Final Project Proposal

Source: Kaggle Competition Spaceship Titanic

https://www.kaggle.com/competitions/spaceship-titanic/overview

Introduction:

Our final project is based on a Kaggle competition "Spaceship Titanic," which focuses on predicting which passengers were transported to an alternate dimension during the spaceship's journey. The evaluation is based on the accuracy of these predictions, aiming to develop a predictive model capable of accurately identifying passengers affected by the anomaly in the "Spaceship Titanic" Kaggle competition. Objectives include a thorough analysis and understanding of the provided dataset, implementing data cleaning, and preprocessing techniques to prepare the data for modeling, investigating the influence of various features on the prediction outcomes, and comparing the performance of multiple machine learning algorithms to determine the most effective approach for this unique challenge. This structured approach will enable us to address the problem with scientific rigor and precision.

Data Description:

The data given has three files, “train.csv”, “test.csv”, and “sample\_submission.csv”. The training dataset has around 8700 observations while the testing dataset has roughly 4300 observations. Key variable that can be used for training our model include “PassengerId ”, “HomePlanet ”, “CryoSleep ”, “Cabin ”, “Destination ”, “Age ”, “VIP ” etc.

Possible Algorithms:

We tend to utilize feature engineering for processing the data and handling missing data. On the classification task, we might use classification algorithms taught in class such as KNN, random forest, decision tree, SVM, XGBoost, and additional algorithms Resnet, Unet. For each model, we perform a grid search using cross-validation to find the best parameters for the corresponding model, further employing ensemble and stacking techniques to incorporate various models.

Evaluation:

Our final result will be evaluated based on our [classification accuracy](https://developers.google.com/machine-learning/crash-course/classification/accuracy), the percentage of predicted labels that are correct.