## Problem Identification :

It's a binary class classification problem because the label of the train set has only two unique and discrete values. We used classification techniques to get prediction y_pred on test set X_test and suppose to submit y_pred on kaggle in order to get a score.

**Given data set :** 1. Train.csv
2. Test.csv

## Pre-processing :

The data is preprocessed via standard scaling because applied models needed to interpret features of the dataset on the same scale. That's why we applied one of the feature scaling techniques. One more feature engineering technique was also applied that was Recursive feature elimination. It selected the best six features and removed the rest of them.

## Models and their best kaggle scores :

1. Model : Logistic Regression
   Score is 90%

2. Model : Decision Tree
   Score is 76.66

3. Model : Random Forest
   Score is 86.66%

4. Model : Gaussian Naive Bayes
   Score is 86.66

## Explanation :

Once we were done with pre-processing of the dataset then we moved to implement different models. Firstly, we applied logistic regression as problem demands, then went to Decision Tree, then Random forest classifier and then Gaussian Naive Bayes. And what we observed was that ? We observed logistic regression scored highest out of all models we applied. Decision tree failed because it consists of only one decision tree and in Random forest collection trees there and average out their results in order to obtain the final result. Gaussian bayes were also performed quite decently but not as logistic regression. Given a smaller dataset, meanwhile logistic regression had performed significantly better than all models. We also applied grid search in order to get the best parameters but it didn't significantly improve the score and scored the same as default models.