UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Proyecto Final: Spaceship Titanic

Fernando Valdivia

Departamento de Matemática
Universidad Técnica Federico Santa María

December 5, 2023

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Definición del Problema

Contextualización

Nos situamos en el año 2912, donde la nave espacial Titanic fue lanzada hace un mes, y al sufrir un accidente, alrededor de la mitad de los pasajeros fueron transportados a una dimensión alternativa.

Problema

Se solicita una predicción del destino de los pasajeros mediante Machine Learning, para lo cual es necesario tanto analizar como procesar los datos para aplicar modelos.

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Dataset

Conjunto de Datos

Se cuenta con dos archivos que almacenan datos: El conjunto de entrenamiento (*"train_df.csv"*) y el conjunto de prueba (*"test_df.csv"*), cuya diferencia radica en la presencia de una columna denominada *"Transported"*, la cual indica si el pasajero fue transportado o no.

Atributos

Ambos archivos con los conjuntos de datos poseen los atributos:

Atributos Categóricos: "PassengerID", "HomePlanet", "CryoSleep", "Cabin", "Destination", "VIP", "Name".

Atributos Continuos: "Age", "RoomService", "FoodCourt", "ShoppingMall", "Spa", "VRDeck".

Estadística Descriptiva

Comparación de promedios de atributos continuos:

	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

Figure: Archivo "train_df"

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	4186.000000	4195.000000	4171.000000	4179.000000	4176.000000	4197.000000
mean	28.658146	219.266269	439.484296	177.295525	303.052443	310.710031
std	14.179072	607.011289	1527.663045	560.821123	1117.186015	1246.994742
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%	26.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	37.000000	53.000000	78.000000	33.000000	50.000000	36.000000
max	79.000000	11567.000000	25273.000000	8292.000000	19844.000000	22272.000000

Figure: Archivo "test_df"

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

HomePlanet, CryoSleep, Destination, VIP

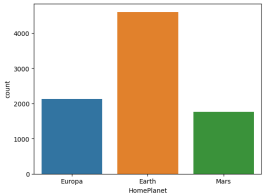


Figure: HomePlanet

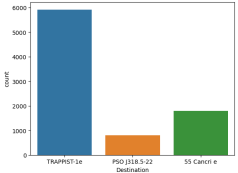


Figure: Destination

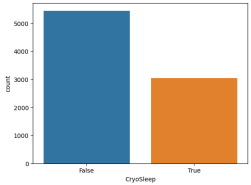


Figure: CryoSleep

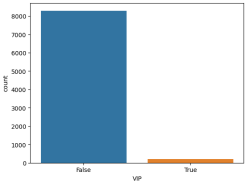


Figure: VIP

Age, RoomService, FoodCourt vs Transported

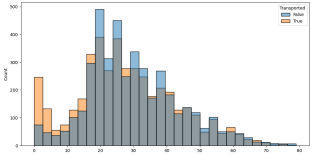


Figure: Age

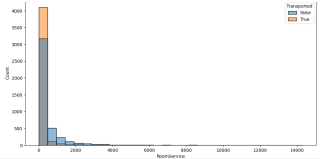


Figure: RoomService

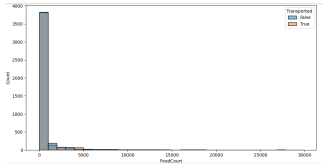


Figure: FoodCourt

ShoppingMall, Spa, VRDeck vs Transported

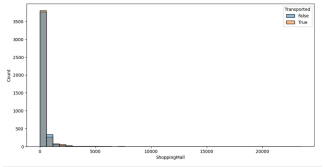


Figure: ShoppingMall

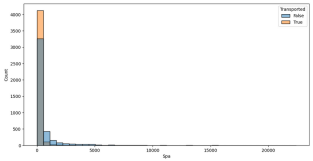


Figure: Spa

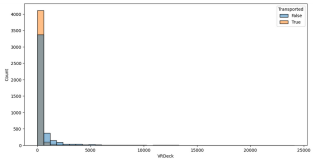


Figure: VRDeck

Transported

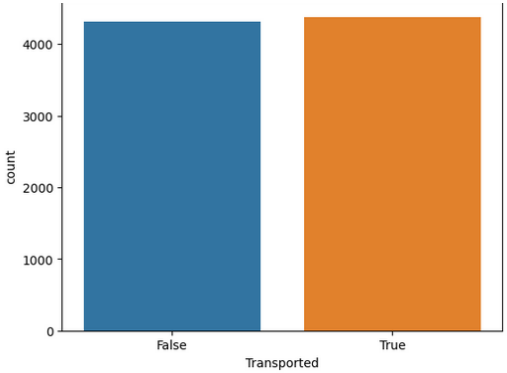


Figure: Transported

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Limpieza de Datos

Para la limpieza de datos, se seguirán los siguientes pasos:

Fraccionar el conjunto entre datos de entrada y salida.

Se revisan los datos que presentan **NaN**.

Atributos categóricos con NaN: Se revisan las clases por atributo, se eliminan los atributos que no aporten, luego en los demás se reemplazan por la moda de la columna.

Atributos continuos con NaN: Se reemplaza el valor por la media de la columna.

Fracción de Conjuntos

Se fracciona train/test en:

train cat/test cat: Se agrupan atributos categóricos que deben ser procesados en un *encoder* para luego ser utilizados.

train cont/test cont: Se agrupan atributos continuos que deben ser procesados por un *scaler* para luego ser utilizados en un modelo.

En otros conjuntos particulares se archivaron los atributos "VIP" y "CryoSleep".

Encoders y Scalers

Encoder: One Hot Encoder, este encoder sirve para atributos categóricos que no tengan un orden entre ellos a costa de aumentar la dimensionalidad del problema.

Scaler: Standard Scaler, este scaler sirve para eliminar la media y cambiar la varianza a 1, con lo que no habrán problemas para entrenar un modelo.

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Modelos a Entrenar

En este proyecto se usaron 3 modelos de la librería sklearn de Python y una red neuronal a partir de la librería tensorflow:

- Support Vector Classifier (SVC)
- Decision Tree Classifier
- Random Forest Classifier
- Red Neuronal Densa

Optimización de Hiperparámetros

Se ejecutó la función train test split de sklearn para así entrenar y probar el modelo con el conjunto *train_df.csv*.

Entrenamiento de modelos: SVC y Decision Tree Classifier

En cuanto al modelo SVC, se probaron los siguientes valores:

C: 0.1, 1, 10.

kernel: linear, RBF

Resultando los mejores parámetros C=10 y **kernel**='RBF'.
Para el modelo DTC, se analizaron los siguientes valores:

max depth: 1, 5, 10, 15, 20, 25, 30.

min samples split: 20, 50, 100, 150, 200.

Resultando como valores óptimos max depth=10 y min samples split=100.

Entrenamiento de Modelos: Random Forest Classifier

Para el modelo RFC se probaron los siguientes valores:

n estimators: 50, 100, 150.

max depth: 1, 5, 10, 15.

min samples split: 20, 50, 100.

Resultando como mejores parámetros n estimators = 100, max depth = 15 y min samples split = 50.

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Métricas en *train_df.csv*

Como métrica se usó la precisión pues ésta es la usada por kaggle para este desafío, de forma que, las métricas obtenidas para los modelos en train df fueron las siguientes:

SVC: Se obtuvo una precisión de 0.77918.

Decision Tree Classifier: Se obtuvo una precisión de 0.77976.

Random Forest Classifier: Se obtuvo una precisión de 0.78838.

Red Neuronal: Se obtuvo una precisión de 0.77378.

Métricas en *test_df.csv*

Se realizaron las predicciones de cada modelo sobre *test_df.csv*, se subieron a **Kaggle** y se obtuvieron los siguientes resultados:

SVC: Se obtuvo una precisión de 0.79167.

Decision Tree Classifier: Se obtuvo una precisión de 0.78115.

Random Forest Classifier: Se obtuvo una precisión de 0.79003.

Red Neuronal: Se obtuvo una precisión de 0.79261.

Contenidos

1 Introducción

2 Estadística Descriptiva

3 Visualización Descriptiva

4 Preprocesamiento

5 Selección de Modelo

6 Métricas y análisis de resultados

7 Visualizaciones del Modelo

8 Conclusiones

Matrices de confusión

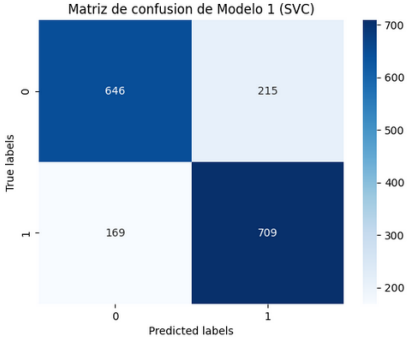


Figure: MatrizSVC

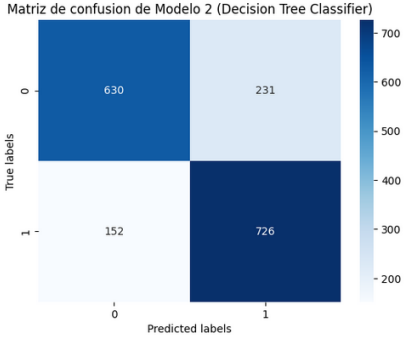


Figure: MatrizDTC

Matrices de Confusión

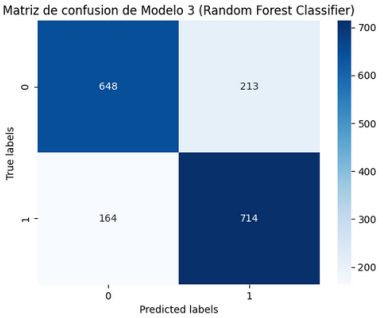


Figure: MatrizRFC

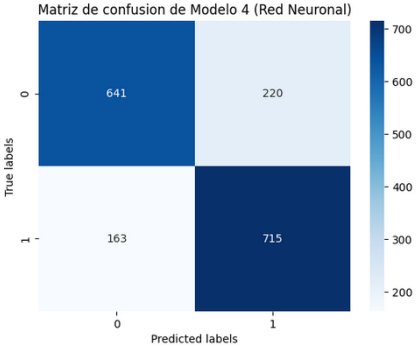
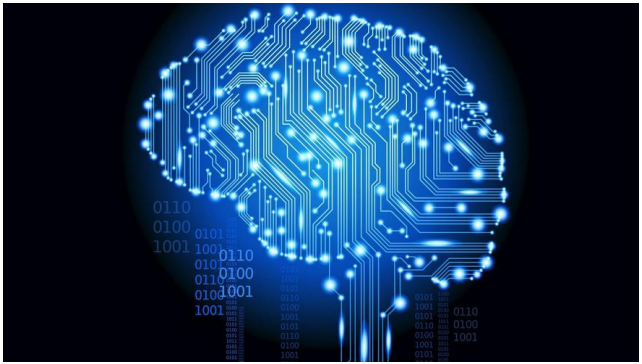


Figure: Matriz Red Neuronal

Contenidos

1. Introducción
2. Estadística Descriptiva
3. Visualización Descriptiva
4. Preprocesamiento
5. Selección de Modelo
6. Métricas y análisis de resultados
7. Visualizaciones del Modelo
8. Conclusiones

Conclusión



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Proyecto Final: Spaceship Titanic

Fernando Valdivia

Departamento de Matemática
Universidad Técnica Federico Santa María

December 5, 2023