Indian Institute of Technology Kanpur

IME692A: Advanced Statistical Methods for Business Analytics

Course Project Report on

Classification Models

Date of Submission: Oct 9, 2020

Submitted By: Group 10
Aayush Ostwal- 170452 | Parth Pandey- 170462

Submitted to:
Dr. Shanker Prawesh,
Dept. of Industrial and Management Engineering, IIT KANPUR

CONTENT

Problem Statement

Use the datasets uploaded with this document to complete this assignment. There are two files: train.csv and test.csv. Use train.csv to train your prediction model and test.csv to only test the performance of the prediction model using the misclassification error rate. There are twenty predictors labeled $x_1,x_2,....,x_{20}$, and the dependent variable y has binary class labels (0 and 1). Your project report should contain the following information.

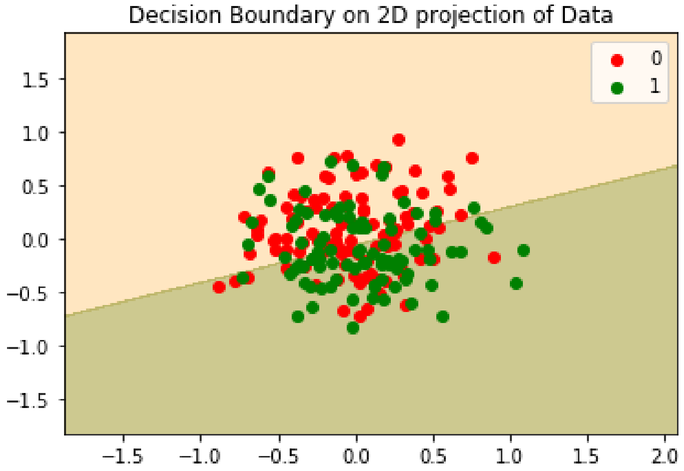
(a) Develop a classification model that has the lowest misclassification error rate on the test data. What is the training error rate for this model? Also, provide a brief explanation for the good performance of this model. (Hint: a good classification model should have a test misclassification error rate below 0.20).

Answer:

Logistic Regression with optimized threshold probability gave the best results on train and test data set. This accuracy would have been more if we had more data points. The threshold Probability is 0.62 for which accuracy is maximum.

Why is Logistic Regression performing better?

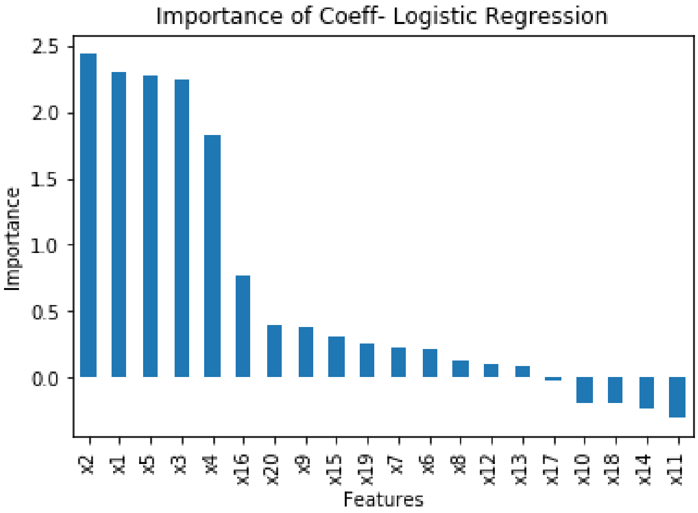
1. **Data is Linearly Separable**: When we project the data with 19 features into 2D plane using Principal Component Analysis, then data points are linearly separable upto some extent.



A clear boundary can be seen which separates the data into two classes upto a great extent.

2. Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets. As our data only contains 19 dimensions, chances of **overfitting the data is very less**.

3. It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).

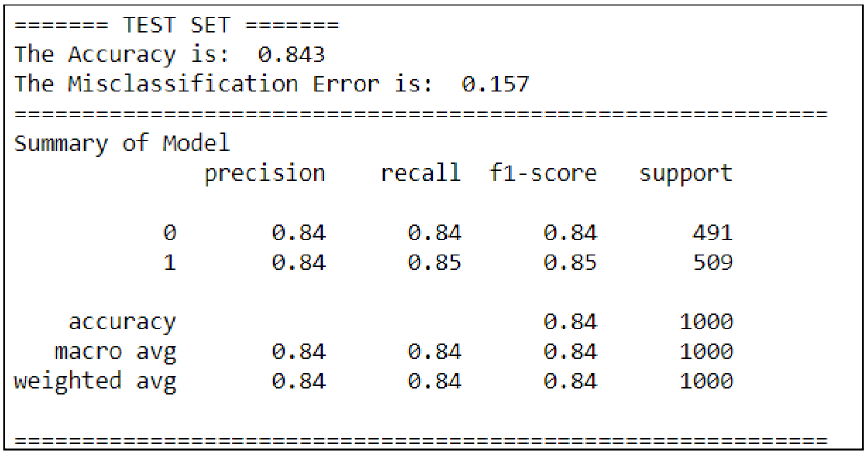


Feature x2 is the utmost important feature and x11 is the least important feature.

4. The data contains equal frequency of classes, hence its a **balanced set**. This is a plus for logistic regression as it is not used when we have an unbalanced dataset.

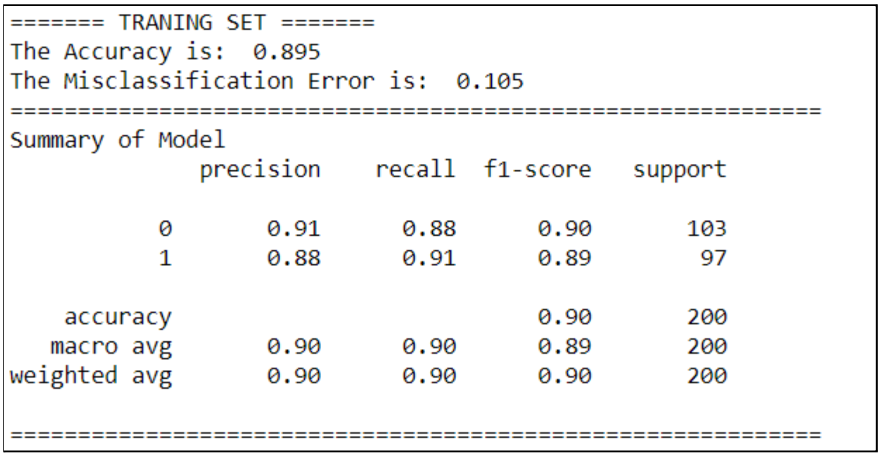
Error in the model:

Test Data set



```
======= TEST SET =======
The Accuracy is:  0.843
The Misclassification Error is:  0.157
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.84      0.84      0.84       491
           1       0.84      0.85      0.85       509

    accuracy                           0.84      1000
   macro avg       0.84      0.84      0.84      1000
weighted avg       0.84      0.84      0.84      1000

============================================================
```

Training Data Set



```
======= TRANING SET =======
The Accuracy is:  0.895
The Misclassification Error is:  0.105
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.91      0.88      0.90       103
           1       0.88      0.91      0.89        97

    accuracy                           0.90       200
   macro avg       0.90      0.90      0.89       200
weighted avg       0.90      0.90      0.90       200

============================================================
```

(b) Document all models you experimented with to evaluate the performance of the test data. The code used for these models should be uploaded along with your submission. The report should include a brief description of all classification models you have used to complete the assignment. If you are using a classification model that is not covered in the course, they provide an appropriate reference for this model.

Answer:

We have trained a number of models and their description is given below:

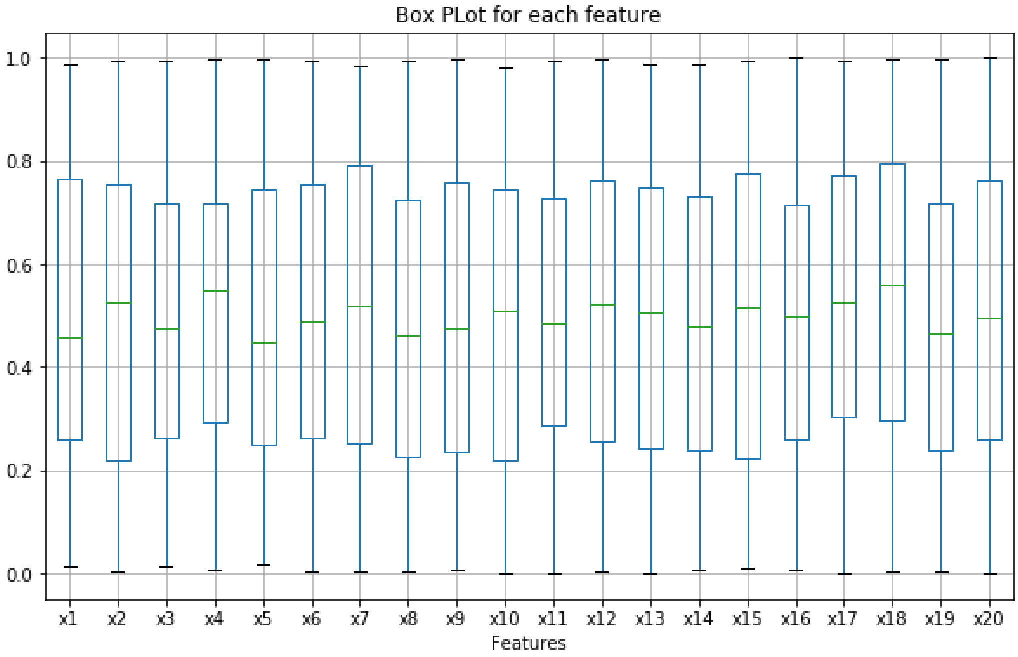
Model Description

(a) Principal Component Analysis:

PCA is used for dimensionality reduction, and data visualization. PCA orthogonally projects the data onto a lower dimensional linear space, such that the variance of the projected data is maximized.

Though it is not a model, it is very important to apply Principal component Analysis to understand how we can reduce the dimension of Data.

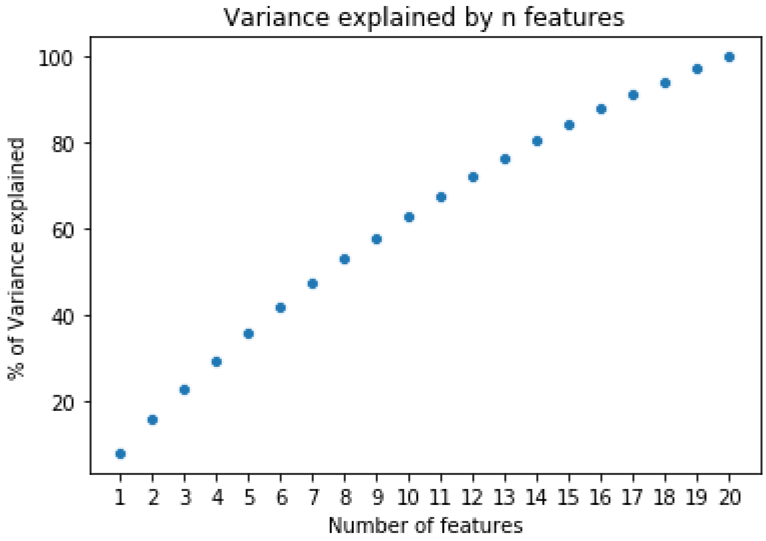
Box plot of all the features in data:



Observation:

a) There are no outliers in the data
b) Values of all features lie between 0 and 1.
c) Distribution of all the features are quite similar.

Variance explained by variables when projected to ‘n’ dimension space:



Observation:

a) It can be seen that the first principal component is responsible for 8.06% variance. Similarly, the second principal component causes 7.44%
b) There is no features set that explains a good variance as the correlation between the features are very less.

Summary:

Hence PCA cannot be implemented here as the correlation between features are less and when the data is projected to lower spaces, variance explained is very low.

Reference:

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://en.wikipedia.org/wiki/Logistic_function

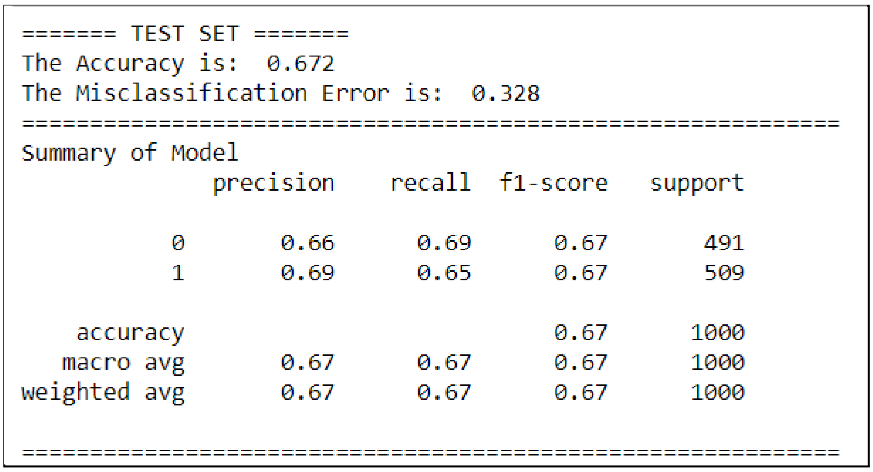
(b) Decision Tree classifier:

Description:

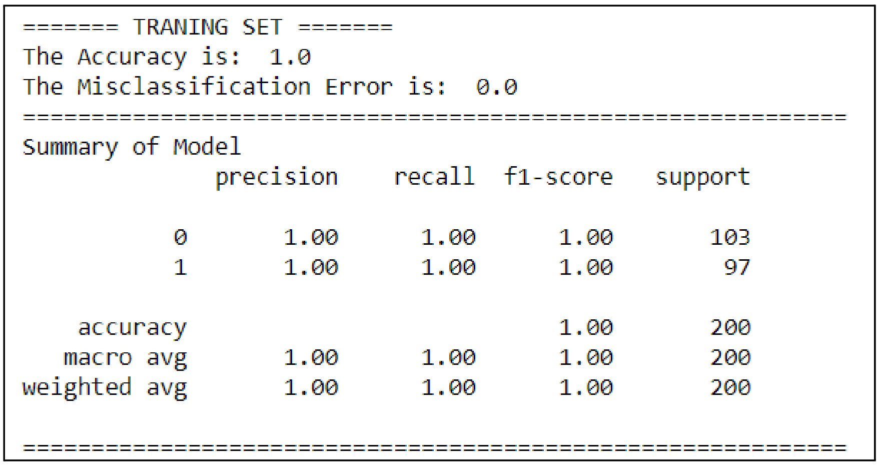
Decision Trees are a non-parametric supervised learning method used for classification and regression. Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

It splits that data based on given input features and those features are selected by the information gain on splitting the data based on a particular variable. More the information gained, the more useful and important the feature is.

Model Performance:

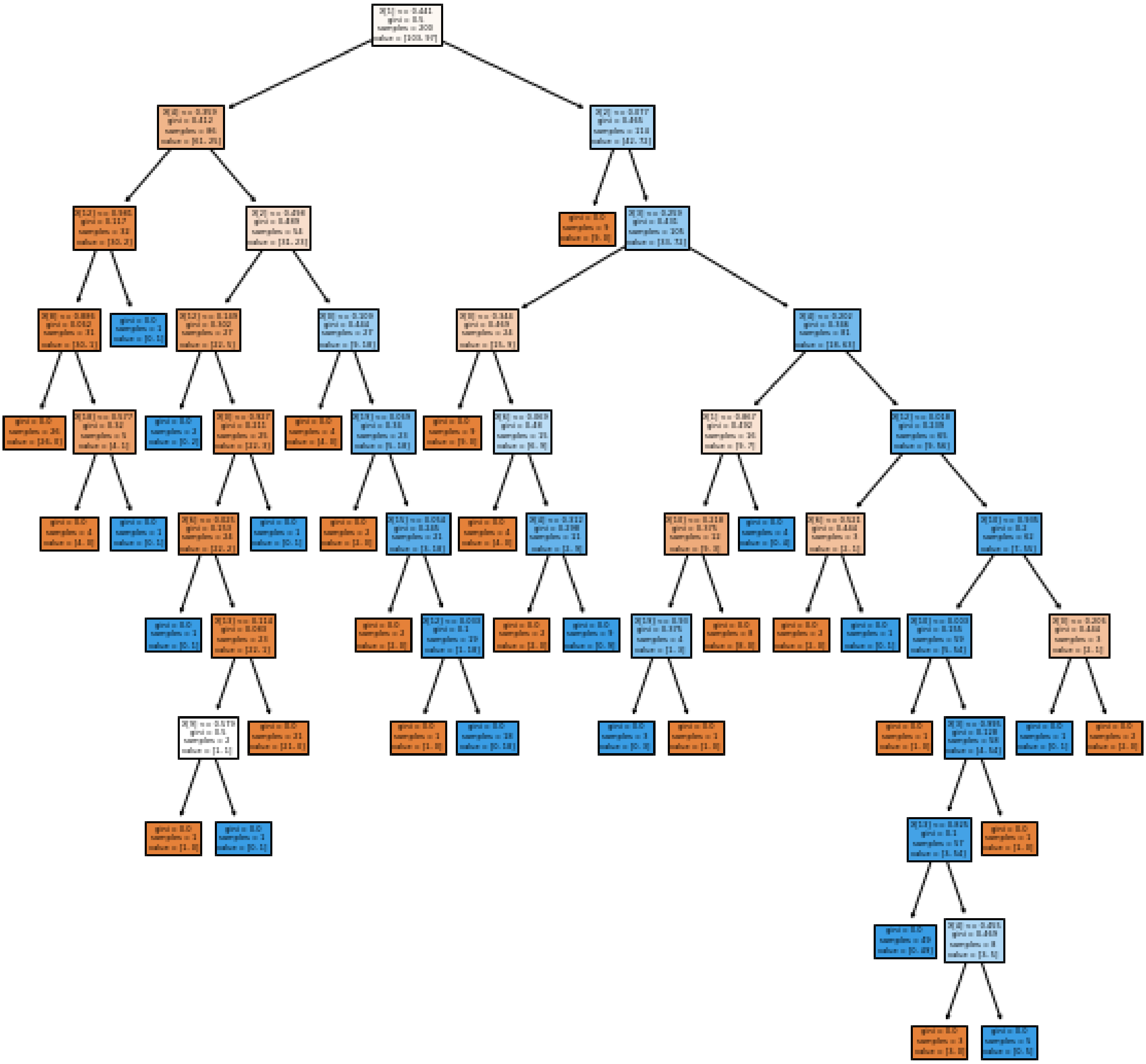


```
======= TEST SET =======
The Accuracy is:  0.672
The Misclassification Error is:  0.328
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.66      0.69      0.67       491
           1       0.69      0.65      0.67       509

    accuracy                           0.67      1000
   macro avg       0.67      0.67      0.67      1000
weighted avg       0.67      0.67      0.67      1000

============================================================
```

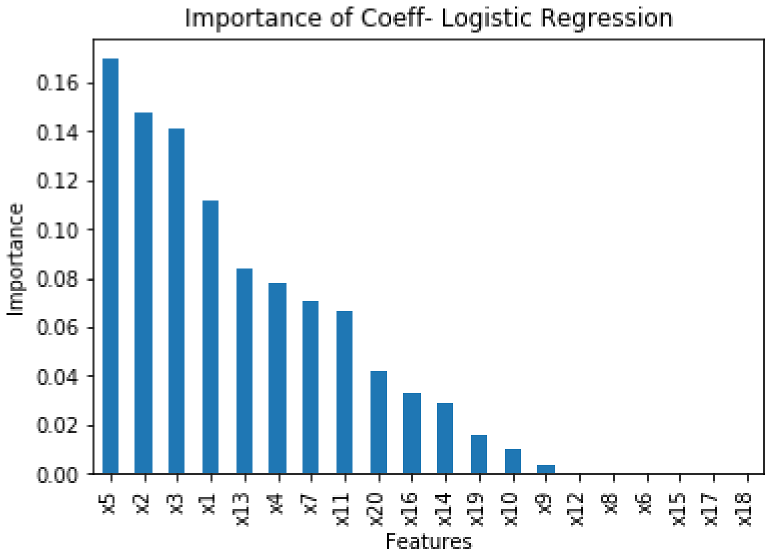


```
======= TRANING SET =======
The Accuracy is:  1.0
The Misclassification Error is:  0.0
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       103
           1       1.00      1.00      1.00        97

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

============================================================
```

Tree:



Importance of Features according to Decision trees are given by:



Summary:

Decision Trees **do not perform well** when the data set is very small. So in this case decision trees are not very good options to use as the data set contains **only 200 points.**

Reference:

https://scikit-learn.org/stable/modules/tree.html

https://en.wikipedia.org/wiki/Decision_tree

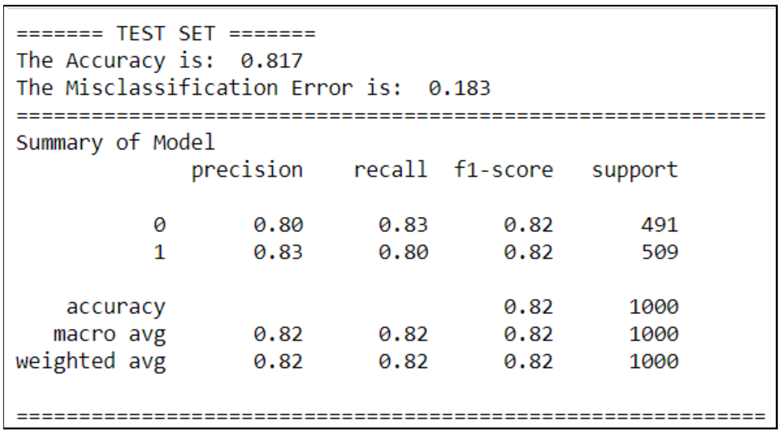
(c) Support Vector Machines:

Description:

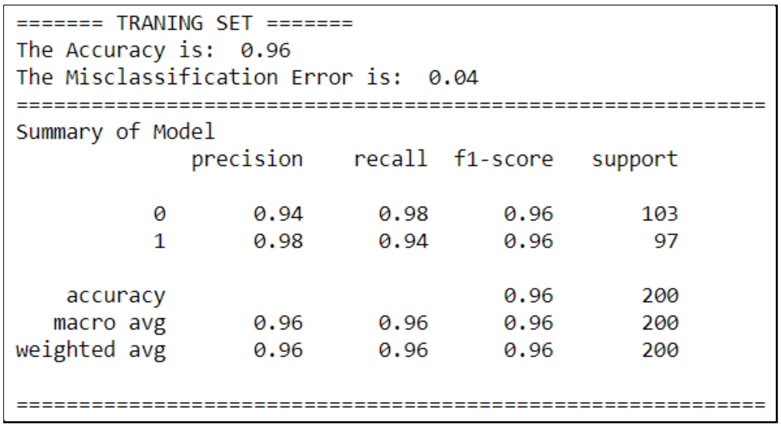
Support vector machines are particular linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance.

The SVM accomplishes the classification task by constructing, in a higher dimensional space, the hyperplane that optimally separates the data into two categories.

Model Performance



```
======= TEST SET =======
The Accuracy is:  0.817
The Misclassification Error is:  0.183
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.80      0.83      0.82       491
           1       0.83      0.80      0.82       509

    accuracy                           0.82      1000
   macro avg       0.82      0.82      0.82      1000
weighted avg       0.82      0.82      0.82      1000

============================================================
```

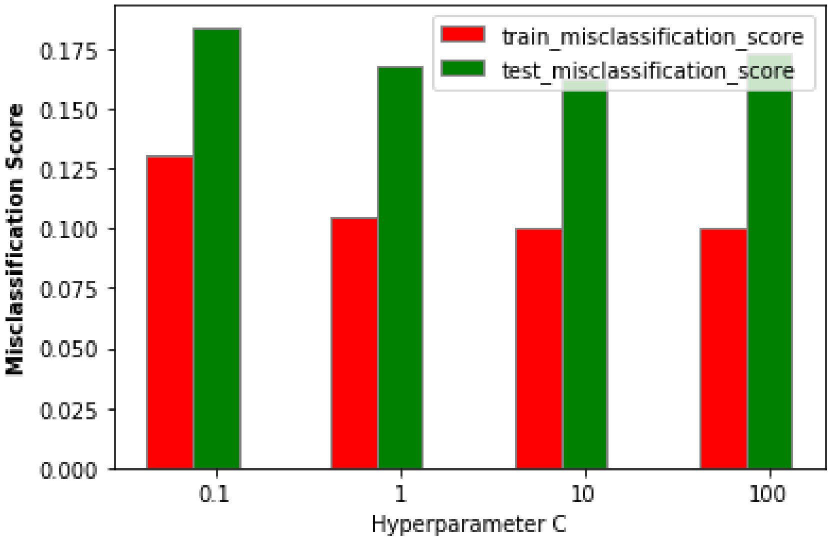


```
======= TRANING SET =======
The Accuracy is:  0.96
The Misclassification Error is:  0.04
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       103
           1       0.98      0.94      0.96        97

    accuracy                           0.96       200
   macro avg       0.96      0.96      0.96       200
weighted avg       0.96      0.96      0.96       200

============================================================
```

Mode Optimization

We can optimize the accuracy of the model with optimizing the ‘C’ values.
C is the penalty parameter of the error term. It controls the trade off between smooth decision boundaries and classifying the training points correctly.
It is a regularization parameter. The strength of the regularization is inversely proportional to C. It must be strictly positive. Its default value is 1.

Error vs Values of C:



Summary:

SVM is a good model and its accuracy is considerably higher than 80%. Hence this can be one of the choices for the final model. But Logistic regression (explained later) has outperformed Support vector machines.
References:

https://en.wikipedia.org/wiki/Support_vector_machine

https://scikit-learn.org/stable/modules/svm.html#svm-classification

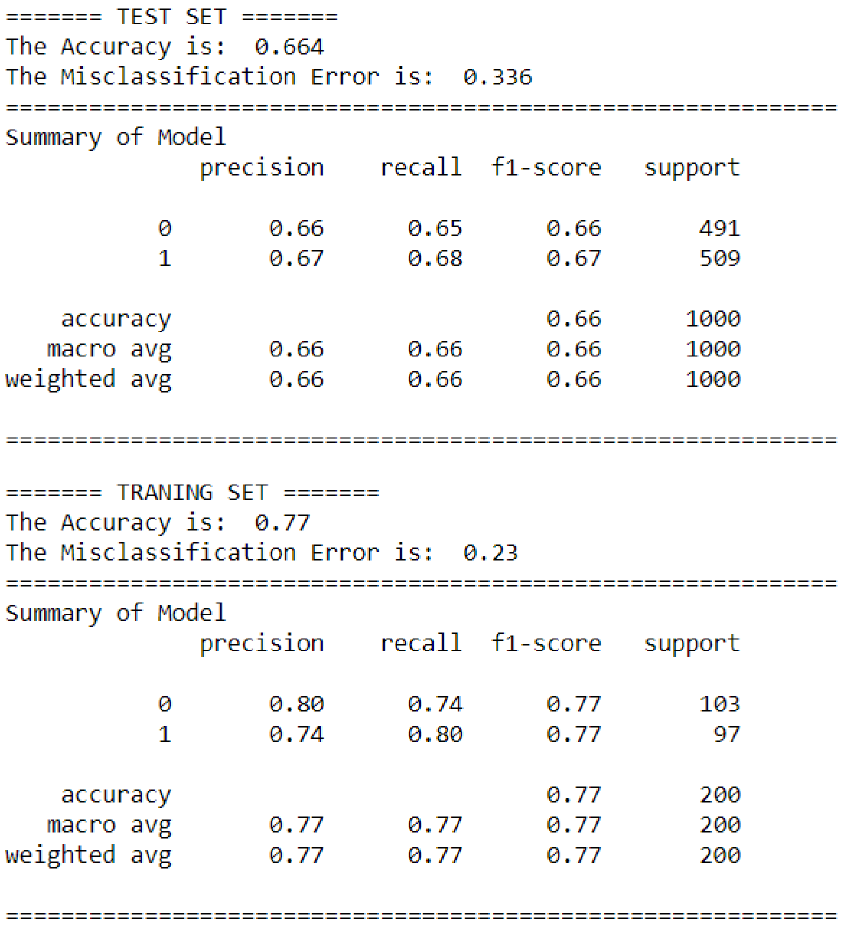
(d) K-Nearest neighbors:

Description:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

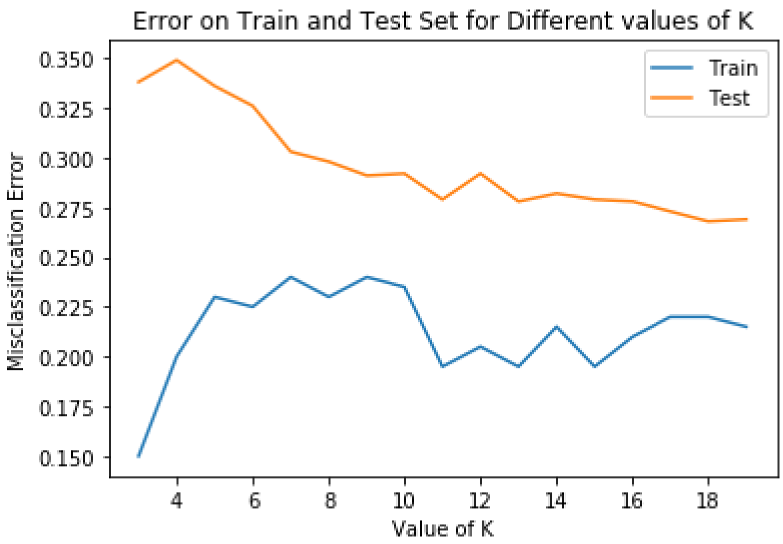
Model Performance:



```
======= TEST SET =======
The Accuracy is:  0.664
The Misclassification Error is:  0.336
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.66      0.65      0.66       491
           1       0.67      0.68      0.67       509

    accuracy                           0.66      1000
   macro avg       0.66      0.66      0.66      1000
weighted avg       0.66      0.66      0.66      1000

============================================================

======= TRANING SET =======
The Accuracy is:  0.77
The Misclassification Error is:  0.23
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.80      0.74      0.77       103
           1       0.74      0.80      0.77        97

    accuracy                           0.77       200
   macro avg       0.77      0.77      0.77       200
weighted avg       0.77      0.77      0.77       200

============================================================
```

Model Optimization:

We can optimize the model based on the "K" value i.e. the number of neighbours used for classification,



As we can observe that there is no significant accuracy any for values of K.

Summary:

In the section of PCA, we can clearly see that data points when projected to 2D space are randomly distributed hence this is the reason why KNN is not performing upto the mark.

Though earlier we have stated that the data is linearly separable, upto some extent, KNN underperforms as we have some data points which lie in opposite decision boundaries.

Reference:

https://scikit-learn.org/stable/modules/neighbors.html#classification

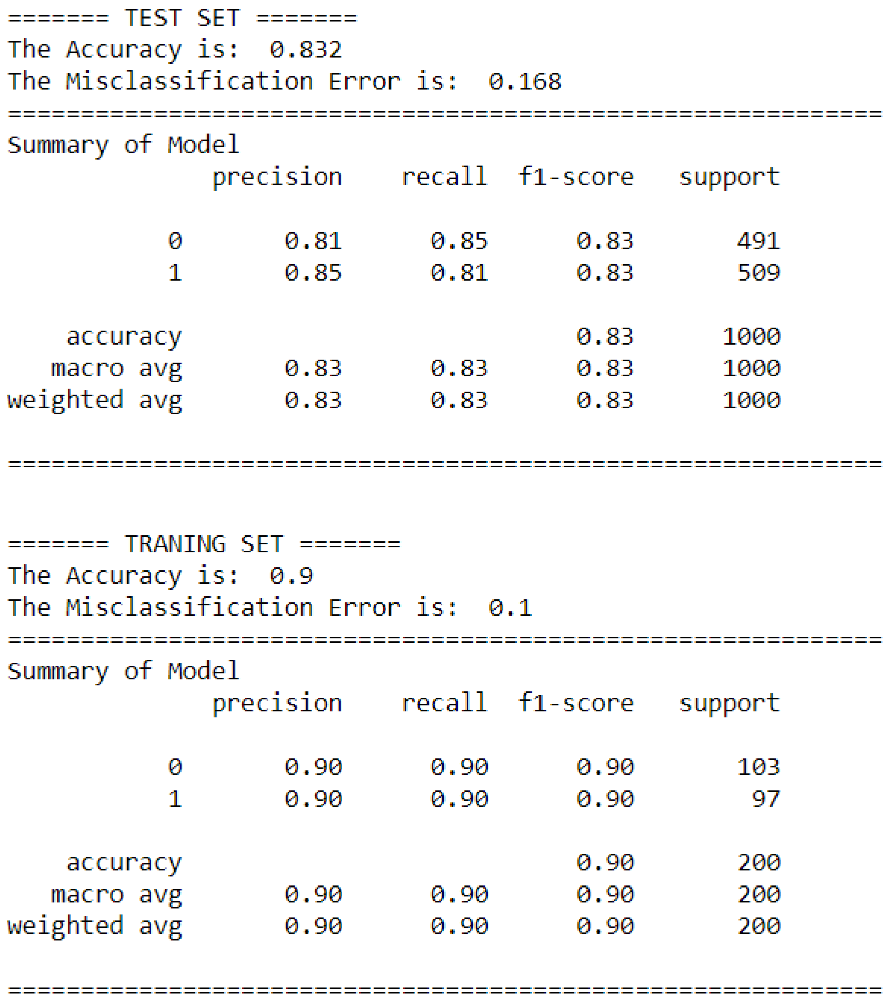
(e) Logistic regression

Description:

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes.

Model Performance:

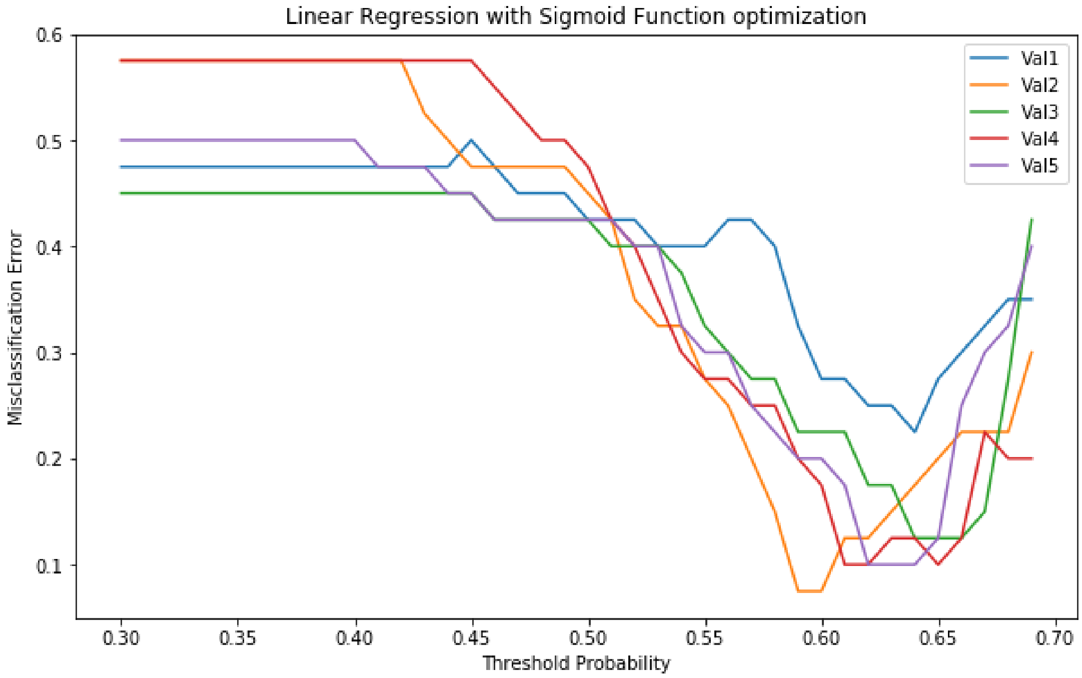


```
======= TEST SET =======
The Accuracy is:  0.832
The Misclassification Error is:  0.168
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.81      0.85      0.83       491
           1       0.85      0.81      0.83       509

    accuracy                           0.83      1000
   macro avg       0.83      0.83      0.83      1000
weighted avg       0.83      0.83      0.83      1000

============================================================


======= TRANING SET =======
The Accuracy is:  0.9
The Misclassification Error is:  0.1
============================================================
Summary of Model
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       103
           1       0.90      0.90      0.90        97

    accuracy                           0.90       200
   macro avg       0.90      0.90      0.90       200
weighted avg       0.90      0.90      0.90       200

============================================================
```

Model Optimization:

We can optimize logistic regression by optimizing the value of threshold probability.

Plot for error on validation data set for a number of threshold probabilities. We use 5 fold cross validation:



Observation:

a) The mean error is minimum for threshold probability = 0.62.

Summary:

Logistic Regression for threshold probability = 0.62 is the model that outperforms all the models. For this probability the error on both test and train set is minimum.

Reference:

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://en.wikipedia.org/wiki/Logistic_function

Model Summary

Model	Misclassification error rate on Training Data	Misclassification error rate on Test Data
Logistic Regression (threshold probability = 0.62 – optimized)	0.105	0.157
Logistic Regression (threshold probability = Not optimized)	0.100	0.168
Decision Tree Classifier	0.00	0.309
SVM	0.40	0.183
KNN	0.230	0.336

Code for Models can be find on:

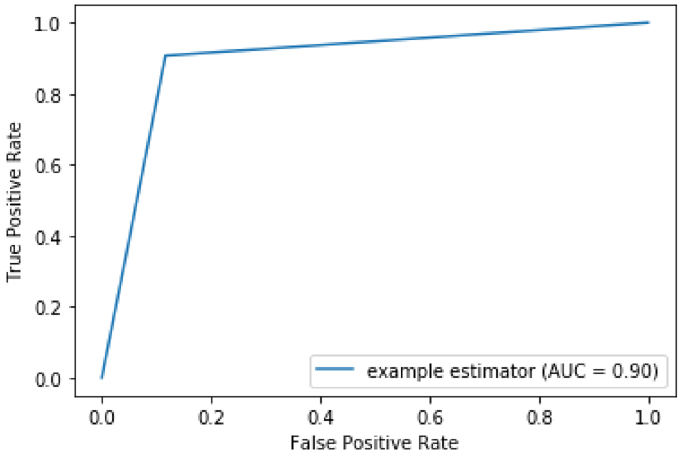
Code for Models: Group 10

(c) Generate the ROC curve for the best performing classification model. How would you interpret the findings from the ROC curve?

Answer:

Receiver Operating Characteristic(ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity (or TPR) and specificity (1 – FPR)

The Roc curve for best performing model is:

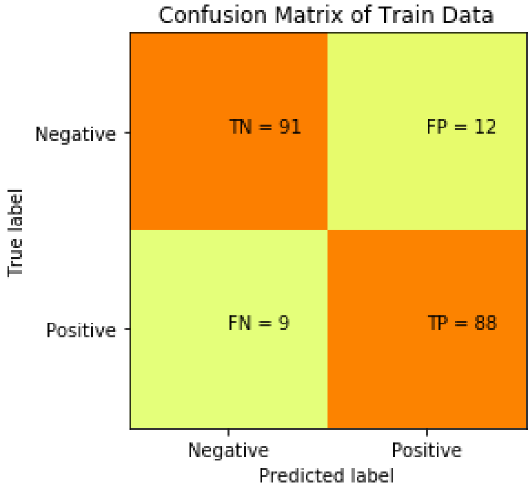


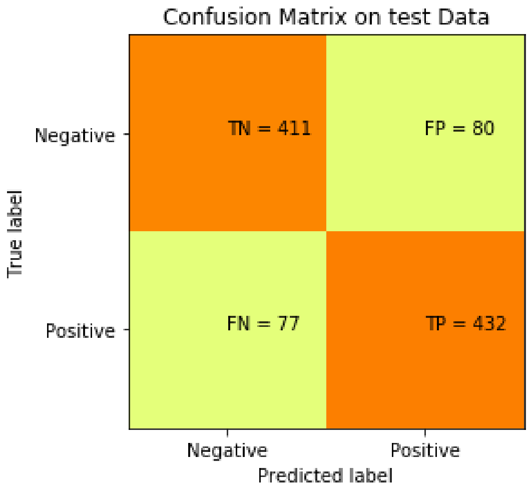
The ROC curve is approaching the ideal scenario where the curves of true positive and true negative do not overlap. In this case the AUC is 1.

Since the AUC score is not exactly 1, there is some overlap between the curve but the model **separates** the two classes with an accuracy of 90%.

As the model distinguishes the two classes, we can infer that this model will predict the novel data with good accuracy.

The confusion matrix on the training data and the test data for given model is given by:





The model classifies major data points as true labels.

(d) Briefly mention the contribution of each team member in the project.

NAME	CONTRIBUTION
Aayush Ostwal (Roll No. 170452)	1. Optimized the Logistic regression 2. Analyzed ROC Curve and confusion matrix 3. Plotted the decision boundary 4. Analyzed each feature in data set 5. Trained Decision trees and Logistic regression 6. Applied PCA for data reduction
Parth Pandey (Roll No. 170462)	1. Cross validation on final model 2. Trained support vector machines, K Nearest Neighbours and calculated misclassification rate. 3. Edited the final Code 4. Explained and documented the trained model 5. Compiled the code 6. Explained and documented the trained model