Data Mining for Fraud Detection in Pawning Services offered by Banks

D.T.D Gamage

179461A

Faculty of Information Technology
University of Moratuwa

2020

Data Mining for Fraud Detection in Pawning Services offered by Banks

D. T. D. Gamage

179461A

Supervisor: Mr. Saminda Premarathna

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the fulfillment of the requirements of Degree of Master of Science in Information Technology.

June 2020

Declaration

We declare that is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student	Signature of Student
D.T.D Gamage	……
	Date: 2020-06-09

Supervised by

Name of Supervisor	Signature of Supervisor
Mr. S.C. Premarathne	…….
	Date: ……...

Acknowledgements

First and foremost, I would like to express my sincere gratitude towards my supervisor, Mr. Saminda Premarathne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for his guidance, supervision, advices and sparing valuable time thorough the research project.

Furthermore, big thank should goes to Prof. Mohamad Ferdihous who taught Research Methodology and Literature Review and thesis writing subjects which were the basis for this research work.

I would also like to thank my family specially my father for the support provided me through the completion of this project and all the batch mates of the M.Sc. in IT degree program and my office mates who gave their valuable feedbacks to improve the results of the research. I must acknowledge Mr. M.C Rajapaksha, my boss, here for the support given me to collect data for this research and my future husband for the support and encouragement given throughout this project.

ABSTRACT

Pawning or gold-pledged loans has been popular instrument/service among the various products provided by banks and other financial institutions in Sri Lanka in recent decades. Major reason for pawning becoming that much popular is gold items is one of the most reliable sources of credit for lower and middle cast households in Sri Lanka. With the increasing demand for pawning services offered by financial institutions exposure for frauds become incremental accordingly. With the increasing market value of gold people who are seeking for chances to do a fraud are activated and find various methods to achieve their goals. This Paper attempts to explore how data mining techniques can be used in detecting the fraudulent Transactions with related to pawning services offered by banking sector in Sri Lanka.This descriptive analysis on discovering frauds in pawning services will analyses the main parameters associated with pawning and how they can be effectively use for detecting fraudulent transactions. Finally, at the end this research will emphasizes how data mining techniques can be used for detecting fraudulent transactions by providing a solution for pawning fraud detection based on classification model.

Keywords— Pawning fraud, Banking fraud, fraud detection, Data Mining

Table of Contents

Declaration ii
ABSTRACT iv
LIST OF TABLES viii
LIST OF FIGURES ix
1 CHAPTER ONE 1
INTRODUCTION 1
1.1 Prolegomena 1
1.2 Background of the Study 2
1.3 Problem Statement 3
1.4 Aims and Objectives of the research 3
1.5 Proposed Solution 4
1.6 Structure of the thesis 5
2 CHAPTER TWO 6
REVIEW OF LITERATURE 6
2.1 Introduction 6
2.2 Pawning procedures in banking sector 6
2.3 Pawning Frauds 7
2.4 Traditional pawning fraud detection procedures in banking sector 7
2.5 Issues with existing pawning fraud detection procedures in banking sector 8
2.6 Data mining for fraud detection History 9
2.7 Data Mining Applications in Use for Fraud Detection 13
2.8 Research problem identification 14
2.9 Summary 14
3 CHAPTER THREE 15
TECHNOLOGY ADAPTED IN PAWNING FRAUD DETECTION 15
3.1 Introduction 15
3.2 What is the data mining? 15
3.3 Reasons for choosing classification for pawning fraud detection 17
3.4 Why Classification? 17
3.4.1 Supervised Learning algorithms 18
3.5 Classification Algorithms 19

3.6 Microsoft SQL SERVER 23
3.7 Rapid Minor Studio 23
3.8 Summary 23
4 CHAPTER FOUR 24
METHODOLOGY FOR DETECTING AND PREDICTING FRAUDS IN PAWNING SERVICES 24
4.1 Introduction 24
4.2 Hypothesis 24
4.3 Input 24
4.4 Output 25
4.5 Process 25
4.5.1 Data collection and selection 25
4.5.2 Data Pre-processing 26
4.5.3 Attribute Selection 26
4.6 Data mining 27
4.7 Users 27
4.8 Features 27
4.9 Summary 28
5 CHAPTER FIVE 29
RESEARCH DESIGN AND ANALYSIS 29
5.1 Introduction 29
5.2 Research Design 29
5.3 Detailed Research Design 30
5.3.1 Sub Research Question One 30
5.3.2 Sub Research Question Two 30
5.3.3 Sub Research Question 3 30
5.4 Summary 30
6 CHAPTER SIX 31
6.1 Introduction 31
6.2 Solution for Research Question One 31
6.3 Solutions for Research Question Two 31
6.3.1 Data Preprocessing 31

6.3.2 Attribute Selection 32
6.3.3 Correlation Matrix 33
6.3.4 Feature Weights 34
6.4 Solution for Sub Research Question 3 35
6.4.1 Model Creation 35
6.5 Summary 36
7 CHAPTER SEVEN 37
7.1 Introduction 37
7.1.1 Evaluation for Classification 37
7.1.2 Confusion Matrix 37
7.2 Evaluation Results 39
7.3 Summary 44
8 CHAPTER EIGHT 45
CONCLUSION AND FUTURE WORK 45
8.1 Introduction 45
8.2 Overview of the research 45
8.3 Limitations 46
8.4 Future work of the project 46
8.5 Summary 46
REFERENCES 47
9 Appendix A 49
Data Preprocessing 49
10 Appendix B 51
Attribute Selection 51
11 Appendix C 53
Model Creation 53
12 Appendix D 55
Evaluation Results 55
Results for Correlation Matrix Attributes 55
Results for Information gain attributes 61

LIST OF TABLES

Table 6-1 Attributes selected by correlation matrix..... 33
Table 6-2 Attributes selected by weight by Information Gain 34
Table 7-1 Classification Evaluation Measurements..... 38
Table 7-2 Comparison of classification techniques dataset1 cross validation 39
Table 7-3 Comparison of classification techniques_dataset1 split validation..... 39
Table 7-4 Comparison of classification techniques dataset2 cross validation 40
Table 7-5 Comparison of classification techniques dataset2 split validation..... 40

LIST OF FIGURES

Figure 2.1 Fraud detection technique……..11

Figure 3.1 KDD Process in datamining……...16

Figure 3.2 supervised learning techniques……...18

Figure 3.3 Structure of Artificial Neural networks……..22

Figure 6.1 Selecting and removing outlier values……....49

Figure 6.2 Converting class variable type to nominal……..49

Figure 6.3 Generate new categorized attributes from the existing attributes 1……50

Figure 6.4 Generate new categorized attributes from the existing attributes 2……50

Figure 6.5 Attribute selection using correlation matrix……...51

Figure 6.6 Attribute selection using correlation matrix results……51

Figure 6.7 Attribute selection using weight by information gain……52

Figure 6.8 Attribute selection using weight by information gain results ……52

Figure 6.9 Model creation cross validation……..53

Figure 6.10 Model creation cross validation parameters……53

Figure 6.11 Model creation split validation……54

Figure 6.12 Model creation split validation parameters……...54

Figure 7.1 Confusion matrix……....38

Figure 7.2 Cross validation Bagging ……...55

Figure 7.3 Cross validation Decision tree ……...55

Figure 7.4 Cross validation KNN……56

Figure 7.5 Cross validation Naive Bayes……....56

Figure 7.6 Cross validation Random Forest……57

Figure 7.6 Cross validation Random Forest……57

Figure 7.8 Split validation Decision tree ……....58

Figure 7.9 Split validation KNN……58

Figure 7.10 Split validation Naive Bayes ……...59

Figure 7.11 Split validation Random Forest……...59
Figure 7.12 Cross validation Bagging…….60
Figure 7.13 Cross validation Decision tree…….60
Figure 7.14 Cross validation KNN……..61
Figure 7.15 Cross validation Naive Bayes……..61
Figure 7.16 Cross validation Random Forest……..62
Figure 7.17 Split validation Bagging……..62
Figure 7.18 Split validation Decision Tree…….63
Figure 7.19 Split validation KNN……...63
Figure 7.20 Split validation Naive Bayes……...64
Figure 7.21 Split validation Random Forest……...64
Figure 7.22 Algorithm Performance comparison for Dataset 1 –Accuracy……41
Figure 7.23 Algorithm Performance comparison for Dataset 2 –Accuracy……42
Figure 7.24 Algorithm Performance comparison for Dataset 1 –Precision ……...42
Figure 7.25 Algorithm Performance comparison for Dataset 2 –Precision ……...43
Figure 7.26 Algorithm Performance comparison for Dataset 1 –Recall ……43
Figure 7.27 Algorithm Performance comparison for Dataset 2 –Recall…….44

CHAPTER ONE

INTRODUCTION

1.1 Prolegomena

This chapter introduces the overall research project on data mining for fraud detection in pawning services offered by banks. Sri Lankan banking sector holds more than 55% of the financial sector assets and so that plays a pivotal role in supporting the growth momentum of the economy. Today banking has become a compulsory event in the lifestyles of the present generation. Almost everyone is dealing with banks every day to fulfill their various requirements. With the rapid advancements in technology and tools as well as the increasing interactions between the customers and bank, occurrence of fraudulent actions has been growing rapidly.

Pawn brokering or gold loan is a popular service provided by almost all the banking institutions today since it is a fixed term facility secured by gold as collateral. With today's busy life style everyone prefers to gain a gold loan for immediate need of money as it can be availed within short duration from a banking institution or other non-banking finance institution. Any other loan schemes provided by financial institutions consumes more time compared to gold loans for the process and paperwork.

There are a number of banking fraud detection tools available and almost of them provide solutions for credit card, online banking fraud, ATM Frauds. But none of the previous researchers has been paid attention for pawning frauds so far. Higher demand for the gold and increasing market value has led to the occurrence of pawning frauds, but due to the security reasons there not imposed to the outside. Time has come to pay much attention to pawning frauds as same as for the other type of banking frauds. Existing pawning fraud detection techniques used in banks are very time consuming and most of the time frauds are identified after a few months later the real transaction has happened. This study is focusing on pawning services in banking sector and will be more beneficial to the bankers and internal auditors who are auditing the banking operations since no one have done proper analysis on how to use to use past transaction data to predict the fraudulent transactions. There are several factors to be

consider for detecting pawning frauds such as first granted advance amount, last date of the payment done for the gold loan, net gold amount, karat content etc.

1.2 Background of the Study

Pawning or gold loan is a quick and fixed term loan facility provided by taking gold as collateral and repayable within a given period with interest accrued. Pawn brokering in Sri Lanka is regulated by pawn brokers audience since 1942 and has become a highly competitive service provided by state banks, commercial banks and other financial institutions. Sri Lankan people have a higher interest towards jewelry items pawning services have become this much popular among middle- and lower-income households. Gold loans are provided by formal financial institutions such as banks and informal private money lenders and other finance companies as well in Sri Lanka.This is an essential source of credit for lower -income households which takes loans against gold as collateral. Gold loans are disbursed more quickly with minimal procedural requirements with compared to other sources of credit available for low-income households. So gold loan products have become extremely popular and reliable due to the consumer-friendly features with compared to other sources of credit. People take gold loans for various reasons including agricultural, commercial, housing and industrial purposes.

Banks gain several unique benefits from providing pawning services which are not likely to get by other loan schemes provided by them. Pawning differs from other forms of secured lending where lender takes gold as collateral at the time of lending. If the initial granted loan amount has not repaid within the due time bank can recover it by auctioning the collateral. This is a secured as it takes gold which has high liquidity in market as collateral. Usually pawning generates high net interest income compared with other products.

In financial sectors a fraud is defines as an unethical act taken with the intension of gaining unauthorized financial benefit. Financial frauds have been a big concern for many organizations across industries and in different countries since it brings huge devastations to business. Billions of dollars have lost yearly due to banking and other financial frauds. In addition to the monetary losses, fraud has the most critical impact towards bank's reputation, goodwill and customer relations. There are very common examples of frauds that take place in banking sector which is known by everyone. Cash Fraud, Billing Fraud, Cheque Tampering Fraud, Skimming, Financial Statement Fraud, Credit/debit card fraud, Online banking fraud

and loan frauds are few of those frauds. Loan frauds can be categorized further, and pawning loans are given highest priority with the increasing value of gold market price.

Using data mining techniques, we can easily identify the useful patterns from large set of data. They also have been used to identify any inconsistent behavior or fraudulent actions over the past few years. Data mining techniques have been successfully used for fraud detection across many industries including financial sector through last few decades. In fraud detection, main focus is to analyze the transactions and learns inconsistent behavior and detect fraud patterns from large data sets. Artificial Intelligence and statistical technique are mainly used for detection of frauds. [1]

1.3 Problem Statement

Majority of the existing research work on banking fraud detection done based on credit/debit card, financial statement fraud and online banking frauds. Although we could not able to find a research work conducted on pawning frauds. This could be mainly due to the unavailability of sufficient datasets for detecting pawning frauds and financial institutions do not expose their fraud data to outside due to legal and competitive reasons. Traditional way of detecting pawning frauds is very time-consuming task and most of them are detected after months or years later though audits conducted by internal audit teams. Time has come to pay attention for detecting pawning frauds earlier as possible since pawning services play a major role in banking sector today by generating profits same as other services provided.

1.4 Aims and Objectives of the research

Aim:

This study aims to determine the ways of using data mining techniques for fraud detection in the pawning services offered by banks and assess new ways to prevent and manage frauds in the pawning services by providing a simulation based on datamining classification techniques.

Objectives

1. To critically evaluate the underlying literature regarding influential factors for frauds and existing fraud detection techniques that are currently used in pawning services offered by banks.
2. To Identify and develop in-depth knowledge on the parameters which has major impact on frauds in pawning services.
3. To assess the most appropriate data mining technique/s that can be used for the fraud detection in pawning services and investigate in which way we can use them.

1.5 Proposed Solution

We propose a simulation which uses a database of historical pawning transactions data to predict frauds when a new transaction comes. Simulation was carried out over several stages to achieve the ultimate goal of this research. As the first stage we have selected 2000 pawning historical transaction data records from 100 branches and labeled as fraud and non-fraud. Rapid minor studio was used for the preprocessing and analyzing part of this research. In the preprocessing stage replacing missing values, removing outliers and attribute categorization steps has been carried out. Attribute selection was done using correlation matrix and information gain techniques. Several data models were built using five different classification techniques and tested them using the split and cross validation in the fourth stage. In the fifth stage overall performance of the data models was measured using confusion matrix and related measures such as accuracy, precision, recall, sensitivity and f-measure of each data model were calculated. Comparison of each data model performance with respective to the above selected measurements was done in sixth stage in order to find out most effective classification technique for predicting and detecting pawning frauds in banking sector.

1.6 Structure of the thesis

The overall structure of the thesis is as follows, first chapter gives an introduction to the project with the objectives, background, problem, and solution and second chapter will critically review the literature in the data mining technology in financial fraud detection with special reference to pawning services in banking sector. Third chapter is about data mining technology and it relevance for the pawning fraud detection. Forth chapter will present our methodology for pawning fraud detection with inputs, outputs and the process while Fifth chapter gives detail description of design and analysis of the research. In sixth chapter we present an implementation for the solution. Seventh chapter provides the evaluation for the methodology by comparing and analyzing the models. Finally, chapter eight concludes the solution with a note on conclusion and future work of this research.

CHAPTER TWO

REVIEW OF LITERATURE

1.1 Introduction

This chapter critically reviews the existing literature for evaluating use of data mining technique for fraud detection in pawning services offered by banks. First, we discuss concepts and traditional techniques for detecting frauds in pawning services. Then we specify the unsolved issues and concerns related with the traditional fraud detection procedures used in banking sector. Finally, we search existing literature to find how data mining techniques have been used to detect frauds in pawning services offered by banking sector. This chapter also identifies the possible data mining techniques need to be used for detecting frauds. The chapter is organized under the heading traditional fraud detection techniques in pawning services, issues with existing fraud detection techniques.

1.2 Pawning procedures in banking sector

When a customer come to take a gold loan first pawning officer get the articles from the customer and check the karatage and weight of the articles. Accurate assaying and valuation of gold jewelry is important due to the nature of the service. Normally article valuation done based on only the value of gold weight in the articles. For the purpose of measuring the gold weight and gold quality of the articles, the electronic metal tester (densi meter) and acid test which uses touch stone and chemicals are used. Densi meter has ability to indicate karatcontent or gold quality, percentage of gold content and density of the article accurately Those figures could be used determine the karatage of the article. However, the measurements taken from densimeter are no accurate if the articles embedded with non-gold items such as diamonds, stones, gems, carvings and complex designs etc. Reason for that is because it is difficult to measure the purity of gold if the article embedded with stones. Most of the banks usually accept jewelry made of 16K – 24K gold as they are the most saleable ones. If the article's weight and karatage matched with their minimum requirement, then they calculate the maximum amount they can grant for the gold articles considering the weight, karatage, market value and the assess value of the articles. Then they take customers basic and contact details and issue a pawning ticket with unique ticket number. All the articles detail along with granted date and granted

amount are included in this pawning ticket. This pawning ticket number is used as the reference for further operations such as part payments, redemptions and repawning. At the end of the transaction process articles are kept safe with the related pawning ticket.

1.3 Pawning Frauds

Pawning Frauds can be divided mainly as:

1.Frauds made by Customers

2.Frauds made by Staff members

Frauds made by customers

- Pawning low karatage articles, (Raise pawning advances either for dud articles or which are indicating low value).
- Pawning articles made out of other materials (Tungsten or mercury filled) to fortify cartage of an article)

Frauds made by staff members

- Issuing advances by over stating value of an article
- Raising pawning advances on forged NIC Numbers (to issue advances beyond the individual customer limit impose by the bank)
- Issuing advances by taking forged articles as a collateral
- Abuse password of the Manager or supervisor
- Removing articles from safe lockers (theft)

1.4 Traditional pawning fraud detection procedures in banking sector

Detection of a fraud depends on the perception of detection, Auditors Skills. fraud can be detected mainly by two ways.

- ✓ Internal Auditing
  - By physically checking pawning transactions and articles.
  - Checking instances where advances given on violating of instructions, standard policies, standard practices, and guidelines
  - Checking - assessing collaterals
  - Checking of relevant documents ex: KYC
  - Periodic verification of pawned articles in bank custody

- ✓ By monitoring activities and analyzing Data

- Analyzing pawning advance growth, monthly for unusual increase in advances of a branch
- Analyzing income over pawning business
- Analyzing violations of individual customer limit, Branch Limits, Officers authority limits
- Analyzing Genuineness of NIC and other information's
- Staff behavioral patterns (involved in lending activities)
- Checking of transections issued to a single customer from several branches
- Checking of transections which are having pawned bulk of articles
- Checking of transections falls under Non performing category (expired, not repaid)

1.5 Issues with existing pawning fraud detection procedures in banking sector

Since internal audit investigations are the most used fraud detection technique in current banking pawning operations, it has several issues which reduce the chance of detecting a fraud in real time.

Normally internal bank audit investigations are carried out about twice a year to check whether the banking operations are operating according to the bank's rules and regulations. In that kind of audit investigation audit officers cross check the weight and karatage of the articles with the amount printed in the pawning ticket. So the detecting a fraudulent transaction in audit investigation happened may be several months later than the transaction date.

In a normal bank audit investigation validity of the pawning transactions are checked through a manual process which consumes a lot of time. Gold weight, karatage and article count of randomly selected pawning transactions are manually checked with the pawning ticket by the professional audit team. So the checking process is lot of time consuming and difficult task. Some time there is a chance that fraudulent transactions may not be checked and detected by the audit team at the investigations.

Another problem associated with the pawning fraud detection is the measuring issues related to densi meters. In most of the banks densi meters are used for measuring gold weight and karatages. But some of those have measuring issues and would not present the actual weight and karatage. Furthermore, when the articles embedded with stones, diamond and other non-gold items, measuring the correct gold weight become very difficult and there is no guarantee

that densimeters would give the correct gold weight of that kind of articles. So the real gold weight will not be revealed. However, there is no proper method for detecting pawning frauds at real time unless if the article is a dud article. In present most of the dud articles are gold-plated and has similar weight as pure gold articles. So those fake articles cannot be identified at first glance as before.

1.6 Data mining for fraud detection History

Financial fraud detection has been one of the favourite areas of researchers since types of frauds associated with fraud detection has been increased in great scale in past few years. Most of the researchers have selected this area as it is very interesting and challenging area of study within the financial domain. Fast technological enhancements have led the fraud seekers to find new methods and techniques to do frauds. Most frauds have been reported on financial institutions as it is an easy way for fraud seekers to get benefits.

Frauds in the banking sector have increased in large scale all over the world with the enormous changes in technology. Technological advancements open up new paths for fraudsters. All banks have begun to realize the importance of fraud detection with the aim of discovering traditional frauds as well as new types of frauds recently.

Most of the recent research work related to financial fraud detection has been done based on data mining techniques. Not only financial frauds but the other types of frauds also successfully discovered using various data mining techniques. It has become new trend in discovering frauds using data mining as it gives quick and accurate solutions for practical issues occurred in traditional fraud detecting techniques.

A comparative analysis on how various classification techniques are used in fraud detection was done by Ankur Rohilla [1]. Author has used several datamining techniques such as K nearest neighbours, decision tree C5.0 and neural networks for the analysis of data sets. Dataset of credit card data from UCI Machine Learning repository was used. Relevant input parameters were identified clearly and various classification techniques Including neural networks were applied on the data set to find out which technique gives most accurate predictive result for detecting frauds. [1]

Many researchers worked with transaction data and identified between patterns from genuine behaviour of the transaction with higher efficiency and accuracy. They have proposed various fraud detections frameworks and aiming higher efficiency and accuracy. A framework called "i-Alertor" proposed by Wei etal in for major Australian banks while a semi-supervised decision support system named Bank Sealer was proposed in for an Italian bank. Hybrid DM method to predict network intrusions and detect fraud activities was proposed by the authors of "A Hybrid Data Mining Method for Intrusion and Fraud Detection in E-Banking Systems" in 2014.A model for frequent item set mining called "fraudMinor" was introduced in 2014 by the authors of "A novel credit card fraud detection model based on frequent item set mining. [2]

Fraud detection has become a popular application area of data mining, which searches for patterns indicative of fraud among the transaction data. Improving fraud detection will result to reducing the loss and maintaining the viability of the financial systems. Construction of algorithms or models which aims to learn how to recognize a variety of fraud patterns is the major task in fraud detection. These algorithms or models of fraudulent behaviour can have used in decision support systems to prevent frauds as well as to plan audit strategies. Authors of [3] says that a BP neural network with three layers is used in HNC software for detecting credit card fraud. The input layer receives information from an external source, while the hidden layer presents common features of the information and the output layer learns combinations of the features that are associated with known behaviours or decisions. [3]

Another Fraud detection system called Metrix, combines neural networks, statistical methods, along with enhanced classifier systems and advanced fuzzy logic rule induction systems in order to provide customers with the ultimate fraud detection product. [4]

Sanjay et al [5] showed that bagging algorithm performs better as a credit card fraud detection classification technique as compared to Ad boost, Logit boost, CART and Dagging with 0.877 correct classification rate and 0.123 correct misclassification rate.

Masoumeh et al [6] have proved that out of various classification techniques like Naïve Bayes, Support Vector Machines, bagging classifier based on decision tree, K-Nearest neighbour, the bagging classifier based on decision tree is the best classifier which can be used to construct credit card fraud detection model.

It is true that in the practical scenarios, when compared to non-fraudulent ones, frequency of fraudulent transactions is too less. Main reason for that is most of the financial frauds have been kept a secret and these confidentiality issues is one of the biggest challenges in designing a fraud detecting data-mining algorithm. He had also showed that it is necessary to choose attributes from the database which are relevant to the task. These attributes should be selected very carefully after a comprehensive analysis since they play important role as input vectors to the fraud detection system. Each attribute should be analysed and select only the most relevant attributes for implementing more effective fraud detection system.[7]

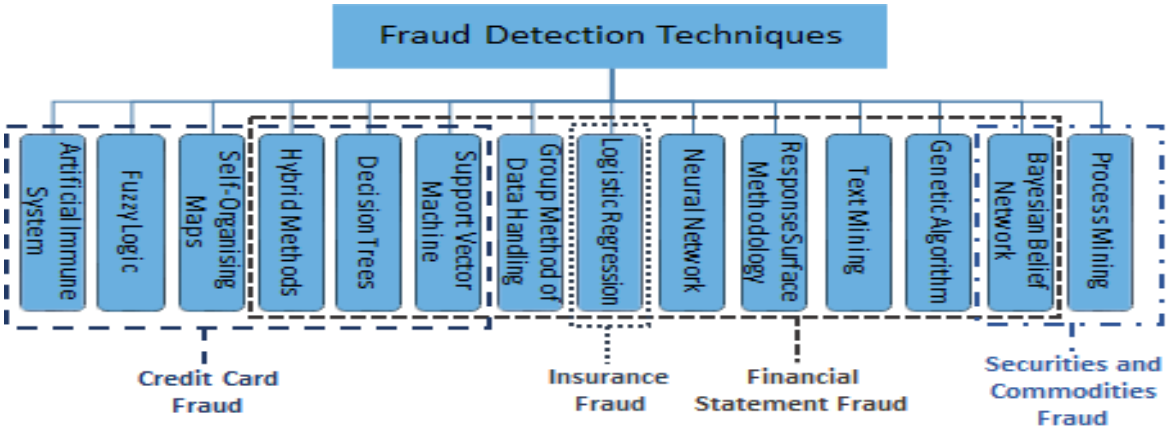


Figure 2.1 -KDD Process in datamining

Authors of "Intelligent Financial Fraud Detection Practices: An Investigation"
[8] in have provided a comprehensive investigation on existing practices and techniques used in financial fraud detection. They have presented a comprehensive classification report by analysing financial fraud detection practices based on several criteria's such as performance, detection algorithm, fraud type etc. by enabling the readers to identify trends in current practices, including which have been successful. [8]

Sevda and Mohammad [9] have presented an analysis for detecting customers behavioural patterns in order to improve the process of fraud detection in financial institutions. They have emphasized that customer's behaviour analysis is very important for providing a better service to them and maintaining the relationship with them. Analysing the behaviours of the customers help to predict their future behaviours and will decrease the risk of fraud. They have chosen clustering techniques to analyse the behaviours of customer groups and each cluster represented a certain type of customer. Suspicious behaviours are detected when the behaviour

account exceed the normal range. Decision tree and neural network algorithms have used for building models to predict the fraud on each cluster group. Model created based on decision tree algorithm has given higher accuracy and speed than neural network. [9]

Author of" Fraud Detection using Supervised Learning Algorithms" [10] has proposed a system which detects credit card frauds using supervised learning algorithms. Author has compared the decision tree learning and naïve Bayes learning algorithms and says that both learning algorithms can be used to detect different types of financial frauds and their ability to learn the different methods of fraudsters is highly effective. In order to minimize the depth of the decision tree before applying on the learning data set Information gain technique has been used. Entropy for each attribute was calculated to select most appropriate attribute to split the tree. By applying these techniques in detecting credit card frauds can predict frauds soon after transactions are done. [10]

One of the major criticism associated with fraud detection based on data mining is not having sufficient publicly available datasets and lack of known methods and techniques.so number of real datasets available for performing experiments are very few. Clifton Phua etal. [11] have done a comprehensive categorisation based on the different approaches including supervised, unsupervised and hybrid methods and techniques. They have also provided alternatives for possible datasets which enables new experiments. Researchers also have highlighted the new ways of fraud detections towards different domains through the research. This helps the new researchers to get an overall description regarding existing data mining based fraud detection techniques. [11].

Anomalies detection is important aspect to be considered in the financial fraud detection. Financial Fraud detection system which is very adaptive for changes has been proposed in the research. [12]. Proposed system is consisting of two stages and identifying the anomalies using BOAT algorithm is the first stage. Checking suspected anomalies with the fraud transaction database is in the second stage. Most of the previous researchers has been choose either supervised or unsupervised learning algorithms for financial fraud detection. Authors have proposed a method which uses a sampling process with various proportions and an automatic attribute selection process by solving the problem of having unbalanced dataset. The proposed fraud detection model based on supervised and unsupervised learning algorithms and gives

more accurate classification result for fraud detection. According to the research applying clustering based unsupervised algorithms and classification based supervised algorithms have given more accurate results for financial fraud detection. [12]

Authors of [13] have proposed and implemented an analytical framework for detecting frauds. Researchers have used ID3 for deciding a specific transaction is fraudulent or not Model is built based on credit card dataset and accuracy of the model evaluated using the confusion matrix. Out of the models which have evaluated model created using random rest given the best result with respect to accuracy, precision and recall. [13].

A two-tier classification model-based solution has been introduced for Financial fraud detection based on Liner Discriminant Analysis(LDA). Researchers have designed the model using three discriminant functions which developed by learning on training and test data. Final classification decision has been taken according to the output given by the functions by evaluating the performance of the model. Model performance also has compared with other approaches for better clarification. By using better optimization techniques this model can be further extended to fit for any type of financial fraud. But the limitation of this model is model only gives best result for linearly separable problems. [14]

1.7 Data Mining Applications in Use for Fraud Detection

Open ML Engine: A service implemented on micro services architecture which includes SDDK for python, R and java languages. This provides integration to commonly used data science and machine learning tools like H20, R studio and data robot.

Card Watch: A system used for credit card fraud detection which has implemented based on neural network machine learning techniques. It provides a very user comfortable GUI and also provide interface to several commercial databases. This fraud detection model with neural network has presented high succession rates in credit card fraud detection. [15]

Data Visor: This system is a predictive analytics based fraud detection solution for banks. This provides risk scores for e-commerce transactions, loan applications and financial services payments. This software has helped one of the U.S' largest bank to detect fraud methods in their loan applications which include stolen and fake identities in customer data.

1.8 Research problem identification

Although literature reveals that all detailed information available under the topic financial frauds are about credit/debit card, online banking, cash and financial statement frauds, we could not found any comprehensive analysis done on techniques and tools which can be used to detect frauds in pawning services. Furthermore, no evidence found about an analysis which describes how datamining techniques can be used for fraud detection in pawning services offered by banking sector. So the goal we try to accomplish through this research is to conduct a comprehensive analysis to find how data mining techniques can be effectively utilized for overcoming above circumstances.

1.9 Summary

This chapter provided a comprehensive review on use of data mining techniques in the domain of financial and banking fraud detection. We have reported the issues related to traditional pawning fraud detection procedures which are currently using in the banking sector. We have defined the research problem and identified the possible technologies addressing the research problem.

CHAPTER THREE

TECHNOLOGY ADAPTED IN PAWNING FRAUD DETECTION

2.1 Introduction

We have discussed different approaches and techniques used in the area of financial and banking fraud detection, its issues and challenges. We also defined our research problem and also identified classification in data mining as a best technological approach to address the problem. This chapter emphasizes how the selected technology can achieve the targeted goal by evaluating different classification techniques in the context of fraud detection.

2.2 What is the data mining?

Today the we all are dealing with large amount of complex data which were generated by different sources such as computers, networks and humans, different institutions, working places, scientific and financial institutions etc. Although large amount of data available everywhere there is a problem in managing them and analyzing those data to use them in effective manner. So identifying hidden data patterns from the large data sources is very important task in this era of technology. That's how datamining came into the screen.

In simple words process of discovering useful hidden data from large set of raw data is known as "Data Mining". Different computer-based technologies are need to apply on the data for analyzing and extracting data patterns from it. Data Mining has wide area of usage in multiple fields such as marketing, finance and medicine. For Example, different product buying patterns of the customers can be used to develop effective ways of storing products in supermarkets. In such a way can make use of different data mining techniques for developing effective business strategies which helps business to improve faster than ever. Data Mining process includes several steps to be followed and also known as knowledge discovery of Data (KDD).

Formal definition for the term data mining is "non-trivial process of identifying valid novel, potentially useful and ultimately understandable patterns in data". The overall process of finding useful patterns in raw data includes clearly identifying the application domain, creating

target data set by integrating several sources, data cleaning and preprocessing, data reduction, choosing the data mining methodology, choosing the data mining algorithm, interpreting identified patterns knowledge. This process of discovering useful knowledge is technically called as “Knowledge Discovery in Databases(KDD)”.

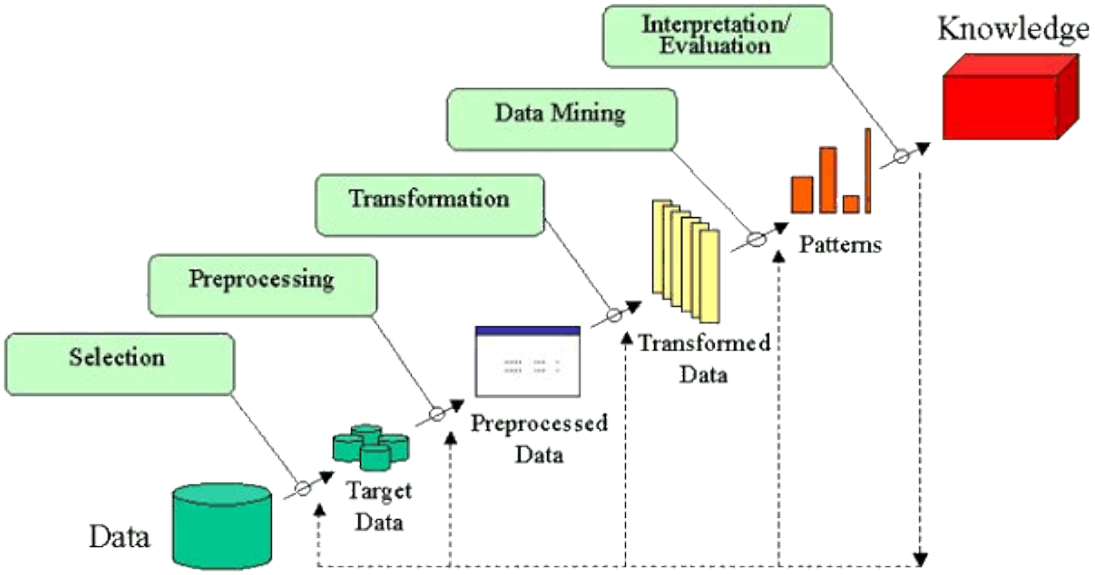


Figure 3.1 -KDD Process in datamining

Steps 1: Clearly identify the application domain and the targeted goals of the end users.

Step 2: Create the target data set by integrating different datasets by focusing on data samples.

Step 3: Clean the data set by removing noisy data or outliers, redundant and unwanted data. Data preprocessing is done by filling the missing values.

Step 4: Data Reduction is done by selecting specific data range depending on the goal. Feature reduction and data transformation is take place in this step to reduce effective number of parameters.

Step 5: Choose the datamining task (classification /Association/Clustering, etc.) according the purpose of the task.

Step 6: Choose appropriate datamining algorithms depending on the data mining task related to the task and apply selected algorithms to the data set. Creating the models and deciding the most effective parameters should be done carefully.

Step 7: Discovering the data patterns using the selected data mining algorithms such as decision trees, regression is done in this step.

Step 8: Presenting the mined patterns in a human understandable format using graphs, pies etc. for decision making purposes.

Data mining has its unique features and successful applications so that identified as interesting and practical method for knowledge discovery among the researchers all over the world. Financial fraud detection, Basket analysis in marketing, Customer relationship management and bio informatics are among the remarkable business applications of datamining today.

2.3 Reasons for choosing classification for pawning fraud detection

Data mining can be classified into major five areas including association rule mining, classification, clustering, regression and sequential patterns etc. According to previous research the classification techniques has wide range of use in financial fraud detection and prediction. Classification techniques have capability of processing a large amount of data and they can be used to predict class labels based on training set or it can be used for classifying newly available data. Simply classification techniques can be applied for any context where we need to do any forecast or take decisions based on previously stored history data. However, this is a very effective way of using history data for decision making purposes with a highest succession rate.

2.4 Why Classification?

Classification techniques which uses a set of data where the output classes are already known are called supervised learning techniques. For example, identifying a disease based on the patient's symptoms to give immediate treatments, predicting the weather (cloudy, sunny, rainy) based on prior knowledge of parameters such as humidity, temperature etc. These techniques have a proven success in handling varies types of real world problems and can be applied to any practical situation very effectively.

As we already know classification techniques are used for predicting an outcome by analyzing the history data. In this research our main target is to predict if a pawning transaction is fraudulent or not by m analyzing the previous pawning transaction data. So considering these

factors, out of the machine learning techniques, classification, clustering and association rule mining we have selected the classification as most appropriate datamining technique for achieving the targeted goal.

2.4.1 Supervised Learning algorithms

Supervised learning algorithms provide a very effective set of functions to process the history data and classify the processed data using learning the data patterns. Here we have to feed previously known labeled data in to the supervised learning techniques. Supervised algorithm then takes the data set and learn the data patterns and built a model. Then the model in trained in order to generate a prediction for a new data or test dataset. There are two important supervised learning algorithms which they used to develop predictive models.

Linear Regression

Linear Regression is a supervised learning technique that have developed to reproduce output value. This method gives an elegant path for determining the descriptors for the parameters in a specific model. This is a leaner model that built a leaner relationship between the one or multiple input parameters (x) and the singe output parameter(Y). When there is a single input parameter this is called as simple liner regression whereas when there are multiple input parameters it's called as multiple linear regression. This Liner regression technique is an attractive technique which has very simple representation.

For example, form a liner regression model would be like following:

$Y=B+B1*x$

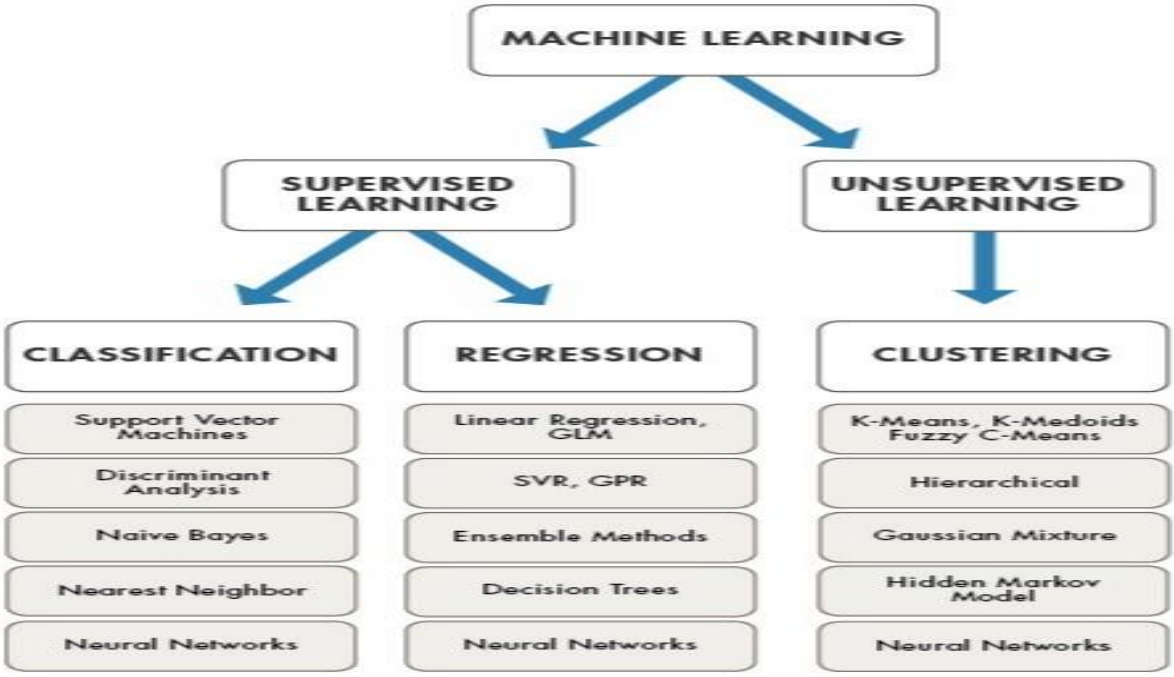


Figure 3.2- Supervised Learning techniques

In classification techniques, classifiers can be either binary or multiclass which are determined according to the number of classes they have. In this research by using the classification algorithms we could get significant results in detecting pawning frauds. Our dataset represents several independent attributes and one dependent attribute. Since our dependent attribute has only two distinct values as fraud and non-fraud binary classification has provided expected result. Under the classification we have several classification algorithms which we can use to get an effective analysis result. Description of several classification algorithm has presented below.

2.5 Classification Algorithms

K'NN algorithm

The KNN (k-nearest neighbors) algorithm is a simple, non-parametric, lazy learning supervised machine learning algorithm that can be used to solve both classification and regression problems. However, it is more widely used in classification problems in the industry. It’s easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows.

In Rapid Miner there is an operator called KNN which generates a K nearest neighbor model from the given dataset. The k-Nearest Neighbor algorithm is based on learning by analogy. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training examples that are closest to the unknown example. "Closeness" means distance metric, for example Euclidean distance or Manhattan distance.

Decision Tree

The Decision tree algorithm used to build classification or regression models in the form of a tree structure. algorithm breaks down data set into smaller subsets while developing a decision tree step by step. It results a tree with decision nodes and leaf nodes where decision node has two or more branches and a leaf node represents a classification or decision. Root node is the topmost node of the tree and it represent the best predictor node.

Decision Tree operator in the rapid minor learns creates decision Tree with decision and leaf nodes. Both nominal and numerical data sets can be applying on this operator. The generated decision tree can be used for classifying unseen examples. This operator can be very efficient when boosted with operators like the Ado boost operator. The dataset can have several attributes and every data row belongs to a class (yes or no). The class name is defined by a leaf node of the decision tree whereas a decision node is defined by a non-leaf node.

Naive Bayes

Naive Bayes is a classification technique based on independence among predictors. Naive Bayes classifier can be used to assume that attribute in a class is unrelated to any other attribute. Despite of the fact that these attributes depend on each other or upon the existence of ant other attribute, all of them are independently contribute to the probability. Creation of model using Naïve Bayes classifier is easy and useful for very large data sets.

Naive Bayes operator in rapid miner generates a Naive Bayes classification model on the given dataset. A Naive Bayes classifier is a probabilistic classifier which provides strong (naive) independence assumptions. The advantage of the Naive Bayes operator is that it only needs a small amount of training data to determine the means and variances of the attributes that are necessary for classification. Because of the assumptions made on independent attributes

variances of the attributes for each class label need to be determined and not the entire covariance matrix.

Random forest

Random forest is an algorithm which generates ensemble model by classifying based on "votes" of multiple trees. Data item is assigned to a class that has most votes from all the trees. As its name implies this algorithm consists of large number of decision trees that operates as an ensemble. Each individual tree spits out a class prediction while the class with the most votes becomes the classification model's prediction. The low correlation between models can produce ensemble predictions that are more accurate than any of the individual predictions. It happens because the trees protect each other from their individual error.

The Random Forest operator in rapid miner generates a set of random trees. The random trees are generated in the same way as how a random tree is generated. The resulting forest model contains a specified number of random tree models. The number of trees parameter specifies the required number of trees.

Bagging

Bagging is an ensemble technique which used to reduces the variance of predictions by combining the result of multiple classifiers modelled on different sub-samples of the same data set This technique is like divide and conquer method. Group of predictive models run on multiple subsets and combined to get a better accuracy and model stability.

Main Steps associated with bagging algorithm are:

1.Creating multiple datasets:

2.Building multiple classifiers: same classifier is built on all the datasets.

3.Combining Classifiers: The predictions of all the individual classifiers combined to give a better classifier

The Bagging operator in rapid miner is a nested operator with a sub process. The sub process must have a learner which generates a model. This operator tries to build a better model using the learner provided in its sub process.

Artificial Neural Networks

Artificial Neural networks are electronic network like structure inspired by the neural structure of the human brain. These are effectively used for high dimensionality problems. But they are theoretically very complex and hard to implement. This is a collection of input units called neurons connected with each other and has a weight associated with it. Those neurons are activated via a function and divided into several layers. Each layer takes the input variables and passes the results into next layer. A neural network algorithm is capable of learning any complex model with computational power.

Neural Net operator in the rapid minor learns a model by using feed-forward neural network trained by a back propagation algorithm. Feed-forward neural network is also an artificial neural network where connections between the input units do not create a directed cycle.

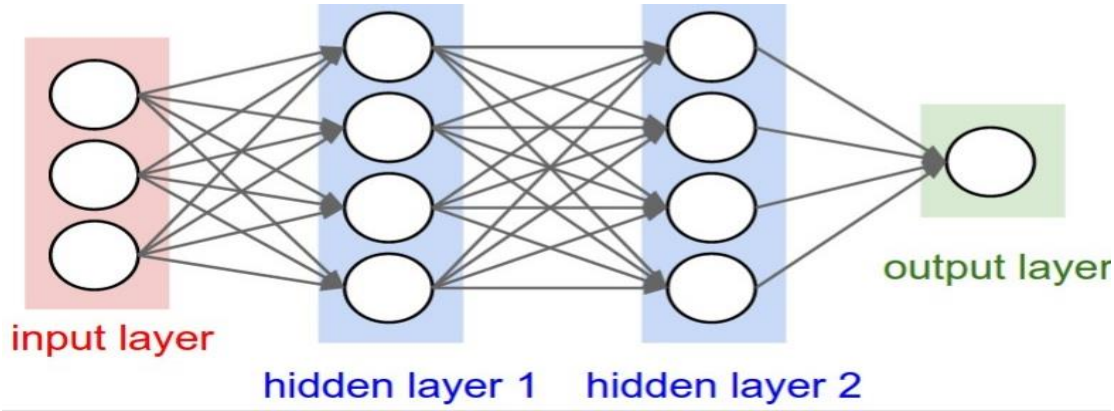


Figure 3.3- Structure of Artificial Neural networks

Support Vector Machines (SVM)

Support vector machine technique is used to find the hyperplane in an N-dimensional space which classifies the data points. Hyperplane is a boundary which separates data points into several classes. To divide the data points in to classes there are many hyperplanes available which can be chosen. Support vectors are kind of data points which are located nearby the hyperplanes and supports to change the position of hyperplane. These support vectors are the data points which uses for building support vector machine models.

Support Vector Machine operator in the rapid minor provides a fast algorithm and good results for many learning tasks and supports different kernel data types including dot, radial, polynomial, neural, anova, epachnenikov, gaussian combination and multiquadric.

2.6 Microsoft SQL SERVER

Microsoft SQL SERVER is a relational database management system provided by Microsoft cooperation while Microsoft SQL SERVER Management Studio is the IDE provided with SQL SERVER for configuring, managing, and administering all components within SQL Server Database in client side. Pawning Transactions were recorded in a SQL SERVER Database and using SQL server management studio required data set was exported into Excel Worksheet.

2.7 Rapid Minor Studio

Rapid Miner is an application which provides an integrated environment for designing visual data science workflows to build machine learning, data mining, text mining, predictive analytics and business analytics models. It is commonly used for business and commercial purposes as well as for research, education, training and prototyping. Application also supports all steps of the data mining process including data pre-processing, model creation, validation and optimization. The Rapid Miner (free) Basic Edition is limited to 1 logical processor and 10,000 data rows is available under the AGPL license. Huge set of operators available in all tasks of data transformation and analysis.

2.8 Summary

This chapter presented data mining as the technology proposed to accomplish the goal of detection of frauds in pawning services offered by banking sector. Detecting pawning frauds in real time using data mining techniques is very effective for preventing the pawning frauds. In this approach rapid minor used for data analysis and SQL server for data storing. This chapter also gives detailed description on how the classification data mining algorithms offers efficient and accurate solution for fraud detection pawning services. Out of the different classification algorithms KNN, Naive Bayes, Decision tree, Bagging and random forest are selected for further analysis in upcoming chapters.

CHAPTER FOUR

METHODOLOGY FOR DETECTING AND PREDICTING FRAUDS IN PAWNING SERVICES

3.1 Introduction

In this chapter we have presented the approach we have taken to address the research problem. Furthermore, we described which techniques we have used to solve the targeted problem of detecting the frauds in pawning services offered by banks. We present our approach by highlighting hypothesis, input parameters, output, processes and datamining techniques used to achieve the underlined goal.

3.2 Hypothesis

We hypothesis that we can use classification techniques in datamining for solving the problem of not having proper mechanism to detect frauds in real-time in pawning services offered by banking sector. Idea behind choosing the hypothesis was influenced by the facts that the increasing number of banking frauds related to pawning services and issues related to the traditional fraud detection mechanisms used in banks.

The hypothesis of this research is that the classification algorithms in data mining can be used to detect the pawning frauds real time in banking sector. We have used several classification techniques including KNN, Naïve Bayes, Decision tree, Random Forest and Bagging and finally choose the best method out of them.

3.3 Input

We have collected pawning transaction data of last 10 years which covers about 100 branches from SQL server database by querying the database though SQL server management studio. We have selected two different sets of parameters from the pawning dataset and applied the same datamining techniques on both sets for analysis purposes.

3.4 Output

Main output of this research will be a simulation for detecting the pawning frauds which can be utilized by the existing pawning applications used in banks. Different classification models will be created using the selected classification algorithms and after testing each of the model using testing set, the model with highest accuracy will be selected. Selected algorithm could be used for predicting fraudulent transactions in pawning services in banking sector.

3.5 Process

In our research, main process is to identify the most appropriate classification algorithm for detection frauds in pawning service offered by banks. In this process several sub processes are also included. Throughout the process for the purpose of fraud data analysing, several classification techniques have been used. Data pre-processing, selection and evaluation are another sub processes that are included in the main process.

3.5.1 Data collection and selection

In order to determine the frauds in pawning services offered by banks we need to collect set of pawning transactions. For that purpose, pawning transaction data were collected from a bank database by querying the database through SQL server management studio.

Initially selected 10000 records with fraud and non-fraud transactions. The dataset we choose need be a balanced dataset in order to gain a correct output from the analysis. Since we have only limited fraud transactions and after preprocessing data set up to 2000.Our final data set reduced up to 1800. Final dataset includes 923 non fraud and 877 pawning fraud transactions.

All most all of the banking institutions and other financial institutions offer pawning services to customers in present. They get customer information such as customer identification number, address and the articles related data such as article, article type. Most common attributes related to the pawning services in banking sector includes customer address, Customer Id, Purpose, Pawning ticket no, Interest rate, Article weight, Article Karat content and Amount granted. Our dataset has 14 attributes including those common attributes.

3.5.2 Data Pre-processing

In this research pawning transaction dataset gathered from a bank is used as the primary data source which includes pawning transaction data from 100 branches in the country. Normally it's obvious that real-world transaction data is incomplete, inconsistent or have extra ordinary behaviors or trends.

Noise in data is major issue related with real world data which reduces the quality of data. Noisy data may come from human or computer error at data entry, and errors in data transmission. So the transaction data have many errors associated with them including user typing errors and software errors. These errors have direct impact on the quality of the data. Another issue is inconsistent data which may come from functional dependency violation in linked data. Duplicate records also need data cleaning which lead to poor data quality.

Accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility are some of the characteristics that should be associated with data taken to a research to draw a well-accepted conclusion. Data Preprocessing is an important step that must take before the any other steps begin. This step involves transforming data consists of noisy and unwanted data into an understandable format. In this research have applied few data preprocessing techniques in order to make the dataset complete and consistent.

3.5.3 Attribute Selection

Attribute selection is a method which is used for data reduction in datamining as a pre-processing technique. In data reduction it is obvious that it reduces the data and keep only the relevant data according the problem domain since it increases the effectiveness of the datamining process. Attribute selection is very important as it reduces the no of attributes to those that are believed to be most relevant to a model to predict the target class variable. So, in our research we should consider attribute selection is an essential step we must follow before proceeding further because having too much attributes in the data set makes the analysis task complex and difficult to understand. The goal of the attribute selection techniques is to discover most relevant minimum set of attributes in order to reduce the cost of data analysis and improve the performance of the model as well. In the pawning transaction dataset out of the 16 attributes

need to select most relevant attributes towards detection of frauds. In next chapter we will present the techniques we have used to select attributes which are most appropriate for the detection of pawning frauds.

3.6 Data mining

Datamining is the most essential section of the overall research in which we use data mining concepts practically on the dataset in order to extract useful data patterns. Not only finding hidden patterns, it includes finding associations, anomalies from the dataset. In Association rule mining, the relationship of a particular item with other items in the same transaction is used to predict patterns. In Classification around 20 methods are embedded for learning different models that map each data item into one of the predefined classes.

Predictive mining is another important concept associated with datamining. Key idea behind the predictive mining is to discover relationship between independent and dependent variables. In sequential pattern analysis it searches for similar patterns over a period in transaction data. These patterns can be used to identify relationship between data for decision making purpose.

Descriptive cluster analysis is another aspect in data mining which takes unorganized data and organize them into groups using mathematical model. We should select either predictive or descriptive data mining according to the goal we are trying to achieve through the research.

3.7 Users

Pawning and internal audit officers in government and private banks can gain help of this pawning fraud analysis and pay attention to patterns represented by the mining process. By this analysis they can use this as an advanced and on time method to detect frauds via the applications.

3.8 Features

The simulation addressed by this research can be used to analyse huge volume of pawning transaction data which can be collected from different government and private banking institutions. Classification (Predictive mining) techniques allows to make predictions for future transactions in order to detect fraudulent ones. This solution provides the output by extracting

previously unknown patterns in pawning transactions while facilitating to check any pawning transaction is going to be fraudulent or not at the run time.

3.9 Summary

This chapter presented our approach to analyse pawning transactions in order to identify the frauds associated with them. In this sense, it is pointed out how this methodology offers an efficient and accurate solution for pawning fraud detection using data mining. The next chapter presents the design of our approach presented in this chapter.

CHAPTER FIVE

RESEARCH DESIGN AND ANALYSIS

4.1 Introduction

Chapter five presents the approach to analyze the use of different datamining techniques to detect pawning frauds in real-time. This chapter emphasizes our approach and focuses on high level design and sub modules within in the design.

4.2 Research Design

In this research we have used datamining for analyzing the pawning transactions for discovering hidden patterns among them to identify the fraudulent ones as earlier as possible. We have followed the steps in scientific research method to achieve our goal successfully. Main steps in the scientific research approach are observation, background study, constructing hypothesis, testing hypothesis draw conclusions, reporting and evaluating.

Our research study starts with doing observation and background study regarding the existing problems in pawning fraud detection procedures in banking sector and limitations in previous research work under this topic. After that study was proceed to find out the solutions that have been used to solve the similar kind of problems and how data mining was used as a solution for similar problems. In the next step we have tried to get information about the influencing factors related with pawning frauds and identified the parameters which could affect most for detecting pawning frauds. Study was carryout further until selecting sub research questions and After selecting sub research questions to be addressed, parameters or factors have recognized by observing the related work in that problem domain. Then we have developed the hypothesis and it was tested using the dataset collected using rapid minor studio tool. Different classification techniques are used to test the hypothesis using two different parameter sets. The results generated by the rapid minor studio using different classification techniques are further analyzed considering the performance or confusion matrix results. Final conclusions were developed using those results.

4.3 Detailed Research Design

Analyzing the issues and limitations in traditional pawning fraud detection in banks has identified as the primary research question of this research study. According the annual financial reports, profits earned from pawning in banking sector has been increased in last decade and however frauds happened in those pawning services in banks have major impact on the financial status of the banks.

4.3.1 Sub Research Question One

What are the existing mechanisms and procedures used in banks to identify the problems and influential factors associated with pawning frauds?

4.3.2 Sub Research Question Two

What are the parameters which are most related to the frauds happened in pawning services?

4.3.3 Sub Research Question 3

How does classification in data mining can be the most appropriate solution for detection of pawning frauds real time in banks and in which way we can use them?

4.4 Summary

This chapter has provided details on research design and analysis for the research. Furthermore, this chapter focused on detailed design for the research and how primary and sub research questions are structured within the research. Next chapter will be discussing about the implementation details according to this design.

CHAPTER SIX

IMPLEMENTATION

5.1 Introduction

In chapter six presents the overall implementation through giving solutions for of each sub research questions regarding the analysis, algorithms, methods we have used. Implementation of the solution has been described in detailed in terms of how the proposed design is implemented, what attributes are used to detection of frauds and which algorithms and techniques are used.

5.2 Solution for Research Question One

As the solution for the first sub research question existing fraud detection mechanisms and procedures used in banks has been analyzed in order to identify the problems and influential factors associated with pawning frauds. Findings of the analysis are presented in the chapter.

5.3 Solutions for Research Question Two

As the solution for this research question different feature selection techniques were used and compared the weight given for each attributes to select the most related attributes for the class attribute fraud.

5.3.1 Data Preprocessing

Data pre-processing is an essential step we should follow before establishing a theoretical framework using the parameters. Removing noisy and inconsistent data is the main purpose of the data pre-processing. Data cleaning, data integration, data transformation discretization are main techniques associated with data pre-processing.

In this research in the pawning transaction database there were some missing data which have occurred due to user typing errors. Consequently, tuples had insignificant value for attributes such as district, first granted amount, weight article, net gold. Also there were attributes with missing values. In this research in order to solve this issue ignoring the tuple or using a global constant to fill in the missing values have been used. Data transformation has been done for better organization of data by categorization of several nominal data and real data into integer value. For example, gender, status, purpose and pay method attribute values were categorized

into integer values from 1 to 5. City attribute was categorized into districts denoted by integers from 1 to 25. Attribute or feature construction is used for new attributes construction from the given attributes. As an example, first granted date and last paid date attributes were combined to create new integer attribute called activeloanmonths. In addition to data cleaning and transformation steps data reduction methods also applied to select most relevant data for the model creation. In order to accomplish that attribute selection methods were applied and found out most relevant attributes for the class attribute "Fraud" in the pawning transaction dataset.

5.3.2 Attribute Selection

In this research for the data model construction, correlation matrix is used to select the most relevant independent attributes out of the 16 attributes. Attribute "Fraud" is the dependent variable which is nominal. In our dataset there is 14 attributes and Overall attribute description presented below.

Class Attribute: Fraud (0: No Fraud 1: have fraud)

Input Attributes

ActiveLoanMonths: Number of months from first granted date to last paid date

Gender: gender of the customer (values →1,2)

District: Living district of the customer (values →1-25)

Purpose: Purpose for borrowing the gold loan (values →1-5)

WeightArticleCat: Total Weight of the articles pawned (values →1,2,3)

NetGoldCat: Total Gold weight of the articles pawned (values →1,2,3)

KaratContent: Quality of the gold in the articles (values →1-9)

ValueAssessedCat: Assessed gold value for the articles (values →1-4)

ValuemarketCat: Market value for the gold contained in the articles (values →1-4)

FirstGrantAmountCat: Gold Loan Amount granted to the customer (values →1-4)

AmountGrantedCat: Balance loan amount (values →1-4)

Pay method: pay method (values →1,2)

Status: current status of the loan (values →1-4)

RateOfInt: interest rate for the gold loan

5.3.3 Correlation Matrix

The correlation matrix is table which represents the correlation coefficients between set of attributes. It can produce a weights vector based on these correlations. Correlation is a statistical technique which shows how strongly pairs of attributes are related. This allows to see which attribute pairs have the highest correlation.

Correlation is a number which ranges between -1 and +1 and indicate the degree of association between two attributes. Positive correlation number implies that the two attributes have positive relationship while as a negative number implies a negative one. Diagonal of the table is always a set on ones since correlation between attribute and itself is always one.

In this research we have used correlation matrix for selecting the attributes which have strongest correlation with the class attribute. Correlation matrix operator in rapid minor studio used to measure correlation between the dependent and independent variables in the dataset. Attributes with correlation values higher than 0.1 have selected for the further analysis. Attributes selected by considering the correlation values are gender, district, purpose, karatcontent, status, paymethod and rateofint. Table 6.1 presents the description of attributes selected by weight by correlation matrix technique. Please refer appendix B for the correlation values gained by the correlation matrix.

Table 5.1 - Attributes selected by correlation matrix

Attribute	Type	Values
Gender	Integer	Female →1 male →2
District	Integer	1-25
Purpose	Integer	Housing → 1 Agriculture → 2 Commercial → 3 Construction →4 Industrial → 5
KaratContent	Integer	24KT →1 23KT→2 22KT →3 21KT→ 4 20KT→5 19KT→6 18kT→7 17KT→8 16KT→9

Status	Integer	Time Expired→ 1 Redeem → 2 UnRedeemed →3
Paymethod	Integer	Cash →1 Repawning→ 2
RateofInt	Integer	0.14,0.15,0.16,0.17,0.18

5.3.4 Feature Weights

Different feature weight calculation methods can be used to measure the relevance of the attributes with respect to the class attribute. These methods calculate the relevance by using different techniques such as information gain, Gini index, support vector machine(SVM), ... etc. Values of relevance of each attribute towards the class attribute are assigned as weight and higher the weight of an attribute, the more relevant it is considered.

There are set of operators available in rapid minor studio to measure the relevance of the attributes and we have used weight by information gain operator to find out most relevant attributes as well. Attributes selected by considering the correlation values are activeloanmonths, gender, district, purpose, karatcontent, status, paymethod and retrofit. Table 6.2 presents the description of attributes selected by weight by Information Gain technique.Please refer appendix b for the values gained by the weight by Information gain technique.

Table 5.2 -Attributes selected by weight by Information Gain

Attribute	Type	Values
Gender	Integer	Female →1 male →2
District	Integer	1-25
Purpose	Integer	Housing → 1 Agriculture → 2 Commercial → 3 Construction →4 Industrial → 5
KaratContent	Integer	24KT →1 23KT→2 22KT →3 21KT→ 4 20KT→5 19KT→6 18kT→7 17KT→8 16KT→9
Status	Integer	Time Expired→ 1 Redeem → 2 UnRedeemed →3
Paymethod	Integer	Cash →1 Repawning→ 2
RateofInt	Real	0.14,0.15,0.16,0.17,0.18
ActiveLoanMonths	Integer	1-24

5.4 Solution for Sub Research Question 3

As the solution for the sub research question 3 different classification data mining techniques has been applied on the two different data sets with selected attribute sets.

5.4.1 Model Creation

Rapid minor studio has a collection of machine learning algorithms for data mining tasks. Holds tools for classification algorithms which we can use for creating data models. Rapid minor studio has tools for data analysis tasks such as classification, regression, clustering, association rules, and visualization.

For this research we have used five different classification data modelling algorithms and created data models separately. Bagging, decision tree, KNN, naiveBayes and randomforest are the different algorithms selected for this research. Before applying the selected classification techniques dataset was spitted into two parts using 'SplitData' operator, training set and testing set. Then cross and split validation operators were used for training and testing the data models.

For the analyzing purposes, two datasets with the two attribute sets mentioned in above section are processed separately by applying same classification technique. Stratified sampling technique was used for dividing the subsets since it ensures that the class distribution in the subsets is the same as in the whole dataset. Data models were created and tested using both cross and split validation techniques to encounter a further analysis.

5.4.1.1 Cross Validation

Cross validation is one of the statistical method that used in data mining to estimate the performance of the machine learning models. Cross validation use a parameter called K fold that refers to number of subsets that a given dataset s split into. This is generally used to estimate how the specific learned model will perform when it used to make predictions on data not used during the training the data model.

Cross validation Steps:

1. Shuffle the datasets
2. Split data sets in to K number of subsets

3. For each unique subset: keep the dataset as test set and use remaining datasets to train the data model.
4. Build a model on training set and test it in the test dataset.
5. Get the performance score and discard the model
6. Summarize performance of the model using each subset evaluation scores.

This validation technique is popular as its simple to understand and generally gives less biased estimate regarding the performance of the model.

5.4.1.2 Split Validation

Split validation performs a simple validation by randomly splitting the dataset into training set and test set and evaluates the model. In rapid minor split validation has a parameter called split which denotes how the split should be done. Either relative or absolute split can be selected by determining split ratio or training/test dataset sizes. Split ratio specifies the relative size of the training set.

The Split Validation can also use several types of sampling techniques for building the subsets. Split validation usually used to estimate the performance of a learning model on unseen data sets. However, it is mainly used to estimate how accurately a model will perform in real cases.

5.5 Summary

This chapter presented the full process in constructing data models for addressing research sub questions. Furthermore, this chapter gives detail description about using rapid minor studio tool to build the data model and to perform attribute selection. Next chapter will be on discussion about evaluation and conclusion.

CHAPTER SEVEN

EVALUATION

6.1 Introduction

The chapter seven discusses the details of the results got from the analysis done for detecting pawning frauds using the data mining techniques mentioned in the proposed solution. This chapter justifies and evaluates the overall solution, data mining techniques and data models used in our research.

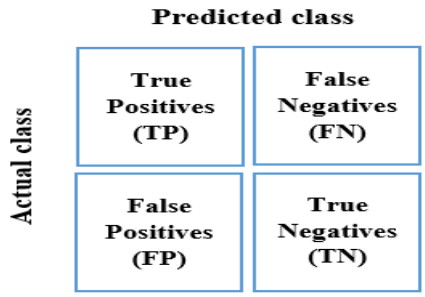
6.1.1 Evaluation for Classification

As a solution for our research problem we have trained collected data set using five different classification techniques namely KNN, Naïve Bayes, decision tree, random forest and bagging with the help of rapid minor studio tool. After building different models using those algorithms they have tested via the data set we have and generated performance of each algorithm in order to compare the algorithms we have.

For evaluating the classifier quality, we can use confusion matrix which can be used to evaluate the performance of the classifier using various measurements such as accuracy, recall and precision.

6.1.2 Confusion Matrix

Confusion matrix is kind of summery table which includes prediction results of classification problems. Both correct and incorrect predictions are summarized with count values and broken down by each class. Confusion matrix also given an insight about errors made by classifier and types of errors that are being made. Figure 7.1 presents a simple confusion matrix for two class classification model.



Measure	formula
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Precision	TP/ (TP+FP)
Recall	TP/ (TP+FN)
F-Measure	2*Precision*Recall/ (Precision+Recall)

Figure 7.1- Confusion Matrix

1. True Positive(TP)

Predicted positive and it's true.

A transaction is predicted as fraudulent and actually it is.

2. True Negative(TN)

Predicted negative and it's true.

A transaction is predicted as non-fraudulent and actually it is not.

3. False Positive: (FP)

Predicted positive and it's false.

A transaction is predicted as fraudulent and actually it is not.

4. False Negative: (FN)

Predicted negative and it's false.

A transaction is predicted as non-fraudulent and actually it is.

Table 6.1 Classification Evaluation Measurements

Measurement	Formula	Description
Precision	TP/(TP + FP)	The percentage of positive predictions those are correct.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive.

Specificity	TN / (TN + FP)	The percentage of negative labeled instances that were predicted as negative.
Accuracy	(TP + TN) / (TP + TN+ FP + FN)	The percentage of predictions those are correct.

Table 7.1 demonstrates the different classification model evaluation measurements associated with confusion matrix and how they are calculated.

6.2 Evaluation Results

The classification results of each technique for datasets containing two different attribute sets which we have defined in previous chapter are given in following tables.

Dataset1→ Dataset with attributes selected by correlation matrix

Dataset2→ Dataset with attributes selected by Information gain

1. Data set 1 classification Results – Cross Validation

Table 6.2 - Comparison of classification techniques dataset1 cross validation

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	96.35	95.08	97.56	32.60	96.30
Naïve Bayes	86.98	89.47	83.06	63.41	86.14
Decision Tree	97.14	96.02	97.23	47.22	96.62
RandomForest	95.16	98.25	91.69	83.60	94.65
Bagging	95.87	96.99	94.46	65.38	95.70

2. Data set 1 classification Results – Split Validation

Table 6.3 - Comparison of classification techniques_dataset1 split validation

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	96.56	96.22	96.74	46.15	96.47
Naïve Bayes	89.95	89.75	90.22	47.36	89.73

Decision Tree	98.68	98.38	98.91	40	98.64
RandomForest	97.88	100	95.65	100	97.97
Bagging	97.62	96.79	98.37	33.33	97.57

3. Data set 2 classification Results – Cross Validation

Table 6.4 - Comparison of classification techniques dataset2 cross validation

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	96.90	95.71	98.05	30.76	96.86
Naïve Bayes	90.24	91.54	88.11	59.34	89.79
Decision Tree	97.78	97.72	97.72	50	97.72
RandomForest	96.59	98.15	94.79	74.41	96.44
Bagging	96.27	97.97	94.30	74.46	96.09

4. Data set 2 classification Results – Split Validation

Table 6.5 - Comparison of classification techniques dataset2 split validation

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	95.50	95.63	95.11	52.94	95.36
Naïve Bayes	93.12	92.47	93.48	46.15	92.91
Decision Tree	97.35	98.88	95.65	80	97.23
Random Forest	97.09	100	94.02	100	96.91
Bagging	98.15	98.36	97.83	57.14	98.09

Table 7.2 shows that ‘Decision Tree’ model represent highest accuracy. Its accuracy of predicting incoming transaction is fraud or not comes to be 97.14 %. Accuracy for KNN has second highest value as 96.35 %. Rapid Minor model creation summaries attached into Appendix D.

Table 7.3 shows that 'Decision Tree' model represent highest accuracy. Its accuracy of predicting incoming transaction is fraud or not comes to be 98.64% which is better than Random forest with accuracy 97.88%. Rapid Minor model creation summaries attached into Appendix D.

Table 7.4 shows that 'Decision Tree' model represent highest accuracy. Its accuracy of predicting incoming transaction is fraud or not comes to be 97.35 %. KNN model has given second highest accuracy as 96.90%. Rapid Minor model creation summaries attached into Appendix D.

Table 7.5 shows that 'Bagging' model represent highest accuracy and 'Decision Tree' model taken the second highest value as 97.35%. Its accuracy of predicting incoming transaction is fraud or not comes to be 97.35 %. Rapid Minor model creation summaries attached into Appendix D.

According to the analysis done using two different parameter sets, it was found that "Decision Tree" produced the best results in predicting fraudulent transactions in most of the test cases. Random Forest, KNN and Bagging algorithms also performed well in our research. Several comparisons done on how algorithms has performed with respect to the accuracy, precision and recall measures are presented as graphically below.

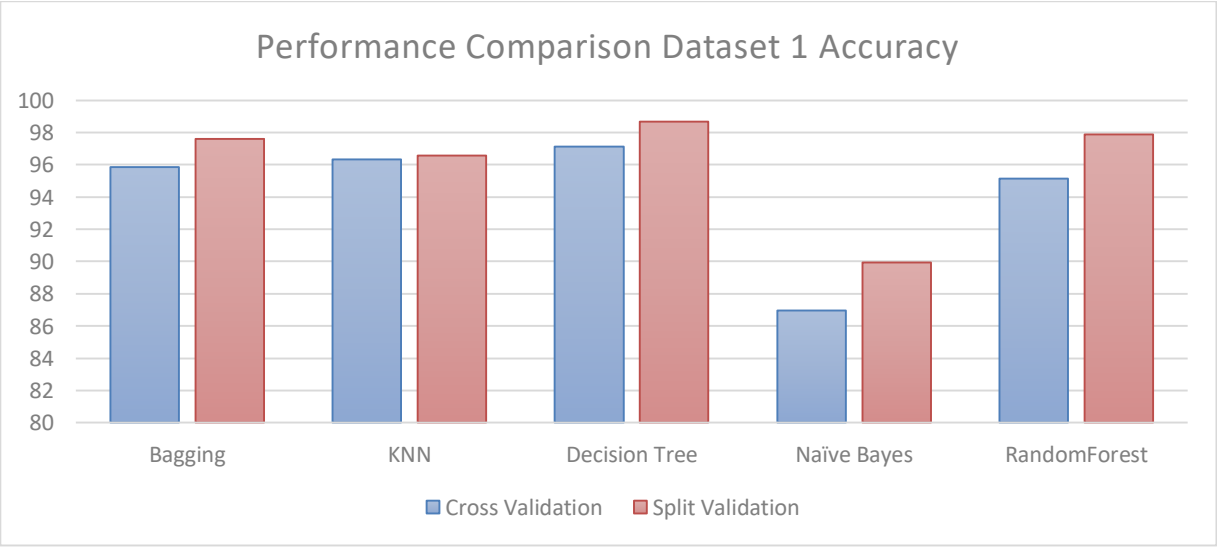


Figure 7.22 -Algorithm Performance comparison for Dataset 1 -Accuracy

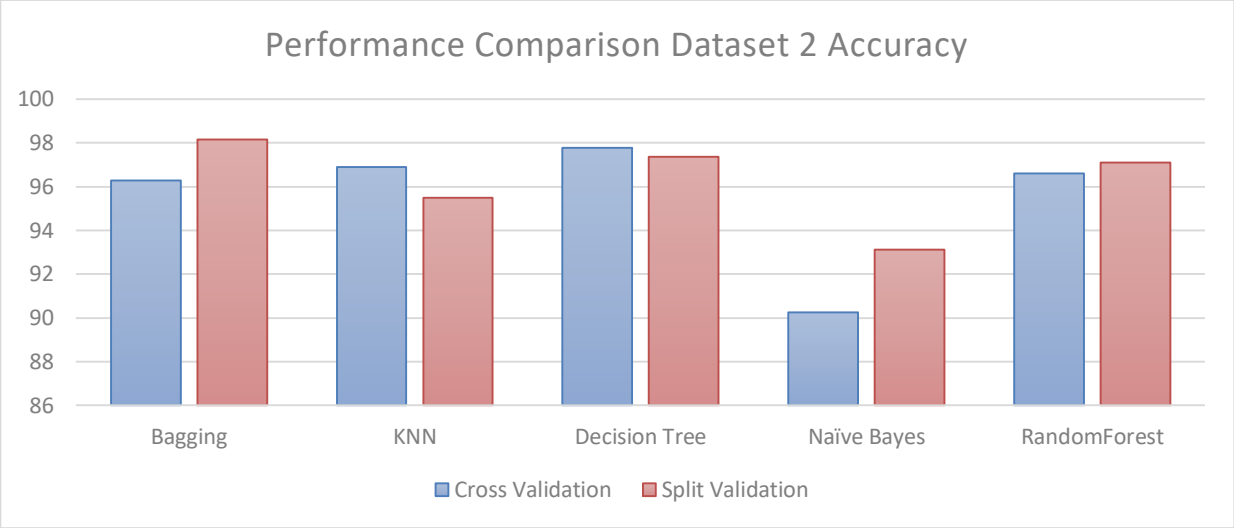


Figure 7.23 - Algorithm Performance comparison for Dataset 2 -Accuracy

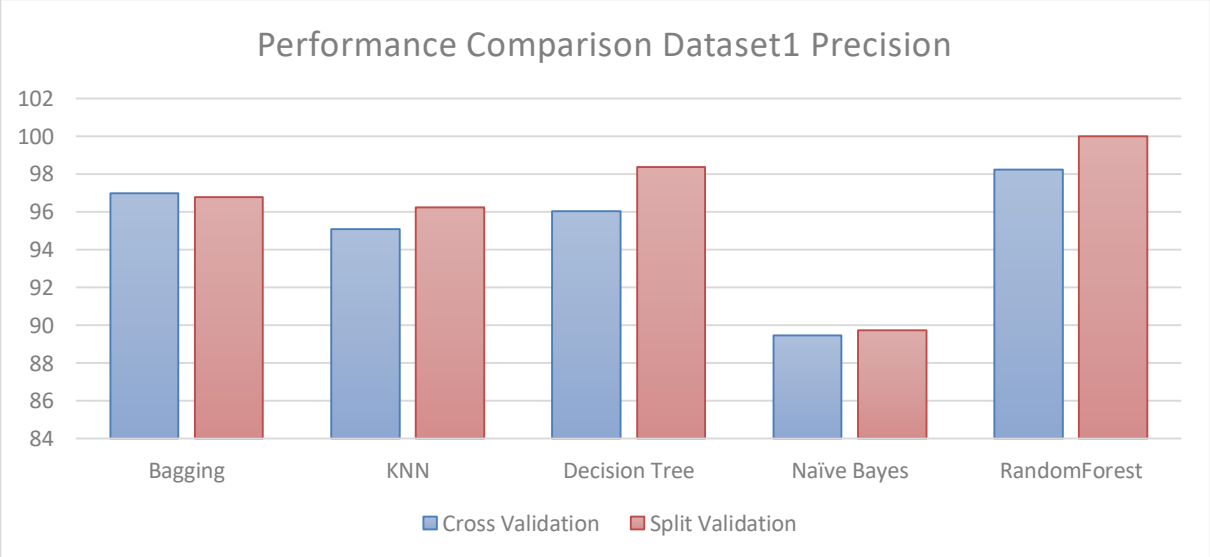


Figure 7.24- Algorithm Performance comparison for Dataset 1 –Precision

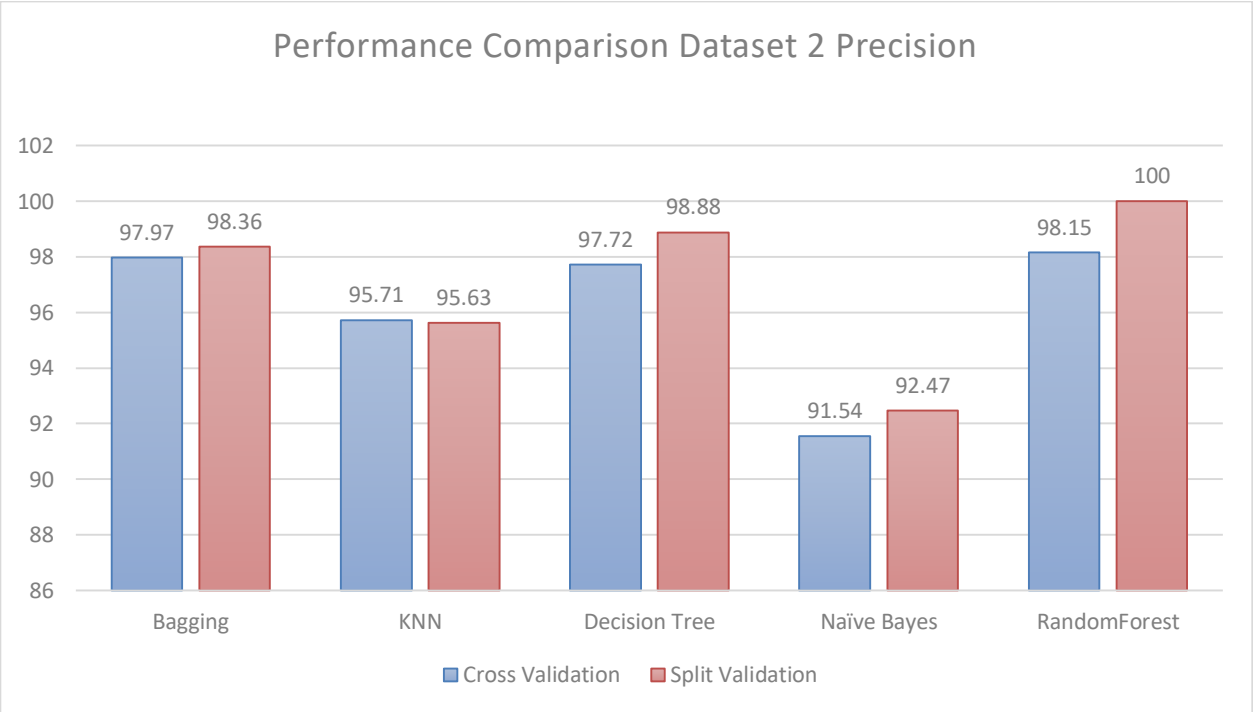


Figure 7.25-Algorithm Performance comparison for Dataset 2 –Precision

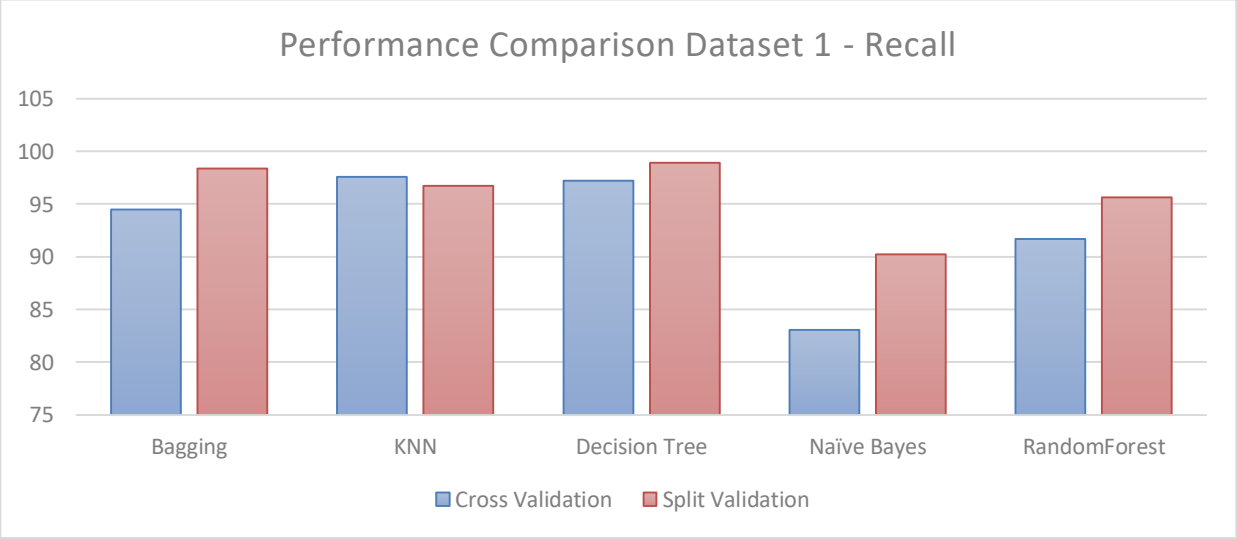


Figure 7.26- Algorithm Performance comparison for Dataset 1 –Recall

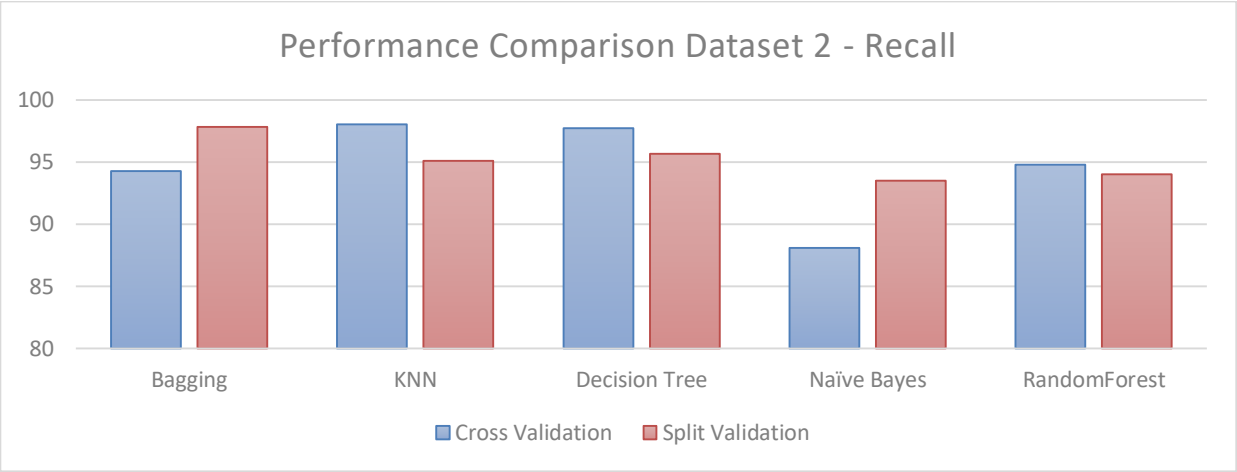


Figure 7.27- Algorithm Performance comparison for Dataset 2 –Recall

6.3 Summary

This chapter evaluated the methodologies and the results discussed in the implementation chapter. This chapter also clearly presented the performance results of each algorithms on different parameter sets. Next chapter discuss limitations and future improvements of the research work.

CHAPTER EIGHT

CONCLUSION AND FUTURE WORK

9.1 Introduction

This chapter provides an overview for the research. It describes how we provide the solution to addressing the problem of detecting pawning frauds in banking sector by using real pawning transactional data. Furthermore, this chapter focuses of limitations and further work of this research.

9.2 Overview of the research

The main goal of this research is to learn a model for predicting frauds happening in pawning services offered by the banking sector. We have selected most relevant attributes using different feature selection techniques and applied classification algorithms on the datasets.

Different fraud learning models were built using those classification algorithms and compare their performances to determine the most appropriate one. Decision tree algorithm is selected out of five classification techniques since it presented highest performance. So decision tree algorithm to predict fraudulent transactions in a given situation. We demonstrated the quality and accuracy of our predictions with an extensive set of experiments and testing done on real pawning transaction data. So, fraudulent transaction prediction can be much help to make strategic decisions to increase the possibility of detecting frauds as soon as possible. The main contributions of our work listed as follows.

- Comparison of five different machine learning classification techniques which revealed that using decision tree algorithms best approach to solve the problem.
- Evaluation of various classifiers over real pawning transaction data which predict frauds in pawning services.

With proper tuning, decision tree classification model simulated by this research can be utilized further for using it by any financial institution which offers pawning services.

9.3 Limitations

This is a secondary research where we use classification data mining as a solution to explore the fraud patterns in pawning transaction data. Dataset includes transaction data from only limited number of banks. Unavailability of sufficient number of fraudulent transactions is a major limitation in this research. If we could have more fraudulent transactions from different banking institutions, we'll be able to discover more hidden fraudulent patterns and improve the effectiveness of the data models.

Parameters were selected considering only the available dataset from a limited number of banking institutions. So the research problems which we can focus on related to the parameters is limited according to the unavailability of sufficient data from different institutions.

We have used only classification data mining techniques for the analysis. According to the previous research work we have concluded that classification is the most used data mining technique when it comes to predictive modelling such as predicting fraud detection. But we can use other type of data mining techniques such as clustering for identifying the distinct clusters within that data set.

9.4 Future work of the project

We can improve the outcome of this research work by using pawning transactions data from both state and private banks for predicting pawning frauds. We can use decision tree algorithm as a base to implement a plugin or tool to predict the pawning frauds which can be used by any banking institution in Sri Lanka. Hence, this research can be extended further to address different issues in Sri Lankan lifestyle.

9.5 Summary

This chapter concludes the thesis by describing the solution given with data mining to analyze pawning transaction data and how it can be enhancing further to improve the level of accuracy in predicating /exploring predicting pawning frauds on time.

REFERENCES

1. Ankur Rohilla, “Comparative Analysis of Various Classification Algorithms in the Case of Fraud Detection”, International Journal of Engineering Research & Technology (IJERT), Vol. 6 Issue 09, PP. 2017

2. Hossein Hassani, Xu Huang and Emmanuel Silva, “Digitalisation and Big Data Mining in Banking”, Big Data Cogn. Comput. 2018

3. Dingdong Zhang and Lina Zhou,” Discovering Golden Nuggets: Data Mining in Financial Application”, IEEE Transactions On Systems, Man, And Cybernetics, Vol. 34, No. 4, 2004

4. Metrix, Fraud Detection, Notch Solution, Inc., Charlotte, NC.

5. Sen, Sanjay Kumar, and Sujata Dash. "Meta Learning Algorithms for Credit Card Fraud Detection.”, Meta 6.6 (2013): 16-20.

6. Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier.”, Procedia Computer Science, Pages 679-685(2015).

7. Aastha Bhardwaj and Rajan Gupta,” Financial Frauds: Data Mining based Detection – A Comprehensive Survey”, International Journal of Computer Applications (0975 – 8887) , Volume 156 – No 10, December 2016

8. Jarrod West, Maumita Bhattacharya and Rafiqul Islam,” Intelligent Financial Fraud Detection Practices: An Investigation”

9. Sevda Soltaniziba, Mohammad Ali Balafar,” The Study of Fraud Detection in Financial and Credit Institutions with Real Data”, Computer Science and Engineering 2015, 5(2): 30-36

10. R. Mallika,” Fraud Detection using Supervised Learning Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2017

11. Clifton Phua, Vincent Lee, Kate Smith, & Ross Gayler,”A Comprehensive Survey of Data Mining-based Fraud Detection Research”,

12. Dahee Choi and Kyungho Lee," An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation", Hindawi Security and Communication Networks ,Volume 2018

13. Suraj Patil*, Varsha Nemade, PiyushKumar Soni, "Predictive Modelling For Credit Card Fraud Detection Using Data Analytics", International Conference on Computational Intelligence and Data Science ,2018

14. Fazlul Hoque, Md. Jahidul Islam, Swakkhar Shatabda," A Two-Tier Classification Model for Financial Fraud Detection", International Journal of Computer, May 2015

15. E.Aleskerov,B.Freisleben and B.Rao ,"CARDWATCH: a neural network based database mining system for credit card fraud detection",IEEE/Computational Intelligence for Financial Engineering(IAFE) ,1997,pp.220-226

Appendix A

Data Preprocessing

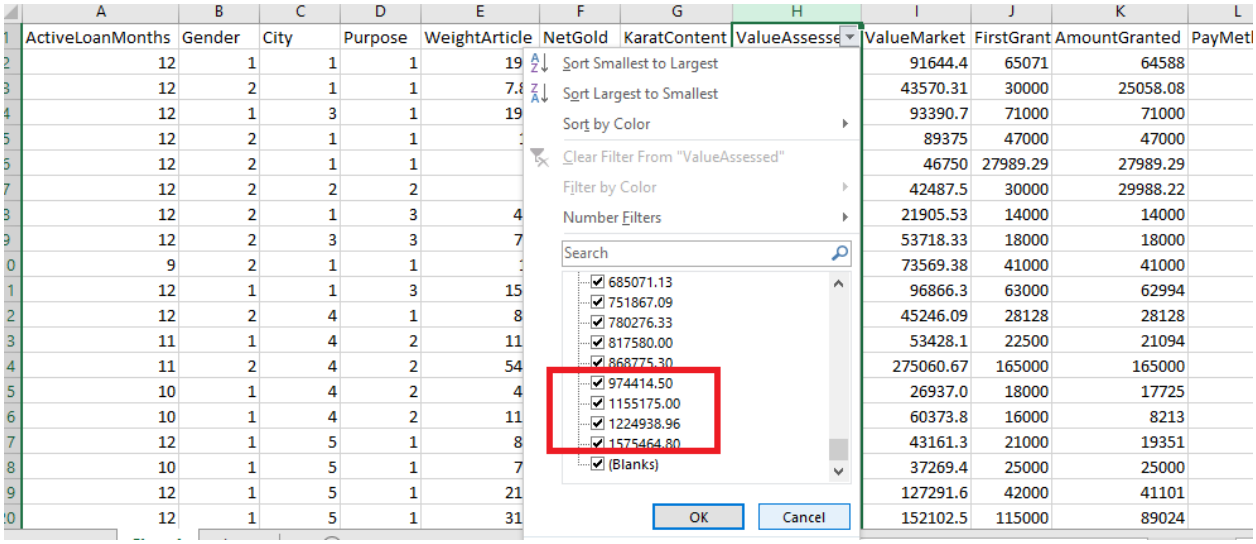


Figure 6.1-Selecting and removing outlier values

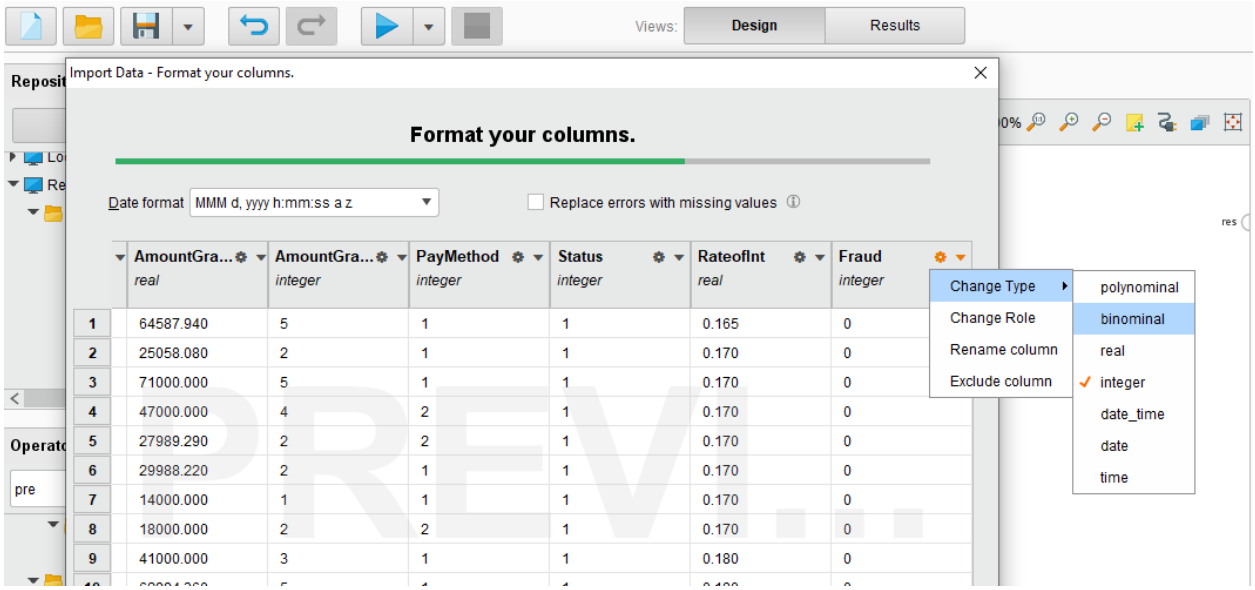


Figure 6.2-Converting class variable type to nominal

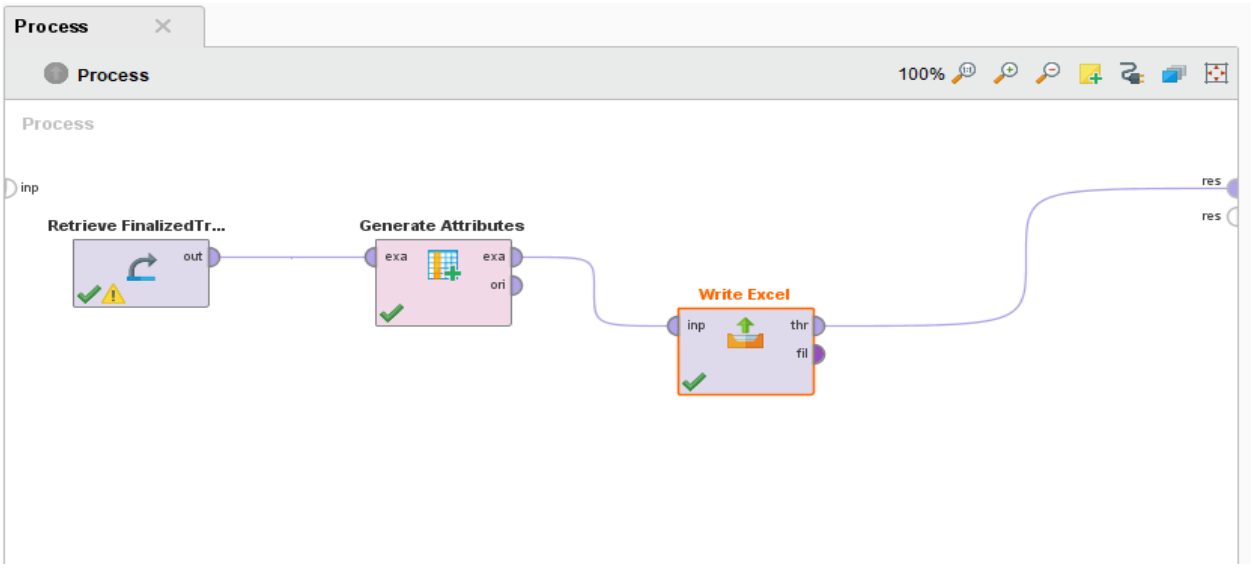


Figure 6.3-Generate new categorized attributes from the existing attributes 1

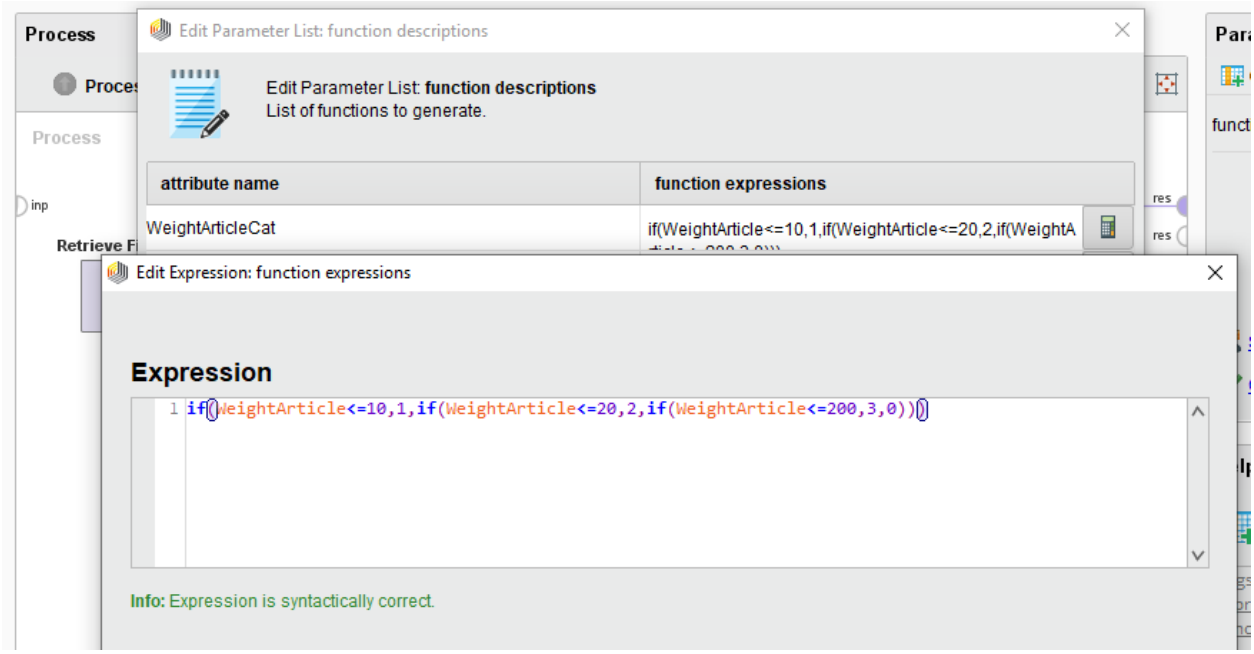


Figure 6.4- Generate new categorized attributes from the existing attributes 2

Appendix B

Attribute Selection

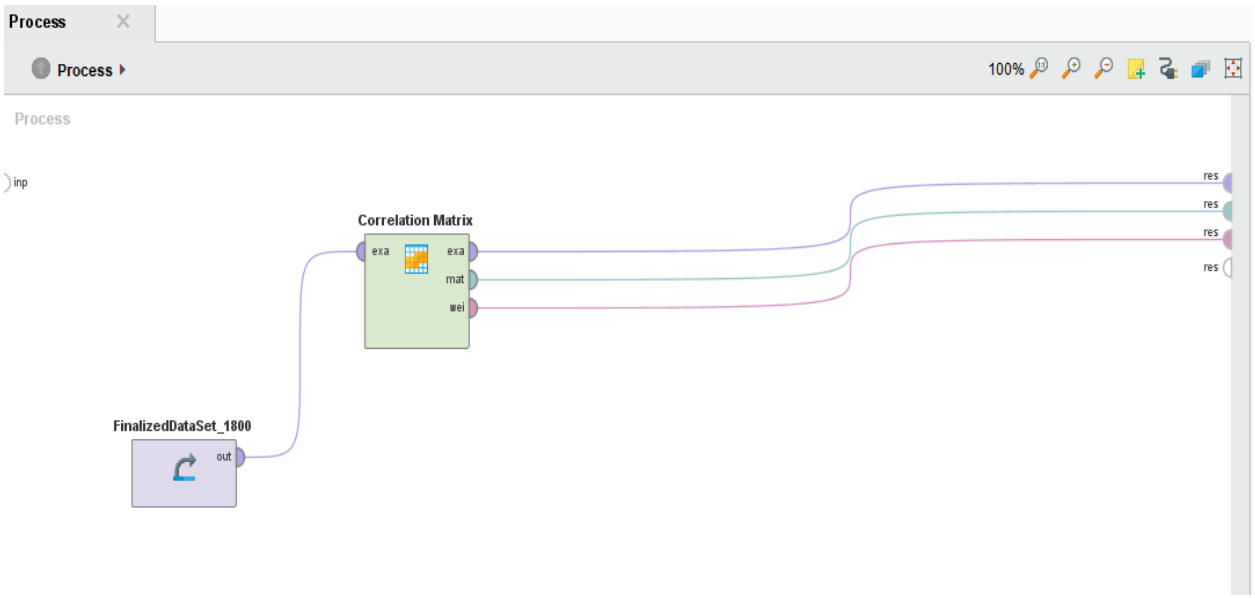


Figure 6.5- Attribute selection using correlation matrix

Data
Pairwise Table
Charts
Annotations

Attribut...	Activ...	Gender	District	Purpose	Weight...	Netgol...	KaratCo...	valueAs...	valuem...	Firstgra...	Amount...	PayMet...	Status	RateofInt	Fraud
ActiveLo...	1	0.001	0.026	0.007	0.020	0.020	0.070	0.020	0.019	0.036	0.046	0.010	0.054	0.002	0.045
Gender	0.001	1	0.074	0.055	0.001	0.001	0.053	0.000	0.001	0.003	0.004	0.017	0.064	0.053	0.186
District	0.026	0.074	1	0.064	0.010	0.009	0.128	0.001	0.000	0.013	0.017	0.103	0.037	0.031	0.348
Purpose	0.007	0.055	0.064	1	0.001	0.001	0.029	0.007	0.008	0.000	0.000	0.061	0.094	0.104	0.189
WeightC...	0.020	0.001	0.010	0.001	1	0.993	0.096	0.696	0.628	0.747	0.664	0.001	0.054	0.015	0.001
Netgoldc...	0.020	0.001	0.009	0.001	0.993	1	0.099	0.696	0.627	0.749	0.667	0.002	0.055	0.016	0.001
KaratCo...	0.070	0.053	0.128	0.029	0.096	0.099	1	0.112	0.102	0.183	0.180	0.013	0.016	0.000	0.192
valueAss...	0.020	0.000	0.001	0.007	0.696	0.696	0.112	1	0.933	0.860	0.759	0.005	0.094	0.029	0.002
valuema...	0.019	0.001	0.000	0.008	0.628	0.627	0.102	0.933	1	0.811	0.713	0.006	0.096	0.033	0.005
Firstgran...	0.036	0.003	0.013	0.000	0.747	0.749	0.183	0.860	0.811	1	0.891	0.000	0.074	0.017	0.008
Amount...	0.046	0.004	0.017	0.000	0.664	0.667	0.180	0.759	0.713	0.891	1	0.000	0.068	0.013	0.015
PayMeth...	0.010	0.017	0.103	0.061	0.001	0.002	0.013	0.005	0.006	0.000	0.000	1	0.031	0.052	0.116
Status	0.054	0.064	0.037	0.094	0.054	0.055	0.016	0.094	0.096	0.074	0.068	0.031	1	0.056	0.187
RateofInt	0.002	0.053	0.031	0.104	0.015	0.016	0.000	0.029	0.033	0.017	0.013	0.052	0.056	1	0.118
Fraud	0.045	0.186	0.348	0.189	0.001	0.001	0.192	0.002	0.005	0.008	0.015	0.116	0.187	0.118	1

Figure 6.6- Attribute selection using correlation matrix results

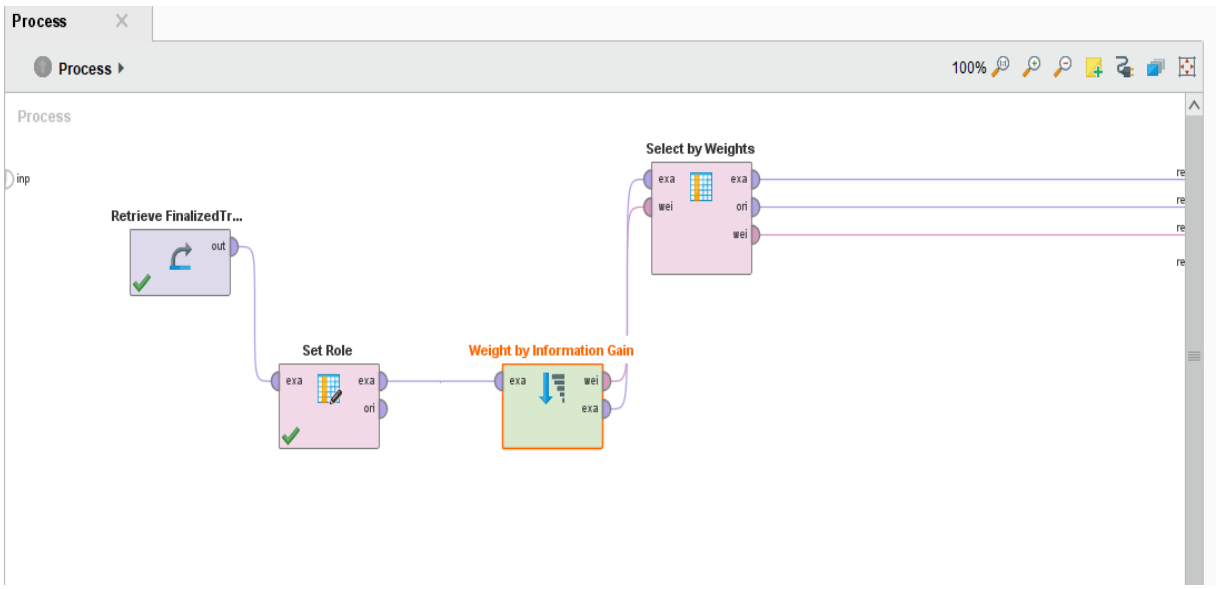
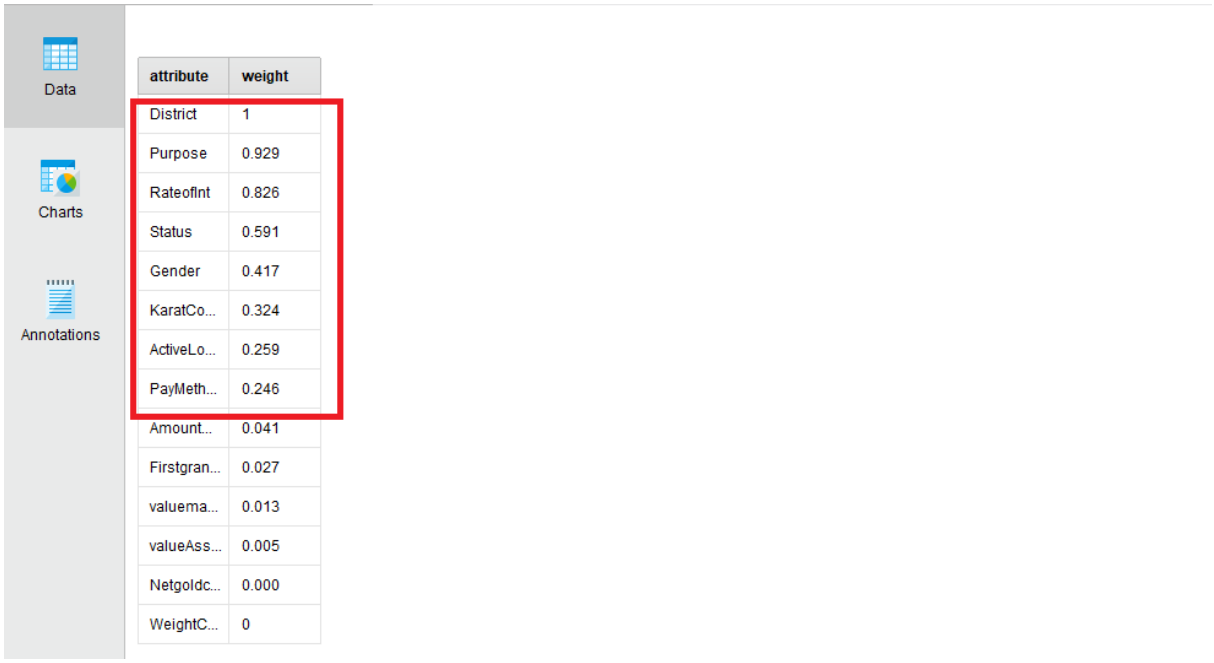


Figure 6.7- Attribute selection using weight by information gain



attribute	weight
District	1
Purpose	0.929
RateofInt	0.826
Status	0.591
Gender	0.417
KaratCo...	0.324
ActiveLo...	0.259
PayMeth...	0.246
Amount...	0.041
Firstgran...	0.027
valuema...	0.013
valueAss...	0.005
Netgoldc...	0.000
WeightC...	0

Figure 6.8- Attribute selection using weight by information gain results

Appendix C

Model Creation

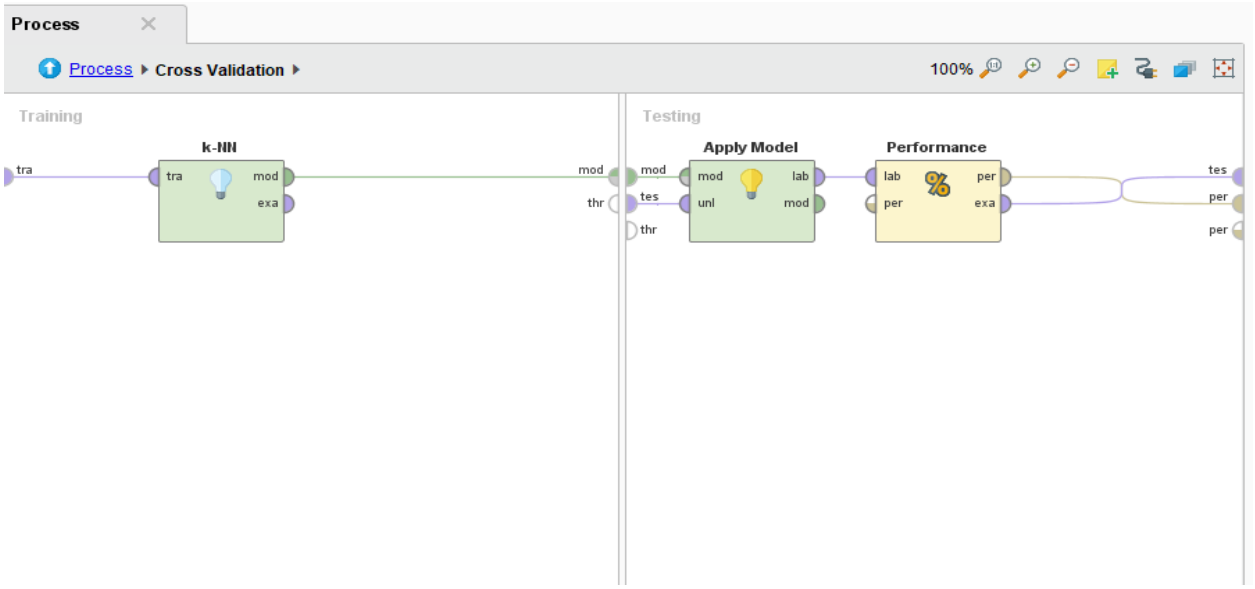


Figure 6.9-Model creation cross validation

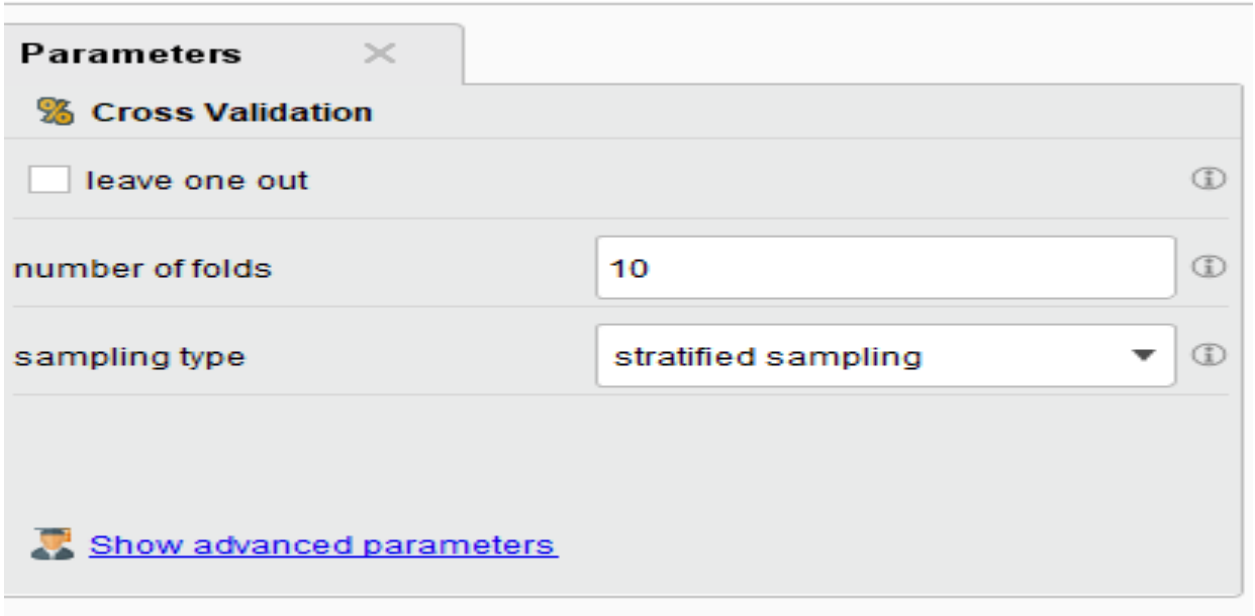


Figure 6.10-Model creation validation cross validation parameters

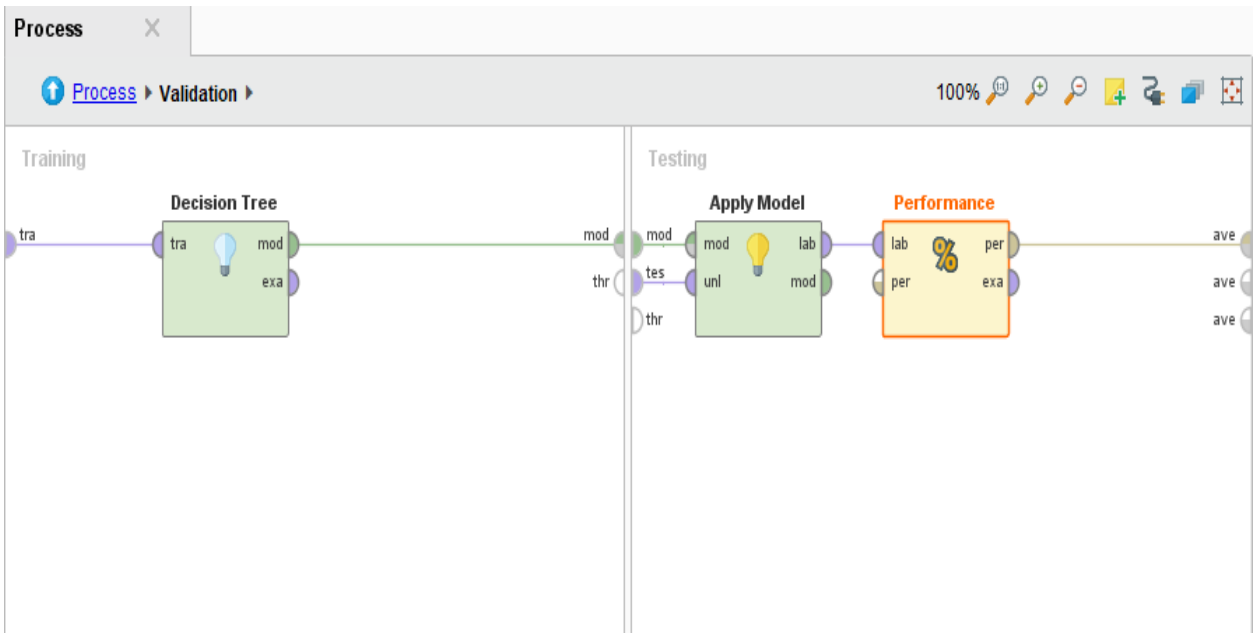


Figure 6.11- Model creation split validation

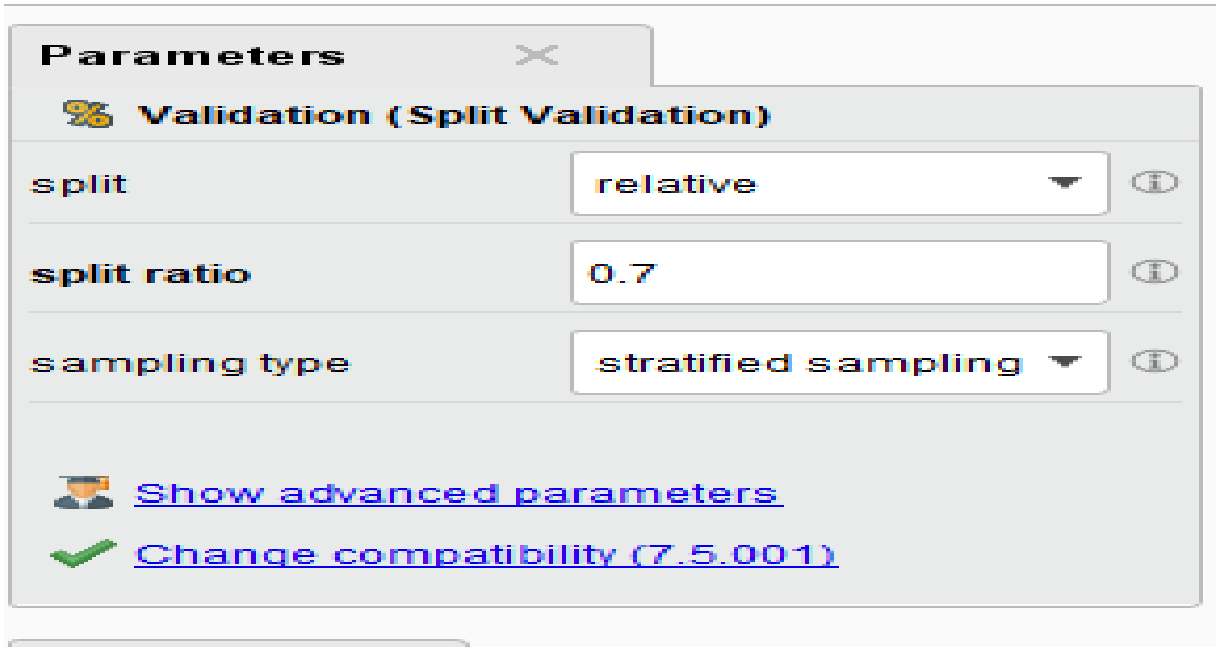


Figure 6.12- Model creation split validation parameters

Appendix D

Evaluation Results

Results for Correlation Matrix Attributes

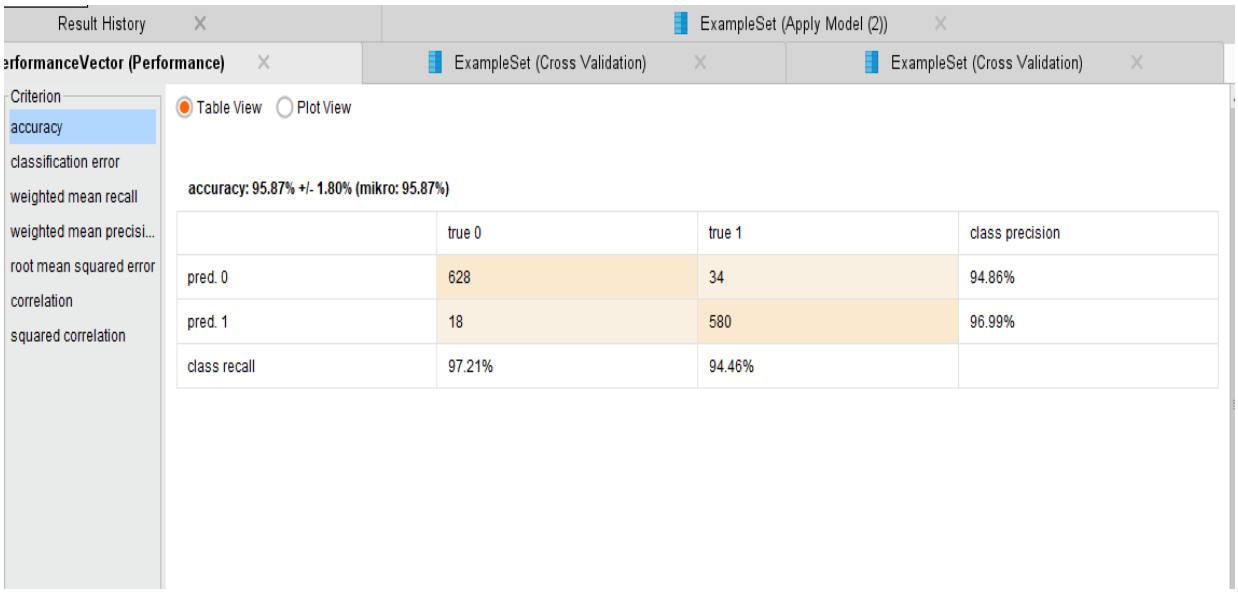


Figure 7.2-Cross validation Bagging

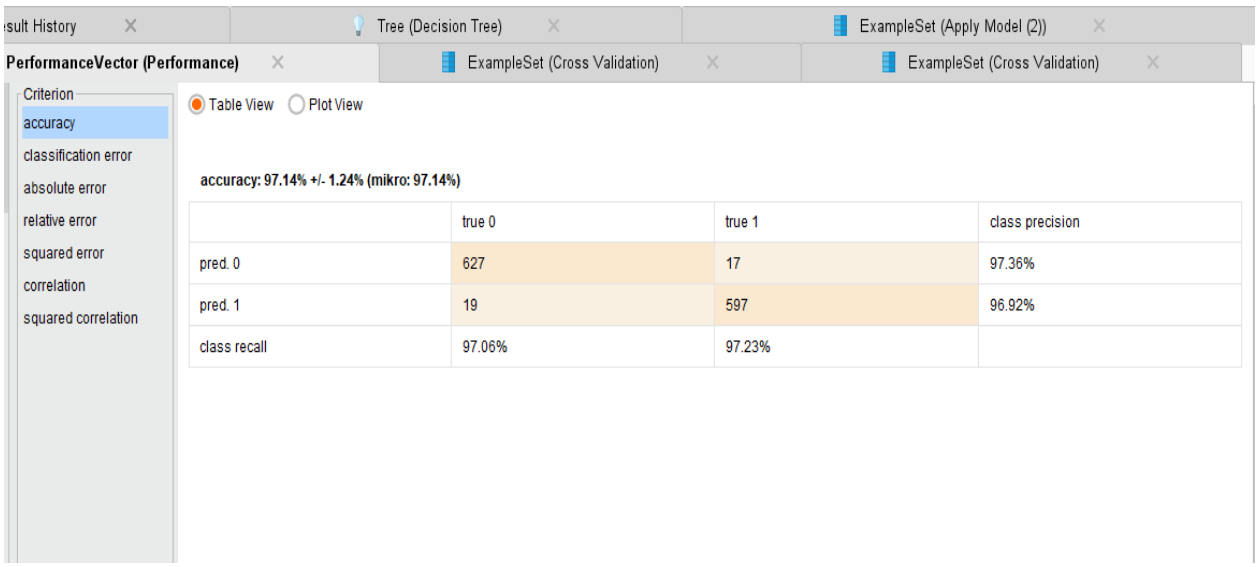


Figure 7.3-Cross validation Decision tree

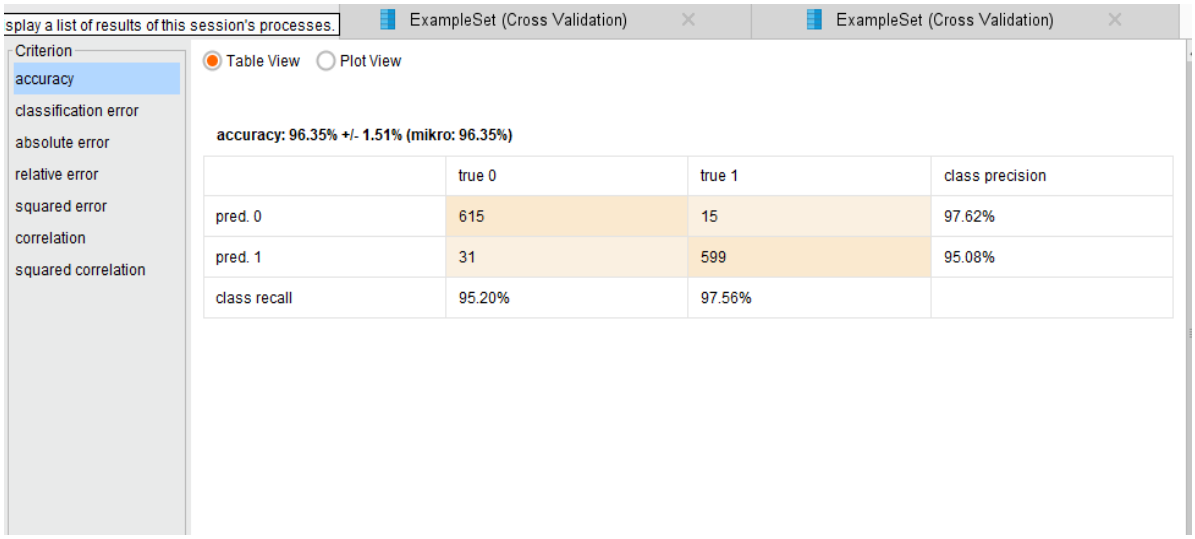


Figure 7.4-Cross validation KNN

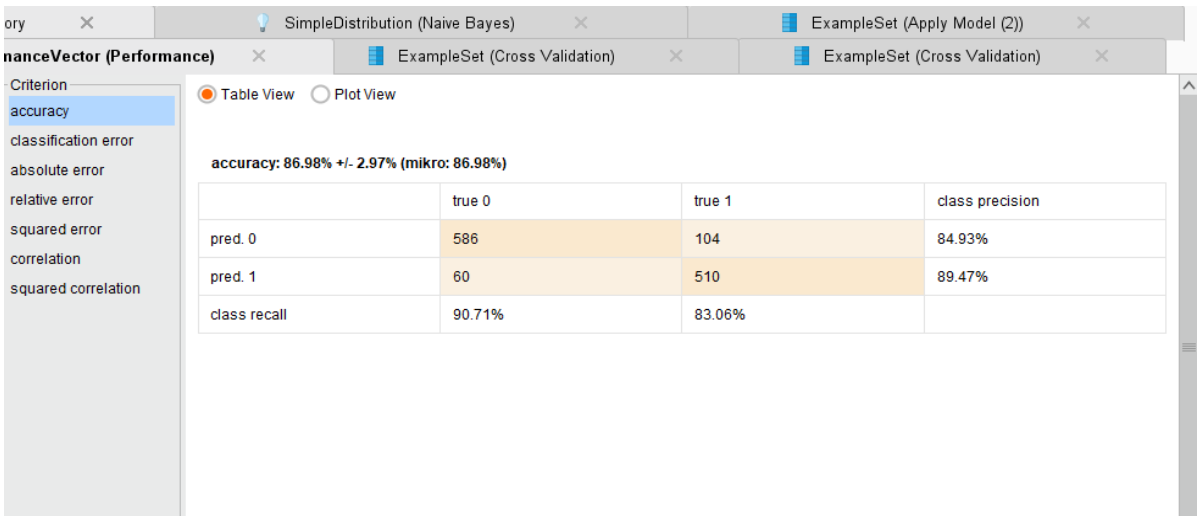


Figure 7.5- Cross validation Naive Bayes

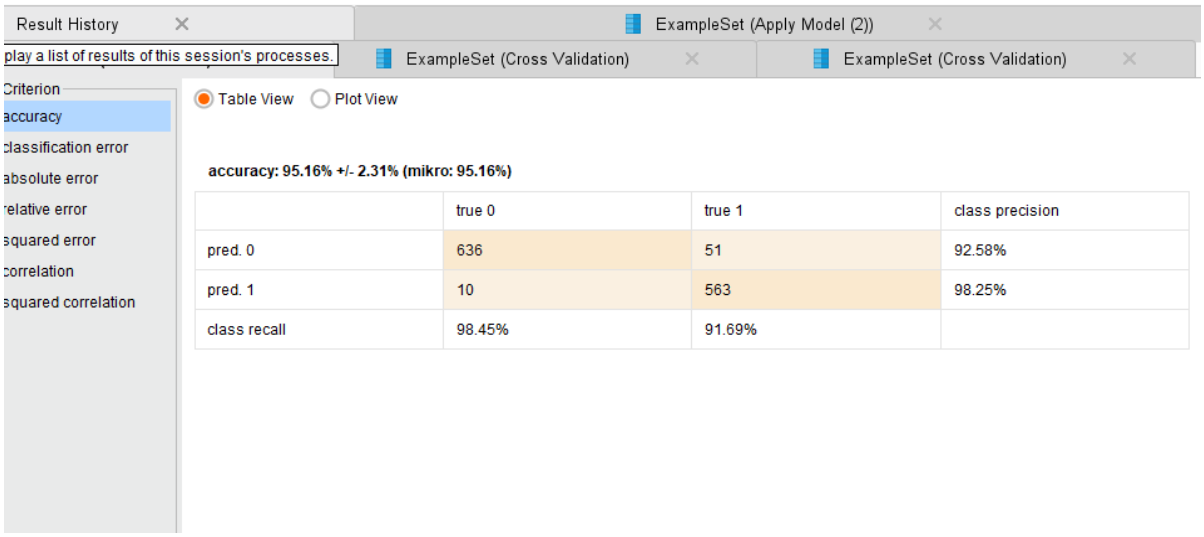


Figure 7.6-Cross validation Random Forest

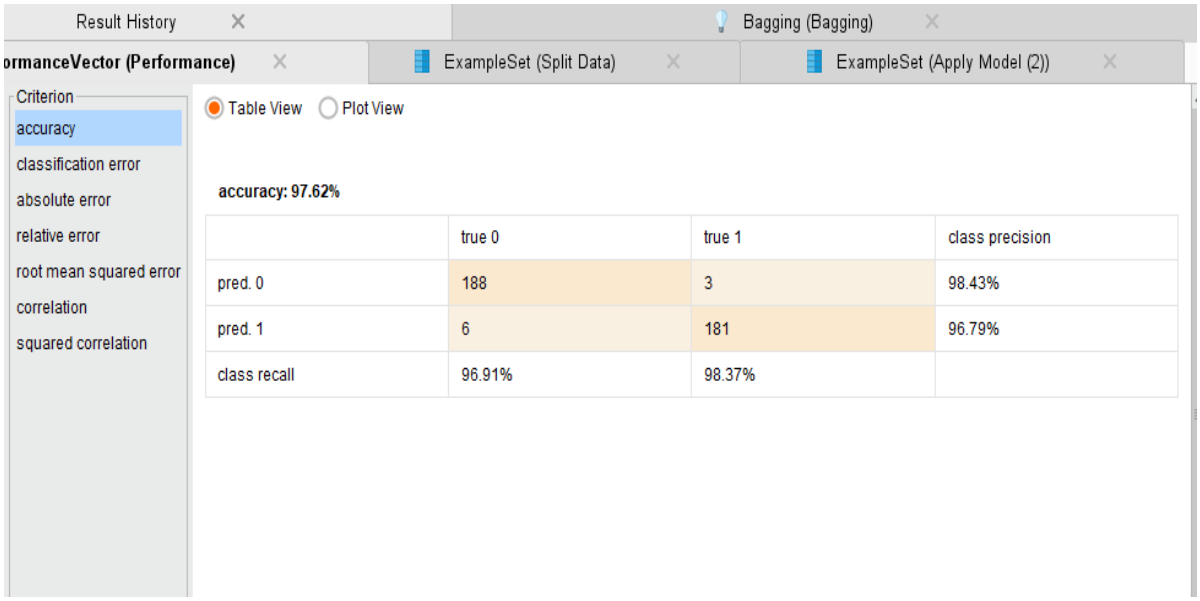


Figure 7.7-Split validation Bagging

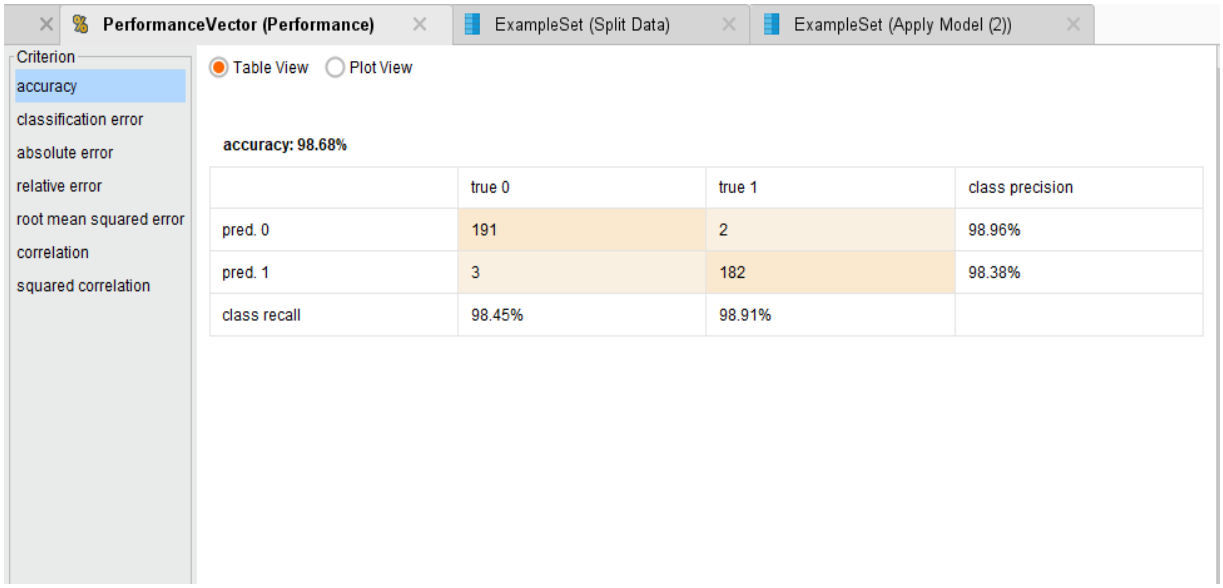


Figure 7.8-Split validation Decision tree

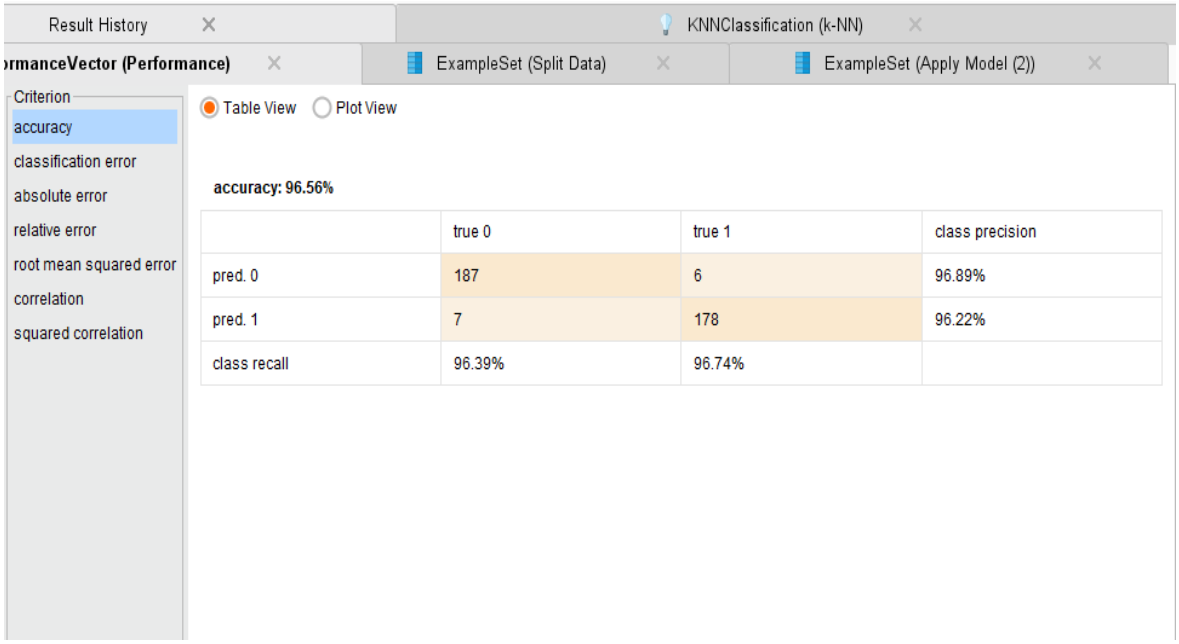


Figure 7.9- Split validation KNN

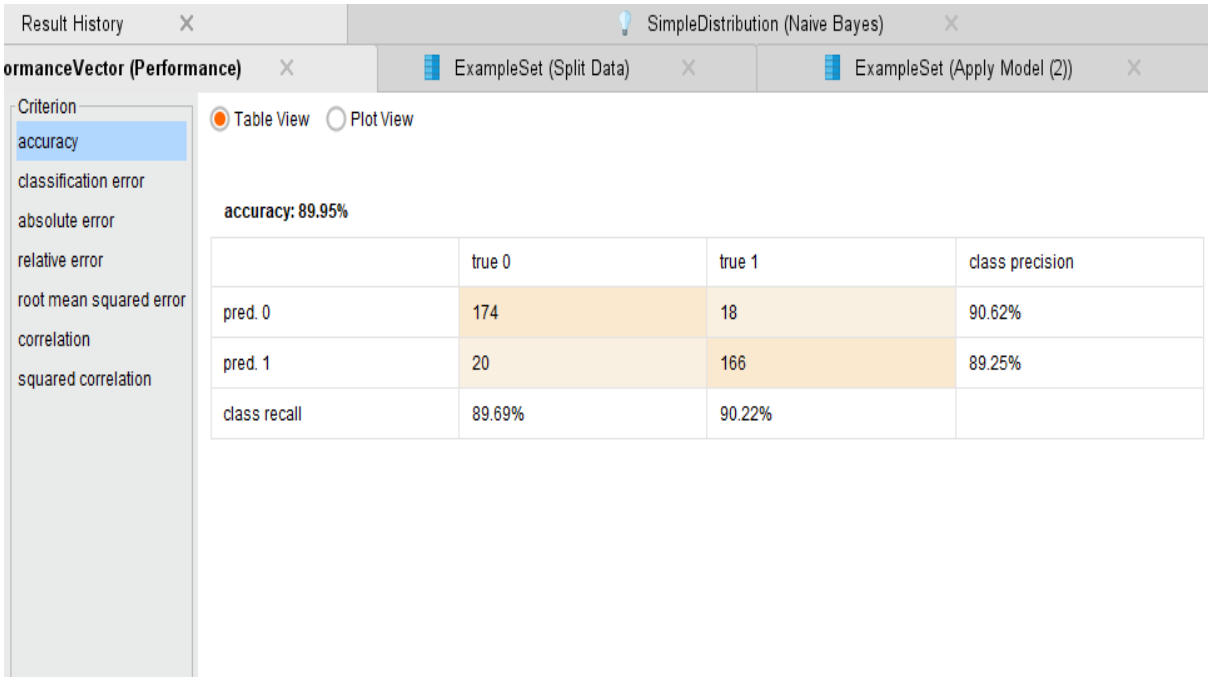


Figure 7.10- Split validation NaiveBayes

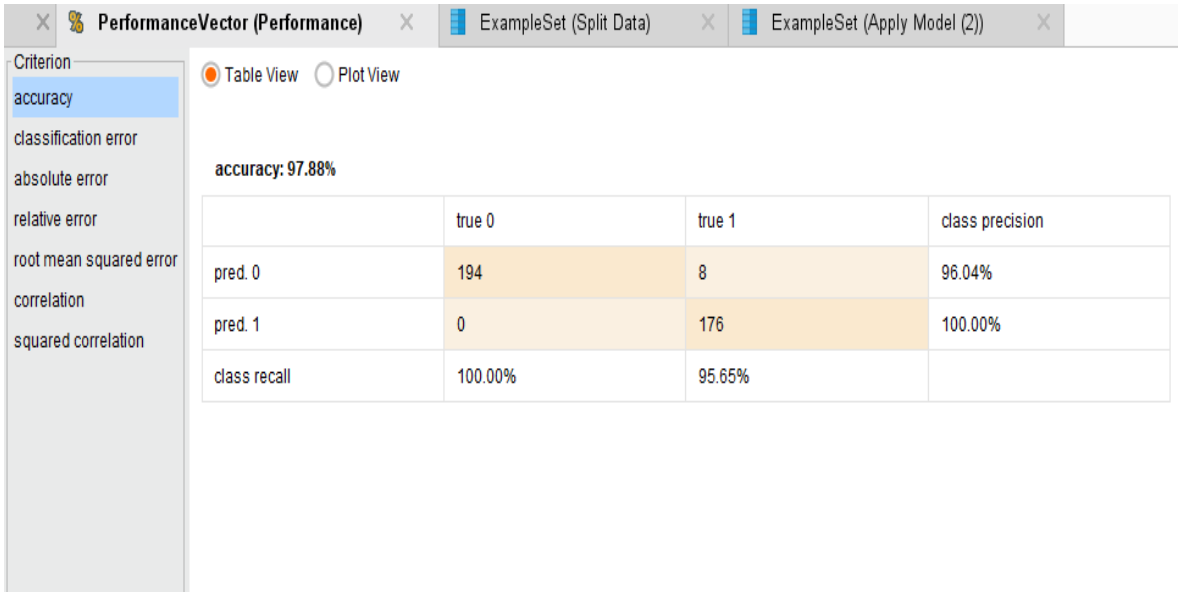


Figure 7.11-Split validation Random Forest

Results for Information gain attributes

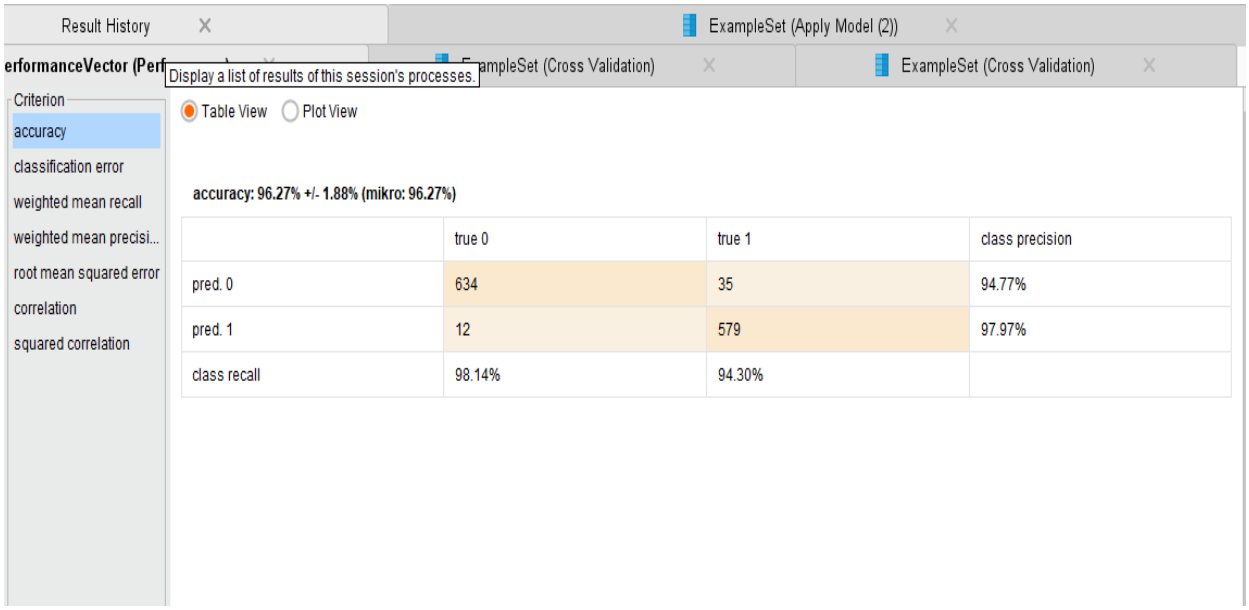


Figure 7.12- Cross validation Bagging

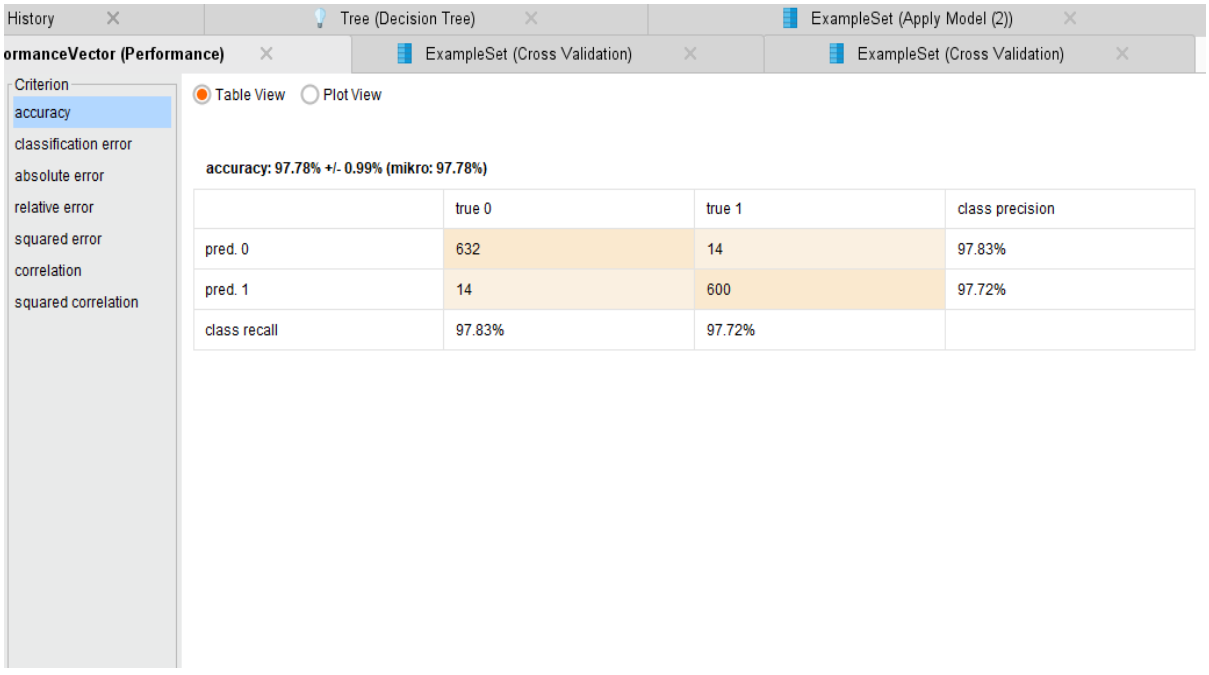


Figure 7.13-Cross validation Decision tree

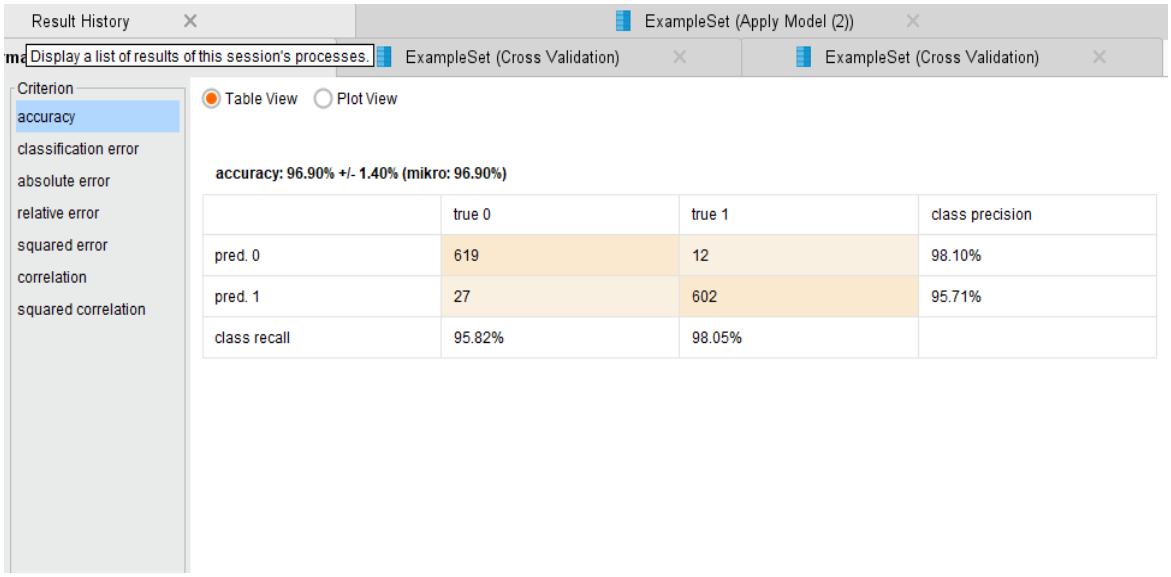


Figure 7.14-Cross validation KNN

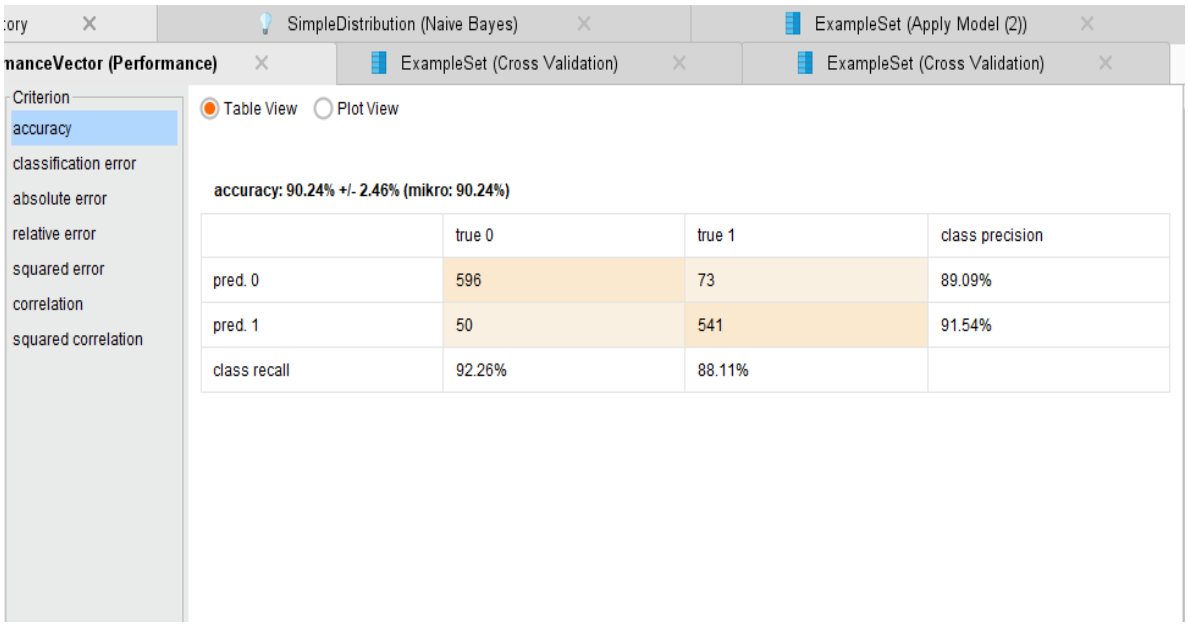


Figure 7.15- Cross validation NaiveBayes

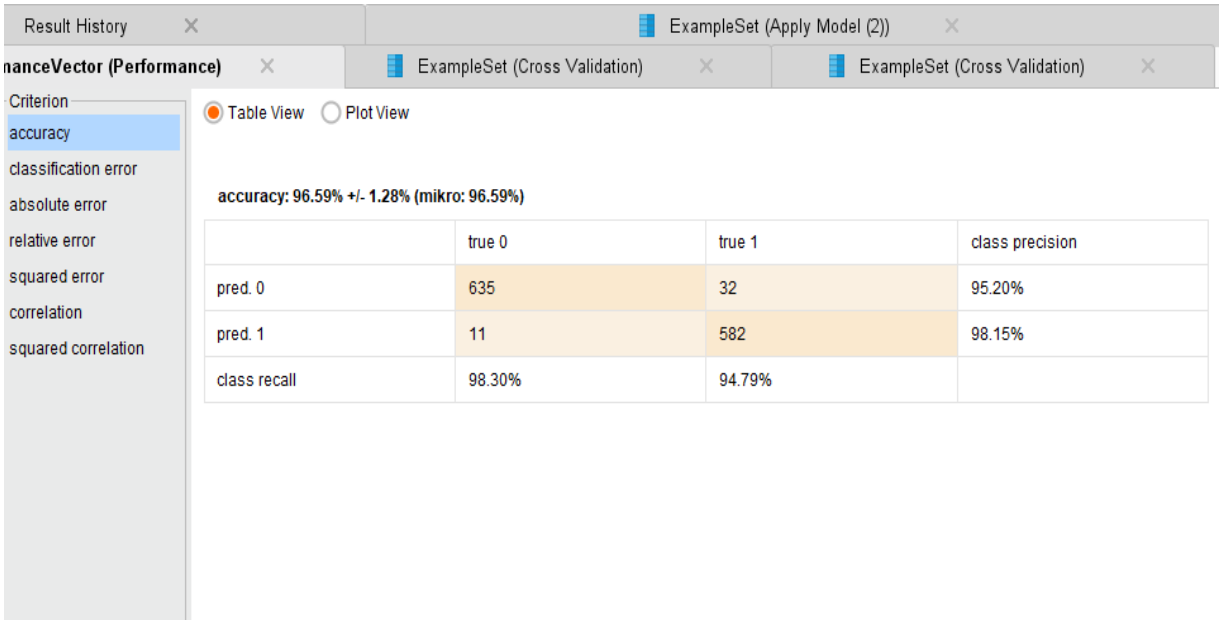


Figure 7.16- Cross validation RandomForest

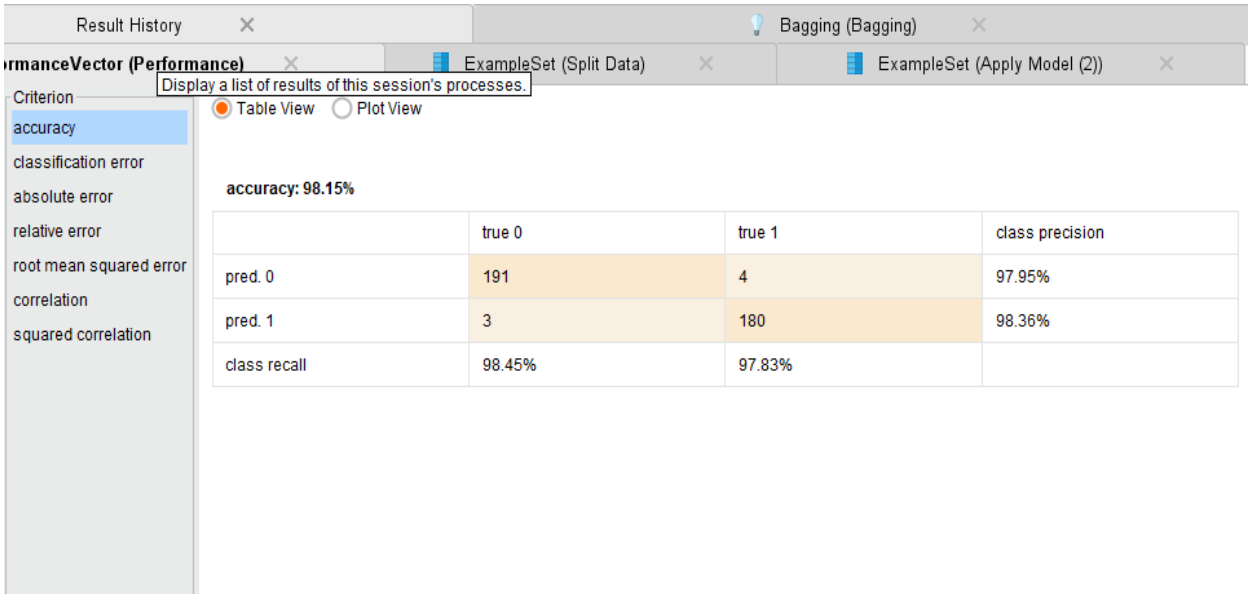


Figure 7.17- Split validation Bagging

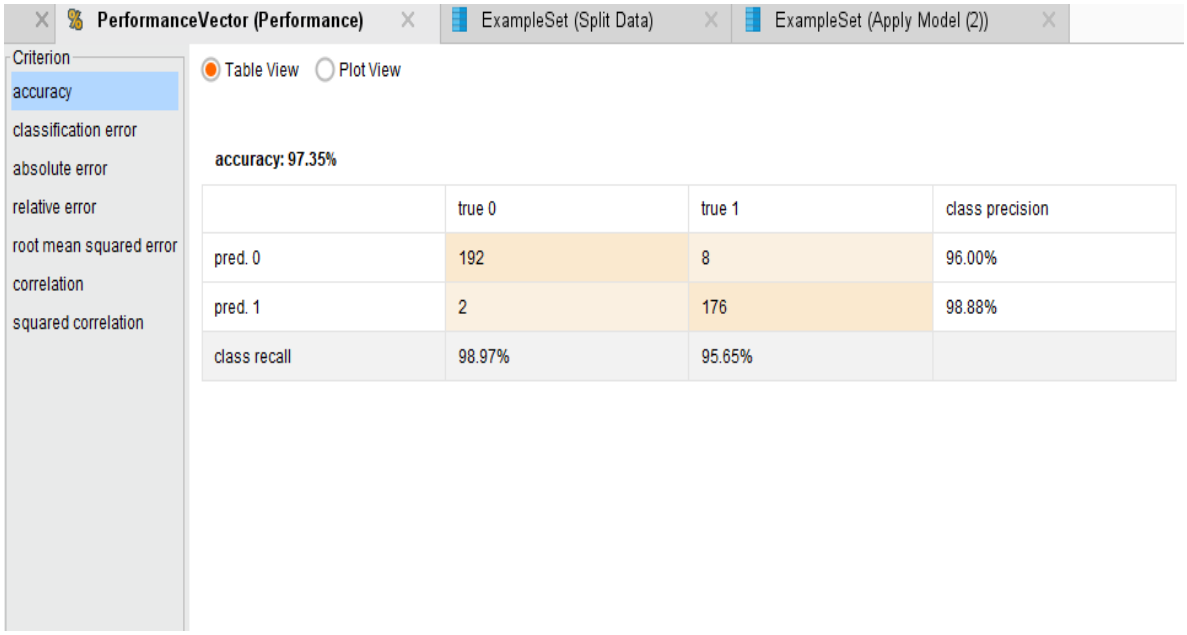


Figure 7.18-Split validation Decision Tree

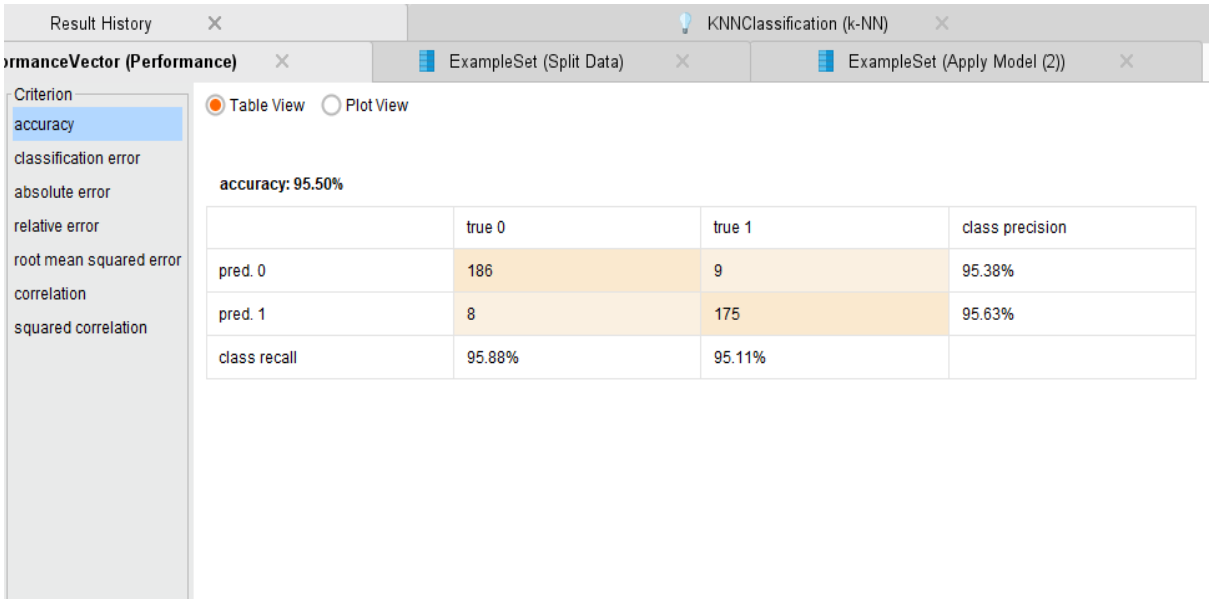


Figure 7.19- Split validation KNN

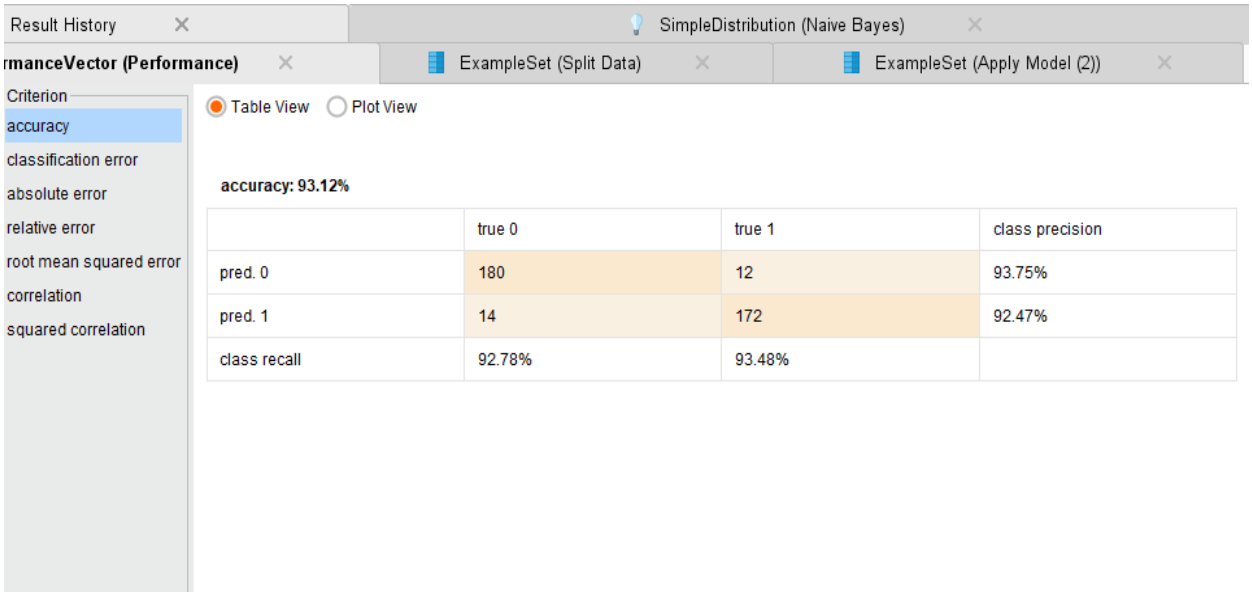


Figure 7.20-Split validation Naive Bayes

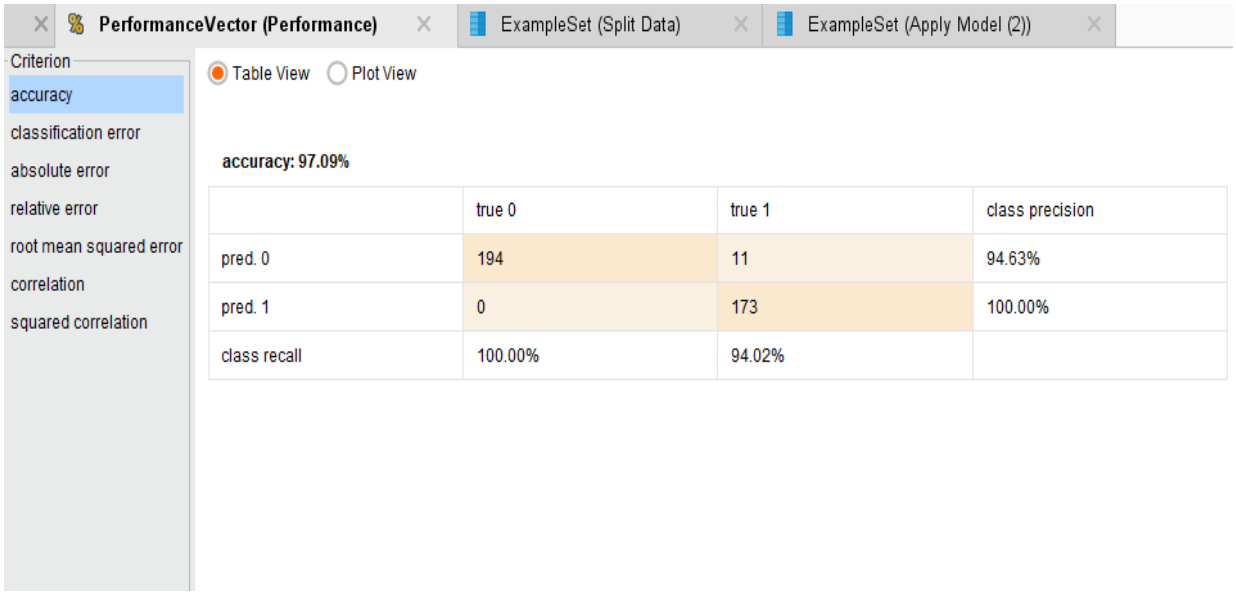


Figure 7.21-Split validation Random Forest