Spaceship Titanic
Proyecto de Machine Learning

ALBERTO LÁZARO TRONCOSO

INDICE

El dataset se distribuye en 8693 entradas con 14 variables que son las que siguen a continuación:

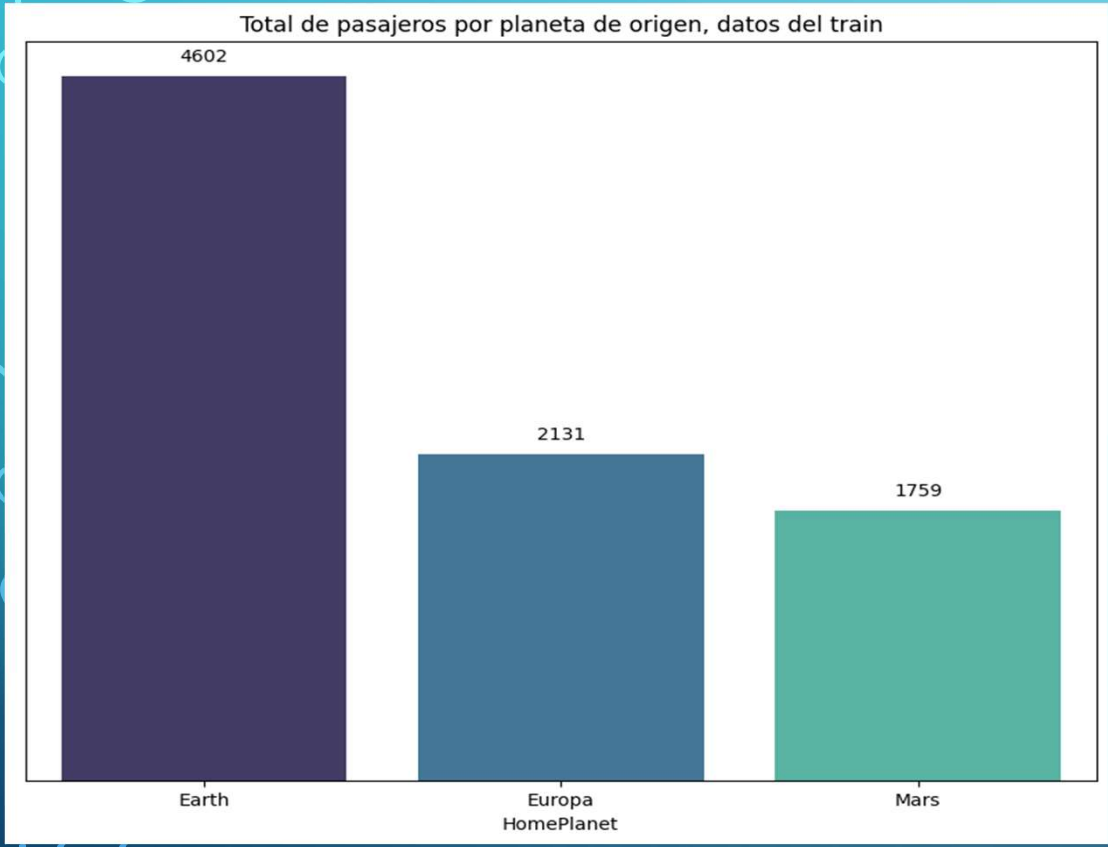
- **PassengerId**: una identificación única para cada pasajero. Cada Id toma la forma XXXX_PP donde XXXX indica un grupo con el que viaja el pasajero y PP es su número dentro del grupo. Las personas en un grupo a menudo son miembros de la familia, pero no siempre.
- **HomePlanet**: el planeta del que partió el pasajero, normalmente su planeta de residencia permanente.
- **CryoSleep**: indica si el pasajero eligió ponerse en animación suspendida durante la duración del viaje. Los pasajeros en criosueño están confinados en sus cabinas.
- **Cabin**: el número de cabina donde se hospeda el pasajero. Toma la forma cubierta/número/lado, donde lado puede ser P para babor o S para estribor.
- **Destination**: el planeta dónde desembarcará el pasajero.
- **Age**: la edad del pasajero.
- **VIP**: si el pasajero ha pagado por un servicio VIP especial durante el viaje.
- **RoomService, FoodCourt, ShoppingMall, Spa, VRDeck** indican el montante que el pasajero ha facturado en cada uno de los muchos servicios de lujo del viaje.
- **Name**: el nombre y apellido del pasajero.
- **Transported**: Es mi Target. Si el pasajero fue transportado a otra dimensión.

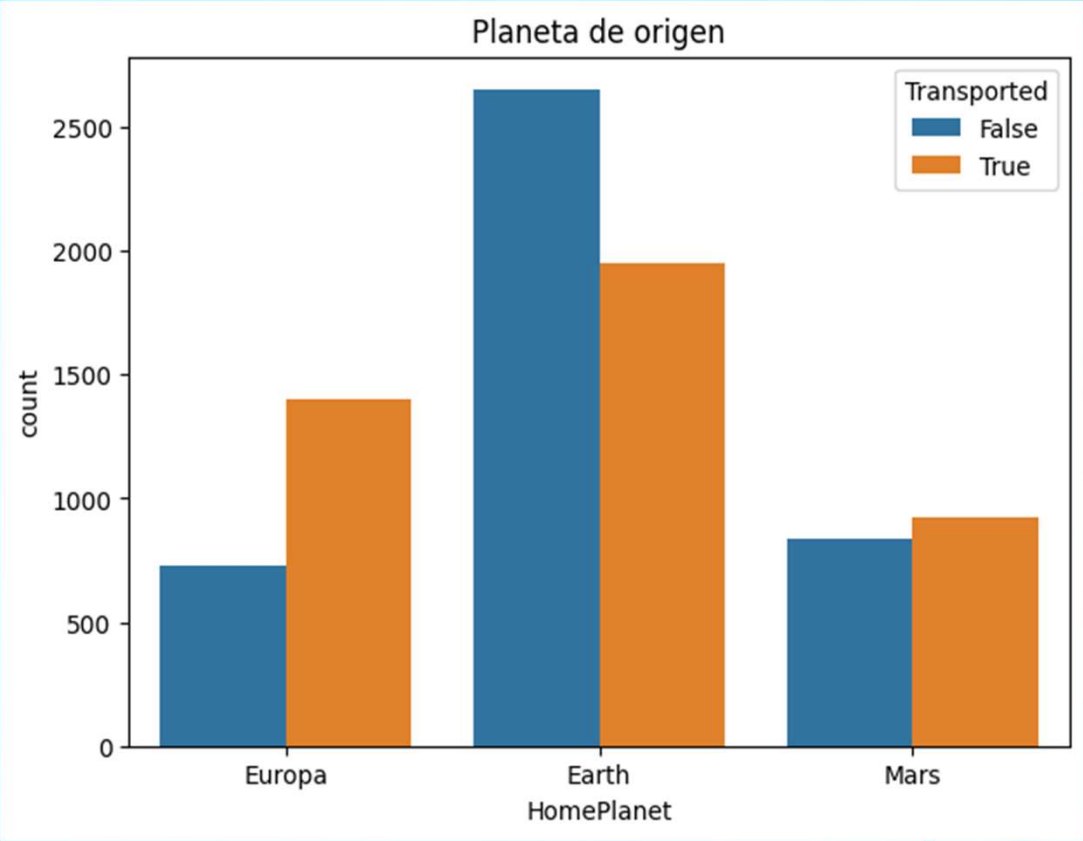
Tratamiento de las variables del dataset:

Destination y Name han sido eliminadas. No las necesito para predecir mi target. No me influye su eliminación.

Passenger_Id se deja tal cual.

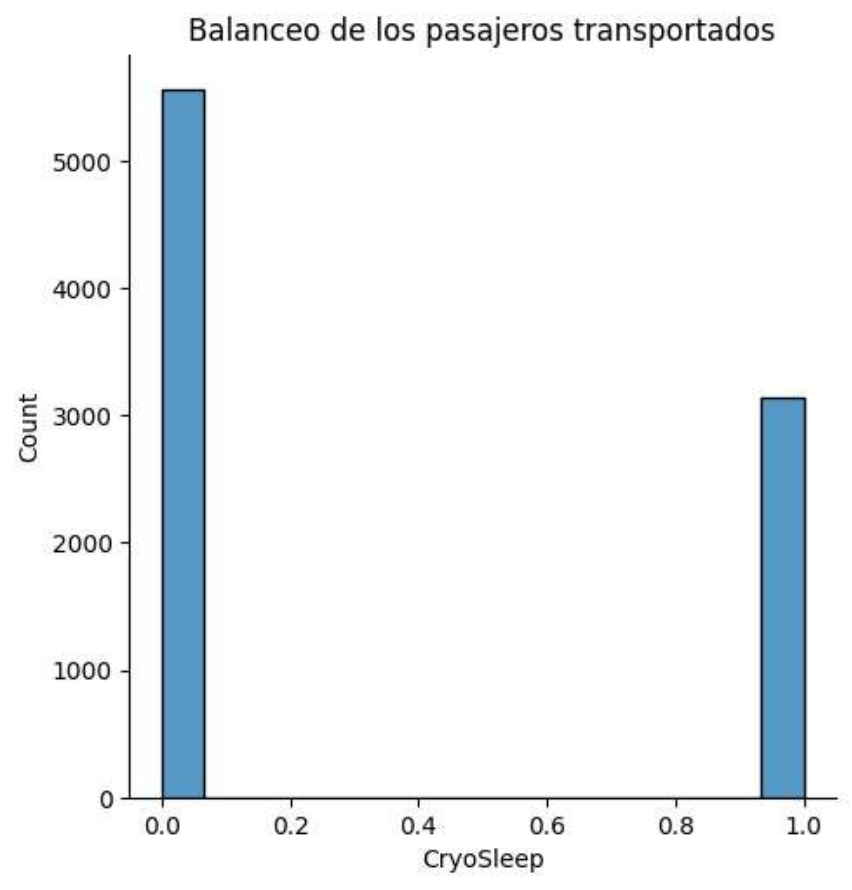
Homeplanet: esta formada por 3 valores: Earth, Europa y Mars. La distribución de ella se muestra a continuación:

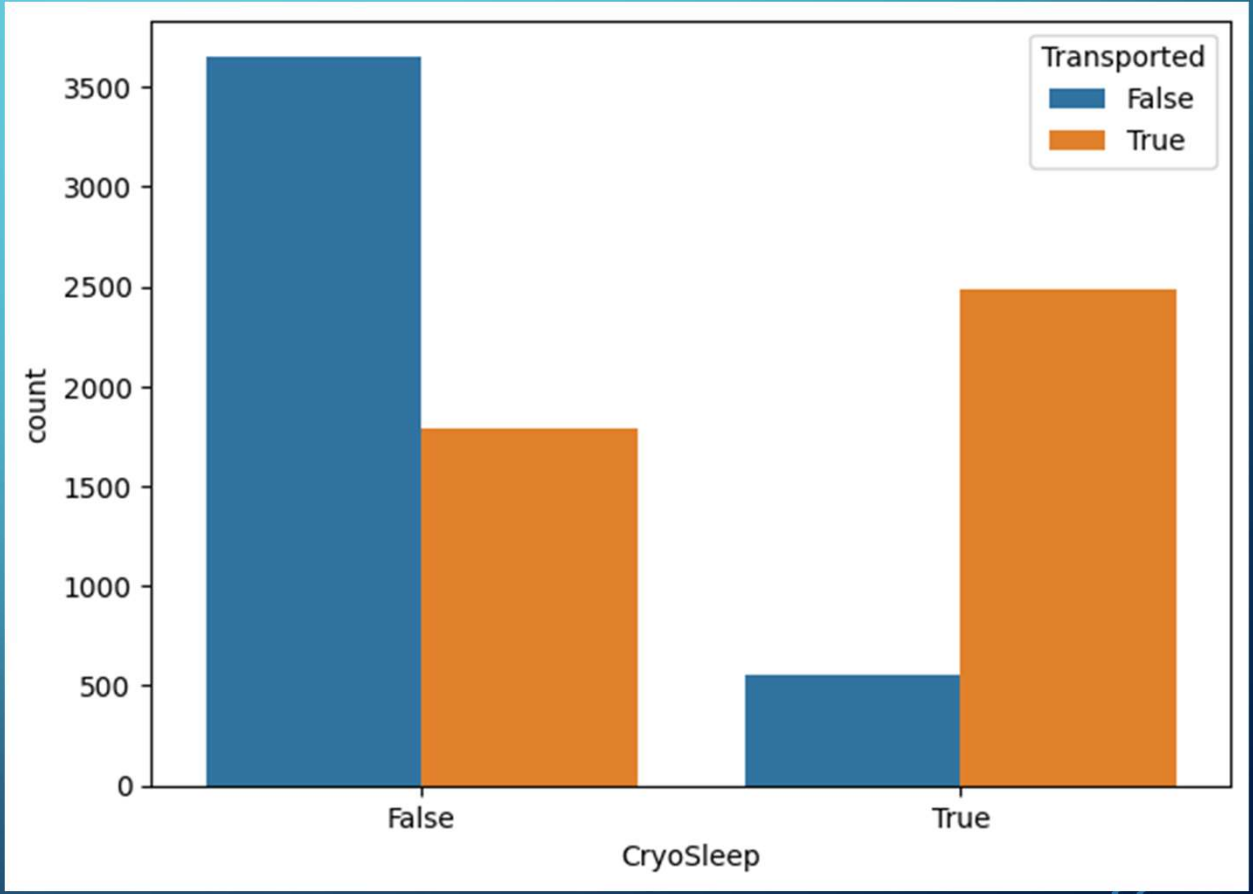




CryoSleep: para estos valores perdidos nos centramos en dos opciones :

- Si un pasajero tiene un gasto distinto de cero en cualquier categoría de gasto, supondremos que este pasajero no esta confinado en su cabina, por tanto rellenamos con un False.
- Si un pasajero tiene un gasto de cero en todas las categorías de gastos, este pasajero estará en animación suspendida en su cabina, por tanto rellenamos con un True.





Cabin : en esta variable rellenamos los missing con la cabina más popular entre los pasajeros. Los datos de cabina vienen dados de la forma cubierta/número/lado, donde lado puede ser P para babor o S para estribor, en mi caso me quedo solo con la informacion de si el pasajero se encuentra en babor o estribor, eliminando así la información adicional. Crearemos así una nueva columna con P o S, eliminando la columna Cabin. A continuación, el antes del procesado y el después:

Cabin
B/0/P
F/0/S
A/0/S
A/0/S
F/1/S
...
A/98/P
G/1499/S
G/1500/S
E/608/S
E/608/S

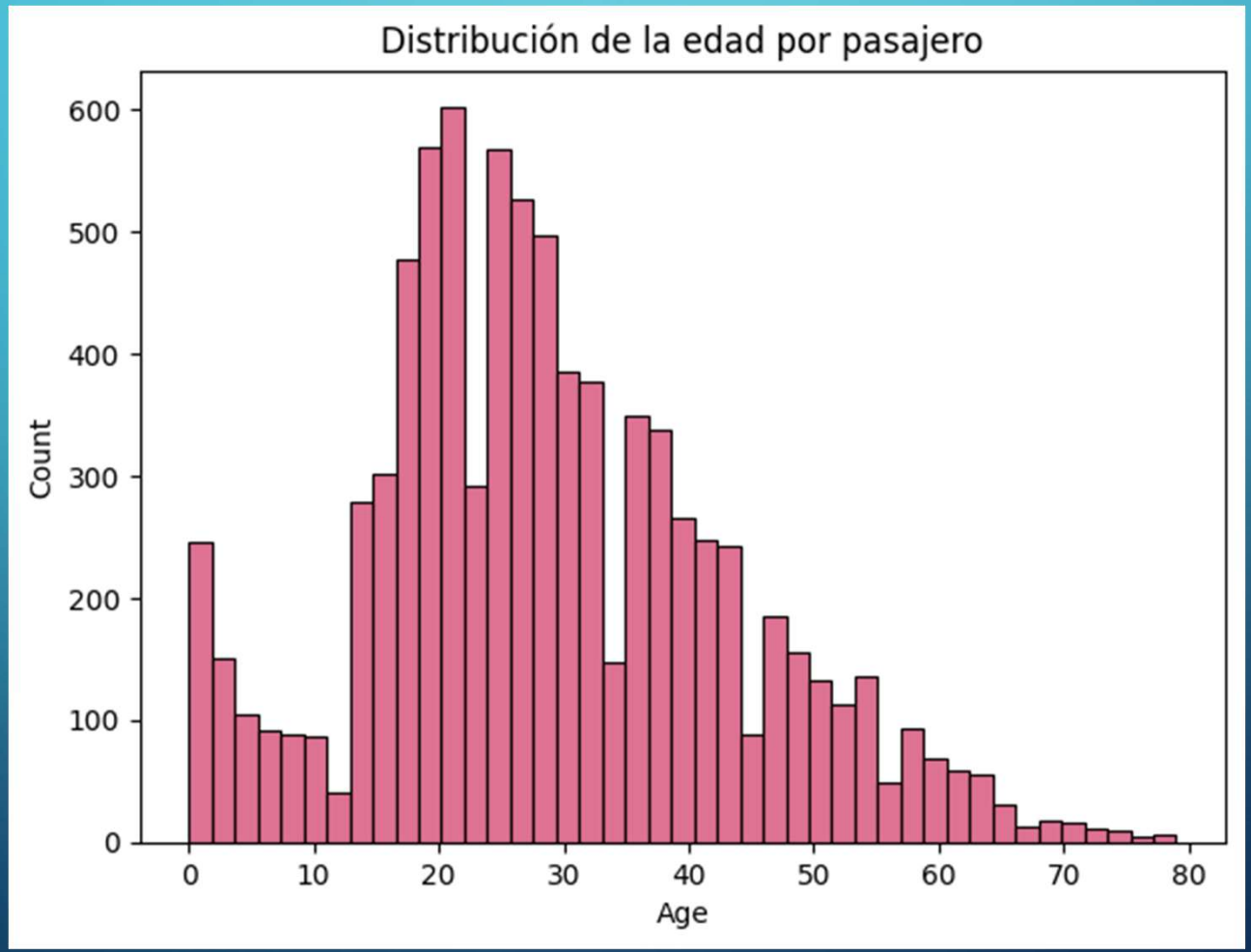
Cabin_Side_P	Cabin_Side_S	Cabin_Deck_A	Cabin_Deck_B	Cabin_Deck_C	Cabin_Deck_D	Cabin_Deck_E	Cabin_Deck_F	Cabin_Deck_G	Cabin_Deck_T
1	0	0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	1	0	0
0	1	1	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...
1	0	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0
0	1	0	0	0	0	1	0	0	0

Age: para esta variable numérica rellenamos aquellos valores missing con la edad mediana de los pasajeros.

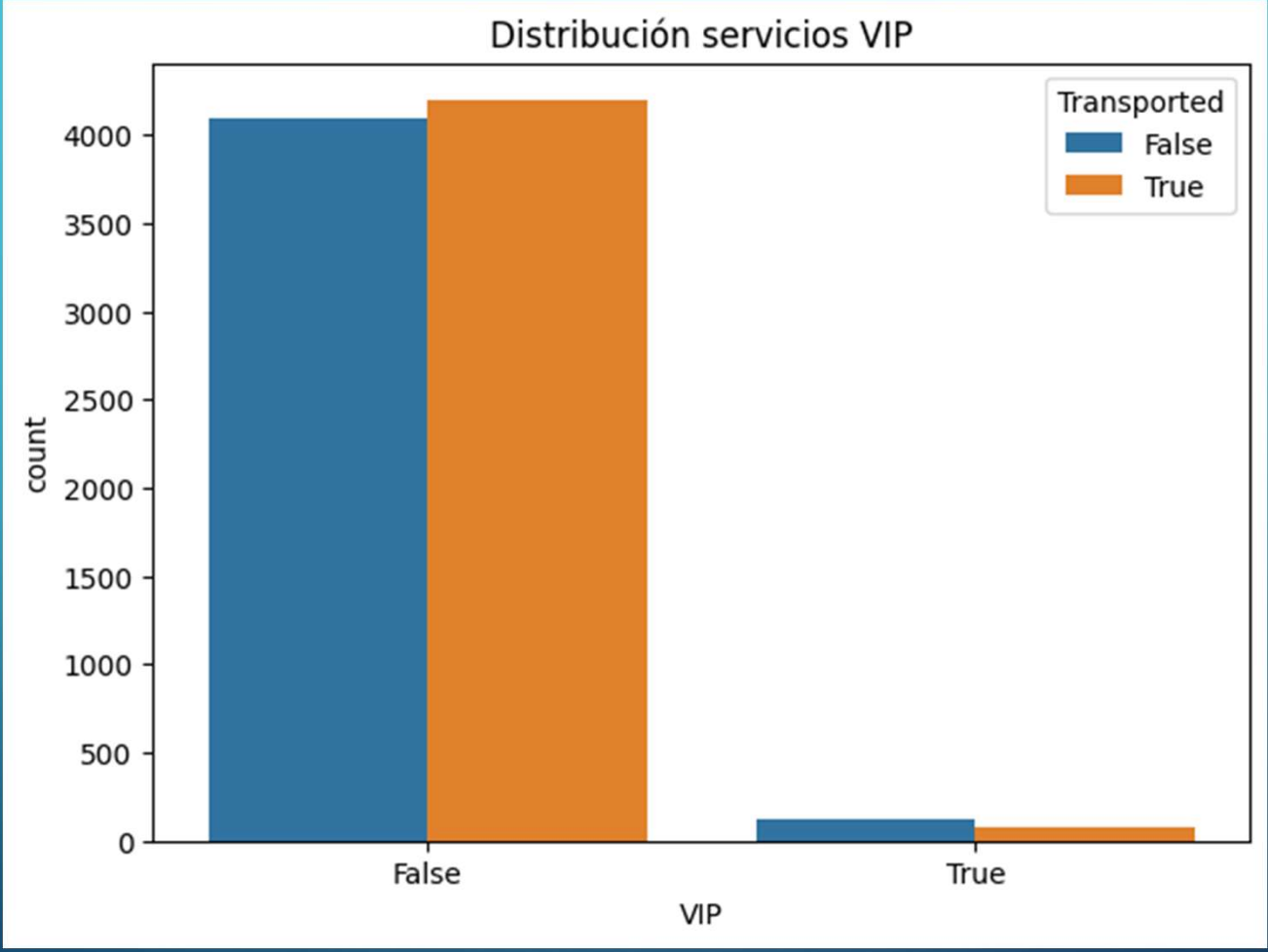
El calculo de la media y la mediana da los siguientes valores:

Edad Mediana: 27.00

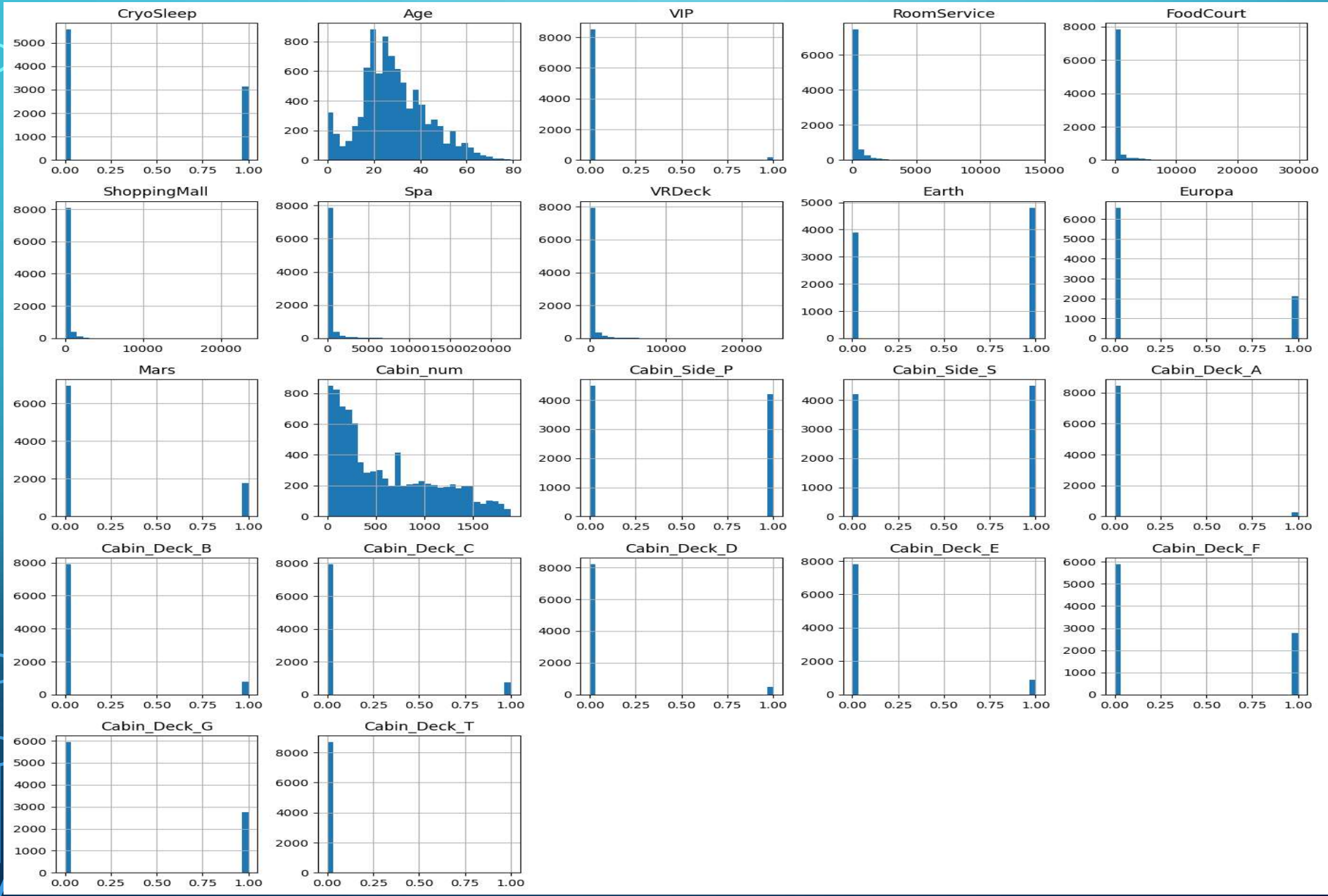
Edad Mediana: 28.82793046746535



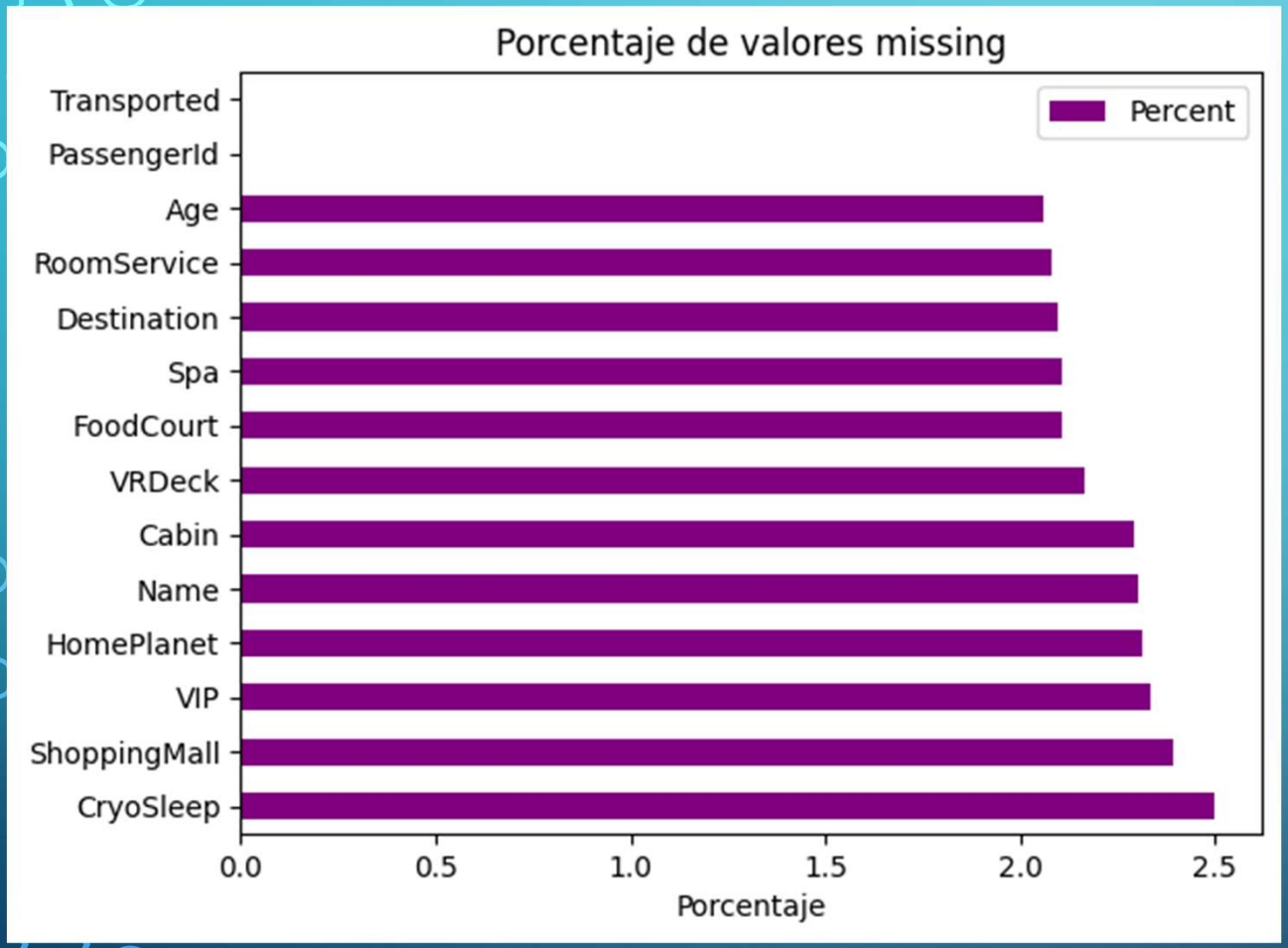
VIP: la mayoría de los pasajeros no tienen contratados los servicios VIP, por lo que rellenamos estos datos perdidos con un False.



Facturas de servicios extras: Completamos todas las facturas faltantes con 0. Tenemos en cuenta que aproximadamente la mitad de las personas a las que les falta alguna factura, están en modo crionizado, por lo que necesariamente tienen que tener en todas sus facturas un importe de 0.

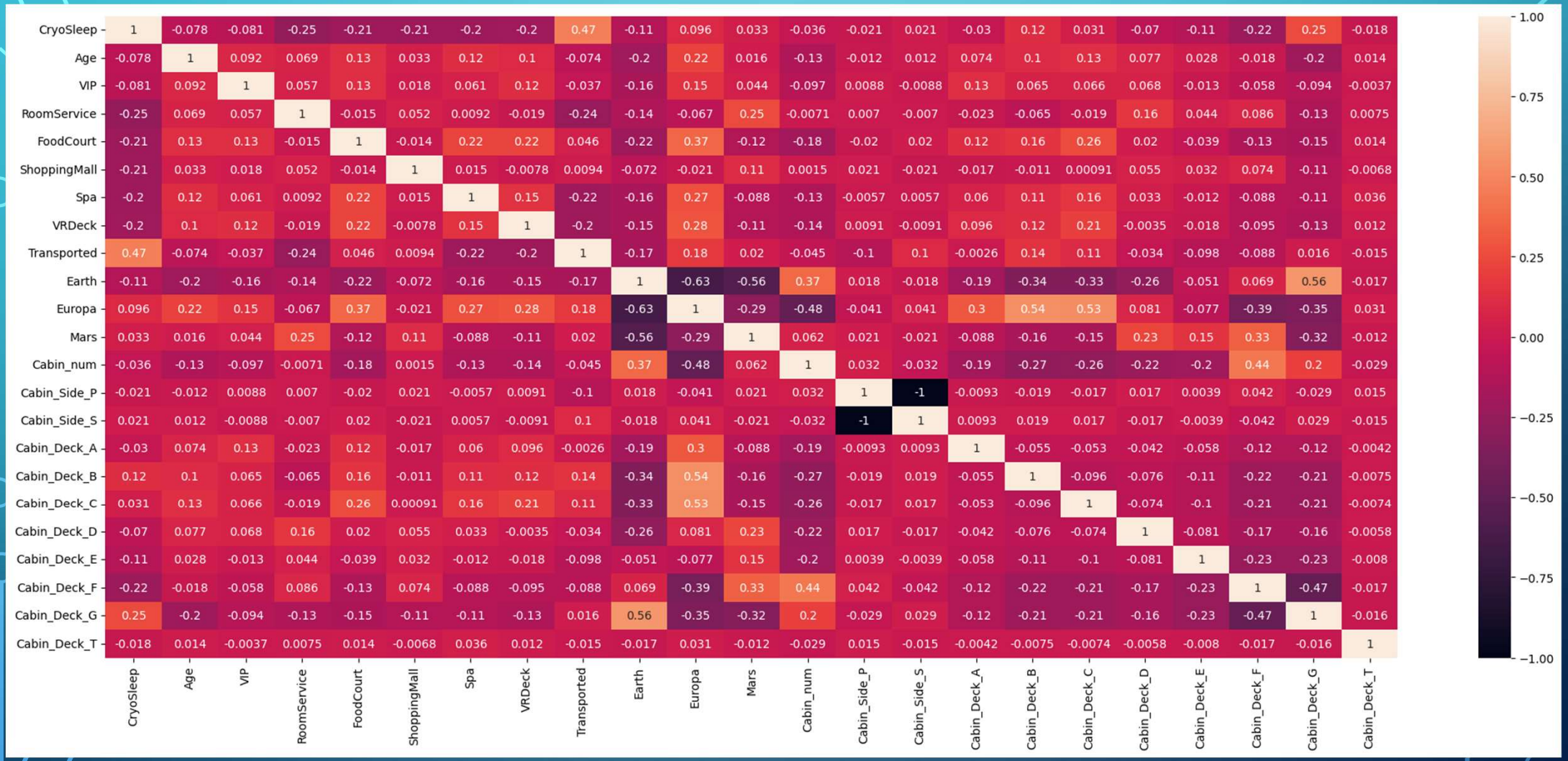


AQUÍ MUESTRO, COMO ES EL NÚMERO Y LA PROPORCIÓN DE VALORES NULOS CON RESPECTO AL TOTAL(8693 FILAS):

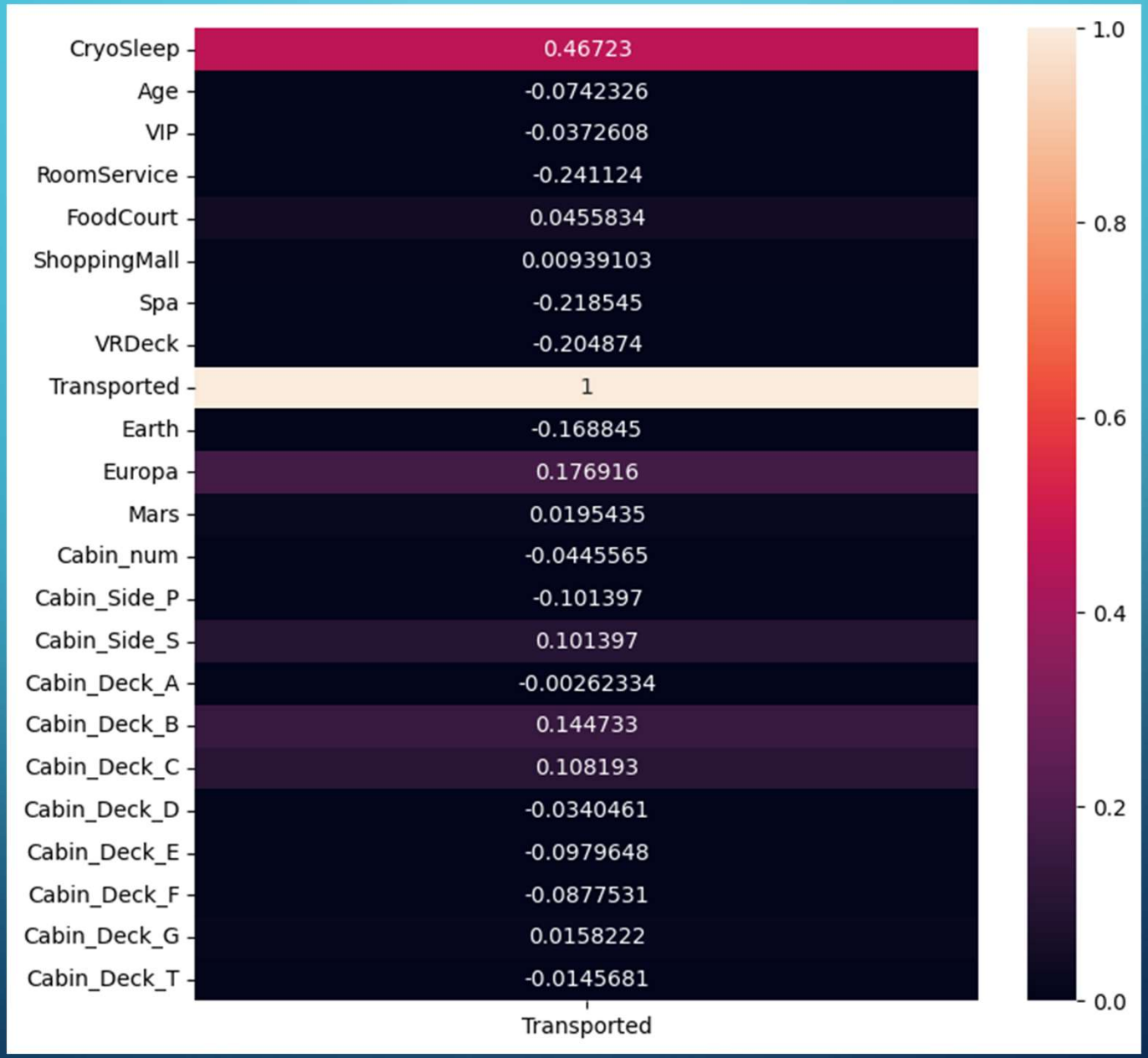


	Total	Percent
CryoSleep	217	2.496261
ShoppingMall	208	2.392730
VIP	203	2.335212
HomePlanet	201	2.312205
Name	200	2.300702
Cabin	199	2.289198
VRDeck	188	2.162660
FoodCourt	183	2.105142
Spa	183	2.105142
Destination	182	2.093639
RoomService	181	2.082135
Age	179	2.059128
PassengerId	0	0.000000
Transported	0	0.000000

CORRELACIONES GENERALES

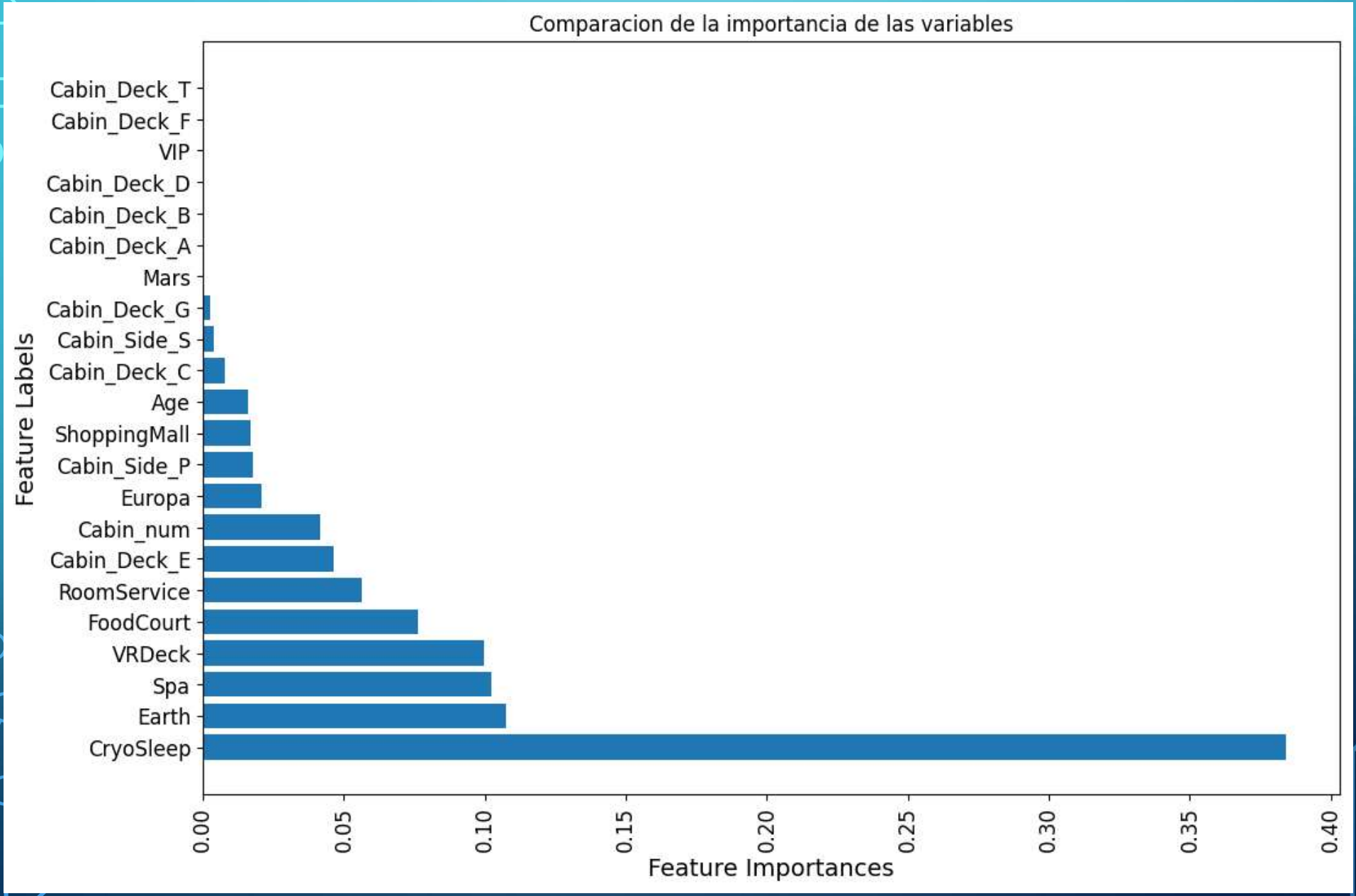


CORRELACIONES CON TARGET

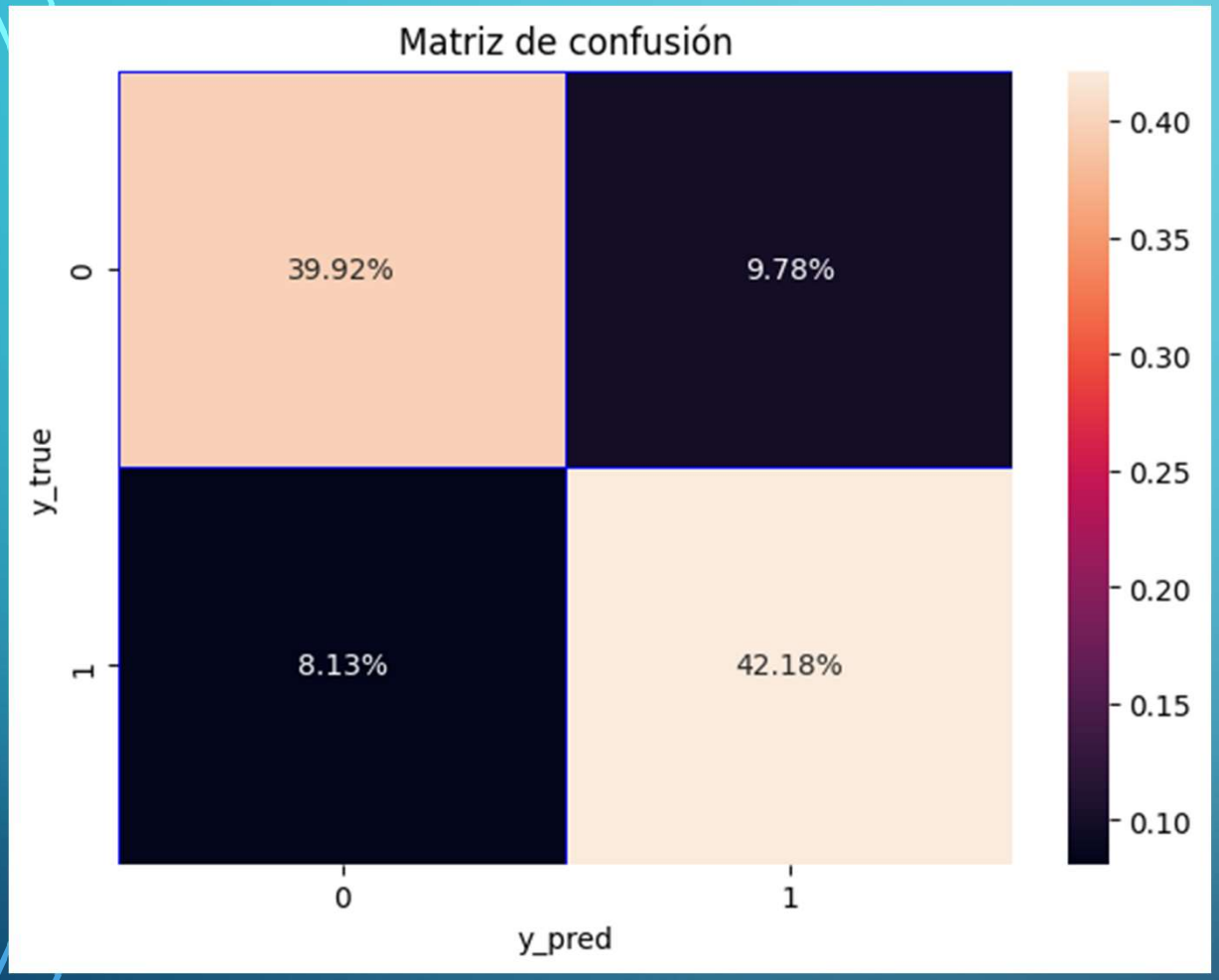


DECISION TREE

FEATURE IMPORTANCE



DECISION TREE



HIPERPARÁMETROS

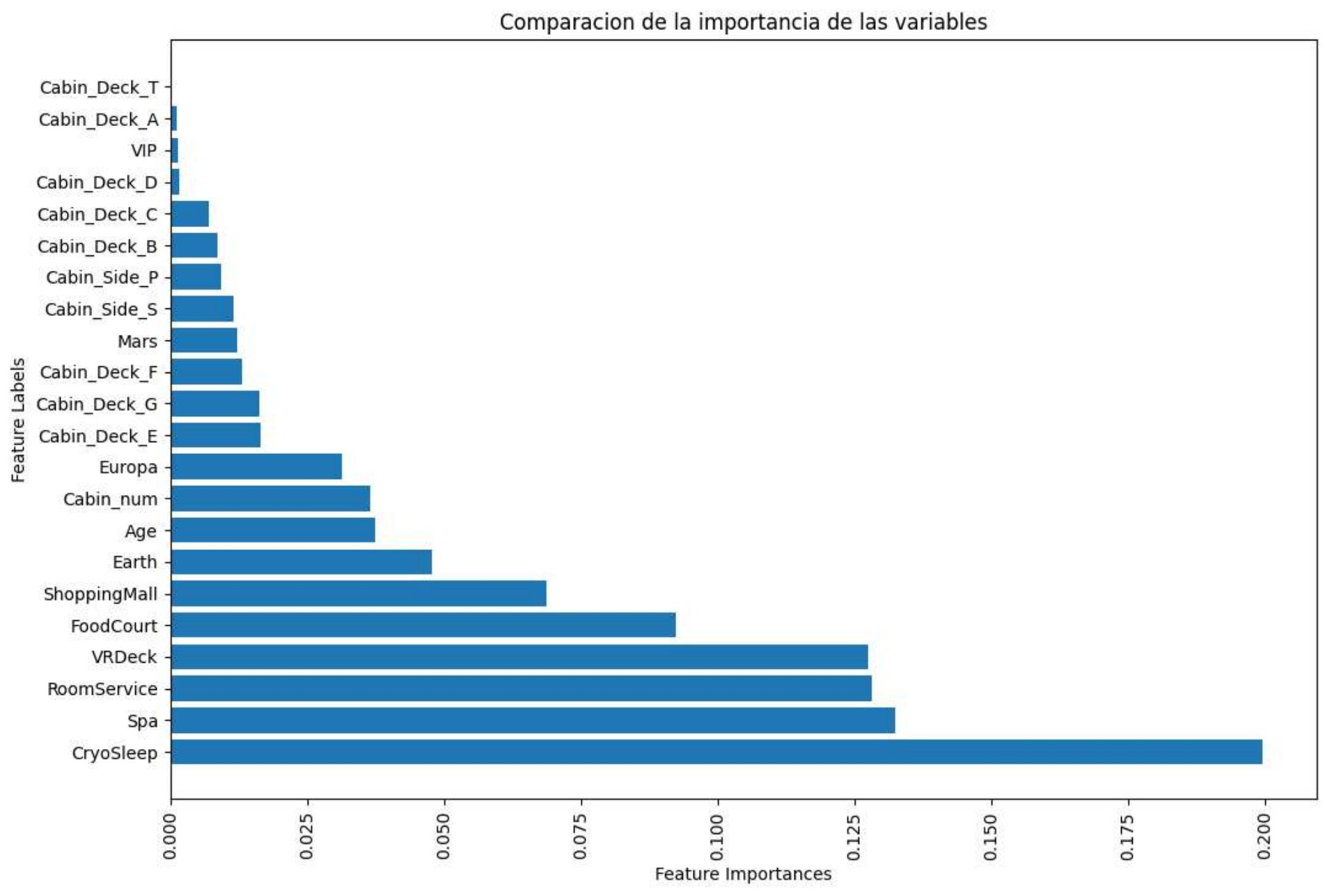
- max_depth=8,
- random_state=17,
- criterion='entropy’
- min_samples_split=65,
- min_samples_leaf=4,
- min_weight_fraction_leaf=0,
- max_leaf_nodes=50

METRICA

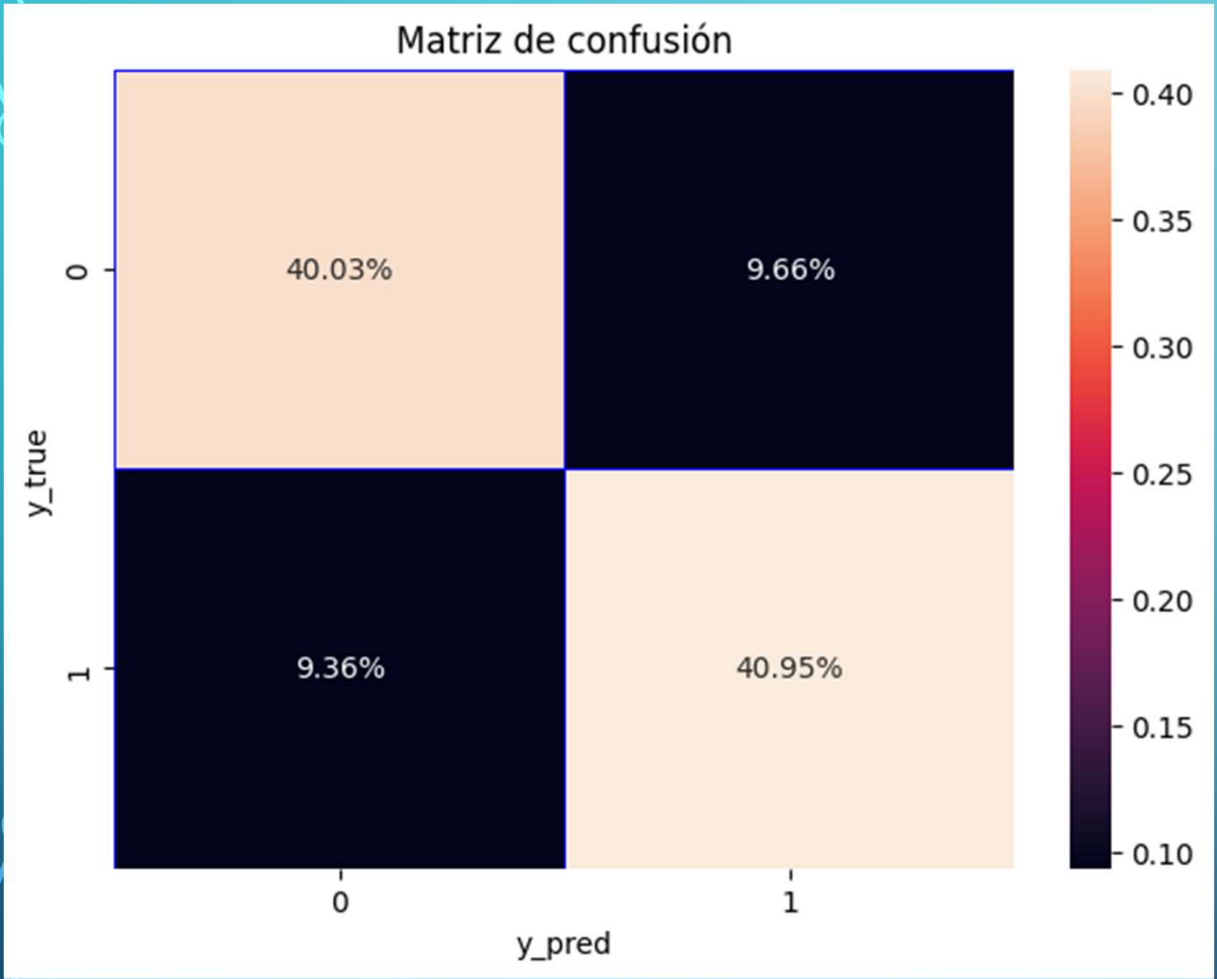
ACCURACY TRAIN: 0.8410846343467543
ACCURACY TEST: 0.8209355828220859

RANDOM FOREST

FEATURE IMPORTANCE



RANDOM FOREST



HIPERPARÁMETROS

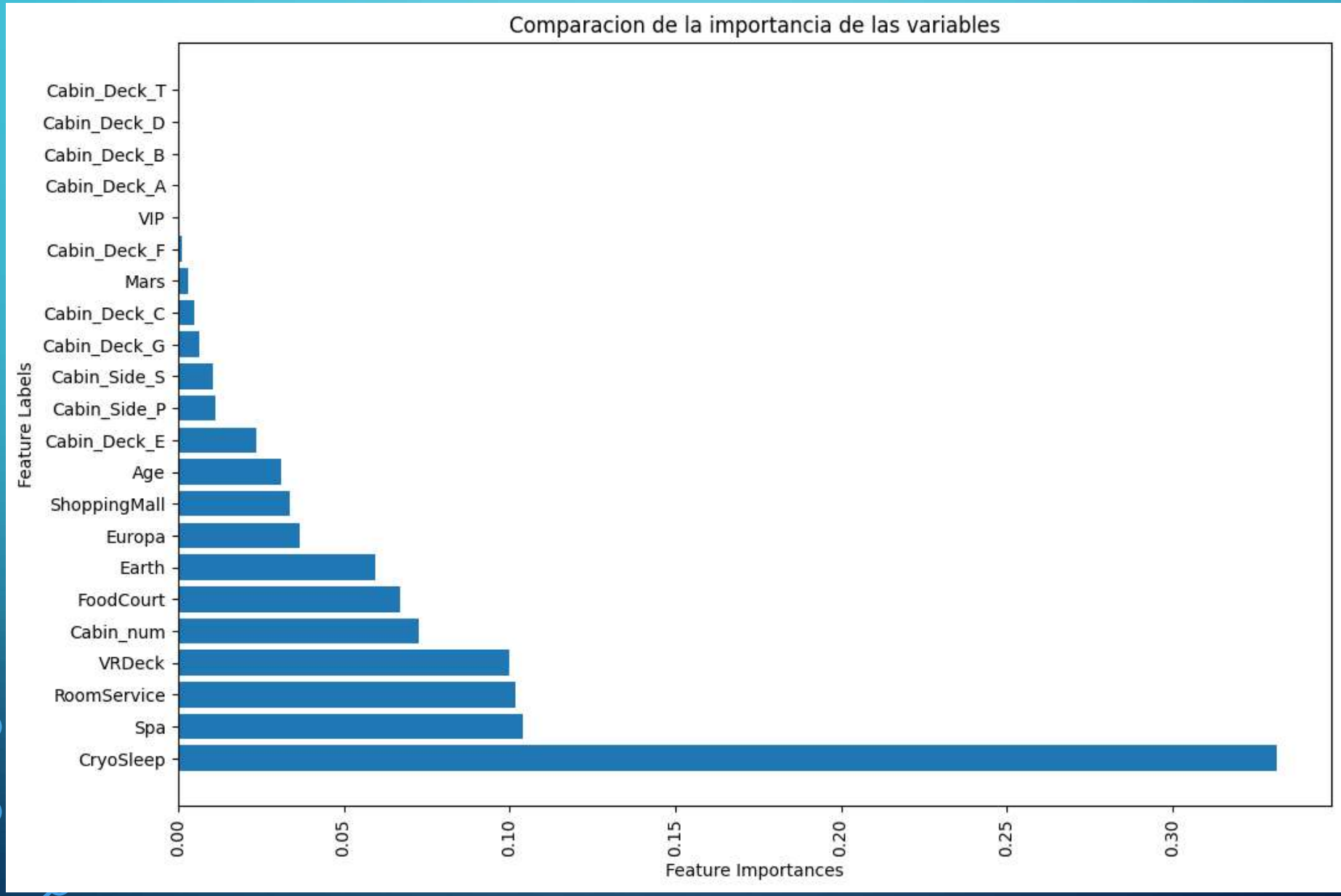
- n_estimators=100
- random_state=17
- min_samples_leaf=1
- max_depth=7
- max_features=5
- bootstrap= True
- n_jobs=-1

METRICA

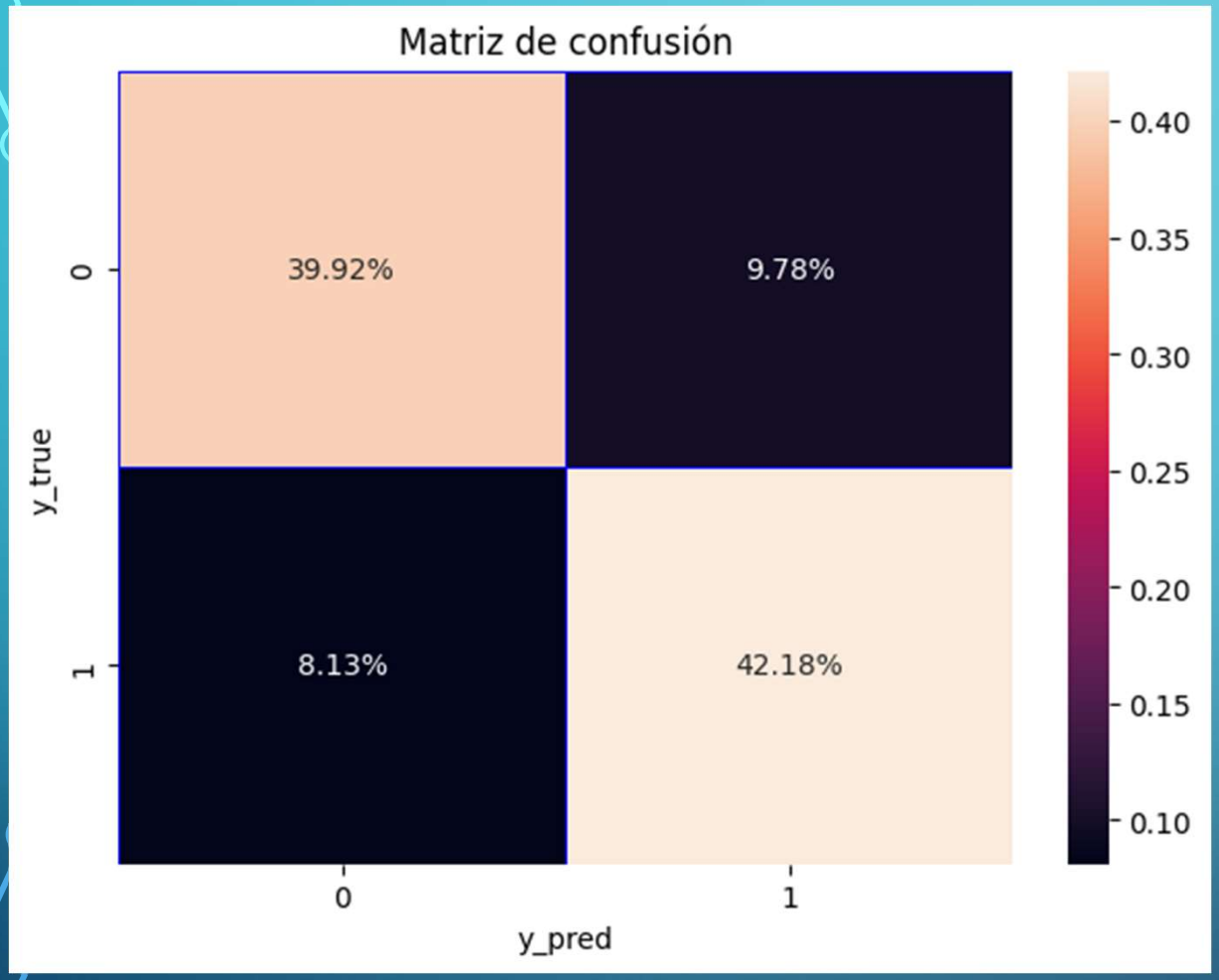
TRAINING SCORE: 0.8210353327855382
TEST SCORE: 0.8098159509202454

GRADIENT BOOST

FEATURE IMPORTANCE



GRADIENT BOOST



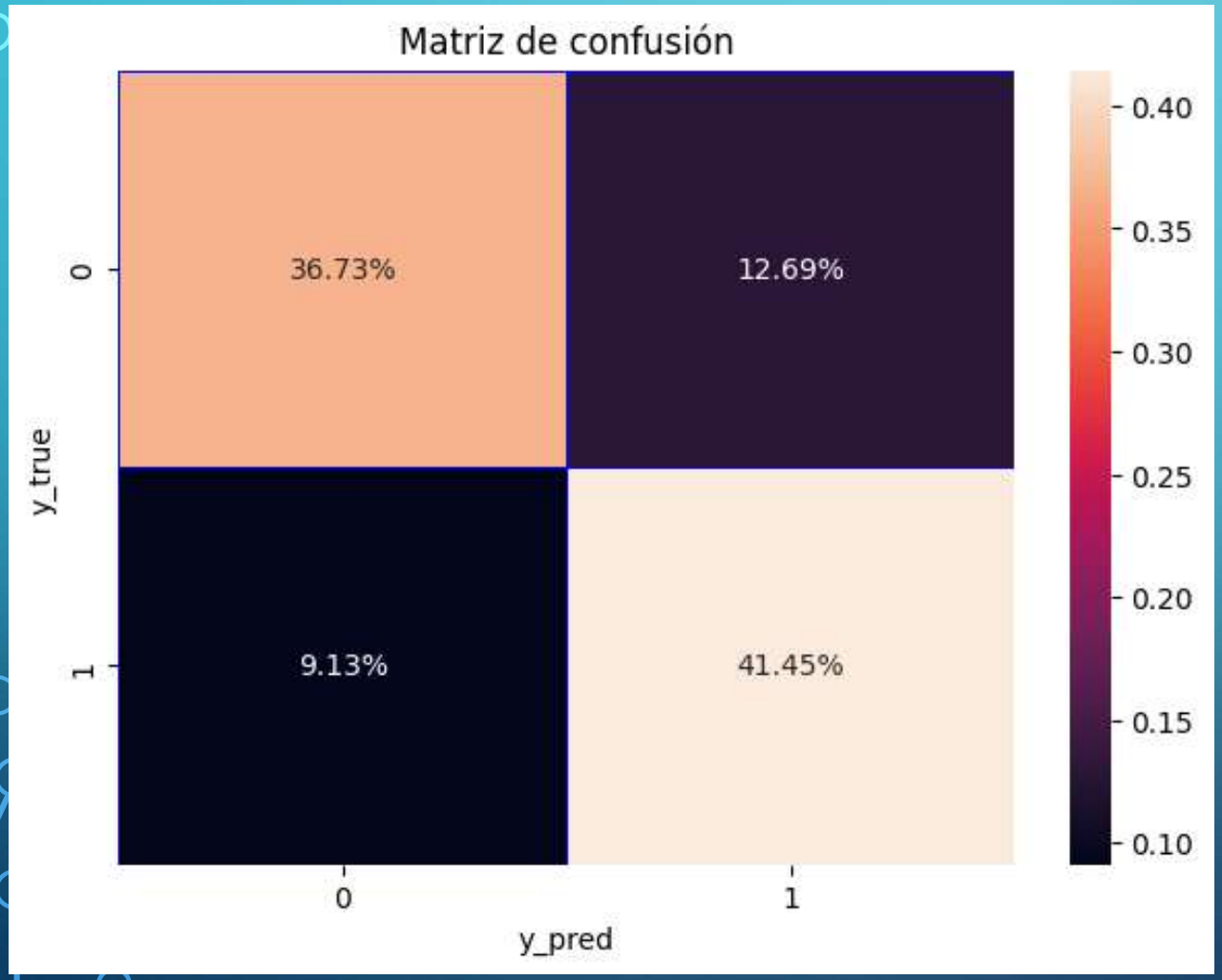
HIPERPARÁMETROS

- max_depth = 2
- n_estimators = 650
- learning_rate=0.1

METRICA

ACCURACY TRAIN: 0.8410846343467543
ACCURACY TEST: 0.8209355828220859

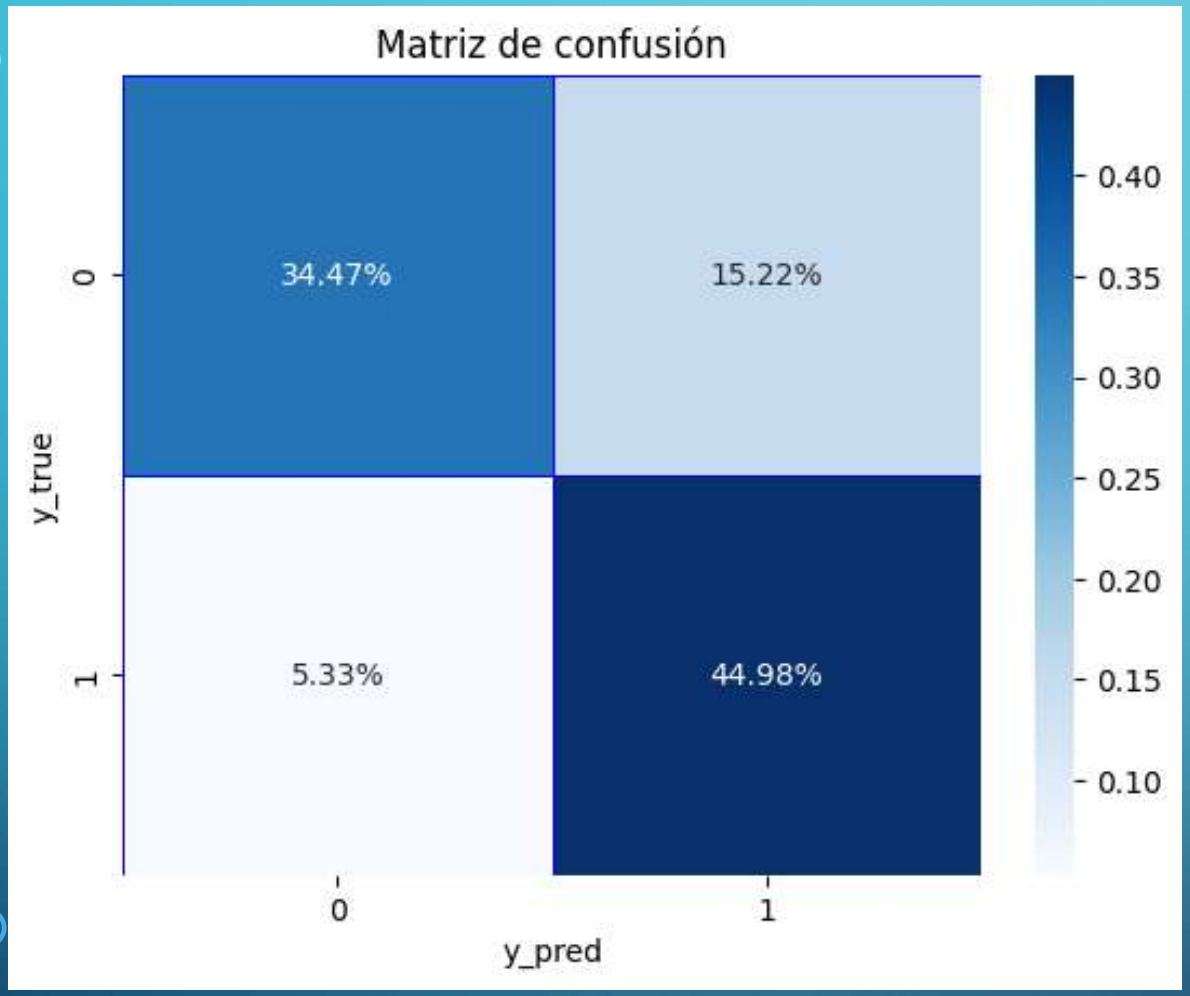
LOGISTIC REGRESSION



HIPERPARÁMETROS

- `max_iter = 83`
- `solver='newton-cg’`
- `random_state=42`
- `multi_class='auto’`
- `C=1.5`

SUPER VECTOR MACHINE



HIPERPARÁMETROS

- C=.1
- class_weight='balanced’
- kernel='rbf'

METRICA

MODELO SVC TRAINING ACCURACY = 0.7715694330320461
TEST ACCURACY = 0.7845092024539877

ENSEMBLE

HIPERPARÁMETROS

- **LogisticRegression**(max_iter = 83, solver='newton-cg', multi_class='auto', C=1.5, random_state=17)
- **RandomForestClassifier**(n_estimators=100, random_state=17) # Se usa en combinación con otro arg
- **SVC**=(kernel="rbf", C=.1, gamma="scale", random_state=17)
- **SGDClassifier**= (loss="hinge", learning_rate="constant", eta0=0.001, alpha=alpha, max_iter=1000, tol=1e-3, random_state=42)

METRICA

LOGISTIC REGRESSION = 0.7971625766871165
RANDOM FOREST CLASSIFIER= 0.8052147239263804
SGD CLASSIFIER = 0.7718558282208589
SVC = 0.7837423312883436
VOTING CLASSIFIER = 0.8059815950920245

CONCLUSIONES

Intenté sumar las variables de facturas, sumarlas todas en una nueva variable y eliminar todas las anteriores para reducir el numero de columnas. Pero predice peor.

El resultado representado en esta presentación, es el que mejor resultados da. Con cada modelo, he probado todos los hiperpárametros posibles que nos muestra la documentación de sklearn. He dejado la combinación que mejor métrica da de todas.

Random forest es el modelo que mejor predice de todos. Y es con el que mejor ranking he obtenido en kaggle.