Question 2

Process Followed and Plots:

1. Import required libraries like pandas, numpy, sklearn and their componants.
2. Get Training and Testing Data from the provided arff files i.e 'training_subsetD.arff' and 'testingD.arff'
3. Pre-processing the data

   3.1 Handling NULL and NaN values

   Since the data set contain many Null and None values, their handling is necessary in getting correct picture of the classification. Technique used for this process is

   1. If Null or NaN value count of any column is more tha 50% of total data points then the column is dropped.
   2. If Null or NaN value count is less than 50%, then the data points are estimated by putting the mode of data points of that specific column.

      Till this point the shape of training data frame is 40000*139 and testing data frame is 25000*139.

   3.2 Labelling

   Since the data available is in the categorical form, it is necessary to convert this into numerical labels. This is done with the help of LabelEncoder of sklearn's preprocessing.

4. Division of Training into two categories where all columns except last column is in one dataframe as 'X' and the last column as 'Y'.

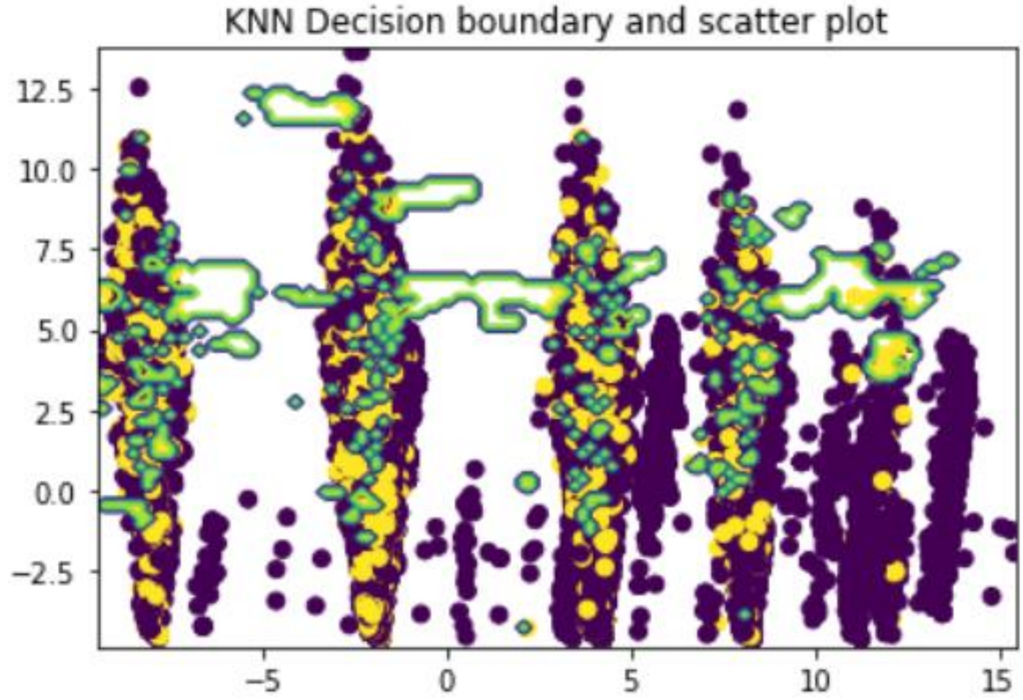
   ```
   X_df=df.values[:,0:138]
   Y_df=df.values[:,138]
   ```

5. Dimensionality Reduction:

   Since we have 139 dimensions in our cleaned data, it is not feasible to plot a graph out of it and decision boundary over it. So, for reducing the dimentionality, I have used Principal Component Analysis Technique.

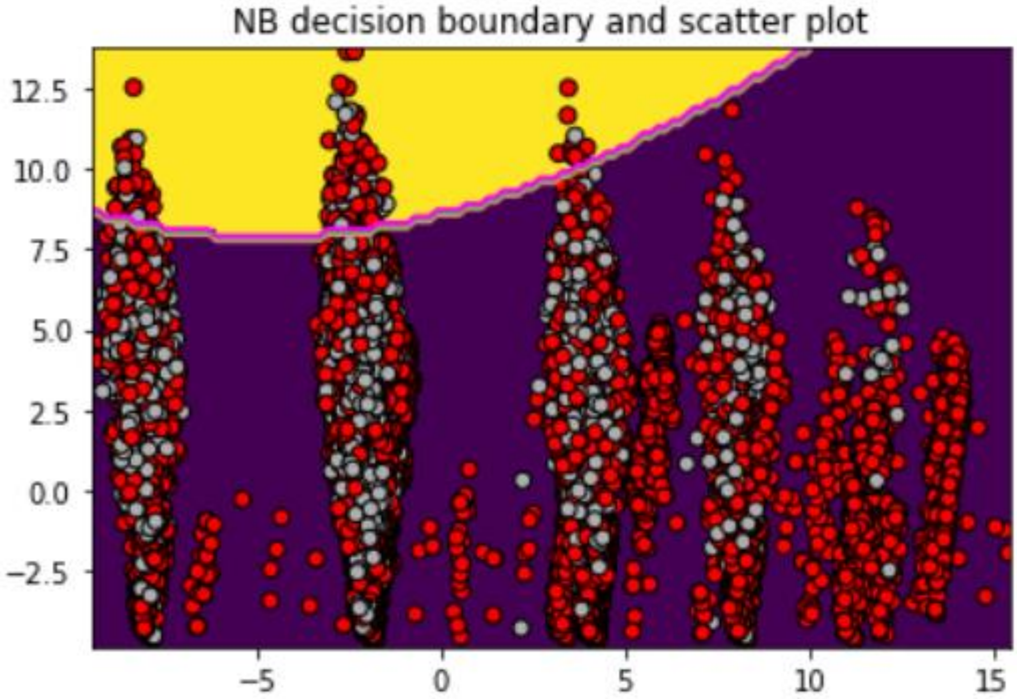
6. <u>KNN Classifier</u>

First, KNN classification is done and the result of the prediction is plot on the scatter plot with the decision boundary. K=10



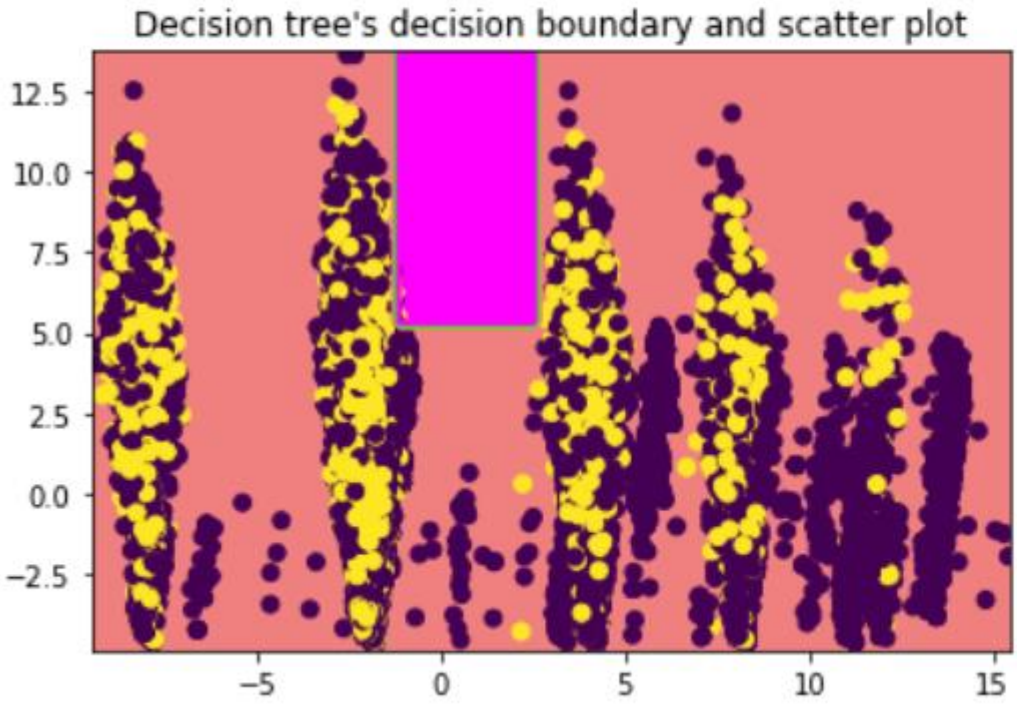
7. <u>Naïve Bayes Classifier</u>

Next, Bayesian classifier is applied and decision boundary plot is created.



8. Decision tree classifier

Decision tree classifier is taken from the Q1 for Depth =6 (optimal) and the plot is created.



Observations and Inferences:

1. Scatter plot of the data points in each of the classifier is showing a great extent of overlapping of data points with each unique values i.e True and False.
   Since, we have large number of data points.(40K)

2. Though I have tried with scaling the Data with RobustScaler of Sklearn but again the overlapping problem was not reduced.

3. When KNN classifier was applied, it is able to create decision boundary which is most clear out of the three classifier used.
   KNN is supervised lazy classifier which has local heuristics

4. NB Classifier decision boundary are clear to some extent.
   As, Naive Bayes is an eager learning classifier.

5. Decision tree Classifier with depth=6 is done but due to large no of overlapping data points, decision boundary is not clear on the plot.
   Also, overfitting is a factor.

6. After observing the accuracy of all these classifier, I think the Random Forest classifier which uses decision tree will give the best accuracy over this data.