AFRICAN CENTER OF EXCELLENCE
IN DATA SCIENCE

PREDICTIVE MODEL FOR LOAN USING DATA MINING TECHNIQUES
CASE STUDY: BUSINESS DEVELOPMENT FUND

by

JUSTIN MURENZI
Registration number: 213001757

A Dissertation submitted in partial fulfilment of the requirement for the award of Master of Data Science in Data Mining

UNIVERSITY OF RWANDA

Supervisor**:** Dr. Christine NIYIZAMWIYITIRA

September, 2020

DECLARATION

I declare that this dissertation entitled “PREDICTIVE MODEL FOR LOAN USING DATA MINING TECHNIQUES, CASE STUDY: BDF” is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Justin MURENZI

Signature 

Date: 30th September, 2020

APPROVAL SHEET

This dissertation entitled "**PREDICTIVE MODEL FOR LOAN USING DATA MINING TECHNIQUES, CASE STUDY: BDF**", written and submitted by **MURENZI JUSTIN** in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Data mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 16% which is less than 20% accepted by ACE-DS.

_____

Supervisor

Dr. Christine NIYIZAMWIYITIRA, PhD

_____

Head of Training

DEDICATION

This research is dedicated to

The Almighty One

My lovely family

Friends and Relatives

ACKNOWLEDGMENTS

This dissertation would not have been a success without the concerted efforts and support from distinguished people that have contributed directly or indirectly to me many thanks to you all for your kind support in various forms.

My thanks go first to the Almighty God for giving me the strength and health to do this research until the end. Then, I thank the Government of Rwanda to promote the science through this Centre, I also thank all the academic administrative staff of University of Rwanda, College of business and Economics. The particular thanks go to the head and staff of African Centre for Excellence for the time, support and innumerable help, and their effort to ensure that the academic research is successfully completed.

Special thanks go to My Supervisor, Dr. Christine NIYIZAMWIYITIRA who agreed to supervise me especially for my research and she gave me valuable guidance and advice that has helped me produce this piece of work

Thank you to all of my colleagues of the promotion, and in particular my everyday companions with whom I shared almost everything during our academic career.

Finally, I thank my sisters and brothers who always wished me success.

Justin MURENZI

ABSTRACT

Data mining is the process of knowledge discovery, attempts to discover useful information or patterns in large data repositories such as databases; that is why the data experts are interested in how data can be collected, stored, accessed and combined for the analysis to extract useful knowledge for the public including financial institutions and other sectors.

The Business Development Funds (BDF) aims to promote SMEs development through the provision of financial services to enhance the lending mechanism of financial institutions. As part of the financial infrastructure to promote SMEs, it was established with the objective of assisting SMEs to access finance with ease, particularly those without sufficient collateral to obtain credit from traditional financial institutions at reasonable rates. The BDF conducts a number of activities including guaranteeing loans for SMEs, and providing financial education services to SMEs in Rwanda. The SME sector, including formal and informal businesses, comprises 98% of the businesses in Rwanda and 41% of all private sector employment (Minicom, 2010; OECD, 2011).

In recent years, machine learning has become a popular field in big data analytics because of its success in learning complicated models. Methods such as decision tree, support vector machines, logistic regression and artificial neural networks can be used for recognizing patterns in the data (with a high degree of accuracy) that may not be apparent to human analysts, The reason why applications of data science using machine learning is important in such organisation and in all financial institutions.

Due to the advanced technology associated with big data, data availability and computing power, most financial or lending institutions are renewing their business models. Loan predictions, monitoring, model reliability and effective loan processing are key to decision-making and transparency. In this research, we will visualize data and build binary classifiers based on machine and deep learning models on real data in predicting loan default probability. The important features from these models are selected and then used in the modeling process to test the stability of classifiers by comparing their performance on separate data. After analysis and visualization of data, we used different models like decision tree, random forest, logistic regression and artificial neural networks to make a real comparison of good predictors in this case.

Keywords: *SMEs, Financial institution, machine learning, data mining, data science, big data*

TABLE OF CONTENTS

DECLARATION i
APPROVAL SHEET ii
DEDICATION iii
ACKNOWLEDGMENT iv
ABSTRACT v
TABLE OF CONTENTS vi
LIST OF FIGURES viii
LIST OF TABLES x
ABBREVIATIONS xi
CHAPTER 1: GENERAL INTRODUCTION 1
1.1. Introduction 1
1.2. Research problem 2
1.3. General objectives 4
1.4. Specific objectives 4
1.5. Research Questions 4
1.6. Significance of the study 4
1.6.1. Personal interest 4
1.6.2. Academic interest 5
1.6.3. Organizational interest 5
1.6.4. Stakeholders 'interest 5
1.7. Justification 5
1.8. Scope of study 5
1. 8. 1. Content scope 5
1. 8. 2. Time scope 5
1. 8. 3. Geographical scope 6
CHAPTER 2: LITERATURE REVIEW 7
2.1. INTRODUCTION 7
2.2. KEYWORDS 8
2.3. INTRODUCTION OF BDF 12
2.3.1. Background of the Organization 12
2.3.2. SOME PRODUCTS PROVIDED BY BDF 12
2.4. USE OF DATA MINING AND MACHINE LEARNING IN FINANCIAL INSTITUTIONS 15
2.4.1. Introduction 15

2.4.2. Data mining....17
2.4.3. Machine learning....20
2.5. PREDICTIVE ANALYTICS....26
2.5.1. Introduction....26
2.5.2. Types of Predictive analytics....26
CHAPTER 3: METHODOLOGY....28
3.1. Introduction....28
3. 2. Research design....28
3. 3. Research approaches....28
3. 4. The population of the study exists....28
3.5. Data collection....29
3.6. DATA PROCESSING....30
3.6.1. Data Cleaning....30
3.6.2. Data Integration....31
3.6.3. Data Selection....31
3.6.4. Data Transformation....31
3.6.5. Data visualization....31
3.6.6. Modelling and Predicting....31
CHAPTER 4: DATA VISUALISATION, ANALYSIS AND INTERPRETATION....37
4.1. DATA VISUALISATION....37
4.2. DESCRIPTIVE ANALYSIS....50
a) Imbalance Data....50
b) Balance Data....51
c) Feature Selection Using Filter Method....53
CHAPTER 5. RESULTS, DISCUSSIONS AND CONCLUSION....55
5.1. RESULTS....55
a) Comparing Machine learning performance on imbalance test data....55
b) Comparing Machine learning performance on balanced test data....56
c) Receiver Operating Characteristic Curve (ROC)....57
5.2. DISCUSSION....58
Confusion Matrices for Machine learning models....59
5.3. CONCLUSION....64
REFERENCES....65

LIST OF FIGURES

Figure 1: Data mining tasks 17

Figure 2: Data mining process 18

Figure 3: Different disciplines of knowledge and the discipline of machine learning. 21

Figure 4: Machine learning techniques 22

Figure 5: Process of data mining 30

Figure 6: Class distribution 38

Figure 7: Distribution by ages 38

Figure 8: Counting using loan status 39

Figure 9: Loan status to gender 39

Figure 10: Principal and age with gender 40

Figure 11: Principal and age with sectors 40

Figure 12: Applicants who pay before due date 42

Figure 13: Principal and loan status 43

Figure 14: Gender and loan status with sectors 44

Figure 15: counting sectors with gender 45

Figure 16: age and loan status 46

Figure 17: Variation of age 47

Figure 18: presentation of sectors with loan status 48

Figure 19: presentation of whole data set 49

Figure 20: Loan eligibility counting on imbalanced data 50

Figure 21: Loan eligibility percentage 51

Figure 22: Loan eligibility counting on balanced data 52

Figure 23: Feature Selection Using Filter Method 53
Figure 24: ROC for five machine learning models 57
Figure 25: Logistic regression 59
Figure 26: Random forest 60
Figure 27: SVM 60
Figure 28: KNN 61
Figure 29: Decision Tree 61
Figure 30: Comparison of models 62

LIST OF TABLES

Table 1: Correlation between Variables and loan status (eligibility) of SMEs in Rwanda 54

Table 2: Machine learning models' comparison on imbalanced data..... 55

Table 3: Machine learning models' comparison on balanced data..... 56

ABBREVIATIONS

- AI: Artificial Intelligence
- AUC: Area under the ROC Curve
- BDCs: Business Development Centres
- BDF: Business Development Fund
- BRD: Development Bank of Rwanda
- CPCs: Community Processing Centres
- FI: Financial Institution
- FN: False Negative
- FP: False Positive
- FPR: False Positive Rate
- ICPC**:** Integrated Craft Production Centres
- KDD: Knowledge Discovery in Databases
- KNN: *K*-nearest neighbors
- MFIs: Micro Finance Institutions
- MINICOM: Ministry of Trade and industry
- NEP: National Employment Program
- OECD: Organisation for Economic Co-operation and Development
- RDDP: Rwanda Dairy Development Project
- ROC: Receiver Operating Characteristic
- RWF: Rwandan francs
- SACCO: Saving and credit cooperative
- SME: Small and Medium Enterprises
- SVM: Support Vector Machines
- TP: True Positive
- TPR: True Positive Rate

CHAPTER 1: GENERAL INTRODUCTION

1.1. Introduction

The era of Big Data has accelerated the use of data mining (Galit, S. et al., 2018). The amount of data in the world, in our lives, seems to go on and on increasing and there's no end in sight (Ian H. W. & Frank, E., 2005). The analysis of data has formed a cornerstone of scientific discovery in many domains (Kamath, C., 2009). Research in a big data era is called data science, which is a profession and a research agenda. The goal of Data Science research is to build systems and algorithms to extract knowledge, find patterns, generate insights and predictions from diverse data for various applications and visualization.

Galit, S. et al (2018), said that a common task in data mining is to examine data where the classification is unknown or will occur in the future, with the goal of predicting what that classification is or will be. Kamath, C., 2009, also said, the original or "raw" data which are provided for data mining often need extensive processing before they can be input to a pattern recognition algorithm.

BDF needs to use big data and data science approaches, such as machine learning and deep learning models, which have a significant role in loan modeling. Data Scientist roles and responsibilities include identifying business trends and changes through advanced big data analytics and using a variety of techniques to interpret results from multiple data sources through statistical analysis, data aggregation, and data mining.

As part of the financial infrastructure to promote SMEs, BDF was established in 2011 as a wholly owned subsidiary of the Development Bank of Rwanda (BRD), with the objective of assisting SMEs to access finance, particularly those without sufficient collateral to obtain credit from traditional financial institutions at reasonable rates. BDF's role was to promote alternative financing avenues at reasonable costs to help small businesses access credit by providing credit guarantees, quasi-equity support to start-up, managing matching grants, SACCO Refinancing, and business development advisory services.

SMEs play an important role in creating employment and wellbeing. Having a strong, vibrant, competitive and resilient base of SMEs is key to enhancing wealth creation and social well-being, but to fully achieve that, SME need to have sufficient and easy access to finance. Yet, in

most developing countries, the majority of SMEs are unable to acquire the financing they need to reach their potential. Governments, particularly financial regulators in Rwanda, have established various measures, programs and schemes aimed at providing SMEs with better access to financing. For the case of Rwanda, BDF was set up to increase access to finance for SMEs. The purpose of this research is to apply data mining techniques, make predictions of future performance according to the past data collected and then assess the role of data mining in BDF. Data science is a method for transforming business data into assets that help organizations improve revenue, reduce costs, seize business opportunities, improve customer experience, and more. The principal purpose of Data Science is to find patterns within data. It uses various statistical techniques to analyse and draw insights from the data. From data extraction, wrangling and pre-processing, a data scientist must thoroughly scrutinize the data. Then, he has the responsibility of making predictions from the data. The goal of a Data Scientist is to derive conclusions from the data. Through these conclusions, he is able to assist companies in making smarter business decisions. (DataFlair, 2018).

It is important to apply data science in BDF, in order to increase the ability to analyse, visualise the data and use that data to predict and/or understand the future. My research will be referred to analysing data, building predictive models for loan facilities given to SMEs from BDF between 2016 and 2019, the dataset is collected and extracted from BDF database.

1.2. Research problem

Despite improvements in the past decade, a considerable section of Rwandan SMEs are faced with several challenges in accessing financing (MINICOM, 2016). A very recent study by Harelimana (2017) showed that almost a half of the firms (38.8%) operate their enterprises using both external and internal sources of finance and only 25.5% operates with external sources while the remaining 35.5% operates with internal sources, implying that some of the SMEs rely on internal revenue sources and most probably due to a myriad of hurdles in access. In addition, most SMEs are faced with difficulties in consolidating capital and creating business plans to qualify for lending from commercial banks and microfinance institutions, they lack skills and capacity and often lack the ability and resources to gather and process market information outside of what is immediately relevant to their current business due to lack of technical

knowledge and training on how to make use of this information (Access to Finance Rwanda, 2012). According to the Ministry of Trade and Industry (2016) SMEs in Rwanda lack an understanding of the local, regional and international market in which they operate, limiting their ability to take advantage of potential market opportunities. They do not have the resources or time to spend gathering and understanding market information that would be useful to their operations.

The fact that the government set up the BDF, whose main objective was to increase and make access to finance easy for small and medium enterprises.

According to that great role in development of the country, BDF must apply data mining to clearly discover the hidden knowledge inside their dataset and improve the working ways and predict for the future outcome and making better decisions using advanced technology and artificial intelligence techniques.

Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask:

- Which products should be promoted to particular customers?
- What is the likelihood that a certain customer will default or pay back a loan?
- Which products are mostly needed by many customers?
- How to identify potential and faithful customers of BDF?
- Which customers are fitting to the products of BDF?

These types of questions can be answered quickly and easily if the information hidden among the huge amount of data in the databases can be located and utilized.

BDF must use data mining in various application areas like marketing, risk management, money laundering detection and potential investment areas. The patterns detected help the BDF to forecast future events that can help in its decision-making processes. More and more financial institutions are investing in data mining technologies to be more competitive and complete their tasks in efficient and effective way with high performance, improve the quality of services and improve profitability

The amount of data collected by financial institution tools has grown rapidly in recent years. Existing statistical data analysis techniques find it difficult to manage with the large volumes of data now available. This explosive growth has led to the need for new data analysis techniques and tools in order to find the information hidden in this data. Financing is an area where vast

amounts of data are collected. This data can be generated from daily transactions, loan applications, loan repayments, etc. It is assumed that valuable information on the financial profile of customers is hidden within these massive operational databases and this information can be used to improve the performance of BDF. Implementing models that automatically teach themselves how to optimize their parameters from available data, that models will help optimize business decisions, increase the value of each customer and communication, and improve customer satisfaction.

1.3. General objectives

The general objective of this study is to apply data mining techniques to discover hidden knowledge from BDF dataset.

1.4. Specific objectives

- To analyse the loan data of BDF and discover hidden knowledge and visualise those outputs.
- Create a model for predicting the future performance.
- To facilitate decision making using output from data mining techniques.

1.5. Research Questions

- Are there hidden knowledge to be discovered in a BDF's loan dataset?
- How BDF loan data can be used to predict for future performance?
- How Predictive model can facilitate BDF management in decision making?

1.6. Significance of the study

This study is significant to four parties of interest such as Personal interest, Organizational interest, stakeholder's interest and Academic interest.

1.6.1. Personal interest

This study will increase my ability to make research while applying different methods and techniques to conduct any academic scientific research especially in this field.

1.6.2. Academic interest

This study will be used by future researchers in this research domain as a reference.

1.6.3. Organizational interest

BDF as a government organization will benefit from the results of this research and they will be advised on what they have to improve in decision making and policy setting.

1.6.4. Stakeholders 'interest

This study results will inform stakeholders of BDF about how the situation is, they will get an insight of all activities and how to deal with the problem through different measures set and prediction of future.

1.7. Justification

This study is expected to increase my understanding in using data mining techniques in problem solving as in this research is to mine data and modelling using algorithms, it can be used by decision makers to control or carry out the problem of loan provided by BDF to Small and medium enterprises based on correlation coefficient of each variable.

1.8. Scope of study

This study was delimited in terms of content, geographical and time.

1. 8. 1. Content scope

This study is limited on mining, analyzing and predicting for the future using indicators found in a dataset like principal, age, years, gender, past due days, education and terms from 2016 to 2019.

1. 8. 2. Time scope

To conduct significant and meaningful research, I need enough time to work on data and read literature, As a result, Researcher will be able to respond to the research questions after creating a useful model; this will take about three months.

1. 8. 3. Geographical scope

The research will be conducted on BDF in Rwanda, especially at head office located in Kigali city. It means that the study will cover the role of data mining in activities of BDF and application of data science in BDF's financial operations.

CHAPTER 2: LITERATURE REVIEW

2.1. INTRODUCTION

Through various approaches, researchers conducted studies of financial fields of research, especially on loan facilitation to SMEs by evaluating its indicators, influencers and many more. This section reviews some research works on use of data mining techniques in these financial operations. The following literature is related to this research, and some gaps will be addressed in this research.

In this review, we are going to discuss important topics that will help us to get the real output that will help organizations to make the data driven solutions that are data mining techniques, machine learning algorithms and predictive analysis then we shall discuss about its application in financial institutions.

Analysis of big data is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of big data using machine learning algorithms involve multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modelling and analysis and interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining. By using data mining to analyse patterns and trends, financial institution executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable.

Most of entrepreneurs started their businesses because they were unable to find a waged employment. This demonstrates that entrepreneurship is a new concept for Rwandans. After graduation, the first choice is seeking for a waged employment. Even some trainees who did not attend school declared to have started their own business because they could not find salaried employment. The BDF is seen as a critical stakeholder in attaining NEP's targets because the government seeks to create the majority of jobs through facilitating SME development and BDF's

main focus is in helping entrepreneurs to find initial operational capital. In order to have a national presence, BDF recently decentralized its structures and created centres in each of Rwanda's 30 districts after it took over what were previously known as Business Development Centres (BDCs), renovated and renamed them as 'Kora Wigire'. As main conclusion, BDF participates in creating jobs for youth of Rwanda as business evaluation and follow up, refinancing grant, coordination of government grant fund, stimulating SMEs growth and advisory and access to financial service etc.

Specifically, data science is enabling these companies to leverage data mining and predictive modeling to personalize offers, reduce risk, create disruptive new products, expand markets, minimize operating expenses, automate traditionally manual processes, and much more. These would be very beneficial business enhancements for traditional banks and insurance companies too, and some are already using data science to achieve them. Let's look at some concrete examples of this, and then we will detail how organizations can best gain competitive competence in data science (Dataiku, 2016).

2.2. KEYWORDS

Here are some key terms definitions:

1. ***SMEs (Small and medium enterprises)***

A small scale entity that is independently managed by an owner manager in both its finance and operation and is also characterized by its small number of staff, limited financial resources and assets. (Bennett, 2007).

2. ***Loan***

It is money, property, or other material goods given to another party in exchange for future repayment of the loan value or principal amount, along with interest or finance charges (Kagan, 2019).

3. ***A financial institution (FI)***

A company engaged in the business of dealing with financial and monetary transactions such as deposits, loans, investments, and currency exchange. Financial institutions encompass a broad

range of business operations within the financial services sector including banks, trust companies, insurance companies, brokerage firms, and investment dealers (Martey, 2018).

4. ***Data mining***

A process of using a computer to examine large amounts of information, for example about customers, in order to discover things that are not easily seen or noticed (Jaseen, 2011).

According to Max Bramer (2016)**, Data mining** is the study of collecting, cleaning, processing, analysing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, "data mining" is a broad umbrella term that is used to describe these different aspects of data processing.

5. ***Artificial intelligence (AI)***

It refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving (Frankenfield, 2010)**.**

6. ***Machine learning***

An application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves (Skiena, 2017).

7. ***Big Data***

It is a term used to identify the datasets whose size is beyond the ability of typical database software tools to store, manage and analyse. Big Data introduces unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors. (Jaseena & Julie, 2016).
Data science has become conflated in the public eye with big data, the analysis of massive data sets resulting from computer logs and sensor devices. In principle, having more data is always

better than having less, because you can always throw some of it away by sampling to get a smaller set if necessary.

The challenges of big data include:

- The analysis cycle time slows as data size grows
- Large data sets are complex to visualize
- Simple models do not require massive data to fit or evaluate

8. *Data Science*

Extraction of relevant insights or knowledge from data, it uses scientific methods, processes, algorithms and other systems, and it uses various techniques from many fields like mathematics, machine learning, computer programing and statistical modeling. (Bramer, 2016)

Data science is commonly defined as a methodology by which actionable insights can be inferred from data. Performing data science is a task with an ambitious objective: the production of beliefs informed by data and to be used as the basis of decision-making.
Data science lies at the intersection of computer science, statistics, and substantive application domains. From computer science comes machine learning and high-performance computing technologies for dealing with scale. From statistics comes a long tradition of exploratory data analysis, significance testing, and visualization. From application domains in business and the sciences comes challenges worthy of battle, and evaluation standards to assess when they have been adequately conquered (Skiena, 2017).

Why data science? According to *Steven S. Skiena*, there is three reasons:

New technology makes it possible to capture, annotate, and store vast amounts of social media, logging, and sensor data. After you have amassed all this data, you begin to wonder what you can do with it.

Computing advances make it possible to analyse data in novel ways and at ever increasing scales. Cloud computing architectures give even access to vast power when they need it. New

approaches to machine learning have led to amazing advances in longstanding problems, like computer vision and natural language processing.

Prominent technology companies (like Google and Facebook) and quantitative hedge funds (like Renaissance Technologies and Two Sigma) have proven the power of modern data analytics. In general, data science allows us to adopt four different strategies to explore the world using data:

1. *Probing reality*. Data can be gathered by passive or by active methods. In the latter case, data represents the response of the world to our actions. Analysis of those responses can be extremely valuable when it comes to taking decisions about our subsequent actions (Laura & Santi, 2017).

2. Pattern discovery. Divide and conquer is an old heuristic used to solve complex problems; but it is not always easy to decide how to apply this common sense to problems. Datified problems can be analysed automatically to discover useful patterns and natural clusters that can greatly simplify their solutions. The use of this technique to profile users is a critical ingredient today in such important fields as programmatic advertising or digital marketing (Laura & Santi, 2017).

3. *Predicting future events*. Since the early days of statistics, one of the most important scientific questions has been how to build robust data models that are capable of predicting future data samples. Predictive analytics allows decisions to be taken in response to future events, not only reactively. Of course, it is not possible to predict the future in any environment and there will always be unpredictable events; but the identification of predictable events represents valuable knowledge (Laura & Santi, 2017).

2.3. INTRODUCTION OF BDF

2.3.1. Background of the Organization

As part of the financial infrastructure to promote SMEs, BDF was established in 2011 as a wholly owned subsidiary of the Development Bank of Rwanda (BRD), with the objective of assisting SMEs to access finance, particularly those without sufficient collateral to obtain credit from traditional financial institutions at reasonable rates. BDF's role was to promote alternative financing avenues at reasonable costs to help small businesses access credit by providing credit guarantees, Quasi-Equity support to start-up, managing matching grants, SACCO Refinancing, and business development advisory services.

The Government also consolidated the different funds provisioned for SME financial support that had been spread across various ministries and agencies under BDF. These included the SME Guarantee Fund, the Agricultural Guarantee Fund, the Rural Investment Facility, the Women's Guarantee Fund and the Retrenched Civil Servants Guarantee Fund. BDF has since harmonized the management of these funds and delivered through comprehensive agreements with the financing institutions.

Given the privatization of BRD, a new ownership structure and strategic plan has been formalized for BDF to deliver on its mandate of supporting SME development. Moreover, with the introduction of the National Employment Program (NEP), BDF has been designated as the key implementing agency for NEP Pillar 2, and its functions need to be aligned with this role.

2.3.2. SOME PRODUCTS PROVIDED BY BDF

2.3.2.1. Guarantee Fund

BDF works with the financial institutions (Banks, MFIs and Saccos) to cover between 50 and 75% of collateral required by the lending institution. The maximum guaranteed amount is 500 million francs for Agriculture campaign and 300 million francs for other sectors within a maturity period of 10 years. It consists: Fixed Assets and Working Capital Loans

This guarantee fund is provided to all Medium Small Enterprises, individuals, associations, cooperatives and Companies.

2.3.2.2. Grants

These are funds meant for easing loan repayment. Currently, there is only one type of grant provided through financial institutions. It consists:

Post-Harvest: is a program that is intended to promote investments in growing beans, cassava, Irish potatoes, and maize.

- Rwanda Dairy Development Project (RDDP): is a program that is intended to promote investments in milk value chain products in the different districts of Rwanda.

2.3.2.3. Sacco Refinancing

This fund was designed by SACCO's for the purpose of increasing SACCOs' capacity in lending to as many Rwandan as possible. It is open to SACCOs from all districts in the country. Women and Youth are encouraged to use this facility.

2.3.2.4 Quasi-Equity

Quasi-Equity is a product designed for small & medium enterprises (both start-ups and those already operating) given through a mixture of debt and equity. The Community Processing Centers (CPCs) are also beneficiaries of Quasi Equity.

The Quasi-Equity is a mix of debts and equity whereby the debt portion pays a subsidized interest rate and Equity portion charges are based on the annual percentage sales that will be paid every year to BDF to ensure self-liquidation of the equity portion up to a certain agreed period. **NB:** Loan amount that BDF provides through Quasi Equity is Rwf 15 Million and above depending on the funding available.

2.3.2.5. BDF Agribusiness Financing

Under this scheme, which is part of Quasi Equity, BDF supports agribusiness projects in combined production and agro-processing through a quasi-equity scheme where BDF co-invests with the project promoter (owner). The package provided to viable agribusiness projects is made up of the promoter's contribution of 10% of the amount requested from BDF, a grant of 30% and BDF convertible shares of 60%. The ceiling of the loan amount under this scheme is 10,000,000Rwf given at an interest of 12%.

2.3.2.6. ICPC Equipment Leasing Facility

This is a leasing facility offered to Integrated Craft Production Centers **(ICPCs)**, otherwise known as Agakiriro, ICPCs in every District established to provide hands on skills training and an environment for technical start-ups.

2.4. USE OF DATA MINING AND MACHINE LEARNING IN FINANCIAL INSTITUTIONS

2.4.1. Introduction

The use of Artificial Intelligence and machine learning in financial services may bring key benefits for financial stability in the form of efficiencies in the provision of financial services and regulatory and systemic risk surveillance. The more efficient processing of information on credit risks and lower-cost customer interaction may contribute to a more efficient financial system. The internal (back-office) applications of AI and machine learning could improve risk management, fraud detection, and compliance with regulatory requirements, potentially at lower cost. In portfolio management, the more efficient processing of information from AI and machine learning applications could help to boost the efficiency and resilience of financial. With use cases by regulators and supervisors, there is potential to increase supervisory effectiveness and perform better systemic risk analysis in financial markets (Bhambri, 2011)

Information is collected almost everywhere in our everyday lives. This leads to the huge increase in the amount of data available. Physical analysis of these huge amounts of information stored in modern databases is very difficult. Data mining provides tools to reveal unknown information in large databases which are stored already. A well-known data mining technique is association rule mining. Association rule mining and classification techniques to find the related information in large databases is becoming very important in the current scenario. Association rules are very efficient in revealing all the interesting relationships in a relatively large database with a huge amount of data. The large quantity of information collected through the set of association rules can be used not only for illustrating the relationships in the database, but also used for differentiating between different kinds of classes in a database.

The financial institution across the world has undergone tremendous changes in the way the business is conducted. With the recent implementation, greater acceptance and usage of electronic transactions, the capturing of transactional data has become easier and, simultaneously, the volume of such data has grown considerably. It is beyond human capability to analyze this huge amount of raw data and to effectively transform the data into useful knowledge for the organization. Data Mining can help by contributing in solving business

problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases (Kazi, 2012).

The financial industry is widely recognizing the importance of the information it has about its customers. Undoubtedly, it has among the richest and largest pool of customer information, covering customer demographics, transactional data, credit cards usage pattern, and so on. As financing is in the service industry, the task of maintaining a strong and effective customer relationship management is a critical issue. To do this, financial institutions need to invest their resources to better understand their existing and prospective customers. By using suitable data mining tools, financial institutions can subsequently offer tailor-made products and services to those customers. There are numerous areas in which data mining can be used in the financial field, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments (Ahmed, 2012).

- **Machine learning and data mining** often employ the same methods and overlap significantly. They can be roughly distinguished as follows:
    - Machine learning focuses on prediction, based on known properties learned from the training data.
    - Data mining focuses on the discovery of unknown properties in the data.

  This is the analysis step of Knowledge Discovery in Databases.

  The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind. On the other hand, machine learning also employs data mining methods as “unsupervised learning” or as a pre-processing step to improve learner accuracy. Much of the confusion between these two research communities comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in Knowledge Discovery and Data Mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data(Le Roux, Nicolas, Bengio, Yoshua, Fitzgibbon & Andrew; 2012).

2.4.2. Data mining

It is also called knowledge discovery in databases, in computer science, is the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyse large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists) (Bramer, 2016).

DATA MINING TASK

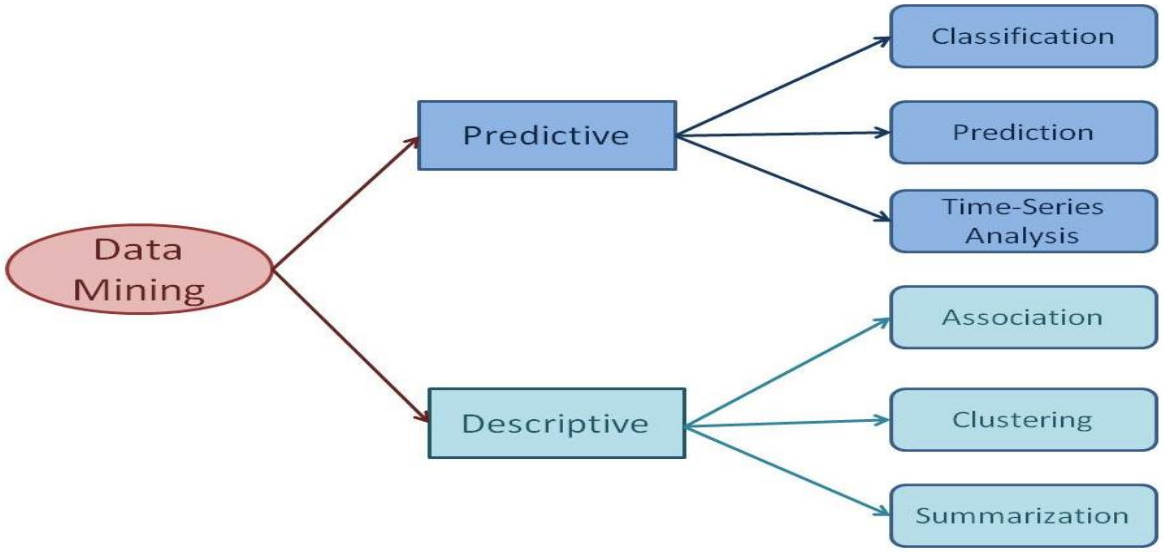

Figure 1: Data mining tasks from https://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks retrieved on 15/05/2020.

The Data Mining Process

The data mining process is a pipeline containing many phases such as data cleaning, feature extraction, and algorithmic design. Knowledge Discovery has been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which data mining forms just one part, albeit a central one. (Bramer, 2016).

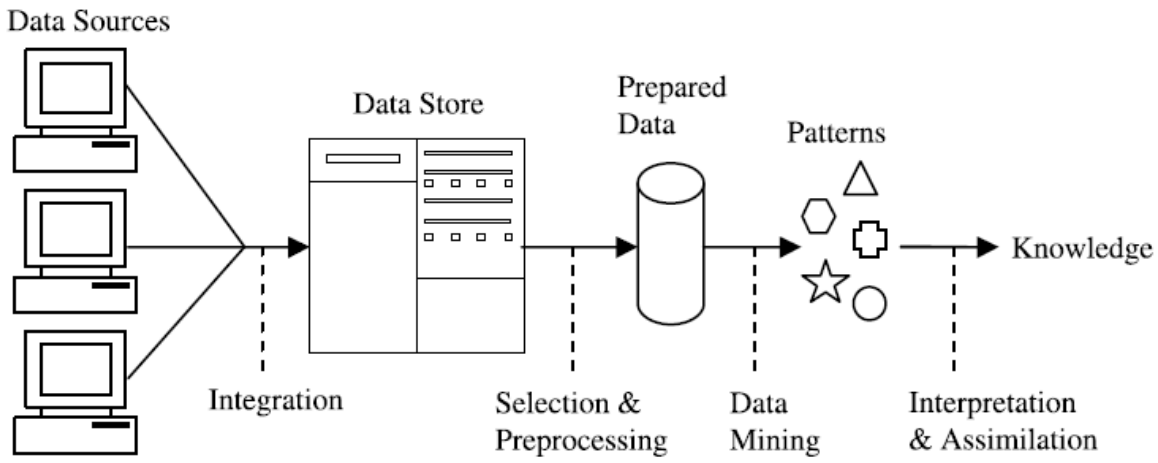

Figure 2: Data mining process (Bramer, 2016).

Following are some examples of how the financial industry has been effectively utilizing data mining in these areas.

a) Marketing

One of the most widely used areas of data mining for the financial industry is marketing. The institution's marketing department can use data mining to analyse customer databases. Data mining carries various analyses on collected data to determine the consumer behaviour with reference to product, price and distribution channel. The reaction of the customers for the existing and new products can also be known based on which banks will try to promote the product, improve quality of products and service and gain competitive advantage. Bank analysts can also analyse the past trends, determine the present demand and forecast the customer behaviour of various products and services in order to grab more business opportunities and anticipate behaviour patterns. Data mining technique also helps to identify profitable customers from non-profitable ones (Desai, 2004). The data mining techniques can be used to determine how customers will react to adjustments in interest rates, the risk profile of a customer segment for defaulting on loans (Kaptan, 2013).

b) Risk Management

Data mining is widely used for risk management in the financial industry. Financial institution executives need to know whether the customers they are dealing with are reliable or not. Offering new customers credits, extending existing customers lines of credit, and approving

loans can be risky decisions for banks if they do not know anything about their customers (Madan 2006).

Financial institutions provide loans to its customers by verifying the various details relating to the loan such as amount of loan, lending rate, repayment period, type of property mortgaged, demography, and income and credit history of the borrower. Customers with banks for longer periods, with high income groups are likely to get loans very easily. Even though they are cautious while providing loans, there are chances for loan defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't (Desai, 2004).

Financial institution executives by using Data mining techniques can also analyse the behaviour and reliability of the customers. It also helps to analyse whether the customer will make prompt or delay payment if the credit cards are sold to them (Desai, 2004). Credit scoring, in fact, was one of the earliest financial risk management tools developed. Credit scoring can be valuable to lenders in the financial industry when making lending decisions. Data mining can also derive the credit behaviour of individual borrowers with instalment, mortgage and credit card loans, using characteristics such as credit history, length of employment and length of residency. A score is thus produced that allows a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a high risk of default. By knowing what the chances of default are for a customer, the financial institution is in a better position to reduce the risks (Madan, 2006).

c) Fraud Detection

Another popular area where data mining can be used in the financial industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a financial institution taps the data warehouse of a third party and uses data mining programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the internal information of financial institutions. Most of the financial institutions are using a “hybrid” approach (Bhasin, 2006).

d) Customer Relationship Management

In the era of cut throat competition the customer is considered as the king. Data mining can be useful in all the three phases of a customer relationship cycle: Customer Acquisition, Increasing value of the customer and Customer retention (Rajanish, 2007). Customer acquisition and retention are very important concerns for any industry, especially the banking industry (Madan 2006). Today customers have a wide range of products and services provided by different institutions. Hence, banks have to cater the needs of the customer by providing such products and services which they prefer. This will result in customer loyalty and customer retention. Data mining techniques help to analyse the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known which will help the banks to perform better and retain its customers (Desai, 2004).

2.4.3. Machine learning

It is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

Machine learning explores the construction and study of algorithms that can learn from and make predictions on data (Kohavi, 2017). It is also a powerful tool used to mine data and discover knowledge with high ability to predict without errors.

Machine learning is a branch of AI that aims at enabling machines to perform their jobs skilfully by using intelligent software. The statistical learning methods constitute the backbone of intelligent software that is used to develop machine intelligence. Because machine learning algorithms require data to learn, the discipline must have connection with the discipline of database. Similarly, there are familiar terms such as Knowledge Discovery from Data (KDD), data mining, and pattern recognition (Kumar, 2018).

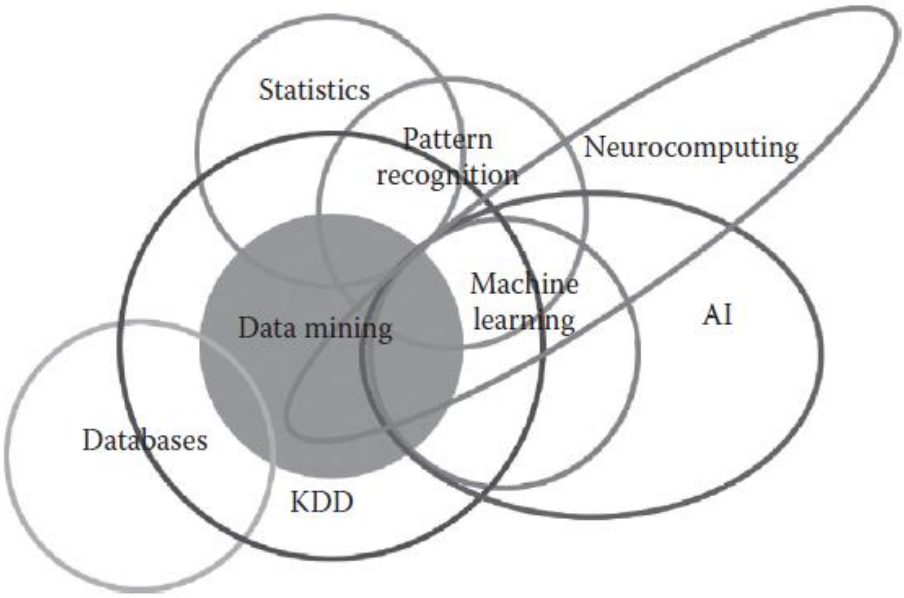

Figure 3: Different disciplines of knowledge and the discipline of machine learning. (From Guthrie, Looking backwards, looking forwards: SAS, data mining and machine learning, 2014, http://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-lookingforwards-sas-data-mining-and-machine-learning/2014. With permission.)

2.4.3.1. Types of problems and tasks of machine learning

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system.

These are: (Russell, 2017)

- **Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
- **Reinforcement learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent (Russell, 2017),

Another categorization of machine learning tasks arises when one considers the desired *output* of a machine learned system (Russell, 2017).

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one (or multi-label classification) or more of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
- Density estimation finds the distribution of inputs in some space.
- Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modelling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar (Russell, 2017).

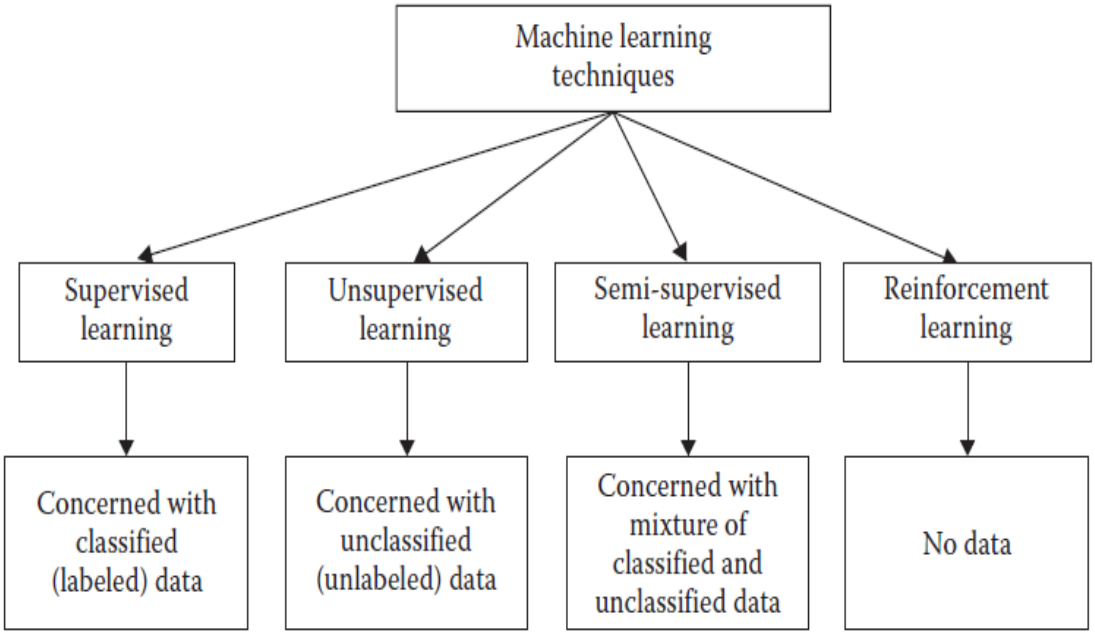

Figure 4: Machine learning techniques (Russell, 2017).

2.4.3.2. Machine learning techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market.

In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques emulate human cognition and learn from training examples to predict future events.
A brief discussion of some of these methods used commonly for predictive analytics is provided below. A detailed study of machine learning can be found in Mitchell (2016).

a) Neural networks

Neural networks are nonlinear sophisticated modeling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control in a wide spectrum of fields such as finance, cognitive psychology/neuroscience, medicine, engineering, and physics.
Neural networks are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they learn the relationship between inputs and output through training. There are three types of training in neural networks (deep learning) used by different networks, supervised and unsupervised training, reinforcement learning, with supervised being the most common one, but deep learning can be supervised, unsupervised or reinforced depends on situation (C. Aggarwal ,2018)

b) Support vector machines

Support Vector Machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are learning machines that are used to perform binary classifications and regression estimations.

They commonly use kernel based methods to apply linear classification techniques to non-linear classification problems. There are a number of types of SVM such as linear, polynomial, sigmoid etc (Mitchell 2016).

c) Naïve Bayes

Naïve Bayes based on Bayes conditional probability rule is used for performing classification tasks. Naïve Bayes assumes the predictors are statistically independent which makes it an effective classification tool that is easy to interpret (Mitchell 2016).

It is best employed when faced with the problem of 'curse of dimensionality' i.e. when the number of predictors is very high.

d) *K*-nearest neighbors

The nearest neighbor algorithm (KNN) belongs to the class of pattern recognition statistical methods. The method does not impose a priori any assumptions about the distribution from which the modeling sample is drawn. It involves a training set with both positive and negative values. A new sample is classified by calculating the distance to the nearest neighboring training case.

The sign of that point will determine the classification of the sample. In the k-nearest neighbor classifier, the k nearest points are considered and the sign of the majority is used to classify the sample (Mitchell 2016).

e) Logistic Regression

It is a Machine Learning algorithm which is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1 (Mitchell 2016).

f) Decision Trees

They are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split (Mitchell 2016).

e) Random forest classifier

It creates a set of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (Mitchell 2016).

2.5. PREDICTIVE ANALYTICS

2.5.1. Introduction

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behaviour patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs (Stephen, 2012). The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguishes it from forecasting. For example, "Predictive analytics—Technology that learns from experience (data) to predict the future behaviour of individuals in order to drive better decisions (Jordan, 2014).

2.5.2. Types of Predictive analytics

Generally, the term predictive analytics is used to mean predictive modelling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modelling and decision modelling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

2.5.2.1. Predictive models

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they

seek out subtle data patterns to answer questions about customer performance, or fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modelling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios (Barkin, 2011).

2.5.2.2. Descriptive models

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behaviour (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modelling tools can be utilized to develop further models that can simulate a large number of individualized agents and make predictions (Barkin, 2011).

2.5.2.3. Decision models

Decision models describe the relationship between all the elements of a decision — the known data (including results of predictive models), the decision, and the forecast results of the decision in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance (Barkin, 2011).

Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years: Cross-sell, Customer retention, direct marketing, Fraud detection, Portfolio, product or economy-level prediction, Risk management, Underwriting etc.

CHAPTER 3: METHODOLOGY

3.1. Introduction

Methodology is explaining the way expected results will be achieved; this means, from basic level to the results by achieving research objectives and solving its related research questions.

In this section, Researcher has shown in detail, tools (software, dataset and algorithm and accuracy measurements) to be used while carrying out this research; Researcher explained the way of gathering data and techniques to be used in dataset formulation, considering the duration and environment of research which all lead to the fulfillment of research objectives.

Those details are provided here below:

3. 2. Research design

Research design is the plan for a study that is used as a guide in collecting and analysing the data. Different studies, field surveys are best approaches for consultation and seek tangible relationships in leaning. In this study Researcher carried out analytical research where he need data to be studied and visualized to obtain trends such as relationship between variables and time series related influences. Dataset for this study consist of data extracted from BDF database, related to loan facilitations provided to SMEs in Rwanda.

3. 3. Research approaches

This research is an analytical study, where evidence to be used have been collected from existing databases, which means the research approach to be followed is quantitative approach, such that it deals with numerical variables; then data will be trained using data mining tools, analysed and presented for results and knowledge extraction.

3. 4. The population of the study exists

This study, as stated before in the geographical and content scope, has focused on data analysis in finding the effect of data mining in BDF operations for data between 2016 and 2019.

BDF is located in Kigali city with numerous workers, but researcher will focus on data stored in data centres mainly those related to financial transactions to get full information that will help me to apply data science and machine learning algorithms in BDF.

3.5. Data collection

It is the process of gathering and measuring information on variables of interest. Data is collected from multiple data sources available in BDF.

Data preparation, known also as data pre-processing, is focused mainly on two issues: firstly, the data must be organized into a proper form for data mining algorithms, and, secondly, the data sets used must lead to the best performance and quality for the models obtained by data mining operations.

As any other statistical and data-based study, data will be collected and used to define results of a statistical study; survey's data can be classified as numerical (quantitative) or non-numerical (qualitative) responses. Depending on the manner they are collected, it distinguishes primary data from secondary data.

3.6. DATA PROCESSING

It is indicated by wideskills.com on the below figure.

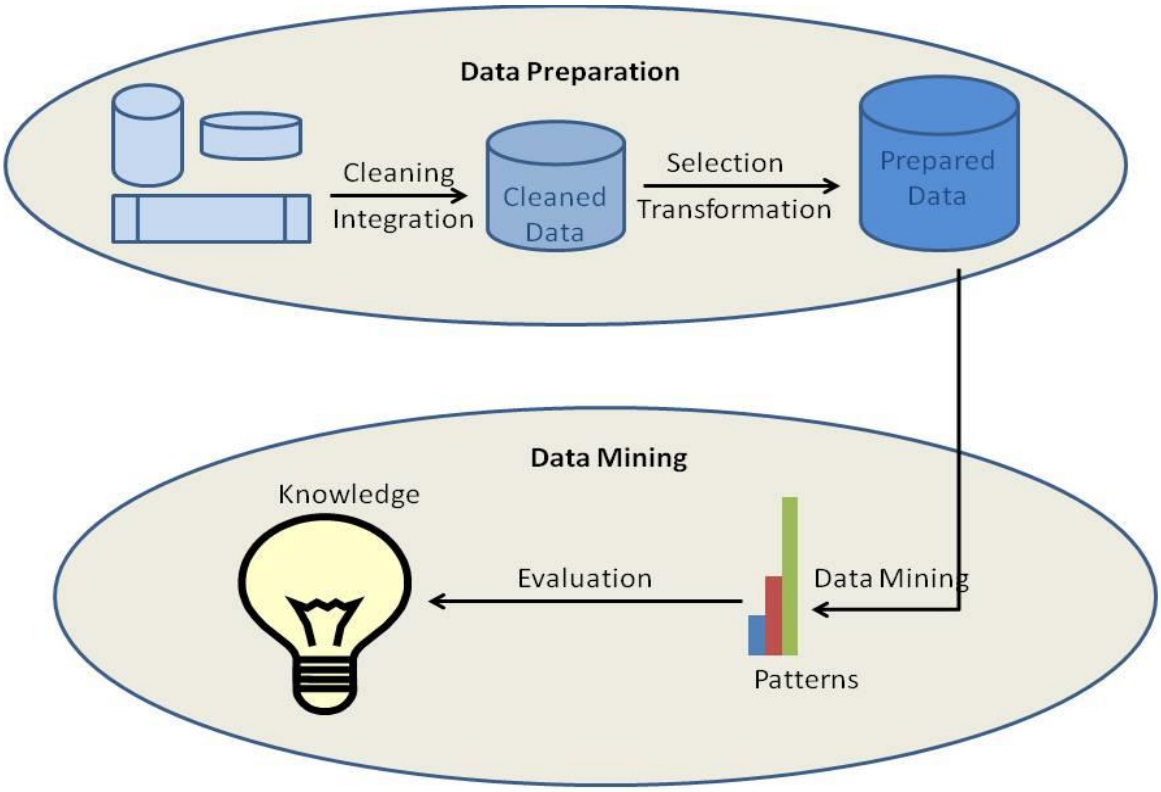

Figure 5: Process of data mining from https://www.wideskills.com/data-mining-tutorial/data-mining-processes retrieved on 20/4/2020.

3.6.1. Data Cleaning

Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent. The data available in data sources might be lacking attribute values, data of interest etc. If the data is not clean, the data mining results would be neither reliable nor accurate.

Data cleaning involves a number of techniques including filling in the missing values manually, combined computer and human inspection, etc. The output of the data cleaning process is adequately cleaned data.

3.6.2. Data Integration

Data integration is the process where data from different data sources are integrated into one. Data lies in different formats in different locations. Data could be stored in databases, text files, spreadsheets, documents, data cubes, Internet and so on. Data integration is a really complex and tricky task because data from different sources does not match normally. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

3.6.3. Data Selection

Data mining process requires large volumes of historical data for analysis. So, usually the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database.

3.6.4. Data Transformation

Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization.

In this process, data is transformed into a form suitable for the data mining process. Data is consolidated so that the mining process is more efficient and the patterns are easier to understand. Data transformation involves data mapping and code generation process.

3.6.5. Data visualization

It is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

3.6.6. Modelling and Predicting

Data mining is the core process where a number of complex and intelligent methods are applied to extract patterns from data. Data mining process includes a number of tasks such as association, classification, prediction, clustering and time series analysis and so on.

Predictive analytics tools are powered by several different models and algorithms that can be applied to a wide range of use cases. Determining what predictive modelling techniques are best for this research is a key to getting the most out of a predictive analytics solution and leveraging data to make insightful decisions.

With machine learning predictive modelling, there are several different algorithms that can be applied for this case but we will focus on Logistic regression, Random forest, Decision tree, *K*-nearest neighbors (KNN) and Support vector machines (SVM) (Those model have been discussed in chapter 2).

With these Predictive Algorithms, they are measured by performance metrics (Accuracy, Precision, Recall & F1 Score). Once you have built your model, the most important question that arises is how good is your model? Therefore, evaluating your model is the most important task in the data science project which delineates how good your predictions are.

According to towardsdatascience.com visited on 24th April 2020, performance metrics are explained as follow:

a) Accuracy

It simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions (the number of test data points).

It is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost the same. Therefore, you have to look at other parameters to evaluate the performance of your model

Formula for accuracy

$$Accuracy = \frac{True\ Positives + True\ Negatives}{All\ samples}$$

b) Precision

The precision score quantifies the ability of a classifier to not label a negative example as positive. The precision score can be interpreted as the probability that a positive prediction made by the classifier is positive. The score is in the range [0, 1] with 0 being the worst, and 1 being perfect. It is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate.

Formula for Precision:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

c) Recall (Sensitivity)

Recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query. It is the ratio of correctly predicted positive observations to all observations in actual class.
It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

This is how Recall is calculated.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

d) F1 score

The F1-score is a single metric that combines both precision and recall via their harmonic mean. The score lies in the range [0, 1] with 1 being ideal and 0 being the worst. Unlike the arithmetic mean, the harmonic mean tends toward the smaller of the two elements. Hence the F1 score will be small if either precision or recall is small.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

The formula is as follows:

$$F1\ Score = \frac{2*(Precision*Recall)}{Precision + Recall}$$

e) Receiver Operating Characteristic (ROC Curve)

In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its prediction threshold is varied. The ROC curve provides nuanced details about the behavior of the classifier. The curve is created by

plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

A ROC curve is a two-dimensional plot that illustrates how well a classifier system works as the discrimination cut-off value is changed over the range of the predictor variable. The x axis or independent variable is the false positive rate for the predictive test. The y axis or dependent variable is the true positive rate for the predictive test. Each point in ROC space is a true positive/false positive data pair for a discrimination cut-off value of the predictive test. If the probability distributions for the true positive and false positive are both known, a ROC curve can be plotted from the cumulative distribution function. In most real applications, a data sample will yield a single point in the ROC space for each choice of discrimination cut-off. A perfect result would be the point (0, 1) indicating 0% false positives and 100% true positives. The generation of the true positive and false positive rates requires that we have a gold standard method for identifying true positive and true negative cases. (Yang & Berdine, 2017).

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

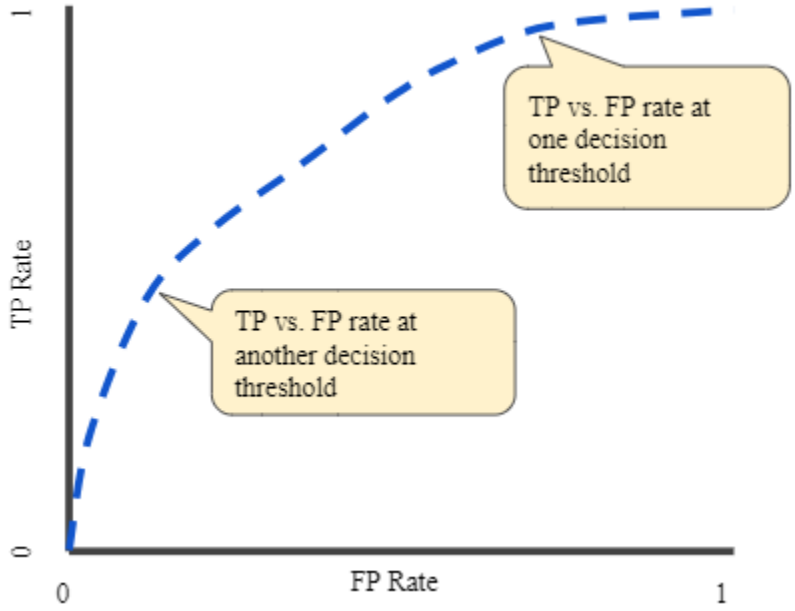

Figure 5: Receiver Operating Characteristic from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc retrieved on 23/05/2020.

CHAPTER 4: DATA VISUALISATION, ANALYSIS AND INTERPRETATION

4.1. DATA VISUALISATION

The information mined from the data must be understandably presented, thus, different knowledge representation and visualization techniques are applied to provide the output of data mining to the users.

Data visualization provides a powerful mechanism to aid the user during both data preprocessing and the actual data mining (Foong, 2001). Through the visualization of the original data, the user can browse to get a "feel" for the properties of that data. For example, large samples can be visualized and analyzed (Grinstein and Ward, 2001). Particularly visualization may be used for outlier detection, which highlights surprises in the data, i.e. data instances that do not comply with the general behavior or model of the data (Han & Kamber, 2001, Pyle, 1999). In addition, the user is supported in selecting the appropriate data through a visual interface. Data transformation is an important data preprocessing step, during data transformation, visualizing the data helps the user to ensure the correctness of the transformation. That is, the user may determine whether the two views (original versus transformed) of the data are equivalent. Visualization may also be used to assist users when integrating data sources, assisting them to see relationships within the different formats.
The following figures represent data visualization in graphs.

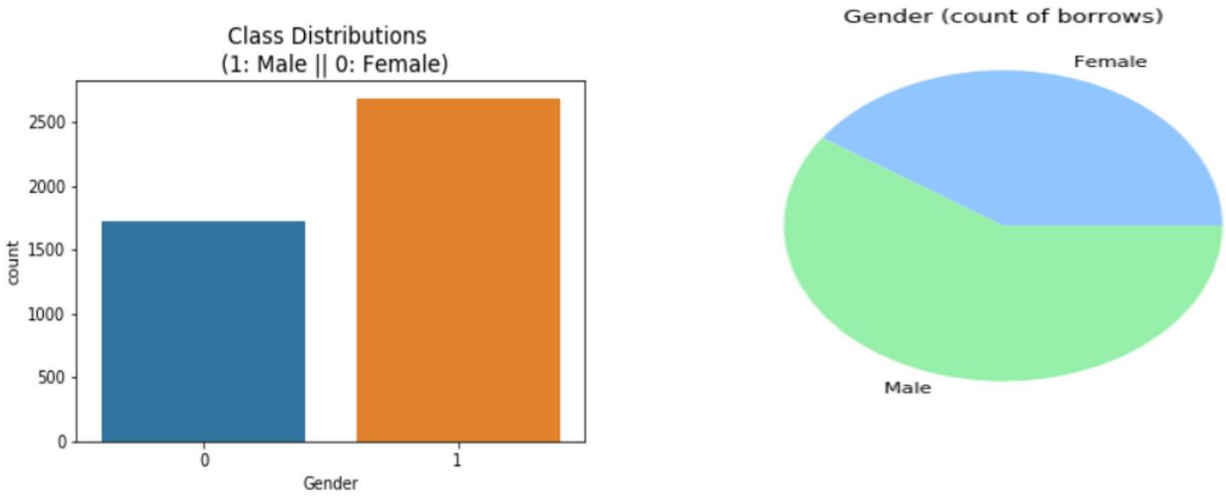

Figure 6: Class distribution

Figure 6 represents the number of applicants where male is 2691 and female is 1732 that contain 60.84% of male and 39.16% of female. The female are presented by blue colour and male are presented by orange and green colour on circle graph.

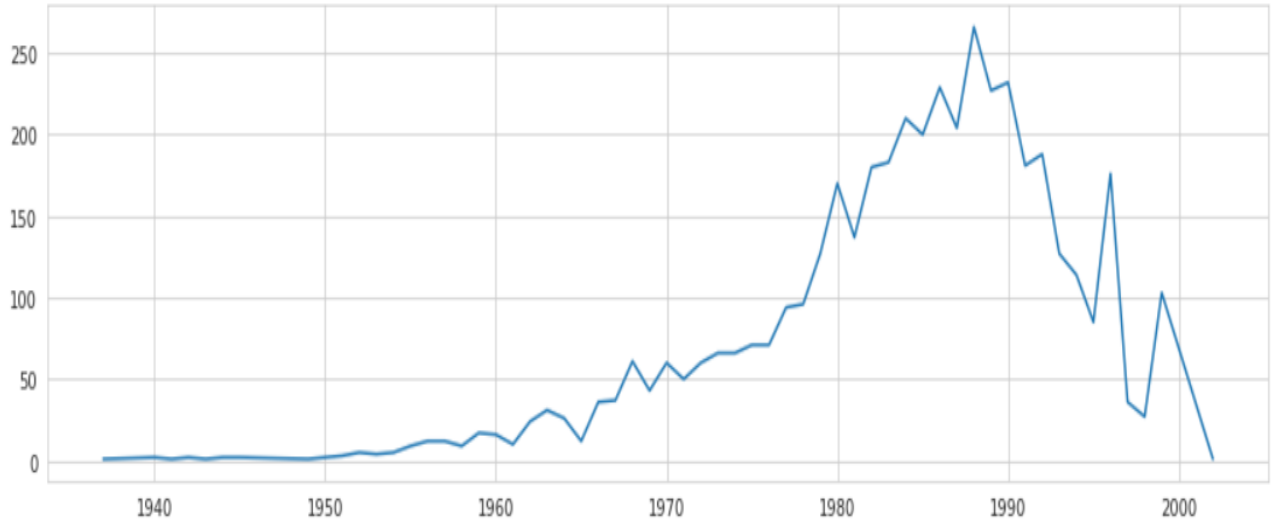

Figure 7: Distribution by ages

Figure 7 shows that the most applicants are born between 1980 and 1990. It means that young people have high attendance at BDF products.

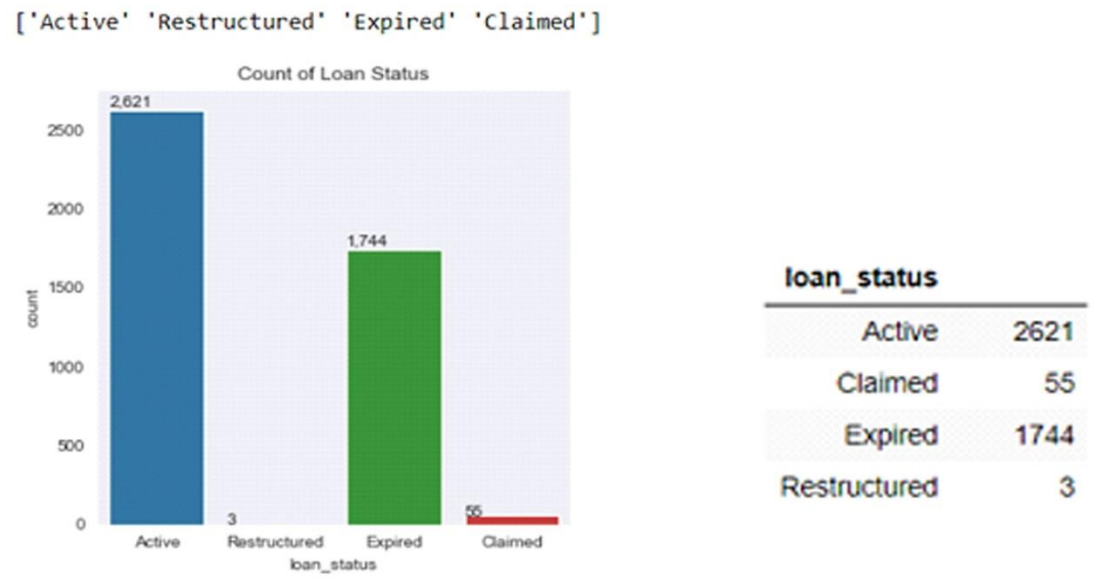

Figure 8: Counting using loan status

Figure 8 shows the number of participants by using loan status where Active are 2621, claimed are 55, expired are 1744 and restructured are 3.

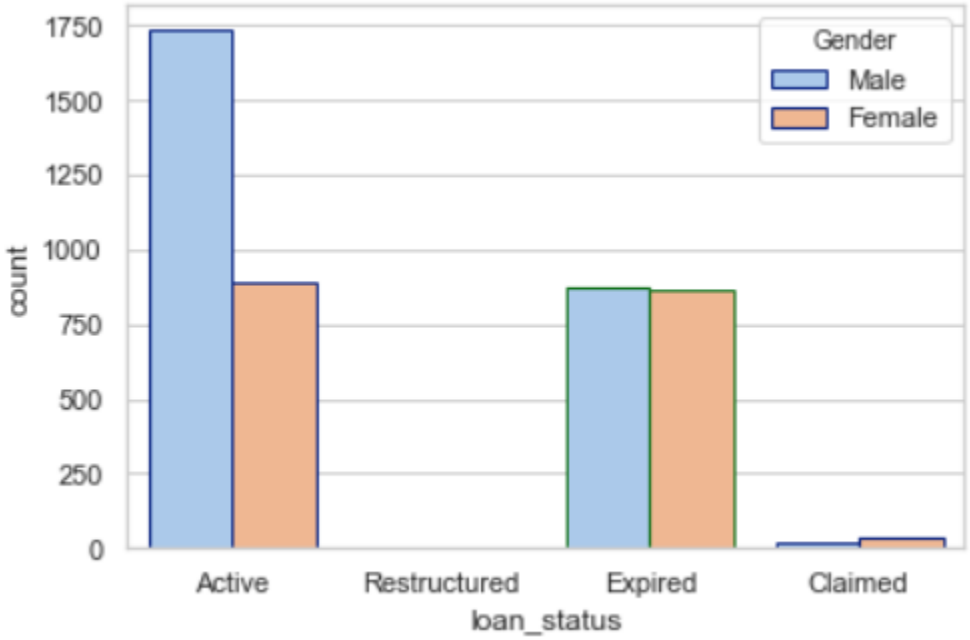

Figure 9: Loan status to gender

Figure 9 shows the relationships between loan status and gender by counting. Then findings show that male with Active status are higher than others.

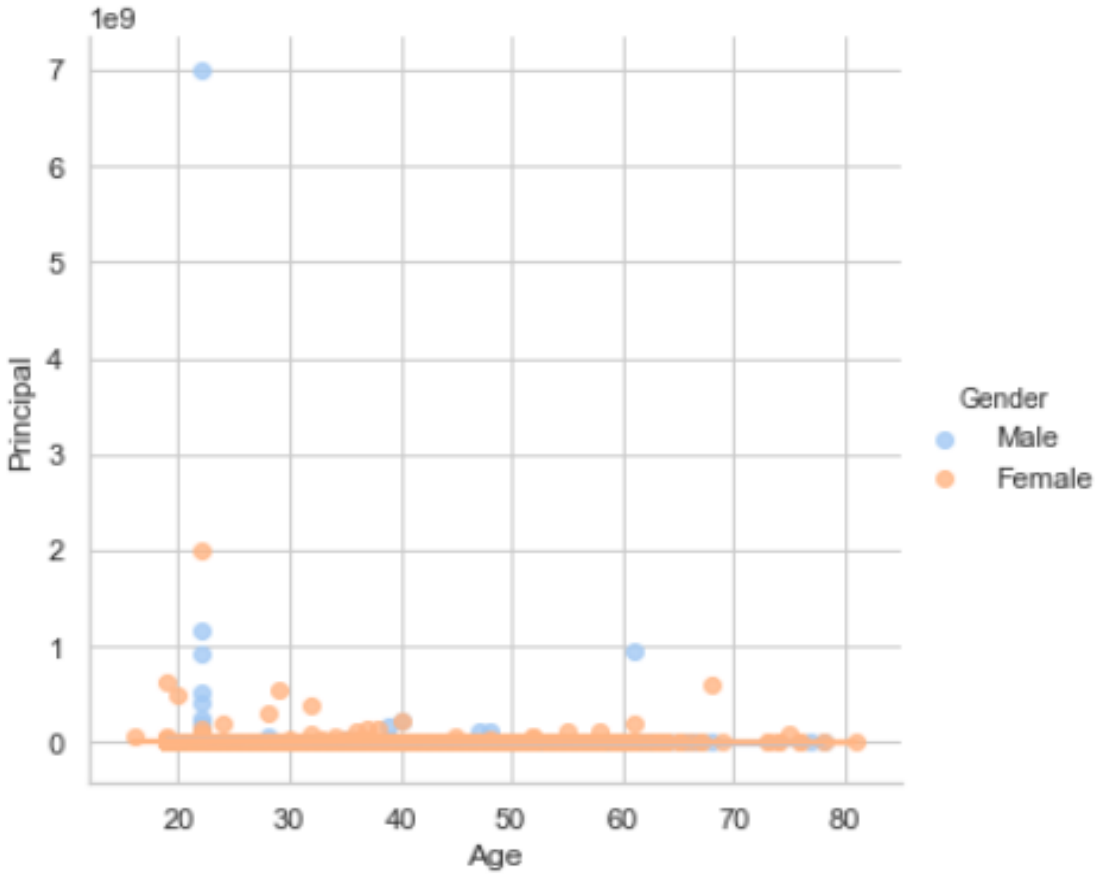

Figure 10: Principal and age with gender

Figure 10 shows how principal and age are related by using the gender of applicants where males are represented by blue colour and red by females.

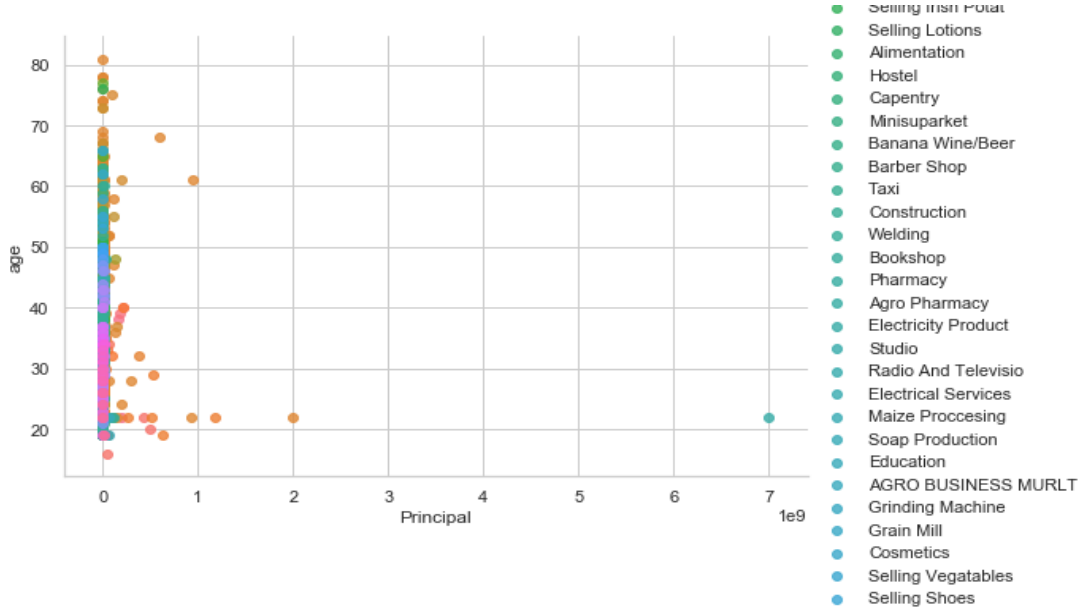

Figure 11.a: Principal and age with sectors

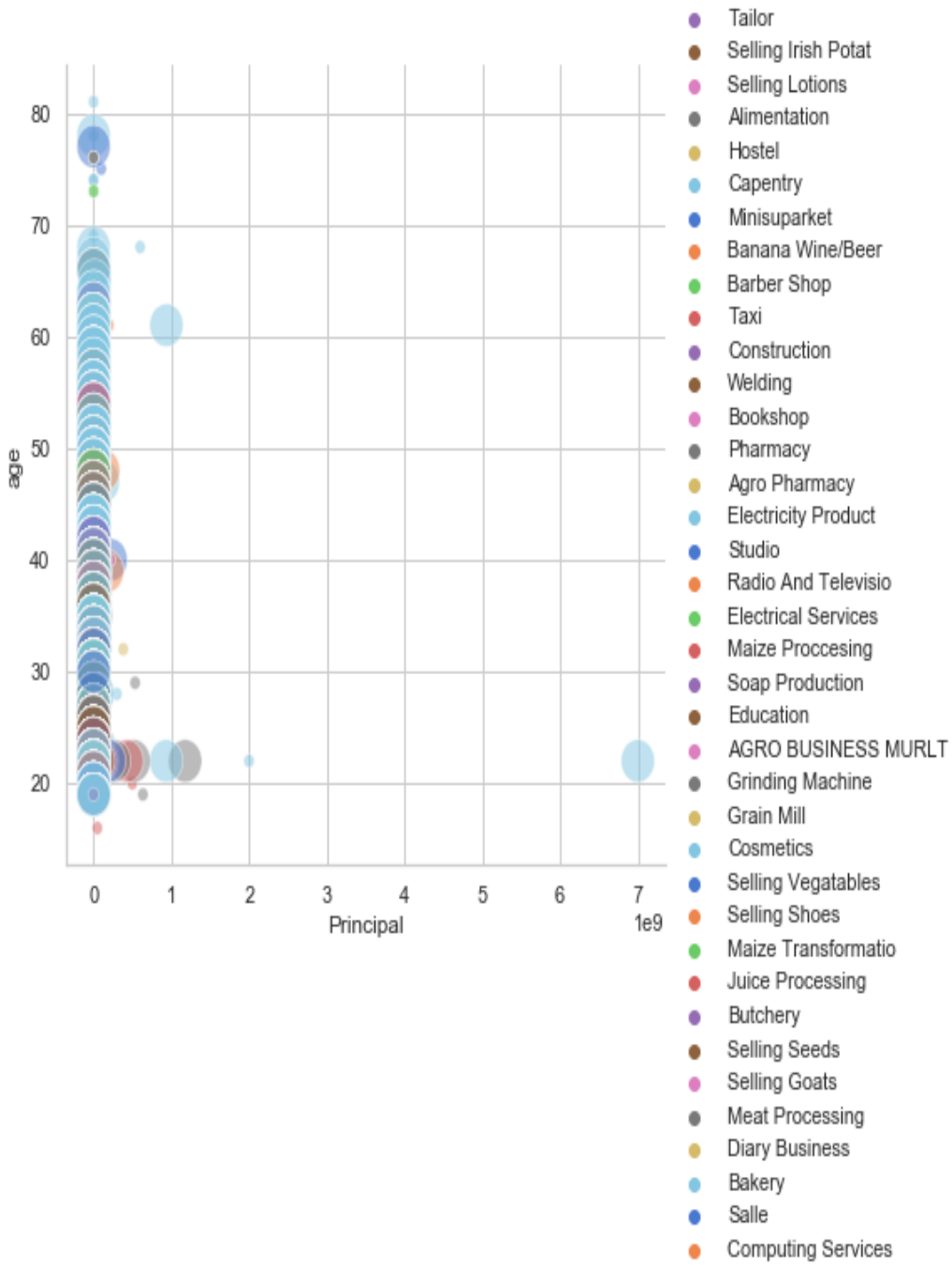

Figure 11.b: Principal and age with sectors

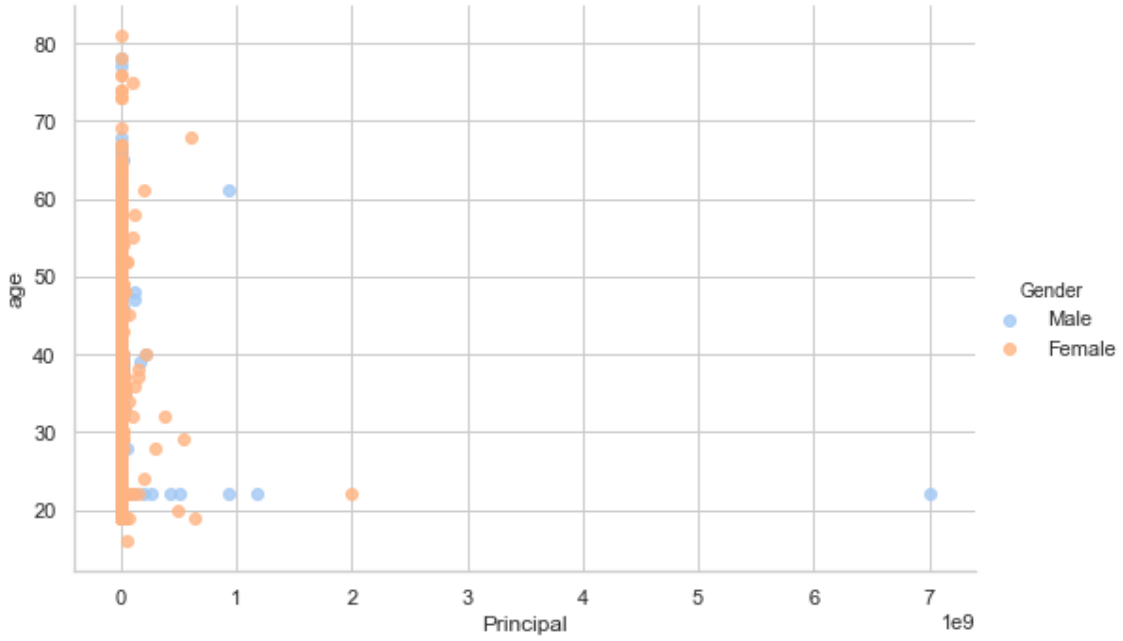

Figure 11.c: Principal and age with gender

Figure 11 (a, b & c) shows how principal and age are related based on sectors (education) and gender of applicants, Researcher can visualise all on different graph because it cannot displayed on one graph according to its height with many sectors but using the data mining software they are clearly visualised all.

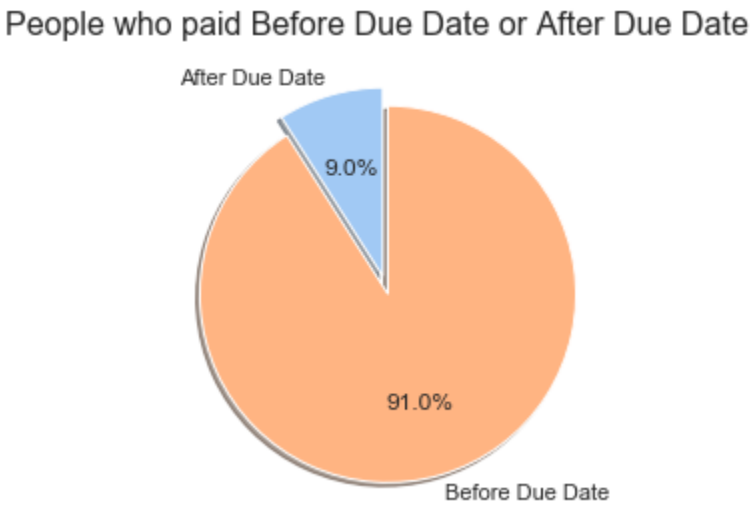

Figure 12: Applicants who pay before due date

Figure 12 shows percentage of people who paid before due date which are equal to 91% (Orange colour) and those who paid after due date are equal to 9% (blue colour).

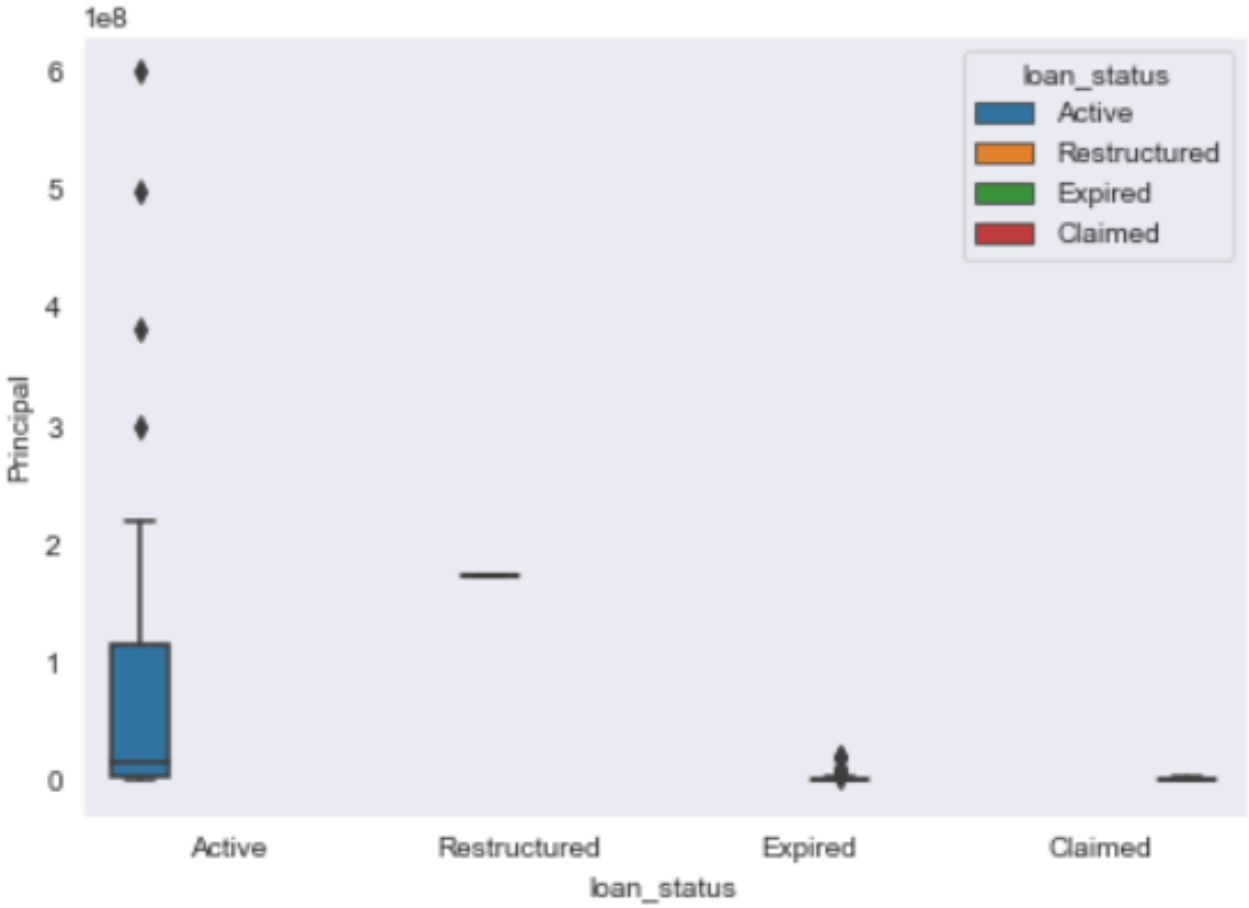

Figure 13: Principal and loan status

Figure 13 shows the correspondence between loan status and principal and how it is related and how they vary.

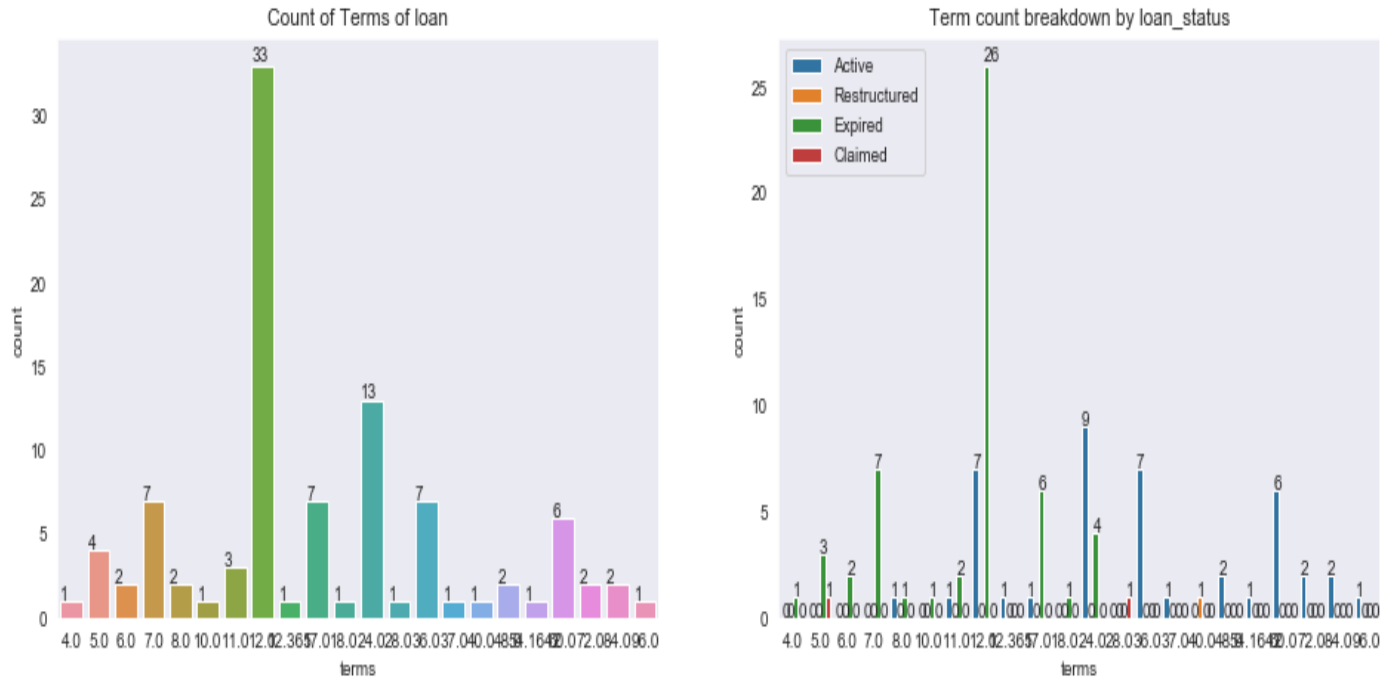

Figure 14: Count of terms of loan

Figure 14 shows terms of payment with its count where term of 12 months has high frequency than others, followed by term of 24 months with 33 and 13 respectively as shown on figure 14. Other side, it shows term count breakdown by loan status where term of 12 is also the one which has higher frequency than others. On figure14 on left part, the green colour represents expired, blue represents active, orange is restructured and red colour represents claimed status as it is shown on right part.

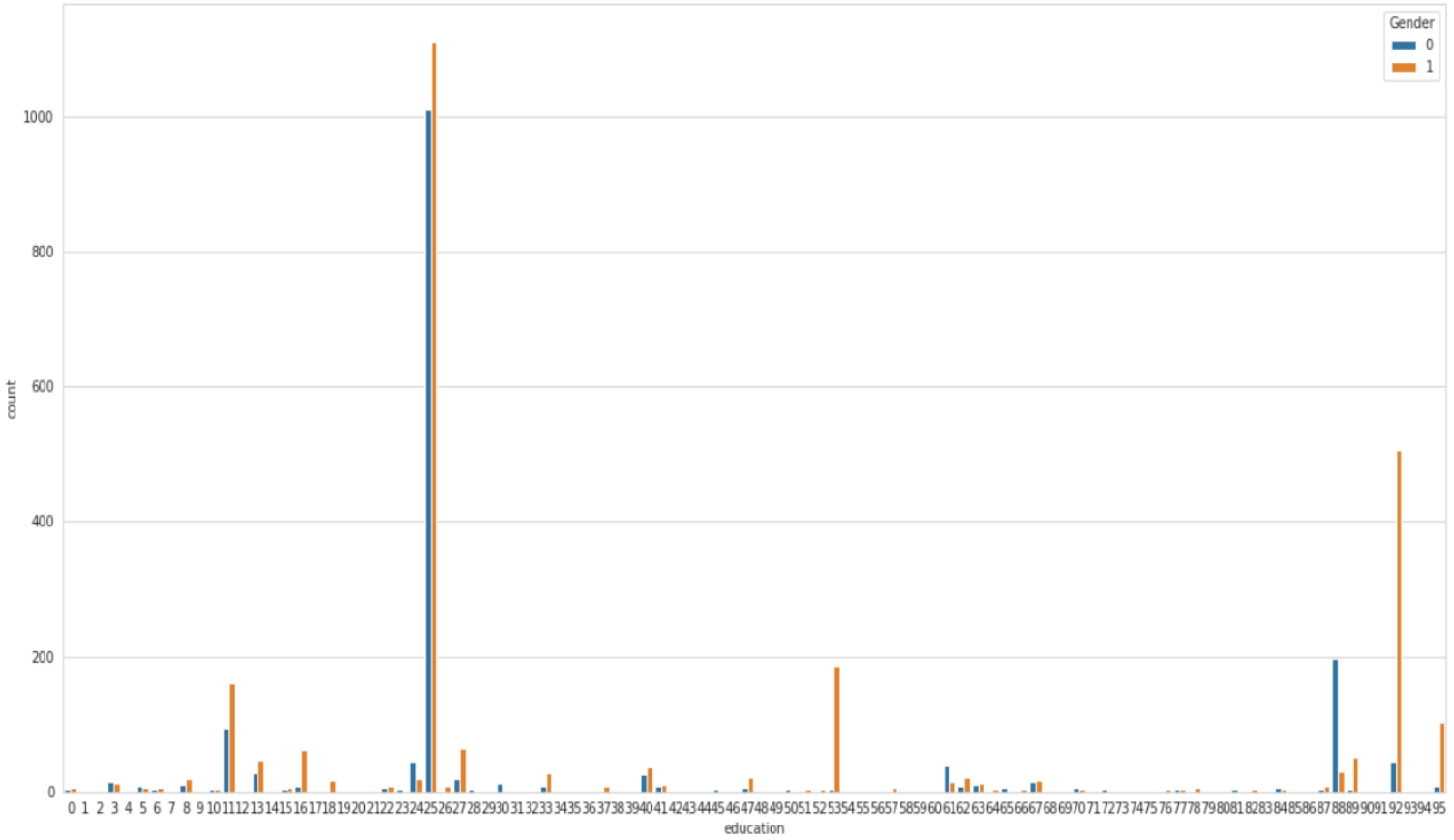

Figure 15: counting sectors with gender

Figure 15 shows the count of customers using education (sector or fields of interest) and gender, referring to the education every number on the graph represents a sector of interest of every participant. On this figure, the number 25 represents commerce is the one with higher frequency where blue is female and red is male.

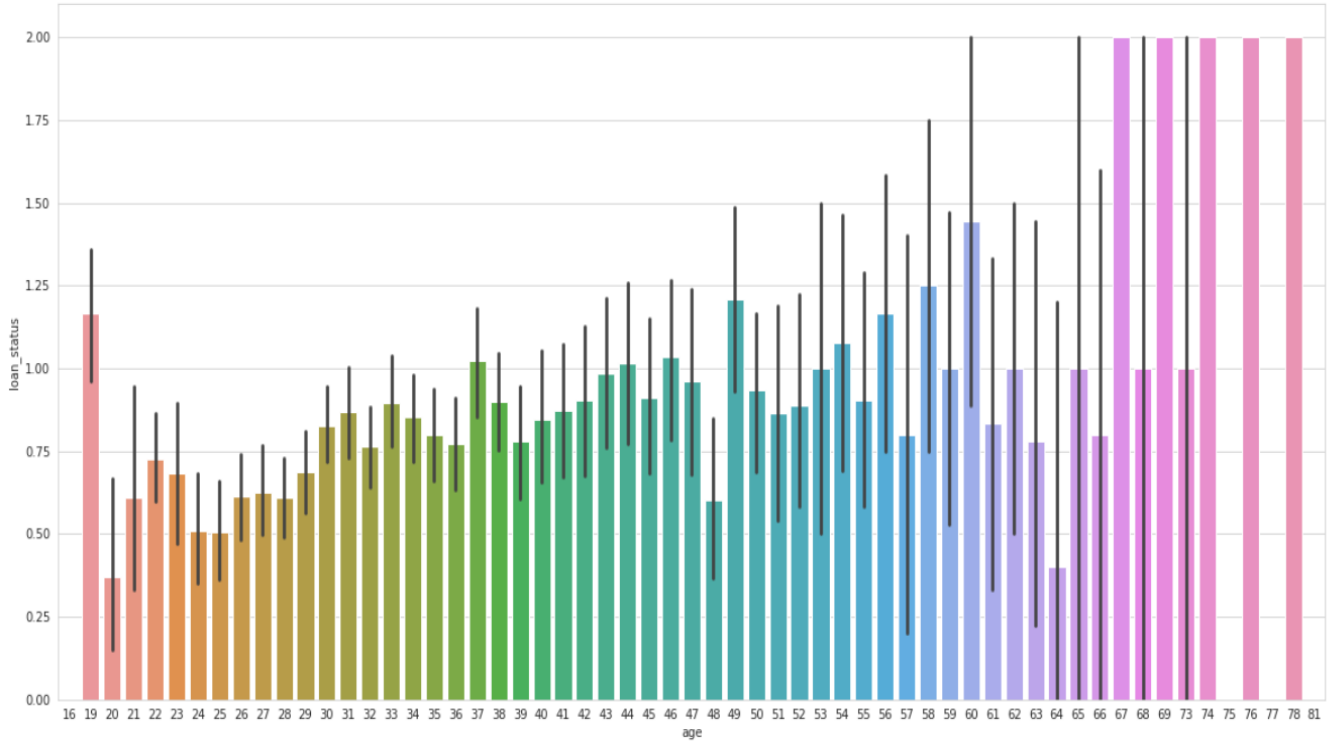

Figure 16: age and loan status

Figure 16 shows the relation between age and loan status. It indicates that ages between 30 and 60 are in good loan status than other rest, the green colour represents expired, blue represents active, orange is restructured and red colour represents claimed status, but there are ages with mixed colour and produces new kind of colour because it is a variation of age to loan status.

Figure 17: Total age distribution across dataset

Figure 17 shows how age is moving according to counting with total age distribution across dataset.

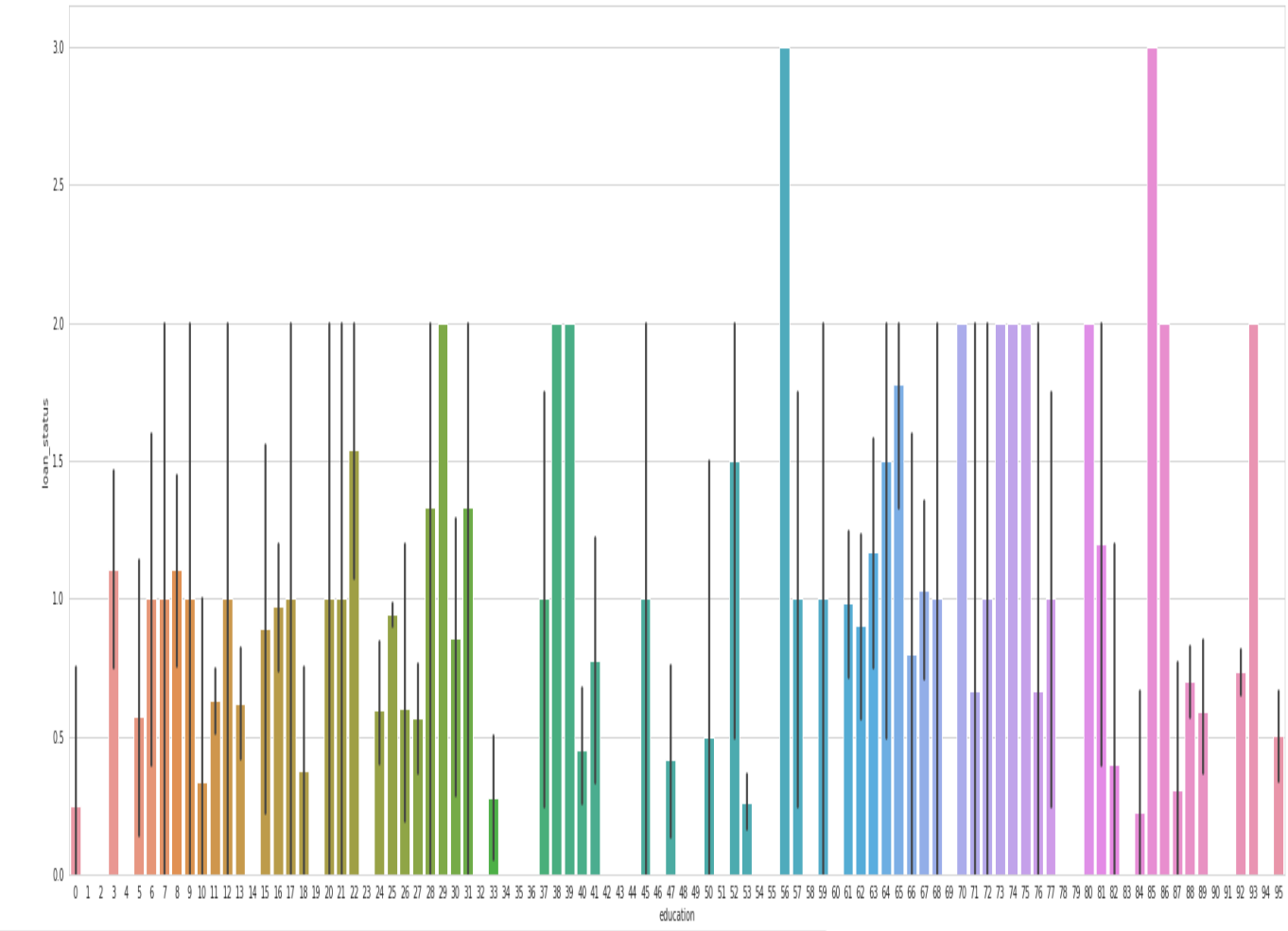

Figure 18: presentation of sectors with loan status

Figure 18 shows how education and loan status are related but sectors are represented by number in order to facilitate visualisation and minimise the size of words where 56 represents commerce and 85 represents selling cattle those two have high frequency compared to others. The green colour represents expired, blue represents active, orange represents restructured and red colour represents claimed status, but there are education (sector of interest) with mixed colour and produces new kind of colour because it possesses different loan status at the same time.

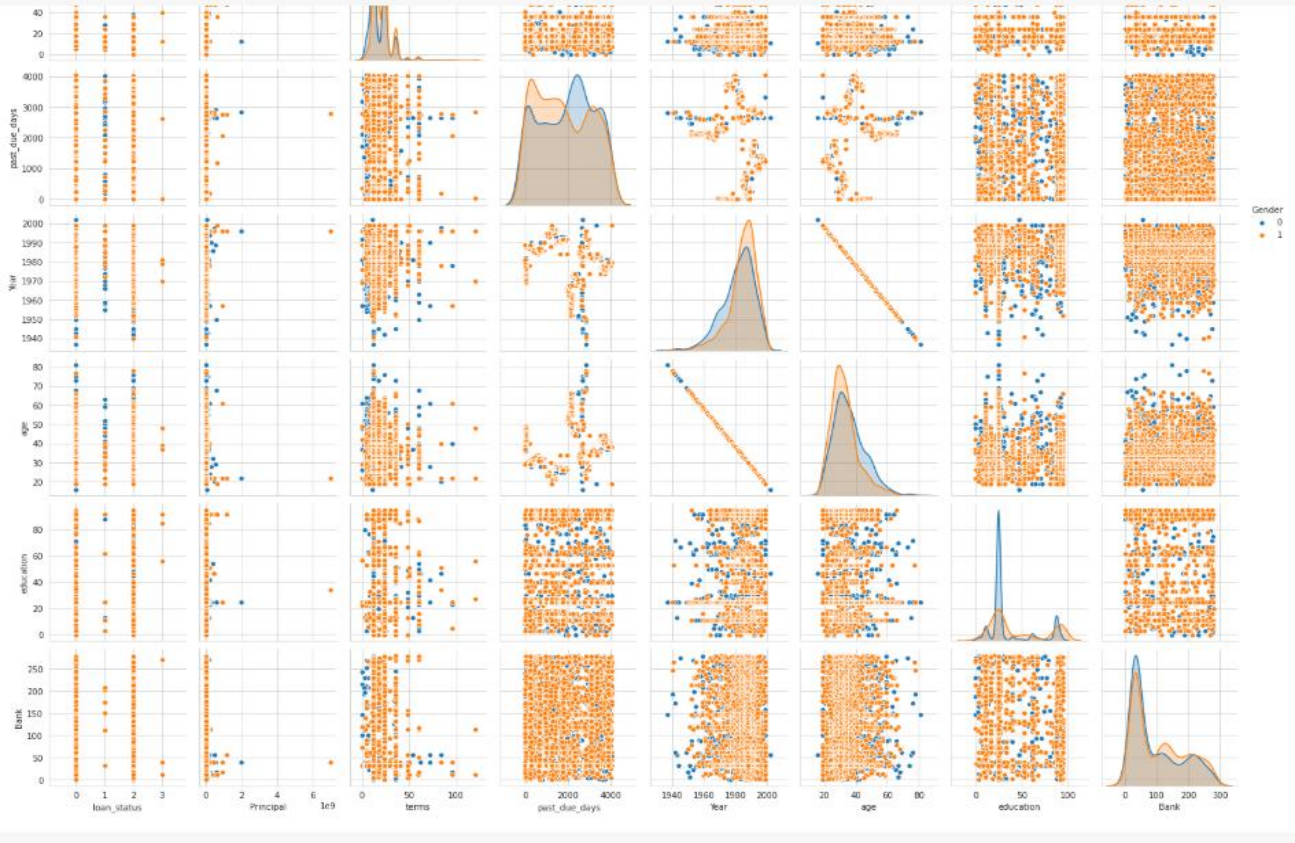

Figure 19: presentation of whole data set

Figure19 shows whole dataset in one graph by using gender where red colour is representing male and blue colour representing female.

4.2. DESCRIPTIVE ANALYSIS

Descriptive analysis is basically used to produce correlation, cross-tabulation, frequency, etc. It is used to determine the similarities in the data and to find existing patterns. Furthermore, descriptive analysis is used to develop the captivating subgroups in the major part of the data available.

The analysis emphasises on the summarization and transformation of the data into meaningful information for reporting and monitoring.

a) Imbalance Data

A dataset is imbalanced if the classification categories are not approximately equally represented. Recent years brought increased interest in applying machine learning techniques to difficult “real-world” problems, many of which are characterized by imbalanced data. Additionally, the distribution of the testing data may differ from that of the training data, and the true misclassification costs may be unknown at learning time (Chawla, 2005).

In our dataset, there is a great imbalance data as it is shown in figure 20.

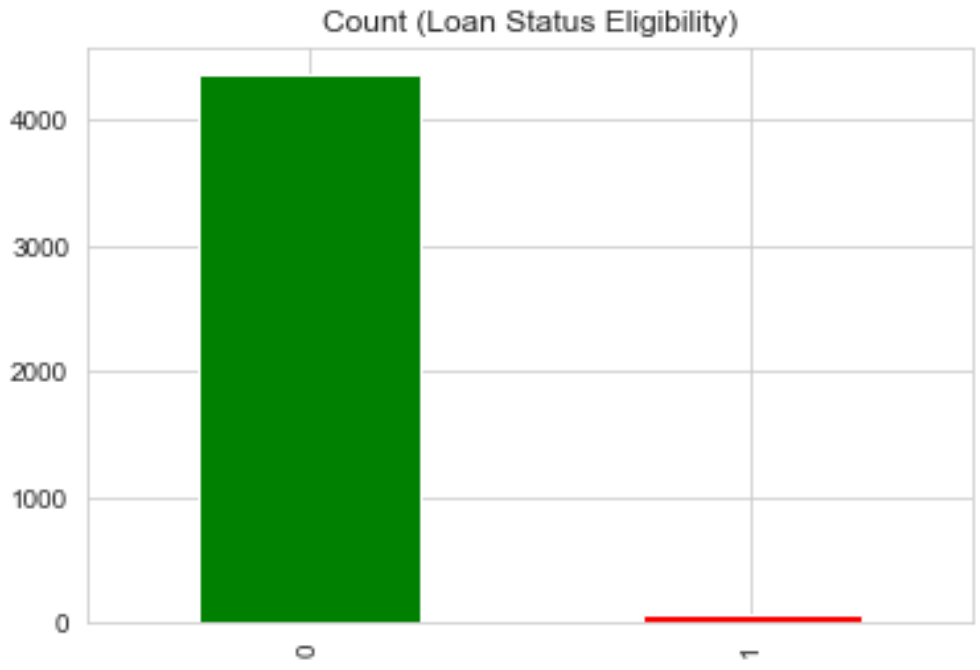

Figure 20: Loan eligibility counting on imbalanced data

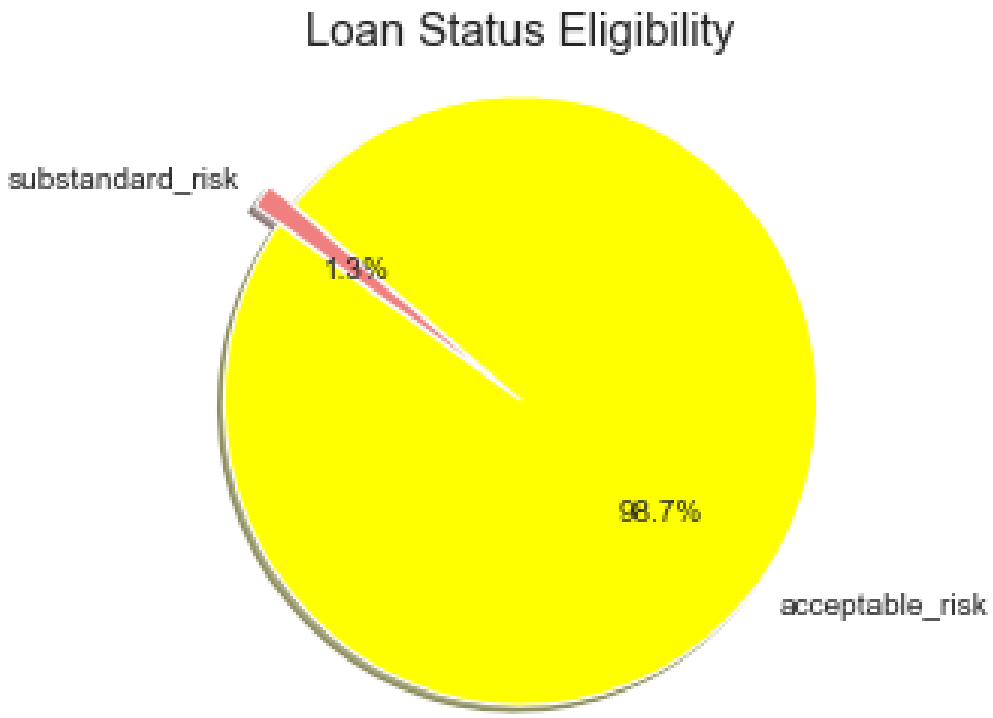

Figure 21: Loan eligibility percentage

Figure 20 and Figure 21 show that there is a high level of data imbalance on response variables. The loan status for SMEs the positive class (acceptable risk) is 98.7% while negative class (substandard risk) is 1.3 percent. The disparity between the positive and negative class leads to poor performance of machine learning algorithms. Figure 20 green colour represents acceptable risk and red colour represents substandard risk. On figure 21 red colour represents substandard risk and yellow colour represents acceptable risk.

b) Balance Data

When conducting a supervised classification with machine learning algorithms such as Random Forests, logistic regression and decision tree. One recommended practice is to work with a balanced classification dataset in order to get a good performance of machine learning algorithms.

Figure 22: Loan eligibility counting on balanced data

Figure 22 shows a balance data on the response variable. The proportions of the loan status for SMEs are 0.5 and 0.5 for the positive class (acceptable risk) which is represented by green colour and negative class (substandard risk) which is represented by red colour respectively. The balance data increases the performance of machine learning models and minimizes misclassification.

c) Feature Selection Using Filter Method.

Relevant features that correlate with loan status eligibility for the SMEs were filtered using Pearson correlation as shown in Figure 23.

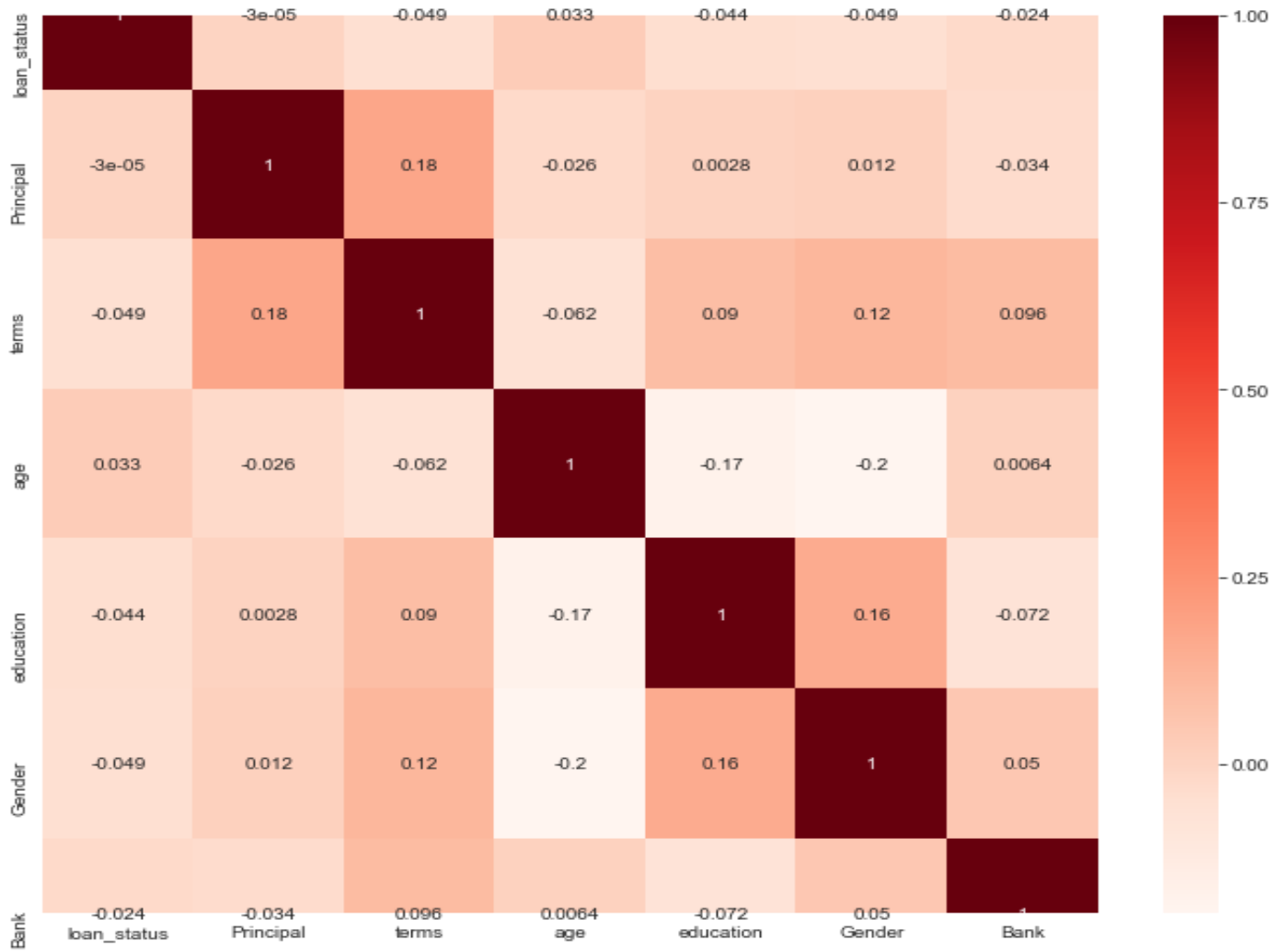

Figure 23: Feature Selection Using Filter Method

Table 1: Correlation between Variables and loan status (eligibility) of SMEs in Rwanda

Features	**Coefficients**
Terms	0.049326
Gender	0.049145
Education	0.044384
Age	0.032718
Bank	0.024052

Table 1 depicts the correlation between selected variables and label (loan status) for SMEs in Rwanda. From Table1 the model with the highest coefficient (most correlated with dependent variable) is terms (0.04933), followed by gender (0.04915), education (0.04438), age (0.03272), bank (0.02405).

CHAPTER 5. RESULTS, DISCUSSIONS AND CONCLUSION

5.1. RESULTS

On one hand the data was split into 80% of the training set and 20% of the test set. Training sets were used in training the model and give a biased sagacity of model effectiveness. On other hand, the test set was held back from training of the model to be used to evaluate model performance.

a) Comparing Machine learning performance on imbalance test data

Metrics were used to compare the performance machine learning models on Table2 such as accuracy, F1 score, recall and precision score. From Table2, the accuracy for each of the models were: SVM (0.9797), Logistic regression (0.9797), Decision tree (0.9795), Random forest (0.9785) and KNN (0.9763).

From Table2, the F1 score for each of the models were: Decision tree (0.5568), SVM (0.4949), Logistic regression (0.4949), Random forest (0.4946) and KNN (0.4940). From Table2, the Recall score for each of the models were: Decision tree (0.5492), SVM (0.5), Logistic regression (0.5), Random forest (0.4994) and KNN (0.4983). Decision tree is a better model in prediction of loan eligibility status based on imbalanced data.

Table 2: machine learning models' comparison on imbalanced data

	Model	**Precision score**	**Recall score**	**F1 score**	**Accuracy**
1	Logistic Regression	0.4898	0.5000	0.4949	0.9797
2	Random Forest	0.4898	0.4994	0.4946	0.9785
3	Decision Tree	0.5677	0.5492	0.5568	0.9695
4	KNN	0.4898	0.4983	0.4940	0.9763
5	SVM	0.4898	0.5000	0.4949	0.9797

b) Comparing Machine learning performance on balanced test data

Metrics were used to compare the performance machine learning models on Table3 such as accuracy, F1 score, recall and precision score. From Table3, the accuracy for each of the models were: The model with the highest accuracy is random forest (0.9909), Decision tree (0.9909), followed by SVM (0.9848), KNN (0.9695) and Logistic regression (0.4726). From Table3, the F1 score for each of the models were: Random forest (0.9908), Decision tree (0.9908), SVM (0.9847), KNN (0.9695) and Logistic regression (0.3209). From Table3, the Recall score for each of the models were: Random forest (0.9913), Decision tree (0.9913), SVM (0.9855), KNN (0.9711) and Logistic regression (0.5). Random forest and decision tree are outstanding models in prediction of loan eligibility status based on balanced data.

Table 3: machine learning models' comparison on balanced data

	Model	**Precision score**	**Recall score**	**F1 score**	**Accuracy**
1	Logistic Regression	0.2363	0.5000	0.3209	0.4726
2	Random Forest	0.9905	0.9913	0.9908	0.9909
3	Decision Tree	0.9905	0.9913	0.9908	0.9909
4	KNN	0.9697	0.9711	0.9695	0.9695
5	SVM	0.9844	0.9855	0.9847	0.9848

c) Receiver Operating Characteristic Curve (ROC)

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.

Figure 24 shows the ROC of machine learning models on test data. The machine learning model with the highest ROC was Random forest (0.999), followed by Decision tree (0.9988), SVM (0.993), KNN (0.9884) and Logistic regression (0.5585).

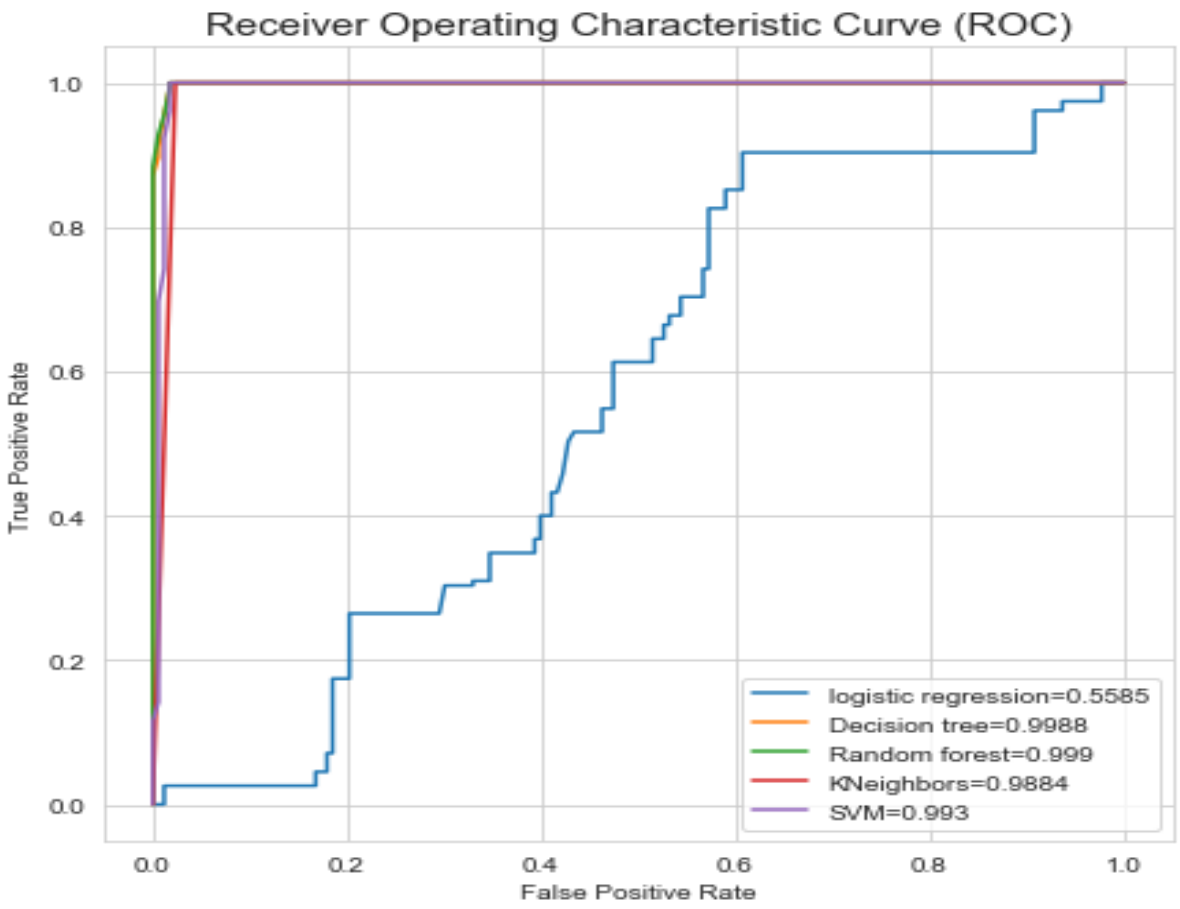

Figure 24: ROC for five machine learning models

5.2. DISCUSSION

In this thesis, machine learning models were compared to their respective performance in predicting loan eligibility for SMEs. The best machine learning model was taken as the one with highest accuracy F1 score, recall score and ROC.

The obtained results for imbalanced data as displayed in Table2 shows that the decision tree was the best performing model based on the precision, recall and F1 score of the prediction model are all above 0.54. This indicates that the decision tree model has a good generalization ability on this imbalanced dataset. The accuracy for the machine learning models on Table2 were comparatively high since all are above 0.96.

From the results in Table2, performance of machine learning models is very poor when trained using imbalance data. Imbalanced data led to low precision score, recall score and F1 score. This is due to high level of misclassification where the minority class (substandard risks) has learned from the majority class (acceptable risks). This leads to poor prediction of SMEs eligible to apply for the loan. The misclassification is attributed to high level of false positive and false negative on predicted outcomes. Therefore, predicting loan eligibility for SMEs using imbalance data in this case would not be suitable

The obtained results for balanced data are displayed in Table3. The results depict that, random forest and decision tree have comparable performance than that of SVM and logistic regression, but the random forest still performs the best, on ROC through results on metric comparison as displayed in Figure 24, though from Table3 evaluation metrics for random forest and decision tree are the same. But the performance for machine learning models using logistic regression improved after balancing the data. The accuracy, precision, recall and F1 score of the random forest, decision tree, SVM and KNN were all above 0.96 while logistic regression performs poorly on balance data where values of metrics was utmost 0.5.

The high level of precision recall and F1 score in Table3 shows that false positive and false negative are tending towards zero as recall and precision score tend to 1. Balancing of data reduces misclassification to almost zero leading to high recall and precision score. **In this thesis random forest and decision trees have dominated and depict outstanding performance in prediction of loan eligibility for SMEs based on balanced data.**

Figure 24 shows the ROC curves of the five models. When the ROC curve is closer to the upper left corner, the recall rate of the model is higher. The point on the ROC curve closest to the upper left corner is the best threshold with the least classification errors, and the total number of false positive examples and false negative examples is the lowest. The machine learning model with the highest ROC was random forest (0.999) while the least was logistic regression (0.5585). We can conclude obviously that the random forest algorithm outperforms the other four machine learning models for this particular case study. Random forest is therefore recommended in this study to identify SMEs which are eligible for loan application based on ROC.

Confusion Matrices for Machine learning models

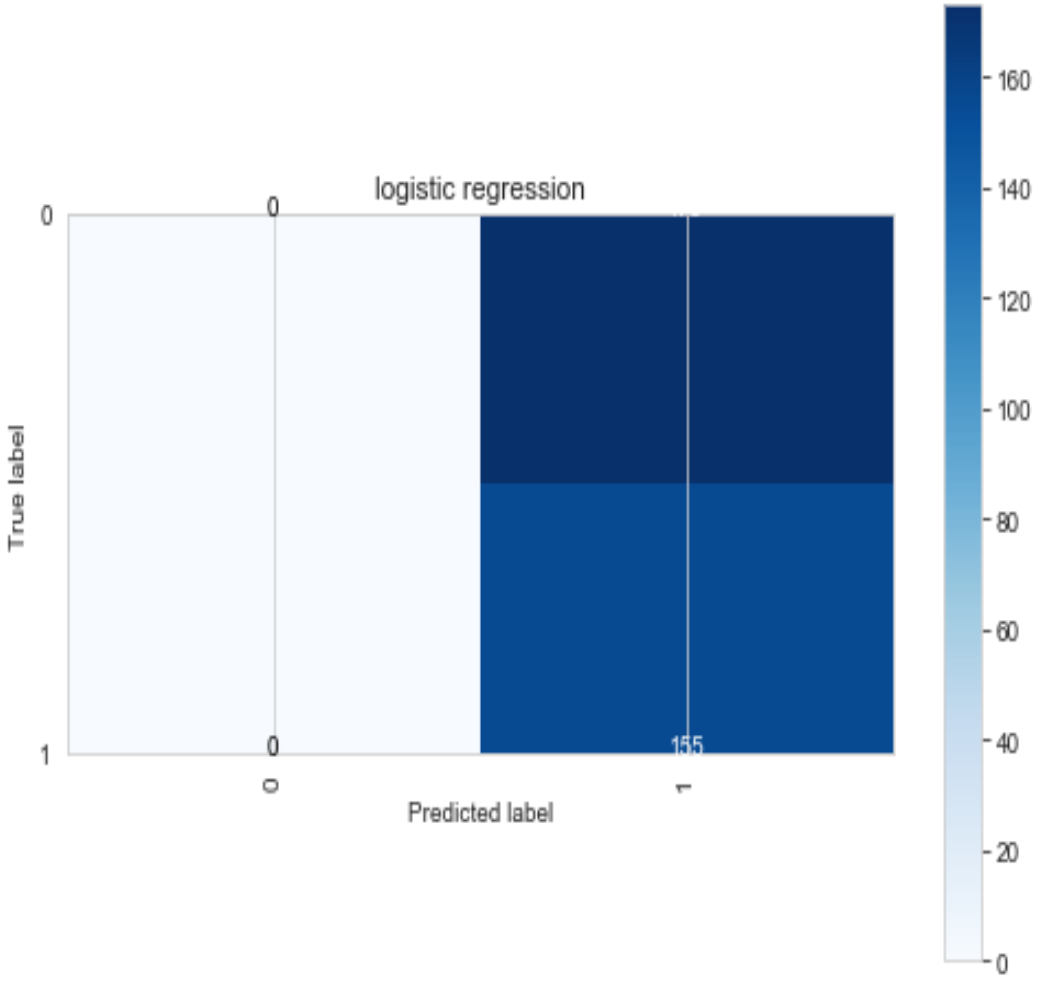

Figure 25: Logistic regression

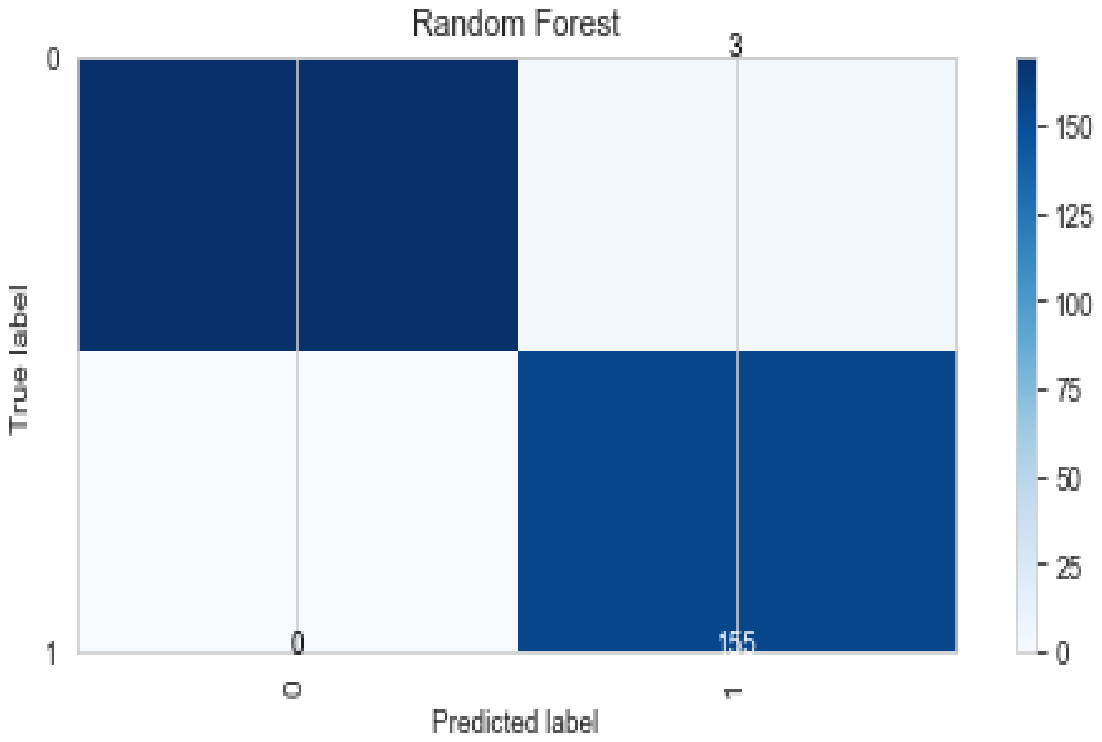

Figure 26: Random forest

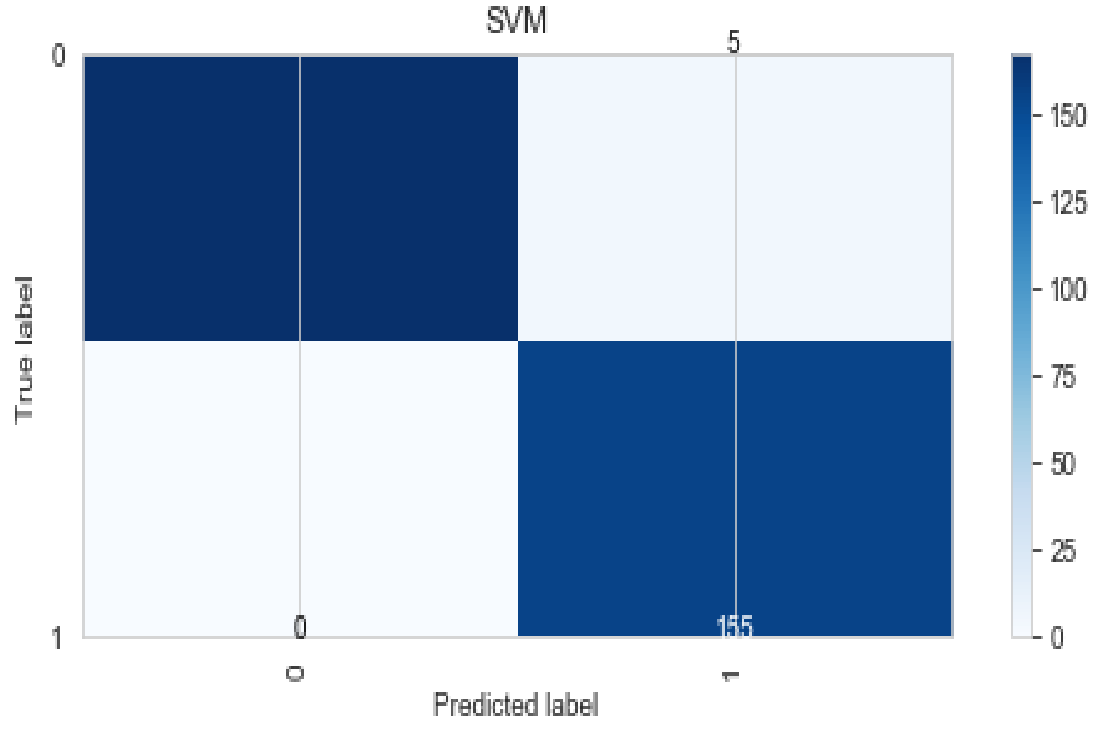

Figure 27: SVM

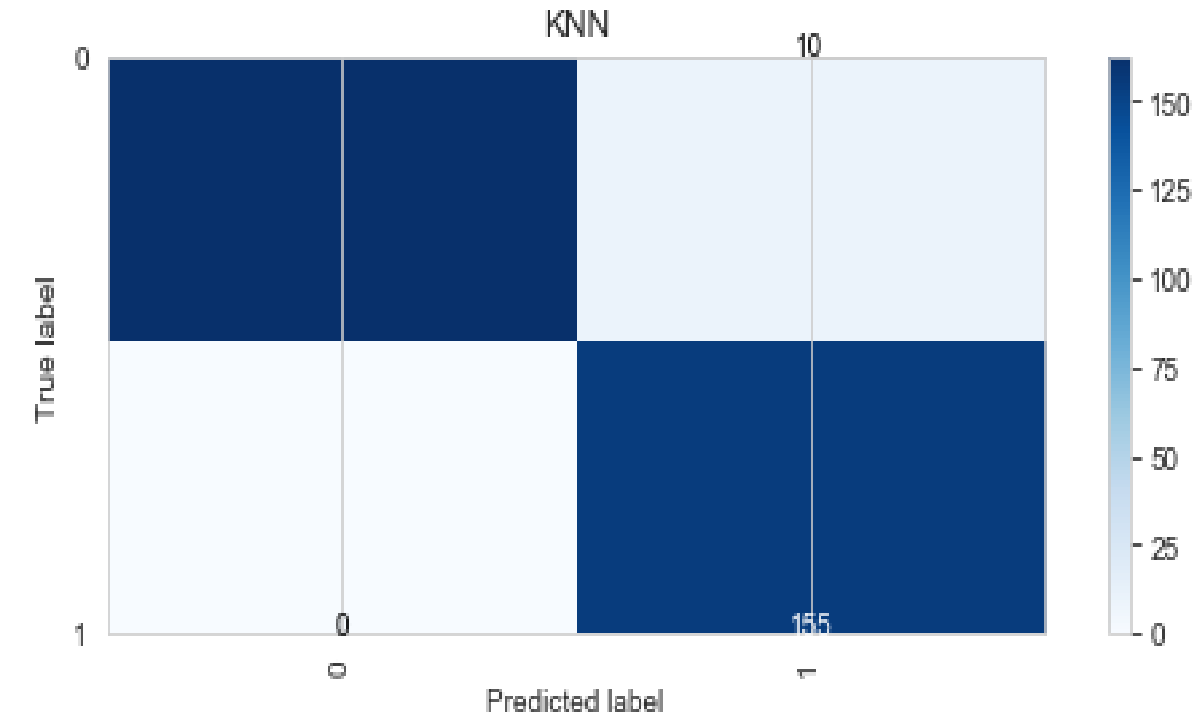

Figure 28: KNN

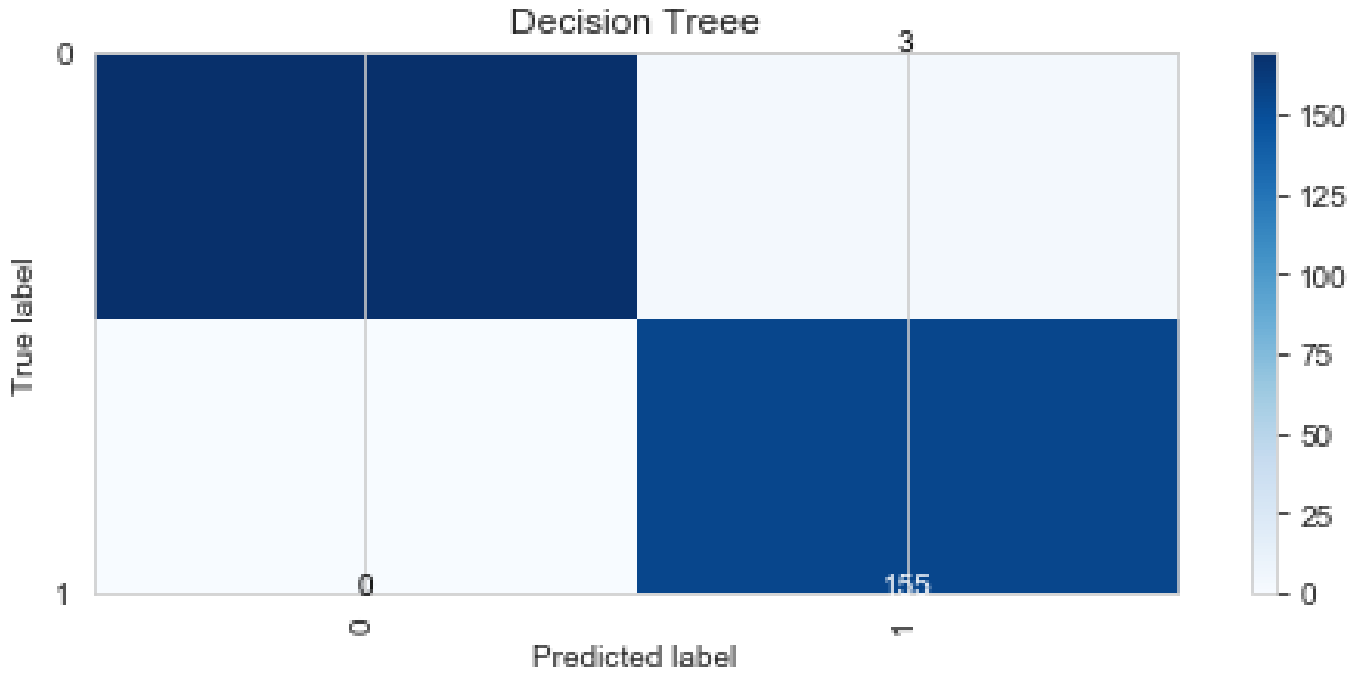

Figure 29: Decision Tree

The confusion matrices above show that random forest and decision tree have lowest false positives and negatives respectively on Figure 26 and Figure 29. Thus, these are suitable models for predicting loan eligibility for SMEs in Rwanda. It also proved by the training and testing score as it is shown,

	Model	Training_Score	Testing_Score
1	Decision Tree	0.998587	0.746893
2	Random Forest	0.998304	0.738983
0	Logistic Regression	0.597230	0.574011

Figure 30: Comparison of models

On research question 1 (Are there hidden knowledge to be discovered in a BDF's loan dataset?), It is responded by chapter 4.1. Where we found data visualisation, we tried to visualise hidden data from BDF dataset, we used different graphs to make data more visible followed by legends. There are a lot of meaningful information on that graph, we can that male has high attendance compared to the female on figure 6. There are other graphs with real meaning for different information from figure 6 to figure 19.

On research question 2 (How BDF loan data can be used to predict for future performance?), this question has responded in part 5.3, the matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making. after considering random forest and decision tree as good models, the matrix on figure 26 and 29 show that True positive is blue at 150 level means the predicted value matches the actual value, The actual value was positive and the model predict a positive value. Our model predict good performance in future, Then we have to encourage BDF because prediction is totally positive with lowest false positives and negatives where the risk is minimal.

On research question 3 (How Predictive model can facilitate BDF management in decision making?) As it is showed by ROC graph random forest's performance is 99.9%, it means that the system used by BDF is effectively performed but we recommend them to make attention when providing loans to old persons above 60 years, they meet with some challenges on payment of loan and to mobilize women for their products because their attendance is still low compared to the men.

Apparently, there is evidence that BDF is contributing towards the ease of access to finance among SMEs in Rwanda, however, this is contribution is significantly premised on one aspect of business and investment services. To further increase ease of access to finance, BDF has to strengthen business advisory services to all SMEs in its docket. That might make access to finance easy even for the SMEs who reported having uneasy access to finance.

To ensure, that all SMEs receive the full package of BDF services, BDF could consider upping its mobilization exercises every time they plan to conduct education sessions for SMEs, BDF has a task to avoid delays of giving a loan to a beneficiaries in order to perform their task accordingly and effectively.

To further make the BDF activities effective, the people charged with educating SME representatives could consider conducting business advisory in order to maintain good management of businesses and avoid losses of the SMEs which lead to fail loan payback.

5.3. CONCLUSION

Machine learning models are better tools in identifying loan eligibility of SMEs. Comparing metrics and performance of machine learning models is vital in order to identify the best model that can give outstanding outcomes in predicting loan eligibility for SMEs. This thesis found out that using imbalance data to predict loan eligibility resulted in low recall and precision score due to high misclassification. Using a balanced data random forest and decision tree produced comparable performance than that of SVM, KNN and logistic regression. However, random forest was preferred to other machine learning models since it had the highest precision, recall, F1 score, accuracy and ROC. Thus, random forest is recommended in this study to identify SMEs which are eligible for loan application. As shown in the model, at least 96% are able to fulfil and pay back loan given, it means the management is working a good job but they have to improve until they eliminate that 4% missing to maximize.

The rise of Big Data and data science approaches, such as machine learning and deep learning models, have a significant role in loan eligibility modelling. In this thesis, we have shown that it is important to make checks on data quality (in the preparation and cleaning process to omit redundant variables), and it is important to deal with an imbalanced training dataset to avoid bias to a majority class.

Business intelligence is a rapidly evolving field carrying tremendous potential for improving efficiency and competitiveness of corporations across the globe. Data mining and machine learning are useful tools in this field that can extract valuable information from big data to aid business strategies. This thesis investigated the adoption of several popular modern day machine learning algorithms for forecasting business failures, i.e. loan eligibility. This information can help governments, investors, managers, and other stakeholders make intelligent economic decisions to avoid financial losses.

In general, machine learning is a powerful toolbox for financial and loan analysts to make predictions and discover patterns in the data with rigour. Many different models and validation techniques exist to aid data mining and decision making. It is difficult to determine if any machine learning technique is superior to others. In fact, as a data scientist and/or financial expert, it is perhaps more beneficial to harvest the strengths of different methods and combine them to make better business judgements

REFERENCES

1. Carol Hargreaves (2017). Machine Learning Application in the Financial Markets Industry, retrieved 20th May 2020 from https://www.researchgate.net/publication/328879367

2. Max Bramer (2016). Principles of Data mining, Retrieved April 20, 2020, from http://www.springer.com/series/

3. Steven S. Skiena (2017). THE MANUAL Data Science Design, Retrieved April 20, 2020, from http://www.springer.com/series/

4. Ivo D. Dinov (2018). Data Science and Predictive Analytics Retrieved April 20, 2020, from http://www.springer.com/series/

5. Charu C. Aggarwal (2015). Data Mining Retrieved April 20, 2020, from http://www.springer.com/series/

6. Laura Igual & Santi Seguí (2017). Introduction to Data Science, Retrieved April 20, 2020, from http://www.springer.com/series/

7. Dr. Jean Paul MPAKANIYE (2016). The role of business development funds (BDF) in creating employment for the youth in Rwanda.

8. Rwanda: Child Undernutrition in Rwanda Implications for Achieving Vision 2020. In *3rd National Food & Nutrition Summit 2014* (pp. 1–31).

9. Acs, Z. (2010). Entrepreneurship in Developing Countries. *Foundations and Trends® in Entrepreneurship*, *6*(1), 1–68. https://doi.org/10.1561/0300000031

10. Byaruhanga, J., Acosta, C., Ruranga, R., Ngabo, F., & Kabera, G. (2014). Cost of Hunger Study in *https://medium.com/@inforobertsmith36/top-use-of-data-science-in-finance-industry-b6b513db8828*

11. Business development funds (BDF) *(n.d.). Retrieved March 20, 2020, from* http://www.bdf.rw

12. Data science. Retrieved March 18, 2020, from *https://towardsdatascience.com/machine-learning-in-finance-why-what-how-d524a2357b56*

13. Data science. Retrieved March 18, 2020, from *https://data-flair.training/blogs/purpose-of-data-science/*

14. Models in data-mining. Retrieved March 18, 2020, from *https://www.educba.com/models-in-data-mining/*

15. Peter Martey Addo, Dominique Guegan & Bertrand Hassani (2018). Credit Risk Analysis Using Machine and deep learning Models

16. Albert Bifet, (2013), "Mining Big data in Real time", Informatica 37, pp15-20

17. Jaseen K.U. & Dr. Julie M. David (2011), issues, challenges and solutions: Big data mining

18. Viktor, Herna L. & Paquet, Eric (2005). Visualization Techniques for Data Mining

19. Banking Software: Data Mining & Banking Intelligence, retrieved 05th May, 2020 from, http://www.stratinfotech.com/banking_software/banking_software_business_intelligence_data_mining.htm

20. Petra Hunziker, Andreas Maier, Alex Nippe, Markus Tresch, Douglas Weers, and Peter Zemp, Data Mining at a major bank: Lessons from a large marketing application retrieved 05th May, 2020 from http://homepage.sunrise.ch/homepage/pzemp/info/pkdd98.pdf

21. Michal Meltzer, Using Data Mining on the road to be successful part III, published in October 2004, retrieved 05th May, 2020 from

22. http://www.dmreview.com/editorial/newsletter_article.cfm?nl=bireport&articleId=1011392&issue=20082

23. Christos, Stergiou and Dimitrios, Siganos, Neural Networks, retrieved 05th May, 2020 from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html #Introduction%20to%20neural%20networks

24. Chun, Se-Hak and Kim, Steven, Data mining or financial prediction and trading: application to single and multiple markets (2003)

25. J. M. Zytkow and W. Klösgen, Handbook of Data Mining and Knowledge Discovery. New York: Oxford, 2002.

26. The Wikipedia Guide, Introduction to Machine Learning, retrieved 05th May, 2020 https://www.academia.edu/41157657/Introduction_to_Machine_Learning_The_Wikipedia_Guide

27. Osmar R. Zaïane (1999), Principles of Knowledge Discovery in Databases retrieved 20th May, 2020 from https://exinfm.com/pdffiles/intro_dm.pdf

Dissetation

ORIGINALITY REPORT

16%	18%	7%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijarcsms.com Internet Source	3%
2	en.wikipedia.org Internet Source	2%
3	www.bdf.rw Internet Source	2%
4	Submitted to Greenwich School of Management Student Paper	2%
5	documents.mx Internet Source	2%
6	www.wideskills.com Internet Source	2%
7	arxiv.org Internet Source	2%
8	link.springer.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		