AML Final Project Report

**Members:**

Girish Gabra (ggabra)

Soumya Achar (soachar)

Data Set Description :

For this project we are using two datasets from UCI machine learning repository, Breast Cancer dataset and Credit card dataset.

I) Breast Cancer dataset

This breast cancer databases was obtained from the University of Wisconsin Hospitals. And it is collected for 3 years. This is a binary classification problem with 699 instances, 241 positive cases and 450 negative cases.

Url: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

**Attribute Information:**

1. Clump Thickness: 1 - 10

2. Uniformity of Cell Size: 1 - 10

3. Uniformity of Cell Shape: 1 - 10

4. Marginal Adhesion: 1 - 10

5. Single Epithelial Cell Size: 1 - 10

6. Bare Nuclei: 1 - 10

7. Bland Chromatin: 1 - 10

8. Normal Nucleoli: 1 - 10

9. Mitoses: 1 - 10

10. Class: (0 for benign, 1 for malignant)

II) Credit Card dataset

This dataset classifies people described by a set of attributes as good or bad credit risks. This is provided by Universit"at Hamburg and has 1000 instances with categorical data. This is also a binary classification problem.

Url : http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

**Attribute Information:**

There are 20 attributes in this dataset. They are as follows

Attribute 1: Status of existing checking account

A11 : ... < 0 DM

A12 : 0 <= ... < 200 DM

A13 : ... >= 200 DM / salary assignments for at least 1 year

A14 : no checking account

Attribute 2: (numerical) Duration in month

We have put it into bins of 6 months

Attribute 3: (qualitative) Credit history

A30 : no credits taken/ all credits paid back duly

A31 : all credits at this bank paid back duly

A32 : existing credits paid back duly till now

A33 : delay in paying off in the past

A34 : critical account/ other credits existing (not at this bank)

Attribute 4: (qualitative) Purpose

A40 : car (new)

A41 : car (used)

A42 : furniture/equipment

A43 : radio/television

A44 : domestic appliances

A45 : repairs

A46 : education

A47 : (vacation - does not exist?)

A48 : retraining

A49 : business

A410 : others

Attribute 5: (numerical) Credit amount

We have put it into bins and converted to discrete data.

Attribute 6: (qualitative) Savings account/bonds

A61 : ... < 100 DM

A62 : 100 <= ... < 500 DM

A63 : 500 <= ... < 1000 DM

A64 : .. >= 1000 DM

A65 : unknown/ no savings account

Attribute 7: (qualitative) Present employment since

A71 : unemployed

A72 : ... < 1 year

A73 : 1 <= ... < 4 years

A74 : 4 <= ... < 7 years

A75 : .. >= 7 years

Attribute 8: (numerical) Installment rate in percentage of disposable income

We have put it into bins and converted to discrete data.

Attribute 9: (qualitative) Personal status and sex

A91 : male : divorced/separated

A92 : female : divorced/separated/married

A93 : male : single

A94 : male : married/widowed

A95 : female : single

Attribute 10: (qualitative) Other debtors / guarantors

A101 : none

A102 : co-applicant

A103 : guarantor

Attribute 11: (numerical) Present residence since

We have put it into bins and converted to discrete data.

Attribute 12: (qualitative) Property

A121 : real estate

A122 : if not A121 : building society savings agreement/ life insurance

A123 : if not A121/A122 : car or other, not in attribute 6

A124 : unknown / no property

Attribute 13: (numerical) Age in years

We have put it into bins and converted to discrete data.

Attribute 14: (qualitative) Other installment plans

A141 : bank

A142 : stores

A143 : none

Attribute 15: (qualitative) Housing

A151 : rent

A152 : own

A153 : for free

Attribute 16: (numerical) Number of existing credits at this bank

We have put it into bins and converted to discrete data.

Attribute 17: (qualitative) Job

A171 : unemployed/ unskilled - non-resident

A172 : unskilled - resident

A173 : skilled employee / official

A174 : management/ self-employed/

highly qualified employee/ officer

Attribute 18: (numerical) Number of people being liable to provide maintenance for

We have put it into bins and converted to discrete data.

Attribute 19: (qualitative) Telephone

A191 : none

A192 : yes, registered under the customer's name

Attribute 20: (qualitative) foreign worker

A201 : yes

A202 : no

## Data Preprocessing:

We converted attributes into numerical categories so as to work with logistic regression. We converted the datasets into two: training and testing dataset.

## Performance results of algorithms:

This section includes confusion matrix, accuracy, TPR, FPR for the tests that we ran using our implementation of Naive Bayes, Logistic regression and Decision tree algorithm.

### **Naive Bayes:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Credit Card DataSet**  Confusion Matrix   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 44 | 30 | | Actual Yes | 23 | 153 |   Accuracy 78.8%  True Positive Rate: 87%  False Positive Rate: 40.5% | **Breast Cancer dataset**  Confusion Matrix   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 110 | 1 | | Actual Yes | 1 | 61 |   Accuracy 98.84 %  True Positive Rate: 98.38 %  False Positive Rate: 0.9 % |

**Logistic Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Credit Card DataSet**  Confusion Matrix   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 4 | 70 | | Actual Yes | 0 | 176 |   Accuracy 72.0%  True Positive Rate: 100%  False Positive Rate: 94.59% | **Breast Cancer dataset**  Confusion Matrix   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 109 | 2 | | Actual Yes | 1 | 61 |   Accuracy 98.27 %  True Positive Rate: 98.38 %  False Positive Rate: 1.8 % |

**Decision Trees**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Credit Card DataSet**  Confusion Matrix (Average)   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 36 | 38 | | Actual Yes | 31 | 145 |   Accuracy 72.4%  True Positive Rate: 82%  False Positive Rate: 51.35% | **Breast Cancer dataset**  Confusion Matrix (Average)   |  |  |  | | --- | --- | --- | |  | Predicted  No | Predicted  Yes | | Actual No | 107 | 4 | | Actual Yes | 7 | 55 |   Accuracy 93.63%  True Positive Rate: 88.70 %  False Positive Rate: 3.60 % |

Performance Analysis

**Naive Bayes Vs Logistic Regression Vs Decision Tree**

i) Credit Card Data Set:

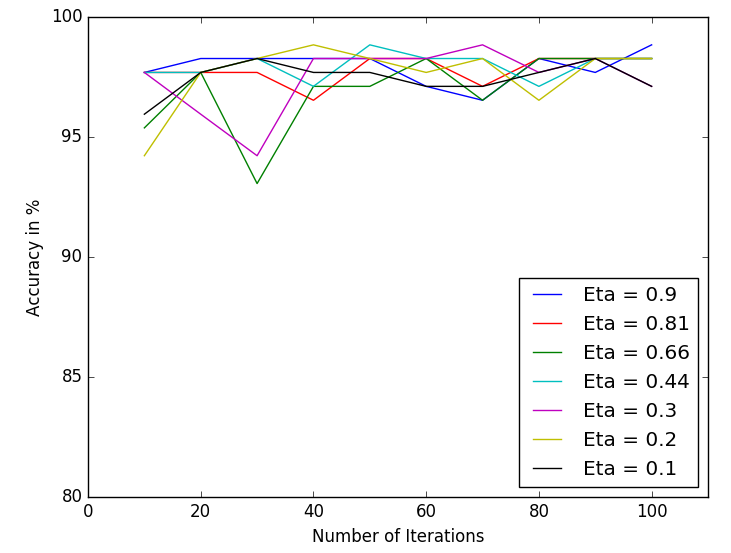
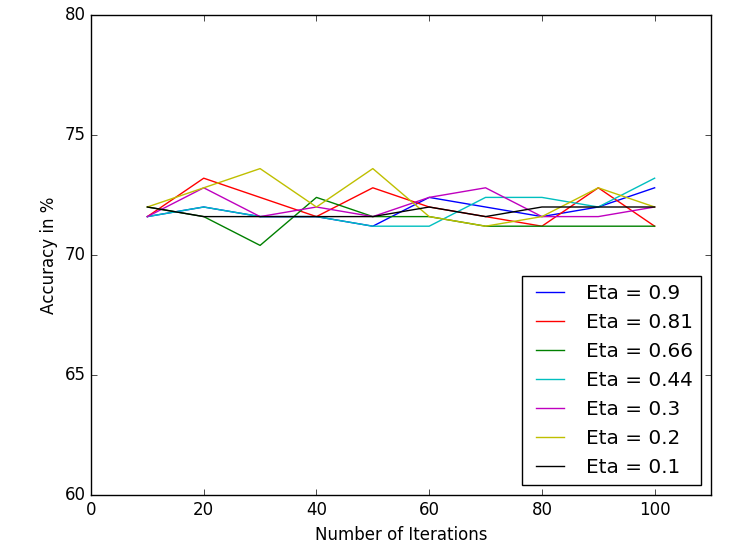
Data was preprocessed and divided into training and testing sets. We made sure that distribution of classes 0 and 1 are same in training and testing set.  
 Both Naive Bayes and Logistic Regression perform very well on this dataset as the dataset is linearly separable. Class label is binary (0 and 1) so both the classifiers perform good on the dataset. Decision Trees also perform good on this dataset as we have labelled training data on which model is built. All three algorithms have accuracy of 70%+ for the dataset.

Naive Bayes has the highest accuracy of 78.8% and is because if we observe features of the dataset, they are independent of each other. There is less correlation among the features of dataset. Naive Bayes makes the assumption of features being independent of each other, this being true in Credit data set gives us the highest accuracy.

ii) Breast Cancer Dataset

This is a binary class dataset so works well on all three algorithms. We have got high accuracy for both naive bayes and logistic regression (98%) however we feel since data is very small we cannot say that these algorithms will work good for every dataset of this type. Decision trees also perform good on this dataset. Overall it was a good fit for supervised learning algorithms that we have implemented

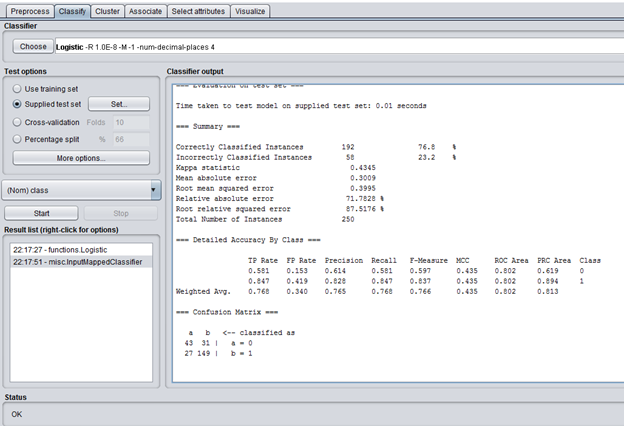
Following are two figures for logistic regression. We have taken different values of eta for iterations (10-100) and plotted obtained accuracies. For both our datasets accuracy is fairly consistent.

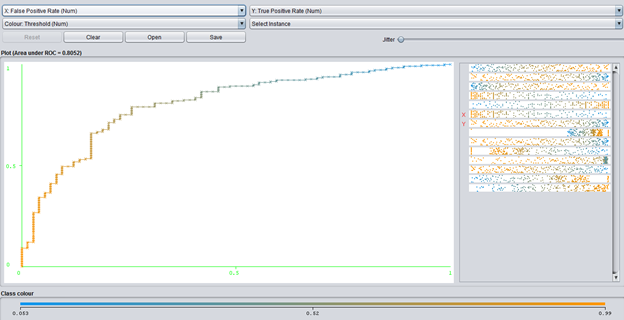
1. Credit card dataset b) Breast cancer dataset

Weka Outputs(Accuracy, Confusion Matrix and ROC Curves)

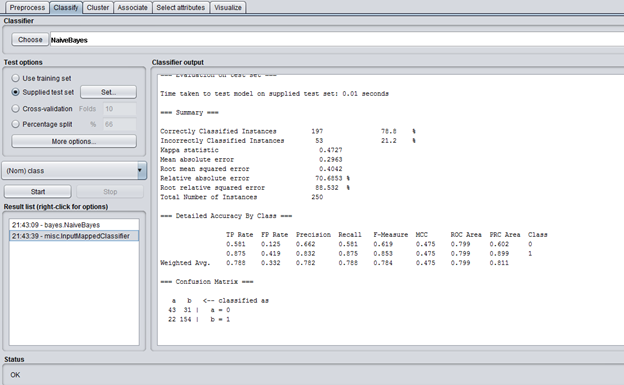
**I)** **German Credit Card Database**

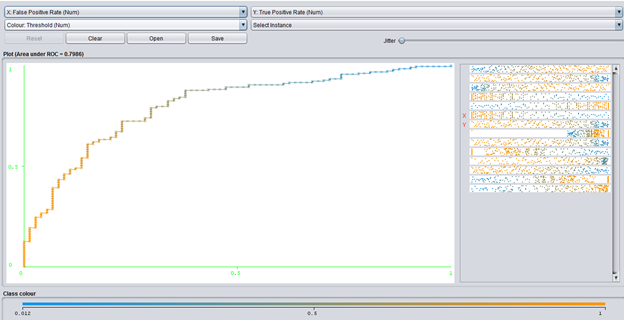
**i) Logistic Regression:**





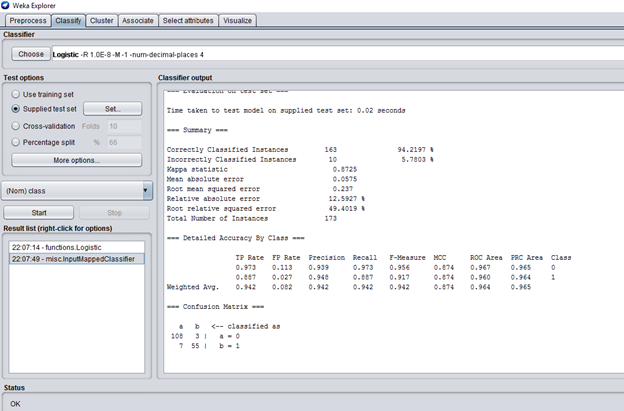
**ii) Naïve Bayes**

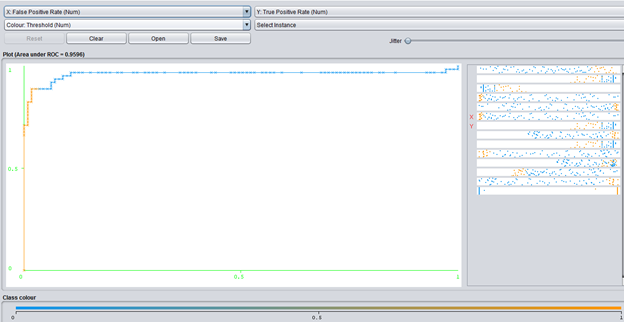




**II) Breast Cancer Dataset**

**i) Logistic Regression**





**ii) Naive Bayes**

