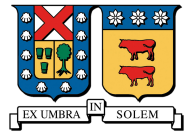


Spaceship Titanic

Claudio González Robles
Universidad Federíco Santa María, Santiago, Chile

November 30, 2023

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Estructura

1. Definición del problema
2. Estadística Descriptiva
3. Visualización descriptiva
4. Preprocesamiento
5. Selección de modelo
6. Visualizaciones del modelo
7. Análisis de resultados
8. Conclusiones

Contexto

La nave espacial Titanic fue un transatlántico de pasajeros interestelar lanzado hace un mes. Con casi 13.000 pasajeros a bordo, la nave emprendió su viaje inaugural transportando emigrantes de nuestro sistema solar a tres exoplanetas recientemente habitables que orbitan estrellas cercanas. Mientras rodeaba Alpha Centauri en ruta hacia su primer destino, el tórrido 55 Cancri E, la desprevenida nave espacial Titanic chocó con una anomalía del espacio-tiempo escondida dentro de una nube de polvo. Lamentablemente, tuvo un destino similar al de su homónimo de 1000 años antes. Aunque la nave permaneció intacta, ¡casi la mitad de los pasajeros fueron transportados a una dimensión alternativa!

Problema

Debido a los problemas del barco solo se logró recopilar dos tercios de la información de los pasajeros, por ello mediante herramientas de Machine Lerning se desea predecir que pasajeros fueron trasportados de la nave para el tercio restante.

Carga de Datos

- Los datos estaban almacenados en github en formato csv, estos fueron trabajados en Google Cobab .
- Importamos varias librerias necesarias, incluyendo las de cada modelo.
- Principalmente se utilizaron Data Frames y Listas de Python.

```
#librerias a utilizar
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt #graficar
import seaborn as sns #graficar
import re
import time #para medir tiempos
from sklearn.preprocessing import StandardScaler #normalizar
from sklearn.linear_model import LinearRegression #modelo
from sklearn.svm import SVR #modelo
from sklearn.ensemble import RandomForestRegressor #modelo
from sklearn.ensemble import GradientBoostingRegressor #modelo
from sklearn.preprocessing import MinMaxScaler
import warnings
warnings.filterwarnings("ignore")
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Estadistica Descriptiva

- Tratamiento de Datos vacios.
- Informacion.
- Descripción.

```
df_train.isnull().sum()
```

```
PassengerId       0
HomePlanet      201
CryoSleep       217
Cabin           199
Destination     182
Age             179
VIP             203
RoomService     181
FoodCourt       183
ShoppingMall    208
Spa             183
VRDeck          188
Name            200
Transported       0
dtype: int64
```

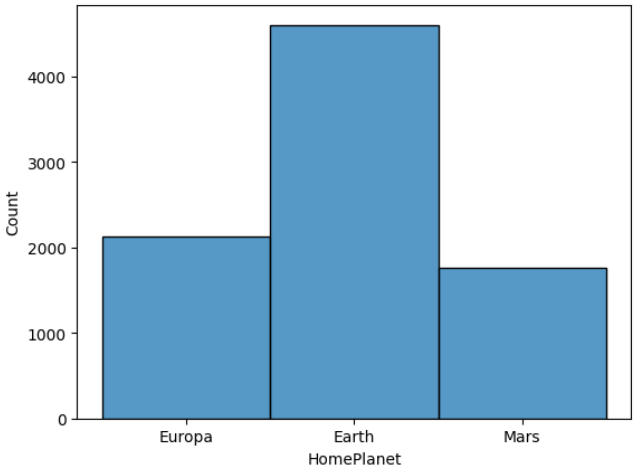
```
[8] df_train.info()
```

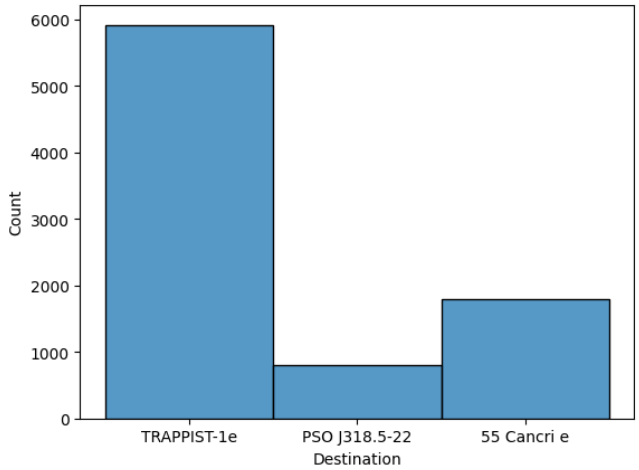
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8693 entries, 0 to 8692
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   PassengerId   8693 non-null   object
 1   HomePlanet    8492 non-null   object
 2   CryoSleep     8476 non-null   object
 3   Cabin         8494 non-null   object
 4   Destination   8511 non-null   object
 5   Age           8514 non-null   float64
 6   VIP           8490 non-null   object
 7   RoomService   8512 non-null   float64
 8   FoodCourt     8510 non-null   float64
 9   ShoppingMall  8485 non-null   float64
 10  Spa           8510 non-null   float64
 11  VRDeck        8505 non-null   float64
 12  Name          8493 non-null   object
 13  Transported   8693 non-null   bool
dtypes: bool(1), float64(6), object(7)
memory usage: 891.5+ KB
```

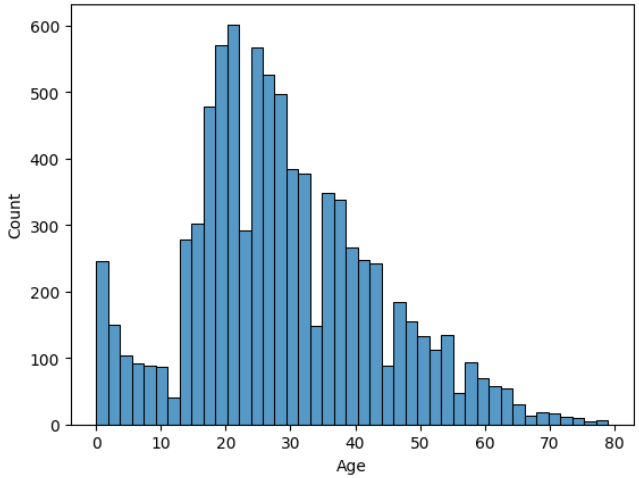
```
df_train.describe()
```

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	8514.000000	8512.000000	8510.000000	8485.000000	8510.000000	8505.000000
mean	28.827930	224.687617	458.077203	173.729169	311.138778	304.854791
std	14.489021	666.717663	1611.489240	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000	27.000000	59.000000	46.000000
max	79.000000	14327.000000	29813.000000	23492.000000	22408.000000	24133.000000

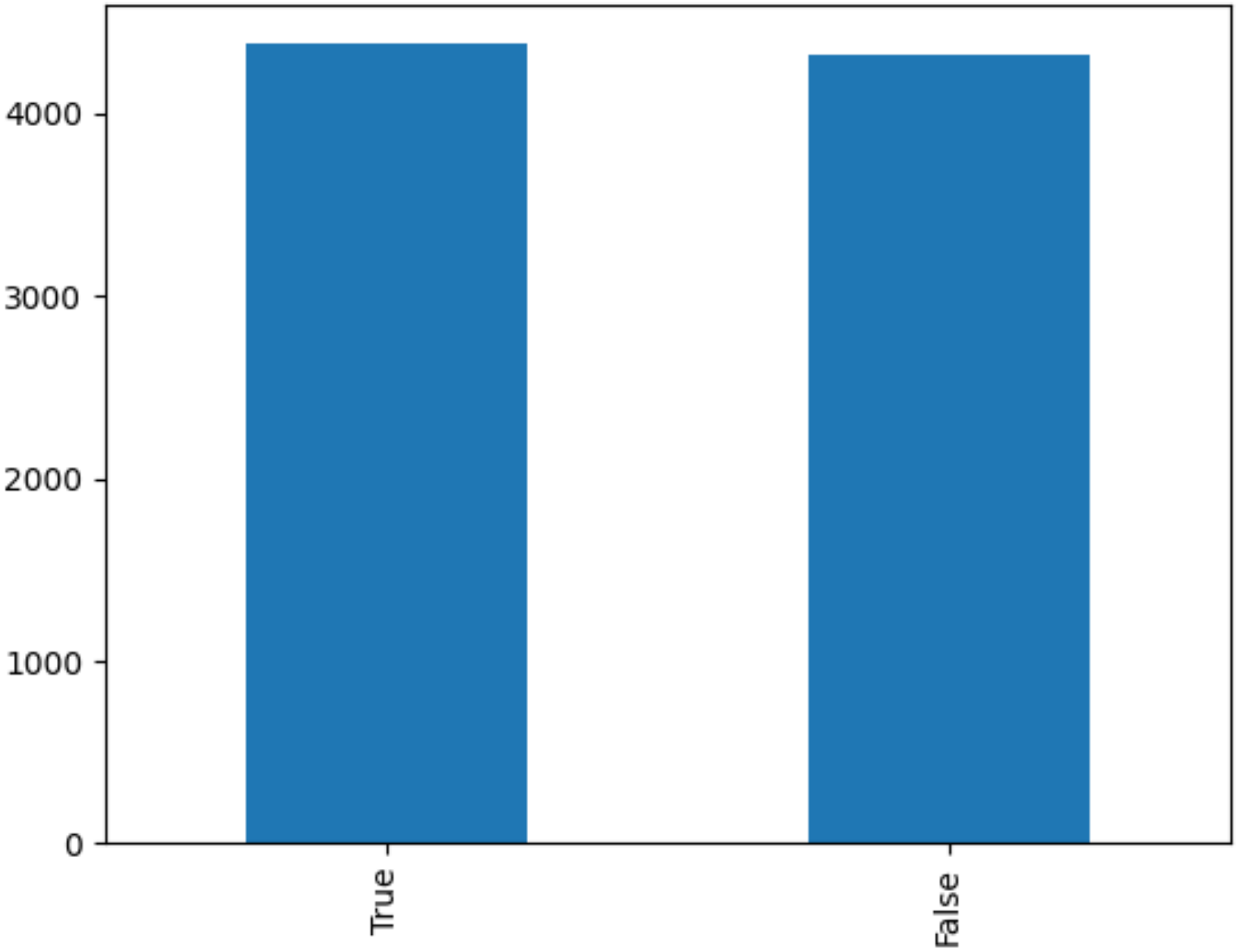
Visualización descriptiva







Transportados



Preprocesamiento

- Tratamiento datos vacios.
- Separación de Datos.
- Trasformación de variables Categóricas a Numéricas.

Vacios

Primero se solucionará para cada columna el problema de los valores vacios segun el siguiente criterio:

- HomePlanet- "N"
- CryoSleep- "False"
- Cabin- "-/-1/-"
- Destination- "N"
- Age-"-1"
- VIP- "False"
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck- "-1"
- Name- "N"

Al rellenar los datos de esta forma no perdemos el hecho de que faltan datos y a la vez no entorpecen el analisis de problema.

Sepación de Datos

- Id Pasajero
  xxxx_yy –> Grupo = xxxx
- Asiento
  xx/yy/zz –> deck = xx ; num = yy ; side = zz

Transformación

Metodo de Clasificacion:

HomePlanet:
>Earth = 0
>Europa = 1
>Mars = 2
>N = 3

CryoSleep:
>0 = 0
>1 = 1

Destination:
>55 Cancri e = 0
>N = 1
>PSO J318.5-22 = 2
>TRAPPIST-1e = 3

VIP:
>0 = 0
>1 = 1

Transported:
>0 = 0
>1 = 1

Deck:
>- = 0
>A = 1
>B = 2
>C = 3
>D = 4
>E = 5
>F = 6
>G = 7
>T = 8

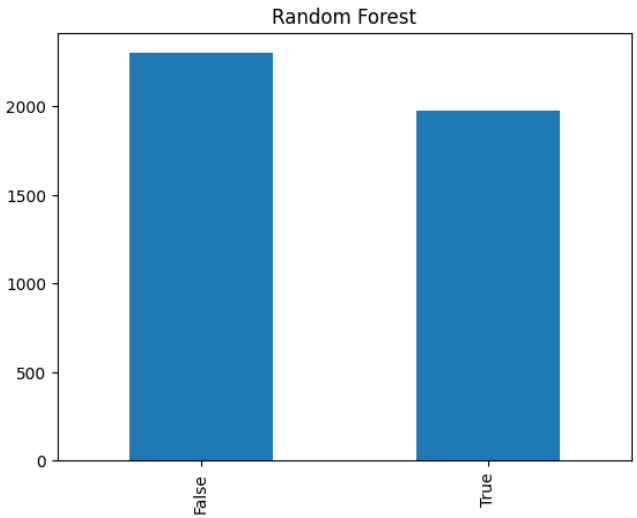
Side:
>- = 0
>P = 1
>S = 2

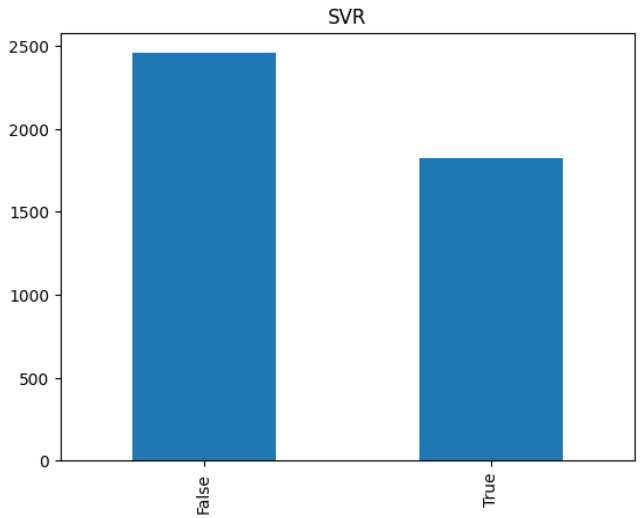
Selección de modelo

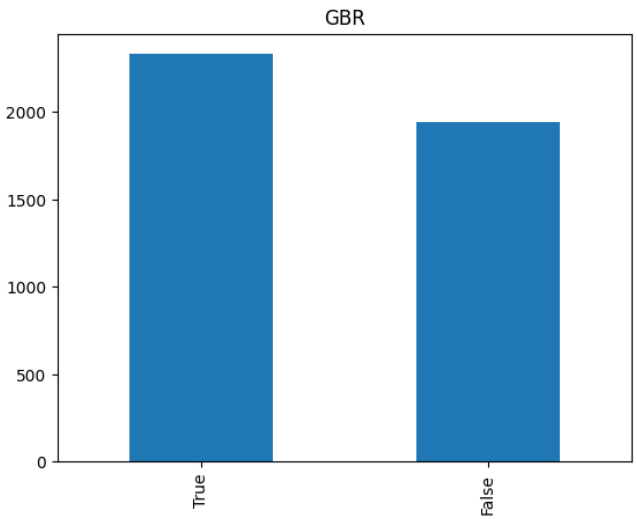
Total de Datos = 4277

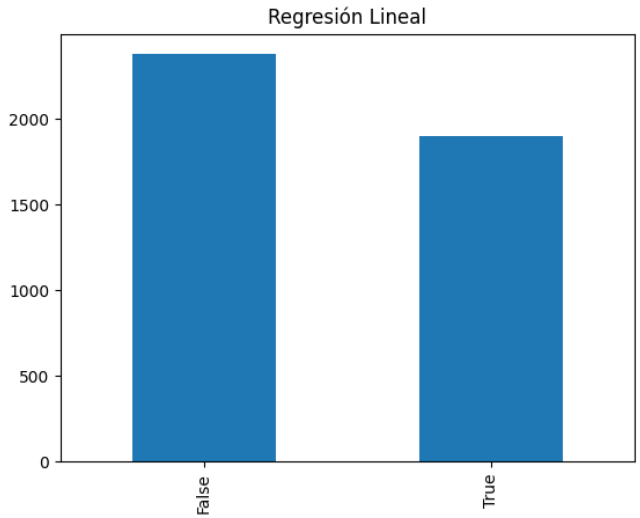
- Random Forest (rf)
  False = 2301
  True = 1976
- Support Vector Machine (svr)
  False = 2455
  True = 1822
- Gradient Boosting (GBR)
  True = 2331
  False = 1946
- Regresión Lineal (linear)
  False = 2375
  True = 1902

Visualizaciones del modelo



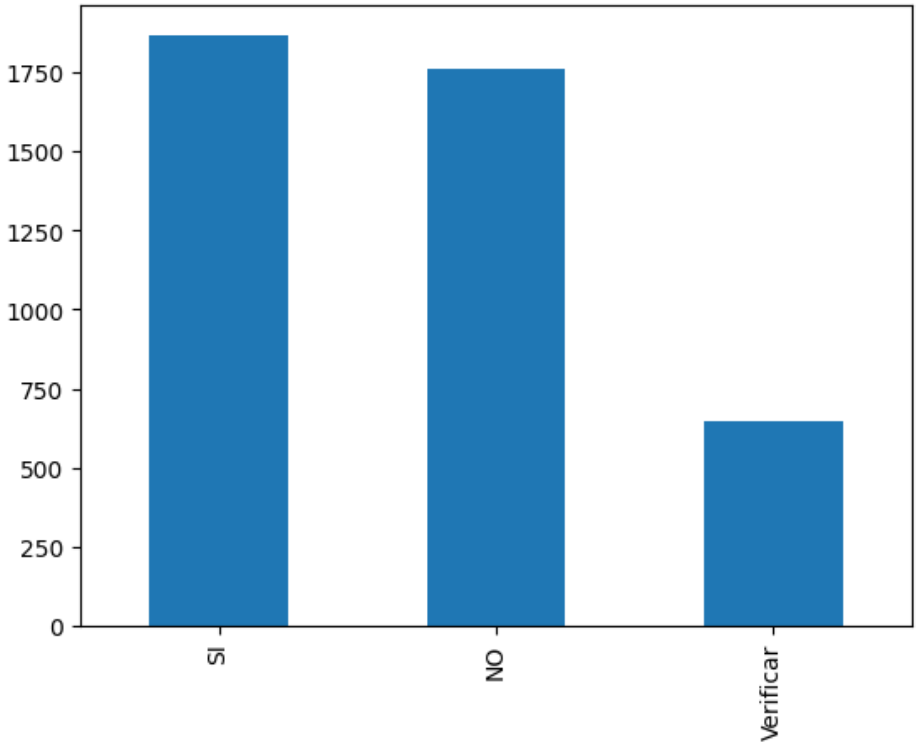






Análisis de resultados

¿Cuanta certeza hay de que fueron tranportados a otra dimensión?



SI = 1868
NO = 1764
Verificar = 645

Conclusiones

Basandonos en el análisis de los cuatro modelos en conjunto, y utilizando como criterio que a lo menos 3 modelos concuerden con su decisión, podemos reducir considerablemente la tarea de busqueda de personas transportadas.
Adicionalmente al anállisis de los datos, se propone como solución extra a la problematica general dos posibles puntos de vistas relacionados directamente al caso:

- A las personas que queden por verificar buscar su numero de asiento para comprobar su estado.
- Relacionar directamente si las teletransportaciones están directamente relacionadas con la ubicación espacial de las personal al momento del incidente, para ellos estudiar itinerarios de actividades del barco.