PartA

1.Processing the data: read csv and transfer to dataframe

2.Run the data for 10 times and shuffle the data each time so that we can have different training sets and testing sets each time.

3.Generate two matrix: yes matrix and no matrix, which will be used for future analysis of conditional probability for naïve bayes model.

4.Get the value of p(yes) and p(no)

5.Get the value of p(f1 | yes) …. P(yes) and p(f1 |no)….P(no)

6.See which value has larger value, then choose the one that has the larger value as our predicted result.

7.Test our data: traverse all the test data and get the average result, run 10 times and get the final overall accuracy.

The overall accuracy is about 0.74

PartB:

1.Processing the data: read csv and transfer to dataframe

2.Prepare our data: First, find all the zero value in column 3, 4, 6, 8
Second, transfer the zero value to NA (R representation)
Third, get the mean value of the column of other data
Fourth, replace the NA value data in the column with the mean value of other values in the current column.

3.Run the data for 10 times and shuffle the data each time so that we can have different training sets and testing sets each time.

4.Generate two matrix: yes matrix and no matrix, which will be used for future analysis of conditional probability for naïve bayes model.

5.Get the value of p(yes) and p(no)

6.Get the value of p(f1 | yes) …. P(yes) and p(f1 |no)….P(no)

7.See which value has larger value, then choose the one that has the larger value as our predicted result.

8.Test our data: traverse all the test data and get the average result, run 10 times and get the final overall accuracy.

The overall accuracy is about 0.81

PartD:

Predict the label using python sklearn.svm library

The overall accuracy is about 0.64