Spaceship Titanic: Proyecto de Aprendizaje Automático

Alejandro Salas, Nicolás Rodriguez, Joaquín Ramirez, Vicente Moreno.

Diciembre 2024

Introducción

Es el año 2912, donde la nave interestelar “Spaceship Titanic” llevaba gente desde el sistema solar terrestre hasta tres exoplanetas recientemente descubiertos, pero mientras se acercaba a Alpha Centauri chocó con una anomalía espacio temporal.

Introducción

El problema que se presenta es predecir si un pasajero será transportado a una dimensión paralela en el impacto basándose en la información obtenida después del accidente.

Datasets

Tamaño: 8693 entradas, 14 columnas.

El conjunto de entrenamiento posee las siguientes columnas:

- PassengerId
- HomePlanet
- CryoSleep
- Cabin
- Destination
- Age
- VIP
- RoomService
- FoodCourt
- ShoppingMall
- Spa
- VRDeck
- Name
- Transported

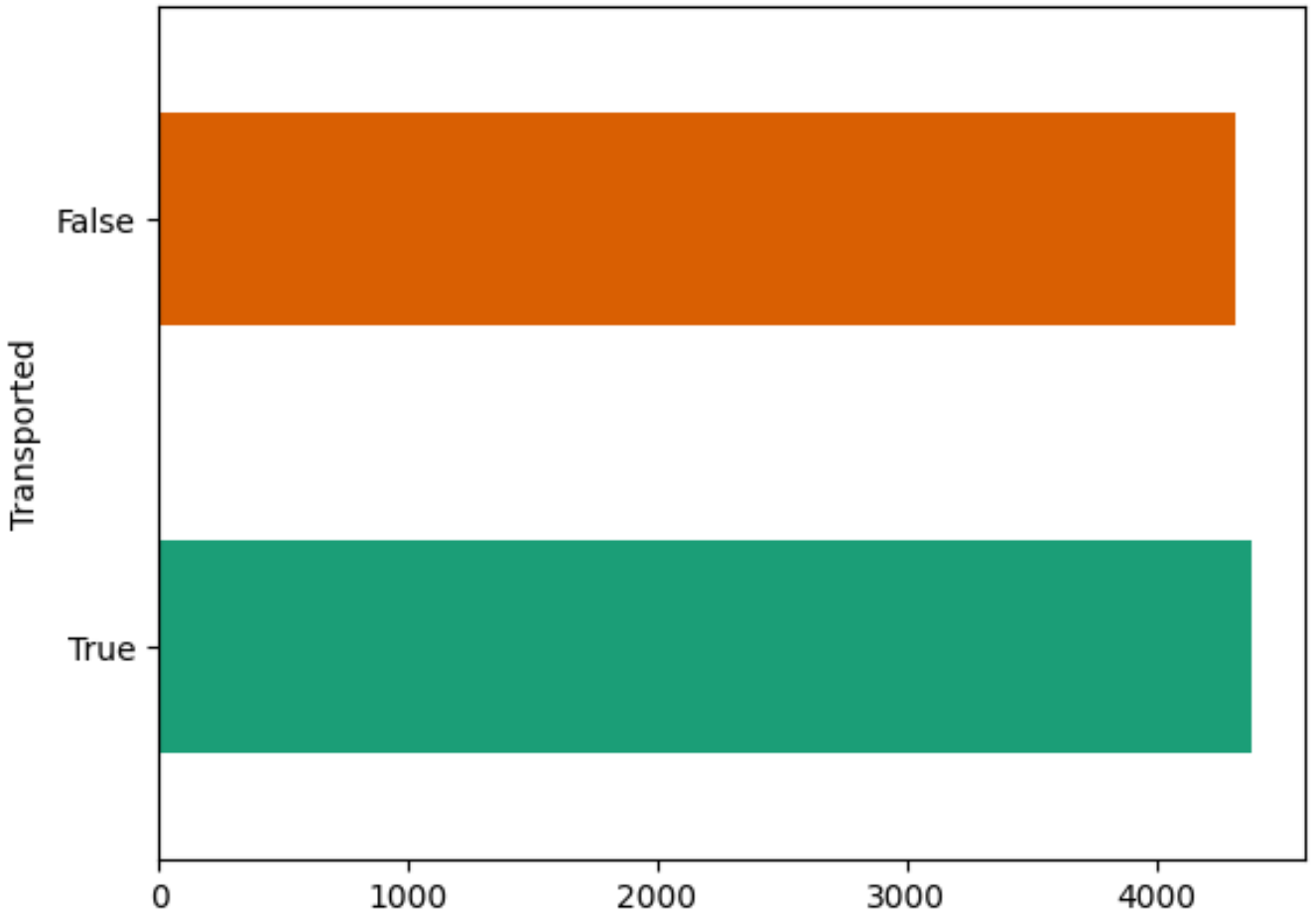
Datasets

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0

FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
0.0	0.0	0.0	0.0	Maham Ofracculy	False
9.0	25.0	549.0	44.0	Juanna Vines	True
3576.0	0.0	6715.0	49.0	Altark Susent	False

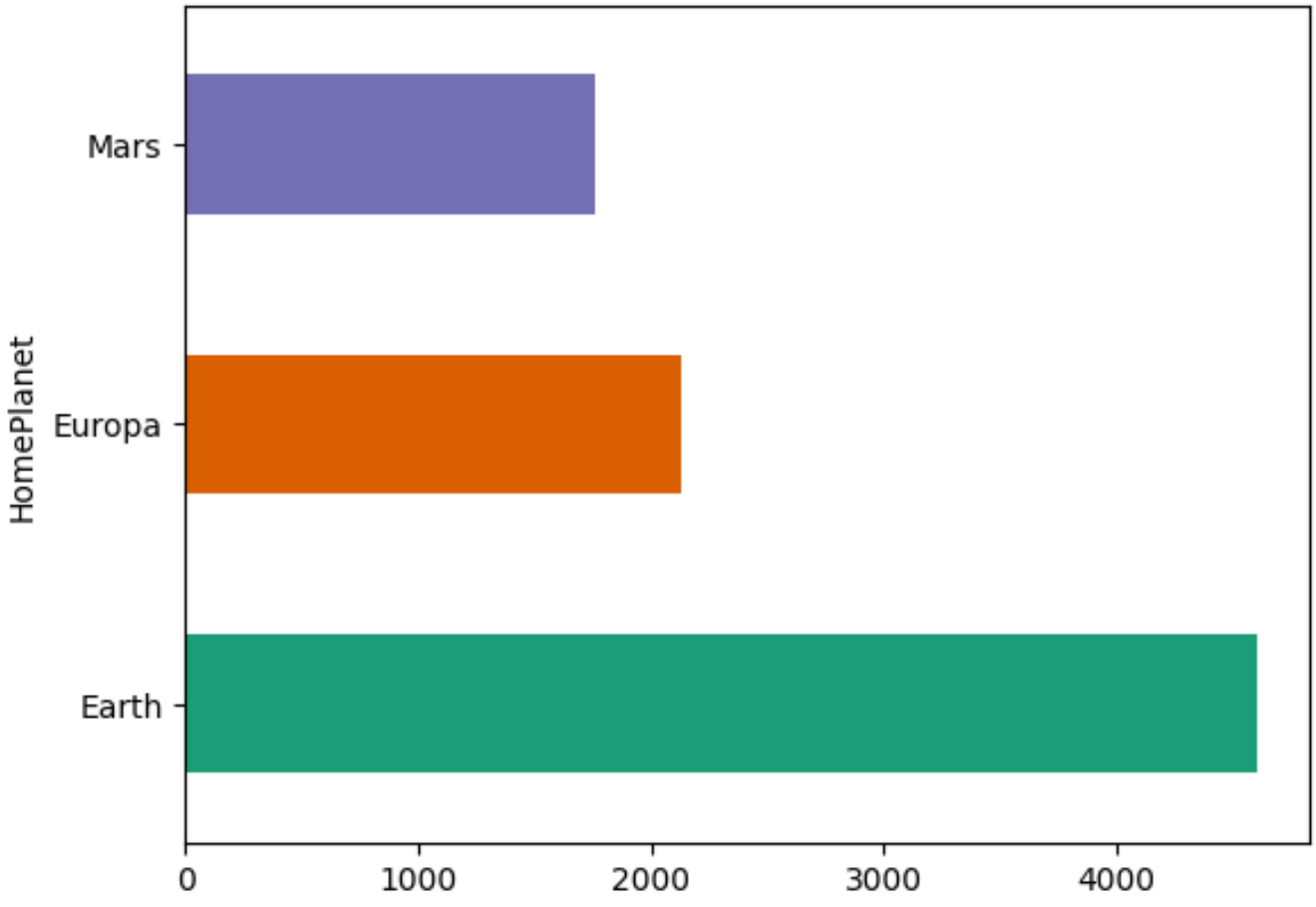
Cabecera de los datos (3 entradas)

Distribución de los Datos



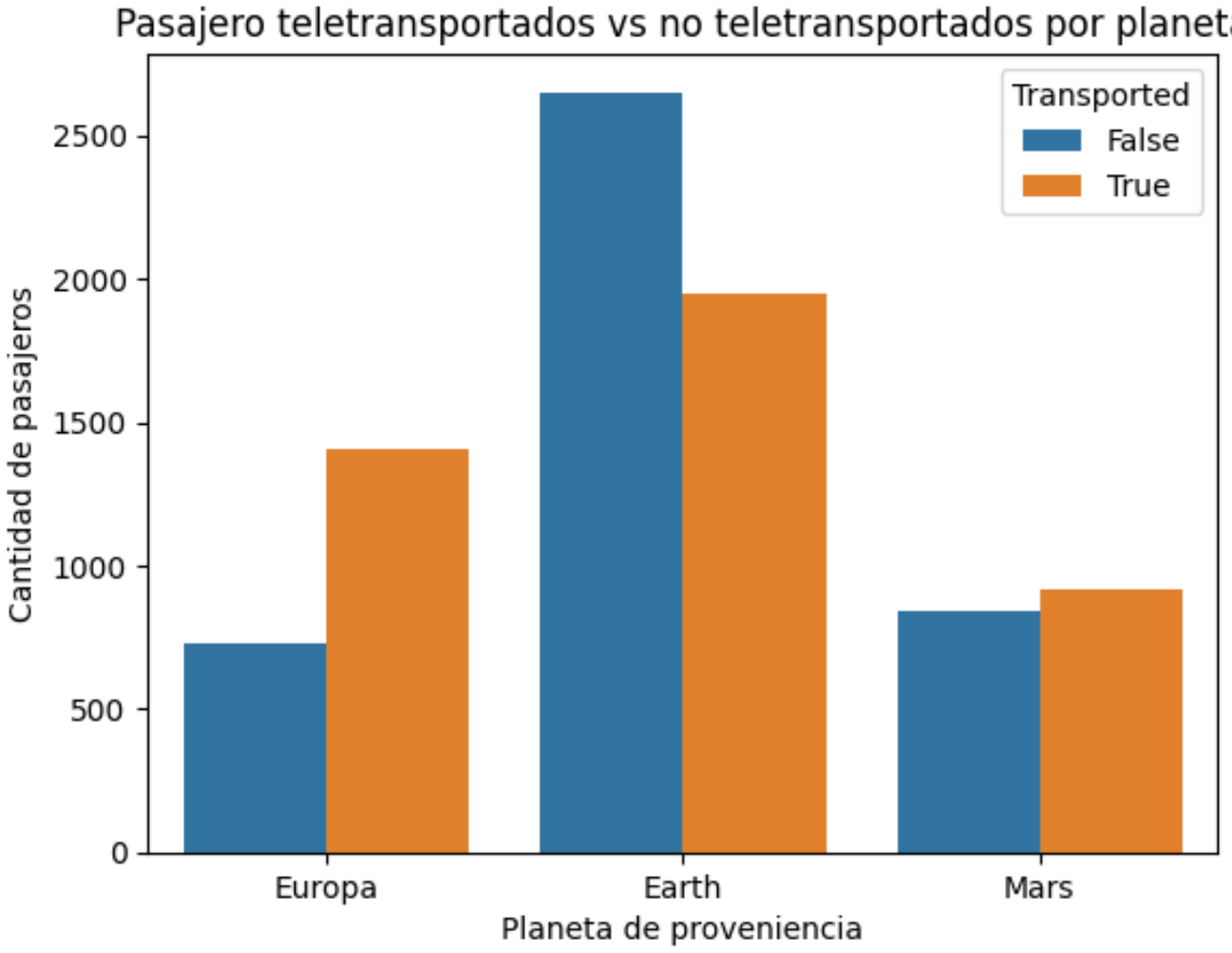
Transportados vs no transportados

Distribución de los Datos



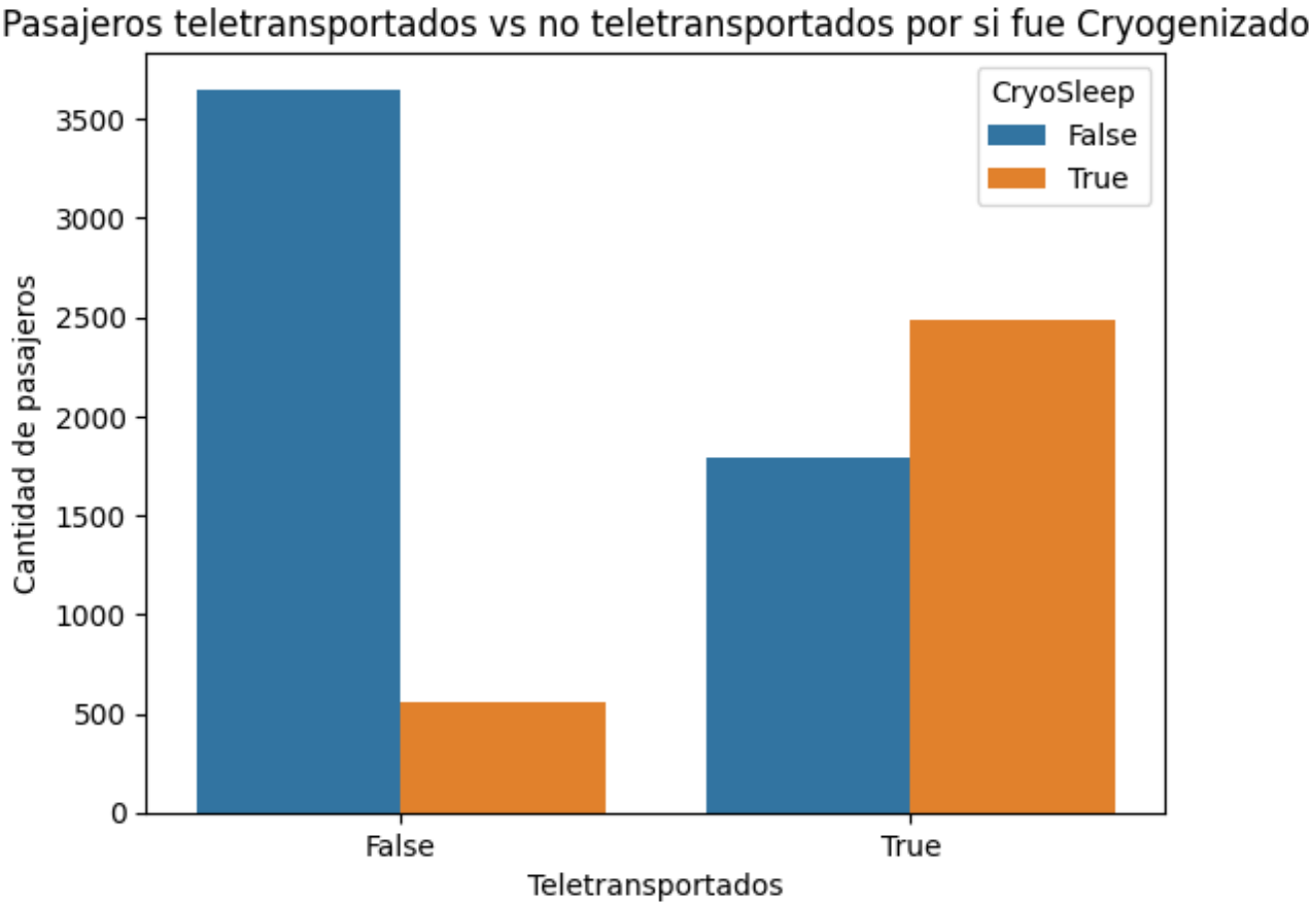
Personas por planeta

Distribución de los Datos



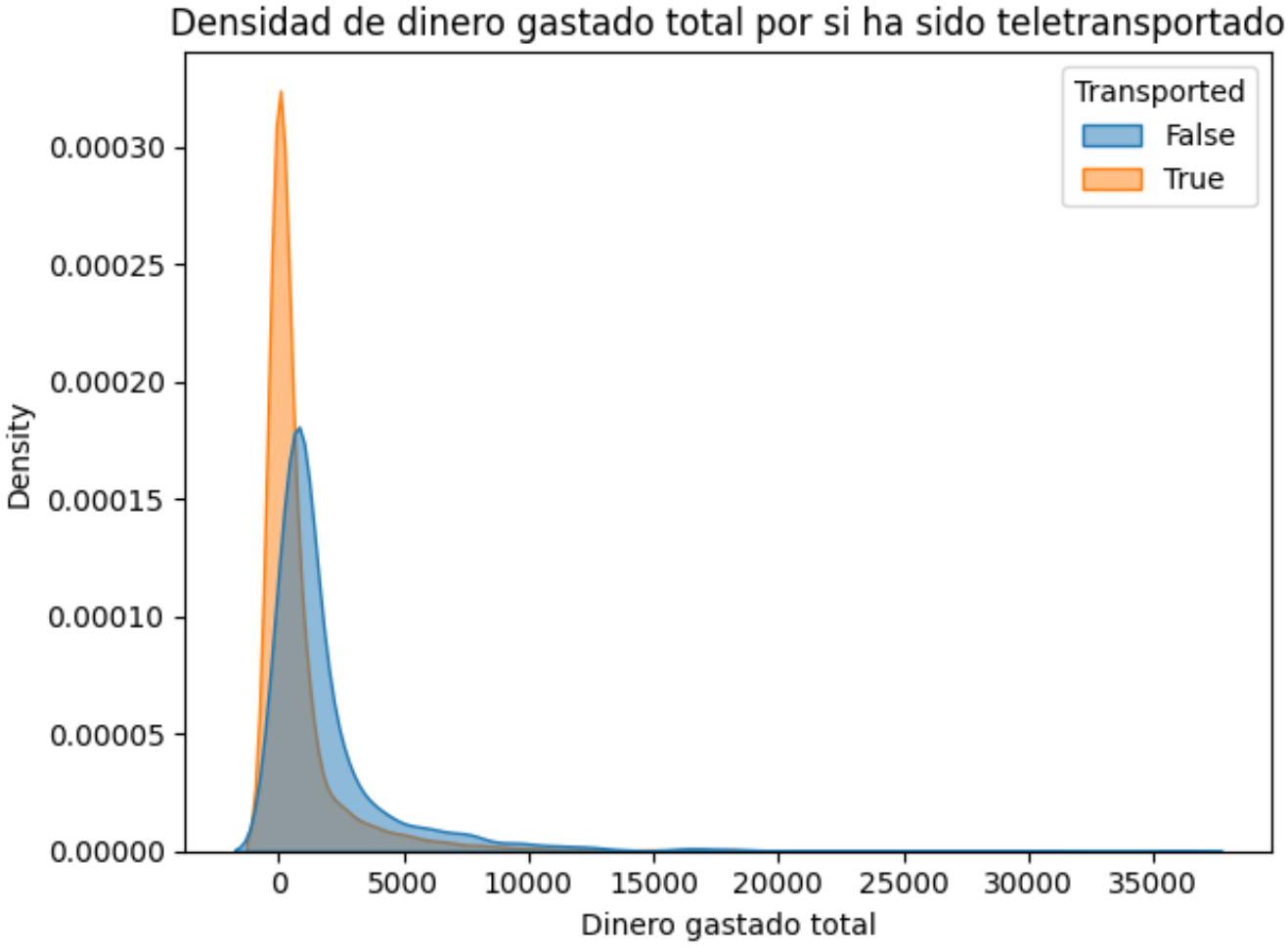
Personas transportadas por planeta de proveniencia

Distribución de los Datos



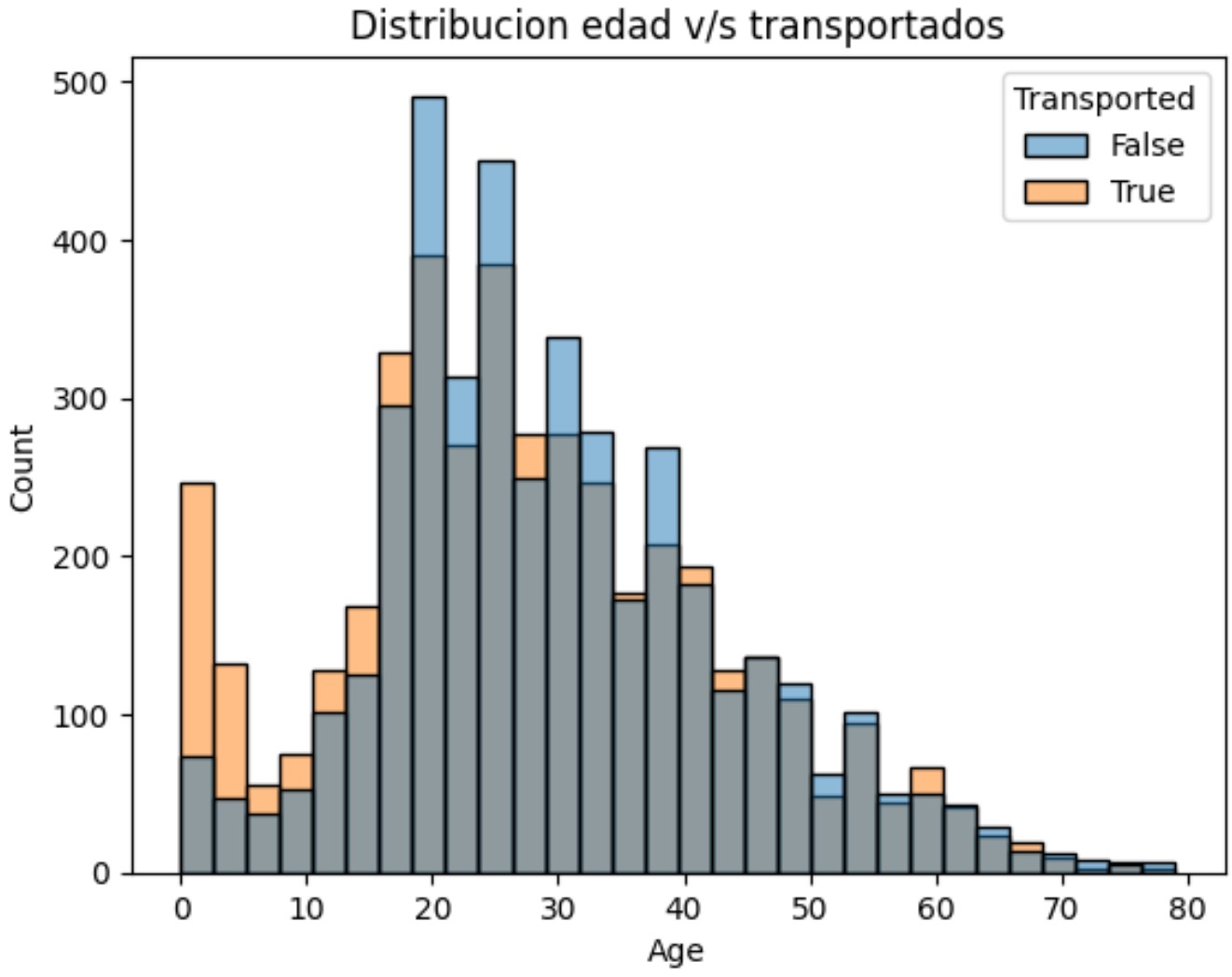
Gente transportada según su estado de Criogenización

Distribución de los Datos



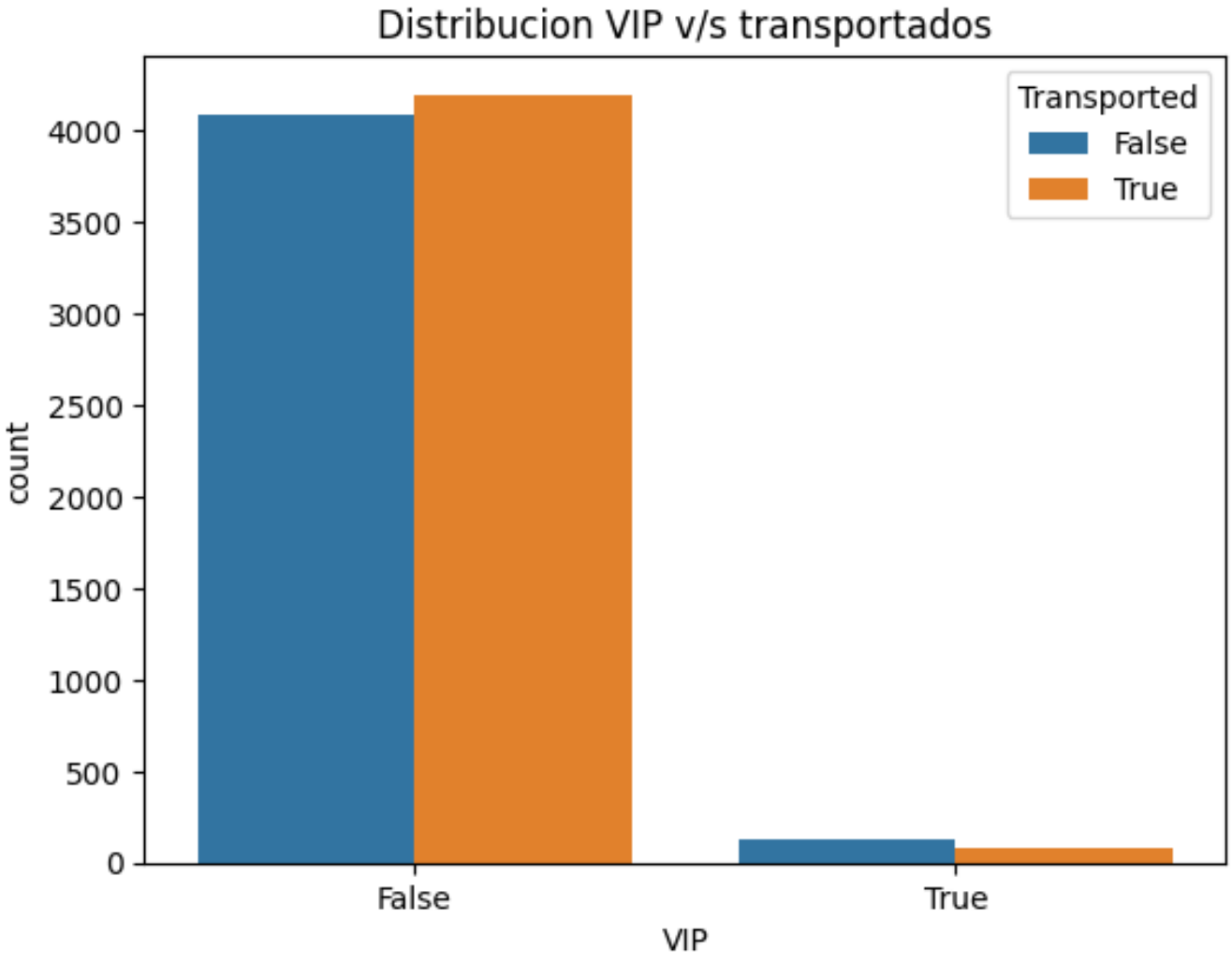
Densidad del dinero total gastado v/s si ha sido transportado o no

Distribución de los Datos



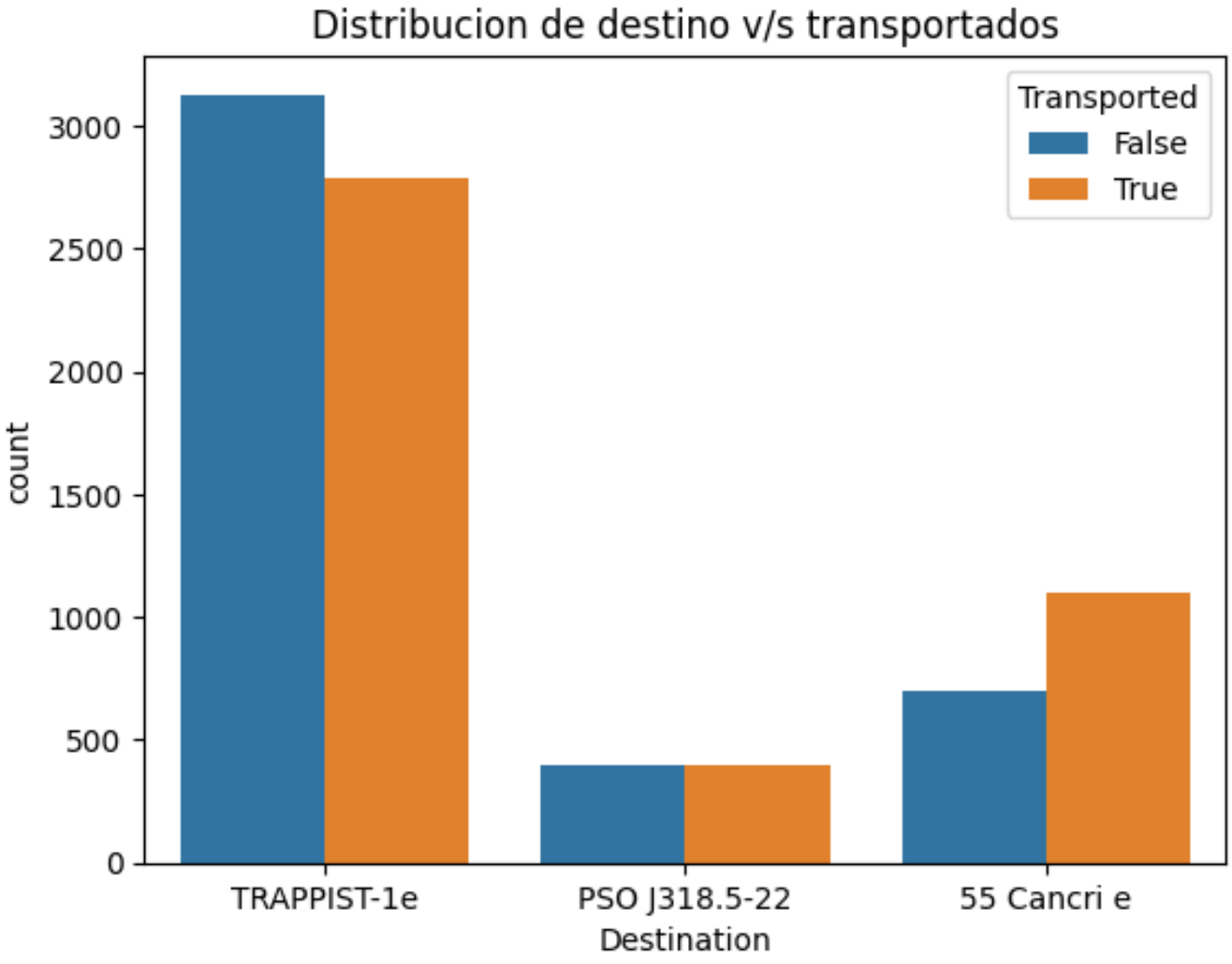
Distribución de edades versus si ha sido transportado o no

Distribución de los Datos



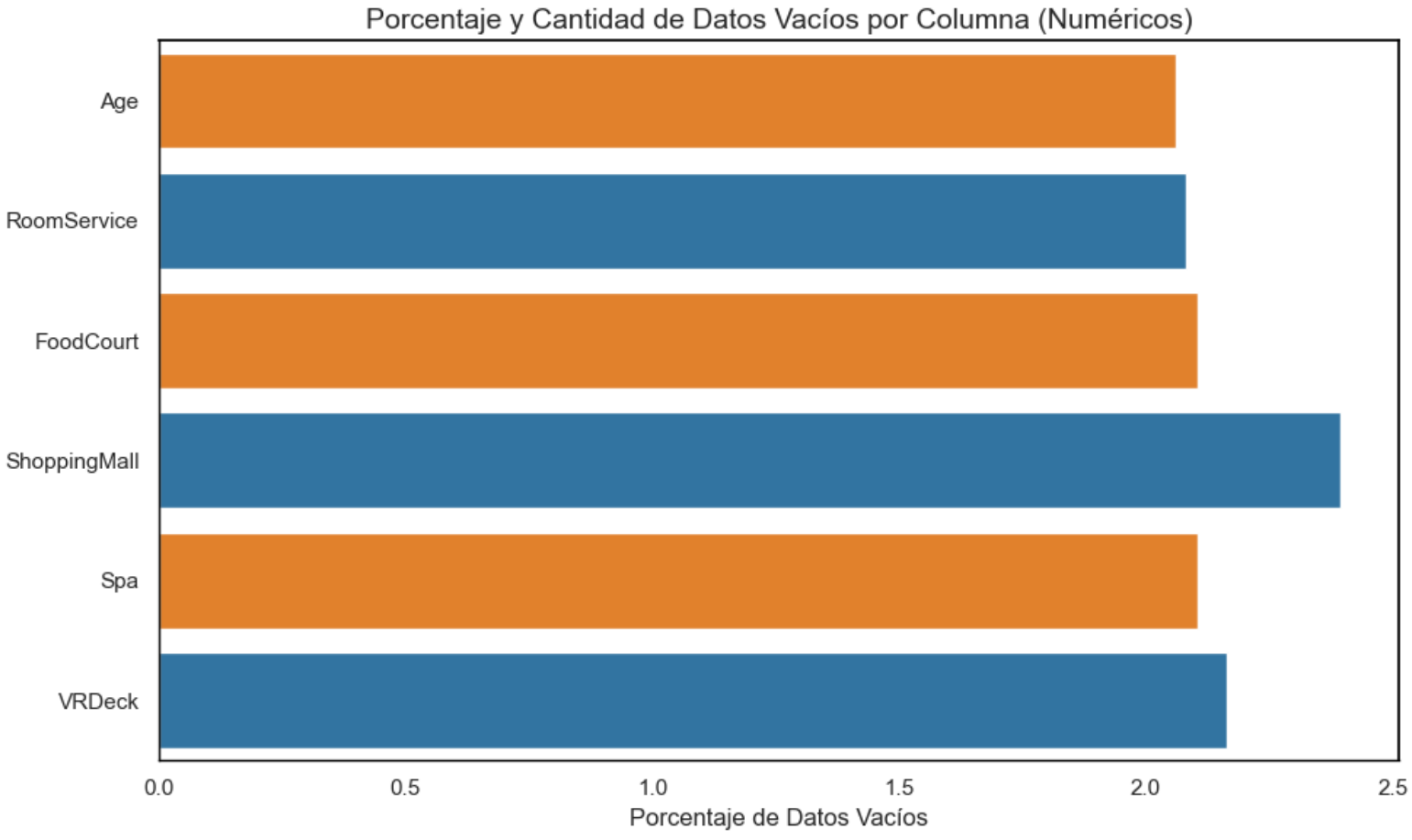
Distribución de si es VIP versus si ha sido transportado o no

Distribución de los Datos

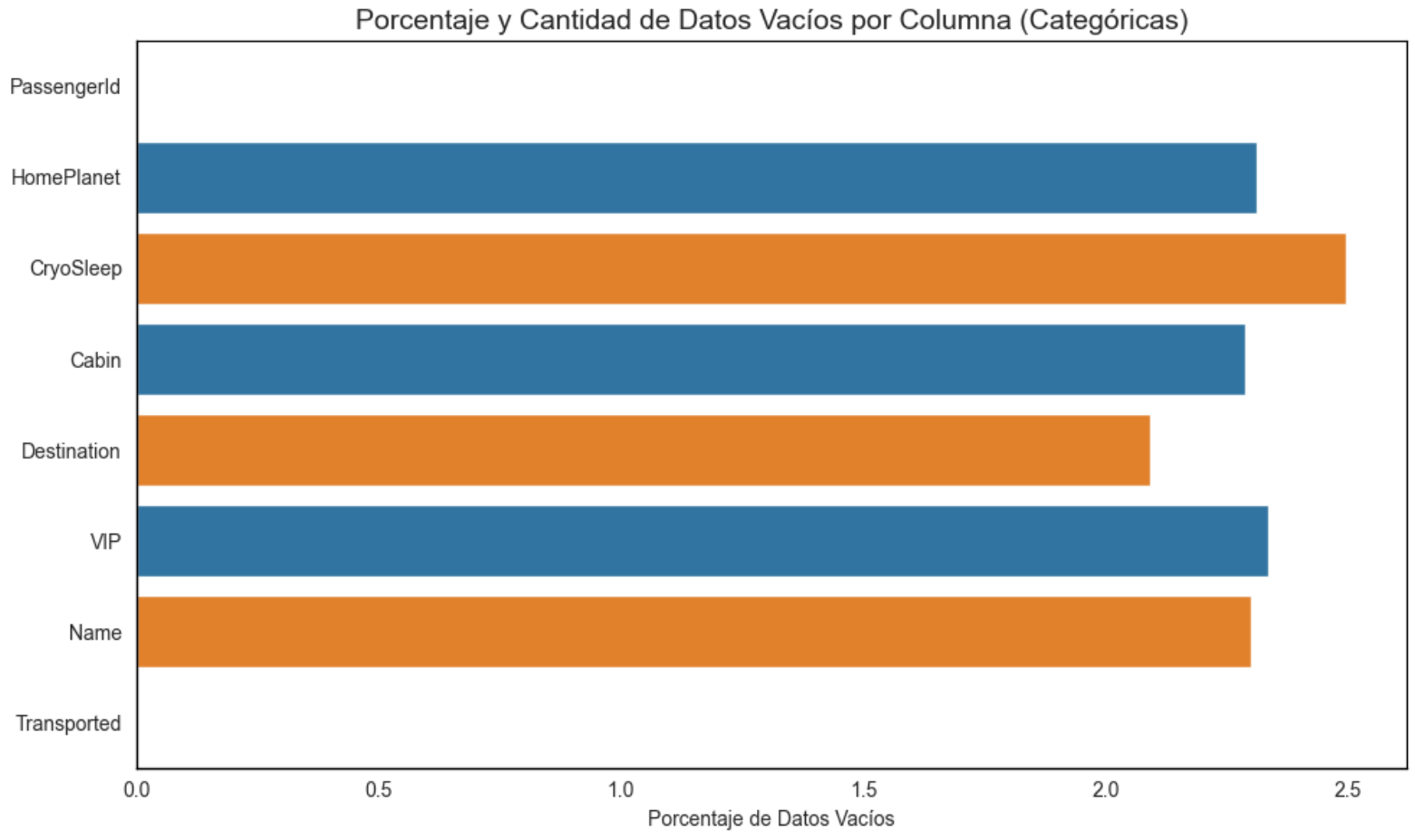


Distribución del destino contra si ha sido transportado o no

Valores Nulos



Valores Nulos



Valores Nulos

La columna que se refiera a la cabina tiene información inutilizable en su estado actual.
Por ello, se desglosó en 3 columnas. Una que se refiere a la cubierta; que hace referencia al numero de la cabina; que hace referencia si esta en babor o estribor.

Valores Nulos

- **Strings:** Se llenaron los datos nulos con 'other'.
- **Booleanos:** Se llenaron los datos nulos por falso.
- **Numericos:** Se llenaron los datos nulos por la mediana.

Datos Categóricos

Lo primero es dividir las columnas por sus categorías:

- **Booleanos:** Para los Booleanos, se cambió a su valores numéricos.
- **Strings:** A cada categoría de le asigna el promedio de su variable objetivo. (Targetencoding)

Datos Categóricos

	HomePlanet	Destination	deck	side
8140	0.654846	0.506329	0.688333	0.547341
4246	0.423784	0.469021	0.519455	0.454221
1526	0.423784	0.604732	0.435817	0.547341
7326	0.423784	0.506329	0.435817	0.547341
438	0.500000	0.469021	0.435817	0.547341

Preprocesamiento

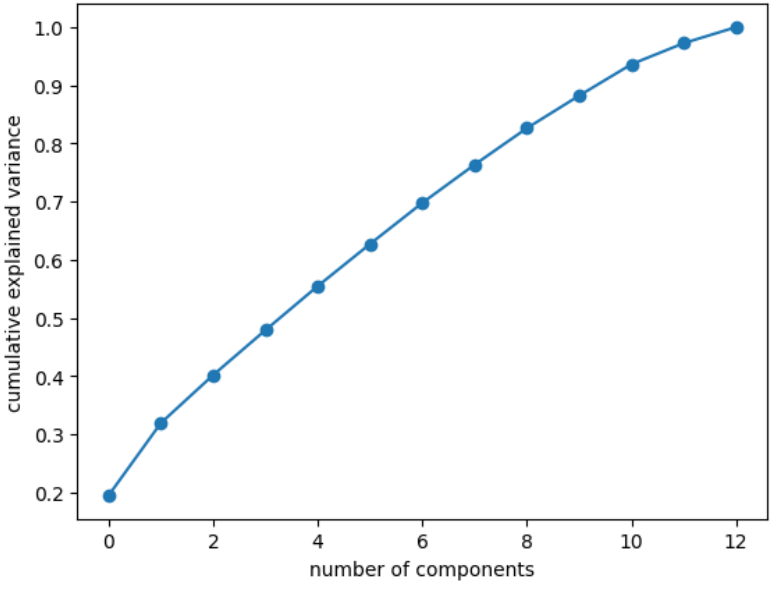
Se sospechó que habían columnas que no aportaban información al modelo. Estas eran

- VIP
- Age
- Destination
- FoodCourt, ShoppingMall, Spa, VRDeck.

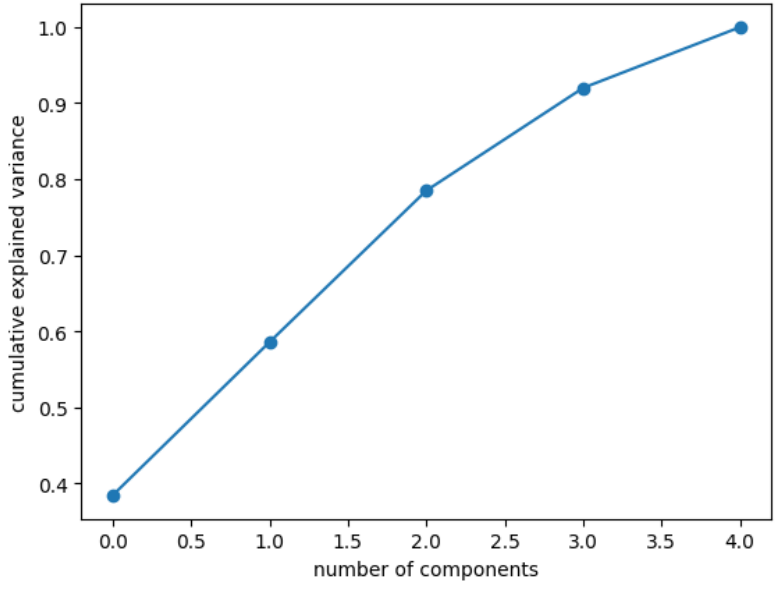
Se analizaron ambos casos al entrenar los modelos.

Preprocesamiento

PCA:



Componentes principales



Componentes principales filtradas

Modelos y grillas de parámetros

Los modelos a probar son:

- **LogisticRegression**
- **DecisionTreeClassifier**
- **RandomForestClassifier**
- **SVM**

Modelos y grillas de parámetros

Para escoger un modelo se utilizó Validación cruzada para entrenar y ver cual de los modelos tuvo mejor desempeño sobre el conjunto de datos sin las características antes mencionadas y todas, con los siguientes resultados:

Con todas las features:

	Modelo	Accuracy	F1 Score	Precision
0	Logistic Regression	0.793434	0.798120	0.782640
1	Decision Tree	0.739844	0.740691	0.740941
2	Random Forest	0.799625	0.793410	0.821804
3	SVM	0.799339	0.802537	0.792781

Sin algunas features:

	Modelo	Accuracy	F1 Score	Precision
0	Logistic Regression	0.725296	0.694047	0.786025
1	Decision Tree	0.660328	0.656218	0.666657
2	Random Forest	0.669115	0.668066	0.672357
3	SVM	0.725007	0.694762	0.783634

Resultados del entrenamiento con Validación Cruzada

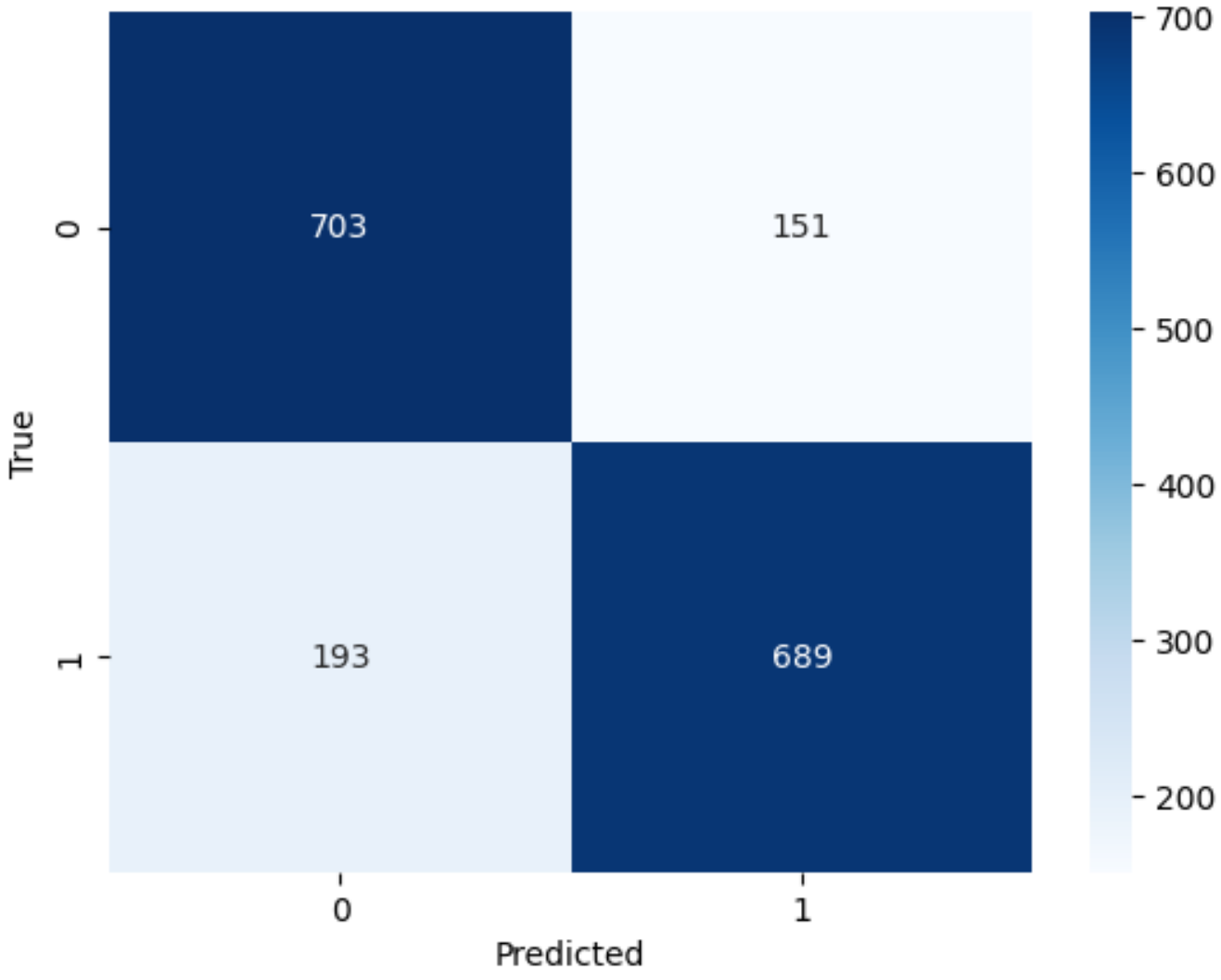
Modelos y grillas de parámetros

De esos datos se seleccionó el modelo "Decisión Tree", el cual se entrenó usando GridSearchCV para encontrar sus mejores hiperparámetros, obteniendo los siguientes puntajes sobre el conjunto de evaluación:

- Accuracy: 80%
- F1 score: 81%
- Precision: 79%

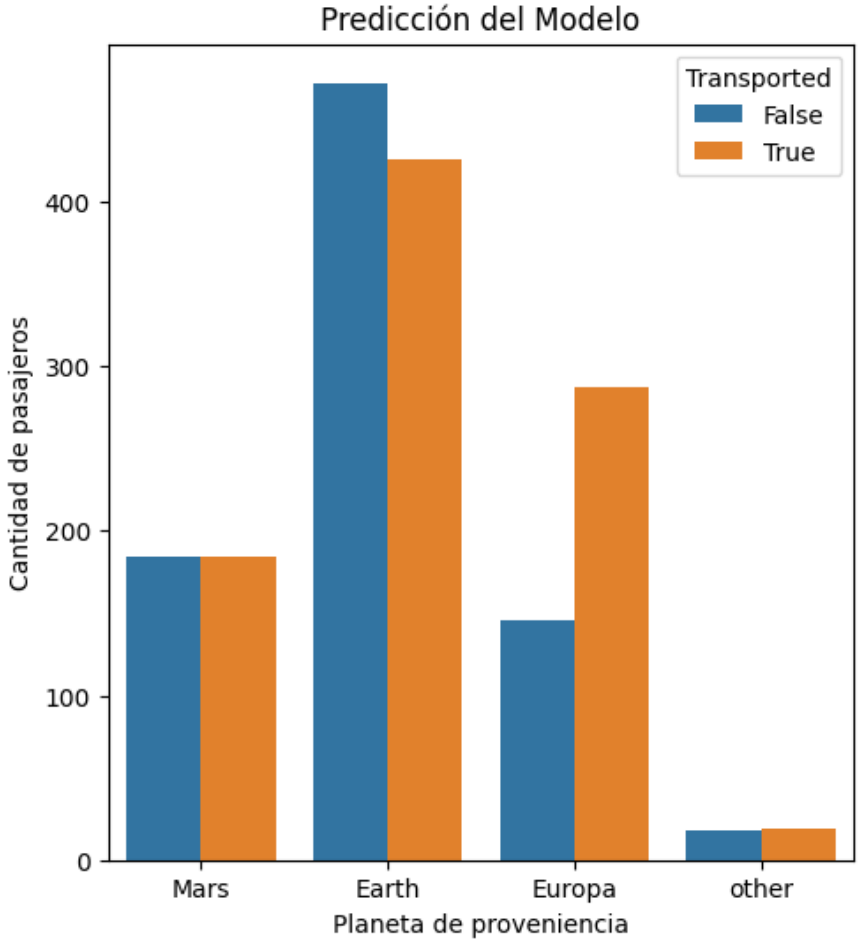
Y un puntaje de 79% de accuracy en el conjunto de prueba

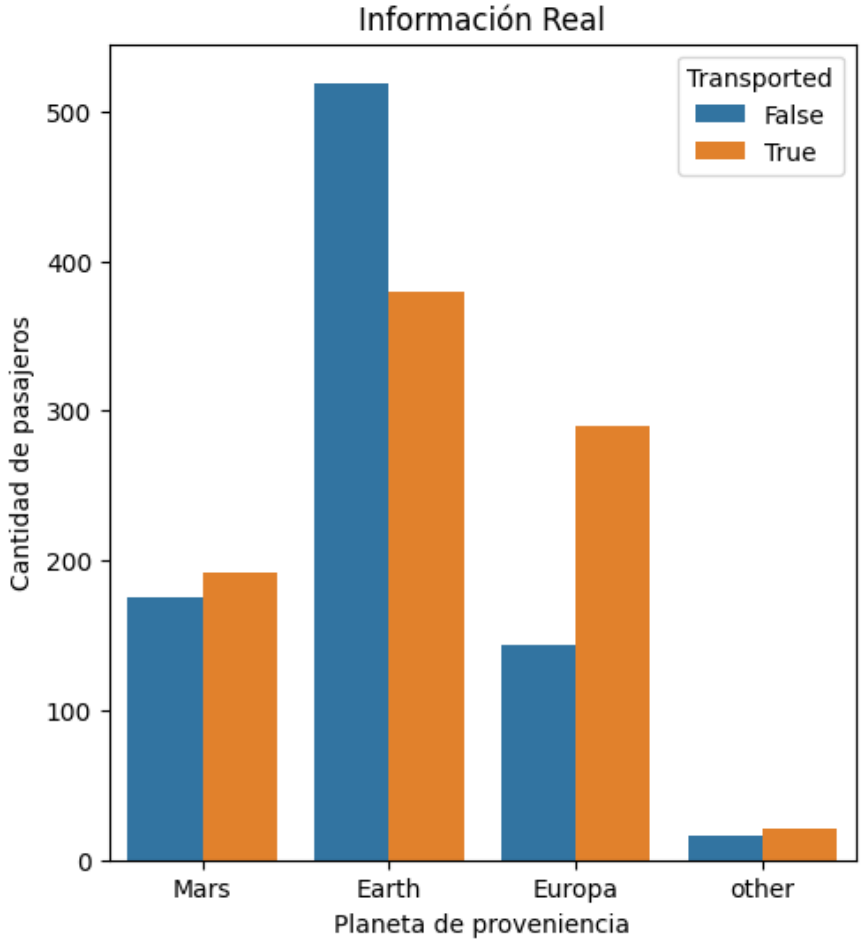
Visualizaciones del Modelo



Predicciones vs datos reales

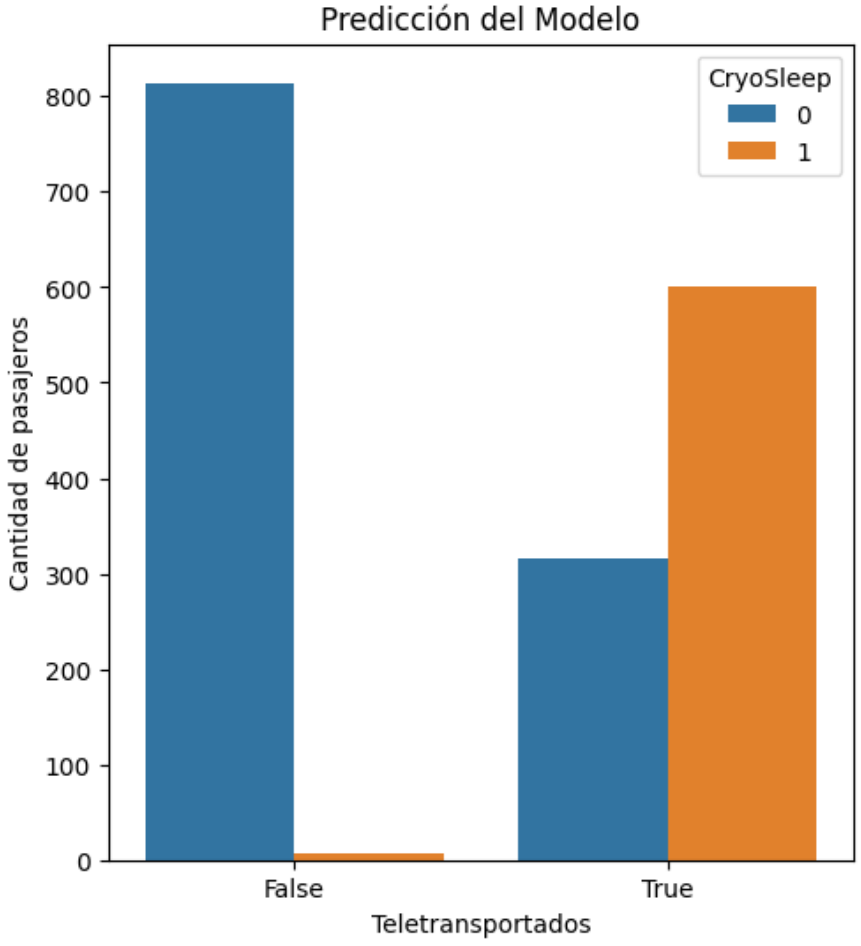
Visualizaciones del Modelo

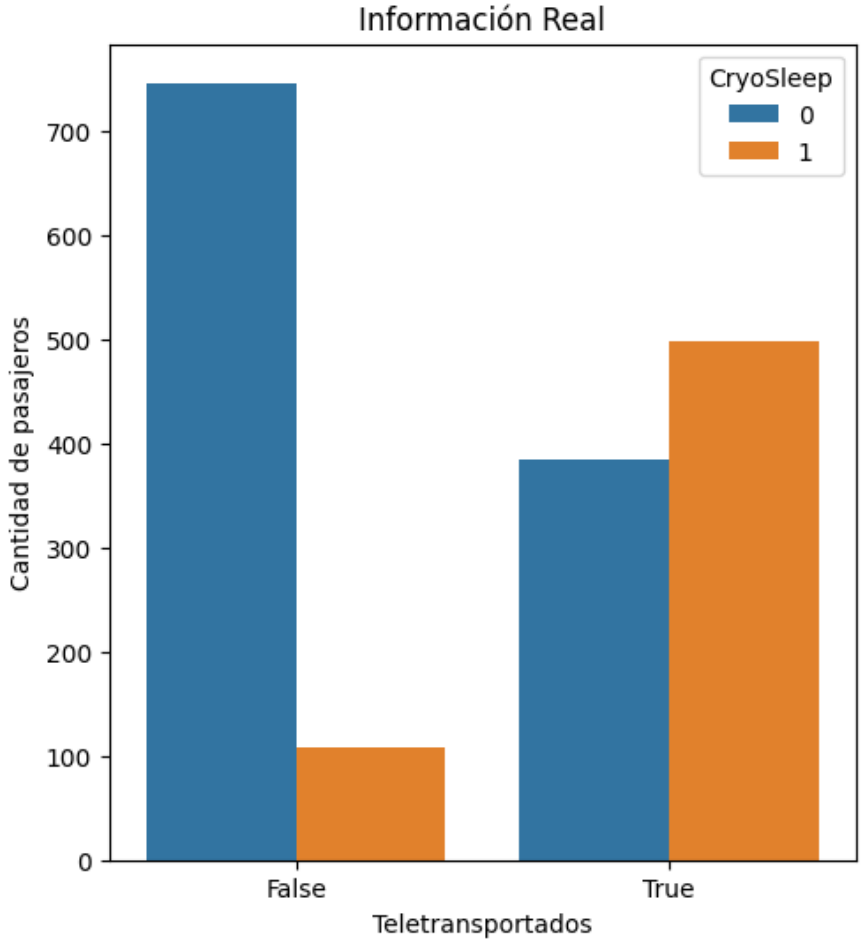




Predicciones vs datos reales por planeta

Visualizaciones del Modelo





Predicciones vs Datos Reales con criogenización

Conclusiones

- El análisis de los datos nos permite detectar ciertas regularidades/irregularidades. Por ejemplo, vivir o haber vivido en la Tierra disminuye la probabilidad de ser transportados.
- Dado que mucha gente criogenizada si fue transportada. El modelo generó un sesgo contra ellos.
- Las características que pensamos que no aportaban, si lo hacían. Por ello, siempre se debe realizar un analisis completo y no dejarse llevar por intuiciones.

Referencias

- Addison Howard, Ashley Chow, and Ryan Holbrook. Spaceship Titanic. 2022. Kaggle.