

Segunda Entrega de Proyecto

Inteligencia Artificial

Código: 2508401

Grupo: 01

Fecha: 22/ 08/ 2022

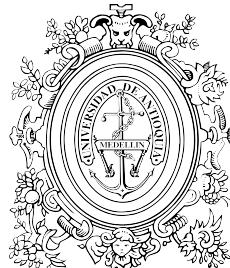
Elaborado por

Sofía Gutiérrez Tangarife, CC: 1001576929

Melissa Galeano Ruiz, CC: 1000416463

Profesor

Raúl Ramos Pollán



**UNIVERSIDAD
DE ANTIOQUIA**

1 8 0 3

Facultad de Ingeniería

Departamento de Ingeniería Mecánica

Universidad de Antioquia

Contents

1 Descripción de la estructura del notebook desarrollado	2
1.1 Cleaning Dataset	2
1.2 Dataset Preparation	3
1.3 Dataset Analysis	3
1.3.1 Correlation Matrix	3
1.3.2 Confusion Matrix	4
1.4 Model Evaluation	5
2 Referencias	6

1 Descripción de la estructura del notebook desarrollado

El notebook desarrollado para este proyecto esta divido en cuatro secciones principales:

1. Cleaning Dataset
2. Dataset Preparation
3. Dataset Analysis
4. Model Evaluation

En cada una de ellas se realiza y se describe detalladamente el análisis llevado a cabo con los datos de entrenamiento para evaluar la información que se tiene del problema, y de esta manera, generar un modelo predictivo de datos funcional que permita predecir cuáles pasajeros podrían ser enviados a la dimensión alterna por la anomalía.

1.1 Cleaning Dataset

En el análisis y procesamiento de datos para modelos de aprendizaje supervisado se debe hacer una buena gestión de la información que se tiene para poder obtener modelos con un alto porcentaje de acierto en las predicciones.

En esta primera sección del notebook, se hace una análisis cualitativo de los datos en donde se observa la información de todas las posibles variables junto con sus valores. Además, en este punto cabe resaltar que se evalúa la cantidad de datos nulos y no nulos, ya que la falta de información se debe ajustar para poder realizar un análisis cuantitativo posterior que permita generar predicciones significativas.

Una vez se determinan cuantos valores nulos y no nulos se tienen en cada una de las columnas del dataset, y a qué tipos de datos corresponden (Discretos o Continuos), se comienza el proceso de reparación de datos. El proceso de reparación de datos consiste en evaluar cada una de la columnas con sus datos faltantes correspondientes (Datos nulos) y reemplazarlos por el mejor tipo de dato que se adapte dado el caso.

Para el dataset que se tenía, se realizaron los siguientes reemplazos de los datos en cada una de las columnas:

- **Name** Se decide eliminar los datos faltantes ya que no representan una cantidad significativa en comparación con los datos totales, además de que al tratarse de nombre personales, no se recomienda reemplazar con datos existentes porque cada persona es única, ni crear nombres aleatorios.
- **Destination** Se decide reemplazar con el dato mas recurrente, pues es razonable que los datos faltantes se reemplacen con el destino mas concurrido por los usuarios.
- **Cabin, CryoSleep, Age** Se decide reemplazar con el valor del dato anterior.
- **RoomService, Foodcourt, ShoppingMall, Spa, VRDeck** Se decide reemplazar con el valor medio.
- **VIP** Se decide reemplazar con el dato que mas se repite.

1.2 Dataset Preparation

Una vez que el dataset esta reparado se procede con su preparación ya que para el proyecto asignado debe cumplir con los siguientes requisitos:

- Al menos 5000 instancias (filas, imágenes, etc.).
- Al menos 30 columnas.
Al menos el 10% de las columnas han de ser categóricas.
- Al menos ha de tener un 5% de datos faltantes en al menos 3 columnas.

Para esto, se desarrollaron las siguientes seis funciones que fueron aplicadas sobre el dataset para generar más columnas de datos que permitieran categorizar de una forma más específica.

1. `split Passenger` Esta función se aplica sobre la columna `passenger` para dividirla en dos columnas nuevas que permitieran diferenciar entre ID y el grupo en el que se encuentran los pasajeros.
2. `one hot` Esta función se aplica sobre las columnas `HomePlanet` y `Destination` para dividirlas en en columnas discretas con cada uno de sus elementos.
3. `split Name` Esta función se aplica sobre la columna `Name` para dividirla en dos columnas diferentes llamadas `Name` y `Last Name`.
4. `TF` Esta función se aplica sobre las columnas `Cryosleep`, `VIP`, `Transported` para crear nuevas columnas booleanas.
5. `taken services` Esta función se aplica para todas las columnas de los servicios que ofrece el spaceship, para crear columnas categóricas con cada uno de ellos.
6. `age range` Esta funcion se aplica sobre la columna `Age`, para subdividirla en tres columnas con tres diferentes rangos de edades `Youth`, `Adult`, `Older`

Finalmente, una vez se confirma que se cumplen con los requisitos se inicia el proceso de análisis de datos para saber como proceder con el modelo predictivo.

1.3 Dataset Analysis

Una vez se tiene el dataset totalmente reparado para el inicio del proceso de creación del modelo, se debe hacer un óptimo análisis de datos para comprender la correlación entre las variables y así definir las variables mas significativas para el modelo predictivo.

1.3.1 Correlation Matrix

Este proceso se inicia con la matriz de correlación ya que esta nos permite visualizar mediante un rango de colores la relación entre cada una de las variables. Como se visualiza en la siguiente figura, en esta matriz se tiene un rango entre $[-1, 1]$ en donde cada numero es equivalente corresponde un color y refleja la relación entre variables, siendo el blanco = 1, la correlación más alta que se puede obtener, y el negro = 0 , la más baja.

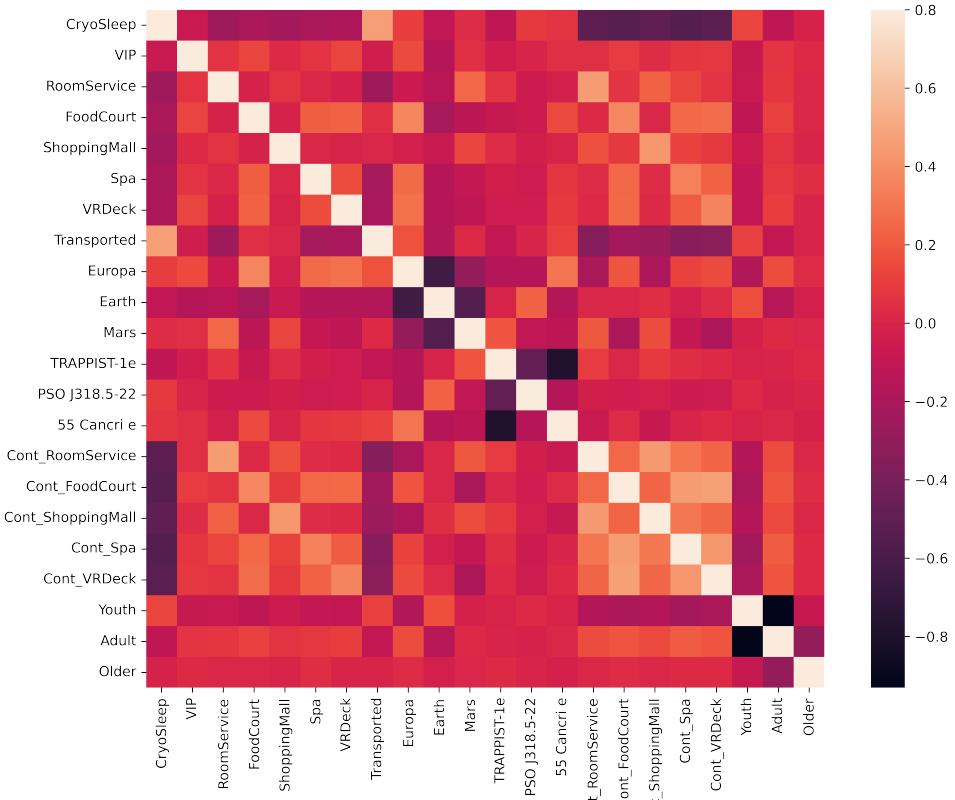


Figure 1: Matriz de Correlación

Es importante destacar que como la variable **Transported** es la que se desea predecir, esta es la que se debe analizar con mas detenimiento pues se necesita verificar cuales son las variables con una correlación mas significativa con respecto a ella. Por tanto, mediante esta matriz se puede visualizar que la variable **Cryosleep** es la variable con la mayor correlación.

Subsecuente a esto se procede a analizar la frecuencia de repetición de cada una de las variables en el dataset mediante histogramas para tener en cuenta otras posibles causas de que los pasajeros sean transportados y asimismo, para obtener una mejor visualización de la distribución de los datos.

1.3.2 Confusion Matrix

Considerando la información obtenida mediante los histogramas de cada una de las columnas del dataset, ahora se analizara la correlación entre la variable **Cryosleep** y **Transported** mediante una matriz de confusión.

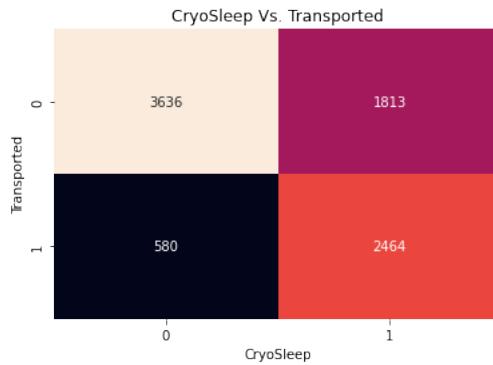


Figure 2: Matriz de Confusión

Como se visualiza, los pasajeros que estaban dentro de las cryosleep son los que tienen una mayor probabilidad de ser transportados a otras dimensiones durante el viaje en el spaceship a comparación de los pasajeros que no las utilizaron.

1.4 Model Evaluation

Para evaluar cuál es el mejor modelo para el dataset del problema se inició con la creación de una copia del dataset para evitar modificar los datos originales. Adicionalmente, se estableció el la columna de **Passenger** como el index del dataset para visualizar cuales de los pasajeros serian los posibles transportados, y se eliminan los datos discretos categóricos ya que estos no contribuyen al proceso de predicción.

Se divide el dataset de entrenamiento inicial en 70% de datos para entrenamiento y el 30% restante para datos de validación.

A continuación, se crea un ciclo *for* para evaluar los datos con 5 posibles estimadores y ver su porcentaje de acierto, con el cual, se decide cuál es el modelo mas óptimo para los datos.

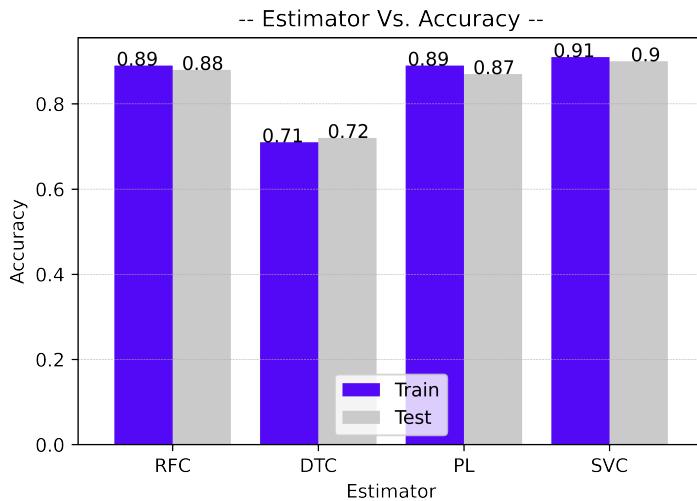


Figure 3: Comparación entre diferentes estimadores

Donde:

- *RFC* : Random Forest Classifier, con `max depth = 30` y `n estimators = 5`
- *DTC* : Decision Tree Classifier, con `max depth = 2`
- *PL* : Pipeline, con `n components = 2` y `SVC(gamma = 1)`
- *SVC* : Support Vector Classifier, con `gamma = 1`

Como se observa en la figura, el mejor porcentaje de acierto se obtuvo con el Support Vector Machine, lo que quiere decir que este es el estimador mas óptimo para el modelo de predicción.

2 Referencias

[1] *Spaceship Titanic | Kaggle*. (n.d.). Retrieved July 1, 2022, from <https://www.kaggle.com/competitions/spaceship-titanic/overview>