PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE

DATA MINING AND WAREHOUSING MINI-PROJECT REPORT
ON

"PREDICTION ON BANKING DATASET USING VARIOUS MODELS"

SUBMITTED BY

Omkar Amilkanthwar	41403
Aniruddha Deshmukh	41405
Atharva Satpute	41409

Under the guidance of
Prof. K. C. Waghmare

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2021-22

Contents

1 Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

2 Abstract

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered)class labels. For example, we can build a classification model to categorize whether client subscribed a term deposit from client data. Such analysis can help provide us with a better understanding of the data at large. In this project we use multiple classification models to analyse the outcome of Banking Dataset to predict whether client subscribed to term deposit or not. Apply suitable data pre-processing steps. We then compare performance of classification models to find which one is the best

3 Hardware and Software Requirements

3.1 Hardware Requirements

1. 500 GB HDD
2. 8 GB RAM
3. Monitor
4. Keyboard

3.2 Software Requirements

1. 64 bit Open Source Operating System like Ubuntu 20.04
2. Python 3
3. Google Colab
4. Libararies like sklearn, pandas, matplotlib, seaborn, numpy

4 Introduction

We have been provided with the data of clients such as age, education, job, etc. The Data fields are

1. age – Age of a person
2. job – Type of job
3. marital_end – Marital status
4. education – Education degree
5. default – Has credit in default?
6. housing – Has housing loan?
7. loan_id – Has personal loan?
8. contact – Contact communication type
9. month – Last contact month of year
10. day_of_weeek – Last contact day of the week
11. duration – Last contact duration, in seconds
12. campaign – Number of contacts performed during this campaign and for this client
13. pdays – Number of days that passed by after the client was last contacted from a previous campaign
14. previous – Number of contacts performed before this campaign and for this client
15. poutcome – Outcome of the previous marketing campaign
16. y – Has the client subscribed a term deposit('yes', 'no')?

The train set contains 32950 records while the test set is made with 20% split. We drop the date column from our analysis.

5 Objective

- To understand data preprocessing
- To perform classification on dataset and predict labels for test dataset.

6 Scope

We select dataset of Term Deposit (Banking). We try to apply many models and compare which one is the best model amongst them.

7 System Architecture

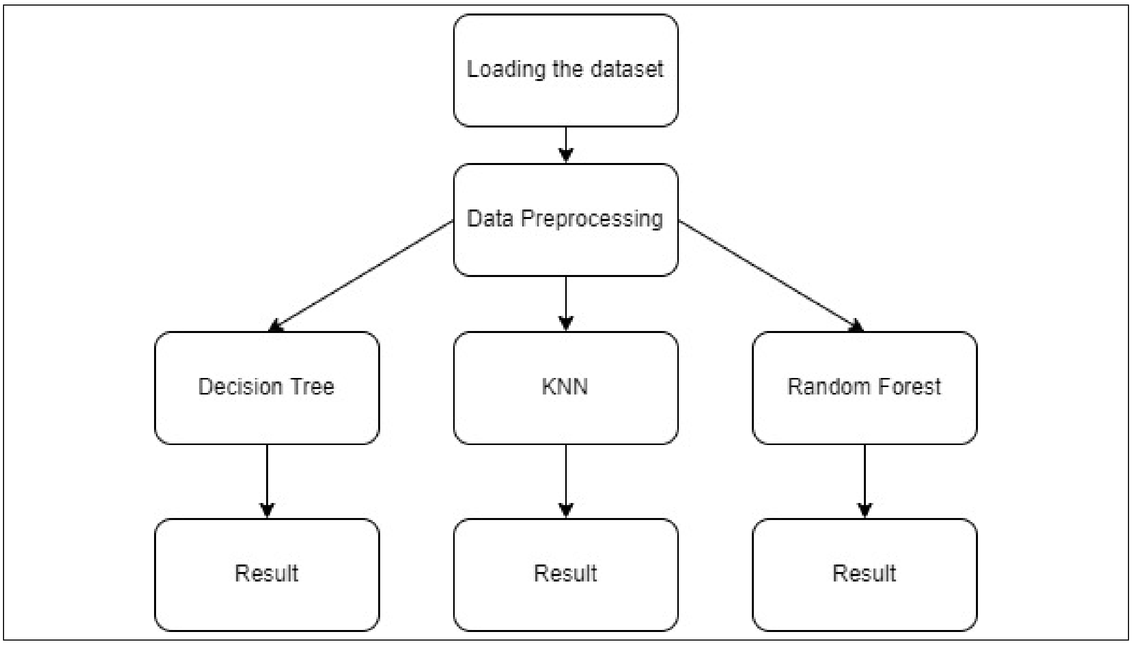
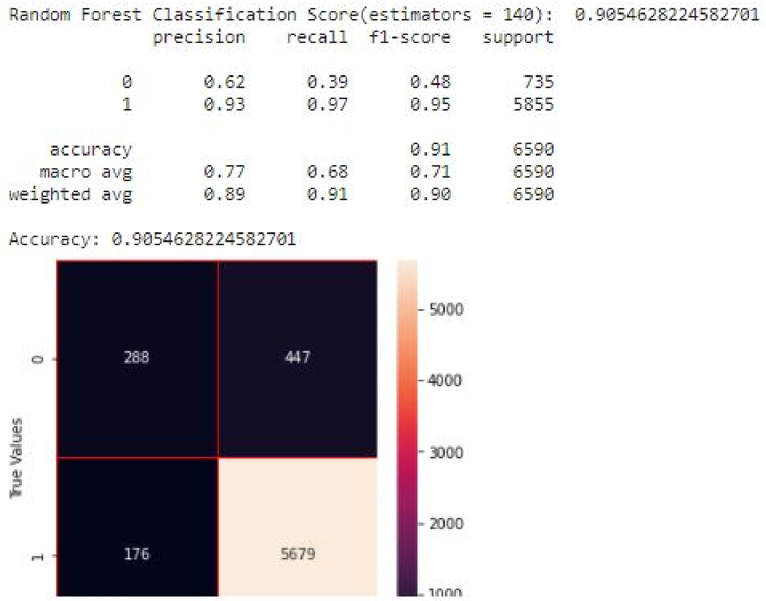


Figure 1: System Architecture

8 Test Cases



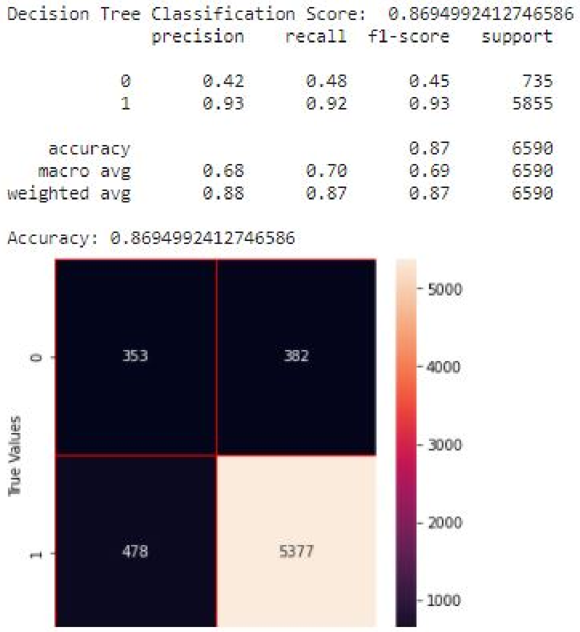
```
Random Forest Classification Score(estimators = 140):  0.9054628224582701
              precision    recall  f1-score   support

           0       0.62      0.39      0.48       735
           1       0.93      0.97      0.95      5855

    accuracy                           0.91      6590
   macro avg       0.77      0.68      0.71      6590
weighted avg       0.89      0.91      0.90      6590

Accuracy: 0.9054628224582701
```

Figure 2: Output for Random Forest Classifier



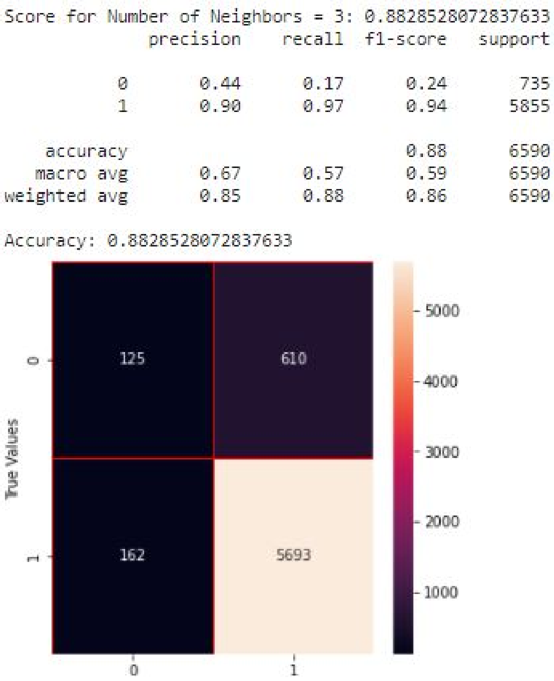
```
Decision Tree Classification Score:  0.8694992412746586
              precision    recall  f1-score   support

           0       0.42      0.48      0.45       735
           1       0.93      0.92      0.93      5855

    accuracy                           0.87      6590
   macro avg       0.68      0.70      0.69      6590
weighted avg       0.88      0.87      0.87      6590

Accuracy: 0.8694992412746586
```

Figure 3: Output for Decision tree



```
Score for Number of Neighbors = 3: 0.8828528072837633
              precision    recall  f1-score   support

           0       0.44      0.17      0.24       735
           1       0.90      0.97      0.94      5855

    accuracy                           0.88      6590
   macro avg       0.67      0.57      0.59      6590
weighted avg       0.85      0.88      0.86      6590

Accuracy: 0.8828528072837633
```

Figure 4: Output for K-Nearest Neighbour

9 Result

The Accuracy for Various models are:

Model	Accuracy
DecisionTree	86.94
RandomForest	90.54
KNN	88.28

Table 1: Accuracy of various Models

We see that Random Forest Classifier gives the best score. We then use this model to perform training and testing of the model. After training, the model gives an accuracy of 90.54 %.

Figure 5: Comparison of various models

10 Conclusion

We have analysed the Banking(term deposit) dataset and performed data pre-processing steps.We have experimented multiple classification models and found out the best performer amongst them. We presented classification of banking(term deposit) results to predict whether client will subscribe to a term deposit. We report a classification accuracy of 90.54%

References

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[4] https://www.kaggle.com/rashmiranu/banking-dataset-classification