Proyecto MAT-281

Titanic Espacial

Martin Alonso Urrea Barros

Universidad Tecnica Federico Santa Maria

Lunes 4 de Diciembre 2023

Tabla de Contenidos

Tabla de Contenidos

Contexto

- La nave espacial Titanic con casi 13.000 pasajeros a bordo. En direccion a 3 exoplanetas.
- Tras chocar con una anomalia del espacio-tiempo, casi la mitad de los pasajeros fueron transportados a una dimension alternativa.

Figura: Imagen del Titanic Espacial

Carga de Datos

Librerias:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Missingno

De Sklearn:

- Modelos
- Metricas
- Preprocesamiento
- Impute

Datos a Manejar:

- Train
- Test

Tabla de Contenidos

Estadistica Descriptiva

Dimensiones:

- Entrenamiento $\rightarrow$ Train: (8693, 14)
- Prueba $\rightarrow$ Test: (4277, 13)

Columnas:

- PassengerId: object
- HomePlanet: object
- CryoSleep: object
- Cabin: object
- Destination: object
- Age: float64
- VIP: object

Columnas:

- RoomService: float64
- FoodCourt: float64
- ShoppingMall: float64
- Spa: float64
- VRDeck: float64
- Name: object
- Transported: bool

Tabla de Contenidos

Visualizacion Descriptiva

Segun los datos del conjunto de Entrenamiento, aproximadamente la mitad de las personas fueron transportadas.

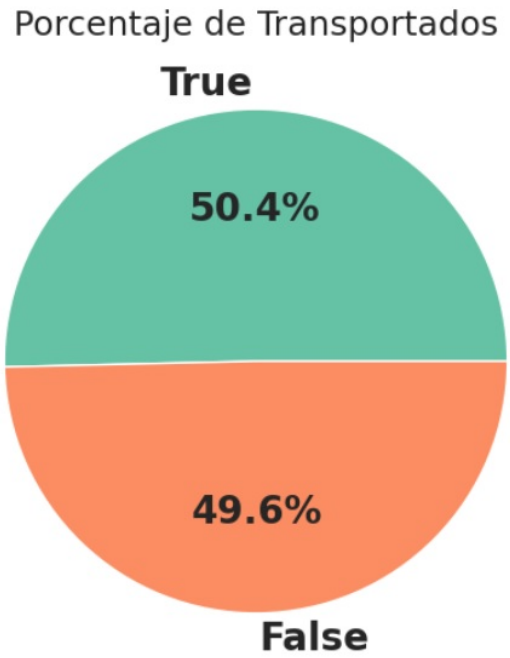


Figura: Porcentaje de Transportados

Visualizacion Descriptiva

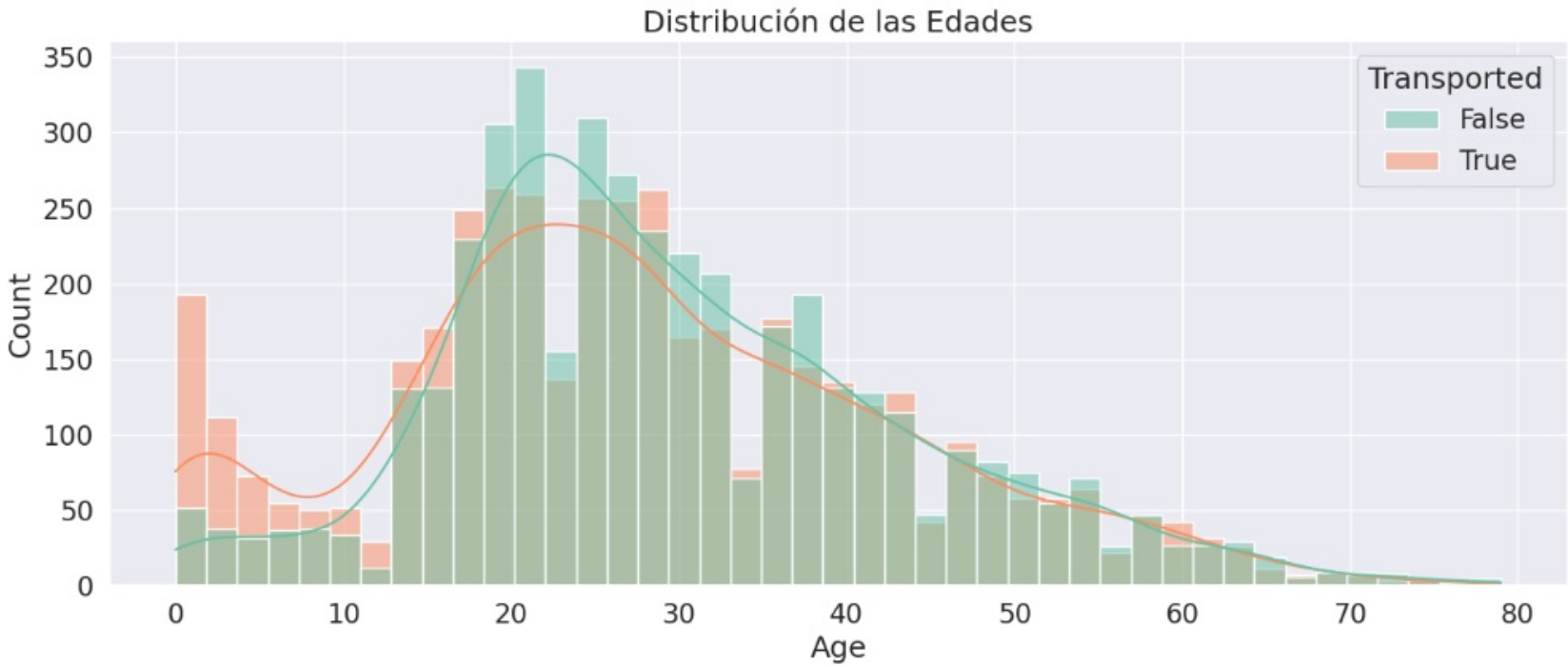


Figura: Distribucion de Edades

Visualizacion Descriptiva

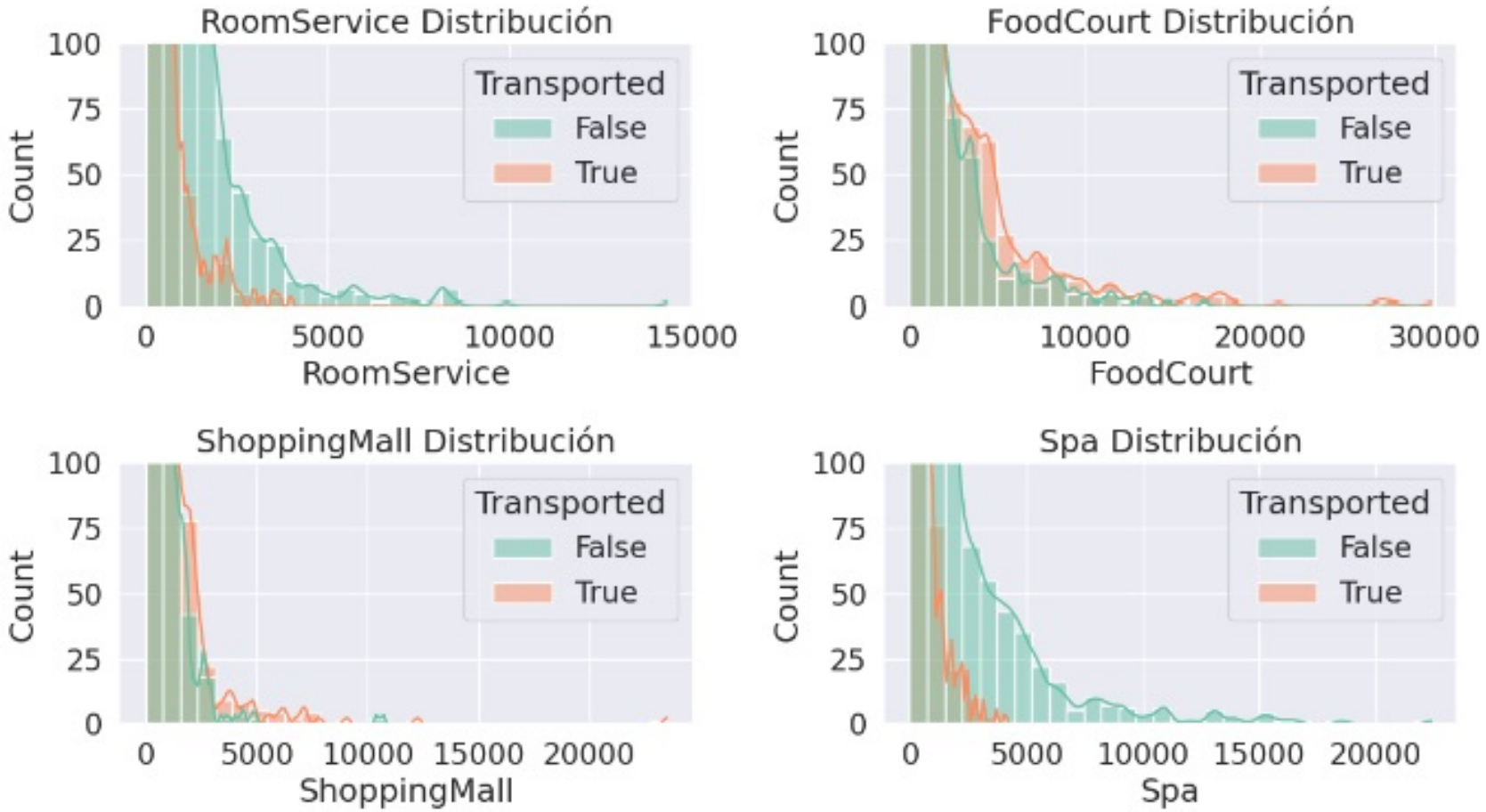


Figura: Distribucion de Gastos

Visualizacion Descriptiva

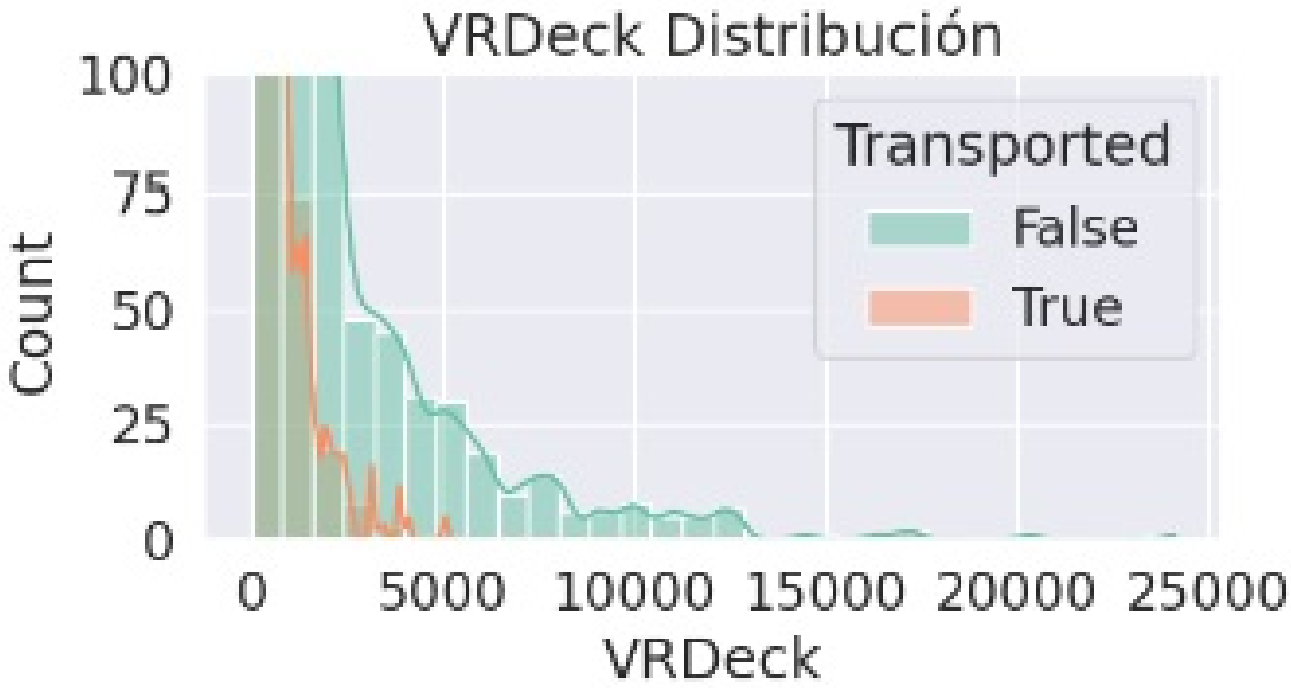


Figura: Distribucion de Gastos

Visualizacion Descriptiva

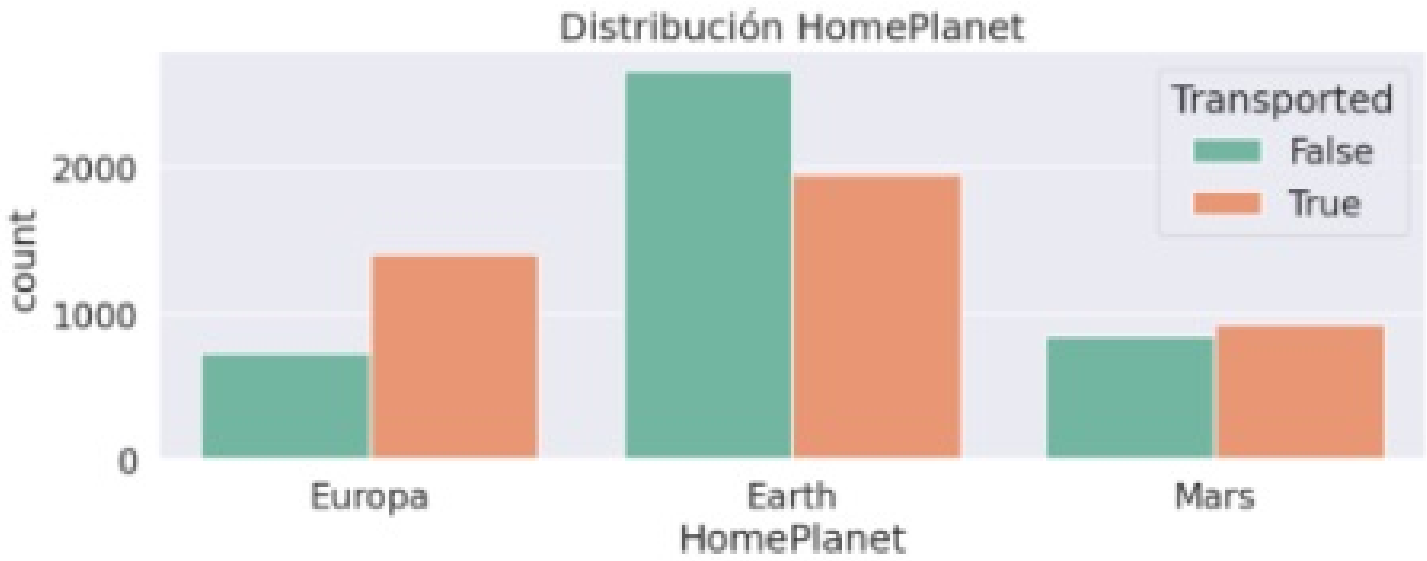


Figura: Distribucion de Lugar de Origen

Visualizacion Descriptiva

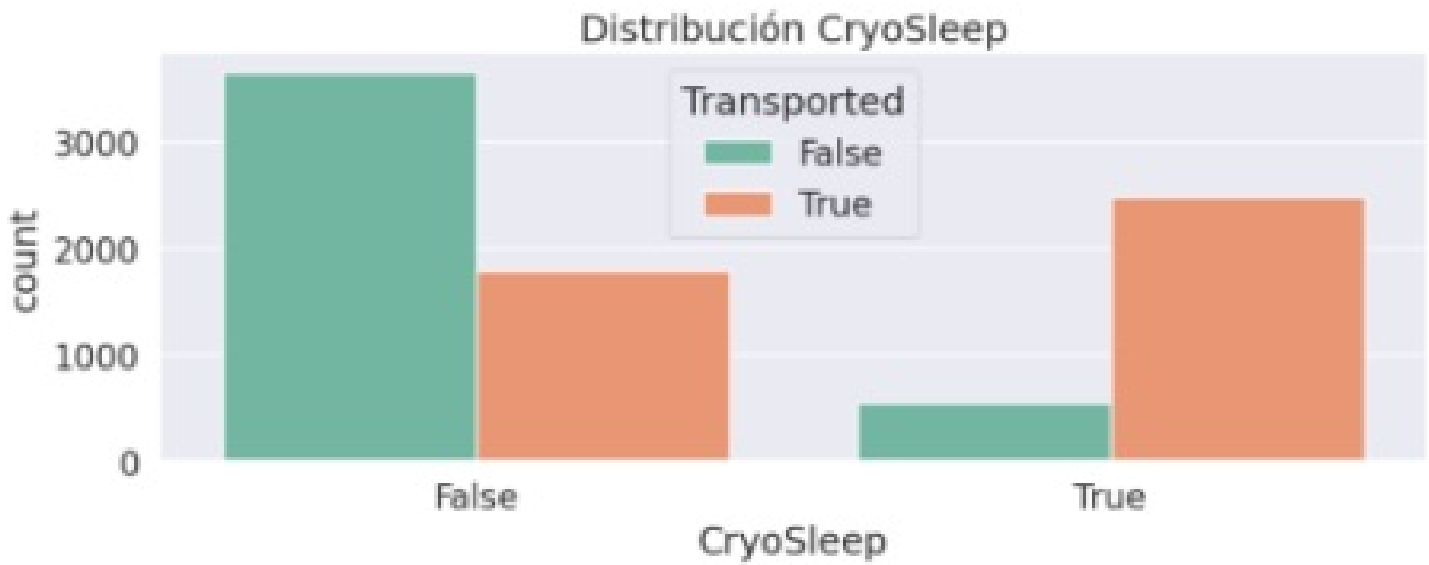


Figura: Distribucion de CryoSleep

Visualizacion Descriptiva

Figura: Distribucion de Destino

Visualizacion Descriptiva

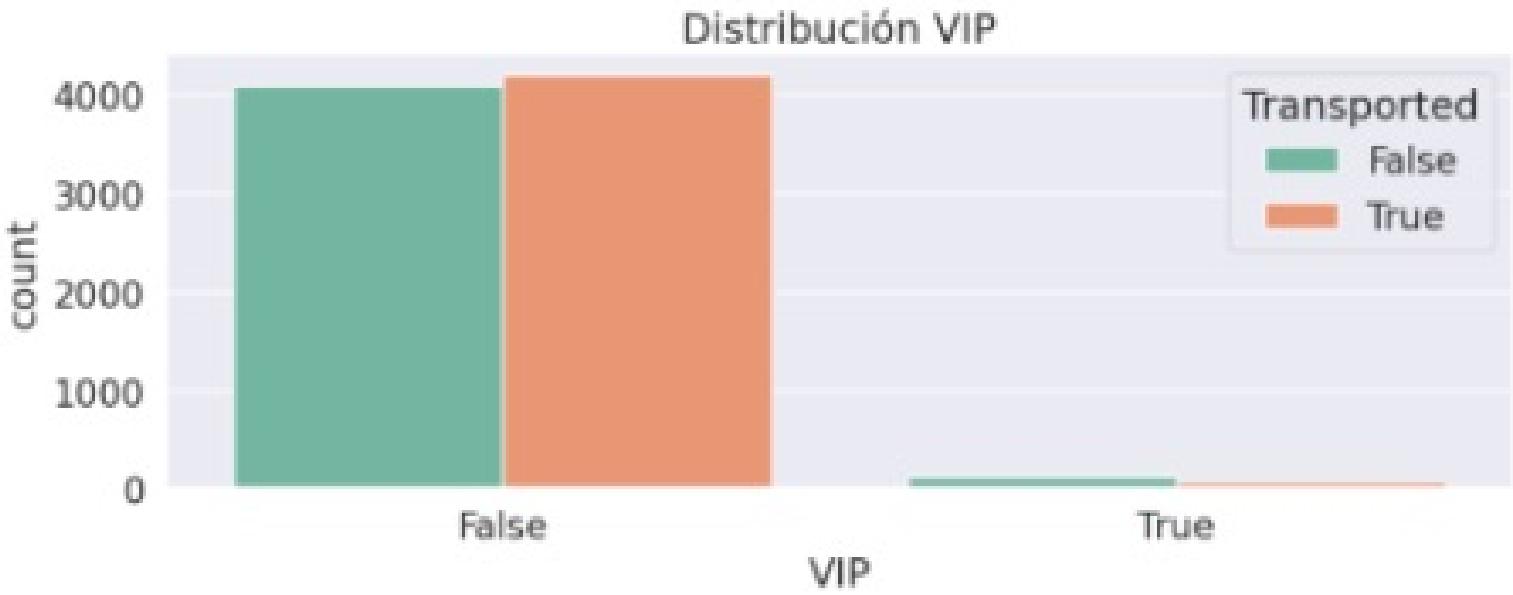


Figura: Distribucion de Pasajeros VIP

Tabla de Contenidos

Ingenieria de Atributos

Creacion de Nuevas Variables:

Reinterpretamos 'PassengerId' que viene en el siguiente formato: gggg-pp

- 'Group_Size"
- "Travelling_Solo"

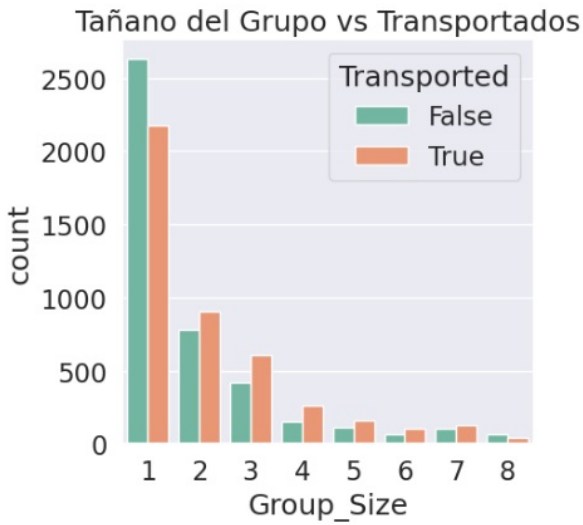


Figura: Tamaño del Grupo vs Transportados

Ingenieria de Atributos

Creacion de Nuevas Variables:

Reinterpretamos ‘cabin’ que viene en el siguiente formato: deck/num//side

- “Cabin_Deck”
- “Cabin_Number”
- “Cabin_Side”

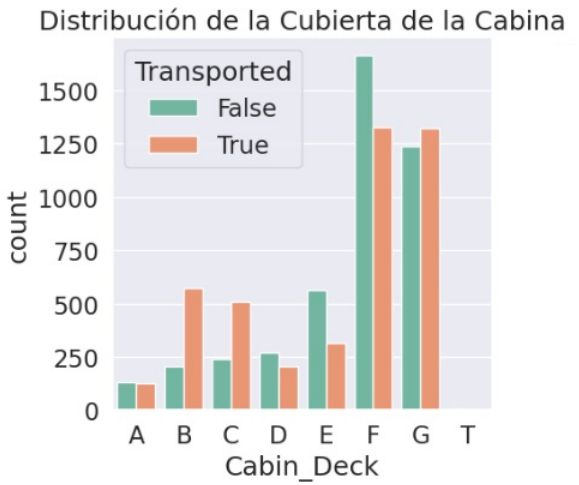


Figura: Distribucion de la Cubierta de la Cabina

Ingenieria de Atributos

Creacion de Nuevas Variables:

- Distribuimos las Cabinas por Regiones del 1 al 6
- Distribuimos las Edades por Rangos Etarios
- Fusionamos los Gastos

Ingenieria de Atributos

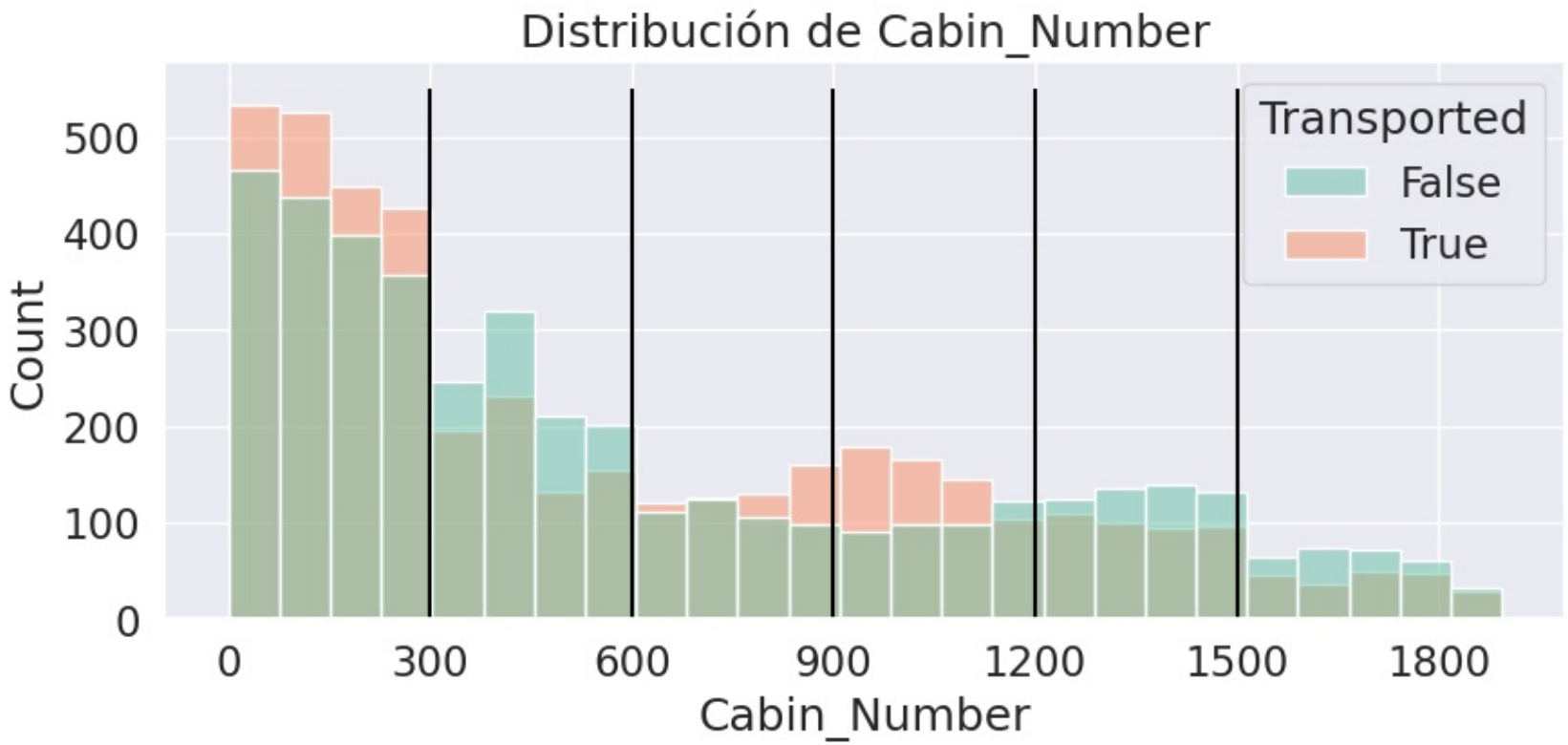


Figura: Distribucion de Regiones de la Cabina

Ingenieria de Atributos

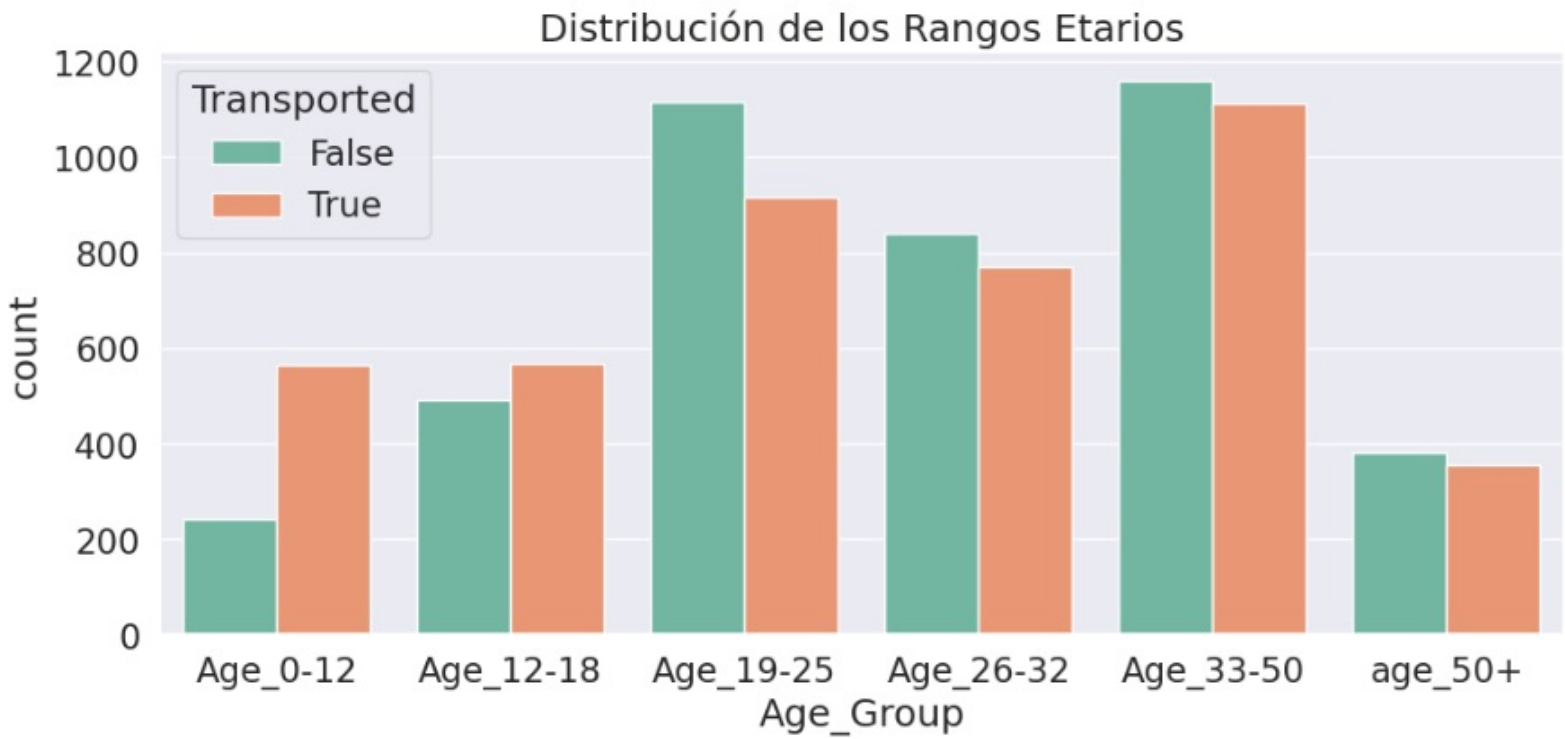


Figura: Distribucion de Rangos Etarios

Ingenieria de Atributos

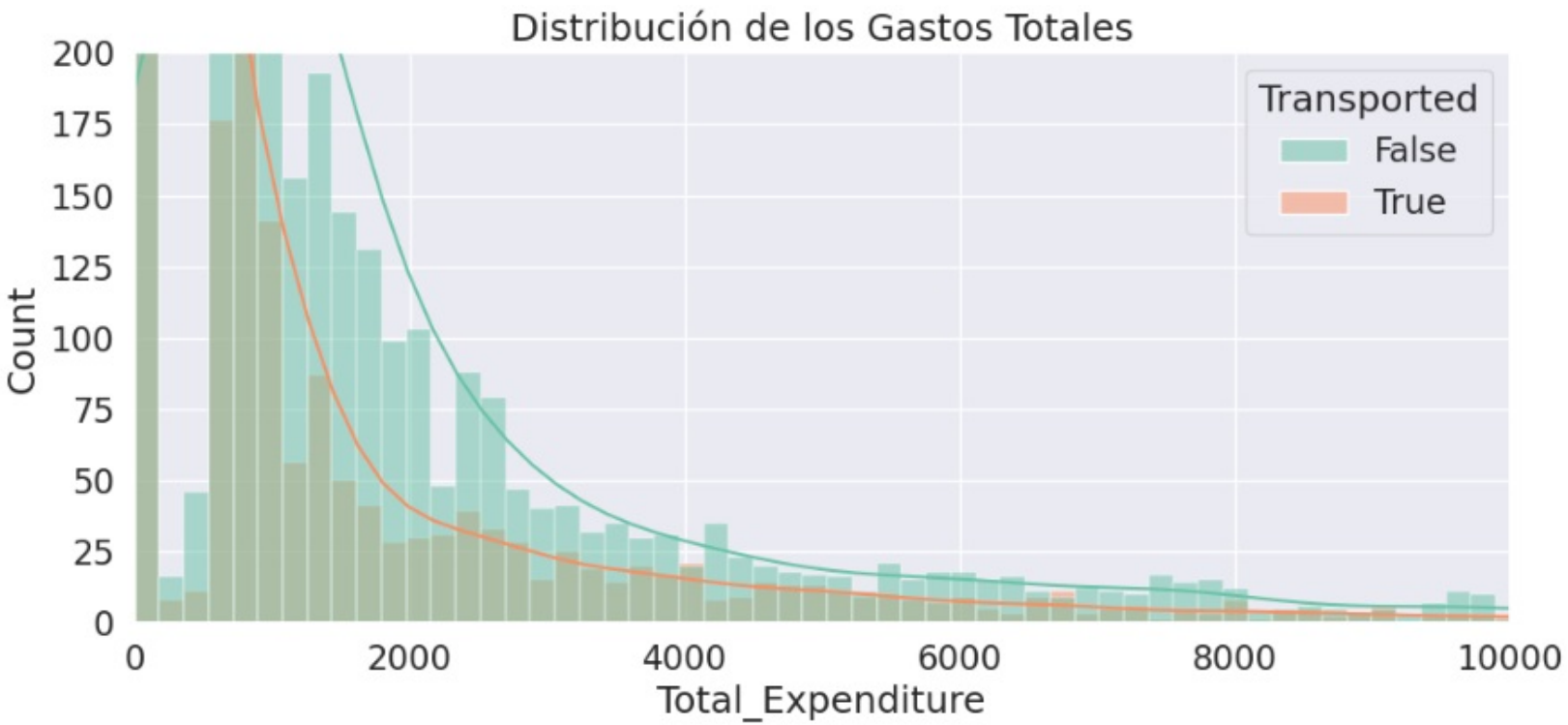


Figura: Gastos Totales

Pre-procesamiento

Informacion:

- Cambiamos funciones de Tipo Objeto a Tipo Booleano
- Cambianos funciones de Tipo Int a Tipo Float

Ademas: (usando missingno)

- Se completaron los datos faltantes con lo mas frecuente (datos categoricos)
- Se completaron los datos faltantes con la mediana (datos numericos)

Pre-procesamiento

Informacion:

- Aplicamos una Transformacion Logaritmica a las variables de Gasto
- Hacemos One Hot Encoding para las variables categoricas nominales.
- Hacemos LabelEncoding para las variables categoricas ordinales.

Ademas:

Tambien pasando a Booleano los datos que falten

Tabla de Contenidos

Modelo para Datos Estandarizados

Pasos:

- Estandarizamos con StandardScaler
- Train Test Split
- Creamos una funcion que nos entregara las metricas

Modelos:

- Regresion Logistica
- Support Vector Machine

Metricas y Analisis de Resultados

Modelo de Regresion Logistica:

- Accuracy_Score del conjunto de Entrenamiento es: 77.86
- Accuracy_Score del conjunto de Testeo es: 77.17
- Precision Score es: 0.75
- Recall Score es: 0.80
- F1 Score es: 0.78

Figura: Matriz de Confusion

Metricas y Analisis de Resultados

Modelo de SVM:

- Accuracy_Score del conjunto de Entrenamiento es: 81.82
- Accuracy_Score del conjunto de Testeo es: 79.70
- Precision Score es: 0.79
- Recall Score es: 0.79
- F1 Score es: 0.79

Figura: Matriz de Confusion

Modelo para Datos NO Estandarizados

Pasos:

- Train Test Split
- Creamos una funcion que nos entregara las metricas

Modelos:

- Random Forest
- Gradient Boosting

Metricas y Analisis de Resultados

Modelo de Random Forest:

- Accuracy_Score del conjunto de Entrenamiento es: 98.51
- Accuracy_Score del conjunto de Testeo es: 80.56
- Precision Score es: 0.82
- Recall Score es: 0.77
- F1 Score es: 0.80

Figura: Matriz de Confusion

Metricas y Analisis de Resultados

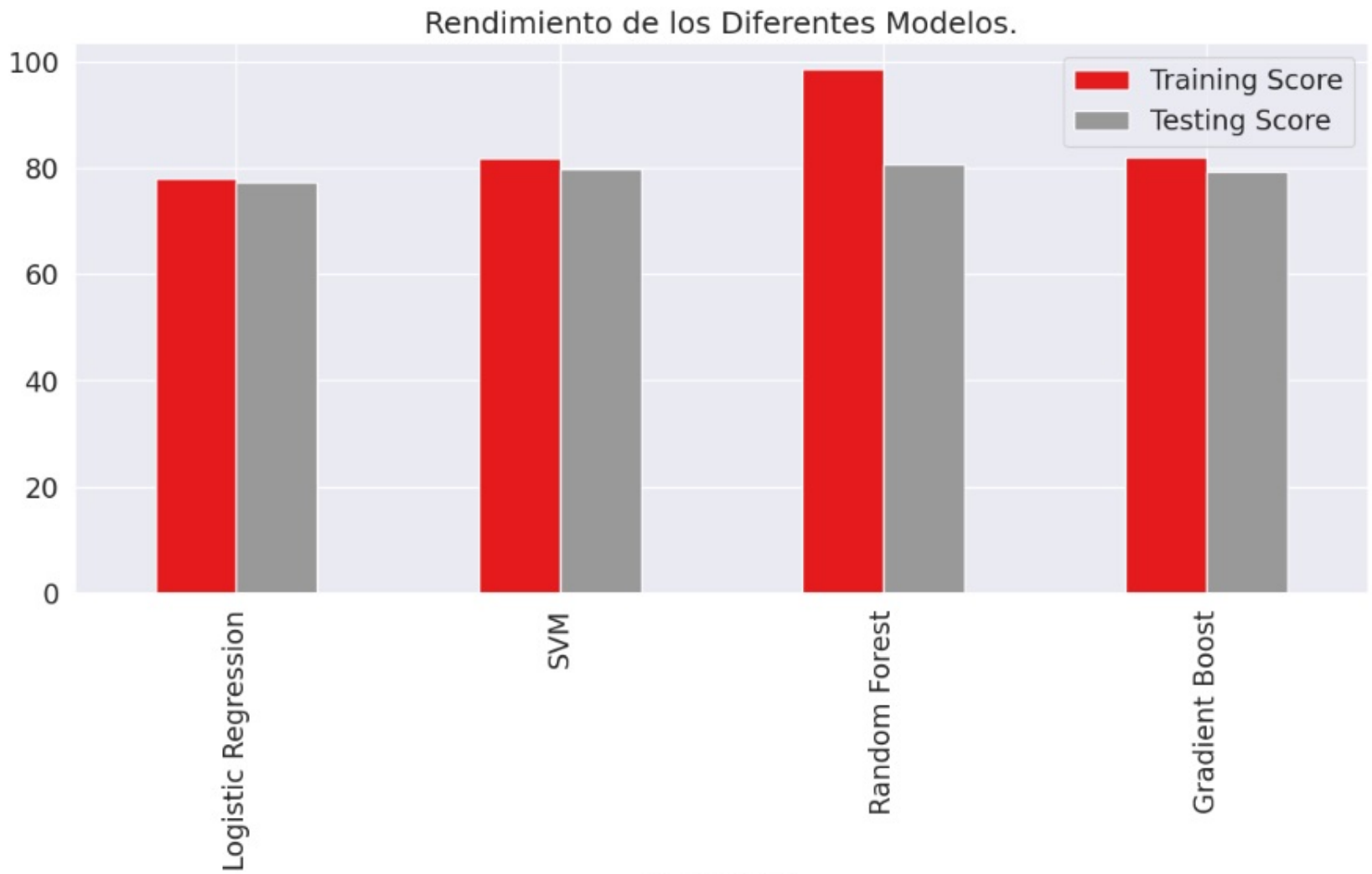
Modelo de Gradient Boosting:

- Accuracy_Score del conjunto de Entrenamiento es: 82.05
- Accuracy_Score del conjunto de Testeo es: 79.29
- Precision Score es: 0.77
- Recall Score es: 0.83
- F1 Score es: 0.80

Figura: Matriz de Confusion

Tabla de Contenidos

Metricas y Analisis de Resultados



Metricas y Analisis de Resultados

Modelo	Training Score	Testing Score
Regresion Logistica	77.868852	77.170788
SVM	81.823411	79.700978
Random Forest	98.518838	80.563542
Gradient Boost	82.053494	79.298447

Cuadro: Tabla de Resultados

Tabla de Contenidos

Conclusiones

Conclusiones Finales:

- Pocas funciones directamente trabajables
- Construimos nuevas funciones relevantes (Evitando perdida de informacion)
- Utilizamos tecnicas de Machine Learning para realizar predicciones
- Comparamos las predicciones y seleccionamos los mejores rendimientos:
  1. Random Forest
  2. Support-Vector Machine
  3. Gradient Boost
- Finalmente, se usa 'Stacking Model' para predecir los datos de prueba.