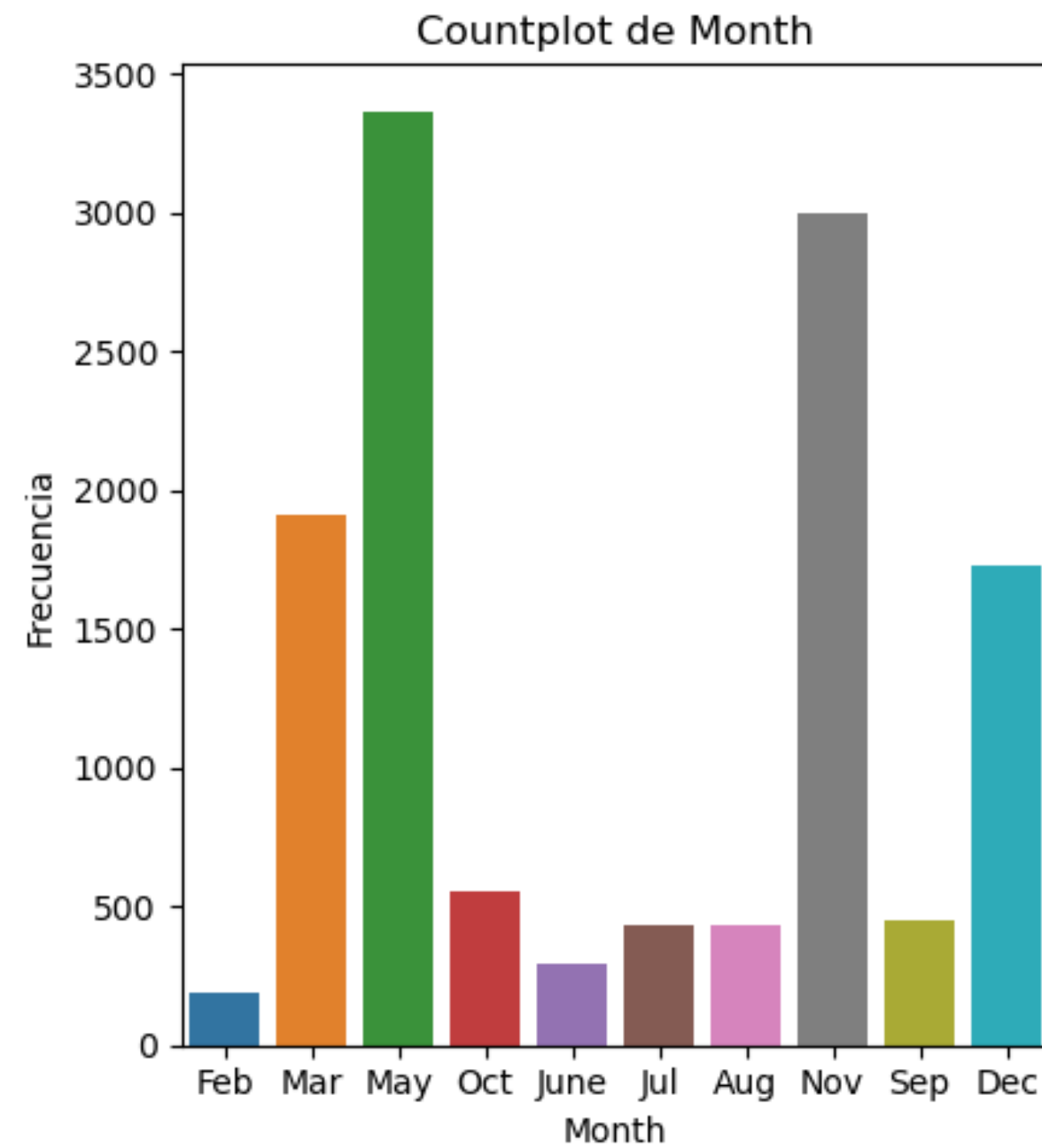
**BOXPLOTS DE VARIABLES
NUMÉRICAS**

Analisis de los Datos



**BOXPLOTS DE VARIABLES
NUMÉRICAS**

Análisis de los datos



COUNTPLOTS DE VARIABLES
CATEGORICAS

Análisis de los datos

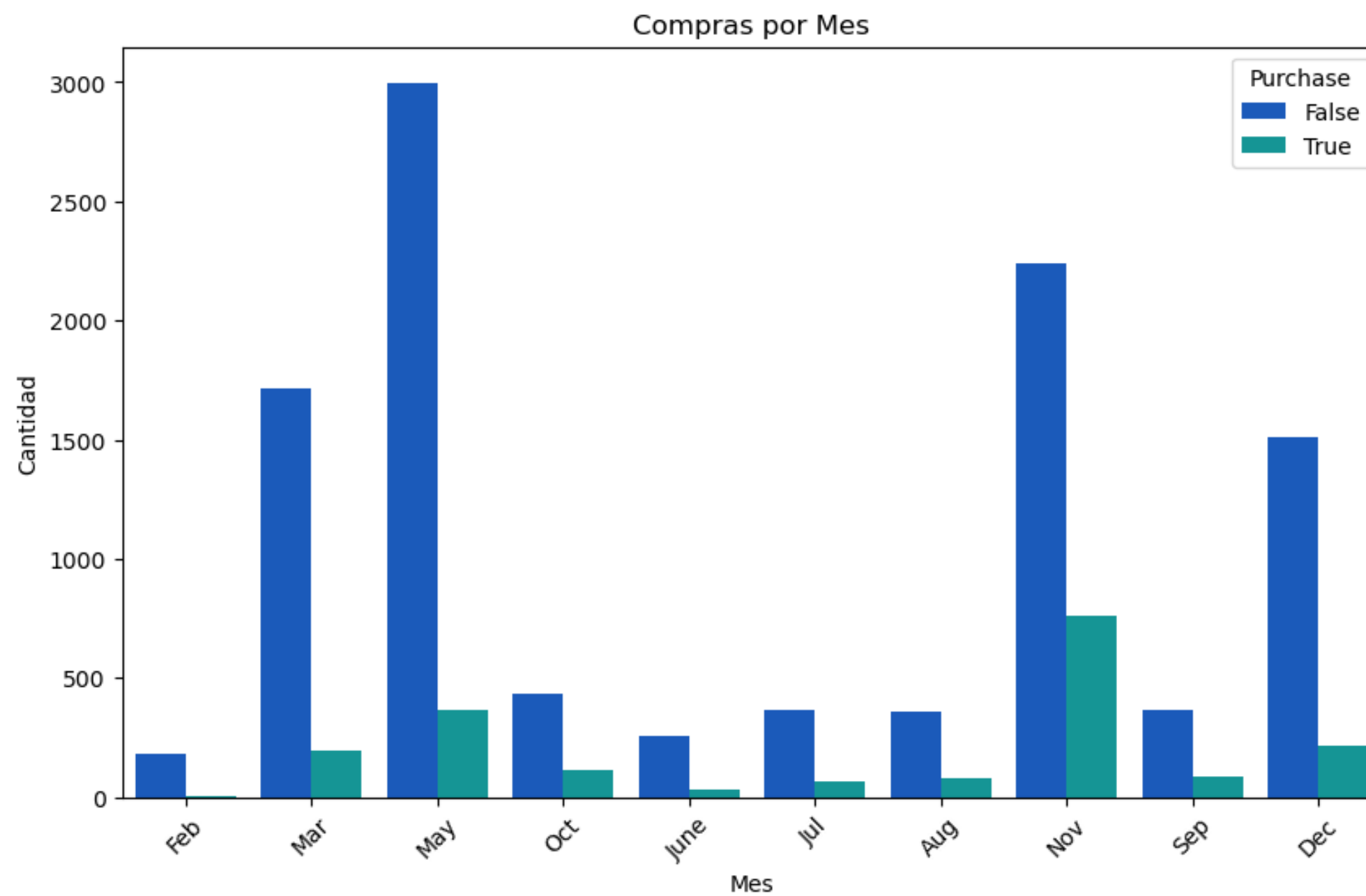


GRÁFICO DE BARRAS DE
COMPRAS POR MES

Análisis de los datos

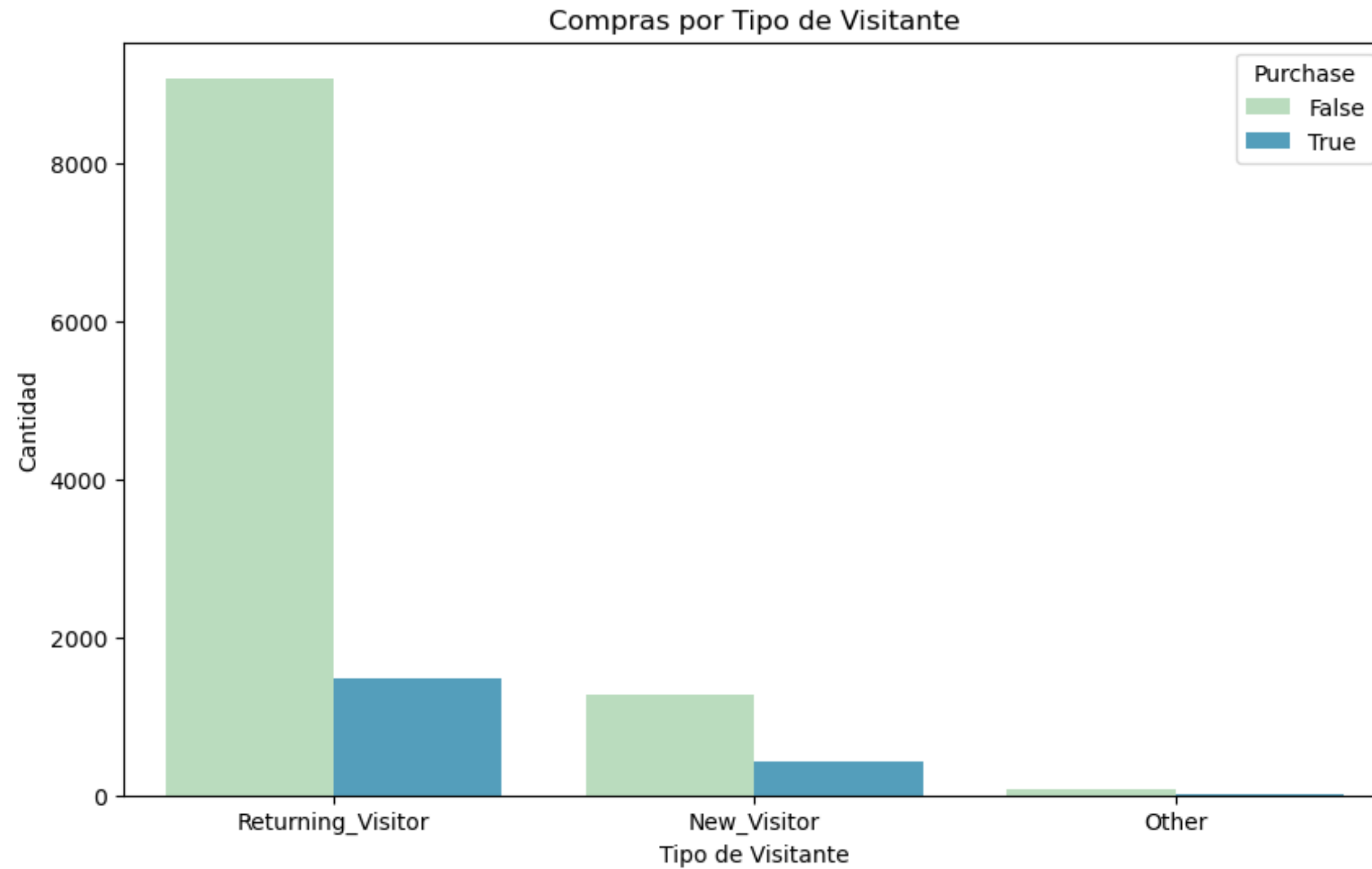


GRÁFICO DE BARRAS DE COMPRAS POR
TIPO DE VISITANTE

Limpieza y transformación de los datos



```
1 # Información general del dataset
2 df_original.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Reviews                12330 non-null  int64
1   Reviews_Duration       12330 non-null  float64
2   Informational           12330 non-null  int64
3   Informational_Duration  12330 non-null  float64
4   ProductRelated         12330 non-null  int64
5   ProductRelated_Duration 12330 non-null  float64
6   BounceRates            12330 non-null  float64
7   ExitRates              12330 non-null  float64
8   PageValues             12330 non-null  float64
9   SpecialDay             12330 non-null  float64
10  Month                  12330 non-null  object
11  OperatingSystems       12330 non-null  int64
12  Browser                12330 non-null  int64
13  Region                 12330 non-null  int64
14  TrafficType            12330 non-null  int64
15  VisitorType            12330 non-null  object
16  Weekend                12330 non-null  bool
17  Purchase               12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

```
1 # Copia del dataset para aplicar transformaciones y limpieza de datos
2 df1 = df_original.copy()
```

```
1 from sklearn.preprocessing import LabelEncoder
2
3 # Codificación de variables categóricas
4 label_encoder = LabelEncoder()
5 df1['Month'] = label_encoder.fit_transform(df1['Month'])
6 df1['VisitorType'] = label_encoder.fit_transform(df1['VisitorType'])
7
8 # Conversión de datos booleanos a numéricos
9 df1['Weekend'] = df1['Weekend'].astype(int)
10 df1['Purchase'] = df1['Purchase'].astype(int)
11
12 print(df1.head())
```

```
1 from sklearn.preprocessing import MinMaxScaler, StandardScaler
2
3 # Seleccionar las características numéricas a normalizar o estandarizar
4 caracteristicas_numericas = df1[names]
5
6 # Inicializar el escalador MinMaxScaler
7 min_max_scaler = MinMaxScaler()
8
9 # Normalizar las características utilizando MinMaxScaler
10 caracteristicas_numericas_normalizadas = min_max_scaler.fit_transform(caracteristicas_numericas)
11
12 # Inicializar el escalador StandardScaler
13 standard_scaler = StandardScaler()
14
15 # Estandarizar las características utilizando StandardScaler
16 caracteristicas_numericas_estandarizadas = standard_scaler.fit_transform(caracteristicas_numericas_normalizadas)
17
18 # Convertir las características normalizadas y estandarizadas de nuevo a un DataFrame de pandas
19 nuevo_df = pd.DataFrame(caracteristicas_numericas_estandarizadas, columns= names)
20
21 # Mostrar las primeras filas del DataFrame con características estandarizadas y normalizadas
22 print("\nCaracterísticas estandarizadas:")
23 print(nuevo_df.info())
```

```
Características estandarizadas:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Reviews                12330 non-null  float64
1   Reviews_Duration       12330 non-null  float64
2   Informational           12330 non-null  float64
3   Informational_Duration  12330 non-null  float64
4   ProductRelated         12330 non-null  float64
5   ProductRelated_Duration 12330 non-null  float64
6   BounceRates            12330 non-null  float64
7   ExitRates              12330 non-null  float64
8   PageValues             12330 non-null  float64
9   SpecialDay             12330 non-null  float64
10  Month                  12330 non-null  float64
11  OperatingSystems       12330 non-null  float64
12  Browser                12330 non-null  float64
13  Region                 12330 non-null  float64
14  TrafficType            12330 non-null  float64
15  VisitorType            12330 non-null  float64
16  Weekend                12330 non-null  float64
17  Purchase               12330 non-null  float64
dtypes: float64(18)
memory usage: 1.7 MB
None
```

- 1.Verificación de datos nulos y tipos de datos:
2. Codificación de variables categóricas:
3. Normalización y estandarización de características:

Preparacion de los datos



```
1 # Comprobación de valores duplicados
2 nuevo_df.duplicated().sum()
```

125

```
1 # Eliminación de valores duplicados
2 nuevo_df = df1.drop_duplicates()
3
4 # Comprobación de valores duplicados
5 nuevo_df.duplicated().sum()
```

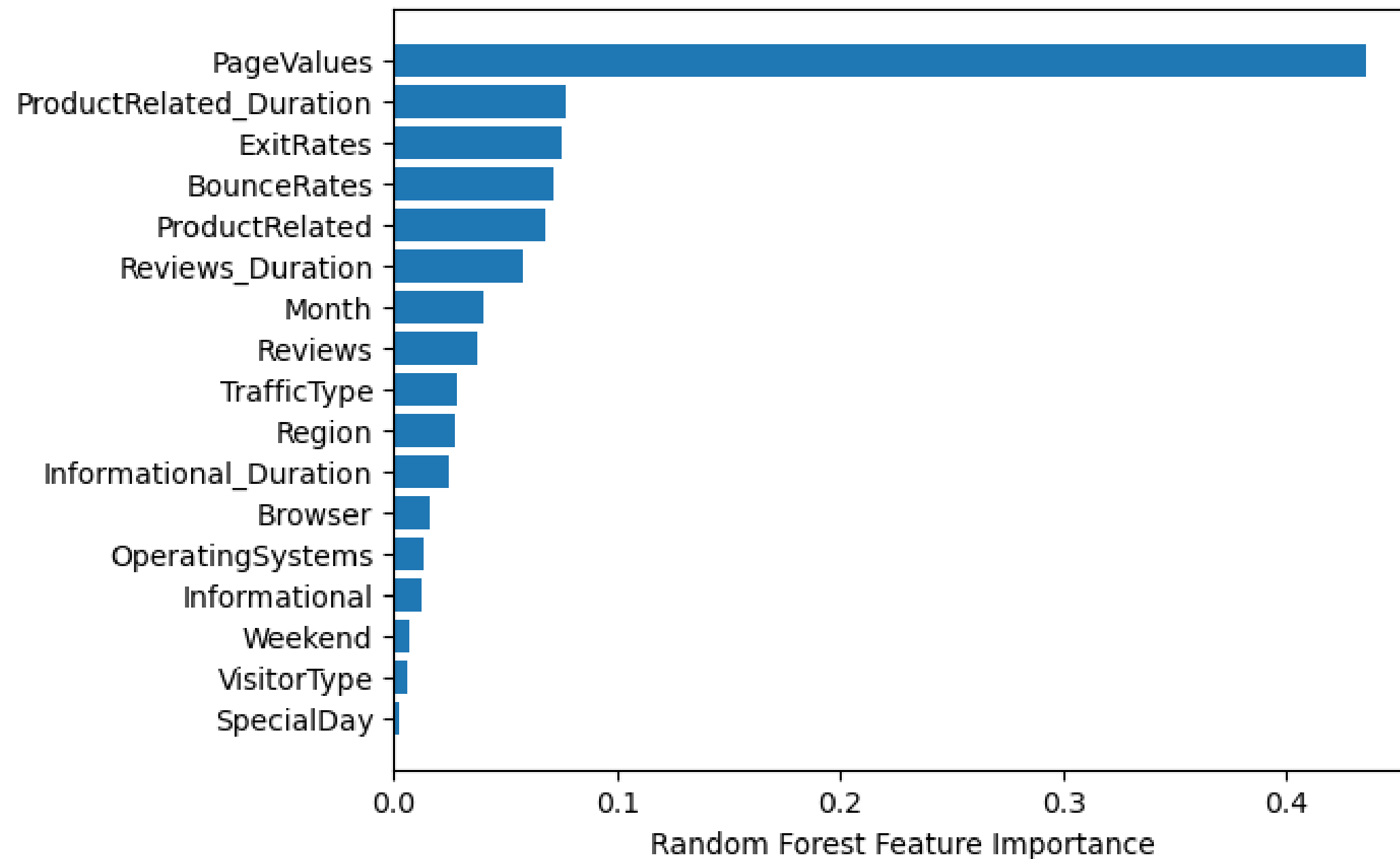
0

```
1 # Separación de características y target (X , y)
2 y = nuevo_df['Purchase']
3 X = nuevo_df.drop(['Purchase'],axis=1)
4
5 # Separación en conjuntos de entrenamiento y validación con 80% de muestras para entrenamiento
6 x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
7
8 #Imprimir Tamaño de dataset
9 print("Tamaño del conjunto de entrenamiento:", x_train.shape)
10 print("Tamaño del conjunto de validación:", x_test.shape)
11
```

Tamaño del conjunto de entrenamiento: (9764, 17)
Tamaño del conjunto de validación: (2441, 17)

- 1.Comprobación y Eliminación de Datos Duplicados:
2. Separación de Características y Objetivo:
3. División de Conjuntos de Entrenamiento y Validación:

Selección de variables



IMPORTANCIA DE
VARIABLES

Selección de variables



Coeficientes del estimador Lasso:

```
[ 0.00000000e+00  1.39668775e-05  0.00000000e+00  0.00000000e+00
 0.00000000e+00  2.35134808e-05 -0.00000000e+00 -0.00000000e+00
 5.13265719e-03 -0.00000000e+00  0.00000000e+00 -0.00000000e+00
 0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00
 0.00000000e+00]
```

```
Index(['Reviews_Duration', 'ProductRelated_Duration', 'PageValues'], dtype='object')
```

Variables seleccionadas:

- Reviews_Duration: que es la cantidad de tiempo dedicado a esta categoría de páginas
- ProductRelated_Duration: es la cantidad de tiempo dedicado a esta categoría de páginas
- PageValues: Métrica arrojada por Google Analytics que representa el valor medio de una página web que un usuario visitó antes de completar una transacción de comercio electrónico

Son las más relevantes según el estimador Lasso y tienen un impacto significativo en la variable objetivo

Selección de variables



Variables seleccionadas:

- Reviews: número de páginas de este tipo (Reviews) que visitó el usuario
- Reviews_Duration: que es la cantidad de tiempo dedicado a esta categoría de páginas
- Informational: número de páginas de este tipo (informativas) que visitó el usuario
- Informational_Duration: cantidad de tiempo dedicado a esta categoría de páginas
- ProductRelated: número de páginas de este tipo (relacionadas con productos) que visitó el usuario
- ProductRelated_Duration: cantidad de tiempo dedicado a esta categoría de páginas
- PageValues: Métrica arrojada por Google Analytics que representa el valor medio de una página web que un usuario visitó antes de completar una transacción de comercio electrónico

Selección de variables



- Month: Mes en el que se realizó la visita al sitio web
- OperatingSystems: Sistema operativo usado por el usuario para navegar en el sitio web
- Browser: Navegador usado por el usuario para navegar en el sitio web
- Region: Región (ubicación geográfica personalizada) desde la cual el usuario navega en el sitio web
- TrafficType: Variable que indica el tipo de trafico al cual pertenece el usuario que navega en el sitio web (por ejemplo, si llegó al sitio desde un anuncio o a través de una búsqueda)
- VisitorType: Tipo de usuario que ingresa al sitio web

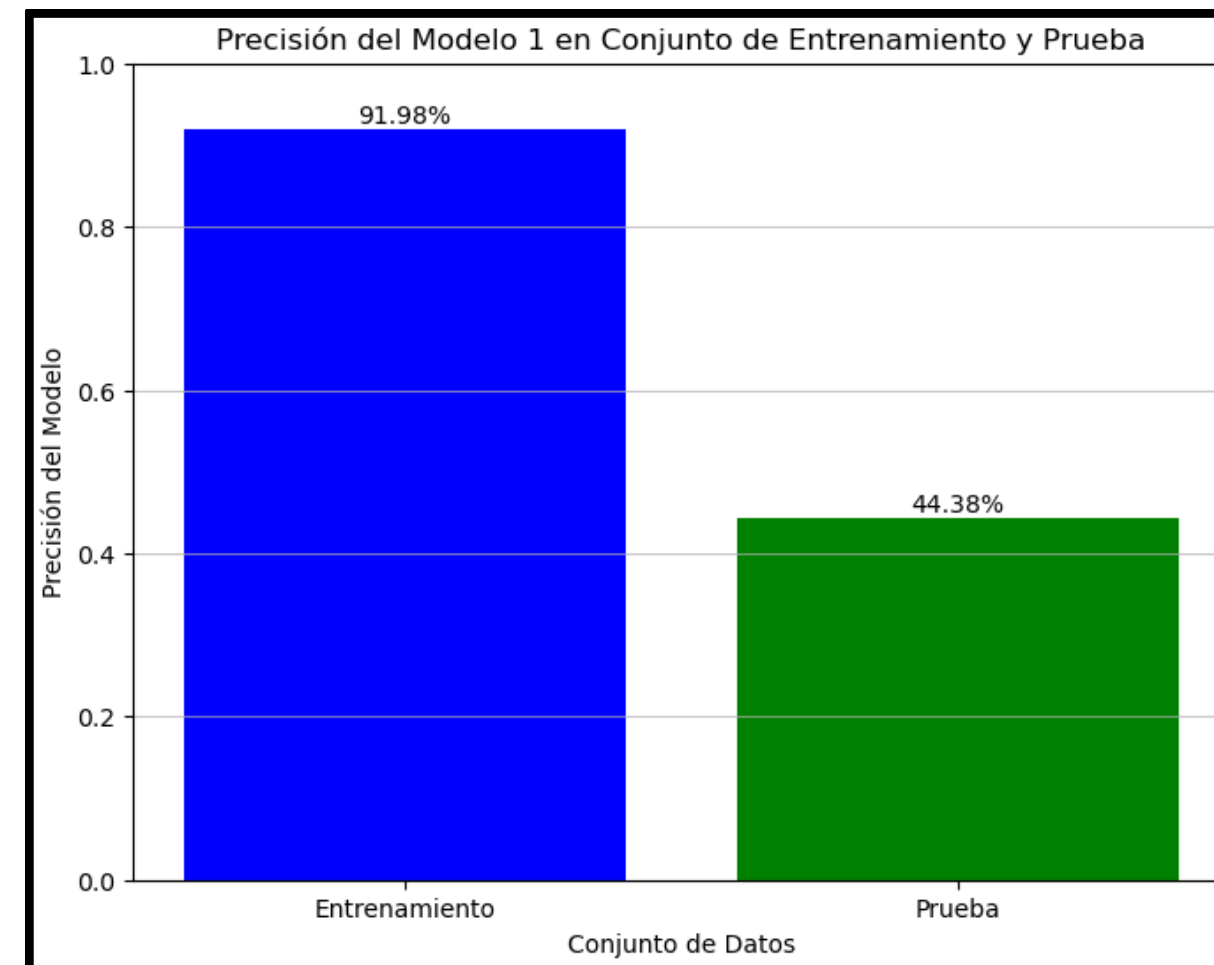
Aplicación y comparación de técnicas de modelado



Modelo 1: RandomForest sin selección de variables

```
RandomForestRegressor  
RandomForestRegressor(max_depth=20, n_estimators=200, random_state=45)
```

Accuracy (Train): 91.97785056254818%
Mean squared error: 0.01
Mean absolute error: 0.05
R2: 91.98
R2-adjusted: 91.96
Accuracy (Test): 44.37664750917888%
Mean squared error: 0.07
Mean absolute error: 0.14
R2: 44.38
R2-adjusted: 43.99



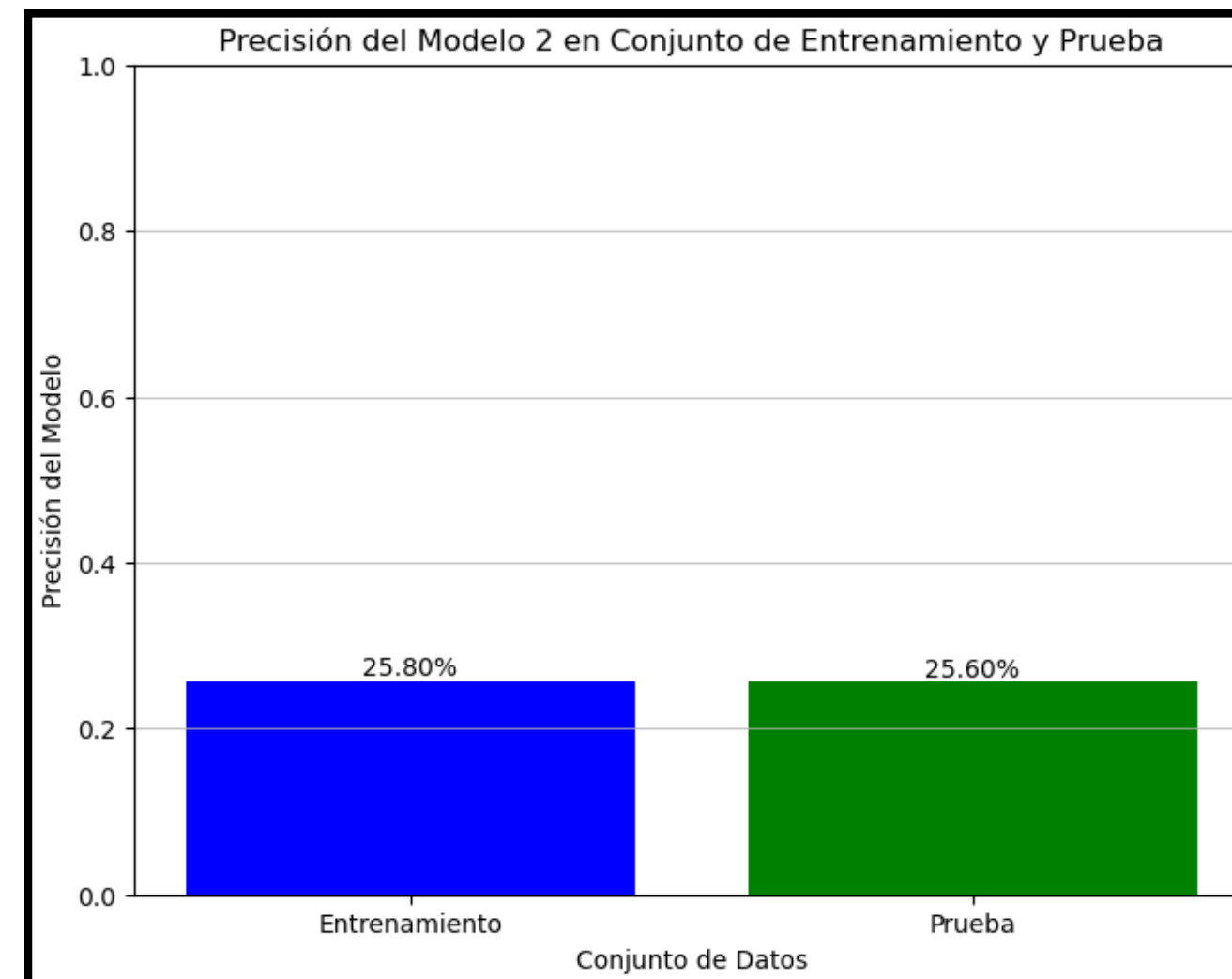
Aplicación y comparación de técnicas de modelado



Modelo 2: Regresión lineal múltiple con selección de variables

Accuracy (Train): 25.796491900956863%
MSE entrenamiento: 0.09805532642273965
MAE entrenamiento: 0.20256115188524898
R2 entrenamiento: 0.25796491900956864

Accuracy (Test): 25.599560145599686%
MSE validación: 0.09737130084768136
MAE validación: 0.20212371432132967
R2 validación: 0.25599560145599687

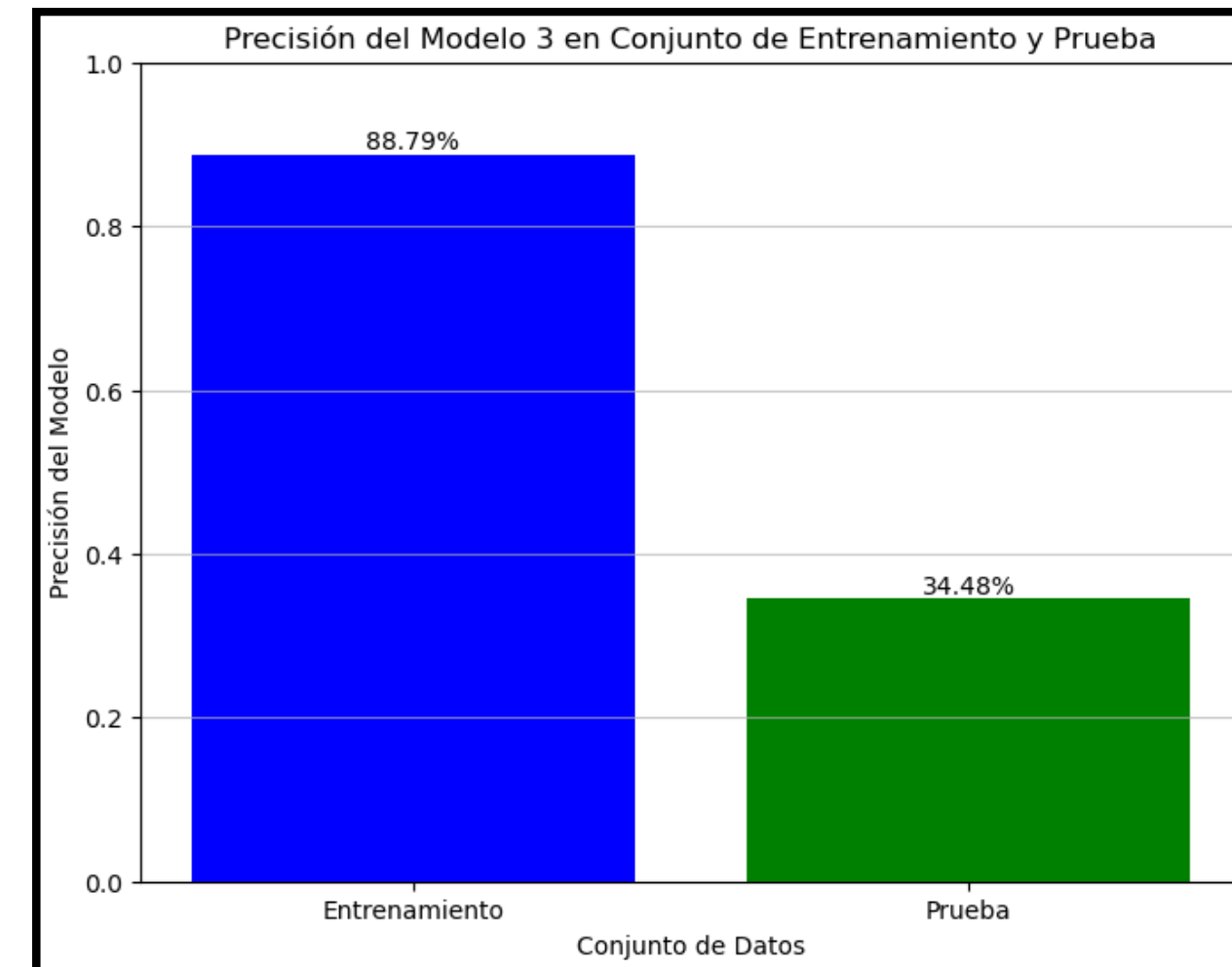


Aplicación y comparación de técnicas de modelado



Modelo 3: RandomForest con selección de variables

Accuracy (Train - Selected Features): 88.78887362221434%
Mean squared error (Train - Selected Features): 0.01
Mean absolute error (Train - Selected Features): 0.06
R2 (Train - Selected Features): 88.79
R2-adjusted (Train - Selected Features): 88.79
Accuracy (Test - Selected Features): 34.48084226155819%
Mean squared error (Test - Selected Features): 0.09
Mean absolute error (Test - Selected Features): 0.15
R2 (Test - Selected Features): 34.48
R2-adjusted (Test - Selected Features): 34.40



Aplicación y comparación de técnicas de modelado



Modelo 4: RandomForest Tuning de Hiperparámetros

Fitting 5 folds for each of 50 candidates, totalling 250 fits

Mejores hiperparámetros encontrados:

```
{'n_estimators': 150, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_depth': 5, 'bootstrap': True}
```

Métricas con el conjunto de entrenamiento:

Accuracy (Train): 51.40521614429934%

Mean squared error (Train): 0.06

Mean absolute error (Train): 0.13

R2 (Train): 51.41

R2-adjusted (Train): 51.32

Métricas con el conjunto de validación:

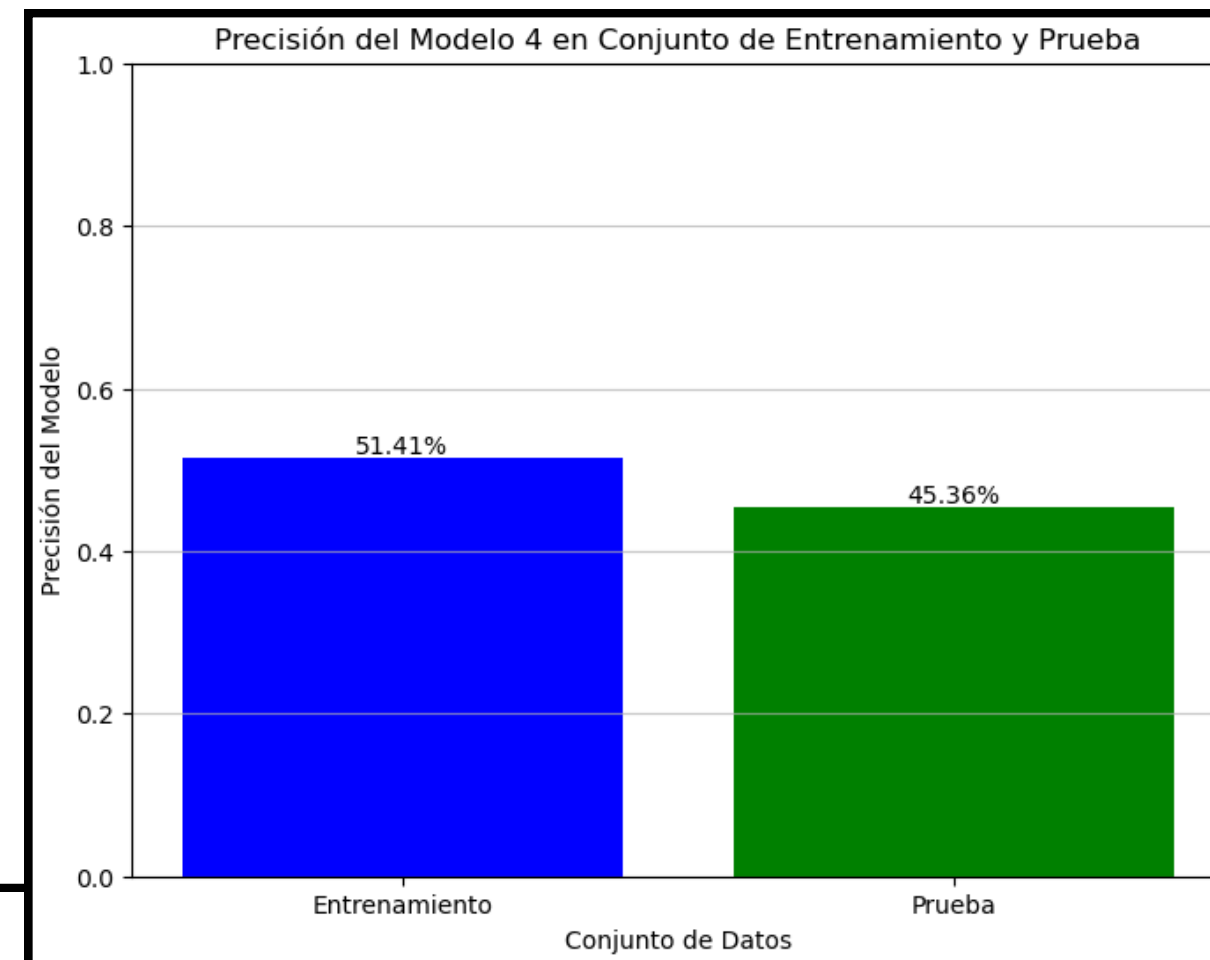
Accuracy (Test): 45.35865054384046%

Mean squared error (Validation): 0.07

Mean absolute error (Validation): 0.14

R2 (Validation): 45.36

R2-adjusted (Validación): 44.98

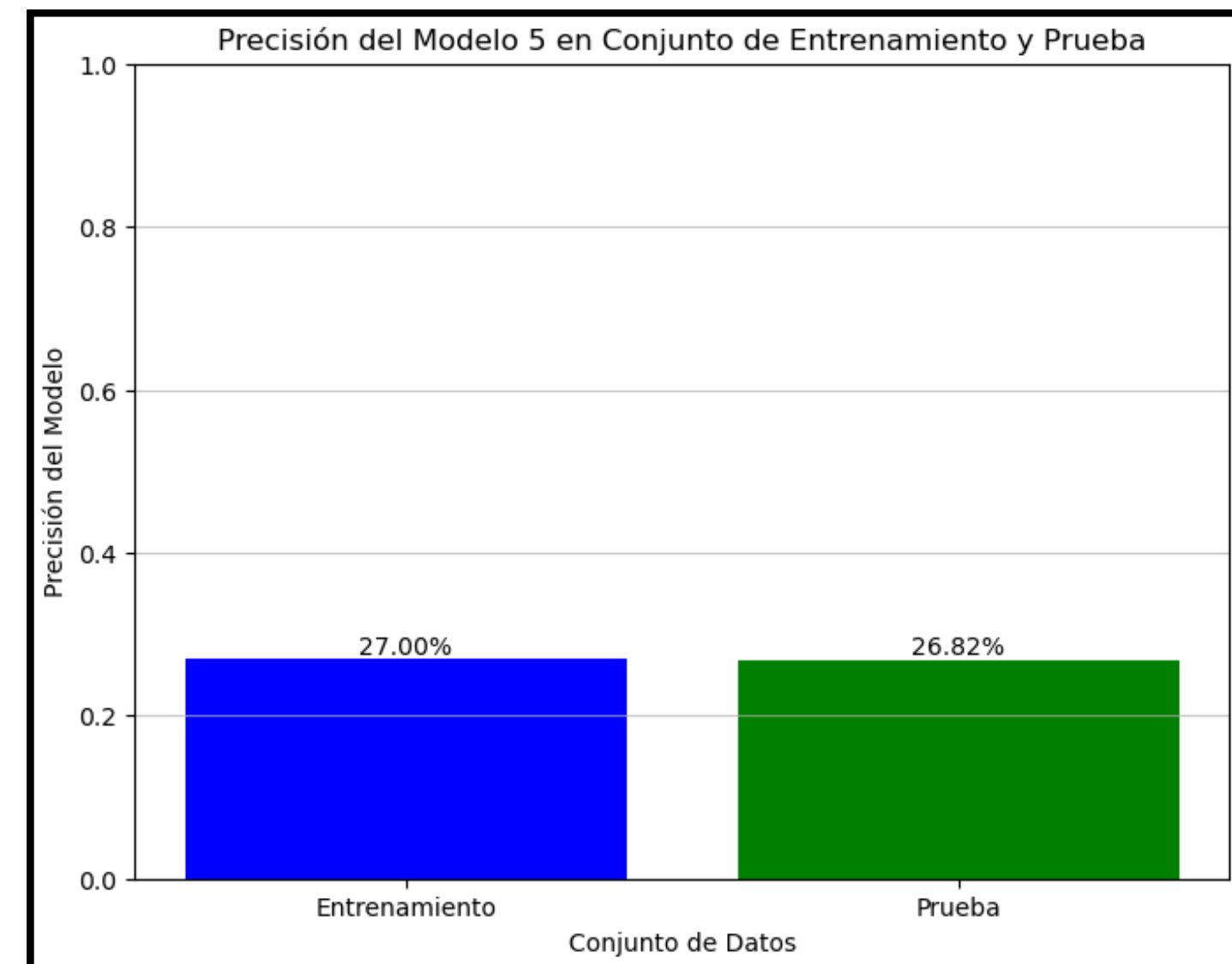


Aplicación y comparación de técnicas de modelado



Modelo 5: Regresión lineal multiples con características Variance Threshold

```
Accuracy (Train): 26.995984506780303%
Accuracy (Test): 26.820169949491902%
Métricas del modelo:
R2 score (Train): 26.995984506780303
R2 score (Test): 26.820169949491902
MAE (Train): 0.20066303634238034
MAE (Test): 0.19526569270688016
MSE (Train): 0.09728969314429613
MSE (Test): 0.09243126517706826
```

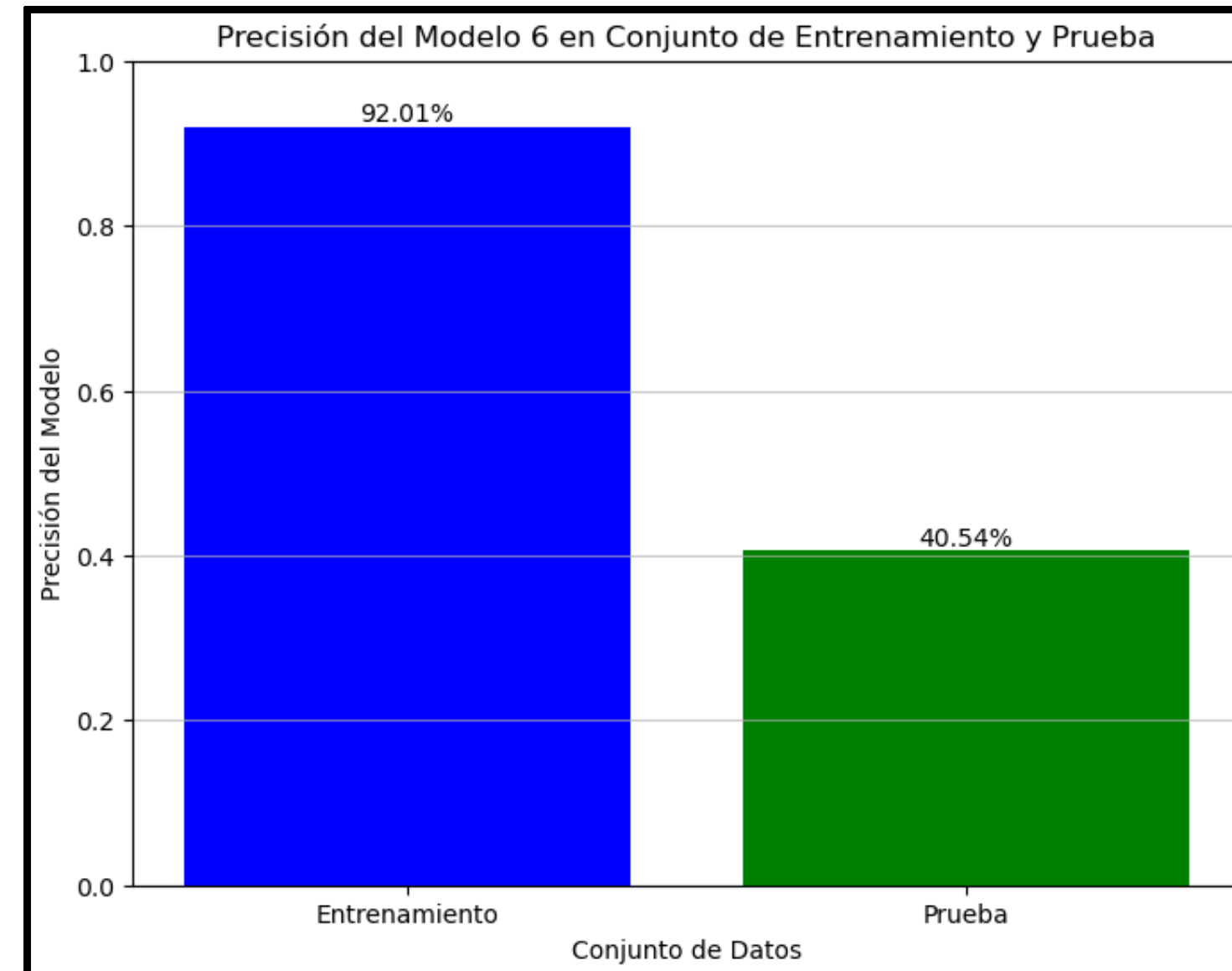


Aplicación y comparación de técnicas de modelado



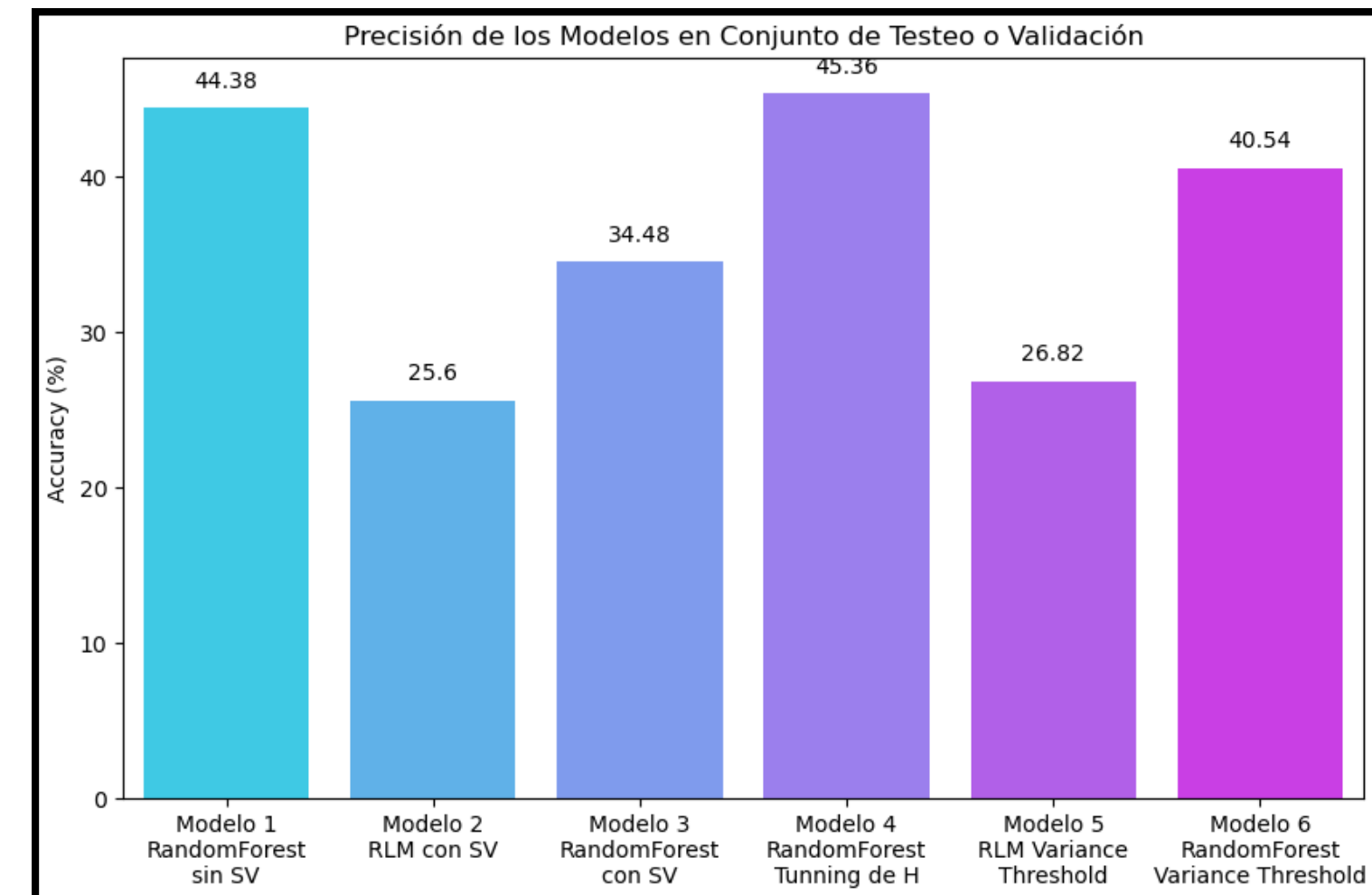
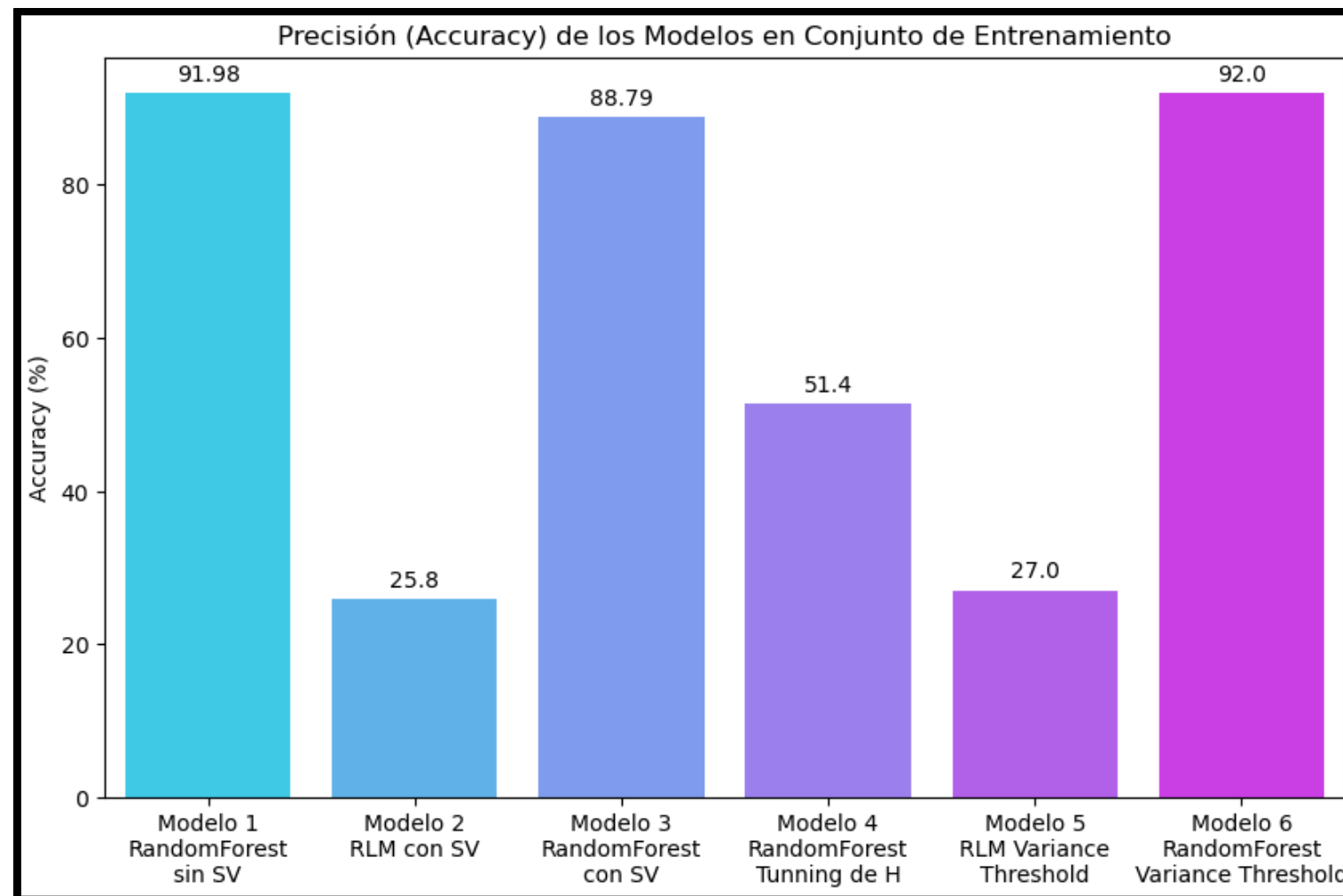
Modelo 6: RandomForest Variance Threshold

Accuracy (Train): 92.00751692636963%
Accuracy (Test): 40.54293961716613%
Métricas del modelo Random Forest Regressor:
R2 score (Train): 92.00751692636963
R2 score (Test): 40.54293961716613
MAE (Train): 0.05440905366653011
MAE (Test): 0.1451249487914789
MSE (Train): 0.010651280213027449
MSE (Test): 0.07509844326095863



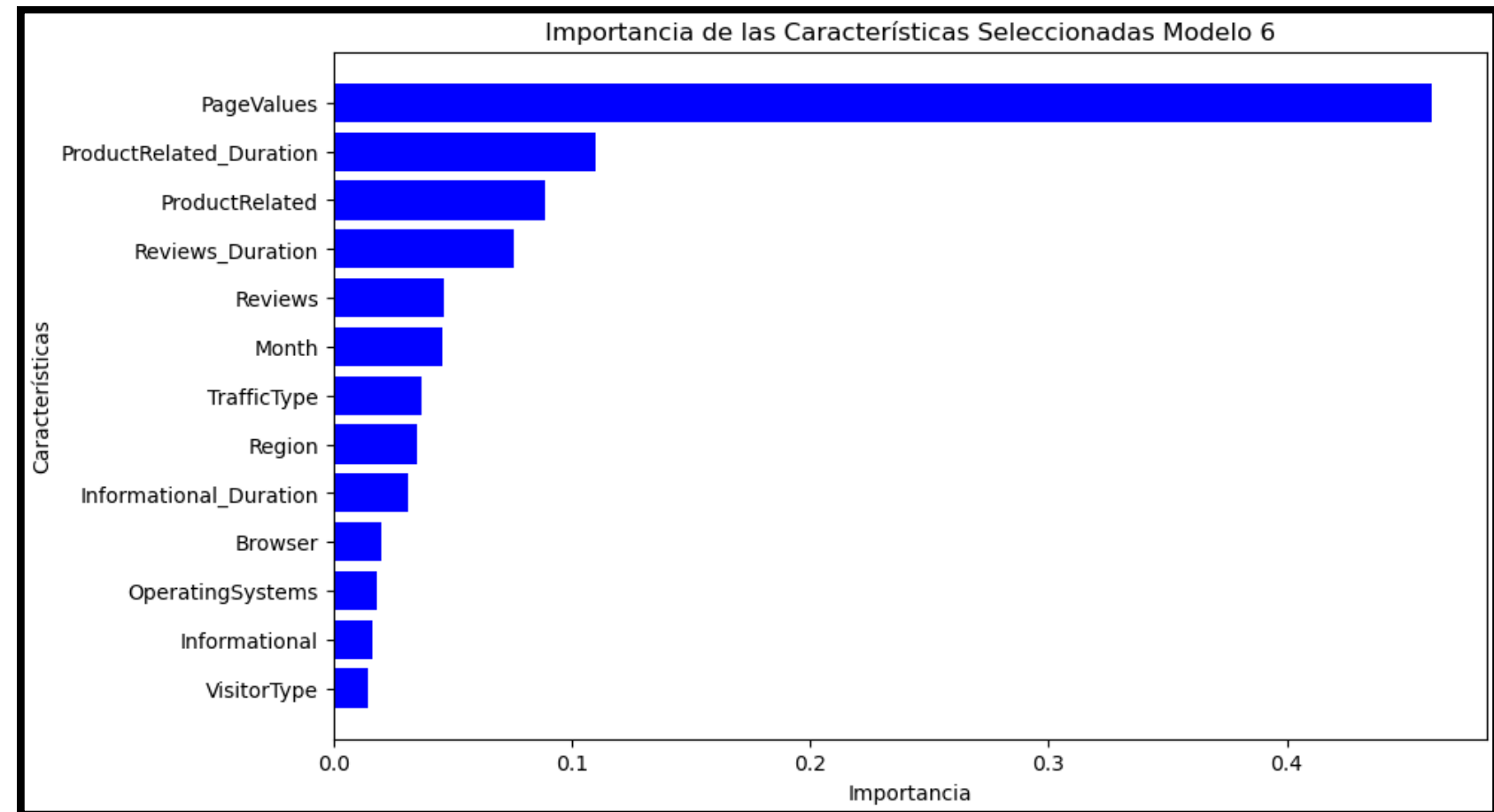
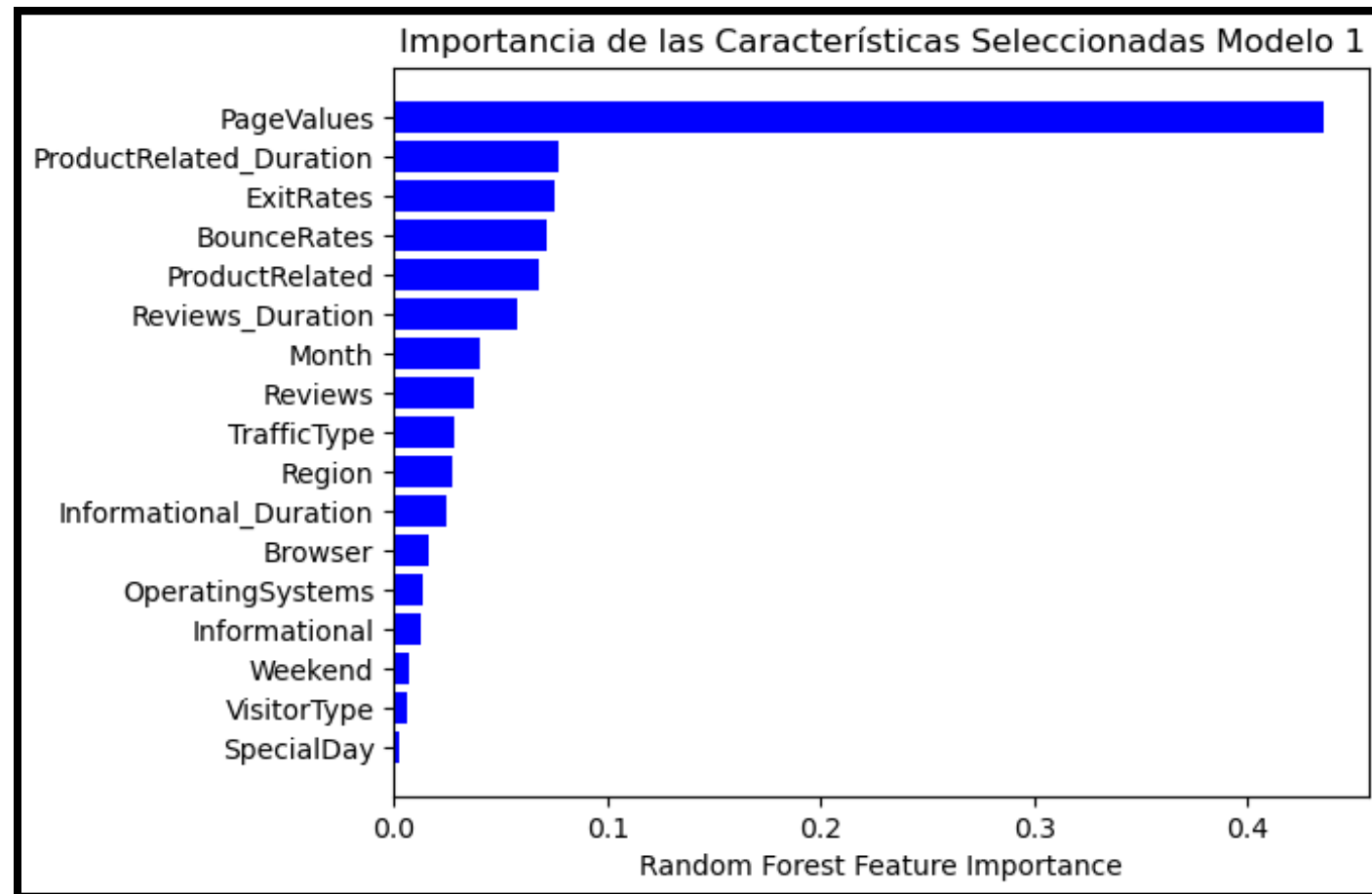


Evaluación del mejor modelo



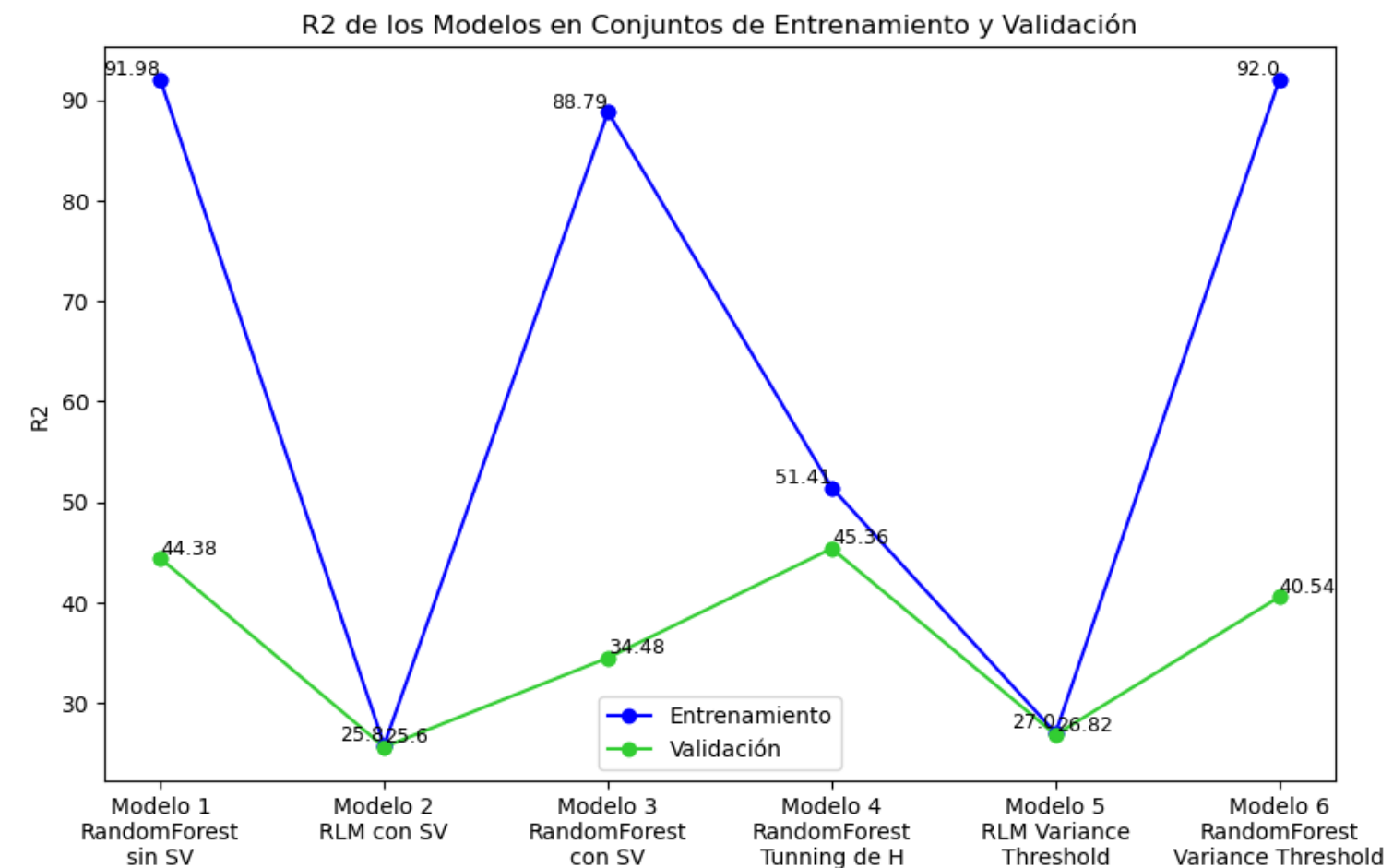
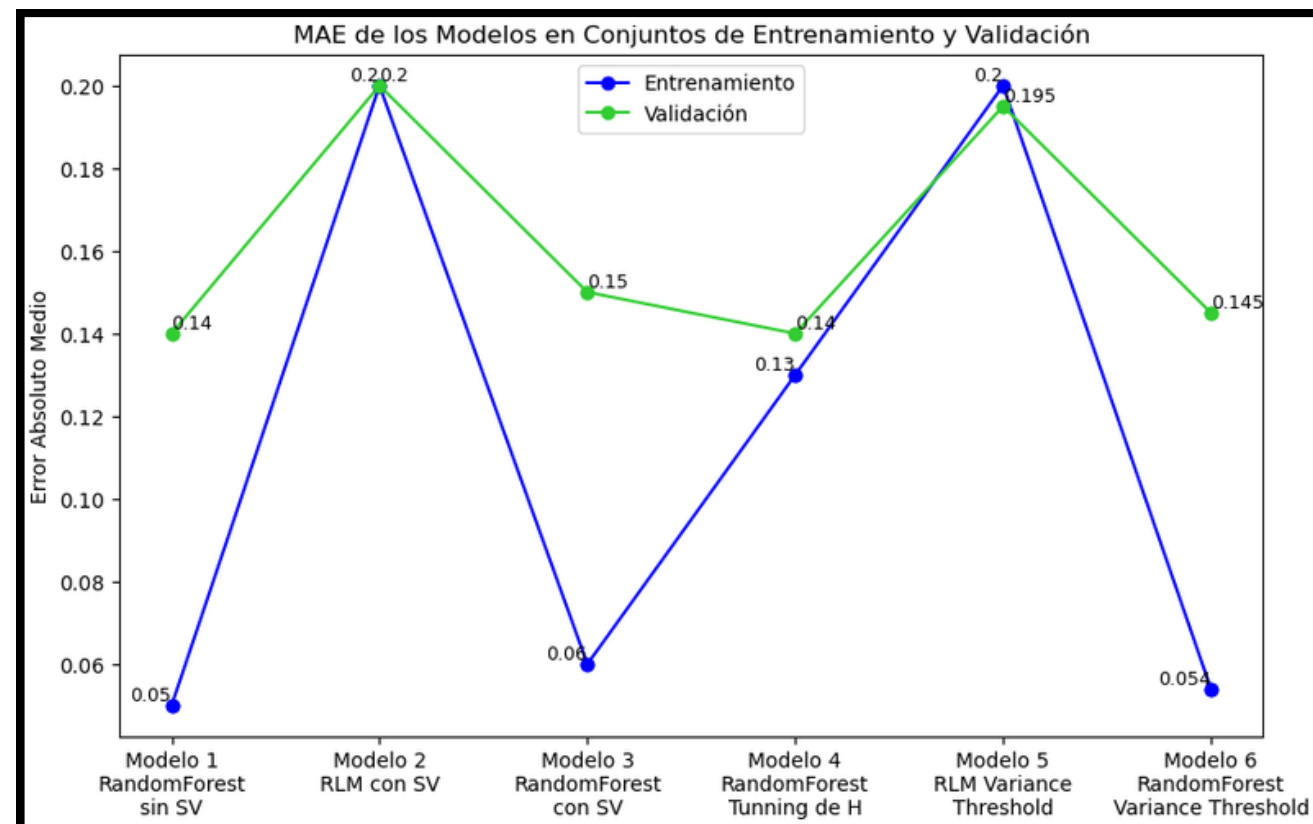
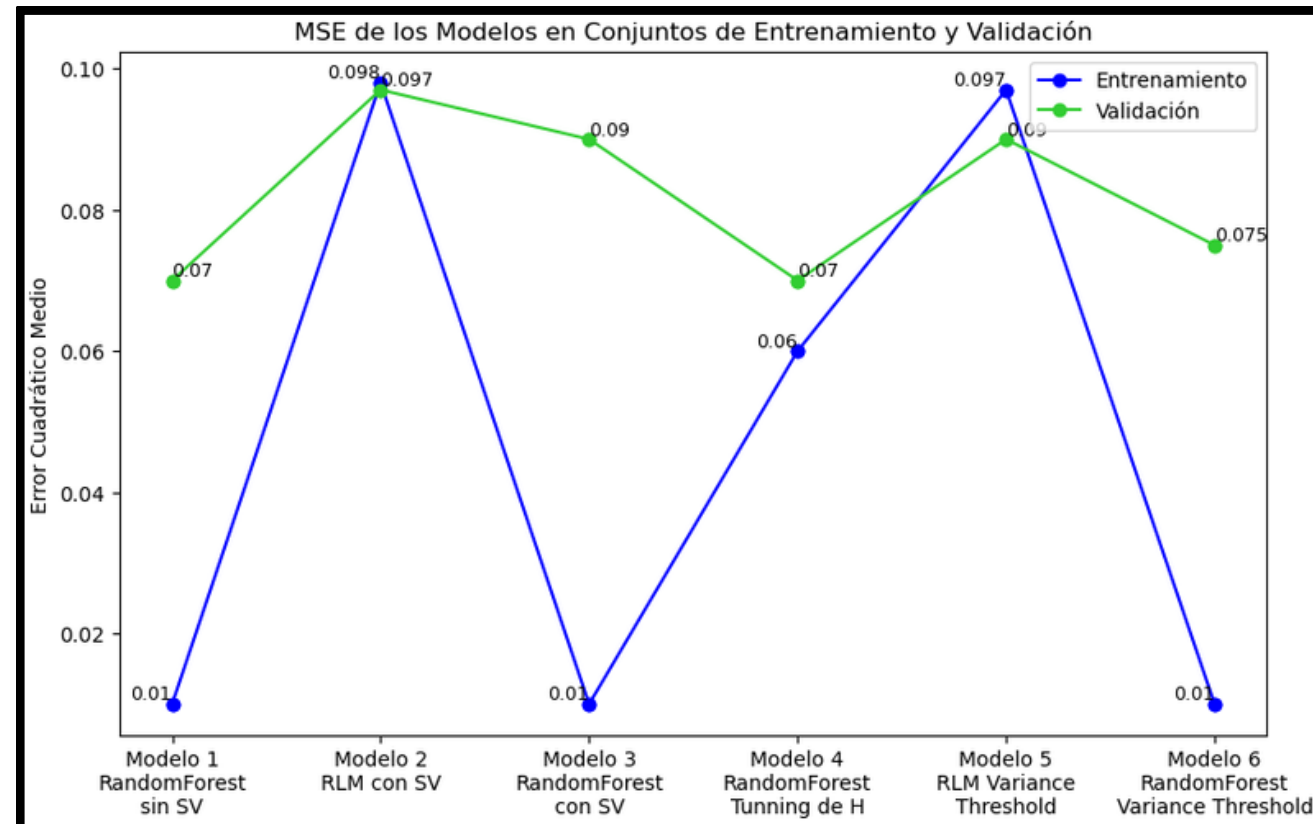
El modelo 1 y el 6 presentan la mayor precisión respecto a los otros modelos para el conjunto de entrenamiento y para el de validación los modelos 1, 4 y 6.

Evaluación del mejor modelo



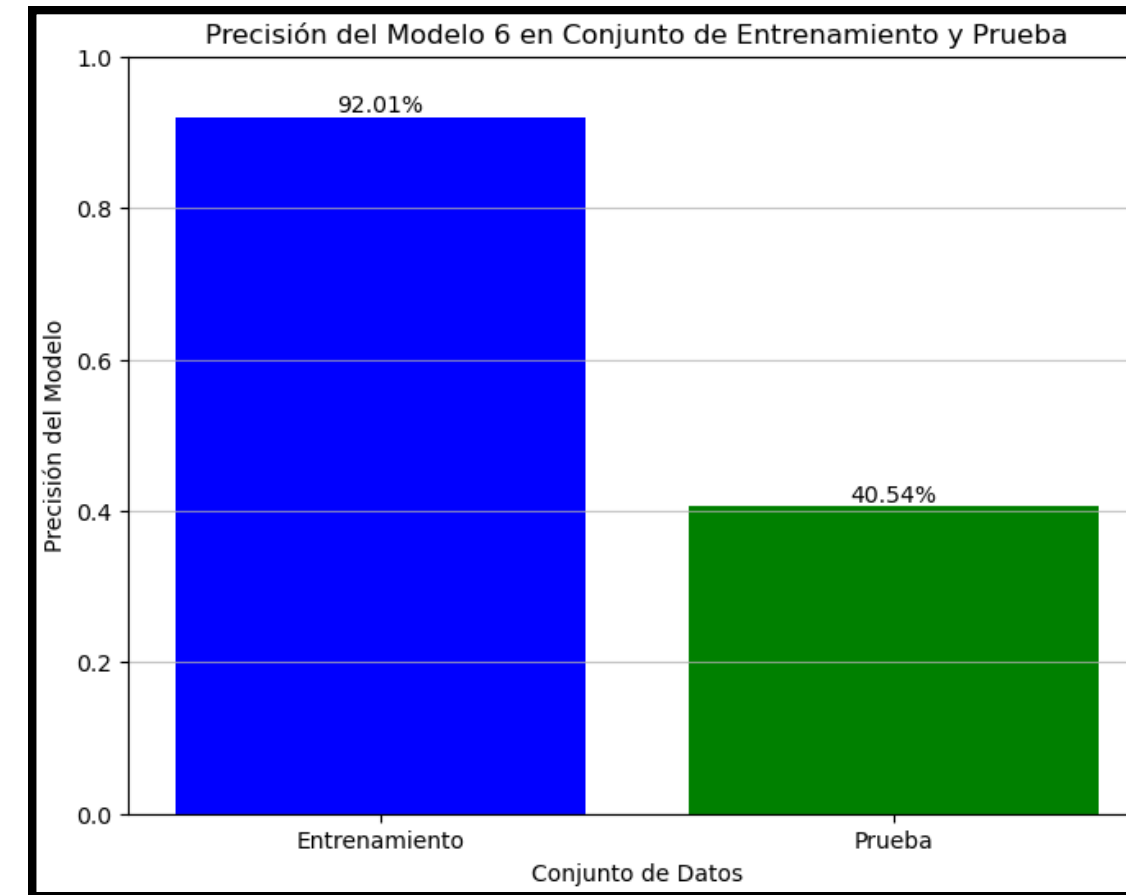
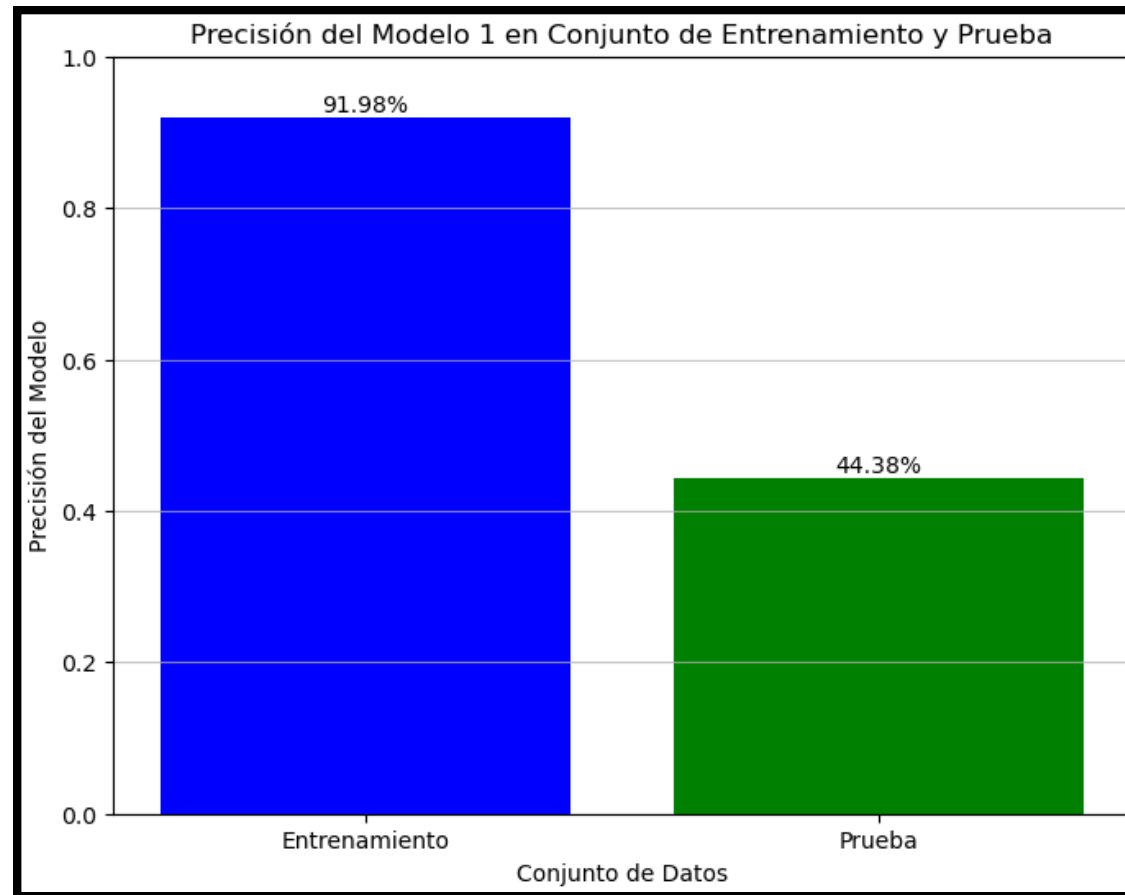
El modelo 1 y el 6 presentan diferencias en las características usadas, mientras el modelo 1 usa todas las 17, el modelo 6 usa 13 y varían su importancia.

Evaluación del mejor modelo



Para el modelo 1 y 6 el MSE y el MAE obtuvieron valores muy similares, la mayor diferencia se encuentra en el R2.

Evaluación del mejor modelo



El Modelo 1 sobresale en capacidad predictiva general con un alto (R^2), mientras que el Modelo 6, a pesar de un (R^2) más bajo, podría ser preferible para identificar clientes potenciales y optimizar la inversión en publicidad digital. La elección entre ambos depende de si se prioriza la precisión predictiva o la identificación de clientes.

Conclusiones y recomendaciones



1. Estrategias de Marketing:

Basado en los resultados de los modelos, hay margen de mejora en las estrategias de marketing digital. Los datos sugieren que la nueva estrategia de marketing, implementando el modelo 6 podría ser más efectiva ya que tiene un enfoque en la personalización y la segmentación para captar nuevos clientes con alta intención de compra.

2. Utilización de Modelos de Machine Learning:

Los modelos de ML en este caso son valiosos para identificar clientes potenciales. El Modelo 6, en particular, es prometedor en la identificación de estos clientes, optimizando así la inversión en publicidad digital. Este modelo destaca por su capacidad predictiva y debería ser el foco de futuras inversiones y desarrollos.

Conclusiones y recomendaciones



3. Refinamiento del Modelo:

Aunque el Modelo 6 parece ser el más efectivo hasta ahora, presenta oportunidades de mejora. Se pueden explorar otras técnicas de ajuste de hiperparámetros y evaluar la implementación de algoritmos alternativos para maximizar la precisión y el retorno de la inversión.

4. Monitoreo Continuo:

El dinamismo del mercado requiere una adaptación constante dado que las preferencias de los clientes pueden cambiar con el tiempo, esto implicaría la toma de datos y nuevas variables que puedan incidir en la compra del cliente. Es crucial implementar un sistema de monitoreo que permita la actualización continua del modelo en respuesta a las nuevas tendencias de consumo y cambios en el entorno de mercado.