MSIM 607: Machine Learning 1

Project 1 Report

Name: Muhammad Enayetur Rahman
UIN: 01216862

Team Members:
Muhammad Enayetur Rahman
Mahmudul Hasan

Part 1

<u>Task 1:</u>

Training dataset scatterplot for the ‘**generated**’ datasets are plotted and given here:

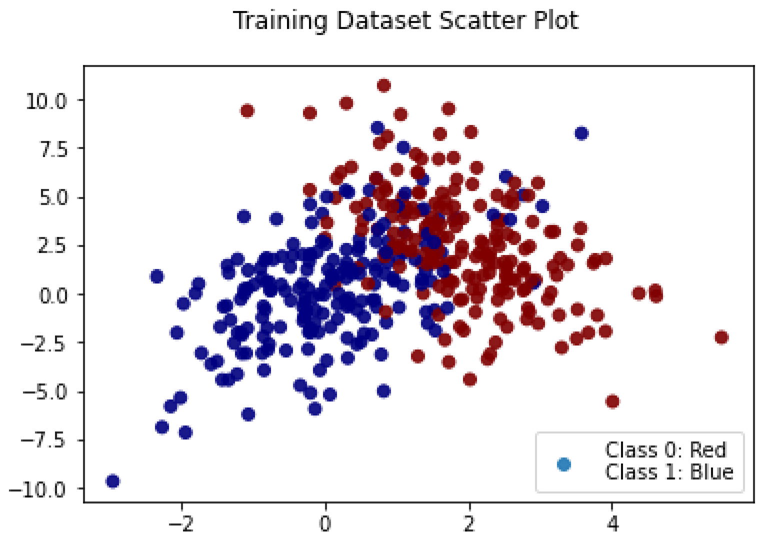


Fig: Training dataset for ‘**generated**’ dataset

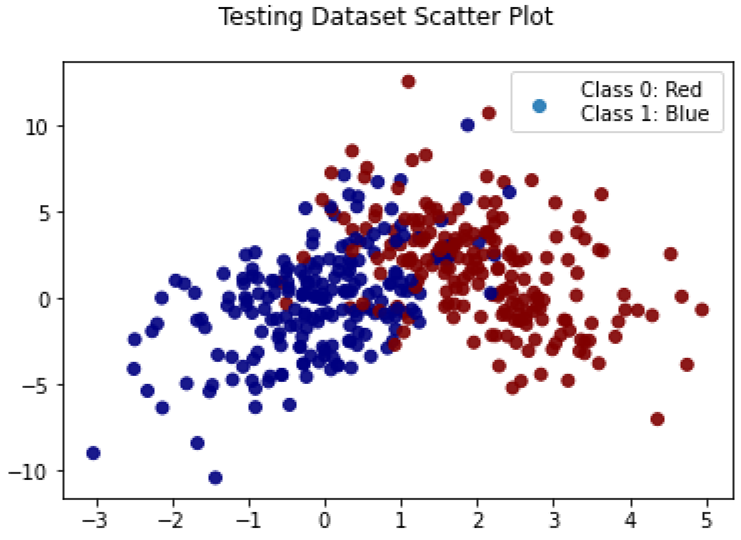


Fig: Test dataset for **‘generated’** dataset

For this task, the scatter plot is generated and class 0 represents Red circles and class 1 represents the blue circles.

Task 2:

For the designing of **Bayes classifier** we got the following results:

For **generated dataset:**

Data type	**Accuracy**
Training data	86%
Testing data	87.5%

For **zipcode dataset:**

Data type	**Accuracy**
Training data	91.3%
Testing data	88.37%

Discussion:

The Bayes classifier is applied to both of the datasets by assuming that the dataset follows *Gaussian distribution*. We can see, for the generated dataset testing accuracy is more than the training dataset, which is unusual.

Task 3:

For designing a ***Naive Bayes classifier*** we got the following results:

For **generated dataset:**

Data type	**Accuracy**
Training data	85%
Testing data	87.25%

For **zipcode dataset:**

Data type	**Accuracy**
Training data	89.367%
Testing data	87.03%

Discussion:
Similarly, we can see the testing data accuracy of ‘generated’ dataset is higher than training data accuracy.

Task 4:

For the Nonparametric estimation technique to estimate the conditional distribution *p(**x**|C$_i$)*, using a Gaussian kernel we got the following results:

For **‘generated’** dataset:

h value	**Accuracy (%)**
0.1	84
0.7	88
1.3	88.5
1.9	88.75

For **‘zipcode’** dataset:

h value	**Accuracy (%)**
0.11	86.83
0.12	86.97
0.13	87.0
0.14	87.1

Discussion:
We can see as the h values are increasing, the accuracy is also increasing for nonparametric estimation for both of the datasets.

Task 5:

For the designing of *k-nearest neighbor* classifier, we got the following results for different k values:

For ‘**generated’** dataset

K value	**Accuracy(%)**
3	86.25
5	87.75
9	88.0
13	87.75

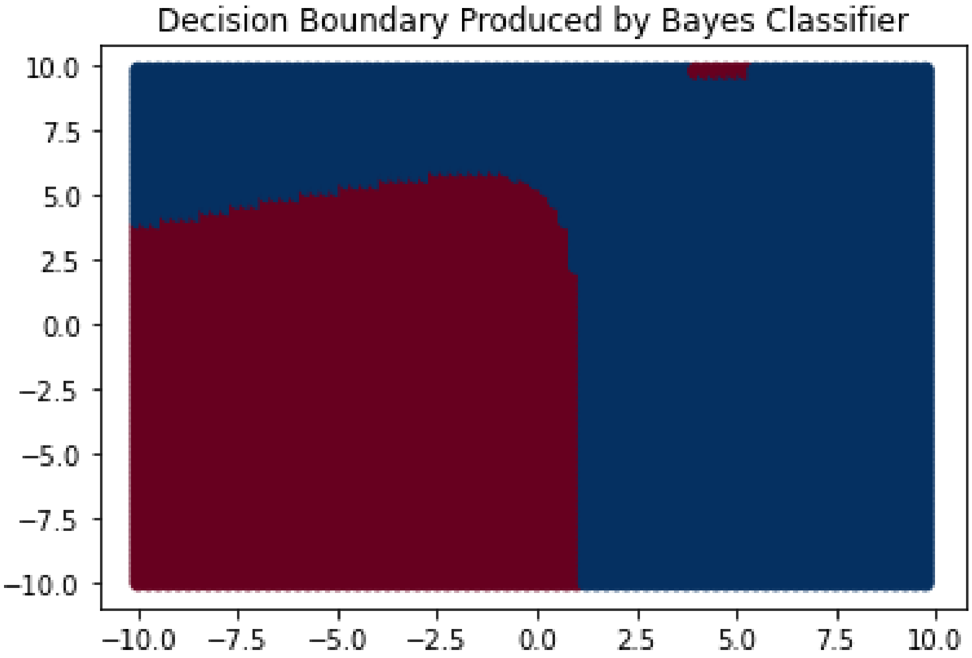
For ‘**zipcode’** dataset we got the following results:

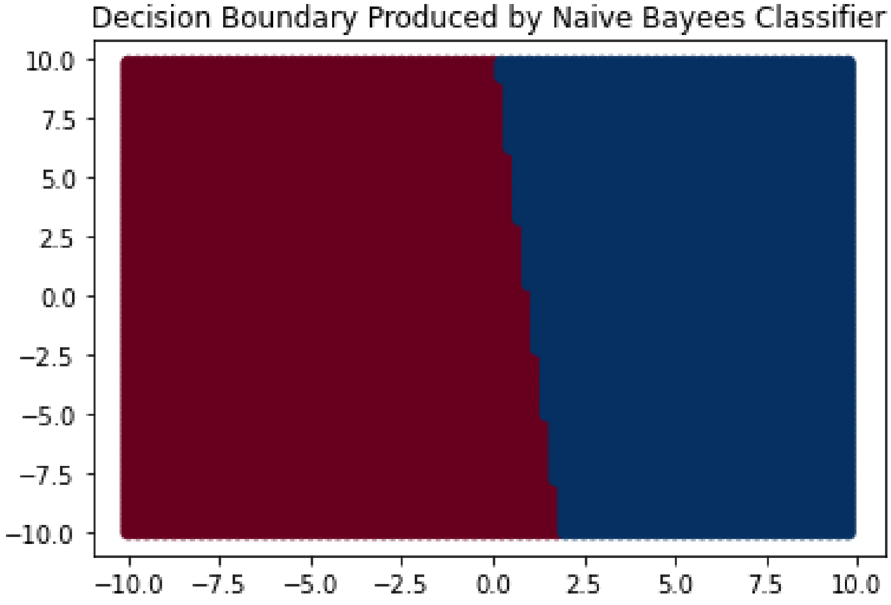
K value	**Accuracy(%)**
3	95.5
7	93.53
11	91.76

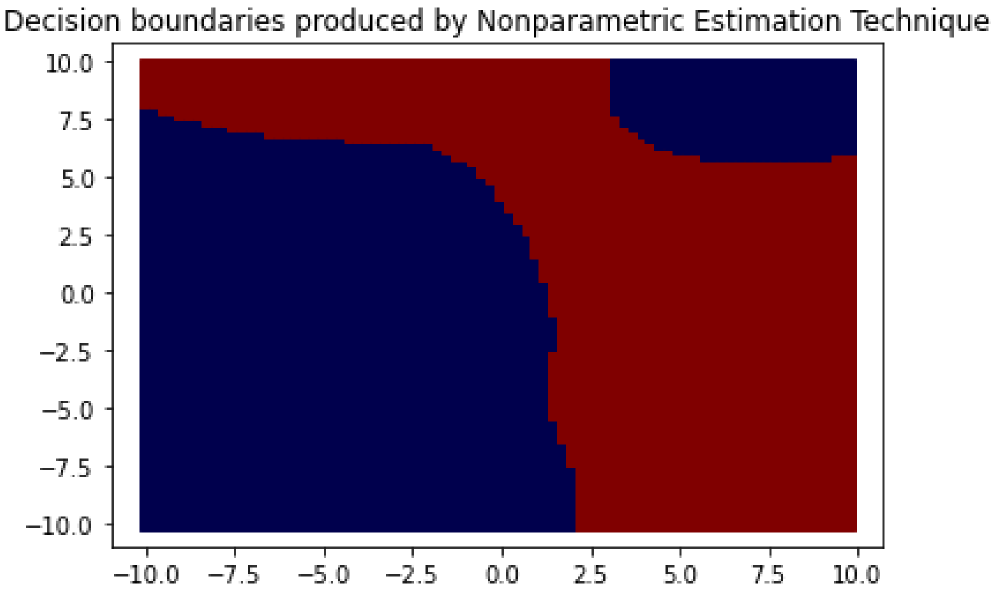
Discussion:
For KNN classifier, if K increases after a certain value, the accuracy is decreased thereafter.

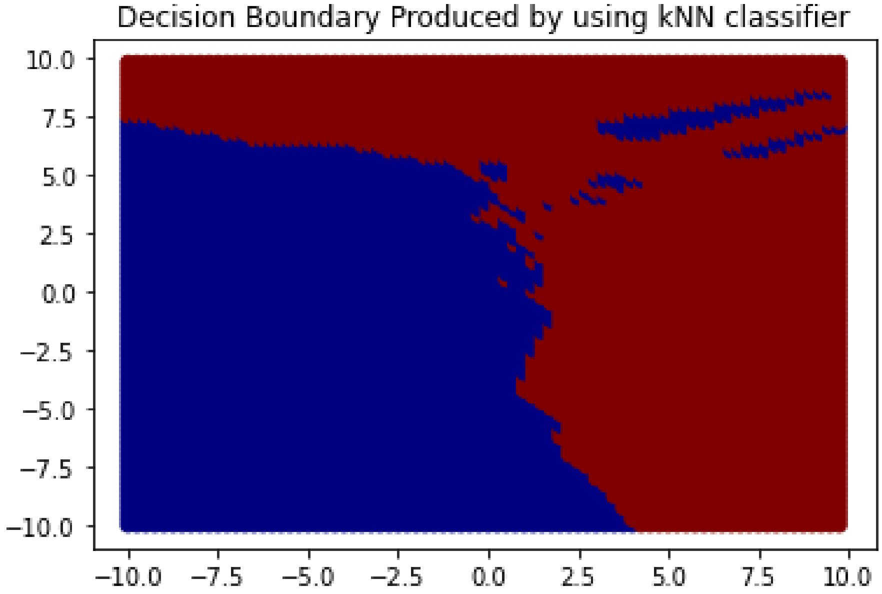
Task 6:

For the '**generated**' dataset we got the decision boundaries for the following classifiers:






Decision boundaries produced by Nonparametric Estimation Technique


Decision Boundary Produced by using kNN classifier

Part 2

Task 1

We designed a Regression model using Matlab and train the model using **‘regression_train.csv’** dataset. The model is applied to the testing: **‘regression_test.csv**’ dataset.

Training Errors:

0.0150
0.0002
0.2361
0.0001
0.0376
0.0363
0.0000

Testing Error:

0.0140
0.0002
0.2567
0.0001
0.0426
0.0359
0.0000

Task 2:

The modified linear regression model so that it can do linear classification on 'generated' datasets:

Training Accuracy	**Testing Accuracy**
85.75%	86.75%

Task 3:

For the six different regularization coefficients, the classification codes on the *zipcode* dataset is given below:

Regularization Co-efficient	**Training Accuracy (%)**	**Testing Accuracy (%)**
0.01	87.8333	85.7667
0.1	87.2333	85.3333
0.5	87.0	85.1667
1	86.7667	84.7333
5	82.5333	80.2667
10	81.57	79.50

We can see, if we increase the regularization co-efficient the training and testing accuracy is decreasing.