Data Exploration:

The data was first explored as done in every machine learning contest.The steps involved to understand, clean and prepare our data for building our predictive model were:

- Variable Identification
- Univariate Analysis
- Bi-variate Analysis
- Missing values treatment
- Outlier treatment
- Variable transformation
- Variable creation

On applying these steps it was noticed that the data was already in a clean format without any outlier and missing values .So after the analysis part we moved directly onto the Training Part which is described below.

Training:

The problem was a classification problem as we were asked to classify the test dataset into 4 different classes. Different models like Logistic Regression, Linear Discriminant Analysis, Kneighbors Classifier, DecisionTree Classifier, Gaussian Naiye Bayes and ensemble methods like Ada Boost, Gradient Boosting, Random Forest and Extra Trees were trained on the training data using **crossvalidatio**n on 80% of the training data. 20% data was kept for testing the performance on unseen data.
Now after the models have been trained an ensemble of all these models was used using the **Voting Ensemble for Classification** with the ensemble of BaggingClassifer, DecisionTreeClassifier,ExtraTreeClassifier seeing which algorithm gave better F1 score for which of the classes.

Accuracy of various algo's used.

- LogisticRegression : 0.827921
- LinearDiscriminantAnalysis : 0.843306
- KNN : 0.944703

- DecisionTreeClassifier : 0.973911
- GaussianNaiveBase: 0.778015
- AdaBoosting : 0.860992 (0.027297)
- GradientBoosting: 0.974698 (0.016369)
- RandomForest : 0.960834 (0.013924)
- ExtraTreeClassifier : 0.963917
- VotingClassifier:0.98

Other Algorithms like SMOTE or MLP could also be used for better results.