Project 1 Report:

Problem Description:
Given ‘traindata’, ‘trainlabel’ and ‘testdata’, the problem can be taken as a supervised learning problem. It is asked to class the samples in ‘testdata’ into two classes, labeled ‘1’ and ‘0’.

Data Preprocessing:
After having a look at the ‘traindata’, it is noticed that the attributes are numeric values in different scale. Instead to weaken the effect of difference between some statistical features, a simple preprocessing is applied, subtracting the mean from every value for a specific attribute.

Classifier Choosing:
As it is not clear that which classifier works well on the data, four usual models are chosen to do the job: logistic, decision tree, naïve bayes and random forest. Since we need to output the predicted labels for the ‘testdata’, we decided to choose only one observation to do the classifying job. In order to do that, a five folds cross validation is applied to all four classifiers. According that, the accuracies for models can be reached as follow:

Logistic: 0.92624223602484479,
Decision tree: 0.90722049689440998,
Naïve bayes: 0.84899068322981364,
Random forest: 0.87150621118012417

We can tell that logistic model outperforms the others, therefore, it would be used to predict the lables.

Label predicting:
Result is saved in ‘project1_20446377.csv’.