Data Analysis

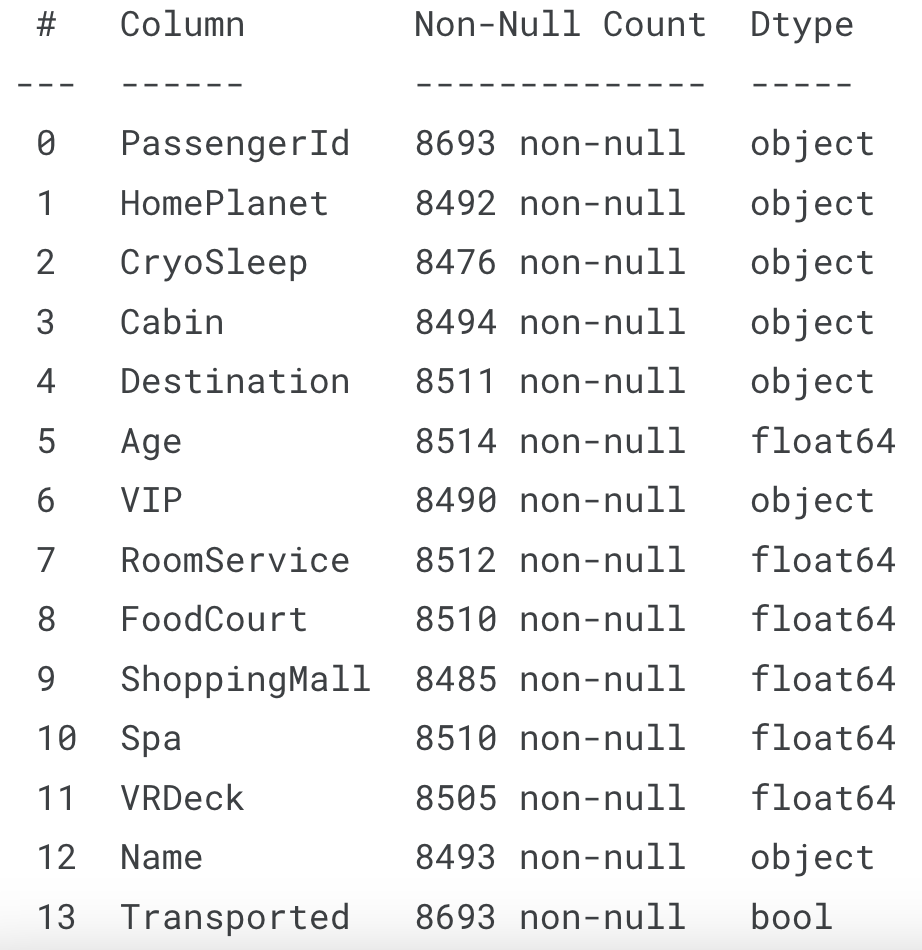
Define Objectives:  
In this competition, the task is to predict whether a passenger was transported to an alternate dimension during the Spaceship Titanic’s collision with the spacetime anomaly.

In order to achieve this goal, I will try several machine learning algorithms including K-Nearest Neighbors, Logistic Regression, Random Forest and deep learning. These techniques perform greate results as to classification problems.

Data Collection:

The data is divided into two segments: a training set and a test set. The training set is subsequently partitioned into a smaller training subset and a validation subset. The validation subset is utilized for tuning hyperparameters, while the actual test set is employed to assess the model's performance. The accuracy metric for the model is determined based on the results from this real test set.

Data Cleaning:



The summary of the dataset indicates that missing values are present in every column except for 'PassengerId' (index column). The data attributes can be classified into two types: categorical and numerical. To address these missing values, the most frequent category is used to fill in the gaps in categorical data. Meanwhile, for numerical data, the missing values are imputed with the feature's average value.

Data Preprocessing

Data dictionary:

PassengerId - A unique Id for each passenger. Each Id takes the form gggg\_pp where gggg indicates a group the passenger is travelling with and pp is their number within the group. People in a group are often family members, but not always.

HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.

CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.

Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard.

Destination - The planet the passenger will be debarking to.

Age - The age of the passenger.

VIP - Whether the passenger has paid for special VIP service during the voyage.

RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.

Name - The first and last names of the passenger.

Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.

Feature Engineering

Upon looking into the dataset, it becomes apparent that the 'Cabin' column includes three distinct features: the deck, the cabin number, and the side of the ship (port or starboard). Thus the next step is extracting these individual elements—'deck,' 'num,' and 'side'—from the 'Cabin' column and imputing any missing values with the most frequently occurring category for each feature.

Moreover, the dataset contains three 'homeplanet' categories and two 'destination' categories. Instead of applying one-hot encoding to these categorical variables, I have assigned numerical representations to each category, as this has yielded a higher accuracy score in this context.

An additional observation is that the 'PassengerId' field is composed of a group name and the number of people in that group, suggesting a potential feature: a count representing the size of each passenger's group. This is predicated on the assumption that passengers are more likely to assist others within their group.

Finally, columns that offer minimal predictive value, such as 'PassengerId,' 'name,' 'Cabin,' and 'Deck,' will be removed from the dataset.

It is also important to note that while I considered categorising 'age' into different groups, this strategy resulted in a decreased accuracy for the models.

Data Transformation

First I separated the features and the target variable. Subsequently, I applied the train\_test\_split() function from the scikit-learn library to partition the data into training and validation sets. Following this division, I applied a Standard Scaler to normalise the numerical features. This step is crucial because numerical features may have varying ranges, and models that are sensitive to the scale of the data could be adversely impacted if the features are not standardised.

Data Modeling

K Nearest Neighbors (KNN):

The KNN algorithm is utilized due to its simplicity and effectiveness in classification problems where the relationship between features is not linear. It excels in scenarios where the decision boundary is irregular. The algorithm's reliance on feature similarity allows it to make predictions based on how closely data points resemble each other, which can be particularly useful in systems where proximity is a strong indicator of equivalence.

Logistic Regression:

Logistic regression is chosen for its efficiency in binary classification tasks. It provides a probabilistic framework that outputs the probability of the target variable belonging to a specific class. This model is beneficial for its interpretability, as it allows for an understanding of the impact of each feature on the likelihood of outcomes. It is particularly useful when the relationship between the independent variables and the log-odds of the dependent variable is approximately linear.

Random Forest:

The Random Forest algorithm is selected due to its robustness and ability to handle overfitting when dealing with large datasets with many features. It operates by constructing multiple decision trees and voting on the most popular output class for classification or averaging the prediction for regression. This ensemble method improves prediction accuracy and generalizability by reducing the variance that individual decision trees might exhibit.

Deep Neural Network (DNN):

A Deep Neural Network is applied to capture complex patterns and interactions within the data. Its multiple layers and non-linear processing units enable it to learn high-level features in data, which is a significant advantage in tasks requiring feature extraction and recognition. DNNs are particularly useful in dealing with large-scale and high-dimensional data, such as image and speech recognition, where the intricate structure of the data can be leveraged for more accurate predictions.