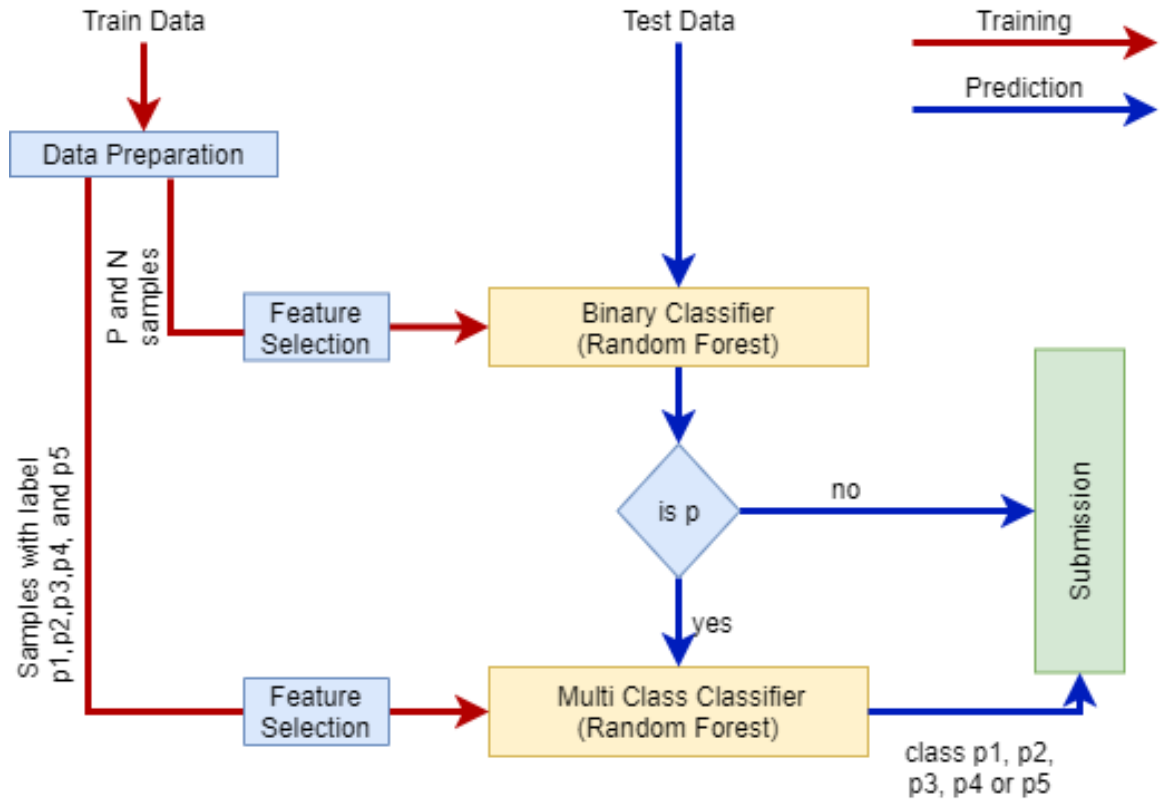
Assignment-9

Question-1 Report

Model Diagram



From the data exploration part it's found that data is unbalanced among different classes. B ut if we take all positive labels as a single class as positive, it will results two classes like nega tive and positive class, so data is balanced among these two classes. Also if we take all positi ve labels separately they are somehow balanced. So my approach is to use two classifiers as my model. The functions each classifier given below,

1. Binary Classifier:
   Classifies the samples in to positive or negative class. During the training two class are nearly balanced (4268 negative samples and 4239 positive samples) so gave good efficiency. If the output during prediction is n, the sample will be in class n and no further classification needed. If output is positive class, the sample is then gives to the multi class classifier for further classification.

2. Multi Class Classifier:
   Classifies the positive samples in to p1, p2, p3, p4, or p5 class. During the training five class are nearly balanced. This classifier, classify the samples only when the output of binary classifier is positive.

I have tried different classification methods like logistic regression, SVM, decision tree and random forest. Random forest gave good efficiency for test data so I have chosen random forest classifier for binary classification and multiclass classification.

Each step of the model given below:

Data Preparation:

In the data preparation step for binary classification, samples with label p1, p2, p3, p4, and p5 treated together as positive samples and it together with negative samples forms a dataset with two classes. Positive label encoded as 1 and negative label encoded as 0.

In multi class classifier data preparation retrieved the samples with labels p1, p2, p3, p4, and p5 only as a dataset and encoded p1=1, p2=2, p3=3, p4=4 and p5=5.

Feature Selection:

Feature selection of two classifiers done separately with decision tree classifier with 'feature_importance_' attribute. Decision tree is trained and fitted with default parameters. After training feature importance values retrieved. Feature importance greater than 0.01 selected as new features.

For binary classifier features 6, 11, 12,15,22,35 and 39 are found as important. For multi class classifier features 6, 8, 9, 10, 22, 23, and 24 are important.

Future selection with random forest classifier gave less efficiency than feature selection with decision tree classifier. Basically each models in the random forest classifier is a decision tree, so feature selection with decision tree classifier gave good efficiency.

Training and Hyper-parameters Tuning Results:

Binary classifier (Random forest classifier):

Efficiency:

Train data = 100% Test data = 99.76%

Hyper Parameters:

Criterion : entropy min_samples_leaf : 1

min_samples_split : 4 n_estimators : 80

Multi class classifier (Random forest classifier):

Efficiency:

Train data = 100% Test data = 99.45%

Hyper Parameters:

Criterion : entropy min_samples_leaf : 1

min_samples_split : 2 n_estimators : 80