Sudan University of Science and Technology

College of Graduate Studies

Microfinance Data Analysis in Banking Sector Using Data Mining Techniques

(Case Study : Agricultural Bank of Sudan)

تحليل بيانات التمويل الأصغر في قطاع البنوك باستخدام تقنية التنقيب عن البيانات

(دراسة حالة : البنك الزراعي السوداني)

A dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology

Submitted by:
Mehad Ibrahim Alamin Abakar

Supervised by:
Dr. Mubarak Mohammed Ahmed

2020

الآيـــة

قال تعالى:

﴿ذَلِكَ مَبْلَغُهُمْ مِنَ الْعِلْمِ إِنَّ رَبَّكَ هُوَ أَعْلَمُ بِمَنْ ضَلَّ عَنْ سَبِيلِهِ وَهُوَ أَعْلَمُ بِمَنِ اهْتَدَى﴾

صدق الله العظيم

النجم (الآيـــة 30)

DEDICATION

This dissertation is dedicated to:

My lovely mother

My Dear father

My brother's

My sisters

All my family, my friends and colleagues

Acknowledgments

In the name of Allah who granted me health and power to accomplish this thesis. It gives me great pleasure to express my deep thanks, sincere gratitude and appreciation to my supervisor Dr. Mubarak for his continued efforts during all phases of the research. I would like to express my sincere thanks to Dr. Hisham Abdullah Mansur for their valuable help, encouragement, follow up and assistance. I would like to thanks El-Dean University for giving me this master program opportunity. Deep thanks to the Agricultural Bank of Sudan for its support in the most important part of the research stages (For their contributions to data collection) in particular, Mr. Mohammed Yahiya, ENG. Abdu El-Hafiz Abdullah Nassir and Ms. Fradus Hassan Taha. Also, I would like to thank my family, my parents and my brothers and sister for support that give me. Finally, my deepest thanks are extended to my colleagues of the batch (9) in Master degree, and all the other individuals who helped me.

ABSTRACT

Banks face lots of challenges associated with the bank loan, Nowadays there are many risks related to microfinance in bank sector. Every year, we face number of cases where people do not repay most of the microfinance amount to the banks which they cause huge losses. The risk associated with making decision on microfinance request approval is massive. In this study a classification model was built based on the microfinance data obtained from an agricultural bank of Sudan to predict the status of microfinance. The dataset has been preprocessed, reduced and made ready to provide efficient predictions. Random forest, NaiveBayes and KNN classification algorithms have been used to build the proposed model. By using Orange application the model has been implemented and tested. The accuracy for the above three techniques is Random forest 94.6%, NaiveBayes 87.4% and KNN 92.3%. Random forest selected as best algorithm based on accuracy. The final model is used for prediction with the test dataset and the experimental results proved the efficiency of the built model.

المستخلص

تواجه البنوك الكثير من التحديات المرتبطة بالقروض المصرفية ، في الوقت الحاضر هناك العديد من المخاطر المتعلقة بالتمويل الأصغر في قطاع البنوك. كل عام ، نواجه عددًا من الحالات التي لا يسدد فيها الأشخاص معظم مبالغ التمويل الأصغر للبنوك مما يتسبب في خسائر فادحة. المخاطر المرتبطة باتخاذ قرار بشأن الموافقة على طلب التمويل الأصغر كبيرة. في هذه الدراسة ، تم بناء نموذج تصنيف بناءً على بيانات التمويل الأصغر التي تم الحصول عليها من البنك الزراعي بالسودان للتنبؤ بحالة التمويل الأصغر. تمت معالجة مجموعة البيانات مسبقًا وجعلها جاهزة لتوفير تنبؤات فعالة. تم استخدام خوارزميات التصنيف Random forestو NaiveBayes و KNN لبناء النموذج المقترح. باستخدام تطبيق Orange ، تم تنفيذ النموذج واختباره. دقة التقنيات الثلاثة المذكورة أعلاه هي94.6 Random forest ٪ و87.4 NaiveBayes ٪ و92.3 KNN ٪ تم اختيار Random forest كأفضل خوارزمية بناءً على الدقة. تم استخدام النموذج النهائي للتنبؤ مع مجموعة بيانات الاختبار ، واثبتت النتائج التجريبية كفاءة النموذج.

Table of Contents

List of table

List of figure

List of Abbreviations

PM	Performance Management
DM	Data Mining
KDD	Knowledge Discovery from Data
KNN	(K -Nearest Neighbor
UCI	International Cycling Union
NLTK	Natural Language Toolkit
KNIME	Konstanz Information Miner
GUI	Graphical User Interface

CHAPTER I: INTRODUCTION

1.1 Research background

Performance management (PM) has become one of the most important initiatives in the microfinance industry today. One dire issue facing the Microfinance institutions is how to link the organization' performance to growth and profitability so that resources can be optimally allocated and fully utilized to meet competition and support increasing demand of quality products/services from customers. Mining customer data to measure productivity and enhance performance management is not only feasible in this information era but also in line with the transformation of a microfinance institution into a "customer driven organization". In this research, we look at application of data mining techniques to performance management in the microfinance industry.

Data Mining or knowledge discovery in databases can be defined as an activity that extracts some new nontrivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios.

With the abundance of existing data stored in databases, and with the proliferation of large storage repositories, it became necessary to find techniques, methods and means to extract information and knowledge from these stored data and to exploit them in problem solving and decision making using modern computer applications.

1.2 Problem statement

Most of the banks use their own credit scoring and risk assessment techniques in order to analyze the microfinance requests and to make decisions on Approval. In spite of this, there are many cases happening every year, where people do not repay microfinance amounts or they default.

Difficulty to estimate the success of projects for the institution provided for funding. And finding out what projects is successful for the applicant to avoiding default.

1.3 Research objectives

This study formulates the following objectives:

- To extract patterns from a common microfinance dataset and build a model based on these extracted patterns, in order to predict the likely microfinance defaulters by using classification data mining algorithms.
- To determining successful finance details, successful projects and avoiding default.
- To easy decision making by determining the success or failure project from the beginning.

1.4 Research importance

This study provides a theoretical reference to the management to analyze their data using data mining techniques to identify the strengths and weaknesses of microfinance.

The study contributes towards understanding the state of microfinance in and information derived from this study will help in making future decisions and policies on how these important microfinance lenders can be well positioned in the country to grow and reach the millions of potential clients who do not currently have access to mainstream financial services.

1.5 Research Methodology

The dataset collected from the Agricultural Bank of Sudan (Investment management) .In this research Random forest, KNN and NaiveBayes classification data mining algorithms are used to build model. The Research realized by using the Orange toolkit.

1.6 Research scope

This research focuses on Appling data mining techniques for supporting bank microfinance by extracting knowledge from banking investment data which obtained from Agricultural Bank of Sudan from investment management from (2002 to 2018).

1.7 Thesis organization

This research contains five chapters organized as follows: Chapter I contains Theoretical background about the domain of the research. Chapter II discusses the literature review and related work. Chapter III describes the research methodology and the implementation of the techniques used. Chapter IV presents the experimental results and their discussion. Lastly Chapter V concluded and presents the Future work.

CHAPTER II: LITERATURE REVIEW AND RELATED WORK

2.1 Introduction

This chapter presents the literature review and the work done by other researchers related to this work.

2.2 Overview of Microfinance

Microfinance, also called microcredit, is a type of banking service that is provided to unemployed or low-income individuals or groups who otherwise would have no other access to financial services. While institutions participating in the area of microfinance most often provide lending (microloans can range from as small as 20,000 to as large as 25,000 SDG), many banks offer additional services, such as checking and savings accounts, and micro-insurance products; and some even provide financial and business education. Ultimately, the goal of microfinance is to give impoverished people an opportunity to become self-sufficient (Vento*, 2007).

2.2.1 Understanding Microfinance

Microfinance services are provided to unemployed or low-income individuals because most of those trapped in poverty, or who have limited financial resources, do not have enough income to do business with traditional financial institutions.

Microfinance allows people to take on reasonable small business loans safely, and in a manner that is consistent with ethical lending practices. Although they exist all around the world, the majority of micro financing operations occur in developing nations, such as Uganda, Sudan, Indonesia, Serbia, and Honduras. Many microfinance institutions focus on helping women in particular (Kokonya, 2014) .

2.2.2 Microfinance Works

Micro financing organizations support a large number of activities that range from providing the basics—like bank checking and savings accounts—to startup capital for small business entrepreneurs, and educational programs that teach the principles of investing. These programs can focus on such skills as bookkeeping, cash-flow management, and technical or professional skills, like accounting. Unlike typical financing situations, in which the lender is primarily concerned with the borrower having enough collateral to cover the loan, many microfinance organizations focus on helping entrepreneurs to succeed .

In many instances, people seeking help from microfinance organizations are first required to take a basic money-management class. Lessons cover understanding interest rates, the concept of cash flow, how financing agreements and savings accounts work, how to budget, and how to manage debt.

Once educated, customers then may apply for loans. Just as one would find at a traditional bank, a loan officer helps borrowers with applications, oversees the lending process, and approves loans .offering education, job training, and working toward a better environment (Hermes et al., 2011).

2.3 Overview of Data mining

The development of information technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction Data mining as a step in the process of knowledge discovery (KDD) (Bharati and Ramageri, 2010) .

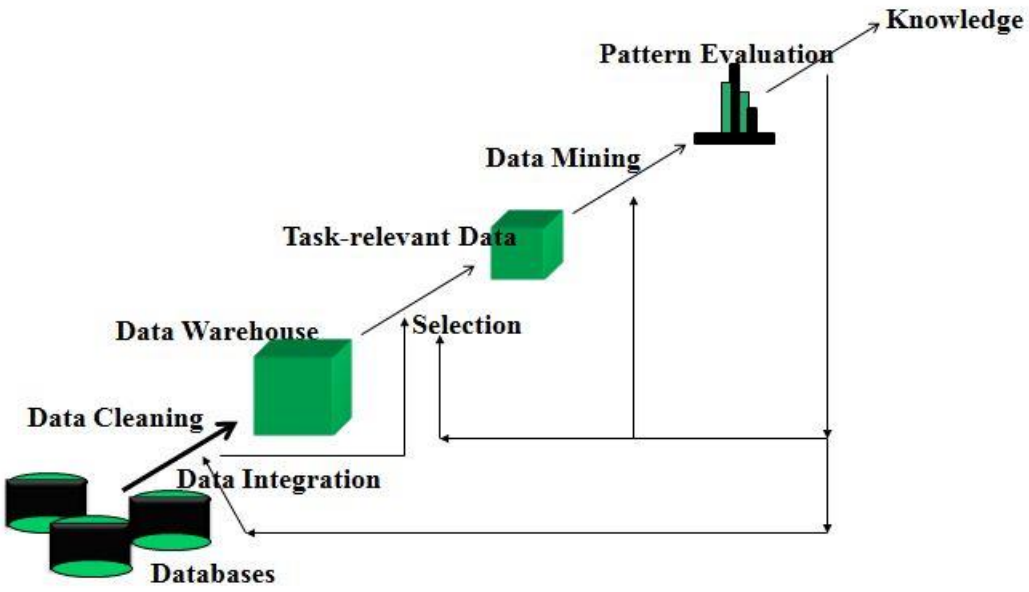


Figure2. 1 data mining as a step in the process of knowledge discovery

Data cleaning (to remove noise and inconsistent data)

Data integration (where multiple data sources may be combined)

Data selection (where data relevant to the analysis task are retrieved from the database)

Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user) (Amala Jayanthi.M*, 2016)

2.4 Data mining functionalities

The KDD process is ultimately data mining methods to extract patterns from data. Each method has different aim, which decides the outcome of the KDD process entirely. The outcome of the KDD process can be any of the following tasks based on the customer desire. These tasks are categorized as predictive and descriptive mining.

2.4.1 Predictive Mining

Supervised learning task where the unknown value of a class or future values of interest is predicted from the existing data. It can also validate a newly invented hypothesis.

2.4.1.1 Classification

Classification results in classification model termed as classifier that classifies the data as classes and concepts .The resultant model is used to predict the class label of the instances for which the class label is unknown. Decision tree induction, bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are the some of the kinds of classification methods used to decide the classifier of the sample.

2.4.1.2 Prediction

Prediction models a predictor that predicts the unknown data and future data from the available data

2.4.2 Descriptive mining

It is a task of summarizing the data and its features as patterns using data mining and data aggregation methods.

2.4.21Clustering

Clustering is a task of grouping data of similar characteristics into a cluster while the different data may group into different respective clusters. Search for the cluster is an unsupervised learning i.e. Class label is unknown. Thus the data are organized into an effective representation that categories the sample data .K-means, k-medoids logic are the some of the kinds of clustering methods.

2.4.2.2 Association rule mining

Association rule mining unwraps the patterns that occur frequently among the data set. It focus in extracting associations, correlations, frequent sequence, frequent item set and frequent patterns with interestingness among the data set in the data repositories.

Apriori is the some of the kinds of method Association rule mining.

2.4.3 Summarization

Summarization is the process of reducing the huge volume of data in a meaningful and intelligent fashion with important and relevant features. Summarization techniques like tabulation of the mean and the standard deviations are often implied to analyze and visualize the data, and to generate the report automatic (Padhy et al., 2012)

2.5 Data mining tools

There are many useful tools available for data mining such as

2.5.1 WEKA

The original version of WEKA was non-JAVA and was developed to analyze data from the agricultural domain. The JAVA version of WEKA, l is very sophisticated and

used in various applications to visualize, analyze and predict. It's a open ware under the GNU General Public License, Users can customize the tool.

2.5.2 Rapid Miner

Rapid Miner is Java based tool that offers advanced analytics through template-based frameworks. This Tool has been offered as a service, rather than local software. Rapid Miner also provides functionalities like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.

2.5.3 R – Programming

R – Programming is developed from C and FORTRAN. It's a freeware that provide software programming language and software environment for statistical computing and graphics. Data miners to develop statistical software and data analysis with the help of R-Programming it is very easy to use. It also provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, and classification and clustering apart from data mining.

2.5.4 Orange

Orange a Python-based, powerful and open ware it has components for machine learning, bioinformatics and text mining. It's wrapped with characteristics for data analytics.

2.5.5 KNIME (Konstanz Information Miner)

KNIME is a Java based. KNIME does all the three process of extraction, transformation and loading of data. It provides a GUI that allows assembling the nodes for data processing it is an open source that is able to do data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept. It is also able to perform business intelligence and financial data analysis. KNIME is easy to extend and to add plug-in.

2.5.6 NLTK (Natural Language Toolkit)

NLTK is python based can be customized. NLTK provides a pool of language processing tools including data mining, machine learning, and data (Hussain, 2017)

2.6 Areas of data mining

There are many areas of data mining such as

2.6.1 Web Mining

As there is huge amount of data and information available in the World Wide Web, the data miners have a fertile area for web mining. Web mining is data mining techniques for extraction of information from web documents and services. The contents of the web are very dynamic. It is growing at a rapid pace, and the information is continuously updated. Web mining may be divided into the following subtasks

1. Resource finding: finding documents intended for the Web.

2. Information selection and preprocessing: Selection and preprocessing of the information retrieved from the Web.

3. Generalization: To discover the general patterns from the individual as well as multiple sites.

4. Analysis: Discovered patterns are interpreted for meaningful knowledge. Web mining may be divided into Web Structure, Web Contents, and Web Access Patterns.

2.6.2 Text Mining

The term text mining or KDT (Knowledge Discovery in Text) was first proposed by Feldman and Dagan in 1996. The unstructured text may be mined using information retrieval, text categorization, or applying NLP techniques as a preprocessing step. Text Mining involves many applications such that text categorization, clustering, finding patterns and sequential patterns in texts, computational linguistics, and association discovery.

2.6.3 Multimedia data mining

Multimedia data mining explores the interesting patterns from databases related to multimedia that manages a large collection of multimedia objects. Multimedia objects include audio, video, image, sequence data and hypertext data containing text, text markups, and linkages.

Multimedia data research focuses on content-based retrieval, similarity search, association, classification and prediction analysis.

2.7 Application of data mining

Data Mining is used in many domains in constant basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many domains like health care, finance insurance, retail stores combines the data mining with as statistics, pattern recognition, and other important tools to perform data analytics. Data mining is used primarily for decision making (Hussain, 2017).

2.8 Challenges of Data mining

Though data mining is considered as a powerful information collection practice, it faces several different challenges for and during its implementation. Such challenges can be related to mining methods, data collection, performance etc. To enable different companies around the world in attaining perfectly calculated data for an even perfect and operational execution, these problems need to be addressed and solved. Some of the widely discussed challenges in the world of data mining are as follows.

2.8.1 Noisy and Incomplete Data

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors. Suppose a retail chain collects the email id of customers who spend more than $200 and the billing staff enters the details into their system. The person might make spelling mistakes while entering the email id which results in incorrect data. Even some customers might not be ready to

disclose their email id which results in incomplete data. The data even could get altered due to system or human errors. All these result in noisy and incomplete data which makes the data mining really challenging.

2.8.2 Distributed Data

Real world data is usually stored on different platforms in distributed computing environments. It could be in databases, individual systems, or even on the Internet. It is practically very difficult to bring all the data to a centralized data repository mainly due to organizational and technical reasons. For example, different regional offices might be having their own servers to store their data whereas it will not be feasible to store all the data (millions of terabytes) from all the offices in a central server. So, data mining demands the development of tools and algorithms that enable mining of distributed data.

2.8.3 Complex Data

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

2.8.4 Data Visualization

Data visualization is a very importance process in data mining because it is the main process that displays the output in a presentable manner to the user. The information extracted should convey the exact meaning of what it actually intends to convey. But many times, it is really difficult to represent the information in an accurate and easy-to-understand way to the end user. The input data and output information being really complex, very effective and successful data visualization techniques need to be applied to make it successful.

2.8.5 Data Privacy and Security

Data mining normally leads to serious issues in terms of data security, privacy and governance. For example, when a retailer analyzes the purchase details, it reveals information about buying habits and preferences of customers without their permission.

2.3 Related works

(Hala Hassan Mahmud Hassan, 2015) Compared and analyzed successful and unsuccessful microfinance projects using decision tree classification machine learning algorithm by weka application .The accuracy obtained 92%.The result had been discussed as, most successful product is the commercial sale with a profit rate of 19%, and a repayment period of 13 months, must this outcome should be implemented to avoid risky funding.The most failing types of products in the field of agriculture are therefore products that should be avoided because they represent a risk to funding. The highest risk is represented in other service projects, which is very large, and the transport project maintenance of carts to some extent.

(Pandit, 2016) proposed model predicts if the customer would be a defaulter or not by using classification data mining algorithms. Dataset used obtained from three different sources (UCI) are gathered together. Naïve Bayes – Decision Tree, Boosting classification algorithms was used. The model implemented using weka. Prediction accuracy was arranged between 70% to 74.22% using all algorithms. The prediction accuracy of defaulter instances is not that good using all the algorithms. The major reason for this could be the class imbalance high number of instances having class as ‘not defaulters, which results in biased output.

(Hamid and Ahmed, 2016) devolved a new model for classifying loan risk in banking sector by using data mining. The model has been built using data from banking sector to predict the status of loans. The number of instance in dataset is 1000 (The dataset divide into two groups training set which represent 80% from all data and testing set which represent 20% of the data set. Three algorithms have been used to build the proposed model: j48, BayesNet and NaiveBayes. By using Weka application, the model has been implemented and tested. The accuracy measure for the above three techniques are J48= 78.3784 % BayesNet =73.8739 % NaiveBayes= 73.8739 %. The results have been discussed and a full comparison between algorithms was conducted. J48 selected as best algorithm based on accuracy.

(Sivasree, 2015) the researchers introduce an effective prediction model for the bankers that help them predict the credible customers who have applied for loan. Decision tree induction data mining algorithm applied to predict the attributes relevant for credibility. A prototype of the model is described in this study which can be used by the institution to making the right decision to approve or reject the loan request of the customers. The dataset used obtained from bank and the data set size (4520) for the experimental analysis. Decision Tree algorithm used for the prediction. The model accuracy is 84% .We noted that the dataset contained a few relevant attributes that may give an inaccurate model, so all important features must be select in the model building stage to give high accuracy .

(Vimala and Sharmili, 2018) compare between tow classifications algorithms Naïve Bayes and Support Vector Machine to predict the status of loans. The dataset obtained from UCI, this study based on accuracy and execution time. The accuracy result obtained: Naïve Bayes accuracy is 77% and Support Vector Machine accuracy is 79% so Naïve Bayes had a low accurate comparing to Support Vector Machine. By comparing execution time of two methods, Naive Bayes had taken more time to execute the model comparing to other

(Sudhamathy, 2016 23) built the model using the data mining functions available in the R package and dataset is taken from the UCI repository with 1000 records and 21 attributes. The tree model is then used to predict the class labels of the new loan applicants several R functions and packages were used to prepare the data and to build the classification model. The work shows that the R package is an efficient visualizing tool that applies data mining techniques comparing by other tools. The model accuracy 94.3% the dataset was balanced to handles unbalanced classification problems.

(Wu, 2010) introduces a case study of applying different data mining technologies in developing a loan risk assessment system for a sub-prime lender. Different data mining methods used in order to produce the results. Weka data mining tool is used. The dataset includes 1000 incessant there are 700 good cases and 300 bad cases. The experiments involved training the models using 70% of a dataset and testing with the remaining 30%. The algorithms used are J48, EM, Naïve Bayes, and FReBE.The

accuracy from the different data mining methods are is 71.44% EM 38.77% Naive Bayes 75.09% FReBE 74.10%. Decision tree is most appropriate data mining technology for developing a loan risk assessment system for sub-prime lenders.

(Batra, 2018)This study introduces a case study of applying different Decision tree learning method and Classification techniques to predict the status of loans this algorithm is ID3, C4.5 and Random Forest. Dataset used for this application is credit approval dataset, obtained from the machine learning repository UCI. Accuracy measure for the above three techniques are: ID3 30% c4.5 53% Random Forest80%Random Forest gives the better prediction result. So among these algorithms, Random Forest is best for accurate classification.

Table2. 1 Summarize the literature review

Authors 'year	Title of paper	Methodology	Result	Gap of Limitation
Hala Hassan Mahmud Hassan 2015	Mining microfinance data using classification and clustering techniques	Classification and clustering using Decision Tree by weka	The details of the successful loans were identified by the results of the Decision tree	92%
Ashish Pandit in 2016	Data Mining on Loan Approved Dataset for predicting Defaulters	Naïve Bayes – Decision Tree, Boosting Classification using weka	Prediction an accuracy of defaulter instances is not good using all the algorithms.	The major reason for this could be the class imbalance i.e. high number of instances having class as 'not defaulters', which results in biased output
Aboobyda Jafar Hamid1 and Tarig Mohammed Ahmed2 2016	DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING	j48, bayesNet and naiveBayes Classification by using weka toolkit	j48 78.3784 % bayesNet 77.4775 % naiveBayes 73.8739 % J48 algorithm is best because it has high accuracy and low mean absolute	attributes selected (Credit history, Purpose, Gender, Credit amount, Age, Housing, Job and the Class (good or bad).)
Hiba Mubarak Musa 2017	Prediction of bank loans by using data mining	Decision Tree j48,Random forest Classification using, Weka, orange	Decision Tree j48 Orange 97%. Weka 89.55 Random forest Orange 97.3 Weka 92	
Sivasree and Rekha Sunny, 2015	Loan Credibility Prediction System Based on Decision Tree Algorithm	Decision Tree Induction Weka tool kit was used	The accuracy achieved is 84%.	the relevant attributes selected is too few that may give an inaccurate model

Vimala and Sharmili, 2018)	Prediction of Loan Risk using Naive Bayes and Support Vector Machine	Naive Bayes SVM Weka used for implementation	Bayes accuracy was obtained 77% and Support Vector Machine accuracy is 79%	
Sudhamathy 2016	Credit Risk Analysis and Prediction Modeling of Bank Loans Using R	Decision Tree R software was used	algorithm Accuracy obtained 94%	
Jia Wu , Karl Dayson , Sunil Vadera 2010	A Comparison of Data Mining Methods in Microfinance	J48, EM, Naïve Bayes, and FReBE Weka used for implementation	is J48 71.44% EM 38.77% NaiveBayes 75.09% FReBE 74.10%	
Shiju Sathyadevan and Remya R. Nair	Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest	Weka used for implementation	Accuracy for the above three techniques are: ID3 30%, c4.5 53% RandomForest80 %	No of instances is 25 No of attributes 21 Less accurate

All of the previous papers focused on the customer data and their details using the science of data mining and classification of data used different algorithms in comparison to high accuracy and performance best in terms of time. This study focused on the details of microfinance operations.

CHAPTER III: METHODOLOGY

3.1 Introduction

This chapter provides a full description of the research methodology. The methodology was carried on three phases in order to achieve the objective of this research. The first phase explains the data preprocessing techniques applied to the dataset under examination. The second presents the tools used for implementation. The third phase presents the tool and algorithm implemented.

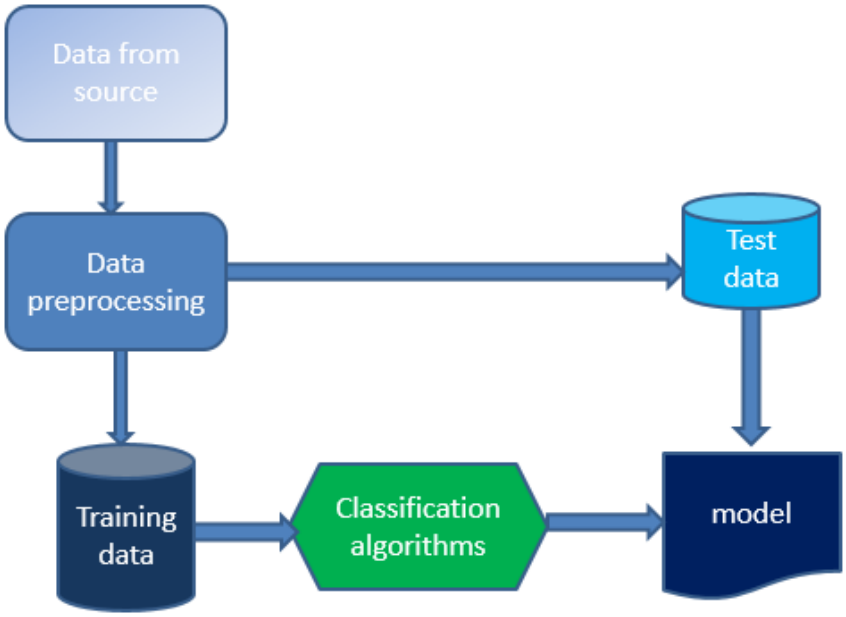


Figure3. 1 Architecture of the methodology

3.2 Data preprocessing

3.2.1 Data Description

Microfinance data is obtained from the Agricultural Bank of Sudan (Investment management) from 2003 to 2018. In this dataset, there are 2199 instances and 19 features the data set format is .Xlsx table 3.1 gives information about the data set.

Table3. 1 information about the data set

NO	The attribute		Data type
1	state	الولاية	Nominal
2	Customer type	نوع العميل	Nominal
3	Authenticator Sum	المبلغ المصدق	Numeric
4	Fringes Profit	هامش الربح	Numeric
5	Premium Monthly	القسط الشهري	Numeric
6	Date	تاريخ المنح	Numeric
7	Finance Sake	أجل التمويل	Numeric
8	Rating Finance	تصنيف التمويل	Nominal
9	Type Finance	نوع التمويل	Nominal
10	Modality Reimbursable	طريقة السداد	Numeric
11	Finance Formula	صيغة التمويل	Nominal
12	Sector	القطاع	Nominal
13	Commodities	السلعة	Nominal
14	Outstanding Balance	الرصيد القائم	Numeric
15	Value Premium Receivable Unsettled	قيمة اقساط مستحقة غير مسددة	Numeric
16	Prescribe Guarantee	وصف الضمان	Nominal
17	Size Finance	حجم التمويل	Nominal
18	Number Premium Receivable	عدد الاقساط المستحقة غير مسددة	Numeric
19	Tripping	التعثر	Nominal

We conducted data exploratory techniques on the dataset in order to understand the nature of the dataset.

3.2.2 Data preprocessing

The data collected for mining process may be contained missing values, noise or inconsistency. This leads to produce inconsistent information from the mining process. A data mining process with high quality of data will produce an efficient data mining results. To improve the quality of data and consequently the mining results, the collected data is to be preprocessed so as to improve the efficiency of data mining process.

In this study some general tasks of the data preprocessing have to be performed on the dataset, such as data integration, data cleaning, data reduction, data transformation.

The first task of the data preprocessing is the Data Filtering the attributes in the bank data set are filtered and the relevant attributes needed for prediction are selected.

The dataset obtained from the bank are not arranged, all the features are nested, so we rearranged similar fields together to make sure they were correct all feature related to money details they putted together, feature relater to Premium also putted together and same to guarantee details. We performed equations on the data to make sure that they were identical the Monthly Premium when multiplied by the Premium Numbers gives the total amount of financing.

The second task of the data preprocessing is handling the missing Data The dataset has missing and imputed data which is replaced in this step, in this study, there are one cases of the missing data will be handling in attribute Last payment which was handled by using the attribute mean for all samples belonging to the same class as the given tuples as shown in Figure 3.4

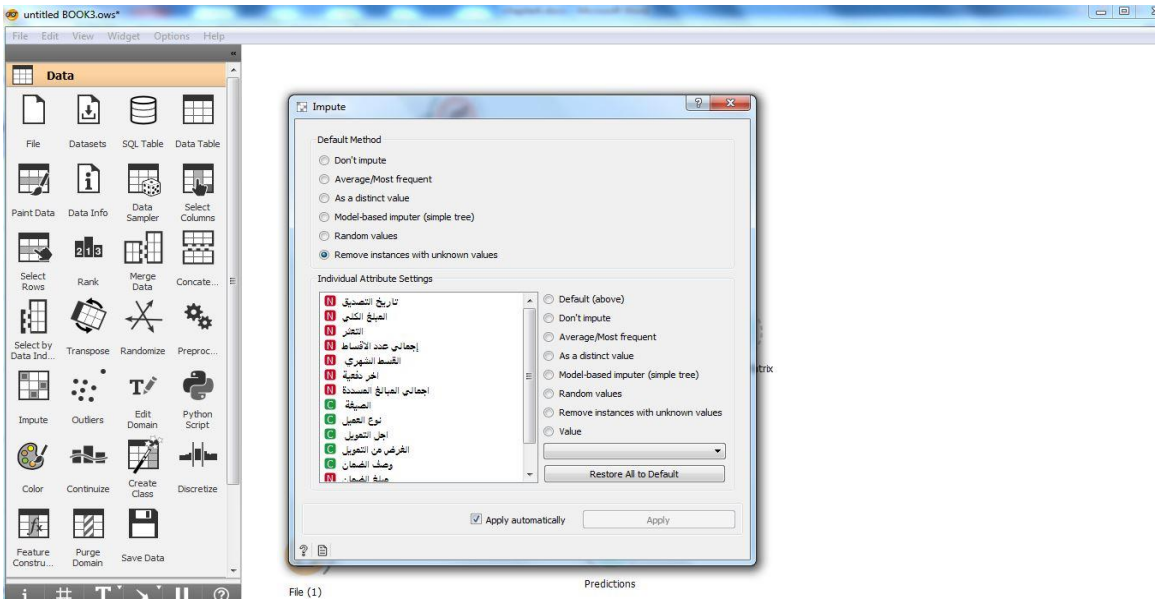


Figure3. 2 handling missing data

The third task of the data preprocessing in this study Data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformations involve many techniques, in this study use normalization technique. Normalizing the courses marks since the Finance Sake is compared by associating the grades with various percentages. The attribute data are scaled so as to fall within a small specified range, as shown in the table: 3.2

Table3. 2 Normalization data by specified range

The range of the finance sake	**Finance Sake**
More than 5 years	long-term
Between two years to 5years	Medium term
Less than one year	short term

E	F	G	H	I	J	K	L	M	N
إجمالي عدد الاقساط	القسط الشهري	اجمالي متعثرا	متعثر ام لا	اخر دفعية	تاريخ نهاية التمويل	اجمالي المبالغ المسددة	الصيغة	نوع العميل	اجل التمويل
1	2,454,796.19	0	NO	8/27/2009	31/05/2009	2454796.19	مرابحة	Individual	متوسط الاجل
5	777,600.00	0	NO	9/14/2009	28/02/2009	777600	مرابحة	Company	قصير الاجل
1	5,130.00	0	NO	2/8/2006	2/1/2006	5130	مرابحة	Individual	قصير الاجل
7	3,560.00	0	NO	7/9/2007	7/10/2006	10680	مرابحة	Individual	قصير الاجل
6	2,702.50	0	NO	12/10/2009	17/04/2009	10810	مرابحة	Individual	قصير الاجل
5	4,086.00	0	NO	2/12/2015	30/03/2015	20430	مرابحة	Individual	متوسط الاجل
6	3,839.83	0	NO	8/17/2016	15/08/2016	23039	مرابحة	Individual	متوسط الاجل
6	5,021.67	0	NO	3/6/2018	5/3/2018	30130	مرابحة	Individual	متوسط الاجل
3	3,538.33	0	NO	1/26/2009	25/11/2008	10615	مرابحة	Individual	قصير الاجل
4	5,408.75	0	NO	2/17/2008	1/1/2008	10817.5	مرابحة	Individual	قصير الاجل
15	10,100.00	0	NO	8/8/2010	3/9/2010	80800	مرابحة	Individual	متوسط الاجل
7	5,160.00	0	NO	9/16/2006	30/11/2005	5160	مرابحة	Individual	قصير الاجل
5	28,903.75	0	NO	1/6/2010	6/1/2010	115615	مرابحة	Company	طويل الاجل
2	4,353.02	0	NO	6/4/2006	6/5/2006	8706.04	مرابحة	Individual	قصير الاجل
3	11,251.22	0	NO	1/21/2007	22/12/2006	33753.66	مرابحة	Individual	قصير الاجل
3	3,606.33	0	NO	12/24/2007	13/08/2007	10818.99	مرابحة	Individual	قصير الاجل
16	1,200.00	0	NO	12/6/2016	26/10/2016	18000	مرابحة	Individual	متوسط الاجل
16	3,594.02	0	NO	7/7/2007	7/7/2007	57504.35	مرابحة	Organizatior	متوسط الاجل
21	13,250.00	0	NO	7/16/2009	30/09/2009	53000	مرابحة	Organizatior	متوسط الاجل
21	2,362.50	0	NO	7/21/2010	27/07/2010	28350	مرابحة	Organizatior	متوسط الاجل
12	2,625.00	0	NO	7/17/2011	27/07/2011	31500	مرابحة	Organizatior	متوسط الاجل
12	4,881.25	0	NO	7/10/2012	27/07/2012	58575	مرابحة	Organizatior	متوسط الاجل
12	4,804.17	0	NO	6/27/2013	27/06/2013	57650	مرابحة	Organizatior	قصير الاجل
12	6,078.67	0	NO	5/18/2014	25/07/2014	72944	مرابحة	Organizatior	متوسط الاجل
1	6,308.00	0	NO	5/18/2014	15/01/2014	6308	مرابحة	Organizatior	قصير الاجل
12	12,996.50	0	NO	5/21/2015	25/05/2015	154878	مرابحة	Organizatior	متوسط الاجل

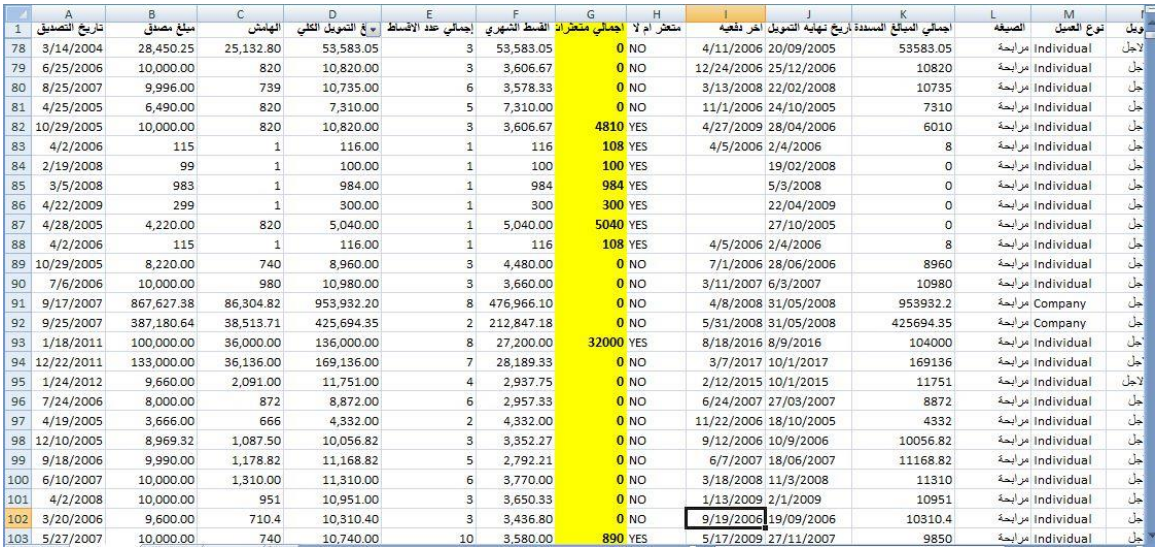
Figure3. 3 Normalize the finance stake

The fourth task of the data preprocessing in this study was Data integration .By merging Authenticator Sum and Fringes Profit we get the total finance Authenticator Sum with Fringes Profit data were merged, and a new attribute named (total finance) was added Figure 3.6 explain merging Authenticator Sum and Fringes Profit .

	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	مبلغ مصدق	الهامش	التمويل الكلي	إجمالي عدد الاقساط	القسط الشهري	اجمالي متعثرا	متعثر ام لا	اخر دفعية	تاريخ نهاية التمويل	اجمالي المبالغ المسددة	الصيغة	نوع العميل	اجل التمويل	
2	2,454,796.19	0	2,454,796.19	1	2,454,796.19	0	NO	8/27/2009	31/05/2009	2454796.19	مرابحة	Individual	متوسط الاجل	راعي
3	720,000.00	57,600.00	777,600.00	5	777,600.00	0	NO	9/14/2009	28/02/2009	777600	مرابحة	Company	قصير الاجل	راعي
5	4,220.00	910	5,130.00	1	5,130.00	0	NO	2/8/2006	2/1/2006	5130	مرابحة	Individual	قصير الاجل	فين
6	10,000.00	680	10,680.00	7	3,560.00	0	NO	7/9/2007	7/10/2006	10680	مرابحة	Individual	قصير الاجل	فين
7	10,000.00	810	10,810.00	6	2,702.50	0	NO	12/10/2009	17/04/2009	10810	مرابحة	Individual	قصير الاجل	فين
8	18,000.00	2,430.00	20,430.00	5	4,086.00	0	NO	2/12/2015	30/03/2015	20430	مرابحة	Individual	متوسط الاجل	ة
9	20,000.00	3,039.00	23,039.00	6	3,839.83	0	NO	8/17/2016	15/08/2016	23039	مرابحة	Individual	متوسط الاجل	ة
10	25,000.00	5,130.00	30,130.00	6	5,021.67	0	NO	3/6/2018	5/3/2018	30130	مرابحة	Individual	متوسط الاجل	ة
11	10,000.00	615	10,615.00	3	3,538.33	0	NO	1/26/2009	25/11/2008	10615	مرابحة	Individual	قصير الاجل	ة
12	10,000.00	817.5	10,817.50	4	5,408.75	0	NO	2/17/2008	1/1/2008	10817.5	مرابحة	Individual	قصير الاجل	ة
13	73,000.00	7,800.00	80,800.00	15	10,100.00	0	NO	8/8/2010	3/9/2010	80800	مرابحة	Individual	متوسط الاجل	فين
14	4,180.00	980	5,160.00	7	5,160.00	0	NO	9/16/2006	30/11/2005	5160	مرابحة	Individual	قصير الاجل	فين
15	95,000.00	20,615.00	115,615.00	5	28,903.75	0	NO	1/6/2010	6/1/2010	115615	مرابحة	Company	طويل الاجل	
16	6,689.28	2,016.76	8,706.04	2	4,353.02	0	NO	6/4/2006	6/5/2006	8706.04	مرابحة	Individual	قصير الاجل	ة
17	31,428.00	2,325.66	33,753.66	3	11,251.22	0	NO	1/21/2007	22/12/2006	33753.66	مرابحة	Individual	قصير الاجل	ة
18	9,999.00	819.99	10,818.99	3	3,606.33	0	NO	12/24/2007	13/08/2007	10818.99	مرابحة	Individual	قصير الاجل	ة
19	15,000.00	3,000.00	18,000.00	16	1,200.00	0	NO	12/6/2016	26/10/2016	18000	مرابحة	Individual	متوسط الاجل	
20	54,521.85	2,982.50	57,504.35	16	3,594.02	0	NO	7/7/2007	7/7/2007	57504.35	مرابحة	Organizatior	متوسط الاجل	ة
21	50,000.00	3,000.00	53,000.00	21	13,250.00	0	NO	7/16/2009	30/09/2009	53000	مرابحة	Organizatior	متوسط الاجل	فين
22	27,000.00	1,350.00	28,350.00	21	2,362.50	0	NO	7/21/2010	27/07/2010	28350	مرابحة	Organizatior	متوسط الاجل	فين
23	30,000.00	1,500.00	31,500.00	12	2,625.00	0	NO	7/17/2011	27/07/2011	31500	مرابحة	Organizatior	متوسط الاجل	فين
24	55,000.00	3,575.00	58,575.00	12	4,881.25	0	NO	7/10/2012	27/07/2012	58575	مرابحة	Organizatior	متوسط الاجل	فين
25	54,905.00	2,745.00	57,650.00	12	4,804.17	0	NO	6/27/2013	27/06/2013	57650	مرابحة	Organizatior	قصير الاجل	فين
26	68,815.50	4,128.50	72,944.00	12	6,078.67	0	NO	5/18/2014	25/07/2014	72944	مرابحة	Organizatior	متوسط الاجل	فين
27	6,184.80	123.2	6,308.00	1	6,308.00	0	NO	5/18/2014	15/01/2014	6308	مرابحة	Organizatior	قصير الاجل	فين
28	150,000.00	1,878.00	154,878.00	12	12,996.50	0	NO	5/21/2015	25/05/2015	154878	مرابحة	Organizatior	متوسط الاجل	فين

Figure3. 4 Merging Authenticator Sum and Fringes Profit

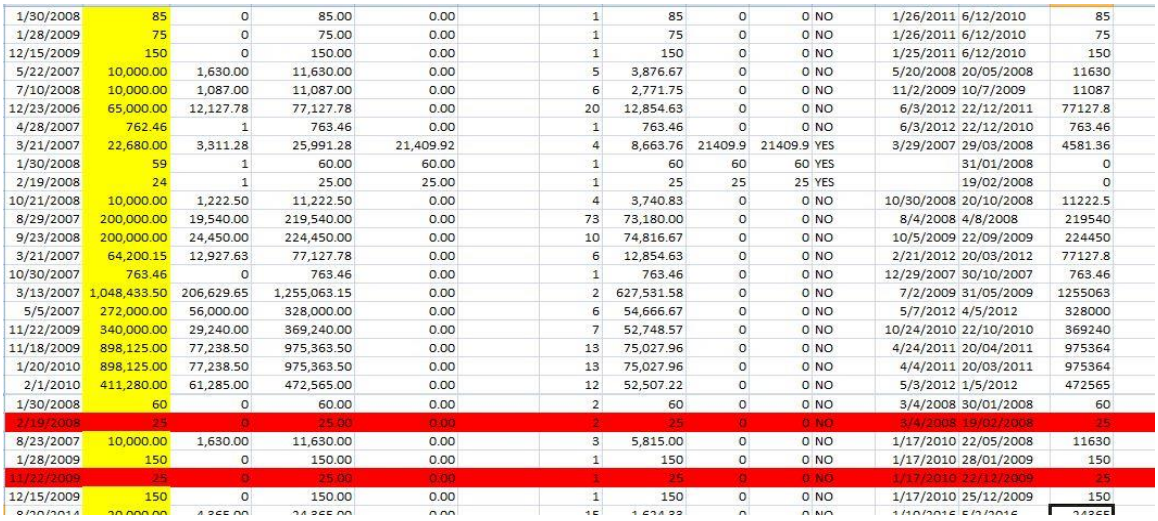
The fifth task of the data preprocessing in this study the dataset doesn't contain the class Tripping (present class label) so we calculated the Tripping by subtract Authenticator Sum from total amount paid as shown in figure 3.7



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	تاريخ التصديق	مبلغ مصدق	الهامش	ع التمويل الكلي	إجمالي عدد الأقساط	القسط الشهري	إجمالي متعثرات	متعثر ام لا	اخر دفعيه	تاريخ نهايه التمويل	اجمالي المبالغ المسددة	الصيغه	نوع العميل	ويل
78	3/14/2004	28,450.25	25,132.80	53,583.05	3	53,583.05	0	NO	4/11/2006	20/09/2005	53583.05	مرابحة	Individual	لاجل
79	6/25/2006	10,000.00	820	10,820.00	3	3,606.67	0	NO	12/24/2006	25/12/2006	10820	مرابحة	Individual	جل
80	8/25/2007	9,996.00	739	10,735.00	6	3,578.33	0	NO	3/13/2008	22/02/2008	10735	مرابحة	Individual	جل
81	4/25/2005	6,490.00	820	7,310.00	5	7,310.00	0	NO	11/1/2006	24/10/2005	7310	مرابحة	Individual	جل
82	10/29/2005	10,000.00	820	10,820.00	3	3,606.67	4810	YES	4/27/2009	28/04/2006	6010	مرابحة	Individual	جل
83	4/2/2006	115	1	116.00	1	116	108	YES	4/5/2006	2/4/2006	8	مرابحة	Individual	جل
84	2/19/2008	99	1	100.00	1	100	100	YES		19/02/2008	0	مرابحة	Individual	جل
85	3/5/2008	983	1	984.00	1	984	984	YES		5/3/2008	0	مرابحة	Individual	جل
86	4/22/2009	299	1	300.00	1	300	300	YES		22/04/2009	0	مرابحة	Individual	جل
87	4/28/2005	4,220.00	820	5,040.00	1	5,040.00	5040	YES		27/10/2005	0	مرابحة	Individual	جل
88	4/2/2006	115	1	116.00	1	116	108	YES	4/5/2006	2/4/2006	8	مرابحة	Individual	جل
89	10/29/2005	8,220.00	740	8,960.00	3	4,480.00	0	NO	7/1/2006	28/06/2006	8960	مرابحة	Individual	جل
90	7/6/2006	10,000.00	980	10,980.00	3	3,660.00	0	NO	3/11/2007	6/3/2007	10980	مرابحة	Individual	جل
91	9/17/2007	867,627.38	86,304.82	953,932.20	8	476,966.10	0	NO	4/8/2008	31/05/2008	953932.2	مرابحة	Company	جل
92	9/25/2007	387,180.64	38,513.71	425,694.35	2	212,847.18	0	NO	5/31/2008	31/05/2008	425694.35	مرابحة	Company	جل
93	1/18/2011	100,000.00	36,000.00	136,000.00	8	27,200.00	32000	YES	8/18/2016	8/9/2016	104000	مرابحة	Individual	جل
94	12/22/2011	133,000.00	36,136.00	169,136.00	7	28,189.33	0	NO	3/7/2017	10/1/2017	169136	مرابحة	Individual	جل
95	1/24/2012	9,660.00	2,091.00	11,751.00	4	2,937.75	0	NO	2/12/2015	10/1/2015	11751	مرابحة	Individual	لاجل
96	7/24/2006	8,000.00	872	8,872.00	6	2,957.33	0	NO	6/24/2007	27/03/2007	8872	مرابحة	Individual	جل
97	4/19/2005	3,666.00	666	4,332.00	2	4,332.00	0	NO	11/22/2006	18/10/2005	4332	مرابحة	Individual	جل
98	12/10/2005	8,969.32	1,087.50	10,056.82	3	3,352.27	0	NO	9/12/2006	10/9/2006	10056.82	مرابحة	Individual	جل
99	9/18/2006	9,990.00	1,178.82	11,168.82	5	2,792.21	0	NO	6/7/2007	18/06/2007	11168.82	مرابحة	Individual	جل
100	6/10/2007	10,000.00	1,310.00	11,310.00	6	3,770.00	0	NO	3/18/2008	11/3/2008	11310	مرابحة	Individual	جل
101	4/2/2008	10,000.00	951	10,951.00	3	3,650.33	0	NO	1/13/2009	2/1/2009	10951	مرابحة	Individual	جل
102	3/20/2006	9,600.00	710.4	10,310.40	3	3,436.80	0	NO	9/19/2006	19/09/2006	10310.4	مرابحة	Individual	جل
103	5/27/2007	10,000.00	740	10,740.00	10	3,580.00	890	YES	5/17/2009	27/11/2007	9850	مرابحة	Individual	جل

Figure3. 5 Tripping calculated

The sixth task of the data preprocessing in this study is trying to solve the outlier problem. The data set contains outlier values in the attribute total finance, these values can effect in result's accuracy, there are 177 rows whose 'authenticator sum' value is 'outliers so these rows would be deleted as shown in figure 3.8



1/30/2008	85	0	85.00	0.00		1	85	0	0	NO	1/26/2011	6/12/2010	85
1/28/2009	75	0	75.00	0.00		1	75	0	0	NO	1/26/2011	6/12/2010	75
12/15/2009	150	0	150.00	0.00		1	150	0	0	NO	1/25/2011	6/12/2010	150
5/22/2007	10,000.00	1,630.00	11,630.00	0.00		5	3,876.67	0	0	NO	5/20/2008	20/05/2008	11630
7/10/2008	10,000.00	1,087.00	11,087.00	0.00		6	2,771.75	0	0	NO	11/2/2009	10/7/2009	11087
12/23/2006	65,000.00	12,127.78	77,127.78	0.00		20	12,854.63	0	0	NO	6/3/2012	22/12/2011	77127.8
4/28/2007	762.46	1	763.46	0.00		1	763.46	0	0	NO	6/3/2012	22/12/2010	763.46
3/21/2007	22,680.00	3,311.28	25,991.28	21,409.92		4	8,663.76	21409.9	21409.9	YES	3/29/2007	29/03/2008	4581.36
1/30/2008	59	1	60.00	60.00		1	60	60	60	YES		31/01/2008	0
2/19/2008	24	1	25.00	25.00		1	25	25	25	YES		19/02/2008	0
10/21/2008	10,000.00	1,222.50	11,222.50	0.00		4	3,740.83	0	0	NO	10/30/2008	20/10/2008	11222.5
8/29/2007	200,000.00	19,540.00	219,540.00	0.00		73	73,180.00	0	0	NO	8/4/2008	4/8/2008	219540
9/23/2008	200,000.00	24,450.00	224,450.00	0.00		10	74,816.67	0	0	NO	10/5/2009	22/09/2009	224450
3/21/2007	64,200.15	12,927.63	77,127.78	0.00		6	12,854.63	0	0	NO	2/21/2012	20/03/2012	77127.8
10/30/2007	763.46	0	763.46	0.00		1	763.46	0	0	NO	12/29/2007	30/10/2007	763.46
3/13/2007	1,048,433.50	206,629.65	1,255,063.15	0.00		2	627,531.58	0	0	NO	7/2/2009	31/05/2009	1255063
5/5/2007	272,000.00	56,000.00	328,000.00	0.00		6	54,666.67	0	0	NO	5/7/2012	4/5/2012	328000
11/22/2009	340,000.00	29,240.00	369,240.00	0.00		7	52,748.57	0	0	NO	10/24/2010	22/10/2010	369240
11/18/2009	898,125.00	77,238.50	975,363.50	0.00		13	75,027.96	0	0	NO	4/24/2011	20/04/2011	975364
1/20/2010	898,125.00	77,238.50	975,363.50	0.00		13	75,027.96	0	0	NO	4/4/2011	20/03/2011	975364
2/1/2010	411,280.00	61,285.00	472,565.00	0.00		12	52,507.22	0	0	NO	5/3/2012	1/5/2012	472565
1/30/2008	60	0	60.00	0.00		2	60	0	0	NO	3/4/2008	30/01/2008	60
2/19/2008	25	0	25.00	0.00		1	25	0	0	NO	3/4/2008	19/02/2008	25
8/23/2007	10,000.00	1,630.00	11,630.00	0.00		3	5,815.00	0	0	NO	1/17/2010	22/05/2008	11630
1/28/2009	150	0	150.00	0.00		1	150	0	0	NO	1/17/2010	28/01/2009	150
11/22/2009	25	0	25.00	0.00		1	25	0	0	NO	1/17/2010	22/12/2009	25
12/15/2009	150	0	150.00	0.00		1	150	0	0	NO	1/17/2010	25/12/2009	150

Figure3. 6 Rows deleted

Data mining algorithms may give poor results due to class imbalance problem, so the data already built with balance consideration in order to improve the accuracy.

The percentage of instances that have missing value after preprocessing is 0% and the dataset size reduced to 1744 from 2199 instances as show in figure 3.9

	متعثر ام لا	تاريخ نهاية التمويل	تاريخ التصديق	المبلغ الكلي	إجمالي عدد الأقساط	القسط الشهري	اجمالي متعثرات
1	NO	31/05/2009	39165.0	2454796.189999...	1.0	2454796.19	0.00
2	NO	28/02/2009	39632.0	777600.0000000...	5.0	777600.00	0.00
3	NO	28/02/2009	39635.0	617760.0000000...	2.0	617760.00	0.00
4	NO	38749.0	38505.0	5130.000000000...	1.0	5130.00	0.00
5	NO	38908.0	38784.0	10680.00000000...	7.0	3560.00	0.00
6	NO	17/04/2009	39646.0	10810.00000000...	6.0	2702.50	0.00
7	NO	30/03/2015	41637.0	20430.00000000...	5.0	4086.00	0.00
8	NO	15/08/2016	42050.0	23039.00000000...	6.0	3839.83	0.00
9	NO	43223.0	42618.0	30130.00000000...	6.0	5021.67	0.00
10	NO	25/11/2008	39590.0	10615.00000000...	3.0	3538.33	0.00
11	NO	39448.0	39260.0	10817.50000000...	4.0	5408.75	0.00
12	NO	40246.0	39785.0	80800.00000000...	15.0	10100.00	0.00
13	NO	30/11/2005	38432.0	5160.000000000...	7.0	5160.00	0.00
14	NO	40330.0	39089.0	115615.0000000...	5.0	28903.75	0.00
15	NO	38873.0	38683.0	8706.039999999...	2.0	4353.02	0.00
16	NO	22/12/2006	38890.0	33753.66000000...	3.0	11251.22	0.00

Figure3. 7 dataset after preprocessing

3.3 Model Implementation

In the Research, the decision trees random forest, KNN and NaiveBayes method which is one of the classification models in data mining used. And also Orange toolkit used.

3.3.1 Classification

There are two forms of data analysis that can be used for extracting models describing important classes and to predict future data trends. These two forms Classification and Prediction **(Lin et al., 2002)**

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

The prediction is how the model will predict the classification of new coming data? (Can the model classify the new data correctly?) (Hand, 2006)

3.3.2 Basic decision tree concept

Decision tree concept is more to the rule based system. Given the training dataset with targets and features, the decision tree algorithm will come up with some set of rules. The same set rules can be used to perform the prediction on the test dataset. In decision tree algorithm calculating nodes and forming the rules will happen using the information gain and Gini index calculations (Han et al., 2011) .

3.3.3 Random Forest

Random forest is an ensemble learning method used for classification, regression and other tasks. It was first proposed by Tin Kam Ho and further developed by Leo Breiman (Breiman, 2001) and Adele Cutler. As the name suggest, this algorithm creates the forest with a number of trees.(Rawate and Tijare, 2017) .

It’s builds a set of decision trees. Each tree is developed from a bootstrap sample from the training data. When developing individual trees, an arbitrary subset of attributes is drawn (hence the term “Random”), from which the best attribute for the split is selected. The final model is based on the majority vote from individually developed trees in the forest. (Rawate and Tijare, 2017) .

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

In random forest algorithm, instead of using information gain or Gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.

3.3.3.1 How does it work?

In Random Forest, we grow multiple trees as opposed to a single tree in CART model. To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest)

It works in the following manner. Each tree is planted & grown as follows:

- Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but with replacement. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on this m is used to split the node. The value of m is held constant while we grow the forest.
- Each tree is grown to the largest extent possible and there is no pruning.
- Predict new data by aggregating the predictions of the n tree trees (i.e., majority votes for classification, average for regression) (Rawate and Tijare, 2017)

3.3.2.3 Advantages of Random Forest

This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.

One of the most benefits of Random forest most is the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. (Rawate and Tijare, 2017) .

3.3.4 (K -Nearest Neighbor)

As mention in {Anuradha, 2015 #3} is a k-nearest-neighbor classifier that uses the same distances metric. The number of nearest neighbors can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbors. A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees”. The distance function used is a parameter of the search method. The remaining thing is the same as for IBL—that is, the Euclidean distance; other options include Chebyshev, Manhattan, and Minkowski distances. Predictions from more than one neighbor can be weighted according to their distance from the test instance and two different formulas are implemented for converting the distance into a weight. The number of training instances kept by the classifier can be restricted by setting the window size option.

3.3.5 Naive Bayes

As mention in {Gokilam, 2016 #2} The Naïve Bayesian classifier is based on Bayes’ theorem with independence assumptions between predictors. Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. It is used the knowledge of prior events to predict future events. It builds, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability,

P(c|x), from P(c), P(x), and P(x|c) .Naive Bayes classifier assume that the effect of the value of a predictor(x) on a given class(c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x) = \frac{P(x|c)\,P(c)}{p(x)}$$

- P [(c/x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x/c) is probability of predictor given class.
- P(x) is the prior probability of predictor class

3.4 Orange

Orange Is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization, and can also be used as a Python library.(Singh and Singh, 2010)

3.4.1 Orange Features

Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. The user can interactively explore visualizations or feed the selected subset into other widgets (Singh and Singh, 2010).

CHAPTER IV: IMPLEMENTITION AND RESULT

4.1 Introduction

This chapter introduces the results which have been conducted through execution of three experiments. The results were carried on three phases, first phases present Measures and Metrics. Second phases explain the result of experiments has been applied. Third phases show models predictions on the data.

4.2 Measures and Metrics

To evaluate the effectiveness of our methods, experiments.

4.2.1 Confusion Matrix

Confusion matrix is the measure of performance of a multi-label/multi-class classification model. It is also referred to as error matrix or a table of confusion. This is used in predictive analytics to understand what type of data is being labeled as 'true' and what kind of data is labeled as 'false' by the classifier or the classification model chosen. In summary, Confusion matrix tells us how the classification algorithm is performing with respect to the ground.

Actual Class	**Predicted Class**	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

Figure4. 1 Confusion matrix

TP refers to positive tuples and TN refers to negative tuples classified by the basic classifiers. Similarly FP refers to positive tuples and FN refers to negative tuples which is being incorrectly classified by the classifiers**.**

4.2.2 Accuracy Measures:

Accuracy measure represents how far the set of tuples are being classified correctly

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

4.3 First Experiment

In this experiment, Random forest algorithm was applied on the data set using all the features and instant. The experiments have been done several times and in each time the training and test sets size have been changed (80% training 20% test set, 60% training 40% test and 70% training 30% test) and we obtained the best result when the splitting is 80% training 20% test set. The accuracy achieved 89.8%.

Table4. 1 Random forest classifier Confusion matrix with full data set

Algorithms	TP %	FN %	FP %	TN %
Random forest with full dataset	86.7	13.3	1.2	98.8

Table4. 2 detailed Random forest accuracy with full data set

Algorithms	**Accuracy %**	**Precision %**	**Recall %**
Random forest with full dataset	89.8	90.7	89.8

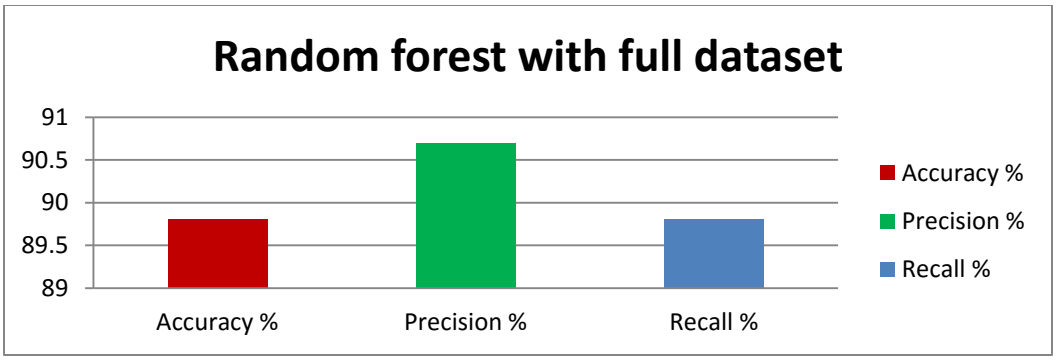


Figure4. 2 detailed Random forest accuracy with full data set

After first experience there are 177 rows whose 'Authenticator Sum' value is 'outliers or have values not align with other dataset so data set was reduced by deleted these rows.

4.4 Second Experiments

In this experiment we built classifier used same algorithms after reduced dataset, Also the experiments have been done several times and in each time the training and test sets size have been changed (80% training 20% test set, 60% training 40% test and 70% training 30% test) and we obtained the best result when the splitting is 80% training 20% test set. The accuracy achieved 94.6 %.

Table4. 3 Random forest Confusion matrix with preprocessed data set

Algorithm	TP %	FN %	FP %	TN %
Random forest	94.4	5.6	4.9	95.1

Table4. 4 detailed Random forest accuracy with preprocessed data set

Algorithm	Accuracy %	Precision %	Recall %
Random forest	94.6	94.6	94.6

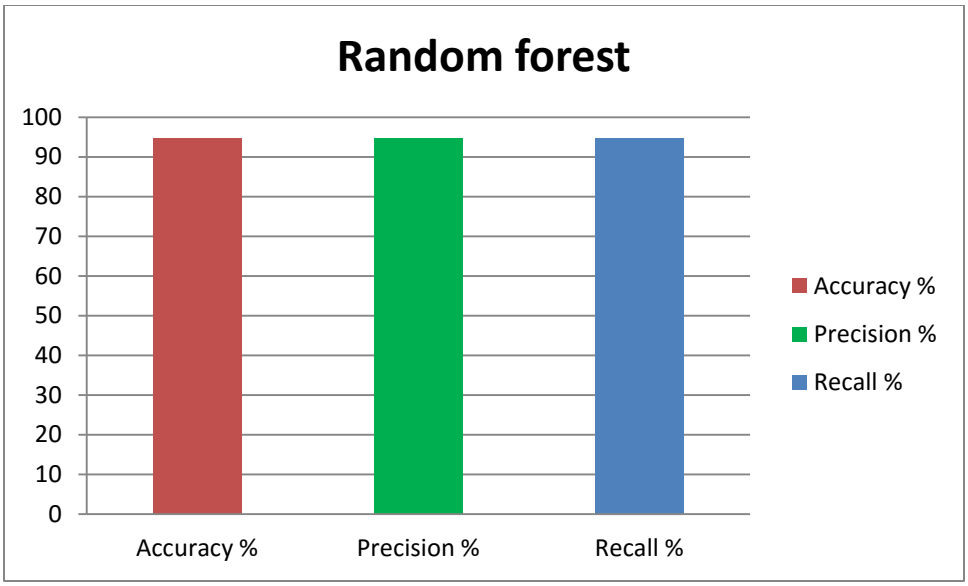


Figure4. 3detailed Random forest accuracy with preprocessed data set

So in this experiment the accuracy was better than before and there was improvement in accuracy and the confusion matrix was very satisfactory. The classifier with preprocessed data set performed well and given better results than using it in data without processing. The preprocessed dataset will be used in the rest of the experiments

4.5 Third Experiments

In this experiment, we conducted three data mining algorithms (Random forest, NaiveBayes and KNN) after reduced dataset, we obtained the best result when the splitting is 80% training 20% test set.

The experiments show that highest accuracy (94.6%) was achieved by Random Forest model. KNN ranked secondly its yielded (92.3 %) and (87. 4%) for NaiveBayes.

Table4. Detailed Confusion matrix between the classifiers

Algorithms	**TP %**	**FN %**	**FP %**	**TN %**
Random forest	94.4	5.6	4.9	95.1
NaiveBayes	92.6	7.4	30.7	69.3.
KNN	92.2	7.8	7.6	92.4

Table4. 5 Comparing accuracy between classifiers

Technique	Accuracy	Precision %	Recall %
Random forest	94.6	94.6	94.6
NaiveBayes	87.4	87.7	87.4
KNN	92.3	91.8	92.3

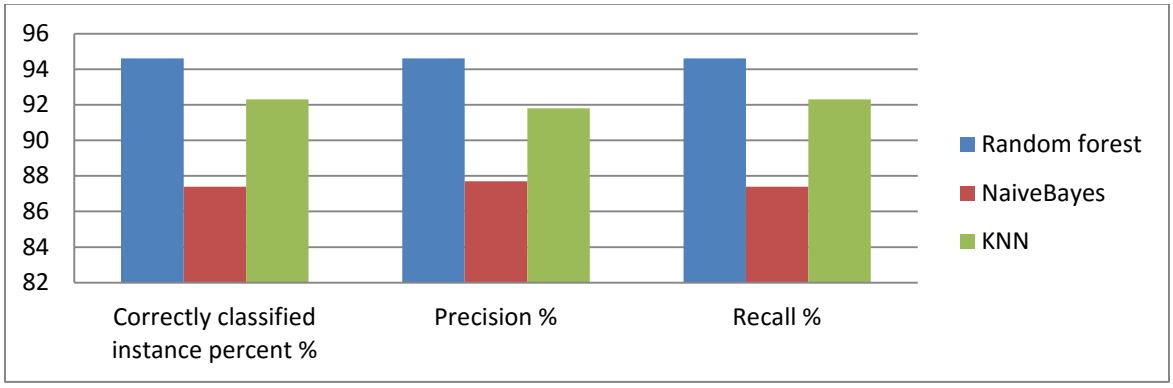


Figure4. 4 Comparing between classifiers

4.6 Predictions

Shows models predictions on the data, Predictions also need the data to predict on. We will use the output of data training for prediction, but this time not the data training, but the remaining data, this is the data that wasn't used for training the model.

	تاريخ نهاية التمويل	تاريخ التصديق	مبلغ مصدق	الهامش	المبلغ الكلي	إجمالي عدد الأقساط	القسط الشهري
1	38749.0	38505.0	4220.00	910.00	5130.00000000000	1.0	5130.00
2	38355.0	38293.0	7692.25	783.75	8476.00000000000	3.0	8476.00
3	18/01/2004	37759.0	77430.40	41437.50	118867.8999999...	2.0	118867.90
4	22/02/2008	39319.0	9996.00	739.00	10735.00000000...	6.0	3578.33
5	28/04/2006	38654.0	10000.00	820.00	10820.00000000...	3.0	3606.67
6	39571.0	39512.0	983.00	1.00	984.00000000000	1.0	984.00
7	31/05/2008	39350.0	387180.64	38513.71	425694.3500000...	2.0	212847.18
8	43009.0	40899.0	133000.00	36136.00	169136.0000000...	7.0	28189.33

Figure4. 5 Testing data without class label

For predictions we need both the training data, which we have loaded in the first datasets widget and the data to predict, which we will load in another datasets widget. We

will use Attrition - Predict data this time. Connect the second data set to predictions. Now we can see predictions for the three data instances from the second data set.

	Random Forest	Naive Bayes	kNN	تاريخ نهاية التعويل	تاريخ التصديق	مبلغ مصدق	الهامش	المبلغ الكلي	إجمالي عدد الأقساط
1	0.83 : 0.17 → NO	1.00 : 0.00 → NO	1.00 : 0.00 → NO	38749.0	38505.0	4220.00	910.00	5130.00000000000	1.0
2	0.39 : 0.61 → YES	0.65 : 0.35 → NO	0.00 : 1.00 → YES	38355.0	38293.0	7692.25	783.75	8476.00000000000	3.0
3	0.27 : 0.73 → YES	0.01 : 0.99 → YES	0.00 : 1.00 → YES	18/01/2004	37759.0	77430.40	41437.50	118867.8999999...	2.0
4	0.88 : 0.12 → NO	1.00 : 0.00 → NO	1.00 : 0.00 → NO	22/02/2008	39319.0	9996.00	739.00	10735.00000000...	6.0
5	0.39 : 0.61 → YES	0.78 : 0.22 → NO	0.00 : 1.00 → YES	28/04/2006	38654.0	10000.00	820.00	10820.00000000...	3.0
6	0.29 : 0.71 → YES	0.34 : 0.66 → YES	0.00 : 1.00 → YES	39571.0	39512.0	983.00	1.00	984.00000000000	1.0
7	0.78 : 0.22 → NO	0.97 : 0.03 → NO	1.00 : 0.00 → NO	31/05/2008	39350.0	387180.64	38513.71	425694.3500000...	2.0
8	0.50 : 0.50 → YES	0.94 : 0.06 → NO	1.00 : 0.00 → NO	43009.0	40899.0	133000.00	36136.00	169136.0000000...	7.0

Figure4. 6 Random forest, NaiveBayes, KNN prediction result in test class

4.7 ROC Analysis

Plots true positive rate against a false positive rate of a test.

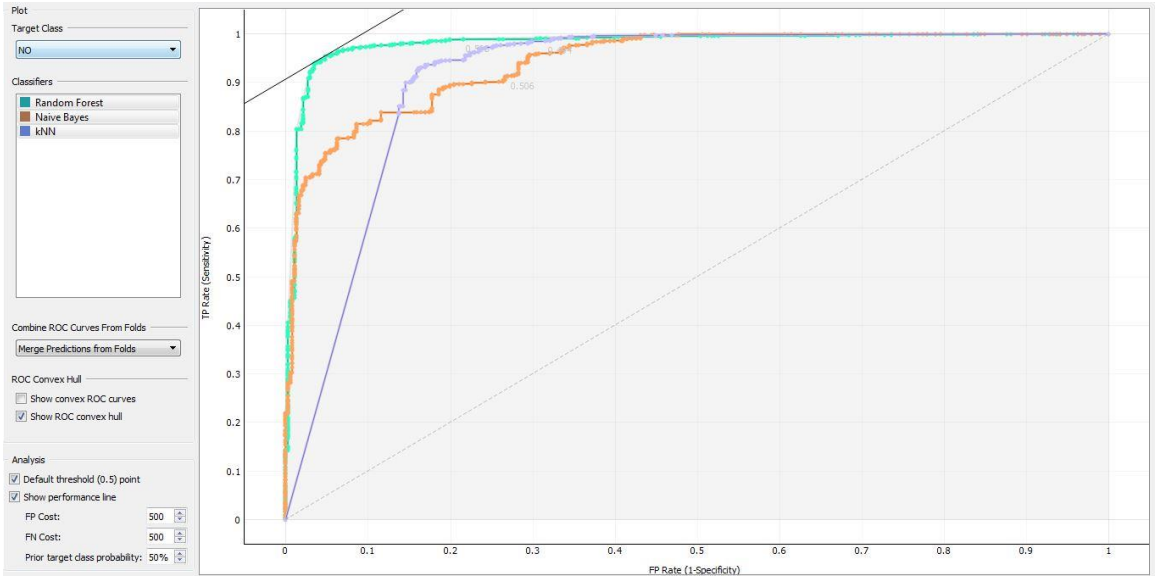


Figure4. 7 Test class No

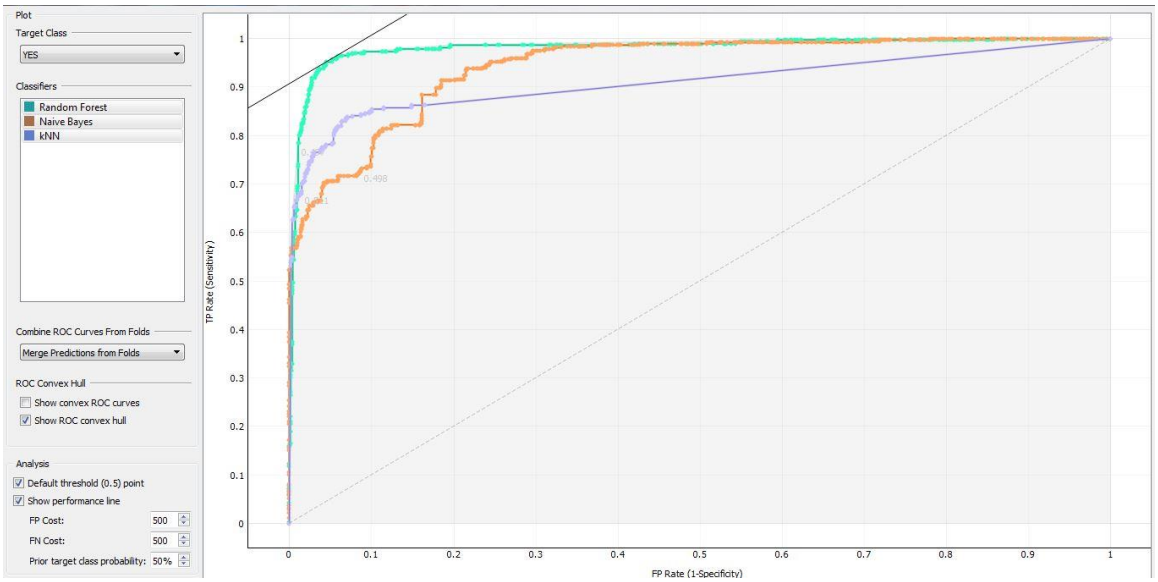


Figure4. 8 Test class yes

4.8 Discussion

According to the results above experiments the second experiment shown that preprocessing techniques is an important issue in classification, because it has a considerable effect on accuracy of the classifier.

And according to microfinance operation details, the most successful microfinance sectors are agriculture a then commercial and the most failing microfinance project have few Premiums. Most of the financing operations in terms of Guarantee are done by personal guarantee. We note more tripping was happened in recent years that may have been due to low economic. There are no field in the dataset refer to follow up of funded projects from beginning to ending of the loan, and make sure it has been implemented, this may be one of the reasons for the failure.

CHAPTER V: CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter explains the research conclusion and recommendation

5.2 Conclusion

The main purpose of this research is to increase the performance of the bank by building models that can be used to predict defaulter's in order to increase the performance in the bank and right decision making (business intelligence).

The study applied on real data which obtained from investment management in Agricultural Bank of Sudan. Some preprocessing phases are applied on data such as handle missing data, data transformation. Random forest, NaiveBayes and KNN algorithms were used to build predictive models that can be used to predict and classify the applications of microfinance. The model has been implemented by using orange application. After applying classification's data mining technique algorithms which are Random forest, NaiveBayes and KNN we find that the best algorithm for Microfinance loan classification is Random forest algorithm. Random forest is best because it has highest accuracy (94. 6%).Orange Program has a simple, clear and creative graphic interface, easy to use and the ability to extract reports.

From the results of the experiment, it can be concluded that the data mining tools and techniques, especially classification techniques, can be effectively applied on the microfinance and financial institutions data in order to generate predictive models with an acceptable level of accuracy. The outcome of the study is highly useful for the microfinance institutes in developing or revising existing loan disbursement and collection policies.

5.2 future work

- Mining large amount of wasted information, and finding links between customers to develop future strategies that increase customer attraction for investment and also help to predict the future of microfinance work in the bank.
- Using other data mining algorithms and tools to discover factors that will attract new types in the microfinance sector.
- Using other data mining tools and comparing them.

5.3 Recommendation

- follow up funded projects from the beginning to avoid defaulter

5.3 References

Credit Risk Analysis and Prediction Modelling of Bank Loans Using R.

AMALA JAYANTHI.M*, S. S., THARAKAI.R 2016. Data Mining– A Survey. ISSN: 2277 128X, Volume 6,.

BHARATI, M. & RAMAGERI, M. 2010. Data mining techniques and applications.

HAMID, A. J. & AHMED, T. M. 2016. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAIJ) Vol, 3.

HAN, J., PEI, J. & KAMBER, M. 2011. Data Mining: Concepts and Techniques, Elsevier Science.

HAND, D. J. 2006. Data Mining. Encyclopedia of Environmetrics, 2.

HERMES, N., LENSINK, R. & MEESTERS, A. 2011. Outreach and efficiency of microfinance institutions. World development, 39**,** 938-948.

HUSSAIN, S. 2017. Survey on Current Trends and Techniques of Data Mining Research. London Journal of Research in Computer Science and Technology (03 2017).

KOKONYA, L. N. 2014. Data Mining And Performance Of Microfinance Institutions In Kenya. University Of Nairobi.

LIN, W., ORGUN, M. A. & WILLIAMS, G. J. An Overview Of Temporal Data Mining. AusDM, 2002. 83-90.

MUSA, H. M. 2017. Prediction of bank loans by using data mining.

PADHY, N., MISHRA, D. & PANIGRAHI, R. 2012. The survey of data mining applications and feature scope. arXiv preprint arXiv:1211.5723.

PANDIT, A. 2016. DATA MINING ON LOAN APPROVED DATSET FOR PREDICTING DEFAULTERS. Rochester Institute of Technology.

RAWATE, K. & TIJARE, P. 2017. Review on prediction system for bank loan credibility. International Journal of Advance Engineering and Research Development, 4**,** 860-867.

SAHU, H., SHRMA, S. & GONDHALAKAR, S. 2011. A brief overview on data mining survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE), 1**,** 114-121.

SINGH, B. & SINGH, H. K. Web data mining research: a survey. 2010 IEEE International Conference on Computational Intelligence and Computing Research, 2010. IEEE, 1-10.

SIVASREE, M. 2015. Loan Credibility Prediction System Based on Decision Tree Algorithm. International Journal of Engineering Research & Technology (IJERT).

SUDHAMATHY, G. 2016. Credit risk analysis and prediction modelling of bank loans using R. Int. J. Eng. Technol, 8**,** 1954-1966.

VENTO*, M. L. T. A. G. 2007. Banks in the Microfinance Market.

VIMALA, S. & SHARMILI, K. Prediction of loan risk using naive bayes and support vector machine. Int Conf Adv Comput Technol (ICACT), 2018. 110-113.

ZHU, X. & DAVIDSON, I. 2007. Knowledge discovery and data mining: challenges and realities, Information Science Reference Hershey, PA.

Cheng-Lung Huang a,*, Mu-Chen Chen b, Chieh-Jen Wang c,"Credit scoring with a data mining approach based on supportvector machines", ,ELSEVIER, C.-L. Huang et al. / Expert Systems with Applications 33 (2007) 847–856.

KetakiChopde, Pratik Gosar, ParasKapadia, NiharikaMaheshwari, Pramila M. Chawan, "A Study of Classification Based Credit Risk Analysis Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-4, April 2012.

MrunalSurve, PoojaThitme, PriyaShinde, Swati Sonawane, SandipPandit, “DATA MINING TECHNIQUES TO ANALYSES RISK GIVING LOAN(BANK)”, IJARIIEISSN(O)-2395-4396 Vol-2 Issue-1 2016.

AboobydaJafar Hamid and Tarig Mohammed Ahmed, “DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING”, Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016.

R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, 2016, pp. 1-7.doi: 10.1109/ICAIPR.2016.7585216

Getachew Hailemariam, Hill Shawndra, and Sintayehu Demissie, “Exploring Data Mining Techniques and Algorithms for Predicting Customer Loyalty and Loan Default Risk Scenarios at Wisdom Microfinance”, October 28-31, Addis Ababa, Ethiopia, 2012.

J. H. Aboobyda, and M.A. Tarig, “Developing Prediction Model Of Loan Risk In Banks Using Data Mining”, Machine Learning and Applications: An International Journal (MLAIJ), vol. 3(1), pp. 1–9, 2016.

https://www.wideskills.com/data-mining/challenges-in-data-mining

https://www.analyticsinsight.net/the-top-10-data-mining-tools-of-2018/