Savitribai Phule Pune University
Modern Education Society's College of Engineering, Pune
19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NAAC WITH "A" GRADE (CGPA – 3.13)

DEPARTMENT OF COMPUTER ENGINEERING

A MINI PROJECT REPORT

ON

"Loan Prediction using Classifiers"

B.E. (COMPUTER)

SUBMITTED BY

KUNAL BHAPKAR (71818401H)

KRATI PATNI (71818389E)

PRADDYUMN WADEKAR (71818640M)

UNDER THE GUIDANCE OF

DR. R. A. KHAN

(Academic Year: 2020-2021)

Savitribai Phule Pune University
Modern Education Society's College of Engineering, Pune
19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NAAC WITH "A" GRADE (CGPA – 3.13)

DEPARTMENT OF COMPUTER ENGINEERING

Certificate

This is to certify that mini project entitled

"LOAN PREDICTION USING CLASSIFIERS"

has been completed by Mr.Kunal Bhapkar,Miss.Krati Patni,Mr.Praddyumn Wadekar of BE COMP Second Shift in the Semester - I of academic year 2020-2021 in partial fulfillment of the Fourth Year of Bachelor degree in "Computer Engineering" as prescribed by the Savitribai Phule Pune University.

Dr.R.A. Khan **(Dr.(Mrs.) N. F. Shaikh)**
H.O.D

Place: MESCOE, Pune.
Date: / /2020

ACKNOWLEDGEMENT

It gives us great pleasure and satisfaction in presenting this seminar on "Loan Prediction Using Classifiers".

We would like to express our deep sense of gratitude towards all the teaching staff of Computer Department which helped us in successfully completing our project work.Also we would like to extend our sincere esteems to all the staff in laboratory for their timely support.

We have furthermore to thank Computer Department HOD **Dr.(Mrs.) N. F. Shaikh** *and Guide* **Dr. R.A. Khan** *to encourage me to go ahead and for continuous guidance. I also want to thank* **Prof. A.S. Kamble** *for all his assistance and guidance for preparing report.*

I would like to thank all those, who have directly or indirectly helped me for the completion of the work during this mini project.

Kunal Bhapkar
Krati Patni
Praddyumn Wadekar
B.E. Computer

Contents

1 INTRODUCTION **1**
1.1 Introduction 1

2 DATASET DESCRIPTION **2**

3 PROBLEM STATEMENT **3**

4 CLASSIFICATION ALGORITHMS **4**
4.0.1 KNN 4
4.0.2 Decision Tree 4
4.0.3 Naive Bayes 5

5 CONFUSION MATRIX **6**

6 SCREENSHOTS OF PROJECT **8**

7 CONCLUSION **15**

List of Figures

6.1 Dataset 8
6.2 Design for training 9
6.3 Design for testing 9
6.4 Model Applied on KNN Classification algorithm 10
6.5 KNN Classification Algorithm 10
6.6 Model Applied on Decision Tree Classification algorithm 11
6.7 Decision Tree Classification Algorithm 11
6.8 Model Applied on Naive Bayes Classification algorithm 12
6.9 Naive Bayes Classification Algorithm 12
6.10 Design for Cross-Validation 13
6.11 Confusion Matrix for KNN Algorithm 13
6.12 Confusion Matrix for Decision Tree Algorithm 14
6.13 Confusion Matrix for Naive Bayes Algorithm 14

List of Tables

Abstract

Loan prediction is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Data mining is a process of investigating data from the different perspectives and summarizing it into useful and valuable information. The main purpose of the data mining process is to identify new patterns from the existing data and to understand the data patterns to present meaningful and helpful information for the users. In this report, data mining techniques were used to analyze loan datasets to predict and compared the results of each classification models performance. These results may help out banks in making accurate decisions in loan prediction.The performances of Naive Bayes KNN, Decision Tree classification algorithm were evaluated and compared. Keywords-Data Mining, Classifiers,Loan Dataset, Data Preprocessing, Feature Selection..

Keywords - ***Data Mining, Decision Tree, K-Nearest Neighbors, Classification, Naive Bayes, Performance***

Chapter 1

INTRODUCTION

1.1 Introduction

Data Mining is one of the most motivating and essential research areas with an objective of discovering significant information from large amounts of data sets. In present period, Data mining methods are becoming favourites in the medical field because there is a requirement for an effective analytical methodology to find the unknown and precious information hidden in medical data. Data mining offers various benefits in health industry, such as detecting the unfair practices in health insurance industry, availability of various medical treatments for curing the diseases to the patients at lower expenditure, finding the reasons for various diseases and detection of best medical treatment procedures. It can also help the healthcare researchers for creating efficient medical policies, designing drug recommendation models, preparing the health records of individual patients etc.. The most common data mining technique it risks.used in both academia and industry for data analysis is Classification. Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

Rapid Miner is a software package that allows data mining, text mining and predictive analytics. The program allows the user to enter raw data, including databases and text, which is then automatically and intelligently analysed on a large scale. Rapid Miner includes a free trial to assess its capabilities

Chapter 2

DATASET DESCRIPTION

A data set is a collection of data. In other words, a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. Loan Prediction Data set is used for analysis using Classification Algorithm.This consists of 324 records and 8 attributes mentioned below.

- Principal
- Terms
- Effective_date
- Due_date
- Age
- Education
- Gender

Chapter 3

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.

Chapter 4

CLASSIFICATION ALGORITHMS

4.0.1 KNN

It is the nearest neighbor algorithm.The k-nearest neighbors algorithm is a technique for classifying objects based on the next training data in the feature space. It is the simplest among all mechanism learning algorithms. This algorithm is initialized by selecting k points in kd as the initial k cluster representatives or centroids. Techniques for selecting the primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k-times .

Then the algorithm iterates between two steps till junction:

- Step1: In Data Assignment each data point is assigned to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.
- Step2: Relocation of means each group representative is relocated to the center of all data points assigned to it. If the data points come with a possibility measure then the relocation is to the expectations of the data partitions K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing etc.

There are various advantages of KNN classifiers. These are: ease, efficacy, intuitiveness and competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data.

A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also be large.

4.0.2 Decision Tree

A decision tree is a flowchart-like tree structure in which each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label.

The topmost node in a tree is the root node.The paths from root to leaf represent classification rules.Decision Tree algorithm belongs to the family of supervised learning

algorithms. Unlike other supervised learning algorithms, decision tree algorithms can be used for solving regression and classification problems too.

The general motive of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data (training data).

Decision trees often mimic the human level thinking so it's so simple to understand the data and make some good interpretations. Decision trees actually make you see the logic for the data to interpret (not like black box algorithms like SVM, KNN, etc)

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

4.0.3 Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate preprocessing, it is competitive in this domain with more advanced methods including support vector machines.It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum Likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

Chapter 5

CONFUSION MATRIX

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:

- The target variable has two values: Positive or Negative
- The columns represent the actual values of the target variable
- The rows represent the predicted values of the target variable

Understanding True Positive, True Negative, False Positive and False Negative in a Confusion Matrix

True Positive (TP)

- The predicted value matches the actual value.
- The actual value was positive and the model predicted a positive value.

True Negative (TN)

- The predicted value matches the actual value.
- The actual value was negative and the model predicted a negative value.

False Positive (FP) – Type 1 error

- The predicted value was falsely predicted.
- The actual value was negative but the model predicted a positive value.
- Also known as the Type 1 error

False Negative (FN) – Type 2 error

- The predicted value was falsely predicted.

- The actual value was positive but the model predicted a negative value.
- Also known as the Type 2 error

Accuracy: The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision can be thought of as a measure of exactness (i.e., what percentage of tuples labeled as positive are actually such).

$$precision = \frac{TP}{TP + FP}$$

Recall: Recall is a measure of completeness (what percentage of positive tuples are labeled as such)

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Chapter 6

SCREENSHOTS OF PROJECT

Row No.	loan_status	education	Gender	Principal	terms	effective_da...	due_date	age
1	PAIDOFF	Bechalor	female	1000	30	Sep 8, 2016	Oct 7, 2016	33
2	PAIDOFF	college	male	1000	15	Sep 8, 2016	Sep 22, 2016	27
3	PAIDOFF	college	female	1000	30	Sep 9, 2016	Oct 8, 2016	28
4	PAIDOFF	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	29
5	PAIDOFF	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	36
6	PAIDOFF	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	28
7	PAIDOFF	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
8	PAIDOFF	college	male	300	7	Sep 10, 2016	Sep 16, 2016	29
9	PAIDOFF	High School ...	male	1000	15	Sep 10, 2016	Oct 9, 2016	39
10	PAIDOFF	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
11	PAIDOFF	college	female	900	7	Sep 10, 2016	Sep 16, 2016	26
12	PAIDOFF	High School ...	male	1000	7	Sep 10, 2016	Sep 16, 2016	27
13	PAIDOFF	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
14	PAIDOFF	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	40
15	PAIDOFF	High School ...	male	1000	15	Sep 10, 2016	Sep 24, 2016	32
16	PAIDOFF	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	32
17	PAIDOFF	college	male	800	30	Sep 10, 2016	Oct 9, 2016	26
18	PAIDOFF	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
19	PAIDOFF	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	25
20	PAIDOFF	college	male	1000	15	Sep 10, 2016	Sep 24, 2016	26
21	PAIDOFF	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	29
22	PAIDOFF	Bechalor	male	800	15	Sep 10, 2016	Sep 24, 2016	39
23	PAIDOFF	Bechalor	male	1000	15	Sep 10, 2016	Sep 24, 2016	34
24	PAIDOFF	college	male	1000	30	Sep 11, 2016	Oct 10, 2016	31

ExampleSet (345 examples, 1 special attribute, 7 regular attributes)

Figure 6.1: Dataset

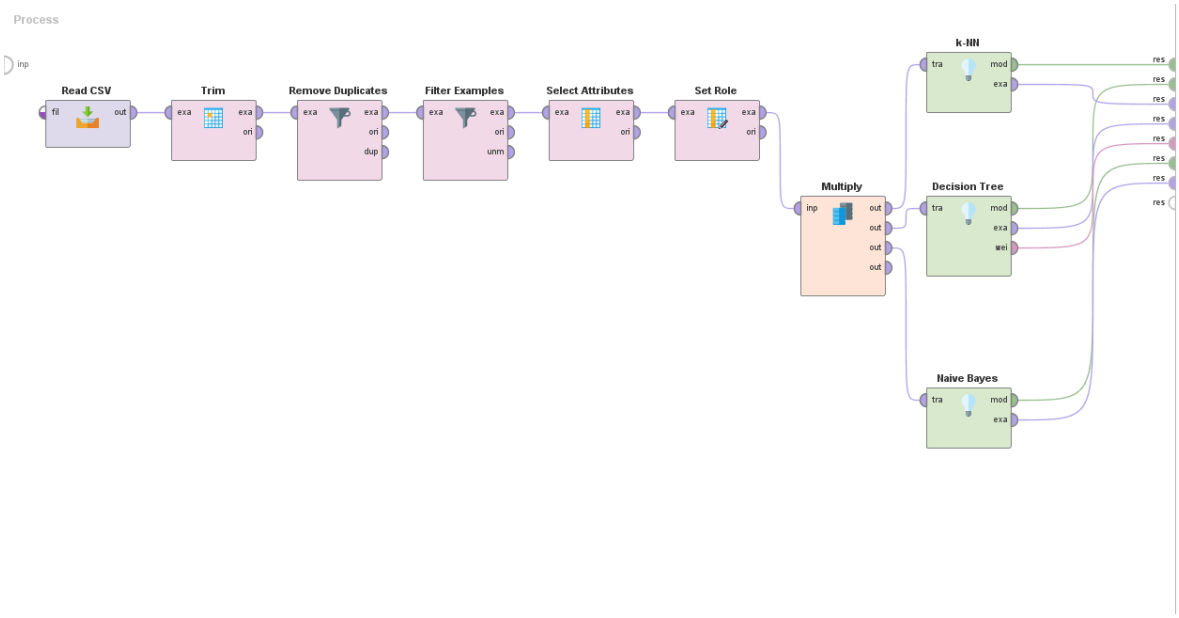

Figure 6.2: Design for training

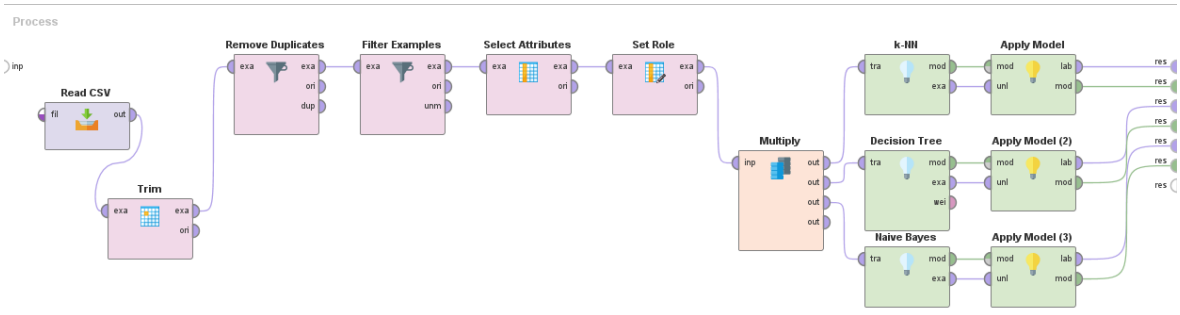

Figure 6.3: Design for testing

Row No.	loan_status	prediction(lo...	confidence(...	confidence(...	education	Gender	Principal	terms	effective_da...	due_date	age
1	PAIDOFF	PAIDOFF	0.833	0.167	High School ...	male	1000	30	Sep 8, 2016	Oct 7, 2016	45
2	PAIDOFF	PAIDOFF	0.833	0.167	Bechalor	female	1000	30	Sep 8, 2016	Oct 7, 2016	33
3	PAIDOFF	COLLECTION	0.400	0.600	college	male	1000	15	Sep 8, 2016	Sep 22, 2016	27
4	PAIDOFF	PAIDOFF	0.866	0.134	college	female	1000	30	Sep 9, 2016	Oct 8, 2016	28
5	PAIDOFF	PAIDOFF	0.865	0.135	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	29
6	PAIDOFF	PAIDOFF	0.764	0.236	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	36
7	PAIDOFF	PAIDOFF	0.869	0.131	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	28
8	PAIDOFF	PAIDOFF	0.757	0.243	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
9	PAIDOFF	PAIDOFF	1	0	college	male	300	7	Sep 10, 2016	Sep 16, 2016	29
10	PAIDOFF	COLLECTION	0.439	0.561	High School ...	male	1000	15	Sep 10, 2016	Oct 9, 2016	39
11	PAIDOFF	PAIDOFF	0.658	0.342	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
12	PAIDOFF	PAIDOFF	1	0	college	female	900	7	Sep 10, 2016	Sep 16, 2016	26
13	PAIDOFF	PAIDOFF	1.000	0	High School ...	male	1000	7	Sep 10, 2016	Sep 16, 2016	27
14	PAIDOFF	PAIDOFF	0.757	0.243	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
15	PAIDOFF	COLLECTION	0.390	0.610	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	40
16	PAIDOFF	PAIDOFF	0.597	0.403	High School ...	male	1000	15	Sep 10, 2016	Sep 24, 2016	32
17	PAIDOFF	COLLECTION	0.428	0.572	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	32
18	PAIDOFF	PAIDOFF	0.587	0.413	college	male	800	30	Sep 10, 2016	Oct 9, 2016	26
19	PAIDOFF	PAIDOFF	0.658	0.342	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
20	PAIDOFF	PAIDOFF	0.615	0.385	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	25
21	PAIDOFF	COLLECTION	0.422	0.578	college	male	1000	15	Sep 10, 2016	Sep 24, 2016	26
22	PAIDOFF	COLLECTION	0.393	0.607	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	29
23	PAIDOFF	PAIDOFF	0.758	0.242	Bechalor	male	800	15	Sep 10, 2016	Sep 24, 2016	39
24	PAIDOFF	PAIDOFF	0.593	0.407	Bechalor	male	1000	15	Sep 10, 2016	Sep 24, 2016	34

ExampleSet (346 examples, 4 special attributes, 7 regular attributes)

Figure 6.4: Model Applied on KNN Classification algorithm

KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 346 examples with 7 dimensions of the following classes:
  PAIDOFF
  COLLECTION
```

Figure 6.5: KNN Classification Algorithm

Row No.	loan_status	prediction(lo...	confidence(...	confidence(...	education	Gender	Principal	terms	effective_da...	due_date	age
1	PAIDOFF	PAIDOFF	1	0	High School ...	male	1000	30	Sep 8, 2016	Oct 7, 2016	45
2	PAIDOFF	PAIDOFF	1	0	Bechalor	female	1000	30	Sep 8, 2016	Oct 7, 2016	33
3	PAIDOFF	PAIDOFF	1	0	college	male	1000	15	Sep 8, 2016	Sep 22, 2016	27
4	PAIDOFF	PAIDOFF	0.598	0.402	college	female	1000	30	Sep 9, 2016	Oct 8, 2016	28
5	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	29
6	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	36
7	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	28
8	PAIDOFF	PAIDOFF	0.598	0.402	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
9	PAIDOFF	PAIDOFF	1	0	college	male	300	7	Sep 10, 2016	Sep 16, 2016	29
10	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	15	Sep 10, 2016	Oct 9, 2016	39
11	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
12	PAIDOFF	PAIDOFF	1	0	college	female	900	7	Sep 10, 2016	Sep 16, 2016	26
13	PAIDOFF	PAIDOFF	1	0	High School ...	male	1000	7	Sep 10, 2016	Sep 16, 2016	27
14	PAIDOFF	PAIDOFF	0.598	0.402	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
15	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	40
16	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	15	Sep 10, 2016	Sep 24, 2016	32
17	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	32
18	PAIDOFF	PAIDOFF	0.598	0.402	college	male	800	30	Sep 10, 2016	Oct 9, 2016	26
19	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
20	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	25
21	PAIDOFF	PAIDOFF	0.598	0.402	college	male	1000	15	Sep 10, 2016	Sep 24, 2016	26
22	PAIDOFF	PAIDOFF	0.598	0.402	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	29
23	PAIDOFF	PAIDOFF	0.598	0.402	Bechalor	male	800	15	Sep 10, 2016	Sep 24, 2016	39
24	PAIDOFF	PAIDOFF	0.598	0.402	Bechalor	male	1000	15	Sep 10, 2016	Sep 24, 2016	34

ExampleSet (346 examples, 4 special attributes, 7 regular attributes)

Figure 6.6: Model Applied on Decision Tree Classification algorithm

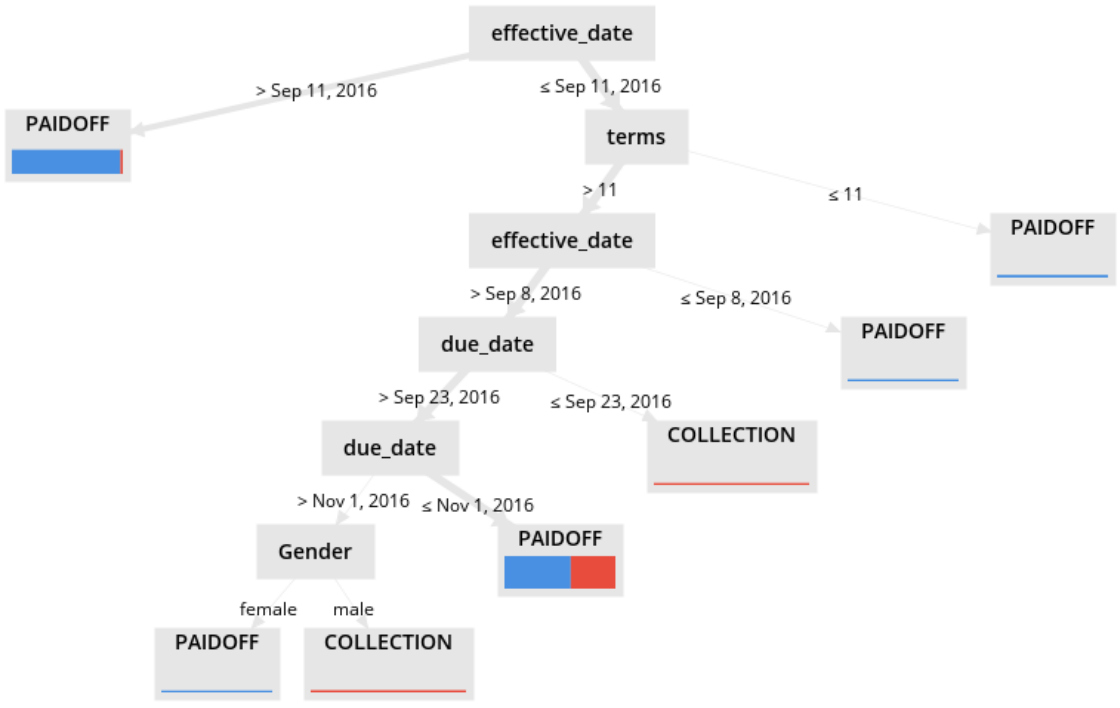

Figure 6.7: Decision Tree Classification Algorithm

Row No.	loan_status	prediction(lo...	confidence(...	confidence(...	education	Gender	Principal	terms	effective_da...	due_date	age
1	PAIDOFF	PAIDOFF	0.998	0.002	High School ...	male	1000	30	Sep 8, 2016	Oct 7, 2016	45
2	PAIDOFF	PAIDOFF	0.999	0.001	Bechalor	female	1000	30	Sep 8, 2016	Oct 7, 2016	33
3	PAIDOFF	PAIDOFF	0.999	0.001	college	male	1000	15	Sep 8, 2016	Sep 22, 2016	27
4	PAIDOFF	PAIDOFF	0.941	0.059	college	female	1000	30	Sep 9, 2016	Oct 8, 2016	28
5	PAIDOFF	PAIDOFF	0.873	0.127	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	29
6	PAIDOFF	PAIDOFF	0.880	0.120	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	36
7	PAIDOFF	PAIDOFF	0.872	0.128	college	male	1000	30	Sep 9, 2016	Oct 8, 2016	28
8	PAIDOFF	PAIDOFF	0.816	0.184	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
9	PAIDOFF	PAIDOFF	1.000	0.000	college	male	300	7	Sep 10, 2016	Sep 16, 2016	29
10	PAIDOFF	PAIDOFF	0.527	0.473	High School ...	male	1000	15	Sep 10, 2016	Oct 9, 2016	39
11	PAIDOFF	COLLECTION	0.405	0.595	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
12	PAIDOFF	PAIDOFF	0.871	0.129	college	female	900	7	Sep 10, 2016	Sep 16, 2016	26
13	PAIDOFF	PAIDOFF	0.677	0.323	High School ...	male	1000	7	Sep 10, 2016	Sep 16, 2016	27
14	PAIDOFF	PAIDOFF	0.816	0.184	college	male	800	15	Sep 10, 2016	Sep 24, 2016	26
15	PAIDOFF	COLLECTION	0.403	0.597	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	40
16	PAIDOFF	PAIDOFF	0.567	0.433	High School ...	male	1000	15	Sep 10, 2016	Sep 24, 2016	32
17	PAIDOFF	COLLECTION	0.392	0.608	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	32
18	PAIDOFF	PAIDOFF	0.686	0.314	college	male	800	30	Sep 10, 2016	Oct 9, 2016	26
19	PAIDOFF	COLLECTION	0.405	0.595	college	male	1000	30	Sep 10, 2016	Oct 9, 2016	26
20	PAIDOFF	COLLECTION	0.371	0.629	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	25
21	PAIDOFF	PAIDOFF	0.580	0.420	college	male	1000	15	Sep 10, 2016	Sep 24, 2016	26
22	PAIDOFF	COLLECTION	0.384	0.616	High School ...	male	1000	30	Sep 10, 2016	Oct 9, 2016	29
23	PAIDOFF	PAIDOFF	0.821	0.179	Bechalor	male	800	15	Sep 10, 2016	Sep 24, 2016	39
24	PAIDOFF	PAIDOFF	0.582	0.418	Bechalor	male	1000	15	Sep 10, 2016	Sep 24, 2016	34

ExampleSet (346 examples, 4 special attributes, 7 regular attributes)

Figure 6.8: Model Applied on Naive Bayes Classification algorithm

SimpleDistribution

```
Distribution model for label attribute loan_status


Class PAIDOFF (0.751)
7 distributions

Class COLLECTION (0.249)
7 distributions
```

Figure 6.9: Naive Bayes Classification Algorithm

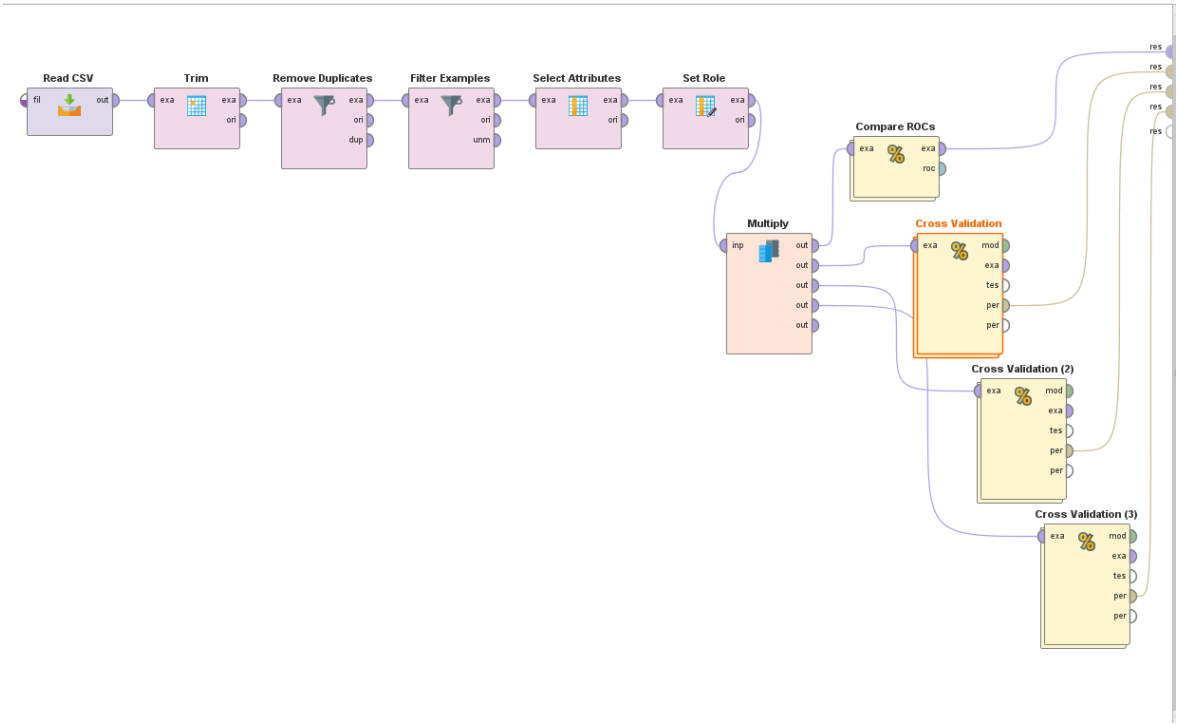

Figure 6.10: Design for Cross-Validation

accuracy: 70.76% +/- 4.78% (micro average: 70.72%)

	true PAIDOFF	true COLLECTION	class precision
pred. PAIDOFF	228	70	76.51%
pred. COLLECTION	31	16	34.04%
class recall	88.03%	18.60%	

Figure 6.11: Confusion Matrix for KNN Algorithm

accuracy: 74.76% +/- 4.41% (micro average: 74.78%)

	true PAIDOFF	true COLLECTION	class precision
pred. PAIDOFF	251	79	76.06%
pred. COLLECTION	8	7	46.67%
class recall	96.91%	8.14%	

Figure 6.12: Confusion Matrix for Decision Tree Algorithm

accuracy: 71.08% +/- 10.57% (micro average: 71.01%)

	true PAIDOFF	true COLLECTION	class precision
pred. PAIDOFF	200	41	82.99%
pred. COLLECTION	59	45	43.27%
class recall	77.22%	52.33%	

Figure 6.13: Confusion Matrix for Naive Bayes Algorithm

Chapter 7

CONCLUSION

For the dataset, we developed a number of models from diverse data mining techniques. This was helpful because it gave us a variety of models and indicated which model is superior by evaluating the accuracy, sensitivity, precision and other measures.We choose the model with the highest overall accuracy, sensitivity and precision. If we just go by that criterion, then the best model is Decision Tree. The results of the classification algorithms applied to the loan dataset were compared. Decision Tree outperformed Naive Bayes and KNN based on accuracy and other performance measures. And the performance of the classifiers increased after removing the unnecessary attributes.

Bibliography

[1] Jake VanderPlas(Mar. 2013) Python Data Science Handbook Chapter 3. Data Manipulation with Pandas by O'Reilly

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, Chapter 8. Classification: Basic Concepts

[3] https://data-flair.training/blogs/machine-learning-classification-algorithms/

[4] Andreas C. Müller & Sarah Guido(Mar. 2013) Introduction to Machine Learning with Python by O'Reilly

[5] https://www.udemy.com/course/machinelearning/learn/lecture/5772258?start=0#overview

[6] Aurélien Géron(Mar. 2017) Hands-On Machine Learning with Scikit-Learn and TensorFlow

[7] https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

[8] https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021