**Group 3 plan**

The Spaceship Titanic data is the dataset we will be using for the assignment. Upon further Exploratory of the dataset on kaggle it was noticed that most of the machine learning model trained using the dataset got an accuracy of around 80%.

In this competition your task is to predict whether a passenger was transported to an alternate dimension during the Spaceship Titanic's collision with the space-time anomaly. To help you make these predictions, you're given a set of personal records recovered from the ship's damaged computer system.

* 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.

**Exploratory Data Analysis**

Each group member was task on preparing the exploratory data analysis of the spaceship titanic competition in order to gain more insight to the datasets being used. To get the best use of the datasets further preparations were done, which is convert CryoSleep and VIP to a Boolean, Map categorical variables to number.

**Distribution of features**

Most passengers are young within the age of 18 – 25 however the baby population were fairly high, most passengers are going to TRAPPIST and there are very few VIPs. All the amenities services have a skewed distribution caused by few very high values with Cabin numbers are more concentrated in the first values

**Initial training with base model**

Three (3) machine learning model was used to determine thedirection we are going to take for more advanced training. The models used so far are as follows:

Linear regression

Logistic regression

Deep learning model.

These models output similar result to the average results on kaggle with around 80% accuracy.

**Further Training**

Further training will involve more feature engineering and more complex model being used. An ensemble model is one of the models that will be used in further training of the model. The dataset is a very interesting dataset in overall.

**Group memberTask**

Each group member has a development branch on the group kaggle repository. Each group member has been task on training 3 models for the dataset. The best model result will be used for the final submission.