**Assignment\_arya.ai**

**Load data**

Panda library is used to load the training and testing data.

**EDA**

First for easiness of referring to column name, a list is created with 1 to 57 name.

Outlier are seen with the jointplot and treated since "X4" gives 99 percentile value 0 so this column is excluded for treatment.

No null value so no treatment needed here.

"y" value are only 0 and 1 so no treatment needed here.

**Feature selection**

This is a classification predictive modeling problem with numerical input variables. For this type of problem statistical correlation based models are preferred.

Anova – F test based correlation best for linear model

Kendall Tau – Null hypothesis based best for non linear model

Mutual Information gain – Entropy gain model

For feature selection two feature selection technique are used Mutual\_information\_gain and other is Kendalltau these feature work best for classification model since data is non linear so Anova is not chosen . According to these two feature two subset are created.

**Test-train data split and Scaling**

Validation and train data are created according to Given instruction.

**Model selection**

For binary classification Logistic regression a basic model is chosen with min max scaler standardization.

Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problem in decision trees and also reduces the variance and therefore improves the accuracy and since this data have large number of feature so this model is selected.

**Prediction and submission**

In this scenario Kendal tau feature selection technique with Random forest produce better accuracy with validation accuracy of 96.035%.