Evaluation of Machine Learning-Based Algorithm to Predicting Loan Default in Nigeria

Kingsley Oghenekaro EFEKODO
LCU/PG/003077

Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria

In Partial Fulfillment of the Requirements for the Award of Master of Science Degree (MSc) in Computer Science

2024

Certification

This is to certify that this study was conducted under my supervision by **Kingsley Oghenekaro EFEKODO** with Matric No. LCU/PG/003077 for the award of Master of Science (M.Sc) Degree in Computer Science in the Department of Computer Sciences, Faculty of Natural and Applied Sciences, Lead City University Ibadan, Oyo State, Nigeria under my supervision.

……..…….. ……..……

Prof. S. O. Akinola **Date**
Supervisor

……..…… ……..……

Dr. Sakpere Wilson **Date**
Head of Department

Dedication

This Project is dedicated to the highest God for the grace and strength to embark on this work and to my beloved family, whose unwavering support, encouragement, prayers and understanding have been the cornerstone of my journey to the glory of God. Your love has given me the strength to overcome challenges and pursue my dreams. The guidance, expertise, and mentorship of my supervisor have been invaluable throughout this research endeavour.

Acknowledgement

I wish to acknowledge and express my sincere gratitude to the Management of Lead City University Ibadan, Oyo State for their support, aligned academic facilities, and ambient environment provided to facilitate the achievement of this programme.

I also in no small measure appreciated the contribution and effort of my able supervisor Prof. S. O. Akinola for his unrelenting support and constructive criticisms in ensuring that this project is actualized. Thank you, sir, for your fatherly love. May Almighty God come to your rescue at the point of need.

I cannot help but appreciate the effort and support of the HOD [Dr. Wilson Sakpere]. I equally appreciate the other members of lecturers in the department of computer science department, Lead City University, Ibadan for their impacts in my academic journey.

I also equally appreciate my coursemates during the course of this study, these include people like Adegboro Samuel, Samuel Babafemi, Akinmoluwa Oluseye, Ajani Olaniyi, Sodeinde Adebisi, Bello Latifat, Adegoke Yinka, Allen Akintan, Olalekan Sunday, Odeyemi Tosin, Oyekunle Rotimi, Ajubade Olalekan, Ademuyiwa Olubunmi, Onalaja Olabisi, Folorunsho Temitope, Makinde Kemi, Extraordinary, Aje and host of others for their support and friendship role displayed. My God Almighty direct our part and show us the way to success.

Although, all the aforementioned Institutions and persons mentioned helped in one way or the other, meanwhile all errors in the research report if found are sorely mind.

Abstract

In the financial sector, accurately predicting loan defaults is critical. Traditional creditworthiness assessment methods, while thorough, often do not capture the dynamic and complex interactions within financial data. This necessitates advanced solutions like machine learning (ML). Traditional credit scoring systems are frequently unable to handle high-dimensional, non-linear data effectively, leading to significant financial losses due to inaccurate predictions of loan defaults. This study aims to harness advanced machine learning techniques to enhance the accuracy of predicting loan defaults, aiming to outperform traditional statistical models. Various machine learning algorithms including Logistic Regression, Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian Naive Bayes were applied to a dataset comprising diverse borrower characteristics and loan details. The selected dataset was an open source containing different datasets for both train and test Demographic data, Performance data and Previous loans data. It contained 3 different datasets for both train and test. The sample submission has 2 outcomes- good (1) or bad (0). The dataset systematically divided into two. 70% for the training set, 30% was the test set. These models underwent rigorous training and validation processes to ensure their robustness and reliability. The Gradient Boosting Classifier emerged as the most effective model, with an accuracy of 78.8%. This model significantly outperformed others by effectively capturing complex patterns in the dataset, thereby substantially reducing both false positives and false negatives. The study confirms that machine learning models, particularly the Gradient Boosting Classifier, offer superior predictive power in the context of loan default risk assessments. Financial institutions should consider integrating these models into their credit evaluation processes to enhance decision-making accuracy and minimize risks. Additionally, future research should explore the integration of more diverse data sources, including non-traditional variables that could affect credit risk assessments, and the application of deep learning techniques to further refine prediction accuracies.

Keywords: Accuracy, Classifier, Defaults, Financial, Machine Learning Models, Predicting, Cross-Validation, Data Imputation, Customer Segmentation, Nigerian Lending Market, Class Imbalance

Word Count: 300

Table of Contents

Content	Page
Title Page	
Certification	ii
Dedication	iii
Acknowledgement	iv
Abstract `	v
Table of Contents	vi
List of Tables	ix
List of Figures	x
List of Acronyms	xi
Chapter One: Introduction	
1.1. Background to the Study	1
1.2. Statement of the Problem	7
1.3. Aim and Objectives of the Study	9
1.4 Methodology Overview	**9**
1.5. Significance of the of Study	11
1.6. Scope of the Study	11
1.7. Limitation of Study	12
1.8 Operational Definition of Terms	13
1.9 Conclusion	14
Endnotes	

Chapter Two: Literature Review

2.1. Conceptual Review 25

2.1.1. Credit Worthiness 25

2.1.2. Loan 28

2.1.3. Artificial Intelligence in Loan Prediction 33

2.1.4. Machine Learning 36

2.1.4.1 Supervised Learning Algorithm 38

2.1.4.2 Unsupervised Learning Algorithm 51

2.1.4.3 Semi Supervised Learning Algorithm 52

2.1.4.4 Reinforcement Learning Algorithm 54

2.1.4.5 Performance Metrics 57

2.2. Methodological Review 64

2.2.1 Gradient Boosting Classifier 64

2.2.2 Gaussian Naive Bayes (GNB) 66

2.2.3 Random Forest 68

2.2.4 Decision Tree 72

2.3 Review of Related Work 75

2.4 Chapter Summary and Gap in Literature Reviewed 106

Endnotes

Chapter Three: Methodology

3.1. Research Approach 124

3.2.1 Hardware Minimum Requirements 125

3.2.2 Software Requirements 125

3.3. Research Design 126

3.3.1 Data Collection 126

3.3.2 Dataset Details 126

3.3.3 Dataset Description 127

3.3.4 Data Preprocessing and Balancing 128

3.3.5 Correlation Analysis 128

3.3.6 Data Splitting 129

3.3.7 Algorithm Used for Model Building 129

3.4. Model Evaluation and Performance 130

Endnotes

Chapter Four: Results and Discussion of Findings

4.1 Result on Dataset Processing 133

4.2. Model Building 138

4.2.1 Decision Tree 139

4.2.2 Gradient Boosting Classifier 142

4.2.3 Random Forest Classifier 145

4.2.4 Gaussian Naive Beyes 148

4.3 Discussion of Findings 152

Chapter Five: Conclusion

5.1. Summary of Findings 156

5.2. Conclusion 157

5.3 **Recommendations** **158**

5.4 Contribution to Knowledge 159

5.5 Suggested Area of Further Studies 160

Bibliography 162

Appendices 179

Bio-data 195

The University Compliance Certification 197

List of Tables

Table	Title	Page
4.1	Classification Report of Decision Tree	126
4.2	Classification Report of Gradient Boosting Classifier	129
4.3	Classification Report of Random Forest Classifier	132
4.4	Classification Report of Gaussian NB Classifier	134
4.5	Classification Report of the Models	137

List of Figures

Figure	Title	Page
1.1	Simple Illustration of Machine Learning Techniques	7
2.1	Machine Learning Working Process	27
2.2:	Types of Machine Learning Algorithms	28
2.3:	General RF Algorithm	34
2.4	Overview of Reinforcement Learning	46
2.5	Random Forest Flow Chart	59
2.6	Random Forest Training Flow Chart	60
2.7	Decision Tree Flow Chart	63
3.1	Conceptual Model of the Proposed Design	115
4.1	Columns Remaining after Data Cleaning	121
4.2	Plots Showing Loan Amount Distribution	122
4.3	Plots Showing Loan Amount by Good/Bad	123
4.4	Plots Showing Educational Level By Good/Bad	124
4.5	Confusion Matrix of Decision Tree	128
4.6	Confusion Matrix of Gradient Boosting Classifier	131
4.7	Confusion Matrix of Random Forest Classifier	134
4.8	Confusion Matrix of Gaussian NB Classifier	137

List of Acronym

Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the Curve
BC	Bayesian Classifier
CAGR	Compound Annual Growth Rate
CBN	Central Bank of Nigeria
DMBs	Deposit Money Banks
DT	Decision Trees
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes
PR	Precision-Recall
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines

Chapter One

Introduction

1.1 Background to the Study

Expenditure by consumers significantly influences the broader economic climate and inherent financial risks. As such, scrutinizing consumer credit is relevant since individuals often secure loans to fulfill their spending requirements[1]. The transaction value within the Consumer Marketplace Lending sector is projected to hit $78.57 million in 2023, with an expected growth at a Compound Annual Growth Rate (CAGR) of 5.29% from 2023 to 2027[2]. This growth trajectory suggests a rise to a total of $96.57 million by 2027. In the year 2023, it is forecasted that the average transaction value per user in the Consumer Marketplace Lending space will amount to $48.75 million. When examining the data on an international scale, the United States is set to record the highest transaction value, topping at $26.18 billion in 2023[3].

The Nigerian credit market is predominantly governed by the Central Bank of Nigeria (CBN), which is the apex regulatory body in the banking system and is therefore responsible for the DMBs[4]. Obviously, there are credit lenders that are not governed or supervised by the CBN. These include the Primary Mortgage Institutions, which report to the Federal Mortgage Bank, and the leasing corporations that operate under the Equipment Leasing Association of Nigeria's self-regulatory body[4].

The credit market in Nigeria falls under the jurisdiction of the Central Bank of Nigeria (CBN), the pinnacle regulatory authority for the nation's banking sector, thereby overseeing

Deposit Money Banks (DMBs)[4]. However, some credit providers operate outside CBN's purview, such as the Primary Mortgage Institutions, which are answerable to the Federal Mortgage Bank, and leasing firms that adhere to the self-regulatory framework of the Equipment Leasing Association of Nigeria[4].

A loan represents a pact where the lender extends credit to the borrower in the form of money, property, or other valuable items, contingent on the lender's confidence in the borrower's capacity to return the amount with interest. Presently, the issuance and management of loans constitute the central operation of nearly every banking institution. The interest earned from these loans is a critical asset for banks, often representing a substantial portion of their income.

Granting loans is an intrinsic and crucial activity within many financial institutions. These lenders constantly seek innovative business strategies to appeal to potential borrowers[5]. Nonetheless, some borrowers default on their loan repayments. A default can happen within the loan's tenure if the borrower falls short in fulfilling their payment commitments. Consequently, gauging the likelihood of default over time is crucial for robust risk management. Traditionally, credit officers manually review a borrower's credit history to evaluate their payment capabilities. However, the last decade has seen a shift in this practice, largely due to technological progress in the field[5].

Assessing credit risk stands as a cornerstone of financial risk management, with banks facing pivotal choices about extending credit to parties. The formidable challenge within financial circles is the prediction of insolvency or default. With the myriad of potential clients,

leveraging models and algorithms that minimize human error in consumer credit application analysis is imperative[1]. Many leading global financial entities have crafted sophisticated automated systems to model credit risk, equipping decision-makers with essential insights. Evaluating a borrower's creditworthiness before issuing credit is vital, as defaults carry considerable risk. Nonetheless, due to a scarcity of adequate data for deploying machine learning methodologies, some institutions still depend on traditional evaluation techniques[6].

The study of credit risk and the likelihood of default is an area of financial research rich with historical context. Credit risk describes the potential financial losses that lenders may face if borrowers do not fulfill their credit obligations[7]. This risk assessment is fundamental in setting the terms of credit approval and the interest rates charged by lending institutions.

While automatic credit scoring cannot replace the expertise of credit professionals, it has the potential to expedite the decision-making process during the initial stages of determining whether a case should be denied or require further analysis[8]. The presentation of risk associated with certain factors by credit rating agencies is not entirely accurate. Credit scores and ratings are constructed through a process of implicit assumptions made by agents, rendering them erroneous in nature. Consequently, individuals are exposed to a range of factors that have the potential to either underestimate or overestimate their creditworthiness[8]. The variability of credit ratings across different agencies, banks raises concerns about the suitability of using them as an independent variable in classifier development.

The contemporary utilisation of machine learning in credit risk analysis has been made possible by the accessibility of big data and the latest developments in computer processing capabilities. In contrast to conventional credit analysis methods that rely on statistical regression techniques and discriminant analysis of established variables, machine learning facilitates the use of algorithms to analyse datasets and generate a procedure that can predict the classification of an observation. The execution of such analytical techniques necessitates the utilisation of substantial datasets[7].

Automated loan default models have emerged as a popular credit risk scoring tool among financial lending institutions for granting loans to prospective borrowers in recent times[5]. Machine learning (ML) algorithms have proven to be an effective technique utilised by financial lending institutions to evaluate the credit risk of borrowers[9]. Effective strategies for managing credit risks are crucial for mitigating losses and enhancing profitability. Previous studies have approached the task of predicting loan default as a discrete binary classification problem, whereby an individual is categorised as either creditworthy or non-creditworthy. Linear Discriminant Analysis (LDA) and Logistic Regression (LR) are two commonly employed techniques for constructing credit scoring models. Following that, various classification algorithms, including Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and Bayesian classifier (BC), have been widely utilised for the purpose of forecasting the likelihood of default among borrowers[11,12,13]. However, the recent trend in Nigeria on various loan default scenarios especially in the online lenders need the embarrassing of machine learning algorithms to reduce bad debt. In the context of loan prediction, the occurrence of two types of errors can

result in a decrease in prediction accuracy. These errors are commonly referred to as "false positives" and "false negatives". A false positive refers to the rejection of an applicant who is actually credit worthy, while a false negative refers to the acceptance of an applicant who is not credit worthy. Both of these errors can lead to inefficiencies in the loan prediction process[5].

The majority of loan default prediction models are centred on the reduction of one or both of these types of errors. Despite the progress made in automating credit decision-based systems, there persist concerns regarding the overestimation of defaults[14]. Numerous scholars have utilised ensemble methodologies and the combination of two or more machine learning algorithms to address the issue of overestimation of defaults. The researchers were unable to determine the variables that effectively reduce the incidence of misclassifying creditworthy individuals as non-creditworthy or non-creditworthy individuals as creditworthy, commonly referred to as the false positive and false negative rates[15,16].

The field of machine learning falls under the umbrella of Artificial Intelligence (AI) and is concerned with the analysis of data structure and subsequent modelling for human utilisation[17]. Machine learning algorithms can be utilised by computers to train data and apply statistical techniques to generate values that fall within a predetermined range. The utilisation of machine learning techniques enables the creation of models through the utilisation of existing data, which in turn allows for informed decision-making based on said data inputs[18]. Supervised learning and unsupervised learning are the most prevalent

machine learning techniques, although there are also reinforcement learning and semi-supervised learning approaches available[19].

The supervised learning methodology involves providing the computer programme with pre-labelled sample inputs that correspond to the desired outputs[20]. Therefore, the machine learning algorithm is trained through the process of comparing its actual output with the original data in order to identify any errors or inaccuracies. This procedure optimises the model effectively. The categorization process is known as supervised learning due to the necessity of a third-party supervision of the computer programme. Classification, as perceived by data miners, is equivalent to prediction and forecasting since it utilises similar methodologies.

The objective of supervised learning techniques is to generate a model that represents the distribution of the response variable class based on sampled predictor labels[20]. The classifier that has been derived can now be utilised to assign class values to hypothetical scenarios in which the predictor labels are provided, but the corresponding class attribute is unknown. Diverse machine learning classification techniques have been formulated in the field of artificial intelligence in response to this.

Unsupervised learning refers to the utilisation of Artificial Intelligence algorithms for the purpose of identifying and comprehending patterns within datasets that lack any form of labelling or classification[21]. The method of classification can be broadly categorised into three distinct types, namely binary, multi-label, and multi-classification. Binary classification

is the most commonly utilised among the three types, owing to the fact that the majority of real-world tasks involve distinguishing between two distinct groups[22]. The efficacy of a machine learning system is contingent upon the calibre of data and the selection of representation features employed for its training. The efficacy of features is dependent upon the nature of the task at hand. However, it is commonly posited that specific features or feature sets are indicative of a given dataset and thus should be employed as input for the purpose of classification.

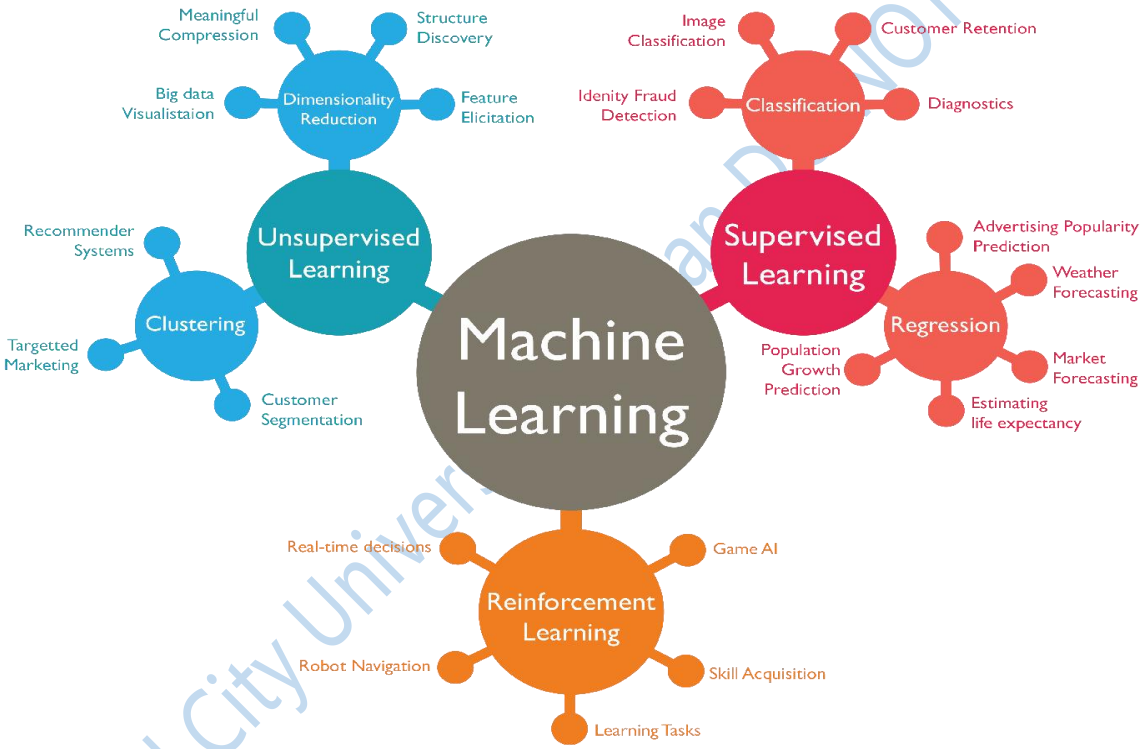


Figure 1.1: Simple Illustration of Machine Learning Techniques[23].

1.2 Statement of the Problem

One of the most critical risks that a financial organization must manage is credit risk. Because there is no profit without loan repayment, the issue of credit risk management affects all financial organizations that lend to individuals and legal companies. The

financial sector,banks, microfinance banks and the online lenders offers a variety of services to customer such as access to loan facility to be repaid in a specific period. Recently, most of the financial institutions especially online lenders in Nigeria has experienced a significant increase in the rate of lending caused by rising cost of living, inflation, poverty, unemployment, and other economic challenges facing the country.

However, many subscribers usually do not pay back their loans in time and many of the financial institutions usually call as soon as the repayment is due and will go to any length to embarrass and ensure you pay them in due time, which many loan eventually resulted into bad debt. To minimize the loss due to bad debt, this study proposed **Evaluation of Machine Learning-Based Algorithm** that seek to predict the eligibility of an individual for loan approval based on the evaluation of certain attributes such as educational level, employment status, loan history amongt others.

Also, precious literature reveals that several empirical research studies pertaining to this study have been conducted[1,2,5,16]. Nevertheless, previous empirical investigations employing alternative algorithms have demonstrated lower levels of accuracy, lower f1 scores, and decreased precision. There is a paucity of prior research examining the utilisation of a comparative analysis including the combination of five distinct machine learning algorithms (Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB) and other performance metrics such as confusion matrix, Precision-Recall curve and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) which

gives a detailed better performance assessment that accuracy in the context of predicting the creditworthiness of borrowers.

This study also aim to determine the variables that effectively reduce the incidence of misclassifying creditworthy individuals as non-creditworthy or non-creditworthy individuals as creditworthy. This paper provides a thorough comparison and analysis of five algorithms: Logistic Regression, Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB.

1.3 Aim and Objectives of the Study

The aim of this study is to develop a model to predict credit risk, using machine learning algorithms to predict the creditworthiness of borrowers and determine their likelihood of defaulting on loans in Nigeria. The specific objectives were to:

i. develop a model for loan defaulting prediction using Logistic Regression, Decision Tree, Gradient Boosting Classifier, Random Forest and Gaussian NB Classifier models
ii. test the models in (i)
iii. evaluate the models' performance using precision, F1-score metric and confusion matrix, Precision-Recall curve and Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

1.4 Methodology Overview

The methodology of this study outlines the approach used to develop a machine learning-based system for predicting loan defaults in Nigeria. It aim to provide a structured framework that guides the research through data collection, processing, model training, and evaluation.

The study adopts a **supervised learning approach**, where models are trained on historical, labeled datasets to predict the likelihood of a loan default. A selection of machine learning algorithms—Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes (NB)—are used to classify loans as "good" or "bad." The choice of these models is based on their diverse capabilities to handle complex data relationships and provide interpretable results.

Data Collection involves gathering datasets that include demographic data, performance data, and previous loan records. This data is sourced from publicly available databases and contains detailed information about borrowers, including their financial histories and demographic characteristics.

Data Preprocessing is conducted to ensure high data quality before analysis. The preprocessing steps include data cleaning to handle missing values, normalization, and data transformation. These steps help to create a consistent dataset for training and testing the models, improving the accuracy of predictions.

The **dataset is split** into training and testing sets, with 70% used for training and 30% for testing. This split allows the models to learn from historical data and then validate their predictive performance on unseen data.

Model training and evaluation utilize performance metrics such as accuracy, precision, recall, F1-score, and the ROC curve to assess each model's effectiveness. The confusion matrix provides insights into how well each model distinguishes between defaulting and non-defaulting loans.

This methodology aims to develop a robust predictive model, offering a systematic approach for improving loan approval processes in the Nigerian financial sector. It emphasizes the importance of data quality, careful model selection, and thorough evaluation to achieve reliable predictions and support data-driven decision-making in lending practices.

1.5 Significance of the Study

Predicting the creditworthiness of borrowers using Machine Learning will play a critical role in the financial sector. This project provides solution and helps to reduce the occurrences of bad debt due to loan defaulting from customers. It will help financial sector especially in Nigeria to predict accurately the credit worthiness of borrowers before granting loans, thereby reducing risk. Additionally, the proposed design will be able to provide detailed information about customers demographic data, performance data and previous loans data. These details can aid management decision-making.

Academically, the study will contribute to the body of knowledge and proffer intelligent solutions to issues relating to loan disbursement. The findings of this study will also serve as a reference for students or academicians or professionals as reference document about the application of machine learning and big data in financial industries; but not limited to Bank loan prediction using machine learning techniques and related work.

1.6 Scope of the Study

The purpose of this research is to predict the credit worthiness of borrowers utilising machine learning algorithms. The selected dataset is an open source from different datasets for both train and test Demographic data, Performance data and Previous loans data. **T**he sample

submission has 2 outcomes- good (1) or bad (0). The evaluation of the design that has been developed will be based on the metrics of precision, recall, and F1-score. The performance of the models on each class will be visualised using the confusion matrix. The accuracy of the models will be used as an evaluation metric. The results will be presented and analysed descriptively.

1.7 Limitation of Study

This study, while comprehensive and insightful, carries certain limitations that could influence the interpretation and applicability of the findings:

i. Despite attempts to address class imbalance, the models still exhibited bias toward the majority class. The techniques used, such as resampling, may not have been sufficient to fully balance the classes or to reflect the complexities of real-world data distributions.

ii. The study focused on a specific set of machine learning models (Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes). While informative, the inclusion of a broader array of models, including deep learning approaches, could provide a more comprehensive view of potential performance across different algorithmic strategies.

iii. The findings were derived from a single, specific dataset, which was segmented into demographic information, performance metrics, and historical borrowing records. Consequently, the results might not be generalizable across different datasets with varying characteristics or in different contexts or industries.

iv. While multiple metrics were used to evaluate model performance, the reliance on conventional metrics like accuracy, precision, recall, and F1-scores might overlook other

important aspects of model behavior, such as interpretability, reliability under varied conditions, and performance on extremely rare events.

1.8 Definition of Operational Terms

Credit Risk: It **r**efers to the financial loss that creditors may incur as a result of debtors failing to meet their credit obligations.

Credit Risk Evaluation: It is a crucial aspect of financial risk management used to make critical decisions regarding whether or not to lend to a counterparty.

False Negative: False negative refers to the acceptance of an applicant who is not credit worthy

False Positive: False positive refers to the rejection of an applicant who is actually credit worthy.

Loan: Loan is essentially an agreement between two parties, the lender and the borrower, whereby the lender grants the borrower credit in the form of cash, property, or other tangible goods if the lender believes that the individual who receives a loan is capable to repay the borrowed funds with interest.

Machine Learning (ML): Machine learning is a field of artificial intelligence (AI) whose aim is to comprehend how data is structured and model it so that it can be utilised by people, facilitates the development of models from available data that will enable decisions to be made based on these data inputs.

Neural Network: Neural Network are computational model utilised in the domains of problem-solving and machine learning. Neural networks (NNs) have been widely applied to

real-world problems in various domains such as business, education, economics, and other areas of life

Supervised Learning: Supervised learning methodology involves providing the computer programme with pre-labelled sample inputs that correspond to the desired outputs where ML algorithms trained through the process of comparing its actual output with the original data in order to identify any errors or inaccuracies.

Unsupervised Learning: Unsupervised learning refers to the utilisation of Artificial Intelligence algorithms for the purpose of identifying and comprehending patterns within datasets that lack any form of labelling or classification.

1.9 Conclusion

This study explored the use of machine learning algorithms to predict loan default and assess creditworthiness in the Nigerian financial sector. By applying models such as Logistic Regression, Decision Trees, Gradient Boosting, Random Forest, and Gaussian NB, the research aimed to improve the accuracy of credit risk assessments and reduce loan default rates. The results demonstrated that machine learning can effectively classify creditworthy and non-creditworthy borrowers, providing a more efficient and data-driven approach to lending decisions.

Although the study faced challenges like class imbalance and dataset limitations, it highlights the potential for machine learning to enhance credit risk management, reducing losses for financial institutions. Future research could build on these findings by incorporating more

diverse datasets and advanced algorithms, further strengthening the decision-making process in the lending industry.

Chapter Two provides a **literature review** that examines the existing research and foundational concepts relevant to using machine learning for predicting loan defaults. It explores the theoretical underpinnings of loan risk assessment, machine learning methodologies, and data preprocessing techniques, laying the groundwork for the study's approach.

The chapter begins by **defining loan default** and discussing its implications in the financial sector, particularly for lending institutions. It emphasizes the critical need for accurate prediction models to minimize risk and improve decision-making processes in loan approvals. This section sets the context for why machine learning has become a valuable tool for addressing the complexities of credit risk assessment.

The **application of machine learning in credit risk assessment** is then explored, covering various algorithms that have been applied to predict loan defaults. The review highlights popular classification models such as Decision Trees, Random Forests, Gradient Boosting, and Gaussian Naive Bayes, emphasizing their strengths, limitations, and suitability for different types of datasets. This provides a comprehensive understanding of why these models were selected for the study.

In addition to discussing specific algorithms, the chapter delves into **data preprocessing techniques**, such as data cleaning, handling missing values, feature engineering, and data transformation. It underscores the importance of these steps in ensuring high-quality input

data, which is essential for building accurate predictive models. The section also addresses the challenges of class imbalance in datasets, a common issue in loan default prediction, and reviews techniques like oversampling and undersampling to mitigate this problem.

The chapter further reviews **previous studies** that have applied machine learning to predict loan defaults, comparing their methodologies, datasets, and findings. It highlights key studies that have influenced the current research, such as those using clustering and classification methods, ensemble models, and hybrid approaches. This comparative analysis allows the study to position itself within the existing body of work and identify areas for further exploration.

Additionally, the chapter addresses **evaluation metrics** commonly used in the domain, such as precision, recall, F1-score, and accuracy, as well as the importance of using tools like confusion matrices and ROC curves. This section highlights the need for a balanced approach in evaluating models, especially when dealing with imbalanced classes, to ensure both accuracy and fairness in predictions.

Chapter Two concludes by identifying **gaps in the literature** and presenting the **research rationale** for the study. It points out the limitations in existing approaches, such as the need for improved handling of class imbalance and a deeper understanding of model behavior on minority classes. These insights justify the study's focus on evaluating multiple machine learning models and using advanced techniques to enhance the prediction of loan defaults.

Overall, Chapter Two provides a comprehensive review of the theoretical and empirical foundations of the study, offering a clear perspective on the current state of research in loan default prediction and setting the stage for the methodology outlined in Chapter Three.

Chapter Three focuses on the **methodology** used in the study, outlining the approach, tools, and processes for building and evaluating a machine learning-based model to predict loan defaults. It describes the systematic steps taken to ensure the accuracy, reliability, and applicability of the predictive models developed during the research.

The **research approach** involves a supervised learning method, where various classification algorithms, including Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes, are applied to a labeled dataset. This approach is designed to assess the predictive capabilities of each model in distinguishing between borrowers who will likely default on their loans and those who will not.

The chapter also details the **requirement specifications**, including the hardware and software needs. The study utilizes a personal computer with 8GB RAM and a 2.2 GHz Intel Core i3 processor. The software environment includes tools like Jupyter Notebook, Scikit-Learn, Pandas, and Matplotlib, providing the necessary infrastructure for data manipulation, visualization, and model building.

The **research design** is structured around the collection, preprocessing, and analysis of data. The dataset, divided into training and testing subsets, includes demographic information, performance data, and records of previous loans. The preprocessing steps involve cleaning the data, handling missing values through imputation, and converting categorical variables

into a numerical format. These steps are crucial for preparing a high-quality dataset suitable for training the machine learning models.

Data analysis techniques like correlation analysis and data splitting are used to understand the relationships between different variables and to prepare the data for model training. The correlation analysis uses visual tools such as heat maps to identify the strength and direction of relationships between features, while the dataset is split into training (70%) and testing (30%) sets to evaluate the models effectively.

The chapter outlines the **algorithms used for model building**, providing a rationale for choosing each. The study examines the strengths and weaknesses of each model, aiming to identify the one best suited for loan default prediction. It leverages ensemble methods like Random Forest and Gradient Boosting to improve predictive accuracy, while the Decision Tree and Gaussian Naive Bayes models offer insights into simpler yet effective classification methods.

Lastly, the **model evaluation** section describes how the models' performance is assessed using metrics such as precision, recall, F1-score, and accuracy. Tools like confusion matrices and ROC curves are employed to visualize and interpret the models' performance on both the majority and minority classes, ensuring a comprehensive analysis of their predictive power.

Overall, Chapter Three provides a thorough explanation of the study's methodology, emphasizing the importance of a structured approach to data preprocessing, model selection, and evaluation to achieve accurate and reliable predictions in loan default analysis.

Chapter Four focuses on the **presentation, analysis, and discussion of the research findings** related to predicting loan defaults using a machine learning-based approach. This chapter outlines the performance of various classification models, providing insights into their effectiveness in assessing the likelihood of loan defaults based on the prepared dataset.

The chapter begins by describing the **data processing and preparation** undertaken before model training. It highlights key steps such as data cleaning, handling missing values, feature engineering, and data transformation. These processes ensured the dataset was of high quality, which is essential for accurate modeling. The pre-processed dataset is divided into training and testing sets, allowing the models to be trained effectively and tested for their predictive accuracy.

Next, the chapter discusses the **evaluation of machine learning models** used in the study, including the Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes classifiers. It presents the results of each model in terms of performance metrics like accuracy, precision, recall, F1-score, and confusion matrices. These metrics help in understanding the strengths and limitations of each model, especially regarding their ability to predict loan defaults accurately.

A major focus of the analysis is on **model comparison**, where the performance of different classifiers is compared to determine which model best predicts loan defaults. For instance, while the Decision Tree model showed moderate accuracy, it struggled with classifying minority cases accurately. The Random Forest and Gradient Boosting models performed better, demonstrating robustness with higher accuracy scores, but still faced challenges with

class imbalance. The Gaussian Naive Bayes model achieved the highest accuracy, suggesting that its assumptions align well with the dataset's characteristics.

The chapter also explores the **issue of class imbalance** encountered during the analysis, which refers to the unequal distribution of classes (e.g., 'good' vs. 'bad' loan outcomes). This imbalance affected the models' ability to generalize across different types of loan cases. The discussion highlights the impact of this imbalance on predictive accuracy, with models generally favoring the majority class (more frequent outcomes). This observation underscores the need for techniques like oversampling, undersampling, or cost-sensitive learning, which could improve the prediction of minority class outcomes.

Furthermore, Chapter Four addresses the **implications of the findings** for real-world loan default prediction. It discusses the strengths and potential risks associated with deploying the models in practical scenarios. For example, while the models showed strong predictive power for majority cases, their weaker performance on minority cases could lead to issues like underestimating the risk of certain types of loan defaults. This insight is critical for stakeholders looking to apply these models in financial institutions, as it can inform strategies for risk management and credit assessment.

The chapter concludes by **summarizing the key insights** derived from the model evaluation, emphasizing the relative strengths and limitations of each machine learning model in predicting loan defaults. It provides a foundation for the recommendations in Chapter Five, particularly in relation to improving model accuracy and handling class imbalance. Overall, Chapter Four serves as a detailed analysis of the study's results, offering valuable information for both the research community and practitioners in the field of financial risk management.

Chapter Five of this study presents the **conclusion** of the research, summarizing the findings, offering practical recommendations, and highlighting the study's contributions to knowledge. It also identifies potential **areas for further research** to improve the application of machine learning models in loan default prediction.

The **summary of findings** outlines the effectiveness of the machine learning-based approach in predicting loan defaults. The study utilized Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes models on a thoroughly preprocessed dataset. The findings showed varying levels of accuracy and highlighted the challenge of class imbalance, which influenced model performance, particularly in distinguishing between defaulting and non-defaulting loans.

In the **conclusion**, the study emphasizes the importance of accurate data preprocessing and the careful selection of machine learning algorithms to improve predictive capabilities. It acknowledges the strengths and limitations of each model, noting that Gaussian Naive Bayes achieved the highest accuracy. However, all models struggled with class imbalance, which remains a critical area for improvement.

The **recommendations** focus on practical strategies to enhance model performance, such as implementing resampling methods like SMOTE, using advanced evaluation metrics, and exploring more sophisticated ensemble techniques. The study also advises the development of cost-sensitive models and regular monitoring of the model's performance once deployed, ensuring it adapts to new data patterns over time.

In terms of **contributions to knowledge**, the study underscores the importance of high data integrity and the role of data preprocessing in successful predictive modeling. It highlights the challenges posed by class imbalance, offering a foundation for further exploration of solutions like resampling techniques. Additionally, the research contributes to the methodology of model evaluation, advocating for comprehensive metrics beyond standard accuracy.

Finally, **suggested areas for further study** include investigating advanced methods for addressing class imbalance, such as Generative Adversarial Networks (GANs) and hybrid models. The study also suggests exploring the use of machine learning models across different domains to evaluate their generalizability and effectiveness.

Overall, Chapter Five synthesizes the study's key insights, emphasizing its practical applications and providing guidance for future research, thus contributing valuable perspectives to the field of predictive analytics in loan default prediction.

Endnotes

1. O.B Alaba, E.O Taiwo & O.A Abass. *Data mining algorithm for development of a predictive model for mitigating loan risk in Nigerian banks*. **Journal of Applied Sciences and Environmental Management**. Dec 28;25(9): 2021 1613-6

2. M.C Aniceto, F Barboza & H Kimura. *Machine learning predictivity applied to consumer creditworthiness*. **Future Business Journal**. Dec;6(1): 2020 1-4.

3. F Isa & R Isa. *Treatment of toxic asset by deposit money banks in Nigeria: A review of literature*. **TSU-International Journal of Accounting and Finance**. Dec 15;1(1): 2021 42-50.

4. A.A Egwa, H Bello, A.A Ahmad & M.S Bizi. *Default prediction for loan lenders using machine learning algorithms*. **SLU Journal of Science and Technology**. Dec 29;5(1&2): 2022 1-2.

5. S Moradi & F.M Rafiei. *A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks*. Financial Innovation. Dec;5(1): 2019 1-27.

6. A.I Ahmed & P.R Rajaleximi. *An empirical study on credit scoring and credit scorecard for financialinstitutions*. **International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)**. Jul;8(7): 2019 2278-1323.

7. M Nilsson & Q Shan. *Credit risk analysis with machine learning techniques in peer-to-peer lending market*.Stockholm Business School Master's Degree Thesis Master's Programme in Banking and Finance.2018

8. V.B Djeundje & J Crook. *Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards*. **European Journal of Operational Research**. Dec 1;271(2): 2018 697-709.

9. Y Guo, J He, L Xu & W Liu. *A novel multi-objective particle swarm optimization for comprehensible credit scoring*. Soft Computing. Sep 1;23: 2019 9009-23.

10. H.A Alaka, L.O Oyedele, H.A Owolabi, V Kumar, S.O Ajayi, O.O Akinade & M Bilal. *Systematic review of bankruptcy prediction models: Towards a framework for tool selection.* Expert Systems with Applications. Mar 15;94: 2018 164-84.

11. N Gulsoy & S Kulluk. *A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. May;9(3): 2019 e1299.

12. S Carta, A Ferreira, D.R Recupero, M Saia, & R Saia. *A combined entropy-based approach for a proactive credit scoring*. Engineering Applications of Artificial Intelligence. Jan 1;87: 2020 103292.

13. C Jiang, Z Wang. & H Zhoo. *A Prediction-driven Mixture cure model and its Application in Credit Scoring*. **European Journal of operational Research**. 277, pp20-31. 2019

14. C Begum. & U Deniz. *Comparison of Data Mining Classification Algorithms: Determining the Default Risk*. Research Article, Hindawi Scientific Programming Volume 2019

15. A.E Awuza, K.A Habeebah, A.A Ahmad, M.B Abubakar & A.M Muhammad. *Prediction Model for Loan Default Using Machine Learning*. **The International Journal Of Science & Technoledge**. DOI No.: 10.24940/theijst/2022/v10/i2/ST2202-009.2022

16. J Alzubi, A Nayyar & A Kumar. *Machine learning from theory to algorithms: an overview*. **InJournal of physics: conference series** 2018 Nov (Vol. 1142, p. 012012). IOP Publishing.

17. K.R Varshney. *Trustworthy machine learning and artificial intelligence*. XRDS: Crossroads, The ACM Magazine for Students. Apr 10;25(3): 2019 26-9.

18. I.H Sarker. *Machine learning: Algorithms, real-world applications and research directions*. SN computer science. May;2(3): 2021 160.

19. A.A Jamali, R Ferdousi, S Razzaghi, J Li, R Safdari & E Ebrahimie. *DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins*. Drug discovery today. May 1;21(5): 2016 718-24.

20. A.S Heinsfeld, A.R Franco, R.C Craddock, A Buchweitz & F Meneguzzi. *Identification of autism spectrum disorder using deep learning and the ABIDE dataset*. NeuroImage: Clinical. Jan 1;17: 2018 16-23

21. J Hagenauer & M Helbich. *A comparative study of machine learning classifiers for modeling travel mode choice*. Expert Systems with Applications. Jul 15;78: 2017 273-82.

22. Dan Shewan, Companies using machine learning in cool ways. November 14, 2023

Chapter Two

Literature Review

2.1 Conceptual Review

2.1.1 Credit Worthiness

Creditworthiness refers to the assessment of an individual's or entity's ability to fulfill their financial obligations, particularly those related to borrowing and credit transactions[1]. It is a crucial factor that lenders, such as banks, financial institutions, and credit card companies, consider when evaluating whether to extend credit to a borrower. A borrower's creditworthiness helps lenders determine the level of risk associated with lending money and whether the borrower is likely to repay the borrowed funds as agreed. Several factors contribute to determining the creditworthiness of borrowers:

Credit History: A borrower's credit history is a record of their past borrowing behavior, including credit cards, loans, and repayment patterns. A positive credit history with timely payments and responsible credit management enhances creditworthiness[2].

Credit Score: A credit score is a numerical representation of a borrower's creditworthiness. It is calculated based on various factors such as payment history, credit utilization, length of credit history, types of credit, and new credit inquiries. Higher credit scores indicate lower credit risk[3].

Income and Employment Stability: Lenders assess the borrower's income level and stability of employment to ensure they have the means to repay the borrowed funds. A stable and sufficient income increases creditworthiness.

Debt-to-Income Ratio: This ratio compares the borrower's total debt payments to their total income. A lower ratio indicates that the borrower has a manageable level of debt and is more likely to meet their repayment obligations[4].

Current Financial Obligations: Lenders consider the borrower's existing debts and financial commitments to determine their ability to take on additional credit.

Assets and Collateral: Borrowers with valuable assets or collateral that can be used to secure the loan may have higher creditworthiness, as it provides a safety net for lenders in case of default.

Payment Behavior: Regular payments of bills and obligations, including rent and utility bills, demonstrate responsible financial behavior and enhance creditworthiness.

Length of Credit History: A longer credit history provides lenders with more information about a borrower's financial habits and behaviors.

Public Records: Negative public records, such as bankruptcies, foreclosures, or tax liens, can significantly impact creditworthiness.

Credit Utilization: This ratio compares the borrower's credit card balances to their credit limits. A lower ratio indicates responsible credit management.

Lenders typically use the information collected from credit reports, credit scores, and application forms to assess the creditworthiness of borrowers. Based on this assessment, they make decisions on the terms of the loan, such as interest rates, loan amounts, and repayment schedules. Maintaining good creditworthiness is important for borrowers, as it allows them to access credit at favorable terms and lower interest rates. It's essential for individuals and businesses to be aware of their creditworthiness and take steps to improve it if needed, as it can significantly impact their financial opportunities and choices.

In terms of both macroeconomic factors and systemic risk, consumer spending is a key factor. Therefore, the analysis of consumer credit is pertinent, as individuals may obtain loans to satisfy their consumption needs[1]. It is expected that the transaction value of the Marketplace Lending (Consumer) segment will reach $78.57 million in the year 2023 and is anticipated to grow at a 5.29% Compound Annual Growth rate (CAGR) over the period 2023-2027, resulting in a total of US$96.57m by the year 2027[5]. In 2023, the average transaction value per Marketplace Lending (Consumer) user is anticipated to reach $48.75m. From a global comparison standpoint, the United States will reach the maximum transaction value (US$26,180,000,000) in 2023[5].

The Nigerian credit market is predominantly governed by the Central Bank of Nigeria (CBN), the apex regulatory body in the banking system and is therefore responsible for the DMBs[6]. Obviously, there are credit lenders that are not governed or supervised by the CBN. These include the Primary Mortgage Institutions, which report to the Federal Mortgage Bank, and

the leasing corporations that operate under the Equipment Leasing Association of Nigeria's self-regulatory body[7].

Credit risk evaluation is a crucial aspect of financial risk management because banks must make critical decisions regarding whether or not to lend to a counterparty. The most significant issue in finance is predicting bankruptcy or default[8]. In consumer lending, the large number of potential consumers necessitates using models and algorithms that minimize or eradicate errors caused by human actions in analysing credit applications[1]. Several largest global institutions have developed advanced automated systems for modelling credit risk, providing decision-makers with crucial data. Before extending credit to borrowers, it is essential to ascertain their creditworthiness, as payment defaults can be extremely risky. Due to the dearth of appropriate data for machine learning techniques, some financial institutions continue to rely on the traditional strategy[9].

2.1.2 Loan

A loan can be defined as a contractual agreement between two entities, namely the lender and the borrower[1]. In this agreement, the lender provides the borrower with a certain amount of funds, assets, or other physical resources, with the understanding that the borrower would repay the loan together with accrued interest within a predetermined timeframe, contingent upon the lender's confidence in the borrower's capacity to fulfill their repayment obligations[1]. In the context of personal loans, the absence of specific criteria poses a challenge for lenders in assessing the borrower's ability to return the loan amount along with the accrued interest within the designated timeframe[10,11]. In contemporary times, the primary focus of nearly all

financial institutions revolves around the process of evaluating and disbursing loan applications. The revenue generated from loans disbursed by a financial institution is a substantial portion of its total assets.

The acquisition of loans for various purposes, such as home loans, education loans, vehicle loans, and business loans, has become a commonplace occurrence in our daily lives. These loans are typically obtained via financial institutions such as credit unions and banks. Nevertheless, a significant portion of individuals face challenges in accurately assessing their capacity to repay the entirety of their financial obligations. The assessment of creditworthiness holds significant importance for banks and other financial institutions in order to sustain operations within the fiercely competitive market and ensure profitability. It is imperative to establish explicit and well-defined criteria for the provision of loans. The aforementioned criteria should be deemed satisfactory and appropriate in order to furnish the necessary details pertaining to the credit structure, borrowers, and payment method. Financial organizations, such as banks and microfinance institutions, are inundated with a substantial volume of credit applications on a daily basis. It is important to note that not all loan applicants will gain approval from these institutions. The majority of banks employ their own credit scoring algorithms and risk assessment procedures when evaluating loan applications in order to determine whether to approve or reject them.

The main objective within a banking context is to allocate assets to reliable entities, ensuring a high probability of generating a guaranteed return. Numerous financial institutions and banks have implemented a stringent loan approval procedure; nonetheless, there is no

assurance that the selected applicant is genuinely deserving of the loan or possesses the lowest risk of loan default and full repayment by the designated due date. A loan is essentially an agreement between two parties, the lender and the borrower, whereby the lender grants the borrower credit in the form of cash, property, or other tangible goods if the lender believes that the individual who receives a loan is capable of repaying the borrowed funds with interest[8]. Almost every bank's primary focus today is on approving and distributing debts. A significant component of a bank's assets comprises profits generated from disbursed loans.

Loan Prediction is a highly valuable tool for both employees of financial institutions, like banks and money lending organizations, and borrowers who are in the process of submitting loan applications. In the lending/banking business, the assessment of borrower risk is of paramount importance, prompting the following key inquiry: a) Does the borrower possess a high-risk profile? b) Considering the borrower's risk profile, should the bank extend a loan to them?[11].

The interest rate of the borrower is contingent upon the response to the initial inquiry. The assessment of the borrower's level of risk is quantified through the interest rate, which is influenced by various factors, including the concept of the time value of money[12,13]. There exists a positive correlation between the interest rate and the level of risk associated with the borrower. The bank can thereafter determine the eligibility of the applicant for the loan by considering the interest rate.

Borrowers obtain loans from investors, who act as lenders, in return for the promise of repayment with interest. In this context, it can be observed that the lender's financial gain,

specifically in the form of interest, is contingent upon the borrower fulfilling their obligation to return the loan. In contrast, the lender incurs financial losses in the event that the borrower fails to fulfill the repayment obligations associated with the loan.

Loans provided by banks have emerged as a significant external financing option for both enterprises and families, mostly driven by the financial limitations faced by both entities in their pursuit of company development and expansion. Lending to the economy is a very advantageous activity for commercial banks, as loans constitute a significant portion of their assets[13]. Nevertheless, the escalation in loan disbursement is connected to various potential hazards, including the risk of default or credit risk. This risk pertains to the borrower's incapacity to fulfill the loan repayment obligations within the predetermined timeframe and under the agreed-upon terms and conditions.

In the event that the debtor fulfills their obligation by repaying the loan, the creditor would realize a financial gain derived from the borrowed funds. Nevertheless, in the event that the debtor is unable to fulfill their obligation to repay the borrowed funds, the creditor incurs a loss in terms of both their financial interest and the capital invested. Hence, creditors are confronted with the challenge of predicting the likelihood of a debtor's inability to fulfill loan repayment obligations.

Credit risk is widely recognized as one of the primary factors contributing to financial instability. Commercial banks view lending as a means of generating profit, but they also recognize the inherent risks involved. To mitigate the risk of default, commercial banks employ two key strategies: evaluating the borrower's ability to repay the loan and requiring

collateral as a condition for loan approval[14]. This operation is accomplished by the utilization of extensively skilled personnel inside commercial banking institutions, who assess the eligibility of loan applicants by scrutinizing several factors to determine their suitability for loan acceptance or denial, resulting in the assignment of a numerical score.

In recent times, advancements in technology have led to the creation of machine learning algorithms and neural networks that can autonomously forecast an individual's credit score by analyzing their historical data[15]. This enables the identification of potential credit defaulters prior to loan approval. The insolvency of banks has been attributed to the accrual of delinquent debt stemming from poor creditworthiness. The borrower's refusal to repay funds intended for the benefit of depositors and to fulfill their financial obligations is the underlying cause.

Subsequently, depositors initiate the process of withdrawing their funds from the bank, so causing a depletion of cash reserves and consequently resulting in financial losses for the bank. In the event of a comparable occurrence transpiring across all financial institutions, the overall economy would experience significant repercussions. Key personnel, such as the management, are unjustly terminated as a consequence of loan defaults.

The responsibility of approving loans typically rests with the manager. Due to the manager's inability to accurately forecast the loan's potential for repayment, they assume full responsibility for the consequences of their decision. Consequently, banks sometimes resort to staff layoffs as a means to alleviate their financial burdens. By leveraging the acquired information and data, it becomes possible to identify and analyze trends, such as the age

group that exhibits a lower propensity to repay debts. The banking industry has amassed a significant amount of data mostly derived from consumer interactions, business financial statements, and payment records, among other sources. The utilization and analysis of this data can be instrumental in obtaining a competitive edge and forecasting the creditworthiness of customers, therefore mitigating the occurrence of precarious transactions[16,17].

The utilization of data mining plays a crucial role in achieving these objectives. Data mining is a continuous process that integrates business intelligence, machine learning techniques, and tools with large amounts of accurate and relevant data to facilitate the discovery of non-obvious insights hidden within an organization's corporate data[18]. The data derived from the observed trends can assist an organization in developing novel tactics aimed at enhancing the firm's interactions with both its consumers and employees. Data mining approaches can be categorized into two primary groups: statistical methods and artificial intelligence[19].

The logistic regression and discriminant analysis are the prevailing statistical methods employed for loan prediction[20]. The logistic regression model posits that the fitted likelihood of an event is a linear function of the observed variables of the explanatory factors[20]. One of the advantages associated with statistical approaches is their ease of execution and their ability to generate easily interpretable outputs[21,22].

2.1.3 Artificial Intelligence in Loan Prediction

Artificial Intelligence (AI) is revolutionizing the way financial institutions assess loan applications and predict the creditworthiness of borrowers. Artificial intelligence techniques

encompass a range of methodologies, such as K-nearest neighbor, Decision Trees, Neural Networks, Naïve Bayesian classifier, Genetic programming, and Support vector machine models[23].The integration of AI technologies in loan prediction processes has led to more accurate, efficient, and data-driven lending decisions. AI algorithms can analyze vast amounts of data from various sources, including credit history, income, employment, transaction history, social media activity, and more[23]. This comprehensive analysis provides a more holistic view of the borrower's financial situation. Also, machine learning models, such as logistic regression, decision trees, random forests, and gradient boosting, can learn from historical loan data to identify patterns and factors that contribute to loan default or repayment[24]. These models continuously improve their accuracy as they learn from new data. AI-driven credit scoring models use advanced analytics to assign credit scores based on a wide range of factors[24]. These models can incorporate more variables and update scores in real-time as new data becomes available.

AI can assess risk more accurately by considering both quantitative and qualitative factors. This helps lenders make more informed decisions and offer competitive interest rates to lower-risk borrowers. It can also automates the underwriting process by analyzing application data, verifying information, and determining whether to approve or decline a loan[23]. This reduces the manual workload for underwriters and speeds up the decision-making process.

AI-driven systems can generate personalized loan offers based on the borrower's profile, needs, and preferences[25]. This enhances the customer experience and increases the likelihood

of loan acceptance. Similarly, AI algorithms can identify loanulent loan applications by detecting anomalies and patterns that indicate potential loan. This helps minimize the risk of lending to loanulent borrowers. After loan approval, AI can monitor borrower behavior and financial status in real-time. This allows lenders to detect early signs of financial distress and proactively offer support or modifications to the loan terms[26].

Advanced AI models can provide insights into the factors that influenced a loan decision[26]. This transparency helps borrowers understand why their application was approved or denied and promotes fairness in lending. AI-powered analytics assist lenders in managing their loan portfolios by predicting potential defaults, identifying high-risk accounts, and suggesting strategies to mitigate risk[23,24]. AI systems can ensure that lending practices comply with regulations, such as fair lending laws and anti-discrimination policies, by identifying and addressing any biases in decision-making.

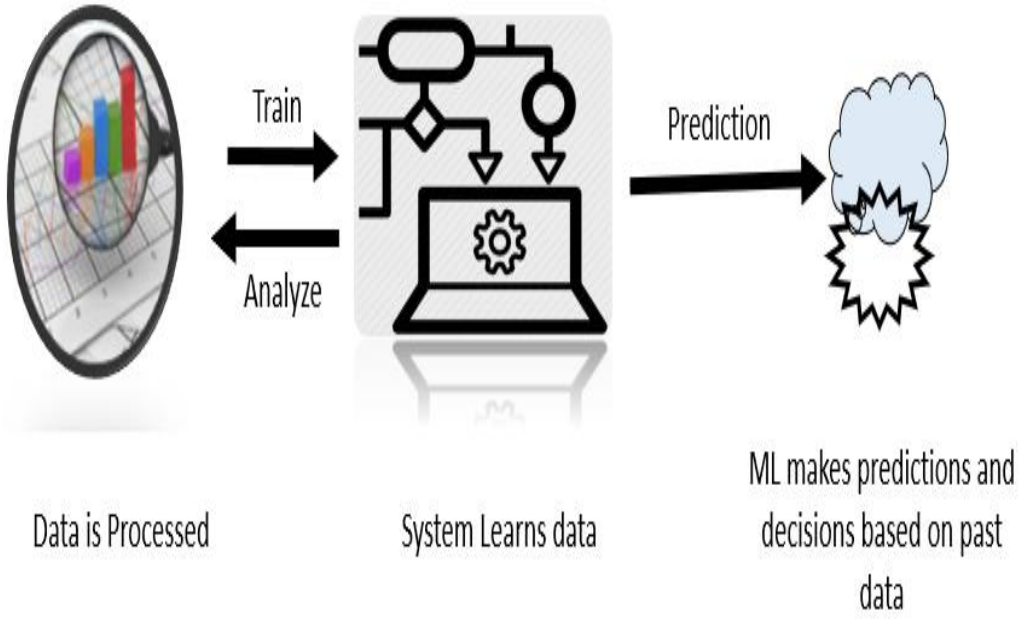
By leveraging AI in loan prediction, financial institutions can make more accurate lending decisions, reduce default rates, improve customer experiences, and streamline their lending processes. However, it's important to carefully design AI models to avoid biases and ensure that ethical considerations are prioritized throughout the loan prediction process**.**

The utilization of AI techniques is prevalent in situations where the relationship between dependent and independent variables is complex and exhibits non-linear characteristic. The issue of loan default prediction holds significant importance for lending institutions such as banks and other financial organizations, as it exerts a substantial impact on their profitability

and growth. Despite the existence of numerous conventional approaches to gather information pertaining to loan applications, it appears that a majority of these methods are exhibiting subpar performance, as seen by reported escalations in the prevalence of non-performing loans. Over the course of time, machine learning techniques have been employed to assess and forecast credit risk through the analysis of an individual's past data.

2.1.4 Machine Learning

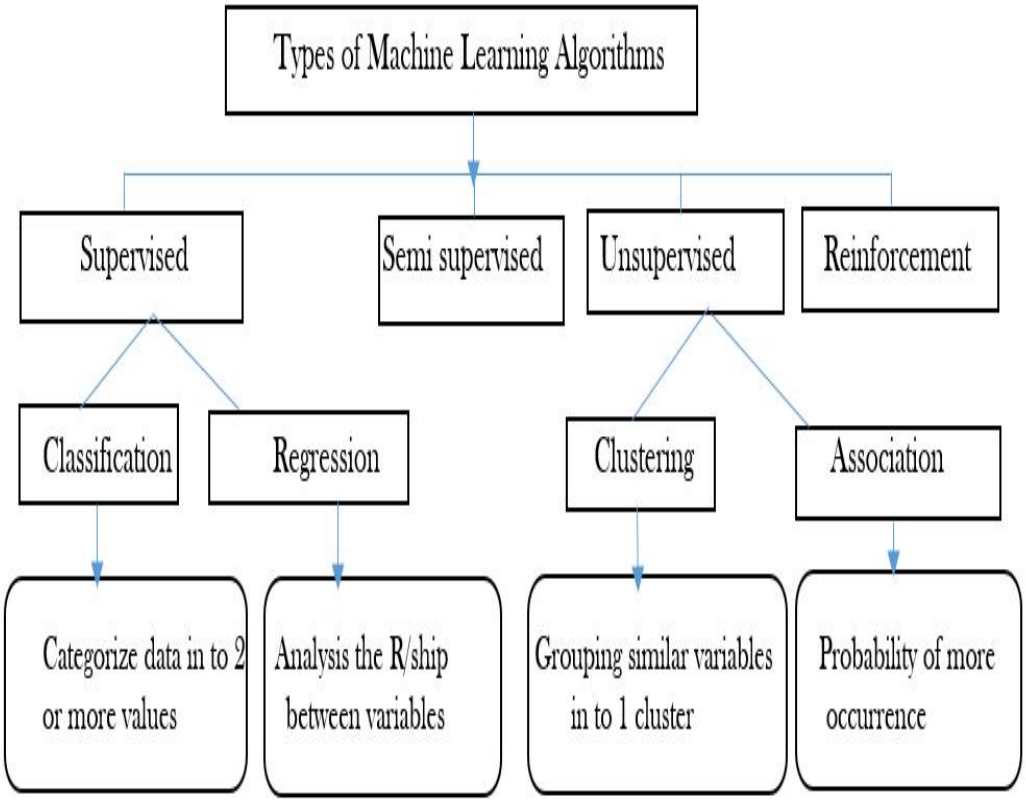
Machine learning is a subfield within the realm of Artificial Intelligence (AI) and computer science, which centers on the use of data and algorithms to replicate human perception and enhance its precision through iterative processes[27,28,29]. Machine learning (ML) is a field of study focused on enabling computers to acquire and process data and information in a manner that allows them to learn and perform tasks without the need for explicit human-like programming[29]. Machine learning employs several algorithms based on the characteristics of the problem and the data at hand. The algorithm is trained using the datasets, enabling the system to learn from the patterns within the data. Consequently, the algorithm is capable of making predictions when fresh data is introduced to the system. This suggests that machine learning has the capability to acquire knowledge from past data and afterwards utilize that knowledge to make decisions on new input data.



Machine Learning Working Process[30].

Machine learning algorithms enable autonomous decision-making in systems without the need for external assistance. The identification of valuable underlying patterns within complex data is the basis for making such decisions[30].

Machine learning can be broadly categorized into two main branches: descriptive analytics and predictive analytics[31]. Descriptive analytics serve the purpose of elucidating the fundamental characteristics and attributes of the data, whereas predictive analytics are employed to get insights into future events or outcomes[31]. Machine learning algorithms are classified into many categories based on factors such as learning capabilities, kind of input and output data, and issue behavior. These categories include supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning[32,33,34].



Types of Machine Learning Algorithms[35].

2.1.4.1 Supervised Learning Algorithm

Supervised learning is a machine learning paradigm employed when data is characterized by input variables or qualities, which are utilized to predict or generate a target or output value[36]. The algorithm accomplishes this by acquiring knowledge from the input data and making projections on the output value. Data pre-processing in supervised learning encompasses several procedures, including but not limited to data cleaning, normalization, transformation, feature extraction, and selection[37]. The ultimate training set is obtained through the process of data pre-processing. Supervised learning can be broadly categorized into two main types: classification and regression[38].

Classification: Classification is a fundamental machine learning technique that involves the utilization of labeled data to train a model, enabling it to categorize new data instances into several predefined categories or classes[39]. The primary objective of classification is to effectively and properly categorize the target class based on the provided input data[39]. The most basic form of classification is binary classification. Several commonly used classification techniques include logistic regression, several forms of decision trees, gradient boosting machines, Naïve Bayes, random forest, Support Vector Machines (SVM), multi-layer perceptron, and K-Nearest Neighbor[40,41].

Regression: Regression is a specific type of supervised learning that focuses on predicting continuous numerical values. In regression, the goal is to establish a relationship between input variables (also known as features) and the output variable, allowing the algorithm to make accurate predictions for new data points[42,43].

Regression algorithms are widely used in various fields, including finance, economics, healthcare, and more, for tasks such as predicting stock prices, estimating sales revenue, and forecasting medical outcomes. In regression, the target variable is continuous and numerica[42].

There are various types of regression techniques, each suited to different types of data and scenarios. Some common types of regression include linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression[42]. During the training phase, the algorithm learns the relationship between the input features and the target variable by fitting a regression function. The goal is to find the best-fitting

function that minimizes the difference between the predicted and actual target values. Regression is a fundamental technique in supervised learning that focuses on predicting continuous numerical values based on the relationships between input features and the target variable. It plays a crucial role in data analysis and decision-making across multiple industries and domains.

Decision Tree

Decision Trees (DT) are a tree-like structure that sorts instances based on feature values to classify them. Each branch of a decision tree indicates a value that the node can accept, and each node represents a feature in an instance to be classified[44]. In a group of observations that make up a data set, decision trees try to establish a strong association between input values and goal values. When a set of input values is found to have a strong link to a target value, all of these values are grouped into a bin, which creates a decision tree branch. It begins with a single root node that divides into several branches, each of which leads to other nodes, each of which can split further or terminate as a leaf node[44].

Decision Trees have several advantages:

Interpretability: Decision Trees are easy to visualize and understand, making them a useful tool for explaining how decisions are made by the model[44].

Non-linearity: Decision Trees can capture non-linear relationships in data without requiring complex mathematical transformations[44].

Handling Missing Values: They can handle missing values in the dataset by using surrogate splits.

However, they also have some limitations:

Overfitting: Deep Decision Trees can overfit the training data, leading to poor generalization on unseen data. This can be mitigated by pruning or using ensemble methods like Random Forests[45].

Instability: Small changes in the data can lead to significantly different trees, making them unstable.

Bias towards Dominant Classes: In classification tasks, Decision Trees can be biased towards dominant classes if not adjusted properly. To address some of these limitations, ensemble methods like Random Forests and Gradient Boosting were developed, which combine multiple Decision Trees to improve performance and generalization[45].

There are key equations and concepts involved in the decision-making process within a Decision Tree.

Information Gain: Information gain measures the reduction in entropy (uncertainty) achieved by splitting the data based on a particular feature[44].

$$\text{Information Gain} = \text{Entropy (S)} - \sum_{V \in (values(A)} \frac{|S_v|}{S} \cdot Entropy\ (S_v) \qquad 2.1$$

Gini Impurity: The Gini impurity measures the degree of impurity in a set of samples. It ranges from 0 (pure set, all samples belong to one class) to 0.5 (maximally impure set, samples are evenly distributed across classes).

$$\text{Gini (P)} = 1 - \sum_{i=1}^{C} P_i^2 \qquad 2.2$$

Logistic Regression

Logistic Regression is a statistical method used for modeling the relationship between a binary outcome variable (dependent variable) and one or more predictor variables (independent variables)[46,47].

Logistic Regression model is a Machine Learning classification method (algorithm) that is used to forecast or predict the probability of a categorical dependent factor[47]. In a logistic regression model, the dependent variable is a binary that contains data coded as 1 (yes, etc.) or 0 (no, etc.)[48]. In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic Regression is one among most popular useful models for categorical data, especially for binary response data in data modelling. Logistic regression models can directly predict probabilities (values that are restricted to the (0,1) interval); furthermore, these probabilities are well-calibrated in comparison to the possibilities predicted by other classifier models, like Naive Bayes[47]. Logistic regression preserves the marginal probabilities of the training data. The multiplier of the model also gives some hints about the relative importance of every input variable.

The equation for Logistic Regression can be described as follows:

Logit Transformation: The logit transformation is used to model the linear relationship between predictor variables and the log-odds of the outcome being in the positive class[49].

consider:

p as the probability of the event (positive outcome).

1 - p as the probability of the non-event (negative outcome).

The logit transformation is the natural logarithm of the odds ratio:

$$\text{logit}(p) - \log\left(\frac{p}{1-p}\right) \quad 2.3$$

Linear Combination of Predictor Variables: In the logistic regression model, the logit of the probability is assumed to be a linear combination of predictor variables[47]:

$$\text{logit}(p) = \beta_0 + \beta_{1x1} + \beta_{2x2} + \ldots\ldots + \beta_{nxn} \quad 2.4$$

where x_1, x_2, ……x_n are the predictors variable
β_0, β_1,…..,β_n are the coefficients associated with the predictor variables
2. Sigmoid (Logistic) Function: The logit values are then transformed back to probabilities using the sigmoid (logistic) function[50]:

$$p = \frac{1}{1+e^{-logit(p)}} \quad 2.5$$

The sigmoid function $\frac{1}{1+e^{-z}}$ takes a linear combination of the features and transforms it into a value between 0 and 1, which represents the probability of the positive class (in binary classification) or the event of interest.

Random Forest

Random Forest (RF) is an additional supervised machine learning technique that may be employed for both classification and regression tasks. Random Forest (RF) generates predictions by aggregating the outputs of several decision trees, as its name suggests. A random tree refers to a tree that is constructed randomly from a given set of trees, where each tree in the set possesses K random properties at every node[51]. Consequently, every tree possesses an equivalent likelihood of being selected for sampling. The whole RF structure was depicted in Figure 2.1 below.

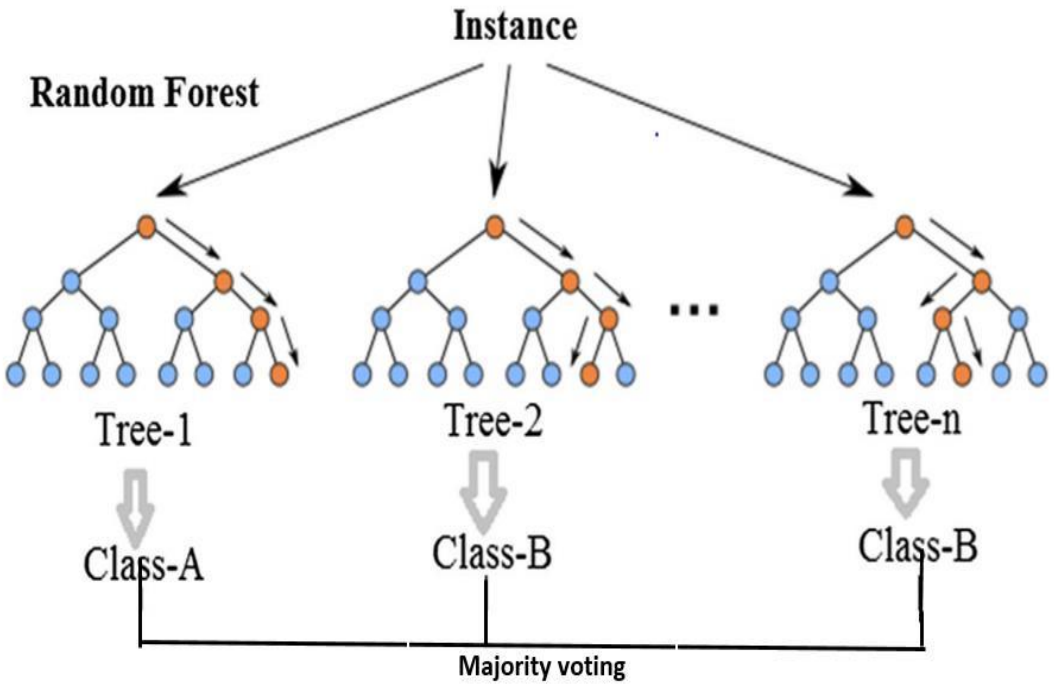


Figure 2.3: General RF Algorithm[52].

In Random Forest, we have a collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class[53]. The forest chooses the classification having the most votes. Each tree is planted & grown as follows If the number of cases in the training dataset is P, then a sample of P cases is taken at random but with replacement. As shown in figure 2.3, the tree consists of one root node, several internal and leaf nodes. The internal node consists several leaf node and this leaf node correspond to decision result. The final model of the random forest is decided by the majority votes produced by all individual decision trees[54].

If there are N input features, a number n<<N is specified such that at each node, n features are selected at random out of the P and the best split on these m is used to split the node[53]. The value of n is held constant during the forest growing

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \quad 2.6$$

Where N is the number of data points in from given dataset, f_i is the value returned by the model and y_i is the actual value for data point i

K Nearest Neighbors classifier

Is a simple machine learning algorithm that stores all available variables and classifies new variables based on a similarity measure (distance) KNN has been used in statistical estimation and pattern recognition as a non-parametric technique[55]. KNN classifiers use the distance to classify to class with its neighbors and this depends on the value of K. If K=1 means that the class is simply assigned to its neighbor by using distance function.

Knowing the correct optimal value K is best by first controlling the data. In general, a high value of K is more precise because it reduces the overall noise in your data but there is no guarantee[55]. Cross-validation is another way to consider a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10[56].

A distance metric, such as Euclidean distance, Manhattan distance, or others, is used to measure the similarity between data points in the feature space[56]. Commonly used distance metrics include

Euclidean Distance: $\sqrt{\sum_{n}^{i=1}(x_1 - y_1)^2}$ 2.7

Manhattan Distance: $\sum_{n}^{i=1}|x_1 - y_1|$ 2.8

Minkowski Distance: $\left(\sum_{n}^{i=1} |x_1 - y_1|^p\right)^{\frac{1}{p}}$ 2.9

KNN is relatively simple and has some advantages: They are Intuitive to understand and implement, can handle multi-class classification and non-linear data, doesn't assume any underlying distribution of data. However, it also has limitations: Sensitive to the choice of K and the distance metric, computationally expensive, especially with large datasets, and doesn't learn a model and might not generalize well in complex scenarios.KNN is often used for tasks where interpretability is important, or when the data distribution is not well-defined and linear methods might not perform well.

GaussianNB classifier

The GaussianNB classifier is an established machine learning technique consisting of two components: the Naive Bayes theorem and a Gaussian distribution[57]. The Naive Bayes formulas are commonly employed in the field of probability, particularly in scenarios where the likelihood of event A needs to be determined, given that event B has previously occurred. The GaussianNB algorithm is utilized to estimate the likelihood of a framework for training dataset model fitting, known as maximum posterior, in the context of developing classification predictive models such as Bayes Naive and Bayes Optimal classifiers[58].

The Gaussian Naive Bayes (GaussianNB) algorithm employs a Gaussian distribution to estimate the probability distribution of each feature inside each class, under the assumption that the feature adheres to a normal distribution[59]. The approach use Bayes' theorem in order to compute the posterior probability of each class, given the input features. The anticipated class is determined by assigning it to the class with the highest posterior probability[60]. The

Gaussian Naive Bayes classifier calculates the class posterior probabilities using the following formula[59]:

$$P(C_K|x) = \frac{P(x|C_K).P(C_K)}{P(x)} \quad 2.10$$

where
$P(C_K|x)$ is the posterior probability of class C_K given features x.

$P(x|C_K)$ is the likelihood of observing features x given class C_K

$P(C_K)$ is the prior probability of class C_K

$P(x)$ is the marginal probability of observing features $P(x)$ (a normalization factor).

Gradient Boosting

Gradient boosting is a machine learning methodology that utilizes decision trees of fixed size as base learners[61]. This approach aims to enhance the quality of fit for each individual base learner. The Gradient Boosting method is a classification technique that involves the sequential construction of trees, which are then evaluated and compared to one another using mathematical scores for splits[62]. In contrast to bagging techniques such as Random Forest, Gradient Boosting is a methodology that prioritizes the iterative enhancement of a model's deficiencies. This is achieved by modifying the weights assigned to training instances, with a particular emphasis on those that were inaccurately identified in the preceding iteration.

The fundamental models employed in Gradient Boosting often consist of shallow decision trees, commonly known as "weak learners[63]." Typically, these arboreal specimens have a diminutive stature, characterized by a limited vertical extent and a sparse distribution of branches. The ensemble in Gradient Boosting is constructed in an iterative manner. During

each iteration of the boosting process, a new weak learner is trained with the objective of rectifying the faults caused by the preceding ensemble of models[63]. During each iteration, the training instances are allocated weights according to their performance in the preceding rounds. Instances that were misclassified or had larger errors are assigned more weights, so directing the new weak learner's attention towards rectifying those specific flaws[63].

Gradient Boosting can be conceptualized as an optimization procedure in function space that resembles gradient descent[60]. The loss function is minimized through an iterative process wherein new models are fitted to the negative gradient of the loss, taking into account the preceding ensemble's predictions.

Gradient Boosting incorporates various techniques to mitigate the issue of overfitting. These techniques encompass constraining the depth of individual trees, employing a learning rate to regulate the influence of each weak learner, and injecting randomness during the creation of trees[63]. Gradient Boosting is capable of operating with a diverse range of loss functions, which are selected based on the specific nature of the task at hand. In regression tasks, the Mean Squared Error (MSE) is frequently employed as a standard metric. Log loss, also known as cross-entropy, is frequently utilized in classification problems. The process of boosting is iteratively executed until a predetermined number of iterations have been accomplished or until a specific level of performance has been achieved[63].

$$gt\,(x) = E_y\left[\frac{\delta\varphi(y, f(x)}{\delta(f(x)}\,|x\right]_{f(x)=f^{t-1}x} \qquad 2.11$$

Support Vector Machine

Support Vector Machine (SVM) is a powerful and widely used supervised machine learning algorithm for classification and regression tasks. Its main focus is on finding the hyperplane that best separates different classes in a feature space[64]. Given a set of labeled training data, where each data point xi is associated with a class label yi (either +1 or -1), SVM aims to find a hyperplane that maximizes the margin between the two classes while minimizing classification errors[66]. The hyperplane is defined by the equation[65]:

$$\omega \,.\, x + b = 0 \qquad 2.12$$

Where:

ω is the weight vector perpendicular to the hyperplane.

x is the input feature vector.

b is the bias term.

The distance between a data point xi and the hyperplane is given by:

$$\frac{|\omega \,.\, x+b=0|}{||\omega||} \qquad 2.13$$

The goal of SVM is to find the hyperplane that maximizes the margin, which is the distance between the two classes' closest data points (support vectors) to the hyperplane.

Regression

Regression function is used to find the relationships between two or more features and use this relationship to classify the target value. For example, when there are two variables and one variable increases the other variable may also increase or decrease, or vice versa. Based on this, each variable has a positive or negative relationship. The regression can be grouped into linear and logistic regression.

Linear Regression

Linear regression is a statistical technique employed to establish a linear association between one or more variables. There are two distinct types of linear regression, namely simple regression and multiple regression[66]. Linear regression is a fundamental and extensively employed technique in the field of machine learning. This approach entails the utilization of mathematical principles to do predictive analysis. Linear regression is a statistical technique that enables the estimation of continuous or mathematical variables[66].

Linear regression is a widely employed supervised machine learning approach that is utilized to make predictions about a continuous numerical output, commonly referred to as the target or dependent variable. These predictions are made by considering one or more input features, which are known as independent variables. The model establishes a linear equation to represent the connection between the input variables and the corresponding output.

The equation for simple linear regression, which involves only one input feature, can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad 2.14$$

Where:

y is the predicted output (target variable).

x is the input feature.

β_0 is the intercept (the value of y when x is 0).

β_1 is the slope of the line (represents how much y changes when x changes by 1 unit).

ϵ is the error term, representing the difference between the actual output and the predicted output. It accounts for unexplained variability and measurement errors.

In the case of multiple linear regression, with p input features, the equation becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \quad 2.15$$

Where:

$x_1, x_2, \ldots, x_p$ are the individual input features.

$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (parameters) associated with each input feature.

2.1.4.2 Unsupervised Learning Algorithm

Unsupervised learning, as opposed to supervised learning, is a form of machine learning that is capable of discovering novel patterns within unannotated data. In order to use the extensive quantity of unannotated data, unsupervised learning methods are utilized to acquire complex, highly non-linear models containing millions of parameters[67,68]. The objective of unsupervised learning algorithms is to acquire knowledge and group unannotated datasets into clusters. These algorithms are capable of detecting concealed patterns or clusters within data sets without the need for human intervention[67]. Unsupervised Machine Learning (ML) techniques can be categorized into two main types: clustering and association, as depicted in Figure 2.3[69]. Unsupervised learning is highly advantageous in the context of exploratory data analysis, data preparation, and extracting valuable insights from extensive and intricate datasets.

Clustering

Clustering refers to the procedure of categorizing similar entities into a unified cluster. In order to consolidate similar entities into a cohesive cluster, the algorithms acquire knowledge of the underlying patterns present within the input data[70]. Cluster analysis is a discipline that involves the systematic examination of techniques and algorithms utilized for the purpose of categorizing or clustering items, taking into consideration their shared characteristics and similarities. The prevalent clustering algorithms in machine learning encompass K-Means Clustering, Mean-Shift Clustering, Hierarchical Clustering, and Spectral Clustering[70,71].

Association

Associative learning is a type of unsupervised rule-based machine learning that identifies significant relationships between features in a dataset[72]. K-means clustering, hierarchical clustering, and Self Organizing Map(SOM) are the most prevalent unsupervised Machine learning techniques[73,74].

2.1.4.3 Semi Supervised Learning Algorithm

Semi-supervised learning refers to a machine learning approach that involves the integration of both labeled and unlabeled data sets[75]. Semi-supervised learning is a computational approach that integrates elements of both supervised and unsupervised learning methodologies[75,76]. The availability of labeled data is constrained whereas the volume of unlabeled data is substantial. The semi-supervised strategy addresses the issue of limited availability of labeled data by initially employing unsupervised learning to cluster the unlabeled data, followed by utilizing supervised learning to assign labels to these clusters based on the labeled dataset[75]. This methodology involves the integration of a limited

quantity of labeled data, which consists of data with predetermined outputs, with a more extensive quantity of unlabeled data, which lacks predetermined outputs, for the purpose of constructing a model. The objective is to utilize the unannotated data in order to enhance the efficacy of the model in tasks that include a scarcity of annotated data[76].

Semi-supervised learning encompasses a diverse range of algorithms and approaches.
Self-Training: Self-training is a simple approach where a model is trained on the labeled data and then used to predict labels for the unlabeled data[77]. The predicted labels for the unlabeled data are then added to the labeled dataset, and the model is retrained. This process iterates until convergence.

Co-Training: Co-training involves training multiple models using different subsets of features and then using each model to predict labels for the unlabeled data. Instances that are confidently predicted by both models are added to the labeled dataset. Co-training is often used for cases where the input features can be divided into distinct and informative subsets[76].
Semi-Supervised Support Vector Machines (S3VM): S3VM extends traditional Support Vector Machines (SVM) to incorporate unlabeled data during training[75]. It aims to find a decision boundary that separates the labeled data while using the unlabeled data to optimize the margin.

Generative Models: Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), can be used in a semi-supervised manner[78]. By

training the model to generate realistic data samples, you can also use it to assign labels to unlabeled data based on their generated features.

Pseudo-Labeling: Pseudo-labeling involves training a model on the labeled data and then using this model to predict labels for the unlabeled data[79]. The predicted labels are treated as pseudo-labels and used as if they were true labels for the unlabeled data during subsequent training.

Transfer Learning: Transfer learning techniques, which often involve pretraining a model on a related task or dataset with a lot of unlabeled data, can be used in a semi-supervised way[80]. The pretrained model can then be fine-tuned on the limited labeled data available for the target task.

Semi-supervised learning is particularly useful when obtaining large amounts of labeled data is expensive or time-consuming, as it allows you to utilize the wealth of unlabeled data to improve model performance. However, it requires careful consideration of how to effectively incorporate the unlabeled data into the learning process while avoiding potential pitfalls like label noise from the pseudo-labeled data.

2.1.4.4 Reinforcement Learning Algorithm

Reinforcement learning is a subfield of machine learning that involves the development of a training process that relies on the application of rewards to reinforce desired behaviors and penalties to discourage undesired actions[81].

During the learning process, an artificial agent is subjected to either rewards or penalties based on its exhibited actions. The reinforcement information exchange is depicted in Figure 2.3[81].

Reinforcement Learning (RL) is a machine learning paradigm that centers on the training of agents to make a series of decisions within an environment with the objective of maximizing a cumulative reward. In the field of Reinforcement Learning (RL), an autonomous agent acquires knowledge and skills by active engagement with its surrounding environment, wherein it interacts with the environment and subsequently receives evaluative feedback in the form of incentives[82]. The primary objective of the agent is to acquire knowledge of a policy that establishes a correspondence between states and actions, with the aim of optimizing the anticipated cumulative reward over a given time period.

Key components and concepts of reinforcement learning include:

Agent: The learner or decision-maker that interacts with the environment.

Environment: The external system with which the agent interacts and learns from. It provides the agent with states, and the agent selects actions based on those states.

State: A representation of the current situation or configuration of the environment.

Action (a): The choices that the agent can take in a given state. The actions may have an impact on the environment.

Policy (π): A strategy that the agent follows to select actions in different states. It defines the agent's behavior.

Reward (r): A scalar feedback signal provided by the environment to the agent after taking an action in a specific state. The reward indicates the immediate benefit or cost of the action.

Value Function (V): The expected cumulative reward that an agent can obtain from a given state while following a specific policy. It helps the agent evaluate the desirability of states[82].

Q-Function (Q): The expected cumulative reward that an agent can obtain from a specific state-action pair while following a specific policy. It helps the agent determine the best action to take in a given [81].

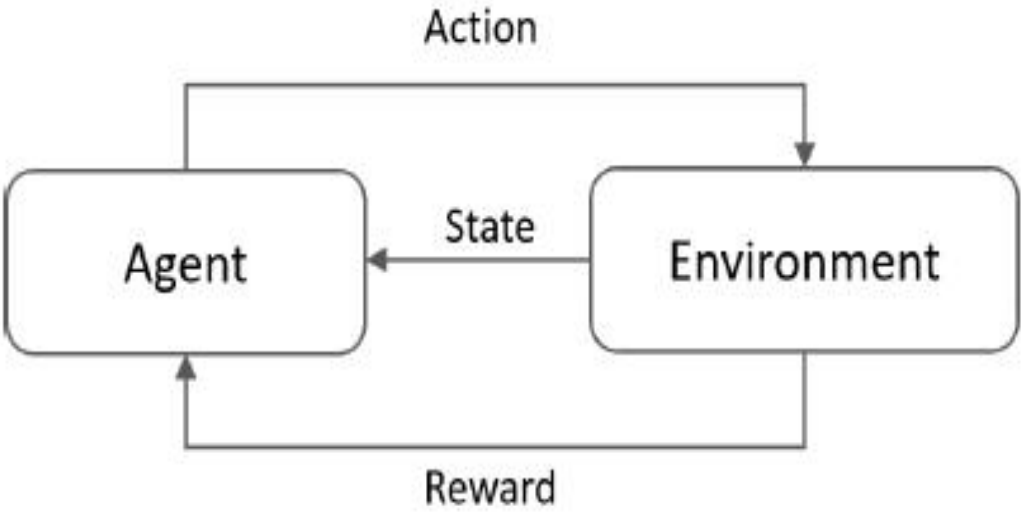
Reinforcement learning algorithms can be categorized into three main types:

Model-Based Algorithms: These algorithms build an internal model of the environment and use it to simulate the effects of different actions[83]. They then use planning or search techniques to find the optimal policy.

Model-Free Algorithms: These algorithms directly learn policies or value functions without explicitly modeling the environment. Q-Learning and SARSA (State-Action-Reward-State-Action) are examples of model-free algorithms[84,85].

Policy Gradient Algorithms: These algorithms directly optimize the policy by adjusting its parameters to increase the expected cumulative reward. They often use gradient ascent techniques to update the policy in the direction of higher rewards[86].

Reinforcement learning has applications in various fields such as robotics, game playing, autonomous driving, recommendation systems, and more. It is particularly suitable for scenarios where the agent learns from trial and error through interactions with the environment.



Overview of Reinforcement Learning[84].

2.1.4.5 Performance Metrics

Machine learning performance metrics are quantitative measures used to assess the effectiveness and quality of a machine learning model's predictions. These metrics provide insights into how well a model is performing and help in making informed decisions about model selection, parameter tuning, and overall model improvement. Different types of machine learning tasks, such as classification, regression, and clustering, require different performance metrics.

Precision

Precision is a performance metric used in binary classification to quantify the proportion of correctly predicted positive instances out of all instances that were predicted as positive by the model[87]. In other words, it focuses on the accuracy of the positive predictions made by the model.

Mathematically, precision (P) is calculated using the following formula:

$$\text{Precision (P)} = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad 2.16$$

Where:

TP (True Positives) represents the number of instances that are actually positive and were correctly predicted as positive by the model.

FP (False Positives) represents the number of instances that are actually negative but were incorrectly predicted as positive by the model.

Precision ranges between 0 and 1, where:

P = 1 indicates perfect precision, meaning that all positive predictions made by the model were correct.

P = indicates that the model's positive predictions were all incorrect.

A higher precision value is desirable when the cost of false positives (predicting positive when it's actually negative) is relatively high.

Accuracy

Accuracy is a common performance metric used in classification tasks to measure the overall correctness of a model's predictions[88]. It represents the proportion of correctly classified instances out of the total instances in the dataset.

Mathematically, Accuracy (ACC) is calculated using the following formula:

$$\text{Accuracy (ACC)} = \frac{\text{Correct Predcitions}}{\text{Total Predictions}} \quad 2.17$$

Where:

Correct Predictions is the number of instances that were correctly classified by the model.

Total Predictions is the total number of instances in the dataset.

Accuracy values range between 0 and 1, where:

ACC = 1 indicates perfect accuracy, meaning that all predictions made by the model were correct.

ACC = 0 indicates that none of the model's predictions were correct.

While accuracy is a straight forward and widely used metric, it might not be suitable for imbalanced datasets where one class is much more prevalent than the other. In such cases, a high accuracy can be misleading because the model might be performing well on the dominant class but poorly on the minority class[88]. Additionally, accuracy might not be the best metric for certain types of problems where the cost of misclassification varies between classes. In such situations, other metrics like precision, recall, F1-score, or area under the ROC curve (AUC-ROC) might provide a more informative assessment of the model's performance[88].

Recall

Recall, also known as Sensitivity or True Positive Rate, is a performance metric used in binary classification to measure the proportion of actual positive instances that were correctly predicted as positive by the model[89]. It focuses on capturing the model's ability to identify all positive instances, regardless of how many false positives it may predict[90].

Mathematically, recall (R) is calculated using the following formula[90]:

$$\text{Recall (R)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad 2.18$$

Where:

TP (True Positives) represents the number of instances that are actually positive and were correctly predicted as positive by the model.

FN (False Negatives) represents the number of instances that are actually positive but were incorrectly predicted as negative by the model.

Recall values range between 0 and 1, where:

R = 1 indicates perfect recall, meaning that the model correctly identified all positive instances in the dataset.

R = 0 indicates that the model failed to identify any of the positive instances.

Recall is particularly important when the cost of false negatives (not predicting a positive when it's actually positive) is high. For instance, in medical diagnostics, failing to diagnose a disease could have serious consequences.

F1-Score

The F1-score is a performance metric used in binary classification that combines both precision and recall into a single value[91]. It provides a balanced measure of a model's accuracy by considering both the true positive rate (recall) and the positive predictive value (precision).

Mathematically, the F1-score (F1) is calculated using the following formula[91]:

$$\text{F1-Score (F1)} = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad 2.19$$

Where:

Precision is the proportion of true positive predictions out of all positive predictions, calculated as $\frac{(TP)}{(TP)+(FP)}$, where (TP) is true positives and (FP) is false positives.

Recall is the proportion of true positive predictions out of all actual positive instances, calculated as $\frac{(\text{TP})}{(\text{TP})+(\text{FN})}$, where (TP) is true positives and (FN) is false negatives.

The F1-score ranges between 0 and 1, where:

(F1 = 1) indicates perfect precision and recall, meaning that all positive predictions are correct and all actual positive instances are identified.

(F1 = 0) indicates that either precision or recall (or both) is zero, representing poor model performance[91].

The F1-score is particularly useful when dealing with imbalanced datasets or when both false positives and false negatives have significant implications. It helps strike a balance between making accurate positive predictions and identifying as many actual positive instances as possible. It's important to consider the F1-score alongside precision and recall to make well-informed decisions about model performance, especially when there's a need to balance the trade-off between different types of errors.

Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a binary classification model's performance across different discrimination thresholds[92,93]. It illustrates the trade-off between the true positive rate (recall) and the false positive rate as the decision threshold for classifying positive and negative instances is varied.

Mathematically, the ROC curve is created by plotting the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis as the threshold for classifying positive instances is adjusted. The formulas for TPR and FPR are as follows[93]:

$$\text{True Positive Rate (TPR)} = \frac{True\ Positives\ (TP)}{True\ Positive + False\ Negatives} \qquad 2.20$$

Where:

TP (True Positives) represents the number of instances that are actually positive and were correctly predicted as positive by the model.

FN (False Negatives) represents the number of instances that are actually positive but were incorrectly predicted as negative by the model.

FP (False Positives) represents the number of instances that are actually negative but were incorrectly predicted as positive by the model.

TN (True Negatives) represents the number of instances that are actually negative and were correctly predicted as negative by the model.

The ROC curve starts at the point (0,0), which corresponds to a threshold where all instances are classified as negative. As the threshold increases, the true positive rate and false positive rate change, and the curve traces the path from the lower-left corner to the upper-right corner of the plot. A diagonal line connecting the points (0,0) and (1,1) represents a random classifier[93].

The area under the ROC curve (AUC-ROC) is a common metric used to quantify the overall performance of a binary classification model. AUC-ROC values range between 0 and 1, where[93]:

AUC-ROC = 1 indicates a perfect classifier.

AUC-ROC = 0.5 indicates a random classifier.

AUC-ROC between 0.5 and 1 indicates varying degrees of classification performance.

The ROC curve provides a visual way to assess the performance of a classification model and helps in comparing the trade-off between true positive and false positive rates at different threshold settings[93]. The AUC-ROC metric summarizes the overall quality of the model's predictions across all possible threshold values.

Area Under the Precision-Recall Curve (AUC-PR)

The Area Under the Precision-Recall Curve (AUC-PR) is a performance metric used in binary classification to quantify the overall quality of a model's predictions, with a specific focus on the trade-off between precision and recall[94]. Unlike the ROC curve, which focuses on the trade-off between the true positive rate (recall) and the false positive rate, the PR curve considers the trade-off between precision and recall. The PR curve is created by plotting precision on the y-axis and recall on the x-axis for different threshold settings. The AUC-PR is then calculated as the area under this curve[95].

Mathematically, the AUC-PR can be approximated using various methods, but one common approach is to use numerical integration, such as the trapezoidal rule[95].

1. Sort the instances based on their predicted probabilities or scores from the model in decreasing order.

2. Calculate the precision and recall for each threshold setting, using formulas:

Precision= $\frac{(TP)}{(TP)+(FP)}$

Recall = $\frac{(\mathrm{TP})}{(\mathrm{TP})+(\mathrm{FN})}$

Calculate the difference in recall (Δ Recall) for consecutive threshold settings.

Calculate the average precision (ΔP) using the precision values and Δ Recall values[95]:

AP= $\sum (Precision \times \Delta \text{ Recal})$

The AUC-PR is then calculated as the sum of the products of precision and Δ Recall values, normalized by the maximum possible value of AP:

AUC-PR = $\frac{AP}{max\,(AP)}$

The AUC-PR value ranges between 0 and 1, where:

AUC-PR = 1 indicates perfect performance, meaning that the model achieves high precision and high recall across different threshold settings.

AUC-PR= 0 indicates poor performance, where either precision or recall (or both) are low.

The AUC-PR is particularly useful when dealing with imbalanced datasets or when there's a focus on positive class prediction[95]. It provides insights into how well the model is performing across various levels of positive class prevalence. AUC-PR summarizes the performance of a binary classification model in terms of precision and recall trade-offs, providing a comprehensive view of its effectiveness, especially in scenarios where the class distribution is skewed.

2.2 Methodological Review

This section gives a theoretical background of the main classification algorithms used in this study.

2.2.1 Gradient Boosting Classifier

The theoretical underpinnings of the Gradient Boosting Classifier entail the iterative optimization of a loss function by the consecutive addition of new weak learners. These learners are specifically designed to rectify faults caused by the ensemble of previous models. The aforementioned procedure yields a robust prediction model that exhibits a high level of accuracy across a diverse set of classification tasks[63].

The Gradient Boosting Classifier is a widely used ensemble learning method that constructs a predictive model through an iterative process of aggregating the outputs of weak learners, such as decision trees[62]. This approach is designed to specifically address and rectify the

faults made by the preceding models. The fundamental principle underlying gradient boosting involves the optimization of a loss function that is differentiable[62,63]. This is achieved by iteratively updating the parameters of the model in a manner that minimizes the gradient of the loss. The following are the fundamental mathematical equations that underlie the Gradient Boosting Classifier.

1. **Loss Function**: Gradient Boosting begins with defining a loss function *(L(y, F(x))*, where y is the true target value and F(x) is the current model's prediction[96]. The goal is to minimize this loss function. Common loss functions for classification include log loss (cross-entropy) for binary classification and multinomial deviance for multi-class classification.
2. **Negative Gradient (Residuals)**: The negative gradient $\left(-\frac{\delta\, L(y, F(x)}{\delta\, F(x)}\right)$ of the loss function with respect to the current prediction represents the residuals, which indicate the errors that need to be corrected[96]. These residuals are used to train the new weak learner.
3. **Weak Learners (Base Models)**: Gradient Boosting uses a sequence of weak learners, often decision trees. Each weak learner is designed to predict the negative gradient of the loss function with respect to the model's current prediction[62].
4. **Initialization**: The process starts by initializing the model's predictions, often with a simple estimator like a decision stump (a single-level decision tree)[63].
5. **Gradient Descent**: In each iteration, a new weak learner is fit to the negative gradient of the loss function with respect to the current model's predictions[97]. The negative gradient points in the direction of steepest decrease in the loss function.

6. **Learning Rate**: A learning rate (also known as shrinkage) scales the contribution of each new model to the ensemble[63]. A smaller learning rate can help prevent overfitting but may require more iterations to achieve good performance.
7. **Update Ensemble Predictions**: The predictions of the ensemble are updated by adding the scaled predictions of the new weak learner. The learning rate controls how much the new model's predictions influence the ensemble (Learning Rate x New Model Prediction)[62].
8. **Iteration**: Steps 3-6 are repeated iteratively, with each new model aiming to correct the errors of the previous ensemble[98].
9. **Final Prediction**: The final prediction of the Gradient Boosting Classifier is the sum of the predictions from all individual models, each scaled by the learning rate.

2.2.2 Gaussian Naive Bayes (GNB)

The Gaussian Naive Bayes (NB) Classifier is a machine learning algorithm that operates on the idea of Naive Bayes. It makes the assumption that the features of the data are distributed according to a Gaussian (normal) distribution within each class[57]. The technique is well-suited for handling continuous data and is characterized by its simplicity and efficiency in classification tasks.The Gaussian Naive Bayes (GNB) classifier is a probabilistic method commonly employed in machine learning for the purpose of classification. However, it is specifically tailored to handle continuous data by making the assumption that the features adhere to a Gaussian distribution. According to[57,58], the Gaussian Naive Bayes (NB) classifier's theoretical model encompasses the following components:

The Probability Density Function (PDF) that characterizes the Gaussian distribution is as follows: The Gaussian Naive Bayes classifier makes the assumption that the feature distributions of each class follow a normal (Gaussian) distribution[99]. The utilization of the Probability Density Function (PDF) of a Gaussian distribution is employed for the estimation of the probability of a feature value, given a specific class.

According to[99], for a feature `X` and a class `C`, the PDF of the Gaussian distribution is given by[99]:

$$P(X|C) = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad 2.21$$

Where:

$\boldsymbol{\mu}$ is the mean of the feature values for class `C`.

$\boldsymbol{\sigma^2}$ is the variance of the feature values for class `C`.

Naive Bayes Principle: The Naive Bayes principle assumes that features are conditionally independent given the class[99]. In other words, the presence or absence of a particular feature doesn't influence the presence or absence of any other feature, given the class label. This assumption simplifies the calculations and allows for efficient estimation.

Likelihood Estimation: For each feature and class, the Gaussian NB classifier estimates the mean $\boldsymbol{\mu}$ and the variance $\boldsymbol{\sigma^2}$ from the training data[63]. These parameters are then used in the PDF to compute the likelihood of a feature value given a class.

Prior Probability: The prior probability P(C) of each class is estimated from the training data. It represents the overall likelihood of each class occurring in the dataset.

Posterior Probability: Using Bayes' theorem[100], the posterior probability P(C|X) of each class given a feature vector X is calculated. This is the probability that an instance with feature values X belongs to class C.

Classification: For a given feature vector X, the Gaussian NB[101] classifier calculates the posterior probabilities for all classes and assigns the class with the highest posterior probability as the predicted class.

Advantages of Gaussian NB Classifier

1. Suitable for Continuous Data: GNB works well with continuous feature data that follows Gaussian distribution.

2. Fast and Simple: It's computationally efficient and doesn't require complex parameter tuning.

3. Handles High Dimensions: GNB can handle high-dimensional data well due to its conditional independence assumption.

Limitations of Gaussian NB Classifier

Strong Assumption: The assumption of Gaussian distribution might not hold for all types of data.

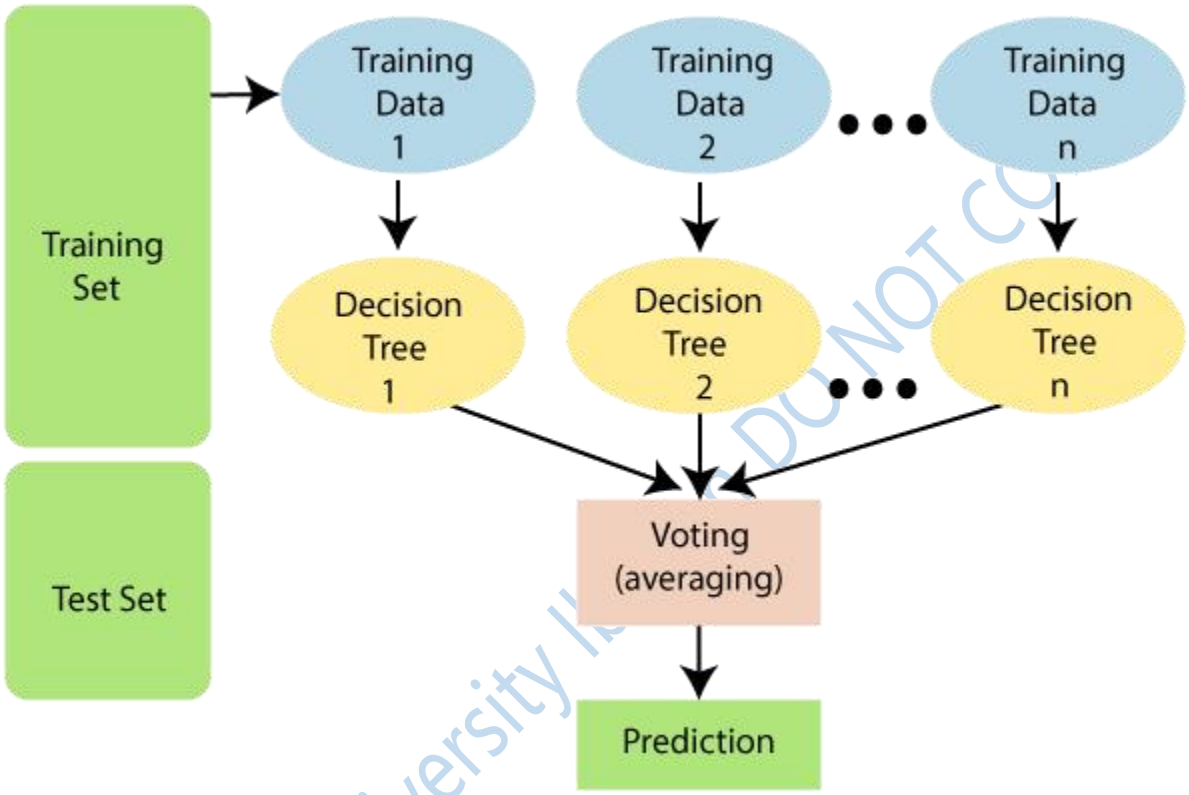
1. Independence Assumption: The independence assumption might not be valid in some real-world scenarios.

2. In ability to Capture Complex Relationships: GNB can't capture complex relationships between features.

2.2.3 Random Forest

The Random Forest algorithm is a supervised ensemble technique that leverages a multitude of decision trees to provide predictions[102]. The Random Forest algorithm is a classification method that comprises a collection of tree-based classifiers. These classifiers are constructed using independent random vectors that are identically distributed. Each tree in the Random Forest assigns a unit vote to the most prevalent class for a given input, denoted as x[103]. A randomly produced vector, which is independent of the previously generated random vectors from the same distribution, is utilized to construct a tree using the training set. From this

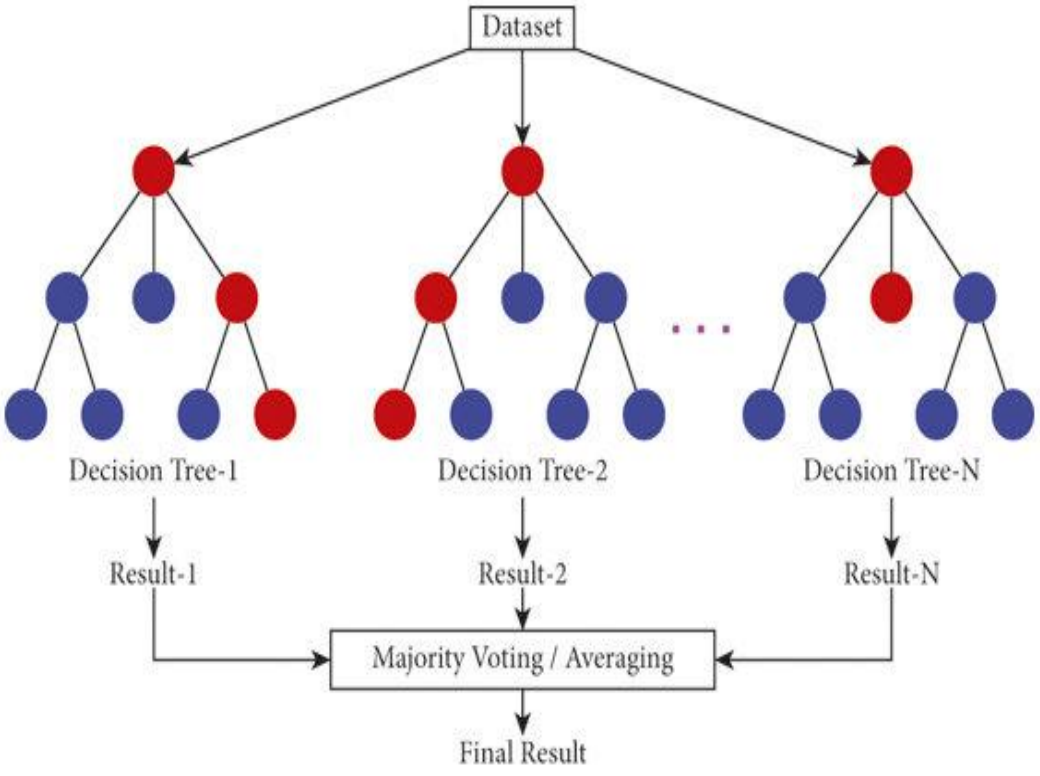
process, an upper bound is derived for Random Forests in order to estimate the generalization error based on two parameters[104,105]. The precision and interconnectedness of separate classifiers have been extensively studied.



Random Forest Flow Chart[105].

In order to obtain various subsets of samples, the bootstrap method is employed. Each subset of samples is then utilized to construct a Decision Tree. Subsequently, several Decision Trees are aggregated to form a Random Forest. The classification conclusion of the sample is determined through a voting process on the Decision Tree[106]. In academic research, it is common for researchers to enhance the precision of a classifier by iteratively refining its parameters and reducing the interdependence across several classifiers.

The Random Forest algorithm is utilized in the classification process to collectively reduce the impact of classification errors from each individual base classifier. This is achieved by employing a common distribution of errors, resulting in the overall decrease of the classification effect. The test features are utilized to apply the rules of each randomly generated Decision Tree in order to predict the outcome and thereafter record the anticipated result (target). Calculate the number of votes received for each projected target[107,108]. The final prediction from the Random Forest algorithm can be regarded as the anticipated goal with the highest number of votes.



Random Forest Training Flow Chart[105].

The RF algorithm is very efficient, as it handles datasets that contain continuous variables, as well as categorical variables robustly. An RF classifier contains subsets of various tree classifiers $\{\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\Theta_k}), \boldsymbol{k} = \boldsymbol{1}, \boldsymbol{2}, ...\}$ where the $\boldsymbol{\Theta_k}$ are independently and identically distributed random vectors, with each tree being able to specify the modal class at input $\boldsymbol{x}$ [109]**.** The performance index, which solely approximates the confidence interval (CI) of the RF model is given as

$$mg(x, y) = av_k I(h_k(x, \Theta_k) = y) - \max_{j \neq y} av_k I(h_k(x, \Theta_k) = j) \qquad 2.22$$

where $I(.)$ denotes an indicator function, and $av(.)$, the average value. It is observed that as the margin increases, the confidence level also increases. The generalisation error becomes

$$PE^* = P_{x,y}(mg(x, y) < 0), \qquad 2.23$$

Advantages of the RF Algorithm

1. **High Accuracy**: Random Forest often provides greater accuracy compared to other algorithms and works effectively with large datasets.
2. **Efficient Handling of High Dimensionality**: It can manage thousands of input variables quickly and effectively.
3. **Feature Importance**: RF provides insights on which variables are most and least significant in classification, aiding feature selection.
4. **Handles Missing Data**: It offers techniques to estimate incomplete data, maintaining model performance despite missing values.
5. **Robust to Missing Details**: RF can handle missing information without a significant loss in accuracy.

6. **Prototype Generation**: Prototypes generated within the model provide metadata on the relationships between variables.
7. **Variable Interaction Analysis**: RF allows for the examination of complex relationships between variables.

Disadvantages of the RF Algorithm

1. **Risk of Overfitting**: Overfitting can occur, particularly with regression tasks on a single dataset if hyperparameters are not carefully tuned.
2. **Challenges with Multi-Dimensionality**: RF struggles with multi-valued or multi-dimensional attributes, favoring categorical variables with distinct levels.

2.2.4 Decision Tree

Decision trees are widely employed in several domains, including machine learning, image processing, and pattern recognition, due to their considerable efficacy. Decision Trees (DT) are a sequential model that effectively and cohesively combines a number of fundamental tests, wherein a numeric feature is compared to a threshold value in each test[110]. Constructing conceptual rules is often considered to be a less complex task compared to determining the numerical weights within the neural network of interconnected nodes. DT is mostly utilized for the purpose of categorization. Furthermore, Decision Trees (DT) are commonly employed as a classification model in the field of Data Mining. The components of each tree consist of nodes and branches[110]. In this context, it can be observed that each individual node serves as a representation of distinct features within a given category that is subject to classification. Furthermore, it is worth noting that each subset within this framework delineates the potential range of values that can be assumed by the respective node. Decision

trees have been widely applied in several fields due to their straightforward analysis and ability to accurately handle diverse forms of input.

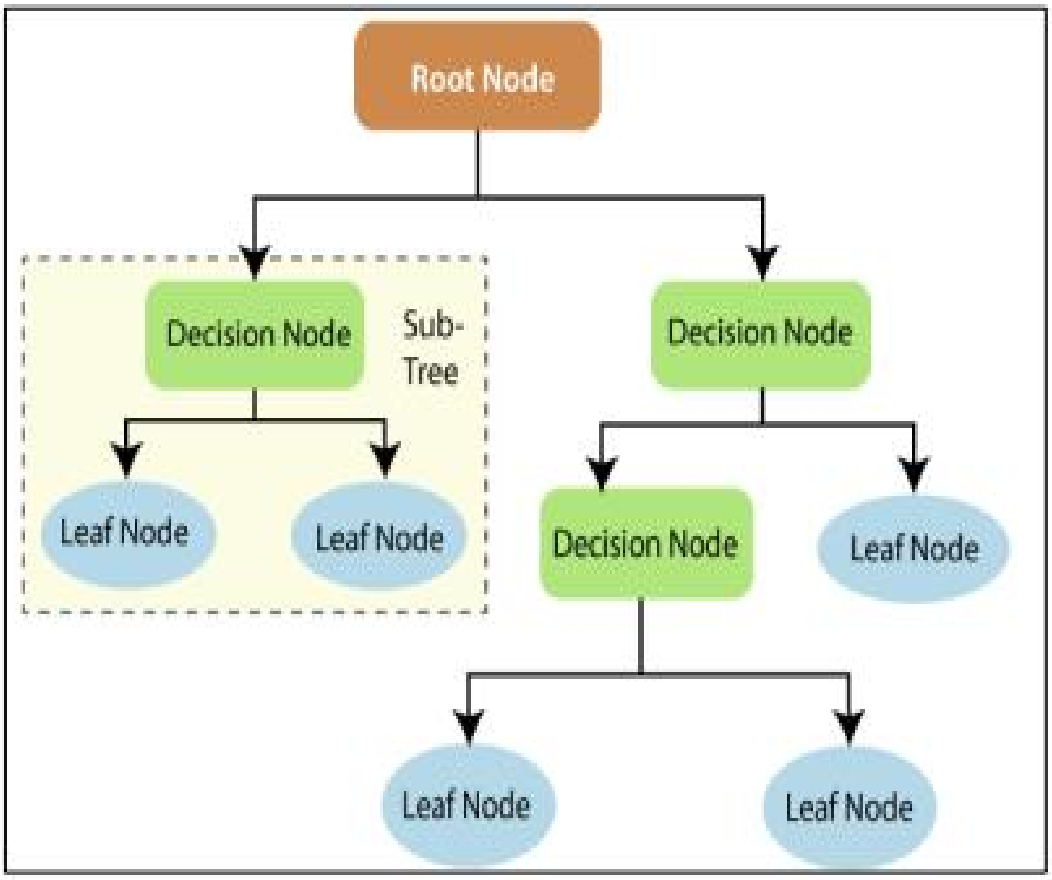


Figure 2.7. Decision Tree Flow Chart[110].

Types of Decision Tree Algorithms:

There exist various types of Decision Tree (DT) algorithms, including Iterative Dichotomies 3 (ID3), Successor of ID3 (C4.5), Classification And Regression Tree (CART), CHi-squared Automatic Interaction Detector (CHAID), Multivariate Adaptive Regression Splines (MARS), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), Conditional Inference Trees (CTREE), Classification Rule with Unbiased Interaction

Selection and Estimation (CRUISE), and Quick, Unbiased and Efficient Statistical Tree (QUEST)[110].

In the context of decision algorithms, entropy and information gain are utilized as metrics to assess the level of impurity or unpredictability present in a given dataset. The entropy value is constrained between the range of 0 and 1. The optimal value is achieved when it equals zero, and the suboptimal value is attained when it deviates from zero. In other words, the closer the value is to zero, the more desirable it becomes. As depicted in Figure 2.7. In the context of the study, if the focus of investigation is the target

$$\text{Entropy (S)} = \sum_{i=1}^{c} P_i \log 2^{p_i} \qquad 2.24$$

Where P_i is the ratio of the sample number of the subset and i-th attribute value.

Benefits of Decision Tree

The DT algorithm is part of the supervised learning algorithm family, and its main objective is to construct a training model that can be used to predict the class or value of target variables through learning decision rules inferred from the training data. The DT algorithm can be used to

i. solve regression and classification problems
ii. Simple to comprehend
iii. Quickly translated to a set of principle for production
iv. Can classify both categorical and numerical outcomes, but the attribute generated must be categorical
v. No a priori hypothesizes are taken with consideration to the goodness of the results

However, DT has some draw backs which include;

i. The optimal decision-making mechanism can be deterred and incorrect decisions can follow
ii. There are lots of layers in the decision tree, which makes it interesting
iii. For more training samples, the decision tree's calculation complexity may increase

2.3 Review of Related Work

This section reveals what several authors have contributed significantly to the development of loan prediction system.

In a study where an Artificial Neural Network algorithm was utilised to construct a loan prediction system, the researchers found that. The system was developed and put into operation utilising Python as the programming language, Hypertext Markup Language (HTML), and Cascading Style Sheet (CSS) for the front end, and then PHP as the backend. When determining the system's level of accuracy, the confusion matrix was also utilised by the system as one of the performance measures. The outcome of the test demonstrated that the system had an accuracy rate of 92%; this demonstrated that the developed system projected correctly and was able to determine whether a potential borrower would be able to repay the loan or not. The technology is also able to determine whether or not a loan will be paid back poorly by the debtor. In the end, the system was evaluated in terms of its accuracy

in comparison to other studies that had been conducted previously, and it was determined that the suggested system performed significantly better than the studies that had been conducted previously[111].

In another study, different machine learning approaches were utilised to determine a customer's eligibility for a loan. These strategies were employed in order to make the prediction. Data on customers is gathered from a variety of banks, and access is granted to customer profiles in order to perform an analysis of the data based on certain factors that are required for integration with machine learning strategies. The machine learning strategy, which involves analysing the data and providing the results based on the customer profile in order to approve loans, is the most sophisticated method compared to the more traditional loan approval-based methods. Cleansing the data, selecting the most important attributes, and evaluating the effectiveness of several machine learning approaches (decision tree, random forest, support vector machine, K-nearest neighbour, and decision tree with AdaBoost, to name a few) in determining whether or not a customer is eligible for a loan are the primary goals of the project. A model is trained using the train dataset, and then the model's performance is evaluated using the test dataset. The data are first partitioned into the training and testing parts. The findings indicate that the ensemble model decision tree with the adaboost technique provided a higher level of accuracy compared to the other models that were deployed[112].

A number of different machine learning algorithms that predict whether or not a loan application will be approved are compared and contrasted in this study. Random Forest

Classifier, K-Nearest Neighbours Classifier, Support Vector Classifier, and Logistic Regression are some of the classification techniques that have been investigated. Exploratory data analysis and feature engineering are the two processes that are used to prepare the dataset.When evaluating the effectiveness of each method, we look at many statistics, such as accuracy score, F1 score, and ROC score.According to the data, the Random Forest Classifier had the best accuracy, scoring 98.04 percent. This was followed by the K-Nearest Neighbours Classifier (78.49 percent), Logistic Regression (79.60 percent), and Support Vector Classifier (68.71 percent). These findings emphasise the potential for algorithms that use machine learning to improve the process of loan approval and lower the likelihood of loan defaults. In general, this study sheds light on the efficacy of several machine learning algorithms for the prediction of loan acceptance, which might be helpful for financial organisations looking to improve their decision-making process[113].

In a study employing machine learning to predict whether or not a loan would be approved. The Logistic regression model was employed for the study. The data is obtained from Kaggle and then used for the purposes of studying and making predictions. Logistic Regression modelling has been carried out, and the various performance metrics have been obtained. Comparisons of the models are made using performance metrics such as sensitivity and specificity to evaluate their relative merits. The culmination of the research has revealed that the model generates a variety of findings. The model is only slightly superior because it takes into account variables (personal characteristics of customers, such as age, purpose, credit history, credit amount, credit duration, etc.) in addition to checking account information, which reveals a customer's level of wealth. These variables should be taken into account in

order to accurately calculate the likelihood that a customer will default on a loan. Therefore, by employing a method known as logistic regression, the appropriate clients to be targeted for the provision of loans can be simply identified by assessing the likelihood that they will default on those loans[114].

In a piece of research that made use of four different classification-based machine learning algorithms, namely Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest, the Support Vector Machine approach was found to be the most accurate in predicting whether or not a loan application would be approved[114].

Another similar study aimed to predict acceptance of the bank loan offers using the Support Vector Machine (SVM) algorithm. In this context, SVM was used to predict results with four kernels of SVM, with a grid search algorithm for better prediction and cross validation for much more reliable results. Research findings show that the best results were obtained with a poly kernel as 97.2% accuracy and the lowest success rate with a sigmoid kernel as 83.3% accuracy. Some precision and recall values are lower than normal ones, like 0.108 and 0.008 due to unbalanced dataset, like for 1 true value, there are 9 negative values (9.6% true value)[116].

In another research, the researchers employed something Modified Synthetic Minority Oversampling Technique (MSMOTE). It is a method of oversampling in which synthetic data of the minority class are generated in order to balance with the data of the majority class. In order to further increase the overall performance of bank loan prediction systems, this is

paired with the ensemble classifier technique. MSMOTE is a modification of the method known as the Synthetic Minority Oversampling Technique (SMOTE). The unbalanced dataset is subjected to the application of bagging and boosting-based ensemble approaches in order to improve the performance of loan prediction. Kaggle is used to collect the dataset that will be used to validate the proposed strategy. The results of the experiments reveal that the proposed model, MSMOTE, achieved 95% of precision and accuracy when combined with adaptive boosting. This result was achieved by combining the two techniques. In contrast, when MSMOTE was used in conjunction with Bagging and Random Forest, the resulting precision and accuracy was 99%[117].

In addition, in a work that proposed a model that aggregates multiple machine learning algorithms with ensemble algorithms such as bagging and voting classifiers. The most important purpose of the work that we are doing is to determine whether or not a specific individual is qualified to receive the loan. The new model that we have presented requires less effort and time from humans to process, and it also generates results that are more accurate than those produced by existing methods. According to the findings of our experiments, the performance of the standard model can be improved by up to 94% by using our approach[118].

Another study seeks to provide a comprehensive review of lending estimation systems and structures that employ prediction methods and techniques flourished and developed after recent years. In this study and paper, researchers studied the learning techniques as well as the raw datasets utilized for training and test sets. The system model's precision is also

discussed. Our work also provides a quick overview of a few datasets that can be used to anticipate loan/mortgage analysis. Recent and future trends are also spotlighted[119].

In this study, ten Machine Learning models, including Decision Tree, Logistic Regression, K Nearest Neighbour (KNN), Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes, XGBoost, Gradient Boosting, Adaboost, and Deep Learning models, including Deep Neural Network (DNN) and Long Short Term Memory network (LSTM), was compared to predict loan applicants who deserve the money. Both Logistic Regression and Linear Discriminant Analysis had the best accuracy of 82.43% after examining all of these models[120].

In a study that was carried out with the objective of determining how well various machine learning and deep learning models can forecast loan eligibility. For this purpose, the authors made use of a Kaggle dataset that contained 614 samples and 13 attributes. After gathering the data, resolving any missing values, and standardising the independent variables, the next step in developing a more accurate model is to put it through training and testing. This process is done after the variables have been standardised. Following this, the results are analysed. In order to accomplish this, they make use of well-known Machine Learning techniques such as Decision Tree, Logistic Regression, K-Nearest Neighbour (KNN), Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis, Gaussian Naive Bayes, XGBoost, Gradient Boosting, and AdaBoost, in addition to Deep Learning techniques such as Deep Neural Network (DNN) and Long Short Term Memory

network (LSTM). When evaluating the models, a number of characteristics are considered, such as their F1 score, their precision, and their accuracy[121].

In this study, the author basically did the thorough investigation on DGHI dataset for the purpose of analysing the customer eligibility using LRD machine learning algorithms (i.e. Logistic Regression, Random Forest, and Decision Trees). The purpose of this inquiry was to determine if the client is eligible for a loan or not. The experimental research that was carried out was split into two parts: training, and then testing, using the data that was readily available. The authors came to the conclusion that logistic regression was the most effective method for predicting the likelihood of a loan being granted to a customer based on the findings of the research that they had carried out. On the basis of factors such as Loan_id, Gender, Married, Education, Self-employed, and so on, the authors were able to get the results they wanted and choose Logistic Regression as the appropriate technique for the given approach. This was done so that they could get the most accurate results possible. It has been agreed that the accuracy and precision of Logistic Regression will be improved in the work that will be done in the future[122].

Another study developed a model with the help of artificial intelligence technology, namely machine learning algorithms, with the purpose of making a prediction about the likelihood of small and medium-sized businesses (SMEs) failing on the repayment of loans. The research utilised the Louvain clustering algorithm to effectively group the loan recipients based on their cumulative repayment amounts over time. Additionally, two distinct machine learning techniques, namely Logistic Regression (LR) and k-Nearest Neighbour (k-NN), were utilised

to evaluate their efficacy in classifying recipients' risk levels, which were either low-risk or high-risk. A mean accuracy score of 100% was attained by the LR model, which indicates a high degree of precision that can accurately estimate the risk of SME loan payback. This is further confirmed by the fact that the LR model obtained an Area Under the Curve (AUC) value of 1.0. This indicates that the model has achieved optimal separation of the two groups, and as a result, it is extremely trustworthy for risk prediction. It is thought that this method will improve the effectiveness and precision of credit risk assessment, which could assist financial institutions (FIs) to optimise their decision-making processes and limit possible losses brought on by loans that go into default[123].

This research was conducted as part of an effort to design a system capable of making such predictions. The generated solution, known as LoanApprovalStatus, was constructed by making use of bank data that had been gathered in the past in order to forecast the approval status of a financial loan. The primary focus of LoanApprovalStatus is classification, and the application makes use of a machine learning model that was created by integrating a voting classifier with a variety of additional methods. A graphical user interface that can be accessed online was developed in order to engage with and direct users. When making a prediction, the inputs provided by the user are taken into consideration. The prediction of the loan approval will be able to forecast individual cases one at a time, with the outcomes displayed individually. Forecasting, Machine Learning, Data Classification, and Credit Approval[124].

In a study that intends to utilise machine learning techniques and algorithms to analyse the applicant's personal information and anticipate if a customer is eligible for a loan and the loan amount based on his financial status, the personal information of the applicant will be analysed. To train and analyse the client creditworthiness and loan approval decision made by the bank employee regarding whether or not the request will be granted, the models were developed using supervised learning techniques such as Logistic Regression, Decision Tree, SVM, and Random Forest. These models were used to determine whether or not the request would be granted. The results of the aforementioned models showed that Random Forest had the highest accuracy, recall, and F1 scores out of the three models. Random Forest also had the largest percentage of correct predictions. The Random Forest machine learning technique was chosen as the best model to forecast the loan approval status, and an additional model was constructed to calculate the suitable loan sanction amount depending on the information provided by the customer. Both of these models are described in more detail below. When deciding how much of a loan to approve a customer for, one of the most important factors to consider is that customer's credit score. Additional research reveals to individuals how they might improve their chances of being approved for a loan in the future by demonstrating a track record of satisfactorily meeting requirements in accordance with the bank's specifications[125].

Another study attempted to establish a system that would allow one to forecast whether or not the applicant that was chosen would be a worthy applicant for the approval of the loan. This study was unsuccessful. The system makes its predictions on the basis of a model that has been trained with the assistance of algorithms for machine learning. The authors have

gone as far as analysing the degree of accuracy achieved by various machine learning methods. We obtained an accuracy rate ranging from 75 to 85%, but the Logistic Regression method produced the highest accuracy, which was 88.70% of the time. The user can enter the information necessary for the model to make a prediction through the user interface web application that is included with the system. The fact that this model takes into account a large number of characteristics is one of its drawbacks. However, in real life, a loan application may occasionally be accepted based on a single compelling characteristic; however, that outcome will not be attainable when utilising this approach[127].

Another study was conducted with the goals of addressing the issue of customers' loans being placed into default by determining whether or not the consumer is qualified for a loan and assisting customers in maintaining a good credit score by advising them to avoid taking out loans that they are unable to repay. The prediction is made using data pre processing techniques, which clean the dataset and provide accurate data for training machine learning models. These techniques are used to clean the dataset. In the course of our job, a great number of machine learning models, the majority of which are employed for classification algorithm training and testing, are utilised to determine whether or not a loan applicant should be approved. The accuracy of models is evaluated through the use of the baseline modelling and the KFold cross validation procedures. The model that achieves the highest level of accuracy is the one that is taken into consideration for the development of the loan approval prediction model. After being trained, this model is then used to take inputs from the user and produce the results of predictions as the output[127].

A related study suggests employing machine learning models in conjunction with ensemble learning techniques in order to evaluate whether or not it is feasible to grant individual loan applications. It is feasible to improve the accuracy of the process of selecting eligible candidates from an existing list by making use of this strategy, which entails applying it. As a result, this procedure can be utilised to address the difficulties outlined above regarding the processes involved in the approval of loans. Because it takes a significantly shorter amount of time to sanction the loan, the concept is beneficial not only to bank staff but also to applicants. Because of this, the authors were able to determine whether or not an application poses a security risk, and hence, the entire process of feature verification is carried out using machine learning techniques. The applicant and the bank workers can both benefit from using the loan prediction tool. In this work, there is a proposition that seeks to supply a method that is speedy, uncomplicated, and immediate for finding individuals who are qualified. There will be a cutoff time for the approval of a loan applicant. With the help of this technology, switching to an application that needs to be checked first is feasible. It is reserved solely for the administrative leadership of the bank or financial institution. The entirety of the process of prediction is carried out in private so that no stakeholders may influence the processing. It is possible to transmit individual loan identification numbers to various banking departments in order to give those departments the ability to take necessary action about the application. However, the bank can only make a limited number of slots accessible, and it must sell them to a select group of customers in order to cover its costs. As a result of this, one of the regular steps is assessing who will be unable to repay the loan and who will demonstrate to the bank that they are a more reliable choice. This study proposes the implementation of a system for the approval of loans, which would use predetermined

criteria to decide whether or not a particular person should be granted a loan. The strategy that we propose to banks will help them identify trustworthy persons who have asked for loans, which will increase the possibility of prompt repayment of the loans. This evaluation is carried out by utilising a wide variety of machine learning algorithms in order to provide the most accurate findings when estimating the potential outcomes of a loan[128].

In a work that presented a loan recommendation system, the author stated that the system would offer an instant and straightforward method for selecting the appropriate applicant based on the validation of attributes. The use of regression modelling is going to be used in this study with the intention of predicting a model for the distribution of loans. Each attribute receives a certain amount of weight based on its relative importance to the bank. This technology is able to make predictions about whether or not an application will be successful in obtaining the loan. It is advantageous not just for those working at the bank but also for those who might be eligible for a loan from the institution. In this section, we are going to extract the essential characteristics from the loan dataset. Within the loan prediction system, the importance of each relevant feature can be analysed and computed. The same properties, together with their appropriate weights, can be processed for newly collected test data. It is possible to set a specific limit on the amount of time that must pass before applicants can check the status of their application. It is possible to incorporate a jumping mechanism into the forecasting system so that loans can be disbursed according to their priority. However, one potential drawback of the method being implemented in a real-world setting is that the recommendation system may have a preference for a single predominating characteristic or quality[129].

Another study's objective is to improve accuracy by performing unique loan approval forecasts using a variety of machine learning algorithms. This will be done in order to achieve the goal.On the dataset that represents the Loan Prediction Problem, machine learning methods are implemented. The dataset includes a train file and a testing file. Logistic regression, Decision tree, Random Forest, and XGBoost were the four groups that were utilised in this process. The sample size was determined to be 35 individuals for each of the groups by utilising a Gpower value of 80%.When compared with Random Forest, XGBoost, and Decision Tree, the accuracy is at its highest when loan approval prediction is done using Logistic Regression (83.24%). This is the case even if there is a statistically significant difference between the classifiers (p0.05). The decision tree yields the lowest level of accuracy, which is 70.34 percent[130].

In a study that offered a loan prediction system that was based on machine learning, it was suggested that the system might automatically select persons who were qualified for loans. In this research, we make a prediction about the effectiveness of various machine learning models for determining whether or not to grant a loan. These models include Logistic Regression, Decision Tree, Random Forest, Extra Trees, SVM, KNeighbors, GaussianNB, AdaBoost, and Gradient Boosting. The accuracy, recall, and f1-score of the model's performance were all taken into consideration during the analysis. The results of the experimental study indicate that the Extra Trees machine learning algorithm is superior than other techniques to machine learning, such as Logistic Regression, Decision Tree, Random Forest, SVM, KNeighbors, GaussianNB, AdaBoost, and Gradient Boosting. The study also

compares the Extra Trees algorithm to the Random Forest and the AdaBoost and Gradient Boosting machine learning methods. Following this, we use an ensemble of machine learning models that have shown to have superior accuracy in order to forecast bank loan defaulters. In addition to this, we also design desktop applications that provide user interfaces (UI). The recommended model achieves a higher level of accuracy than the model-wise strategies that perform best in terms of accuracy, such as Extra Trees, by making use of a voting classifier that is based on ensemble learning[131].

Investigating the process of loan prediction using a variety of machine learning algorithms is the primary objective of this work. The proposed method begins with the preprocessing of the data in order to clean the data, get rid of any outliers, and find the correlation between the features in order to determine which feature is the most significant. Following that, three different machine-learning algorithms, namely Logistic Regression, Decision Tree, and Random Forest, will be trained and evaluated. The originality of this study can be illustrated by contrasting three different machine-learning algorithms in an effort to locate the one that makes the most accurate forecast. The results of the experiments demonstrated that Logistic Regression is superior to the other two algorithms in terms of accuracy, precision, recall, and Area Under the Curve (AUC). The decision tree algorithms were also put through Receiver operating characteristic (ROC) testing, which revealed the capacity of Logistic Regression to forecast the state of the loan based on a variety of criteria[132].

The aim of the study is to provide detailed analysis of previous studies and to propose a predictive model for automatic loan prediction using four classification algorithms.

Exploratory data analysis is performed to obtain correlation between various features and to get insights of banking datasets[133].

In another research, the aim is to determine the status of loans by employing an algorithm called backpropagation. One dependent variable and thirteen independent variables make up the dataset that was employed. Seventy-five percent of the variables were used for data training, while the remaining quarter were used for data testing. There are two primary types of simulation experiments: one simulates all of the predictor variables, and the other simulates only those predictor variables that have a significant association with the goal variable. Both types of simulations involve the target variable. According to the results of the first primary simulation experiment, the first model's top performance metrics are as follows: 94.37% accuracy, 78.57% sensitivity, 98.25% specificity, 91.67% precision, and 84.62% F1 score. The measures of performance for the second simulation are identical to the metrics of performance for the first simulation that performed the best. The findings of this research have the potential to be utilised by financial institutions as a tool to assist in the feasibility evaluation of prospective debtors, with the end goal of minimising the amount of money lost by businesses[134].

For the purposes of this study, the datasets for both the training and the testing phases were obtained from Kaggle. The findings that were obtained from the two different datasets were compared in order to identify which algorithm might be most effectively utilised for predicting loan approval and also to ascertain which characteristics are most essential when it comes to forecasting loan approval. Accuracy, precision, recall, and the F1-score are the

several metrics of performance that were utilised in the process of defining the outcomes. The models were trained with the assistance of eight distinct methods, including the Logistic Regression methodology, the Random forest algorithm, the Decision tree algorithm, the Linear Regression algorithm, the Support Vector Machine (SVM) algorithm, the Naive Bayes algorithm, the K-means algorithm, and the K Nearest Neighbours (KNN) algorithm. The culmination of the research showed that the models produced a wide range of findings. Logistic regression achieved the highest level of accuracy across both datasets, with 83.24 percent, followed by Naive Bayes, which achieved 82.16% accuracy, and Random Forest, which achieved 77.34% accuracy[135].

A study whose is to increase the performance of loan prediction system. This study is focusing on different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, K nearest neighbors, Artificial neural network, Naive Bayes, Adaboost, and Voting classifier to predict the loan approval[136].

In this research, Machine Learning (ML) techniques are utilised to find patterns in anticipating potential loan defaulters and to extract patterns from a common loan-approved dataset. The analysis will make use of the historical data of customers, such as their ages, incomes, loan amounts, and lengths of employment, among other things. Several different machine learning methods, including Random Forest, Support Vector Machine, K-Nearest Neighbour, and Logistic Regression, were utilised in order to ascertain the maximum relevant features, also known as the characteristics that have the most influence on the outcome of the prediction. The aforementioned algorithms are analysed using the

conventional metrics, and the results are compared with one another. The algorithm known as random forest is superior in terms of accuracy[137].

This study makes use of three different machine learning methods, including Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), in order to forecast whether or not consumers would be approved for loans. Based on the findings of the experiments, one may draw the conclusion that the accuracy of the Decision Tree machine learning algorithm is higher when compared to the accuracy of the Logistic Regression and Random Forest machine learning approaches[138].

In a study that proposed three machine learning algorithms, Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF), by using real data collected from Quds Bank with a variables that cover credit restriction and regulator instructions. The algorithm has been implemented to predict the loan approval of customers and the output tested in terms of the predicted accuracy[139].

The purpose of the research presented in this paper is to address this issue and provide assistance to lending institutions in their efforts to prevent making bad loans. In contrast to earlier works, which focused only on particular aspects such as the client's salary, the current one uses the entire customer's details to conduct a background check on each and every customer. This ensures that the check is comprehensive. Models of ensemble learning, such as Random Forest, Gradient Boosting, and XGBoost, amongst others, have been utilised over the course of this work. This has been the key area of concentration. The Gradient Boosting classifier produced the highest quality output out of all the different models. The

aforementioned algorithms have been pieced together to create a voting classifier, which has allowed it to finally be constructed. This contributed to a slight improvement in the metrics such as the F1-score and the roc-auc-score[140].

The authors of a study that proposes a hybrid data mining method that consists of two phases, first cluster the eligibility of customers to be given a loan using the k-means algorithm, and second classify the loan amount using data from the clustering of eligible customers using the k-nearest neighbours algorithm. Both of these steps are included in the study. Due to the findings of this study, they were able to divide their 25 clients into two distinct groups: 10 customers were placed in the "Not Feasible" cluster, and 15 customers were placed in the "Feasible" cluster. The authors were also successful in identifying clients who filed for new loans and whose occupation was Entrepreneur, whose salary was at least IDR 5000000, whose loan guarantees included Proof of Vehicle Ownership, whose account balance was at least IDR 5000000, and whose family size was at least four (4). The findings specifically pertained to loans that were categorized as having a lower loan amount. They discovered that the data validity assessment of each input variable to the target variable reached 97.57%[142].

During the course of a study to establish whether or not certain organisations or people should be granted loans. In order to monitor performance and locate consumers who are qualified for loan approval, the Random Forest Regressor model has been applied. According to the model, banks should not only focus on attracting wealthy customers, but they should also take into account other consumer characteristics that are essential in determining whether or not to give credit and in estimating the likelihood of loan default. The study

investigates a variety of criteria for deciding whether or not to approve a loan application, including gender, educational qualification, type of employment, type of business, loan length, and marital status. In addition, the study examines the number of loans that have been granted, drawn, and refused, which offers important insights into the approval process for loans and loan forecast[143].

The primary objective of this study is to assess the efficacy of the Logistic Regression method in comparison to the Random Forest approach in terms of enhancing the precision of loan prediction. This will be done by comparing the two methods against one another. In order to gather the data, which came from a wide variety of sources, we used the findings of the current research, as well as a confidence interval of 95%, a threshold of 0.05 percent, the mean, and the standard deviation. In order to categorise the data based on a sample size of n equaling ten, the G-power tool's scaling factor is increased by 80%, and two separate kinds of algorithms are applied. The accuracy of the Random Forest approach is just 75.6 percent, however the accuracy of the Logistic Regression method is extremely high at 81.30 percent. A statistically significant gap between the two groups is considered to exist when the difference between them is smaller than 0.05, or 0.001. It illustrates that the method known as Logistic Regression seems to be more accurate when estimating the amount of a customer's loan compared to the alternative method known as Random Forest[144].

In a work that aims to provide a new approach, specifically a Social Border Collie Optimisation (SBCO)-based deep neural fuzzy network for the prediction of loan eligibility, this work's main focus is on the latter. In this approach, the box-cox transformation is applied

to the input loan data in order to generate data that is suitable for subsequent processing. The altered data make use of a wrapper-based feature selection in order to choose appropriate features that will improve the performance of the loan eligibility calculation. After the features have been selected, the naïve Bayes (NB) algorithm is modified so that it may do feature fusion. During the NB training process, the classifier uses the features of the incoming data to construct a probability index table and then categories the values. When assessing the NB classifier in this manner, the posterior probability ratio is used, and the conditional likelihood of normalisation constant with class evidence is also taken into consideration. In the end, a deep neural fuzzy network that has been trained using a customised SBCO is able to predict whether or not a borrower will be eligible for a loan. In this case, the social ski driver (SSD) algorithm and the Border Collie Optimisation (BCO) algorithm are combined to create the SBCO, which is designed to produce the most accurate result possible. The accuracy, sensitivity, and specificity parameters enable the analysis to be completed successfully. When compared to the existing methods, such as fuzzy neural network (Fuzzy NN), multiple partial least squares regression model (Multi_PLS), Instance-based Entropy fuzzy Support Vector Machine (IEFSVM), deep recurrent neural network (Deep RNN), and whale social optimisation algorithm-based deep RNN (WSOA-based Deep RNN), the newly designed method performs with the highest accuracy of 95%, sensitivity and specificity of 95.4 and 97.3%, respectively[145].

In this study, a solution to this challenge is proposed, and it involves the generation of simulated data for AI. The concept of creditworthiness will be examined via the lens of the banking industry as a case study. A loan is seen as the primary source of revenue for the

banking industry, as well as the primary source of risk. As a consequence of this, determining a customer's creditworthiness is an essential step for both the banks and the clients themselves. The authors suggest a system that is geared to lenders so that they may review credit application and consumers so that they can be aware of behaviours that can affect their credit score as a means of addressing this need. The strategy that is presented in this paper attempts to realise realistic datasets for Artificial Intelligence (called IDEA) in order to fulfil particular user requests and cater to the requirements of certain businesses. Using the datasets that are already in existence, we are going to undertake an analysis of the available literature as well as approaches for the development of conceptual models. The strategy that is being presented both pulls from and contributes to such previous research. The application that is planned for this strategy is to implement it in the banking industry with the purpose of evaluating the creditworthiness of consumers who have established financial relationships. Therefore, the use case that is now being considered is to anticipate the likelihood of borrowers defaulting on their loans. The methodology that was used to analyse certain financial datasets for the use case is outlined in the paper. Before employing IDEA to make a prediction regarding credit solvency, a validation of the datasets is carried out with the assistance of the Data Quality Index[146].

The purpose of this study is to determine the probability of approving individual loan applications by integrating machine learning (ML) models and ensemble learning methodologies. This strategy has the potential to improve the degree of precision with which competent candidates are chosen from among a group of applicants. As a consequence of this, this approach can be utilised to solve the issues with loan approval procedures that were

discussed earlier. The dramatically shortened period of time required for loan approval is a benefit that accrues to both the people applying for loans and the workers of the bank. The growth of the banking industry resulted in an increase in the number of persons requesting for loans at financial institutions. We used four distinct algorithms—namely, Random Forest, Naive Bayes, Decision Tree, and KNN—in order to improve the accuracy with which we could forecast whether or not an application individual would be granted a loan. By utilising them, we were able to get a higher accuracy of 83.73%, with the Naive Bayes algorithm emerging as the most successful option[147].

In a work that examines the factors that go into determining a person's credit score, with the goal of assisting financial organisations in determining the terms of loans that are made available to their clients, this topic is covered. The purpose of this article is to provide financial institutions with information regarding a loan prediction solution known as Seven Seas. This article addresses a variety of issues relating to the beginning of the loan process. It has been explained how an application for a loan is processed on a high level as well as an alternative scoring model for credit that uses machine learning. This article also discusses the potential size of the entire market for such a solution and highlights a number of different financial institutions that are able to start the transformation processes they need to with the help of such a disruptive technology. The size of the current market and the potential for it to adopt this technology is amazing, not only in India but also everywhere else in the world[148].

In this study, the authors have constructed a number of models by employing a variety of techniques, such as Deep Support Vector Machine (DSVM), Boosted Decision Tree (BDT),

Averaged Perceptron (AP), and Bayes Point Machine (BPM), in an effort to improve their ability to forecast individuals who will default on their payments. The repository of machine learning at the University of California, Irvine (UCI) was able to provide us with a dataset that had 30,000 different instances and 25 different features. According to the findings of our research, the DSVM is the model that outperforms the other three in its ability to forecast defaulters. The authors were of the opinion that these models may be utilised by credit risk management systems in banking and lending institutions to improve their ability to forecast customers who default on their loans[149].

The authors of this work conduct a comprehensive analysis of a number of significant academic contributions (76 publications) made over the course of the last eight years to address the issues associated with credit risk by utilising statistics, machine learning, and deep learning methodologies. To be more specific, we present a new classification approach for ML-driven credit risk algorithms and a way for assessing the effectiveness of these algorithms using public datasets. They go on to examine the issues, which include the data imbalance, the inconsistent datasets, the lack of model transparency, and the insufficient utilisation of deep learning models. The analysis of their findings demonstrates that 1) the majority of deep learning models perform better than traditional machine learning and statistical algorithms in the evaluation of credit risk, and 2) ensemble approaches provide higher accuracy compared to single models. In conclusion, we give some summary tables on the datasets and models that have been proposed[150].

This work demonstrates the benefits that artificial intelligence may bring to the process of assessing credit risk. with this particular topic, we take a look at the current state of affairs with the advancement of research. In order to manage this evaluation, the writers initially concentrated on the keywords in order to capture and examine the available publications written by specialists. They narrowed the time frame down to 2016–2021 so that they could focus on the most recent developments. Numerous approaches to feature selection, classification, and prediction have been investigated by the research community. The algorithms used in data mining, machine learning (both supervised and unsupervised), and deep learning (artificial neural networks) are highly distinct from one another and target a variety of different areas that need to be investigated. Because of these advancements, banks are now able to become more intelligent, which enables them to provide a service that is both better and more quickly, all while protecting themselves from losses caused by credit defaulters. According to the research that was done, the Support Vector Machine, Catboost, Decision Tree, and Logistic Regression all produced fascinating results[151].

In a study that presented six (6) different machine learning algorithms (Random Forest, Gradient Boost, Decision Tree, Support Vector Machine, K-Nearest Neighbour, and Logistic Regression) for predicting loan eligibility, the researchers found that the Random Forest approach performed the best overall. The models were trained using the historical dataset known as 'Loan Eligible Dataset,' which can be found on Kaggle and is distributed with a licence known as Database Contents Licence (DbCL) v1.0. Processing and analysis of the dataset were carried out in the cloud-based Jupyter Notebook environment provided by Kaggle using Python programming packages. The findings of our study demonstrated great

levels of performance accuracy, with the Logistic regression algorithm receiving the lowest score of 80% and the Random forest algorithm receiving the highest score of 95.55%. In terms of precision-recall and accuracy, our models came out on top, besting two of the three loan prediction models that could be identified in the relevant research[152].

In a study that suggested research to construct a model that would anticipate upcoming business sectors in retail banking, the researchers thought it would be helpful to develop such a model. The research made use of a variety of documents, including those pertaining to business customers of a retail bank. These records came from both the rural and the urban areas of Bangladesh. These records were utilised to analyse the primary transactional drivers of customers and to make a prediction regarding a prototype for likely subdivisions in a retail bank. The challenges that were used to develop the model were analysed with a decision tree data mining method, and then weka was used to put the model through its paces and evaluate how well it performed. The creation of a Credit Scoring Model for implementation by Sudanese financial institutions was the primary objective of the article. Decision Tree (DT) and Artificial Neural Network (ANN) were the two options for the classification of mined data that were ultimately selected. After that, the approaches of Generic Algorithm (GA) and Principal Component Analysis (PCA) were utilised in order to identify features. In order to evaluate the efficacy of the approaches, both the Sudanese credit dataset and the German credit dataset were utilised. According to the results of the categorization, ANN performs significantly better than DT in the vast majority of scenarios. When compared, it was discovered that the GA method is superior to the PCA technique in terms of the selection of features. The accuracy of the German data set came in at 80.67%, whereas the accuracy of

the Sudanese data set was 69.74%. Finally, the conclusion. It was observed that ANN performed better than DT as well as DT's hybrid models, which include PCA-DT and GA-DT[153].

In a different research, data mining was used to provide an innovative method for classifying the levels of credit risk in the banking industry. In order to accurately forecast the state of loans, the data that was used for this model came from many institutions. In order to create the projected models, three different techniques were utilised. These algorithms are j48, bayesNet, and Naive Bayesian. Weka was the application that was utilised for the implementation, and it was then put through testing. The findings of the study indicated that the j48 algorithm performed the best in terms of accuracy. The purpose of this study was to investigate how the accuracy of credit risk prediction could be enhanced by utilising classifier ensembles, as well as the expected behaviour of five different classifiers about how they would react to various types of interference. After that, the result is based on four different credit datasets, and a comparison of how each classifier performed in terms of its predicted accuracy at varying degrees of attribute noise is shown. The results of the experimental evaluation indicate that using a collection of different classifiers has the ability to increase the accuracy of predictions[154].

Another study suggested a prediction model that employed the Artificial Neural Networks (ANN) procedure of Machine Learning to accomplish loan nonpayment forecasting and liken it with the Logistic regression procedure. The authors prepared their archetypal on pre-documented data to estimate the of the debtor and they made effort to yield the greatest likely results[155].

In a work that projected a strategy that utilised vector machines for loan nonpayment forecasting, the system was anticipated to be effective. After comparing the results of the study with those of other classifiers, the researchers came to the conclusion that the support vector machine performed significantly better than many of the older approaches in terms of both throughput and arithmetical implementation in situations where large amounts of data were associated with several descriptive characteristics. The accuracy of the model predicted by the system was 81%[156].

A logistic regression model was used in a study to determine the percentage of individual loans in Kenya that were not repaid. The method of arithmetic analysis was utilised in this study, and the focus was placed on the characteristics of debtors in terms of their failure to repay individual debts. The test data precision of the prototype was 0.7333, whereas the train data precision was 0.7727. The precision of the logistic regression model with the train statistics was 0.8440, while the precision of the test statistics was 0.8244, respectively. The greatest shortcoming of this model is that it produces an excessive number of false positive results[157].

Another study revealed the findings of an exhaustive analysis and outlined a process to account for the possibility of defaulting on a loan. In order to conduct this investigation, the KDD, CRISP-DM, and SEMMA techniques were utilised. Because of its significant physiognomies regarding the estimation of loan nonpayment in the fiscal subdivision, the superior system was thoughtfully selected, clarified, and advised because it was built on

specific restrictions. The accuracy of this plan is 78%, yet it was a failure because its ROC score and zone weren't in the proper place[158].

In a separate but related study, it was suggested that the neural network method be used for the evaluation of loan default. The author provided a basis for merging a neural network method that was utilised to speculate on nonpayment loans, and they advised that this be done. The evaluation was carried out with reference to the economic and public information that was provided by the potential borrower. When calculating the likelihood of a late or missed payment[159].

In order to evaluate how well logistic regression works, a study made use of data provided by a microfinance organisation. The author made use of a number of different predictors, such as age, family status, gender, years of schooling, years of industry experience, and starting capital. The marital status of the individual, number of years spent working in the business sector, and amount of base capital were the relevant prognosticators for the procedure. This strategy was effective about 91 percent of the time; however, the most notable flaw was that the difference between the two scenarios that were predicted was not significant[160].

In a related study, an effort was made to train the archetype using machine learning; the classifier was set to LSVM, and the level of competence was tested using RMSE. The flaw was that it had a lower level of productivity if the descriptive variables were less than 10, which was its minimum threshold. In addition to that, he made an effort to instruct the model by utilising ANN. The accuracy of the Neural Network classifier was 93 percent in this instance[161].

In another research, the researchers used six different classification methods to make a default prediction. These methods included Bayesian Networks, Multilayer Perceptron, NB, LR, RF, and J48. These algorithms were compared with one another using performance metrics such as RMSE, ROC, ACC, Prec, and F-measure. Rec was also considered. Their research demonstrates that LR is the best model to utilise according to the experiments, and it was applied in order to ascertain the effect of the covariant on the variables that had a greater default factor by utilising Chi-square[162].

Another study reviewed and analysed the dataset containing bank loans using RF, LR, SVM, and various other applicable algorithms implemented in Python. Because of its high classification impact, particularly when used to bigger or more highly dimensional data sets, RF method proves to be more ideal for bank credit default prediction model based on five model effect evaluation matrices: ACC, Rec, Prec, F1-score, and ROC. These matrices are used to evaluate the effectiveness of the model[163].

This paper took data on peer-to-peer lending into consideration in order to forecast defaults on credit loans. They used the LR, DT, RF, and KNN algorithms among their collection of naive-bias algorithms. The authors were dedicated to the use of probability of default in a default point to minimise credit risk, and as a result, they achieved an accuracy of 94.6%. According to the findings of the article, the random forest model is the most appropriate one to use for the model that will be implemented[164].

Another research came up with a model for credit risk management in banking institutions, with the goals of improving prediction and standardising pre-lending evaluation. This author implements a default forecast using the RF approach, which is based on information from previously taken out loans. According to the results of the experiments described in this article, the RF approach performs better than other algorithms, such as DT and LR, when measured in terms of their ACC and REC rates[165]

In a study that was carried out, early payback was shown to be a factor that may significantly cut down on classification error, and this factor was employed in the study. On the other hand, their work only makes use of three different methods for developing models, with logistic regression being the most successful classification model. Therefore, the purpose of this work is to further investigate the usage of additional algorithms with more datasets for the purpose of forecasting the likelihood of loan default based on a list of applicants[166]

A study was conducted with the aim of devising a method to effectively discern and authenticate loan applicants. The authors accomplished this with the assistance of a machine learning technique. The method allows for the automatic selection of suitable applications based on the criteria that are currently available. The prediction was achieved with the Decision Tree algorithm. One of the primary challenges associated with this method pertains to the exclusive utilisation of a single machine learning algorithm[167].

The utilisation of a machine learning technique was employed in a study to forecast individuals who are likely to fail on their loans. The Logistic regression model was employed

in conjunction with the dataset acquired from Kaggle to make predictions. The collected results were compared in order to determine their effectiveness, utilising specificity and sensitivity as criteria. The conclusive findings indicate a higher level of performance in comparable initiatives. However, it should be acknowledged that the obtained findings do not accurately reflect the anticipated outcomes due to the utilisation of only one machine learning method, without any valid rationale[168].

A web-based application was developed for the purpose of carrying out extensive and more reliable prediction using logistic regression, which was implemented in the Python programming language. This was done as a result of the realisation that loan prediction plays an important role in the modern banking system. The system is capable of delivering findings with a high accuracy and just a minor loss when used to train or validate data. It should be emphasised, however, that the performance of the system is restricted with regard to specific characteristics, and the users cannot receive assistance beyond those restrictions[169].

In a study that used six different machine learning algorithms for the prediction of android applications, the researchers wanted to reduce the risk element that was behind selecting the safe person in order to save lots of bank work and assets. This was accomplished through the study's findings. The goals of their work were accomplished by mining the profiles of individuals who had previously been offered loans that were comparable to the one in question. The finished product is effective in comparison to other projects of a similar nature. The applicant profiles that are used to determine who is qualified for a loan are quite narrow, and as a result, they cannot possibly represent the entire population[170].

An integrated learning classification model that used Particle Swarm Optimisation (PSO) optimisation Support Vector Machine (SVM) was suggested in a study. While a prediction model was being developed, PSO was utilised in order to improve SVM, and AdaBoost was utilised in order to incorporate SVM's weak classifier. It was discovered that the AdaBoost-PSO-SVM strategy has the potential to successfully increase the level of accuracy. The relatively low number of samples that were used for the classification is the primary obstacle. For the purpose of determining whether or not a bank customer in Nigeria is eligible for credit or a loan, an improved model of the machine learning technique was constructed. A very trustworthy and usable dataset was gathered from the UCL repository in an effort to verify the efficacy of this model. This dataset was used in this endeavour. The usefulness of the concept was demonstrated by the confusion matrix, with accuracy serving as the primary criterion for measuring success[171].

2.4 Chapter Summary and Gap in Literature

The chapter was structured into four distinct sections, including conceptual review, theoretical review/framework and review of empirical works pertaining to the research topic. The conceptual review provided a comprehensive analysis of the underlying concepts explored throughout the study. The themes under discussion encompass credit worthiness, loans, the application of artificial intelligence in loan prediction, and the utilisation of machine learning techniques. The study also provided comprehensive insights into many types of machine learning and classification algorithms, including GBoost, Support Vector

Machines (SVM), Gaussian Naive Bayes (GNB), K-Nearest Neighbour (KNN), Random Forest, and Decision trees. Several performance metrics were identified and explained.

The methodology review provided a comprehensive explanation of the primary categorization algorithms employed in this study. The Gradient Boosting Classifier involves the iterative optimisation of a loss function by the sequential incorporation of fresh weak learners. The Gaussian Naive Bayes (NB) Classifier assumes that the features of the data follow a Gaussian (normal) distribution within each class. The Random Forest algorithm utilises an ensemble of decision trees to generate predictions. In this approach, many decision trees are employed, with each tree making use of various characteristics to classify data and Decision Tree, where each node represents features in a category to be classified and each subset defines a value that can be taken by the node.

The literature study encompassed a number of empirical research that focused on the application of machine learning techniques for the purpose of loan and credit worthiness classifications and predictions. The literature reveals that several empirical research studies pertaining to the subject of investigation have been conducted. Nevertheless, previous empirical investigations employing alternative algorithms have demonstrated lower levels of accuracy, lower f1 scores, and decreased precision. There is a paucity of prior research examining the utilisation of a comparative analysis including four distinct machine learning algorithms in the context of predicting the creditworthiness of borrowers. This study aims to mitigate the financial impact caused by non-performing loans by introducing four distinct machine learning models. These models are designed to predict the likelihood of loan

approval for individuals based on the assessment of specific attributes, including educational attainment, employment status, and loan repayment history, among others. The evaluated empirical studies indicate a dearth of research in the topic area, highlighting a gap in the existing literature that necessitates further investigation.

Endnotes

1. MC Aniceto, F Barboza, H Kimura. *Machine learning predictivity applied to consumer creditworthiness*. **Future Business Journal**. 2020 Dec;6(1):1-4.

2. JM Lee, N Park, W Heo. *Importance of subjective financial knowledge and perceived credit score in payday loan use*. **International Journal of Financial Studies**. 2019 Sep 17;7(3):53.

3. PK Roy, K Shaw. *A multicriteria credit scoring model for SMEs using hybrid BWM and TOPSIS*. Financial Innovation. 2021 Dec;7:1-27.

4. B Gavurova, M Dujcak, V Kovac, A Kotásková. *Determinants of successful loan application at peer-to-peer lending market*. Economics & Sociology. 2018;11(1):85-99.

5. F Isa, R Isa. *Treatment of toxic asset by deposit money banks in Nigeria: A review of literature*. **TSU-International Journal of Accounting and Finance**. 2021 Dec 15;1(1):42-50

6. F Assef, MT Steiner, PJ Neto, DG de Barros Franco. *Classification algorithms in financial application: credit risk analysis on legal entities*. IEEE Latin America Transactions. 2019 Oct;17(10):1733-40.

7. S Moradi, F Mokhatab Rafiei. *A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks*. Financial Innovation. 2019 Dec;5(1):1-27.

8. J Ismuhadi, F Santiago. *Legal protection for default debtors in online loan agreements*. InProceedings of the 2nd International Conference on Law, Social Science, Economics, and Education, ICLSSEE 2022, 16 April 2022, Semarang, Indonesia 2022 Aug 8.

9. N Vardi. *Creditworthiness assessment and other contractual duties as tools of 'responsible credit': The case of consumer loans*. InCreditworthiness and'Responsible Credit' 2022 Aug 4 (pp. 144-214). Brill Nijhoff.

10. TM Nisar, G Prabhakar, M Torchia. *Crowdfunding innovations in emerging economies: Risk and credit control in peer-to-peer lending network platforms*. Strategic Change. 2020 May;29(3):355-61.

11. JS Al Zaidanin, OJ Al Zaidanin. T*he impact of credit risk management on the financial performance of United Arab Emirates commercial banks*. **International Journal of Research in Business and Social Science** (2147-4478). 2021 May 1;10(3):303-19.

12. G Calcagnini, R Cole, G Giombini, G Grandicelli. *Hierarchy of bank loan approval and loan performance*. Economia Politica. 2018 Dec;35:935-54.

13. AZ Woldaregay, E Årsand, S Walderhaug, D Albers, L Mamykina, T Botsis, G Hartvigsen. *Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabete*s. Artificial intelligence in medicine. 2019 Jul 1;98:109-34.

14. E Sudarmaji, NA Achsani, Y Arkeman, I Fahmi. *Credit-worthiness prediction in energy-saving finance using machine learning model*. Studies of Applied Economics. 2021 Oct 18;39(10).

15. A Motwani, G Bajaj, S Mohane. *Predictive modelling for credit risk detection using ensemble method*. **International Journal of Computer Sciences and Engineering**. 2018 Jun;6(6):863-7.

16. P Ghavami. *Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing*. Walter de Gruyter GmbH & Co KG; 2019 Dec 16.

17. KG Al-Hashedi, P Magalingam. *Financial loan detection applying data mining techniques: A comprehensive review from 2009 to 2019*. Computer Science Review. 2021 May 1;40:100402.

18. LX Liu, S Liu, M Sathye. *Predicting bank failures: a synthesis of literature and directions for future research*. **Journal of Risk and Financial Management**. 2021 Oct 8;14(10):474.

19. D Putz, M Gumhalter, H Auer. *A novel approach to multi-horizon wind power forecasting based on deep neural architecture*. Renewable Energy. 2021 Nov 1;178:494-505.

20. A Maheshwari, N Davendralingam, DA DeLaurentis. *A comparative study of machine learning techniques for aviation applications*. In2018 Aviation Technology, Integration, and Operations Conference 2018 (p. 3980).

21. H Sadok, F Sakka, ME El Maknouzi. A*rtificial intelligence and bank credit analysis: A review*. Cogent Economics & Finance. 2022 Dec 31;10(1):2023262.

22. M Anand, A Velu, P Whig. *Prediction of loan behaviour with machine learning models for secure banking*. **Journal of Computer Science and Engineering (JCSE).** 2022 Feb 15;3(1):1-3.

23. N Kshetri. *The role of artificial intelligence in promoting financial inclusion in developing countries*. **Journal of Global Information Technology Management**. 2021 Jan 2;24(1):1-6.

24. P Jindal, J Kaur. *Artificial Intelligence Applications for Lending and NPA Management*. In2021 Asian Conference on Innovation in Technology (ASIANCON) 2021 Aug 27 (pp. 1-6). IEEE.

25. TM Maddox, JS Rumsfeld, PR Payne. *Questions for artificial intelligence in health care*. Jama. 2019 Jan 1;321(1):31-2.

26. RM Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster; 2021 Aug 17.

27. J Alzubi, A Nayyar, A Kumar. *Machine learning from theory to algorithms: an overview*. **InJournal of physics: conference series** 2018 Nov (Vol. 1142, p. 012012). IOP Publishing.

28. S. Sah, "Machine Learning: *A Review of Learning Types*," ResearchGate, no. July, 2020, doi: 10.20944/preprints202007.0230.v1.

29. R Gupta, S Tanwar, S Tyagi, N Kumar. *Machine learning models for secure data analytics: A taxonomy and threat model*. Computer Communications. 2020 Mar 1;153:406-40.

30. H Nozari, ME Sadeghi. *Artificial intelligence and Machine Learning for Real-world problems (A survey)*. **International Journal of Innovation in Engineering**. 2021 Oct 7;1(3):38-47.

31. A Shakarami, M Ghobaei-Arani, A Shahidinejad. *A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective*. Computer Networks. 2020 Dec 9;182:107496.

32. IH Sarker. Machine learning: *Algorithms, real-world applications and research directions*. SN computer science. 2021 May;2(3):160.

33. W Asfaw. *Addis Ababa Institute of Technology School of Electrical and Computer Engineering Telecommunication Engineering Graduate Program* (Doctoral dissertation, Addis Ababa University Addis Ababa).

34. DA Otchere, TO Ganat, R Gholami, S Ridha. *Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative*

analysis of ANN and SVM models. **Journal of Petroleum Science and Engineering**. 2021 May 1;200:108182.

35. K Maharana, S Mondal, B Nemade. *A review: Data pre-processing and data augmentation techniques*. Global Transitions Proceedings. 2022 Jun 1;3(1):91-9.

36. YC Lo, SE Rensi, W Torng, RB Altman. *Machine learning in chemoinformatics and drug discovery*. Drug discovery today. 2018 Aug 1;23(8):1538-46.

37. IH Sarker. Machine learning: *Algorithms, real-world applications and research directions*. SN computer science. 2021 May;2(3):160.

38. SF Sabbeh. *Machine-learning techniques for customer retention: A comparative study*. International Journal of advanced computer Science and applications. 2018;9(2).

39. O Almomani, MA Almaiah, A Alsaaidah, S Smadi, AH Mohammad, A Althunibat. *Machine learning classifiers for network intrusion detection system: comparative study*. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 440-445). IEEE.

40. Y Singh, PK Bhatia, O Sangwan. *A review of studies on machine learning techniques*. **International Journal of Computer Science and Security**. 2007 Jun;1(1):70-84.

41. JA Sidey-Gibbons, CJ Sidey-Gibbons. *Machine learning in medicine: a practical introduction*. BMC medical research methodology. 2019 Dec;19:1-8.

42. M Enayati, O Bozorg-Haddad, M Pourgholam-Amiji, B Zolghadr-Asli, M Tahmasebi Nasab. *Decision tree (DT): a valuable tool for water resources engineering*. InComputational Intelligence for Water and Environmental Sciences 2022 Jul 9 (pp. 201-223). Singapore: Springer Nature Singapore.

43. O Sagi, L Rokach. *Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2018 Jul;8(4):e1249.

44. EY Boateng, DA Abaye. *A review of the logistic regression model with emphasis on medical research*. **Journal of data analysis and information processing**. 2019 Sep 12;7(4):190-207.

45. P Schober, TR Vetter. *Logistic regression in medical research*. Anesthesia and analgesia. 2021 Feb;132(2):365.

46. Y Kumar, S Saini, R Payal. *Comparative analysis for loan detection using logistic regression, random forest and support vector machine*. Random Forest and Support Vector Machine (October 18, 2020). 2020 Oct 18.

47. EC Norton, BE Dowd. *Log odds and the interpretation of logit models*. Health services research. 2018 Apr;53(2):859-78.

48. MA Sheikh, AK Goel, T Kumar. *An approach for prediction of loan approval using machine learning algorithm*. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp. 490-494). IEEE.

49. M Breskvar, D Kocev, S Džeroski. *Ensembles for multi-target regression with random output selections*. Machine Learning. 2018 Nov;107:1673-709.

50. AK Mishra, SV Ramteke, P Sen, A Kumar. *Random Forest Tree Based Approach for Blast Design in Surface Mine*. Geotech. Geol. Eng., 2017, doi: 10.1007/s10706-017-0420-8.

51. U Sharma, S Saran, SM Patil. *Fake news detection using machine learning algorithms*. **International Journal of Creative Research Thoughts (IJCRT)**. 2020 Jun 6;8(6):509-18.

52. A Ouadah, L Zemmouchi-Ghomari, N Salhi. *Selecting an appropriate supervised machine learning algorithm for predictive maintenance*. **The International Journal of Advanced Manufacturing Technology**. 2022 Apr;119(7-8):4277-301.

53. IH Sarker. Machine learning: *Algorithms, real-world applications and research directions*. SN computer science. 2021 May;2(3):160.

54. SD Immaculate, MF Begam, M Floramary. *Software bug prediction using supervised machine learning algorithms*. In2019 International conference on data science and communication (IconDSC) 2019 Mar 1 (pp. 1-7). IEEE.

55. L Segurola-Gil, F Zola, X Echeberria-Barrio, R Orduna-Urrutia. *NBcoded: network attack classifiers based on Encoder and Naive Bayes model for resource limited devices*. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2021 Sep 13 (pp. 55-70). Cham: Springer International Publishing.

56. SK Singh, RW Taylor, B Pradhan, A Shirzadi, BT Pham. *Predicting sustainable arsenic mitigation using machine learning techniques*. Ecotoxicology and Environmental Safety. 2022 Mar 1;232:113271.

57. GF Yeo, D Akman, I Hudson, J Chan. *A stochastic approximation approach to fixed instance selection*. Information Sciences. 2023 May 1;628:558-79.

58. R Chan, M Rottmann, F Hüger, P Schlicht, H Gottschalk. *Application of decision rules for handling class imbalance in semantic segmentation*. arXiv preprint arXiv:1901.08394. 2019 Jan 24.

59. J Wang, P Li, R Ran, Y Che, Y Zhou. *A short-term photovoltaic power prediction model based on the gradient boost decision tree*. Applied Sciences. 2018 Apr 28;8(5):689.

60. C Qin, Y Zhang, F Bao, C Zhang, P Liu, P Liu. *XGBoost optimized by adaptive particle swarm optimization for credit scorin*g. Mathematical Problems in Engineering. 2021 Mar 23;2021:1-8.

61. VA Dev, MR Eden. *Gradient boosted decision trees for lithology classification.* InComputer aided chemical engineering 2019 Jan 1 (Vol. 47, pp. 113-118). Elsevier.

62. AA Aldino, A Saputra, A Nurkholis, S Setiawansyah. *Application of support vector machine (svm) algorithm in classification of low-cape communities in lampung timur*. Building of Informatics, Technology and Science (BITS). 2021 Dec 31;3(3):325-30.

63. WC Leong, A Bahadori, J Zhang, Z Ahmad. *Prediction of water quality index (wqi) using support vector machine (svm) and least square-support vector machine (ls-svm)*. **International Journal of River Basin Management.** 2021 Apr 3;19(2):149-56.

64. D Maulud, AM Abdulazeez. *A review on linear regression comprehensive in machine learning*. **Journal of Applied Science and Technology Trends**. 2020 Dec 31;1(4):140-7.

65. MW Berry, A Mohamed, BW Yap, editors. *Supervised and unsupervised learning for data science*. Springer Nature; 2019 Sep 4.

66. N Li, M Shepperd, Y Guo. *A systematic review of unsupervised learning techniques for software defect prediction*. Information and Software Technology. 2020 Jun 1;122:106287.

67. KK Hiran, RK Jain, K Lakhwani, R Doshi. *Machine learning: Master supervised and unsupervised learning algorithms with real examples (english edition)*. BPB Publications; 2021 Sep 16.

68. S Nielsen. *Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions*. Journal of Accounting & Organizational Change. 2022 Oct 4;18(5):811-53.

69. K Anwar, J Siddiqui, SS Saquib Sohail. *Machine learning techniques for book recommendation: an overview*. InProceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India 2019 Feb 26.

70. N Alexander, DC Alexander, F Barkhof, S Denaxas. *Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning*. BMC Medical Informatics and Decision Making. 2021 Dec;21(1):1-3.

71. S Ilbeigipour, A Albadvi, EA Noughabi. *Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making*. Informatics in Medicine Unlocked. 2022 Jan 1;32:101005.

72. S Yu, M Yang, L Wei, JS Hu, HW Tseng, TH Meen. *Combination of Self-organizing Map and k-means Methods of Clustering for Online Games Marketing*. Sensors & Materials. 2020 Aug 30;32.

73. J Yan, X Wang. *Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology*. The Plant Journal. 2022 Sep;111(6):1527-38.

74. V Rani, ST Nabi, M Kumar, A Mittal, K Kumar. *Self-supervised learning: A succinct review*. Archives of Computational Methods in Engineering. 2023 May;30(4):2761-75.

75. K Sohn, D Berthelot, N Carlini, Z Zhang, H Zhang, CA Raffel, ED Cubuk, A Kurakin, CL Li. Fixmatch: *Simplifying semi-supervised learning with consistency and confidence*. Advances in neural information processing systems. 2020;33:596-608.

76. AA Imran, D Terzopoulos. *Multi-adversarial variational autoencoder nets for simultaneous image generation and classification*. Deep Learning Applications, Volume 2. 2021:249-71.

77. G Li, X Li, Y Wang, Y Wu, D Liang, S Zhang. Pseco: *Pseudo labeling and consistency training for semi-supervised object detection*. InEuropean Conference on Computer Vision 2022 Oct 23 (pp. 457-472). Cham: Springer Nature Switzerland.

78. A Abuduweili, X Li, H Shi, CZ Xu, D Dou. *Adaptive consistency regularization for semi-supervised transfer learning*. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 6923-6932).

79. AT Azar, A Koubaa, N Ali Mohamed, HA Ibrahim, ZF Ibrahim, M Kazim, A Ammar, B Benjdira, AM Khamis, IA Hameed, G Casalino. *Drone deep reinforcement learning: A review*. Electronics. 2021 Apr 22;10(9):999.

80. N binti Ismail, WY Chong. *Robust control strategies for autonomous vehicles in varied traffic conditions*. **Journal of Sustainable Technologies and Infrastructure** Planning. 2023 Jul 8;7(3):1-6.

81. O Ogunfowora, H Najjaran. *Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization*. **Journal of Manufacturing Systems**. 2023 Oct 1;70:244-63.

82. MA Chadi, H Mousannif. *Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization*. arXiv preprint arXiv:2304.00026. 2023 Mar 31.

83. MA Samsuden, NM Diah, NA Rahman. *A review paper on implementing reinforcement learning technique in optimising games performance.* In2019 IEEE 9th international conference on system engineering and technology (ICSET) 2019 Oct 7 (pp. 258-263). IEEE.

84. H Srinath, AK Sharma, MR Akhil. *Reinforcement learning in real-world scenarios: Challenges, applications, and future directions*. **International Journal of Research in Engineering, Science and Management**. 2023 Jul 31;6(7):40-5.

85. SA Hicks, I Strümke, V Thambawita, M Hammou, MA Riegler, P Halvorsen, S Parasa. *On evaluation metrics for medical applications of artificial intelligence*. Scientific reports. 2022 Apr 8;12(1):5979.

86. R Padilla, SL Netto, EA Da Silva. *A survey on performance metrics for object-detection algorithms*. In2020 international conference on systems, signals and image processing (IWSSIP) 2020 Jul 1 (pp. 237-242). IEEE.

87. SA Hicks, I Strümke, V Thambawita, M Hammou, MA Riegler, P Halvorsen, S Parasa. *On evaluation metrics for medical applications of artificial intelligence*. Scientific reports. 2022 Apr 8;12(1):5979.

88. Ž Vujović. *Classification model evaluation metrics*. **International Journal of Advanced Computer Science and Applications**. 2021;12(6):599-606.

89. C Cao, D Chicco, MM Hoffman. *The MCC-F1 curve: a performance evaluation technique for binary classification.* arXiv preprint arXiv:2006.11278. 2020 Jun 17.

90. JH Cabot, EG Ross. *Evaluating prediction model performance*. Surgery. 2023 Sep 1;174(3):723-6.

91. GF Von Borries, AV de Castro Quadros. *ROC app: An application to understand roc curves*. **Brazilian Journal of Biometrics**. 2022 Jun 10;40(2).

92. H Gelbard-Sagiv, S Pardo, N Getter, M Guendelman, F Benninger, D Kraus, O Shriki, S Ben-Sasson. *Optimizing electrode configurations for wearable eeg seizure detection using machine learning*. Sensors. 2023 Jun 21;23(13):5805.

93. H Gelbard-Sagiv, S Pardo, N Getter, M Guendelman, F Benninger, D Kraus, O Shriki, S Ben-Sasson. *Optimizing electrode configurations for wearable eeg seizure detection using machine learning*. Sensors. 2023 Jun 21;23(13):5805.

94. P Nie, M Roccotelli, MP Fanti, Z Ming, Z Li. *Prediction of home energy consumption based on gradient boosting regression tree*. Energy Reports. 2021 Nov 1;7:1246-55.

95. K Wang, J Lu, A Liu, G Zhang, L Xiong. *Evolving gradient boost: A pruning scheme based on loss improvement ratio for learning under concept drift*. IEEE Transactions on Cybernetics. 2021 Oct 6.

96. J Dong, Q Zhang, X Huang, Q Tan, D Zha, Z Zihao. *Active ensemble learning for knowledge graph error detection*. InProceedings of the Sixteenth ACM International Conference on Web Search and Data Mining 2023 Feb 27 (pp. 877-885).

97. KV Vineetha, P Samuel. *A multinomial naïve bayes classifier for identifying actors and use cases from software requirement specification documents*. In2022 2nd International Conference on Intelligent Technologies (CONIT) 2022 Jun 24 (pp. 1-5). IEEE.

98. A Salazar, L Vergara, E Vidal. *A proxy learning curve for the bayes classifier*. Pattern Recognition. 2023 Apr 1;136:109240.

99. K Gohari, A Kazemnejad, M Mohammadi, F Eskandari, S Saberi, M Esmaieli, A Sheidaei. *A bayesian latent class extension of naive bayesian classifier and its application to the classification of gastric cancer patients*. BMC Medical Research Methodology. 2023 Dec;23(1):1-5.

100. AB Shaik, S Srinivasan. *A brief survey on random forest ensembles in classification model*. InInternational Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2 2019 (pp. 253-260). Springer Singapore.

101. I Reis, D Baron, S Shahaf. *Probabilistic random forest: A machine learning algorithm for noisy data sets*. **The Astronomical Journal**. 2018 Dec 20;157(1):16.

102. BO Yigin, O Algin, G Saygili. *Comparison of morphometric parameters in prediction of hydrocephalus using random forests*. Computers in Biology and Medicine. 2020 Jan 1;116:103547.

103. NM Abdulkareem, AM Abdulazeez. *Machine learning classification based on radom forest algorithm: A review*. **International Journal of Science and Business**. 2021;5(2):128-42.

104. LV Utkin, MS Kovalev, FP Coolen. *Imprecise weighted extensions of random forests for classification and regression*. Applied Soft Computing. 2020 Jul 1;92:106324.

105. ML Kolhe, S Tiwari, MC Trivedi & KK Mishra (Eds.). *Advances in data and information sciences: Proceedings of icdis 2019 (vol. 94)*. Springer Singapore. https://doi.org/10.1007/978-981-15-0694-9.

106. K Gajowniczek, I Grzegorczyk, T Ząbkowski, C Bajaj. *Weighted random forests to improve arrhythmia classification*. Electronics. 2020 Jan 3;9(1):99.

107. S Koley, AK Sadhu, P Mitra, B Chakraborty, C Chakraborty. *Delineation and diagnosis of brain tumors from post contrast t1-weighted mr images using rough granular computing and random forest*. Applied Soft Computing. 2016 Apr 1;41:453-65.

108. B Charbuty, A Abdulazeez. *Classification based on decision tree algorithm for machine learning*. **Journal of Applied Science and Technology Trends**. 2021 Mar 24;2(01):20-8.

109. MO Adebiyi, OO Adeoye, RO Ogundokun, JO Okesola, AA Adebiyi. *Secured loan prediction system using artificial neural network*. **Journal of Engineering Science and Technology**. 2022 Apr;17(2):0854-73.

110. CN Kumar, D Keerthana, M Kavitha, M Kalyani. *Customer loan eligibility prediction using machine learning algorithms in banking sector*. In2022 7th International Conference on Communication and Electronics Systems (ICCES) 2022 Jun 22 (pp. 1007-1012). IEEE.

111. PS Saini, A Bhatnagar, L Rani. *Loan approval prediction using machine learning: A comparative analysis of classification algorithms*. In2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) 2023 May 12 (pp. 1821-1826). IEEE

112. Y Diwate, PS Rana, PA Chavan. *Loan approval prediction using machine learning*. **International Research Journal of Modernization in Engineering Technology and Science** (2023): n. pag

113. N Pandey, R Gupta, S Uniyal, V Kumar. *Loan approval prediction using machine learning algorithms approach*. **International Journal of Innovative Research in Technology**. 2021;8(1):898-902

114. MF Akça, O Sevli. "*Predicting acceptance of the bank loan offers by using support vector machines"*. **International Advanced Researches and Engineering Journal** 6 (2022): 142-147

115. SB Babo, AM Beyene. *Bank loan classification of imbalanced dataset using machine learning approach*, 15 March 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2667057/v1]

116. Y Dasari, K Rishitha, O Gandhi. *Prediction of bank loan status using machine learning algorithms*. **International Journal of Computing and Digital Systems.** 2023 May 1;14(1):1-.DOI: http://dx.doi.org/10.12785/ijcds/140113.ISSN: 2210-142X

117. A Sharma, V Kumar. *An exploratory study-based analysis on loan prediction*. Inventive Communication and Computational Technologies: Proceedings of ICICCT 2022. 2022 Nov 14:423-33

118. S Archana, KS Divyalakshmi. *A comparison of various machine learning algorithms and deep learning algorithms for prediction of loan eligibility*.**International Journal for Research in Applied Science & Engineering Technology (IJRASET)** ISSN: 2321-9653; IC Value: 45.98; **SJ Im**pact Factor: 7.538 Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

119. CN Sujatha, A Gudipalli, B Pushyami, N Karthik, BN Sanjana. *Loan prediction using machine learning and its deployement on web application*. In2021 Innovations in Power and Advanced Computing Technologies (i-PACT) 2021 Nov 27 (pp. 1-7). IEEE.

120. H Sharma, I Tyagi, G Agarwal, D Gupta. *An exhaustive investigation on loan prediction in banks using lrd*. **International Journal of Innovative Science and Research Technology**, Volume 8, Issue 3, March – 2023.ISSN No:-2456-2165

121. S Abdullah, Z Othman, R Mohamad. "*predicting the risk of sme loan repayment using ai technology-machine learning techniques: A perspective of* Malaysian *financing institutions"*. **Journal of Advanced Research in Applied Sciences and Engineering Technology** 31, no. 2 (July 28, 2023): 320–326. Accessed August 21, 2023. http://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/article/view/2994.)

122.T Dissanayake. *A machine learning approach to predict bank loan approval* (Doctoral dissertation).2022

123.K Sriranganathan. *Bank loan approval prediction using machine learning approach: Evidence from sri lanka* (Doctoral dissertation).2022)

124.S Nalawade, S Andhe, S Parab, A Sankhe. *Loan approval prediction*.**International Research Journal of Engineering and Technology (IRJET)** e-ISSN: 2395-0056 Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072)

125.R Priscilla, T Siva, M Karthi, K Vijayakumar, R Gangadharan. *Baseline modeling for early prediction of loan approval system*. In2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) 2023 Jan 5 (pp. 1-7). IEEE

126.SK Hegde, R Hegde. *Performance analysis of machine learning algorithms for the loan prediction in the banking sector*. **AIJR Abstracts**. 2022 Oct 10:92-3.

127.AP Behera, SS Rautaray, M Pandey, MK Gourisaria. *Predictive loan approval model using logistic regression*. InSmart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 1 2021 (pp. 715-722). Springer Singapore.

128.KR Prathap, R Bhavani. *Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG Boost, Random Forest, Decision Tree)*. **Journal of Survey in Fisheries Sciences**. 2023 Mar 4;10(1S):2438-47

129.N Uddin, MK Ahamed, MA Uddin, MM Islam, S Aryal. *An ensemble machine learning based bank loan approval predictions system with a smart application*. Available at SSRN 4376481.

130.SM Fati. *Machine learning-based prediction model for loan status approval*. **Journal of Hunan University Natural Sciences**. 2021;48(10)

131.V Sharma, R Sharma. *A systematic survey of automatic loan approval system based on machine learning*. **International Journal of Security and Privacy in Pervasive Computing (IJSPPC)**. 2022 Jan 1;14(1):1-25

132.ES Nugraha, GJ Sitepu. *A backpropagation artificial neural network approach for loan status prediction*.URI: http://repository.president.ac.id/xmlui/handle/123456789/11222\

133.AA Nureni, OE Adekola. *Loan approval prediction based on machine learning approach*. **Fudma Journal of Sciences**. 2022 Jun 24;6(3):41-50.

134. MJ Kannan, AR Nithej. ML *based loan approval prediction system a novel approach*.International **Journal of Innovative Research in Computer and Communication Engineering** | e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | || Volume 11, Issue 3, March 2023 || | DOI: 10.15680/IJIRCCE.2023.1103095 |)

135. P Tumuluru, LR Burra, M Loukya, S Bhavana, HM CSaiBaba, N Sunanda. *Comparative analysis of customer loan approval prediction using machine learning algorithms*. In2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) 2022 Feb 23 (pp. 349-353). IEEE

136. J Tejaswini, TM Kavya, RD Ramya, PS Triveni, VR Maddumala. *Accurate loan approval prediction based on machine learning approach*. **Journal of Engineering Science**. 2020;11(4):523-32.

137. MJ Hamayel, MA Mohsen, M Moreb. *Improvement of personal loans granting methods in banks using machine learning methods and approaches in* Palestine. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 33-37). IEEE

138. K Gupta, B Chakrabarti, AA Ansari, SS Rautaray, M Pandey. *Loanification-loan approval classification using machine learning algorithms*. InProceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021 Apr 24.

139. EP Mandala, E Rianti, S Defit. *Classification of customer loans using hybrid data mining*. JUITA: Jurnal Informatika. 2022 May 31;10(1):45-52.)

140. EP Mandala, E Rianti, S Defit. *Classification of customer loans using hybrid data mining*. JUITA: Jurnal Informatika. 2022 May 31;10(1):45-52.)

141. D Dansana, SG Patro, BK Mishra, V Prasad, A Razak, AW Wodajo. *Analyzing the impact of loan features on bank loan prediction using r andom f orest algorithm*. Engineering Reports. 2023:e12707

142. R Vivek, R Mahaveerakannan. *Analyze the lack of accuracy in loan prediction using logistic regression compared with random forest to improve accuracy*. In2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) 2023 Apr 6 (pp. 1-5). IEEE

143. GL Infant Cyril, JP Ananth. *Deep learning based loan eligibility prediction with social border collie optimization*. Kybernetes. 2023 Aug 3;52(8):2847-67.

144. F Zampino, A Longo, M Zappatore. *A user-centered approach to create realistic datasets for ai. Case study: Creditworthiness in the banking sector*. InCEUR Workshop Proceedings 2022

145. V Viswanatha, AC Ramachandra, KN Vishwas, G Adithya. *"prediction of loan approval in banks using machine learning approach"*. **International Journal of Engineering and Management Research** 13, no. 4 (August 2, 2023): 7–19. Accessed August 21, 2023. https://ijemr.vandanapublications.com/index.php/ijemr/article/view/1318.

146. AS Kadam, SR Nikam, AA Aher, GV Shelke, AS Chandgude. *Prediction for loan approval using machine learning algorithm*. **International Research Journal of Engineering and Technology (IRJET)**. 2021 Apr;8(04).

147. A Shivanna, DP Agrawal. *Prediction of defaulters using machine learning on azure ml*. In2020 11th IEEE annual information technology, electronics and mobile communication conference (IEMCON) 2020 Nov 4 (pp. 0320-0325). IEEE

148. S Shi, R Tse, W Luo, S D'Addona, G Pau. *Machine learning-driven credit risk: A systemic review*. Neural Computing and Applications. 2022 Sep;34(17):14327-39.

149. IR Berrada, FZ Barramou, OB Alami. *A review of artificial intelligence approach for credit risk assessment*. In2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) 2022 Feb 12 (pp. 1-5). IEEE

150. UE Orji, CH Ugwuishiwu, JCN Nguemaleu, PN Ugwuanyi. *Machine learning models for predicting bank loan eligibility*. 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

151. MR Islam, MA Habib. *A data mining approach to predict prospective business sectors for lending in retail banking using decision tree*. arXiv preprint arXiv:1504.02018. 2015 Apr 8.

152. AJ Hamid, TM Ahmed. *Developing prediction model of loan risk in banks using data mining. Machine learning and applications*: **An International Journal**. 2016 Mar;3(1):1-9.

153. S Srivastava, G Saranya, A Pratap, R Agrawal, A Jain. *Loan default prediction using artificial neural networks*. **International Journal of Advanced Science and Technology**, 2020:29(6), 2761-2769. 21.

154.U Aslam, HI Tariq Aziz, A Sohail, NK Batcha. *An empirical study on loan default prediction models*. **Journal of Computational and Theoretical Nanoscience**. 2019 Aug 1;16(8):3483-8.

155.DM Obare, GG Njoroge, MM Muraya. *Analysis of individual loan defaults using logit under supervised machine learning approach*. **Asian Journal of Probability and Statistics**. 2019 May 1;3(4):1-2.

156.HI Tariq, A Sohail, U Aslam, NK Batcha. *Loan default prediction model using sample, explore, modify, model, and assess (semma)*. **Journal of Computational and Theoretical Nanoscience**. 2019 Aug 1;16(8):3489-503.

157.M Kumar, V Goel, T Jain, S Singhal, L Goel. *Neural network approach to loan default prediction*. **International Research Journal of Engineering and Technology (IRJET)**. 2018;5(4):4231-4.

158.C Kwofie, C Owusu-Ansah, C Boadi. *Predicting the probability of loan-default: An application of binary logistic regression*. **Research Journal of Mathematics and Statistics**. 2015 Nov 25;7(4):46-52.

159.M Jayadev, N Shah, R Vadlamani. *Predicting educational loan defaults: Application of machine learning and deep learning models*. IIM Bangalore Research Paper. 2019 Dec 4(601).

160.B Çığşar, D Ünal. *Comparison of data mining classification algorithms determining the default risk*. Scientific Programming. 2019 Feb 3;2019.

161.L Ying. *Research on bank credit default prediction based on data mining algorithm*. **The International Journal of Social Sciences and Humanities Invention**. 2018;5(6):4820-3.

162.V Padimi, ST Venkata, DN Devarani. *Applying machine learning techniques to maximize the performance of loan default prediction*. **Journal of Neutrosophic and Fuzzy Systems (JNFS)**. 2022;2(2):44-56.

163.Q Wu. *Real-time predictive analysis of loan risk with intelligent monitoring and machine learning technique*. In2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS) 2022 Jul 29 (pp. 852-856). IEEE.

164.AA Egwa, HA Kakudi, AA Ahmad, AM Bichi, MA Madu. *Prediction model for loan default using machine learning*. **The International Journal of Science & Technoledge**. 2022 Feb 28;10(2).

165. CK Gomathy, M Charulatha, M AAkash, M Sowjanya. *The loan prediction using machine learning*. **International Research Journal of Engineering and Technology**. 2021 Oct;8(10).

166. MA Sheikh, AK Goel, T Kumar. *An approach for prediction of loan approval using machine learning algorithm*. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp. 490-494). IEEE.

167. CN Sujatha, A Gudipalli, B Pushyami, N Karthik, BN Sanjana. *Loan prediction using machine learning and its deployement on web application*. In2021 Innovations in Power and Advanced Computing Technologies (i-PACT) 2021 Nov 27 (pp. 1-7). IEEE.

168. K Arun, G Ishan, K Sanmeet. *Loan approval prediction based on machine learning approach*. **IOSR J**. Comput. Eng.2016 May;18(3):18-21.

169. Z Zhang, B Li. *Loan prediction model based on adaboost and pso-svm*. In2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018) 2018 May (pp. 733-739). Atlantis Press

Chapter Three

Methodology

3.1 Research Approach

The research employs a supervised learning approach, focusing on training a classification model to predict the severity of road accidents. Supervised learning is well-suited for this task, as it relies on labeled datasets where the input features are used to predict specific target outcomes. By training the model on historical data with known outcomes, the model learns to identify patterns and make predictions about new, unseen cases.

Four classification algorithms are utilized in the study: Decision Tree, Gradient Boosting Classifier, Random Forest, and Gaussian Naive Bayes (NB) Classifier. Each of these models has unique strengths and is selected for their ability to handle classification tasks, making them suitable for predicting accident severity based on a variety of input features. These algorithms differ in complexity and approach, providing a balanced comparison to identify the best-performing model.

The study's approach includes dividing the dataset into training and testing subsets, ensuring that the model is exposed to different data during training and evaluation. Additionally, cross-validation techniques are employed to further ensure the robustness of the model. Cross-validation involves splitting the data into several folds, training on some folds while testing on others, which helps in minimizing overfitting and obtaining a more generalizable model.

3.2 Requirement Specification

3.2.1 Hardware Minimum Requirements: The study is conducted using a personal computer equipped with 8GB of RAM and a 2.2 GHz Intel Core i3 processor. The selected hardware is sufficient to handle the computational requirements of data processing and training models on moderately-sized datasets. The choice of this hardware ensures that the models can be trained and evaluated without experiencing significant performance delays, making it accessible for similar research applications.

3.2.2 Software Requirements: A range of software tools and libraries is essential for implementing the developed model. Integrated Development Environments (IDEs) like PyCharm, Jupyter Notebook, and Visual Studio Code provide user-friendly platforms for coding, debugging, and managing the project. These IDEs facilitate efficient code writing and execution, allowing for seamless integration of various libraries and packages.

The research also uses data manipulation and analysis libraries such as Pandas and NumPy, which are instrumental in managing dataframes, performing data cleaning, and handling large datasets. For data visualization, Matplotlib and Seaborn are employed to create insightful graphs and plots, which help in understanding data distributions and relationships among variables. The Scikit-Learn library is crucial for implementing machine learning models and evaluation metrics, simplifying the process of training models and assessing their performance. The use of Anaconda as a development environment ensures that dependencies are managed effectively, enabling smooth installation and operation of all required libraries.

3.3 Research Design

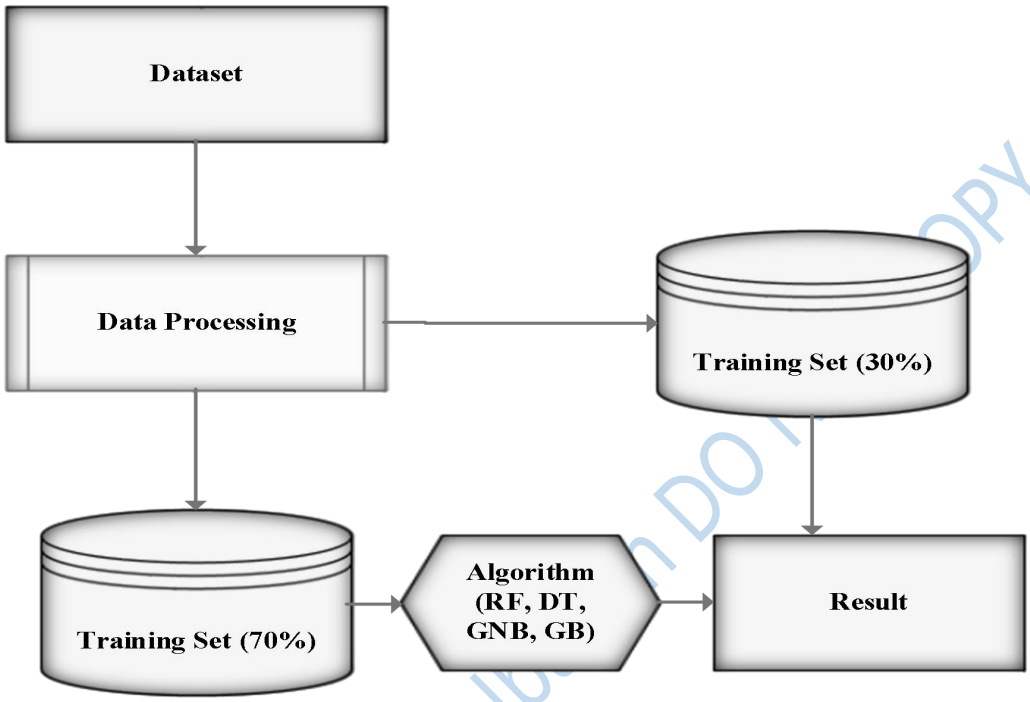


Figure 3.1: Conceptual Model of the Proposed Design

3.3.1 Data Collection: The study uses an open-source dataset containing information relevant to predicting loan defaults, sourced from publicly accessible repositories. This dataset includes multiple data files for training and testing purposes, ensuring that both phases of the machine learning pipeline are well-supported. The data was selected based on its completeness and relevance to the study objectives, covering various aspects of loan applications and repayments.

3.3.2 Dataset Details: The dataset is composed of three key segments: demographic data, performance data, and previous loan data, each of which provides critical insights into

customer behavior. The training and test datasets are structured to facilitate machine learning model development, with the target variable labeled as "good" (1) or "bad" (0), indicating whether a loan is likely to default. This binary classification allows the models to distinguish between high-risk and low-risk loan applicants effectively.

3.3.3 Dataset Description:

- *Demographic Data:* This dataset includes information like customer ID, birthdate, type of bank account, geographical coordinates (longitude and latitude), bank name and branch, employment status, and the highest level of education attained. These features provide insights into the socio-economic background of customers, which can influence their loan repayment behavior.
- *Performance Data:* This subset focuses on the repeat loans taken by customers and the likelihood of their repayment based on historical performance. It helps in assessing whether a customer who has taken a previous loan is likely to default again. This information is pivotal in evaluating the overall risk profile of the customer, taking into account their past behavior with regard to loan repayments.
- *Previous Loans Data:* This dataset records all past loans associated with each customer, with unique identifiers for each loan. It enables tracking the entire borrowing history of a customer, including the amounts borrowed, repayment timelines, and any previous defaults. This historical data is crucial for training the model to recognize patterns and predict future loan performance.

3.3.4 Data Preprocessing and Balancing

Data preprocessing is an essential step to prepare raw data for analysis, ensuring that the dataset is clean, consistent, and suitable for training machine learning models. This process involves steps such as data cleaning, where any irrelevant or erroneous records are removed to maintain the integrity of the dataset. In this study, missing data is addressed through imputation, using mean values for numerical variables and mode values for categorical variables to ensure that all data points are filled appropriately.

The data is then normalized and transformed to ensure that all features are on a similar scale, which is important for algorithms sensitive to feature scales. Scikit-learn's pipeline is used for seamless implementation of these preprocessing steps, allowing the data transformation to be carried out in a streamlined manner. Additionally, categorical variables are converted into numerical formats using dummy variables, making them suitable for model input while retaining their original information.

3.3.5 Correlation Analysis

Correlation analysis is conducted to explore the relationships between different variables in the dataset. It helps identify which features have a strong positive or negative correlation with the target variable, providing insights into which factors most significantly impact loan defaults. By understanding these relationships, the study can focus on the most influential variables during model training, potentially enhancing model accuracy.

The results of the correlation analysis are visualized using a heat map, which provides a graphical representation of the strength and direction of relationships between variables. This visualization is critical for identifying multi-collinearity among features, which can be

addressed by removing or combining highly correlated variables to improve model performance.

3.3.6 Data Splitting

The pre-processed loan dataset is randomly split into two parts: 70% of the data is used for training the model, while the remaining 30% is reserved for testing. This data splitting strategy ensures that the model has ample data to learn from while being tested on unseen data to evaluate its generalization ability. By separating the training and testing sets, the study aims to prevent overfitting and ensure that the model performs well on new data.

The random splitting method is crucial as it helps to maintain the representativeness of both the training and testing sets. This approach ensures that the testing set is a true reflection of the diversity and complexity of the training data, allowing for a more accurate assessment of the model's predictive capabilities.

3.3.7 Algorithm Used for Model Building

- **Naive Bayes (NB):** The Naive Bayes algorithm is based on Bayes' theorem and assumes independence between input features. Despite its simplicity, it is highly efficient and capable of handling large datasets. It is particularly effective when the assumption of feature independence holds true, making it a valuable baseline model in this study.
- **Random Forest (RF):** The Random Forest algorithm is an ensemble learning technique that constructs multiple decision trees during training and outputs the class that is the mode of the classifications of the individual trees. This method is robust

and can handle overfitting better than a single decision tree by averaging the outcomes of many trees, making it a powerful tool for classification tasks.

- **Decision Tree:** A Decision Tree classifier partitions the data into subsets based on the values of input features, leading to a tree-like structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome. This intuitive approach makes it easy to interpret the decision-making process of the model, though it can be prone to overfitting if not properly tuned.
- **Gradient Boosting Classifier:** Gradient Boosting builds models sequentially, focusing on correcting the errors of previous models in the sequence. It is known for its ability to improve accuracy through the gradual reduction of error, making it suitable for complex datasets. The method combines the strengths of weak learners to create a more accurate and robust predictive model.

3.4 Model Evaluation and Performance

Confusion matrix analysis is used as a primary method for evaluating the performance of each classification model. It provides detailed insights into the true positives, false positives, true negatives, and false negatives, helping to assess how well each model distinguishes between the "good" and "bad" outcomes.

Performance metrics such as precision, recall, and F1-score are employed to provide a more nuanced understanding of the model's predictive accuracy. Precision measures the accuracy of positive predictions, while recall assesses the ability to identify actual positive cases. The F1-score balances these metrics, offering a single measure of a model's effectiveness, especially in cases of class imbalance.

Additionally, ROC curves are used to evaluate the models' performance across different classification thresholds, offering insights into the trade-off between true positive and false positive rates. By considering accuracy, sensitivity, and specificity, the evaluation provides a comprehensive view of how well the models perform in predicting loan default risks, guiding the selection of the best model for practical deployment.

Endnotes

1. A Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."; 2022 Oct 4.

2. F Reiss, B Cutler, Z Eichenberger. *Natural language processing with pandas dataframes*. InProc. Of The 20th Python In Science Conf.(Scipy 2021) 2021 (pp. 49-58).

3. W.R Paczkowski. *Data Visualization: The Basics. Business Analytics*: Data Science for Business Problems. 2021:85-126.

4. C.R Harris, K.J Millman, S.J Van Der Walt, R Gommers, P Virtanen, D Cournapeau, E Wieser, J Taylor, S Berg, N.J Smith, R Kern. *Array programming with NumPy. Nature*. 2020 Sep 17;585(7825):357-62

5. G Herda, R McNabb. *Python for Smarter Cities: Comparison of Python libraries for static and interactive visualisations of large vector data*. arXiv preprint arXiv:2202.13105. 2022 Feb 26.

6. https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data

7. A.A Egwa, H Bello, A.A Ahmad, M.S Bizi. *Default Prediction for Loan Lenders Using Machine Learning Algorithms*. **SLU Journal of Science and Technology.** 2022 Dec 29;5(1&2):1-2.

8. J.C Chow. *Analysis of financial credit risk using machine learning. arXiv preprint arXiv*:1802.05326. 2018 Feb 14.

9. I.H Sarker. *Machine learning: Algorithms, real-world applications and research directions. SN computer science*. 2021 May;2(3):160.

10. X Deng, Q Liu, Y Deng, S Mahadevan. *An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences.* 2016 May 1;340:250-61.

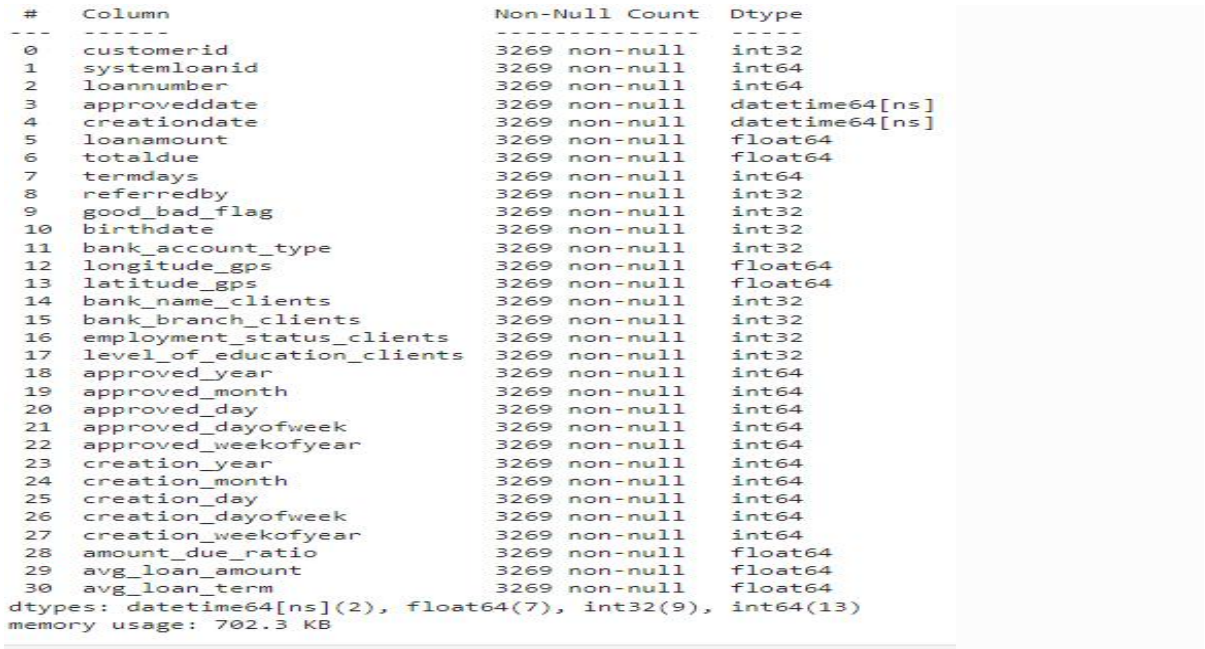
Chapter Four

Result and Discussion of Findings

This section presents results of the methodology employed based on the objectives of this study. This section presents how the dataset is processed and divided into demographic information, performance metrics, and historical borrowing data to prepare it for machine learning applications. Key preprocessing steps include which cleansing to eliminate incomplete records, maintaining data integrity essential for reliable analysis. The cleaned dataset is then used to build various predictive models. Performance is evaluated using precision, recall, f1-scores, and accuracy metrics, with a focus on how different models perform on majority and minority classes, providing valuable insights into each model's predictive capabilities and suitability for deployment in real-world scenarios.

4.1 Result on Dataset Processing

The dataset has been segmented into three distinct categories: demographic information, performance metrics, and historical borrowing records. To ensure data integrity, columns containing null entries were rigorously examined and cleansed if they failed to satisfy an established threshold for validity. This process was critical to ascertain the proportion of data points falling short of accuracy within each column. Subsequent to this refinement, Figure 3.1 presents a detailed classification of the remaining data types within the cleaned dataset, specifically pertaining to demographic details, performance statistics, and antecedent loan transactions, all of which are pivotal for the model's application.



```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   customerid                   3269 non-null   int32
 1   systemloanid                 3269 non-null   int64
 2   loannumber                   3269 non-null   int64
 3   approveddate                 3269 non-null   datetime64[ns]
 4   creationdate                 3269 non-null   datetime64[ns]
 5   loanamount                   3269 non-null   float64
 6   totaldue                     3269 non-null   float64
 7   termdays                     3269 non-null   int64
 8   referredby                   3269 non-null   int32
 9   good_bad_flag                3269 non-null   int32
 10  birthdate                    3269 non-null   int32
 11  bank_account_type            3269 non-null   int32
 12  longitude_gps                3269 non-null   float64
 13  latitude_gps                 3269 non-null   float64
 14  bank_name_clients            3269 non-null   int32
 15  bank_branch_clients          3269 non-null   int32
 16  employment_status_clients    3269 non-null   int32
 17  level_of_education_clients   3269 non-null   int32
 18  approved_year                3269 non-null   int64
 19  approved_month               3269 non-null   int64
 20  approved_day                 3269 non-null   int64
 21  approved_dayofweek           3269 non-null   int64
 22  approved_weekofyear          3269 non-null   int64
 23  creation_year                3269 non-null   int64
 24  creation_month               3269 non-null   int64
 25  creation_day                 3269 non-null   int64
 26  creation_dayofweek           3269 non-null   int64
 27  creation_weekofyear          3269 non-null   int64
 28  amount_due_ratio             3269 non-null   float64
 29  avg_loan_amount              3269 non-null   float64
 30  avg_loan_term                3269 non-null   float64
dtypes: datetime64[ns](2), float64(7), int32(9), int64(13)
memory usage: 702.3 KB
```

Figure 4.1 Columns Remaining after Data Cleaning

Research Design, 2024.

Figure 4.1 shows a screenshot of a dataframe displayed in a Python coding environment using pandas or a similar data manipulation library. This dataframe contains various columns with information about loans, customers, and their demographics. The dataset has 30 distinct columns, each representing a different attribute or feature related to loans and customers. The naming convention is straightforward and descriptive, indicating good data management practices. Each column has a 'non-null' count of 3269, which suggests that there are no missing values across the entire dataframe for the columns displayed. This is a positive sign, indicating that the dataset is complete and may not require further cleaning for missing values. The absence of null values implies that there won't be a need for imputation strategies typically required to handle missing data. The data types are consistent with what one would expect for each column (e.g., dates are in date-time format, identifiers are integers). This consistency is essential for ensuring that data types match their expected format for analysis. The columns represent potential features for predictive

modeling. For instance, factors such as loan amount, interest rates, and customer bank account flags might be used to predict loan default (good_bad_flag). With no null values and proper data types, the dataset exhibits high data integrity, which is conducive to reliable outcomes from data analysis or machine learning models.

Also, loan amount distribution, loan amount by good or bad and Educational level by good/bad to for better understanding and analysis of the data was plotted. Which shows that in terms of loan amount and educational level, the good out performed the bad

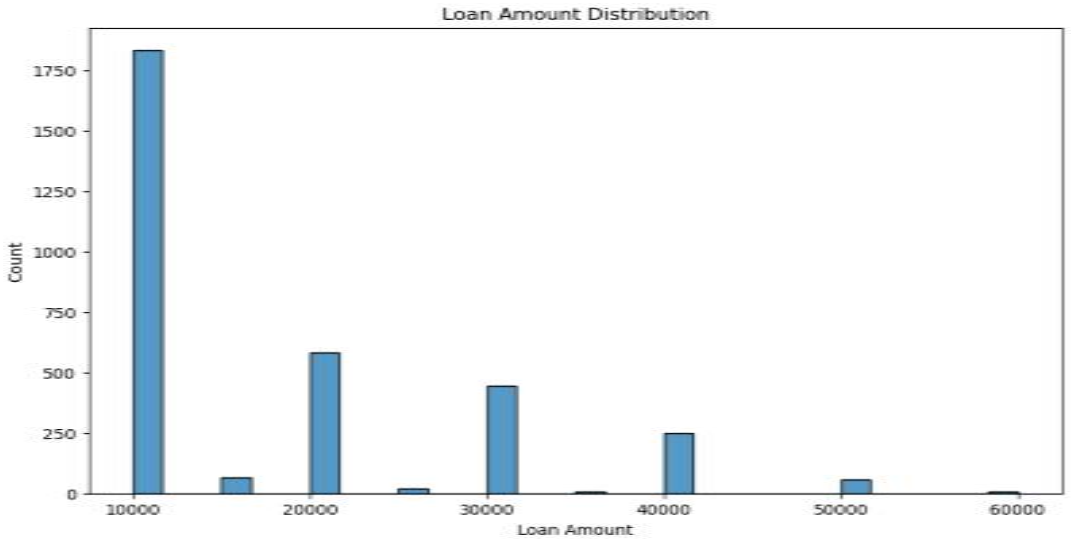


Figure 4.2: Plots Showing Loan Amount Distribution
Research Design, 2024.

Figure 4.2 shows the bar chart which plots the frequency of loans at various loan amount levels. From the chart, it is evident that the most common loan amount is in the lowest bracket shown, 10,000, which has a count significantly higher than any other amount, with over 1750 occurrences. The frequency of loans decreases as the loan amount increases, showing fewer loans distributed in the higher amounts of 20,000, 30,000, 40,000, and 50,000.

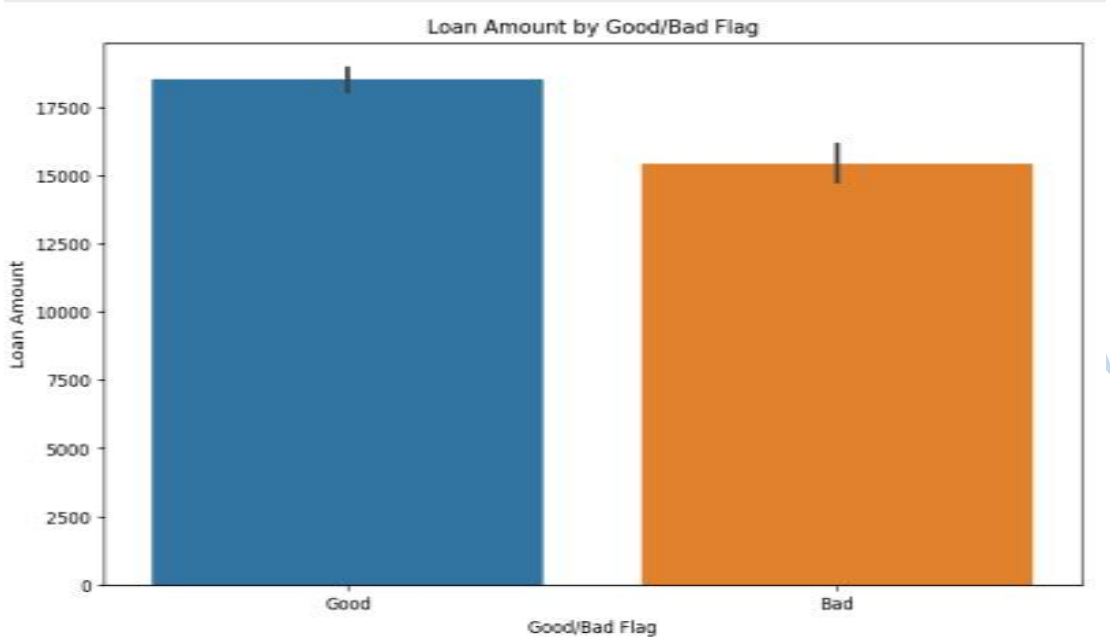


Figure 4.3: Plots Showing Loan Amount by Good/Bad
Research Design, 2024.

Figure 4.3 shows a bar chart comparing the total loan amounts categorized by the "Good/Bad Flag," which likely represents the creditworthiness or repayment history of the borrowers. 'Good' indicating reliable borrowers and 'Bad' indicating those who may have defaulted or are at risk. The blue bar, representing 'Good' borrowers, shows a higher total loan amount compared to the orange bar for 'Bad' borrowers. The presence of error bars on top of each column indicates variability in the loan amounts within each group, suggesting there is some variation in the loan sizes for both 'Good' and 'Bad' borrowers. The higher total loan amount for 'Good' borrowers suggests that the lender's strategy may favour extending more credit to individuals with a positive repayment history. The chart also imply that 'Bad' borrowers are less likely to be approved for larger loans, reflecting a risk-averse lending approach. The borrowers who are classified as 'Good' may generally be more financially stable, allowing them to take out larger loans, while 'Bad' borrowers may either apply for smaller loans or be approved for less due to their credit history.

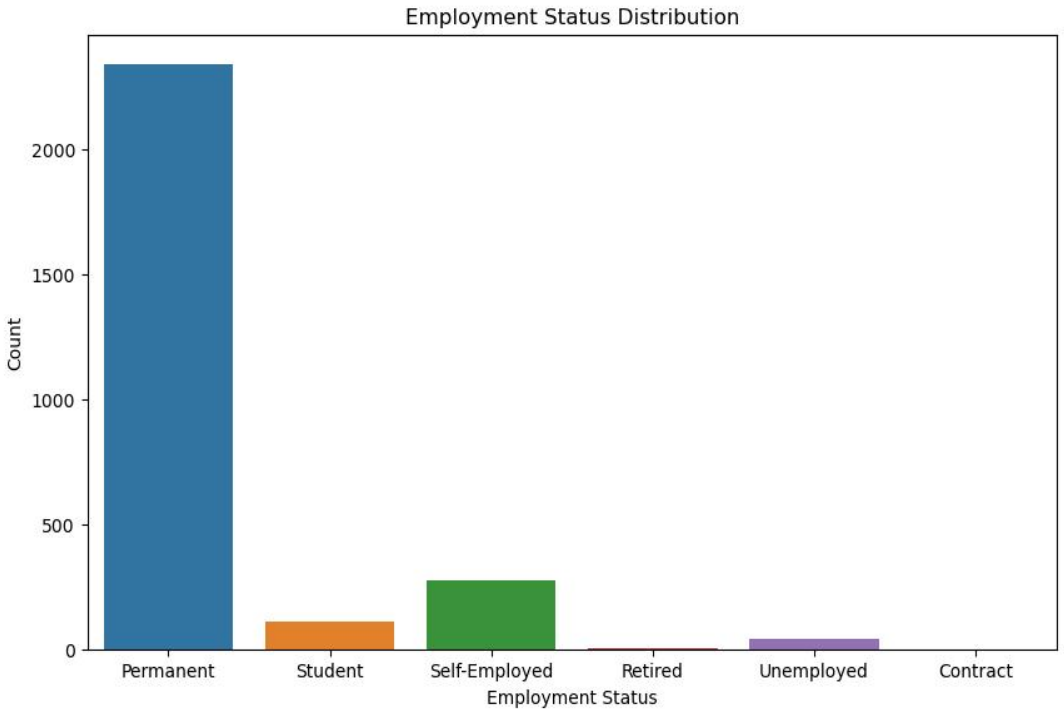


Figure 4.4: Plots Showing Educational Level By Good/Bad
Research Design, 2024.

Figure 4.4 depicts a bar chart titled which presents the count of individuals across various employment categories. These categories include Permanent, Student, Self-Employed, Retired, Unemployed, and Contract. From the chart, the 'Permanent' category has the highest count by a significant margin, indicating that the majority of individuals in this dataset are permanently employed. The counts for 'Student', 'Self-Employed', 'Retired', 'Unemployed', and 'Contract' are substantially lower, with 'Students' and 'Self-Employed' being slightly more than the other categories, but still much less compared to 'Permanent'. The high count of permanently employed individuals shows a lower credit risk for lenders, as these individuals potentially have a stable income source. Also, it reflects the lender's target market, indicating a focus on individuals with permanent employment.

4.2 Model Building

Before model's training, the dataset was systematically divided to form two separate subsets. The Training set was allocated 70% of the total data; this portion is utilized to educate the model, allowing it to discern and learn the intricate patterns and relationships inherent in the data. The remaining 30% was designated as the Test set, serving as a new and unseen dataset for the model to make predictions on. This approach allows for the validation of the model's performance in a realistic scenario, closely simulating how it would perform when deployed in a real-world environment. To ensure a thorough evaluation of the model's predictive prowess, a suite of performance metrics was employed. The Accuracy metric would offer a straightforward proportion of correct predictions over the total predictions made. The F1 Score would provide a balance between precision and recall, especially useful in situations where an equal importance to false positives and false negatives is given. Precision would measure the model's accuracy in terms of the proportion of positive identifications that were actually correct, while the Confusion Matrix would offer a detailed breakdown of the model's predictions, showcasing the true positives, true negatives, false positives, and false negatives. For the construction of the predictive model, four distinct algorithms were selected, each with its own strengths and approaches to learning from data. The use of multiple algorithms is a strategic decision, designed to compare and contrast the different models' abilities to generalize from the training data and accurately predict the outcomes on the test data. This multi-algorithm approach not only enhances the robustness of the model evaluation but also assists in identifying the most effective algorithm that aligns with the data characteristics and the predictive goals of the project.

4.2.1 Decision Tree

A variable for the decision tree classifier was established, and the essential libraries were brought in from Scikit-Learn. Following this, adjustments were made to the data to prepare it for the decision tree model's learning and prediction processes, which would be applied to both the training and test datasets.

Table 4.1: Classification Report of Decision Tree

	Precision	Recall	f1-score	Support
0	0.25	0.26	0.26	152
1	0.77	0.77	0.77	502
Accuracy			0.65	654
Macro avg	0.51	0.52	0.51	654
Weighed avg	0.65	0.65	0.65	654

Research Design, 2024

Table 4.1 shows the classification report of the performance of a decision tree classifier on a dataset with two classes, labeled '0' and '1'.

Class 0 Performance

Precision (0.25): Out of all instances predicted as class 0, only 25% were actually class 0. This indicates a high number of false positives.

Recall (0.26): Of all actual class 0 instances, the model correctly identified 26% of them. A low recall indicates many false negatives.

F1-Score (0.26): The F1-score combines precision and recall into a single metric. A score of 0.26 suggests poor performance for class 0.

Support (152): This is the actual number of occurrences of class 0 in the dataset. The model had 152 instances to learn from.

Class 1 Performance

Precision (0.77): This is considerably higher for class 1, with 77% of the predicted class 1 instances being correct.

Recall (0.77): Similarly, the model correctly identified 77% of the actual class 1 instances.

F1-Score (0.77): A much better F1-score for class 1 indicates a more balanced precision and recall, suggesting good performance.

Support (502): Class 1 had a larger representation in the dataset with 502 instances, which could have contributed to the model's better performance on this class.

Overall Model Performance

Accuracy (0.65): The model correctly predicted 65% of the total instances. While not outstanding, this suggests moderate performance.

Macro Average (Precision: 0.51, Recall: 0.52, F1-score: 0.51): Macro averaging treats all classes equally, giving a better measure of the true performance across the class imbalance. The scores around 0.51 indicate mediocre performance across both classes.

Weighted Average (Precision: 0.65, Recall: 0.65, F1-score: 0.65): The weighted average takes class imbalance into account. Given that class 1 has a higher support, the better metrics of class 1 heavily influence these averages, pushing them to 0.65.

The decision tree model performed well with the majority class (class 1) but poorly with the minority class (class 0). The significant class imbalance (with class 1 having more than three times the instances of class 0) is likely affecting the model's ability to predict class 0 accurately. Given the imbalance, the accuracy might not be the best stand-alone metric. The F1-scores, especially the macro average, give a clearer picture of the model's limitations.

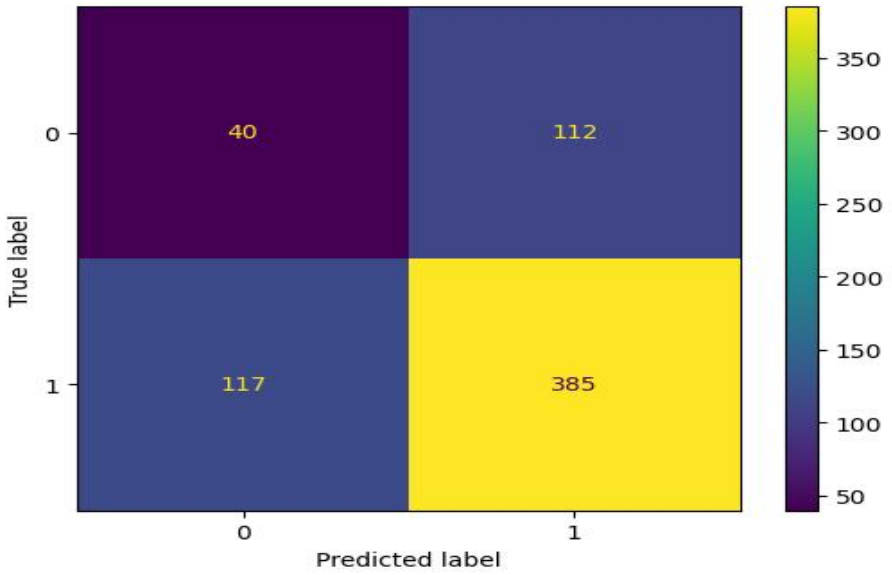


Figure 4.5: Confusion Matrix of Decision Tree
Research Design, 2024

Figure 4.5 shows the confusion matrix of decision tree classifier, to measure the performance of a classification model. The matrix displays the actual versus predicted classifications that a model has made.

True Positive (TP): The yellow square (bottom right) shows the number 385, indicating that the model correctly predicted the positive class ('1') 385 times.

True Negative (TN): The purple square (top left) with the number 40 shows that the model correctly predicted the negative class ('0') 40 times.

False Positive (FP): The purple square (top right), with the number 112, shows the instances where the model incorrectly predicted the positive class ('1') when it was actually the negative class ('0'). False Negative (FN): The blue square (bottom left), showing the number 117, represents the instances where the model incorrectly predicted the negative class ('0') when it was actually the positive class ('1').

The model has a higher number of true positives and true negatives than false positives and false negatives, which generally indicates a model that is performing reasonably well. However, the number of false negatives is close to the number of true negatives, which could be a concern depending on the cost or risk associated with a false negative in the specific application for this model. The relatively high number of false positives suggests that the model may be over-predicting the positive class.

4.2.2 Gradient Boosting Classifier

Table 4.2: Classification Report of Gradient Boosting Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

Research Design, 2024

Table 4.2 gives the classification report for the Gradient Boosting Classifier.

Performance on Class 0 (Typically the 'negative' class)

Precision (0.41): When the model predicts class 0, it is correct 41% of the time.

Recall (0.06): The model correctly identifies only 6% of all actual class 0 instances.

F1-Score (0.10): This low F1-score indicates a poor balance between precision and recall for class 0. It suggests the model is not performing well on the minority class.

Performance on Class 1 (Typically the 'positive' class)

Precision (0.77): The model's predictions for class 1 are correct 77% of the time.

Recall (0.97): The model identifies 97% of all actual class 1 instances, which is very high.

F1-Score (0.85): This score is robust for class 1, suggesting the model is quite good at predicting this class.

Overall Model Performance

Accuracy (0.76): Across both classes, the model accurately predicts 76% of the instances.

Macro Average (Precision: 0.59, Recall: 0.52, F1-score: 0.48): These averages treat both classes equally. The low recall and F1-score suggest the model is not performing equally well across both classes.

Weighted Average (Precision: 0.69, Recall: 0.76, F1-score: 0.69): These metrics account for the support of each class. The weighted scores are skewed towards class 1 due to its larger support, reflecting better overall performance due to the model's effectiveness in predicting the majority class.

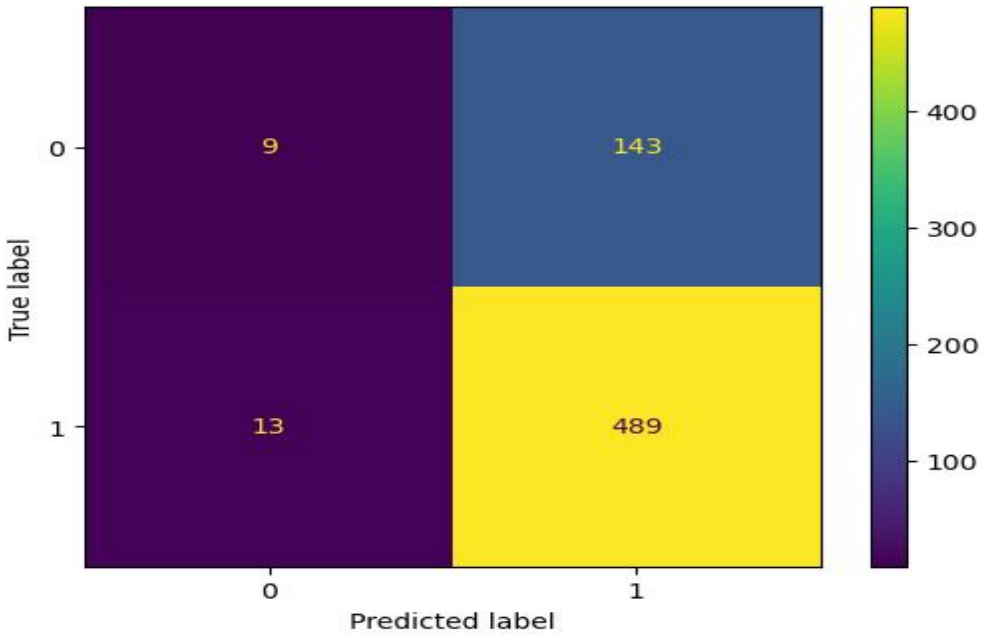


Figure 4.6: Confusion Matrix of Gradient Boosting Classifier
Research Design, 2024

Figure 4.6 shows The confusion matrix for the Gradient Boosting Classifier. The model predicted the negative class '0' correctly 9 times (True Negatives), but it incorrectly predicted the positive class '1' as '0' on 13 occasions (False Negatives). For the positive class '1', the model predicted correctly 489 times (True Positives) and incorrectly predicted the negative class '0' as '1' 143 times (False Positives). The diagonal from the top left to the bottom right shows the correct predictions by the model, with the larger numbers indicating the model's tendency to predict class '1' correctly more often than class '0'. The small number of True Negatives compared to False Negatives suggests that the model has difficulty identifying the negative class. This is consistent with the earlier classification report, which indicated a low recall for class '0'. The high number of True Positives and low number of False Negatives for class '1' indicates that the model is much better at predicting the positive class. The model's strong bias towards predicting class '1' indicate an imbalance in the dataset or in the model's ability to distinguish between the classes.

4.2.3 Random Forest Classifier

Table 4.3: Classification Report of Random Forest Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

Research Design, 2024

Table 4.3 shows the classification report for the Random Forest Classifier

Class 0 (Potentially the less frequent or 'negative' class)

Precision (0.41): This suggests that when the model predicts an instance to be class 0, it is correct about 41% of the time.

Recall (0.06): The model correctly identifies 6% of the actual class 0 instances indicating a very high miss rate for this class.

F1-Score (0.10): The harmonic mean of precision and recall is quite low for class 0, reflecting the poor performance of the model on this class.

Class 1 (Potentially the more frequent or 'positive' class)

Precision (0.77): When the model predicts class 1, it is correct 77% of the time.

Recall (0.97): The model successfully identifies 97% of all actual instances of class 1, which is quite high.

F1-Score (0.85): A strong F1-score for class 1 indicates good model performance for this class.

Overall Model Performance

Accuracy (0.76): The overall accuracy of the model is 76%, meaning it correctly predicts 76% of the time when considering both classes.

Precision (0.59): This average precision considers both classes equally and is moderately low, suggesting imbalanced class performance.

Recall (0.52): The macro average for recall is also moderately low, heavily impacted by the poor recall for class 0.

F1-score (0.48): This score is quite low, indicating poor overall performance across both classes when weighted equally.

Weighted Average

Precision (0.69): This suggests better performance, but it is influenced by the larger class 1.

Recall (0.76) & F1-score (0.69):These reflect the weighted contributions of each class to the overall metric, skewed by the better-performing class 1.

The Random Forest model is significantly better at predicting class 1 than class 0, which may be due to class imbalance or features that do not distinguish well between the two classes. The low recall for class 0 is concerning as it indicates the model misses many actual instances of class 0. The precision and recall are substantially higher for class 1, showing a clear model bias toward the majority class, which often occurs in imbalanced datasets.

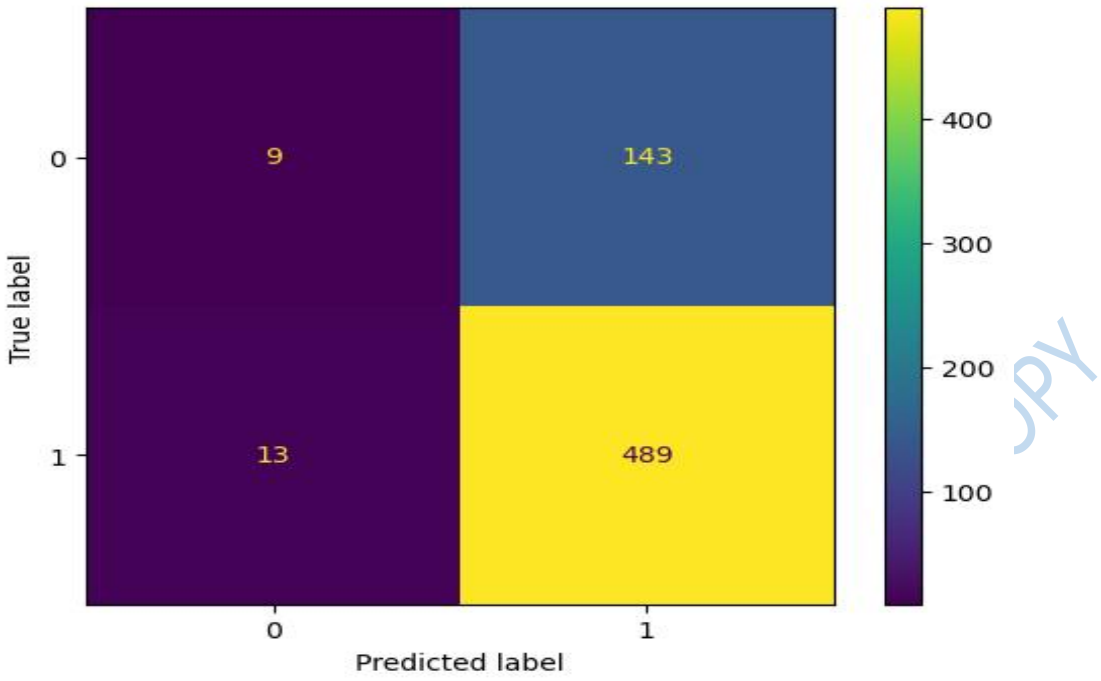


Figure 4.7: Confusion Matrix of Random Forest Classifier.

Figure 4.7 shows the confusion matrix for the Random Forest Classifier with the number of correct and incorrect predictions made by the model. There are 9 true negatives, indicating that the model correctly predicted the negative class '0' nine times. There are 489 true positives, where the model correctly predicted the positive class '1'. However, there are 13 false negatives, meaning the model incorrectly predicted the negative class when it was actually the positive class, and 143 false positives, where the model incorrectly predicted the positive class when it was actually the negative class. The model is substantially better at predicting the positive class than the negative class.

4.2.4 Gaussian Naive Beyes

Table 4.4 : Classification Report of Gaussian NB Classifier

	Precision	Recall	f1-score	Support
0	0.41	0.06	0.10	152
1	0.77	0.97	0.85	502
Accuracy			0.76	654
Macro avg	0.59	0.52	0.48	654
Weighed avg	0.69	0.76	0.69	654

Research Design, 2024

Table 4.4 presents the classification report for the Gaussian Naive Bayes (NB).

Class 0 Metrics

Precision (0.41): This indicates that when the classifier predicts an instance to be in class 0, it is correct about 41% of the time.

Recall (0.06): The classifier only correctly identifies 6% of all actual class 0 instances, which is very low and suggests that most of the class 0 instances are being missed.

F1-score (0.10): The F1-score for class 0 is quite low, which means the balance between precision and recall is poor. This is not ideal, especially if class 0 is critical to identify correctly.

Class 1 Metrics

Precision (0.77): Suggests a relatively high precision, meaning the classifier's predictions for class 1 are correct 77% of the time.

Recall (0.97): The classifier identifies 97% of all actual class 1 instances, which is excellent.

F1-score (0.85): A high F1-score for class 1 reflects a good balance between precision and recall for this class.

Overall Performance

Accuracy (0.76): The model has an overall accuracy of 76%, which indicates that it correctly predicts the class for 76% of the instances across both classes.

Macro Average

Precision (0.59): An average of the precision for both classes, moderately low, indicating imbalanced performance.

Recall (0.52): Also moderately low on average, skewed by the very low recall for class 0.

F1-score (0.48): Similarly, this score reflects the poor performance across both classes when equally considered.

Weighted Average

Precision (0.69), Recall (0.76), F1-score (0.69): These weighted metrics consider the number of instances in each class, thus favoring class 1 due to its larger number of instances.

The Gaussian NB classifier is showing a significant discrepancy in performance between the two classes, with a strong bias towards class 1, likely influenced by the class imbalance. The very low recall for class 0 implies that the classifier is not effective at identifying instances of this class, which could lead to a high number of missed critical cases if class 0 represents an important condition such as a disease in medical diagnosis. Although the

overall accuracy is relatively high, the low macro averages highlight the model's inability to perform equally well for both classes.

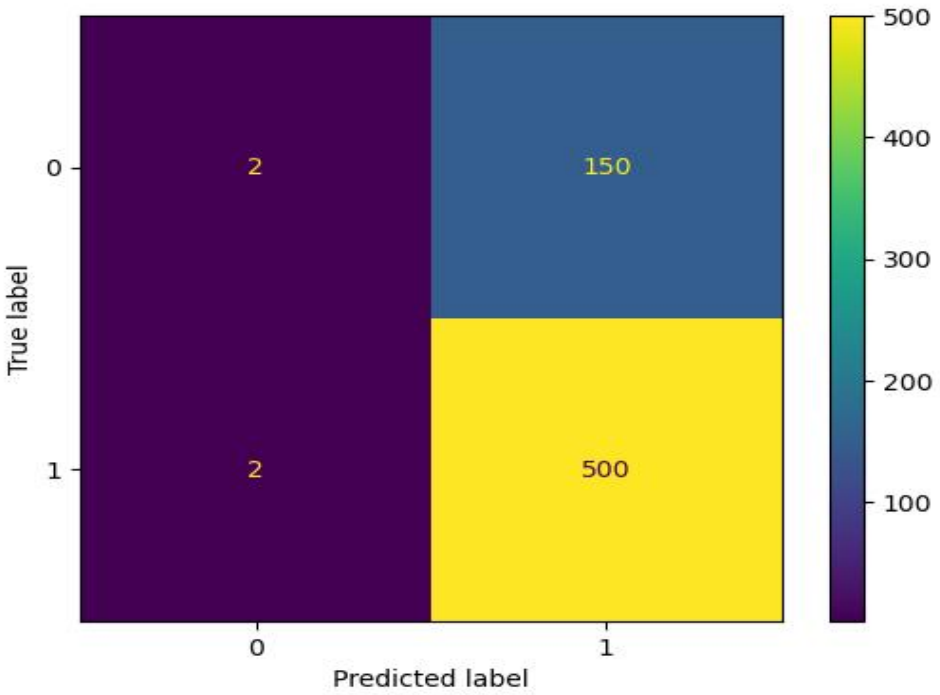


Figure 4.8: Confusion Matrix of Gaussian NB Classifier
Research Design, 2024

The confusion matrix for the Gaussian Naive Bayes Classifier shows that the model has correctly predicted class 0 (True Negatives) only 2 times and class 1 (True Positives) 500 times. It has incorrectly predicted class 0 as class 1 (False Positives) 150 times, and class 1 as class 0 (False Negatives) 2 times. This suggests the model is highly effective at identifying class 1 instances but struggles significantly with class 0, failing to identify the majority of actual class 0 instances correctly. The disproportionally small number of correct predictions for class 0 indicates a possible bias towards class 1.

Table 4.5 : Classification Report of the Models

Model	Score
GaussianNB	0.787584
Random Forest Classifier	0.762997
Gradient Boosting Classifier	0.762997
Decision Tree Classifier	0.647615

Research Design, 2024

The classification report in Table 4.5 present the overall accuracy, of four different models. The Gaussian Naive Bayes (NB) Classifier with a score of approximately 0.788, the Gaussian NB model has the highest accuracy among the four models. This suggests that, despite its simplicity, the Gaussian NB classifier is best at generalizing from the training data to the test data for this particular dataset. Random Forest Classifier has an accuracy score of around 0.763. This ensemble method, which typically performs well on a wide range of classification tasks due to its capacity for reducing overfitting, is slightly less accurate than the Gaussian NB model for this dataset.

Gradient Boosting Classifier also with a score of 0.763, it performs equivalently to the Random Forest model. Gradient Boosting is another ensemble method that focuses on learning from the errors of previous trees. Its performance being similar to the Random Forest suggests that both ensemble methods are benefiting similarly from the dataset's characteristics. Decision Tree Classifier with a score of approximately 0.648 is significantly lower than the other models. As a single decision tree, it's more prone to overfitting and generally less accurate on unseen data compared to ensemble methods. The simplest model, Gaussian NB, outperforms the more complex ensemble models in this

case. This could suggest that the dataset's features have a relationship that aligns with the conditional independence assumption of Naive Bayes. The equivalent scores of the Random Forest and Gradient Boosting models suggest that they are similarly effective for this dataset.

4.3 Discussion of Findings

The findings from this research highlight several insights into data processing and model performance. The analysis of the methodology and the results derived from employing various machine learning models dataset provides a comprehensive insight into the strengths and limitations of each model with respect to the loan prediction data.

The dataset, segmented into demographic information, performance metrics, and historical borrowing records, was cleansed to ensure data integrity, a crucial step for reliable predictive modeling. The preprocessing steps eliminated incomplete records and validated the remaining data, ensuring a robust foundation for building predictive models. In evaluating the models, a range of metrics including precision, recall, F1-scores, and accuracy were used. These metrics highlighted how different models handled the predictive tasks, particularly how they performed across majority and minority classes, shedding light on their potential real-world applicability.

Among the models evaluated, the Decision Tree Classifier showed moderate overall performance with an accuracy of 65%. It performed notably better on the majority class compared to the minority class, suggesting a bias or an overfitting issue towards the majority class due to the class imbalance. The ensemble methods, Gradient Boosting and

Random Forest, displayed better performance with an accuracy of 76%, and a notable strength in handling the majority class but again showed weaknesses in dealing with the minority class. This trend underscores the challenge of class imbalance affecting the model's ability to generalize across different classes. The Gaussian Naive Bayes Classifier stood out with the highest accuracy of approximately 78.8%, indicating its efficiency despite the simplicity of the model. This could suggest that the assumptions of feature independence in Naive Bayes align well with the characteristics of this particular dataset. Despite its higher accuracy, the Gaussian Naive Bayes also demonstrated a similar challenge in accurately predicting the minority class, underscoring a common issue across all models tested. The confusion matrices provided further depth to the analysis, illustrating the actual versus predicted classifications and revealing specific areas where each model faltered, particularly in terms of false negatives and false positives.

These matrices were instrumental in understanding the practical implications of deploying these models, such as the potential risks associated with misclassifications.

Comparing the results of this study, a study opposes this finding that reported the use of three different machine learning methods, including Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), in order to forecast whether or not consumers would be approved for loans. Based on the findings, the accuracy of the Decision Tree machine learning algorithm is higher when compared to the accuracy of the Logistic Regression and Random Forest machine learning approaches[1]. Another study used different machine learning methods, including Random Forest, Support Vector Machine, K-Nearest Neighbour, and Logistic Regression. The random forest outperformed others in terms

accuracy[2]. Another study's used Logistic regression, Decision tree, Random Forest, and XGBoost. Result showed that Logistic Regression (83.24%) out performed When compared with Random Forest, XGBoost, and Decision Tree, the accuracy is at its highest when loan approval prediction is done using[3]. Another study presented eight distinct methods, including the Logistic Regression methodology, the Random forest algorithm, the Decision tree algorithm, the Linear Regression algorithm, the Support Vector Machine (SVM) algorithm, the Naive Bayes algorithm, the K-means algorithm, and the K Nearest Neighbours (KNN) algorithm. Logistic regression achieved the highest level of accuracy across both datasets, with 83.24 percent, followed by Naive Bayes, which achieved 82.16% accuracy, and Random Forest, which achieved 77.34% accuracy[4].

Endnotes

1. J. Tejaswini, T.M. Kavya, R.D. Ramya, P.S. Triveni, V.R. Maddumala. *Accurate loan approval prediction based on machine learning approach.* **Journal of Engineering Science**. 2020;11(4):523-32.

2. P. Tumuluru, L.R. Burra, M. Loukya, S. Bhavana, H.M. CSaiBaba, N. Sunanda. *Comparative analysis of customer loan approval prediction using machine learning algorithms*. In2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) 2022 Feb 23 (pp. 349-353). IEEE.

3. K.R. Prathap, R. Bhavani. *Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree)*. **Journal of Survey in Fisheries Sciences**. 2023 Mar 4;10(1S):2438-47.

4. J. Tejaswini, T.M. Kavya, R.D. Ramya, P.S. Triveni, V.R. Maddumala. *Accurate loan approval prediction based on machine learning approach*. **Journal of Engineering Science**. 2020;11(4):523-3.

Chapter Five

Conclusion

This chapter presents a summary of the findings of the research, conclusion, recommendations, contributions to knowledge, and areas for additional research.

5.1 Summary of Findings

This study aims to use a Machine Learning-Based approach to predict loan default. Using Zindi data which is divided in to three (3); demographic data, performance data and previous loan data, research delve into the thorough data processing and modeling techniques utilized for loan approval prediction. The comprehensive analysis of the dataset, which was carefully segmented into demographic information, performance metrics, and historical borrowing records, revealed significant insights regarding the performance of various machine learning models. The dataset underwent cleaning to eliminate incomplete records, ensuring a high degree of data integrity crucial for the subsequent analysis. The key findings from the model evaluations highlight several important aspects. Firstly, the data preparation demonstrated a high level of completeness with no missing values across all columns, facilitating accurate modeling without the need for imputation strategies. This thorough preparation allowed for the effective application of machine learning techniques.

Regarding model performance, the Decision Tree Classifier achieved moderate accuracy of 65% and displayed better performance on the majority class compared to the minority class, which might indicate potential issues of class bias or overfitting. The ensemble methods, namely the Random Forest and Gradient Boosting models, reported higher accuracy at 76%,

showcasing their robustness. However, these models, similar to the Decision Tree, also performed poorly on the minority class, suggesting difficulties in handling class imbalance. The Gaussian Naive Bayes Classifier stood out by achieving the highest accuracy at approximately 78.8%, despite its simplicity. This performance suggests that the assumptions of feature independence in Naive Bayes might align well with the dataset's characteristics.

A recurring theme across all models was the notable discrepancy in performance between the majority and minority classes, with a strong bias towards the majority class. This emphasizes the impact of class imbalance on model accuracy and the generalizability of predictions, highlighting a critical area for improvement in model training and application. The findings also underscore the implications for deploying these models in real-world scenarios. While the models can effectively predict outcomes for the majority class, their performance on the minority class could lead to potential risks, especially in scenarios where the minority class represents important but less frequent outcomes.

5.2 Conclusion

This study successfully utilised Machine Learning based approach to predict loan default. The aim was to enhance credit risk assessment and predict the likelihood of a borrower failing to meet their payment obligations. The study's comprehensive approach to data preparation and model evaluation has yielded a deep understanding of the performance characteristics of various machine learning models when applied to a rigorously cleansed dataset. This loan default prediction data, was segmented into demographic information, performance metrics, and historical borrowing records, was shown to have high integrity, which is vital for reliable machine learning applications. The absence of missing values due

to thorough data cleansing allowed for straightforward application of machine learning techniques without the complication of data imputation.

The performance analysis of the models Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes revealed that while each model has strengths, they also exhibit significant limitations, particularly in handling class imbalance. The Decision Tree model demonstrated moderate accuracy and highlighted potential overfitting issues, as it performed significantly better on the majority class. Similarly, both ensemble methods, Random Forest and Gradient Boosting, although robust with a higher accuracy of 76%, struggled with the minority class, suggesting that even sophisticated models can falter without strategies to address class imbalance. The Gaussian Naive Bayes model emerged as the top performer with the highest accuracy, suggesting that its underlying assumptions might be particularly well-suited to the dataset's features. However, like the other models, it also showed a bias towards the majority class, indicating a common challenge across the board.

5.3 Recommendations

Based on the findings the following recommendations were made;

i. Implement techniques such as synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling (ADASYN), or other resampling methods to balance the dataset. This can help improve model performance on minority classes and ensure more equitable predictions.

ii. Beyond standard accuracy, it is recommended to incorporate additional evaluation metrics like the Area Under the Curve (AUC) or the Matthew's Correlation Coefficient

(MCC). These metrics can provide a more understanding of model performance, especially in scenarios with imbalanced classes.

iii. Also, modify the learning algorithms to make them cost-sensitive, where misclassifications of the minority class are penalized more heavily than those of the majority class. This approach can help in reducing the bias towards the majority class and focus the model on reducing more critical errors.

iv. Given that ensemble methods like Random Forest and Gradient Boosting showed robust performance, further exploration into more sophisticated ensemble techniques such as stacking or blending might yield better results. These methods can leverage the strengths of multiple predictive models to improve overall performance.

v. It is also recommended to establish a system for continuous monitoring of the model's performance once deployed. Regular updates to the model based on new data and feedback can help in maintaining its relevance and accuracy over time. Additionally, monitoring can quickly identify any shifts in data patterns or performance degradation, allowing for timely adjustments.

5.4 Contribution to Knowledge

This study's contributes to knowledge in the following ways:

i. By ensuring high data integrity through meticulous data cleansing and preparation, the study underscores the importance of data quality in predictive modeling. This contributes to a deeper understanding of the link between data preparation and model performance, reinforcing best practices in data science.

ii. The research highlights the pervasive issue of class imbalance and its impact on model performance, providing a foundation for further studies into effective strategies for managing this challenge. The findings promote a broader application of techniques such as resampling and cost-sensitive learning in machine learning projects.

iii. By employing a variety of metrics (accuracy, precision, recall, F1-score, and confusion matrices), the study contributes to the methodology of model evaluation, advocating for a more comprehensive approach that goes beyond simple accuracy measures. This helps in building more reliable and interpretable models in diverse applications.

iv. The study's exploration of ensemble methods enriches the dialogue around their application, especially in handling datasets with imbalanced classes.

5.5 Suggested Area of Further Studies

Building upon the findings of this research, several areas can be identified for further studies to enhance the understanding and efficacy of machine learning models in handling real-world data complexities.

i. Investigate more sophisticated approaches beyond traditional resampling methods, such as generative adversarial networks (GANs) for generating synthetic data, or advanced anomaly detection techniques that focus specifically on minority class characteristics.

ii. Explore the development of hybrid ensemble models that combine multiple machine learning techniques to optimize performance. Research could focus on how different models complement each other and whether combining models like decision trees with neural networks can yield better performance on imbalanced datasets.

iii. Further explore the practical applications and outcomes of cost-sensitive learning in diverse industries such as finance, healthcare, and public services, where the costs of misclassification can be particularly high.

iv. Further studies can be done to undertake comparative studies to evaluate how the findings from this research apply across different domains with varying data characteristics. This could help in understanding the domain-specific challenges and effectiveness of machine learning models.

v. Deepen the research into the impact of different dimensions of data quality (such as accuracy, completeness, consistency, and timeliness) on the performance of predictive models.

Bibliography

Internatioal Conferences

Abuduweili A, Li X, Shi H, Xu CZ, Dou D. Adaptive consistency regularization for semi-supervised transfer learning. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 6923-6932).

Almomani O, Almaiah MA, Alsaaidah A, Smadi S, Mohammad AH, Althunibat A. Machine learning classifiers for network intrusion detection system: comparative study. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 440-445). IEEE.

Anwar K, Siddiqui J, Saquib Sohail S. Machine learning techniques for book recommendation: an overview. InProceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India 2019 Feb 26.

Behera AP, Rautaray SS, Pandey M, Gourisaria MK. Predictive Loan Approval Model Using Logistic Regression. InSmart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 1 2021 (pp. 715-722). Springer Singapore.

Berrada IR, Barramou FZ, Alami OB. A review of Artificial Intelligence approach for credit risk assessment. In2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) 2022 Feb 12 (pp. 1-5). IEEE

Dong J, Zhang Q, Huang X, Tan Q, Zha D, Zihao Z. Active ensemble learning for knowledge graph error detection. InProceedings of the Sixteenth ACM International Conference on Web Search and Data Mining 2023 Feb 27 (pp. 877-885).

Gupta K, Chakrabarti B, Ansari AA, Rautaray SS, Pandey M. *Loanification-loan approval classification using machine learning algorithms*. InProceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021 Apr 24.

Hamayel MJ, Mohsen MA, Moreb M. *Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine*. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 33-37). IEEE

Immaculate SD, Begam MF, Floramary M. Software bug prediction using supervised machine learning algorithms. In2019 International conference on data science and communication (IconDSC) 2019 Mar 1 (pp. 1-7). IEEE.

Ismuhadi J, Santiago F. *Legal protection for default debtors in online loan agreements*. InProceedings of the 2nd International Conference on Law, Social Science, Economics, and Education, ICLSSEE 2022, 16 April 2022, Semarang, Indonesia 2022 Aug 8.

Jindal P, Kaur J. Artificial Intelligence Applications for Lending and NPA Management. In2021 Asian Conference on Innovation in Technology (ASIANCON) 2021 Aug 27 (pp. 1-6). IEEE.

Kolhe, M.L., Tiwari, S., Trivedi, M.C., Mishra, K.K. (Eds.). *Advances in Data and Information Sciences:* Proceedings of ICDIS 2019 (Vol. 94). Springer Singapore. https://doi.org/10.1007/978- 981-15-0694-9

Kumar CN, Keerthana D, Kavitha M, Kalyani M. *Customer loan eligibility prediction using machine learning algorithms in banking sector*. In2022 7th International Conference on Communication and Electronics Systems (ICCES) 2022 Jun 22 (pp. 1007-1012). IEEE.

Li G, Li X, Wang Y, Wu Y, Liang D, Zhang S. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. InEuropean Conference on Computer Vision 2022 Oct 23 (pp. 457-472). Cham: Springer Nature Switzerland.

Maheshwari A, Davendralingam N, DeLaurentis DA. A comparative study of machine learning techniques for aviation applications. In2018 Aviation Technology, Integration, and Operations Conference 2018 (p. 3980).

Padilla R, Netto SL, Da Silva EA. A survey on performance metrics for object-detection algorithms. In2020 international conference on systems, signals and image processing (IWSSIP) 2020 Jul 1 (pp. 237-242). IEEE.

Priscilla R, Siva T, Karthi M, Vijayakumar K, Gangadharan R. Baseline Modeling for Early Prediction of Loan Approval System. In2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) 2023 Jan 5 (pp. 1-7). IEEE

Reiss F, Cutler B, Eichenberger Z. *Natural language processing with pandas dataframes*. InProc. Of The 20th Python In Science Conf.(Scipy 2021) 2021 (pp. 49-58)

Saini PS, Bhatnagar A, Rani L. Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms. In2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) 2023 May 12 (pp. 1821-1826). IEEE

Samsuden MA, Diah NM, Rahman NA. A review paper on implementing reinforcement learning technique in optimising games performance. In2019 IEEE 9th international conference on system engineering and technology (ICSET) 2019 Oct 7 (pp. 258-263). IEEE.

Segurola-Gil L, Zola F, Echeberria-Barrio X, Orduna-Urrutia R. NBcoded: network attack classifiers based on Encoder and Naive Bayes model for resource limited devices. InJoint

European Conference on Machine Learning and Knowledge Discovery in Databases 2021 Sep 13 (pp. 55-70). Cham: Springer International Publishing

Shaik AB, Srinivasan S. *A brief survey on random forest ensembles in classification model*. InInternational Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2 2019 (pp. 253-260). Springer Singapore.

Sharma A, Kumar V. An Exploratory Study-Based Analysis on Loan Prediction. Inventive Communication and Computational Technologies: Proceedings of ICICCT 2022. 2022 Nov 14:423-33

Sheikh MA, Goel AK & Kumar T. An approach for prediction of loan approval using machine learning algorithm. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp. 490-494). IEEE.

Shivanna A, Agrawal DP. Prediction of defaulters using machine learning on Azure ML. In2020 11th IEEE annual information technology, electronics and mobile communication conference (IEMCON) 2020 Nov 4 (pp. 0320-0325). IEEE

Sujatha CN, Gudipalli A, Pushyami B, Karthik N, Sanjana BN. Loan Prediction Using Machine Learning and Its Deployement On Web Application. In2021 Innovations in Power and Advanced Computing Technologies (i-PACT) 2021 Nov 27 (pp. 1-7). IEEE.

Tumuluru P, Burra LR, Loukya M, Bhavana S, SaiBaba HM & Sunanda N. Comparative analysis of customer loan approval prediction using machine learning algorithms. In2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) 2022 Feb 23 (pp. 349-353). IEEE

U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi, "*Machine learning models for predicting bank loan eligibility*," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

Vineetha KV, Samuel P. A Multinomial Naïve Bayes Classifier for identifying Actors and Use C*ases from Software Requirement Specification documents*. In2022 2nd International Conference on Intelligent Technologies (CONIT) 2022 Jun 24 (pp. 1-5). IEEE.

Vivek R, Mahaveerakannan R. Analyze the Lack of Accuracy in Loan Prediction using Logistic Regression Compared with Random Forest to Improve Accuracy. In2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) 2023 Apr 6 (pp. 1-5). IEEE

Wu Q. *Real-time predictive analysis of loan risk with intelligent monitoring and machine learning technique*. In2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS) 2022 Jul 29 (pp. 852-856). IEEE.

Zampino F, Longo A, Zappatore M. *A user-centered approach to create realistic datasets for AI. Case study: Creditworthiness in the banking sector.* InCEUR Workshop Proceedings 2022

Zhang T, Li B. *Loan prediction model based on AdaBoost and PSO-SVM*. In2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018) 2018 May (pp. 733-739). Atlantis Press.

Journals

Abdulkareem, N.M., & Abdulazeez, A.M. *Machine learning classification based on Radom Forest Algorithm: A review*. I**nternational Journal of Science and Business.** 2021;5(2):128-42.

Adebiyi MO, Adeoye OO, Ogundokun RO, Okesola JO, Adebiyi AA. *Secured loan prediction system using artificial neural network*. **Journal of Engineering Science and Technology**. 2022 Apr;17(2):0854-73.

Ahmed AI, Rajaleximi PR. *An empirical study on credit scoring and credit scorecard for financialinstitutions*. **International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)**. 2019 Jul;8(7):2278-1323.

Akça, M. F. , Sevli, O. "*Predicting acceptance of the bank loan offers by using support vector machines"*. **International Advanced Researches and Engineering Journal** 6 2022: 142-147

Al Zaidanin JS, Al Zaidanin OJ. *The impact of credit risk management on the financial performance of United Arab Emirates commercial banks*. **International Journal of Research in Business and Social Science (2147-4478)**. 2021 May 1;10(3):303-19.

Alaba OB, Taiwo EO, Abass OA. *Data mining algorithm for development of a predictive model for mitigating loan risk in Nigerian banks*. **Journal of Applied Sciences and Environmental Management**. Dec 28;25(9): 2021 1613-6

Alaka HA, Oyedele LO, Owolabi HA, Kumar V, Ajayi SO, Akinade OO, Bilal M. *Systematic review of bankruptcy prediction models: Towards a framework for tool selection.* Expert Systems with Applications. 2018 Mar 15;94:164-84.

Aldino AA, Saputra A, Nurkholis A, Setiawansyah S. *Application of support vector machine (svm) algorithm in classification of low-cape communities in Lampung Timur*. Building of Informatics, Technology and Science (BITS). 2021 Dec 31;3(3):325-30.

Alexander N, Alexander DC, Barkhof F, Denaxas S. *Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning*. BMC Medical Informatics and Decision Making. 2021 Dec;21(1):1-3.

Al-Hashedi KG, Magalingam P. Financial loan detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review. 2021 May 1;40:100402.

Alzubi J, Nayyar A, Kumar A. *Machine learning from theory to algorithms: an overview*. **InJournal of physics: conference series** 2018 Nov (Vol. 1142, p. 012012). IOP Publishing.

Anand M, Velu A, Whig P. *Prediction of loan behaviour with machine learning models for secure banking.* **Journal of Computer Science and Engineering (JCSE)**. 2022 Feb 15;3(1):1-3.

Aniceto MC, Barboza F, Kimura H. *Machine learning predictivity applied to consumer creditworthiness*. **Future Business Journal**. 2020 Dec;6(1):1-4

Archana S, Divyalakshmi KS. *A comparison of various machine learning algorithms and deep learning algorithms for prediction of loan eligibility*. **International Journal for Research in Applied Science & Engineering Technology (IJRASET)** ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Arun K, Ishan G, Sanmeet K. L*oan approval prediction based on machine learning approach*. **IOSR J. Comput**. Eng. 2016 May;18(3):18-21.

Aslam U, Tariq Aziz HI, Sohail A, Batcha NK. *An empirical study on loan default prediction models*. J**ournal of Computational and Theoretical Nanoscience**. 2019 Aug 1;16(8):3483-8.

Assef F, Steiner MT, Neto PJ, de Barros Franco DG. *Classification algorithms in financial application: credit risk analysis on legal entities*. IEEE Latin America Transactions. 2019 Oct;17(10):1733-40.

Awuza AE, Habeebah KA, Ahmad AA, Abubakar MB & Muhammad AM. *Prediction model for loan default using machine learning*. **The International Journal of Science & Technoledge**. DOI No.: 10.24940/theijst/2022/v10/i2/ST2202-009.2022

Azar AT, Koubaa A, Ali Mohamed N, Ibrahim HA, Ibrahim ZF, Kazim M, Ammar A, Benjdira B, Khamis AM, Hameed IA, Casalino G. *Drone deep reinforcement learning*: *A review*. Electronics. 2021 Apr 22;10(9):999.

Begum C & Deniz U. *Comparison of data mining classification algorithms: Determining the default risk*. Research Article, Hindawi Scientific Programming Volume 2019, Article ID 8706505, 8 pages https://doi.org/10.1155/2019/8706505.2019

Berry MW, Mohamed A, Yap BW, editors. *Supervised and unsupervised learning for data science*. Springer Nature; 2019 Sep 4.

Binti Ismail N, Chong WY. *Robust control strategies for autonomous vehicles in varied traffic conditions.* **Journal of Sustainable Technologies and Infrastructure Planning**. 2023 Jul 8;7(3):1-6.

Boateng EY, Abaye DA. *A review of the logistic regression model with emphasis on medical research*. **Journal of data analysis and information processing**. 2019 Sep 12;7(4):190-207.

Breskvar M, Kocev D, Džeroski S. *Ensembles for multi-target regression with random output selections*. Machine Learning. 2018 Nov;107:1673-709.

Cabot JH, Ross EG. E*valuating prediction model performance.* Surgery. 2023 Sep 1;174(3):723-6.

Calcagnini G, Cole R, Giombini G, Grandicelli G. *Hierarchy of bank loan approval and loan performance*. Economia Politica. 2018 Dec;35:935-54.

Cao C, Chicco D, Hoffman MM. *The MCC-F1 curve: a performance evaluation technique for binary classification*. arXiv preprint arXiv:2006.11278. 2020 Jun 17.

Carta S, Ferreira A, Recupero DR, Saia M, Saia R. *A combined entropy-based approach for a proactive credit scoring*. Engineering Applications of Artificial Intelligence. 2020 Jan 1;87:103292.

Chadi MA, Mousannif H. Understanding *Reinforcement Learning Algorithms: The progress from basic q-learning to proximal policy optimization*. arXiv preprint arXiv:2304.00026. 2023 Mar 31.

Chan R, Rottmann M, Hüger F, Schlicht P, Gottschalk H. *Application of decision rules for handling class imbalance in semantic segmentation*. arXiv preprint arXiv:1901.08394. 2019 Jan 24.

Charbuty, B., Abdulazeez, A. *Classification based on decision tree algorithm for machine learning*. **Journal of Applied Science and Technology Trends**. 2021 Mar 24;2(01):20-28.

Chow JC. *Analysis of financial credit risk using machine learning. arXiv preprint arXiv*:1802.05326. 2018 Feb 14.

Çığşar B, Ünal D. *Comparison of data mining classification algorithms determining the default risk*. Scientific Programming. 2019 Feb 3;2019.

Dansana D, Patro SG, Mishra BK, Prasad V, Razak A, Wodajo AW. *Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm*. Engineering Reports. 2023:e12707

Dasari Y, Rishitha K, Gandhi O. *Prediction of bank loan status using machine learning algorithms*. International **Journal of Computing and Digital Systems. 2023** May 1;14(1):1-.**DOI:** http://dx.doi.org/10.12785/ijcds/140113.**ISSN:** 2210-142X

Deng X, Liu Q, Deng Y, Mahadevan S. *An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences.* 2016 May 1;340:250-61.

Dev VA, Eden MR. *Gradient boosted decision trees for lithology classification*. InComputer aided chemical engineering 2019 Jan 1 (Vol. 47, pp. 113-118). Elsevier.

Diwate, Yash, Prashant Singh Rana & Pratiksha Ashok Chavan. "*Loan approval prediction using machine learning.*" *International Research Journal of Modernization in Engineering Technology and Science* (2023): n. pag

Djeundje VB, Crook J. *Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards*. **European Journal of Operational Research**. 2018 Dec 1;271(2):697-709.

Egwa AA, Bello H, Ahmad AA, Bizi MS. *Default prediction for loan lenders using machine learning algorithms*. **SLU Journal of Science and Technology**. 2022 Dec 29;5(1&2):1-2.

Egwa AA, Kakudi HA, Ahmad AA, Bichi AM, Madu MA. Prediction model for loan default using machine learning. **The International Journal of Science & Technoledge**. 2022 Feb 28;10(2).

Enayati M, Bozorg-Haddad O, Pourgholam-Amiji M, Zolghadr-Asli B, Tahmasebi Nasab M. *Decision tree (DT): a valuable tool for water resources engineering*. InComputational Intelligence for Water and Environmental Sciences 2022 Jul 9 (pp. 201-223). Singapore: Springer Nature Singapore.

Fati SM. *Machine learning-based prediction model for loan status approval*. **Journal of Hunan University Natural Sciences**. 2021;48(10)

Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T., Bajaj, C. *Weighted random forests to improve arrhythmia classification. Electronic*s. 2020 Jan 3;9(1):99.

Gavurova B, Dujcak M, Kovac V, Kotásková *A. Determinants of successful loan application at peer-to-peer lending market.* Economics & Sociology. 2018;11(1):85-99.

Gelbard-Sagiv H, Pardo S, Getter N, Guendelman M, Benninger F, Kraus D, Shriki O, Ben-Sasson S. *Optimizing electrode configurations for wearable eeg seizure detection using machine learning*. Sensors. 2023 Jun 21;23(13):5805.

Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."; 2022 Oct 4.

Ghavami P. *Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing*. Walter de Gruyter GmbH & Co KG; 2019 Dec 16.

Gohari K, Kazemnejad A, Mohammadi M, Eskandari F, Saberi S, Esmaieli M, Sheidaei A. *A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients.* BMC Medical Research Methodology. 2023 Dec;23(1):1-5

Gomathy CK, Charulatha M, AAkash M, Sowjanya M. *The loan prediction using machine learning*. **International Research Journal of Engineering and Technology**. 2021 Oct;8(10)

Gulsoy N, Kulluk S. *A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019 May;9(3):e1299.

Guo Y, He J, Xu L, Liu W. *A novel multi-objective particle swarm optimization for comprehensible credit scoring*. Soft Computing. 2019 Sep 1;23:9009-23.

Gupta R, Tanwar S, Tyagi S, Kumar N. *Machine learning models for secure data analytics: A taxonomy and threat model.* Computer Communications. 2020 Mar 1;153:406-40.

Hagenauer J, Helbich M. *A comparative study of machine learning classifiers for modeling travel mode choice*. Expert Systems with Applications. 2017 Jul 15;78:273-82.

Hamid AJ, Ahmed TM. *Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications:* **An International Journal**. 2016 Mar;3(1):1-9.

Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. *Array programming with NumPy. Nature*. 2020 Sep 17;585(7825):357-62

Hegde SK, Hegde R. *Performance analysis of machine learning algorithms for the loan prediction in the banking sector*. **AIJR Abstracts**. 2022 Oct 10:92-3.

Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. *Identification of autism spectrum disorder using deep learning and the ABIDE dataset*. NeuroImage: Clinical. 2018 Jan 1;17:16-23

Herda G, McNabb R. *Python for Smarter Cities: Comparison of Python libraries for static and interactive visualisations of large vector data*. arXiv preprint arXiv:2202.13105. 2022 Feb 26

Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, Parasa S. *On evaluation metrics for medical applications of artificial intelligence*. Scientific reports. 2022 Apr 8;12(1):5979.

Hiran KK, Jain RK, Lakhwani K, Doshi R. Machine learning: Master supervised and unsupervised learning algorithms with real examples (English edition). BPB Publications; 2021 Sep 16.

Ilbeigipour S, Albadvi A, Noughabi EA. *Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making.* Informatics in Medicine Unlocked. 2022 Jan 1;32:101005.

Imran AA, Terzopoulos D. *Multi-adversarial variational autoencoder nets for simultaneous image generation and classification*. Deep Learning Applications, Volume 2. 2021:249-71.

Infant Cyril GL, Ananth JP. *Deep learning based loan eligibility prediction with Social Border Collie Optimization.* Kybernetes. 2023 Aug 3;52(8):2847-67.

Isa F, Isa R. *Treatment of toxic asset by deposit money banks in nigeria: A review of literature*. **TSU-International Journal of Accounting and Finance**. 2021 Dec 15;1(1):42-50.

Islam MR, Habib MA. *A data mining approach to predict prospective business sectors for lending in retail banking using decision tree*. arXiv preprint arXiv:1504.02018. 2015 Apr 8.

Jamali AA, Ferdousi R, Razzaghi S, Li J, Safdari R, Ebrahimie E. *DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins*. Drug discovery today. 2016 May 1;21(5):718-24.

Jayadev M, Shah N, Vadlamani R. *Predicting educational loan defaults: Application of machine learning and deep learning models.* IIM Bangalore Research Paper. 2019 Dec 4(601).

Jiang C, Wang Z & Zhoo H. *A Prediction-driven Mixture cure model and its Application in Credit Scoring*. **European Journal of operational Research**. 277 2019, pp20-31.

Kadam AS, Nikam SR, Aher AA, Shelke GV, Chandgude AS. *Prediction for loan approval using machine learning algorithm*. **International Research Journal of Engineering and Technology (IRJET)**. 2021 Apr;8(04).

Kannan MJ, Nithej AR. *ML based loan approval prediction system a novel approach*. **International Journal of Innovative Research in Computer and Communication Engineering** | e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | || Volume 11, Issue 3, March 2023 || | DOI: 10.15680/IJIRCCE.2023.1103095 |)

Koley, S., Sadhu, A.K., Mitra, P., Chakraborty, B., Chakraborty, C. *Delineation and diagnosis of brain tumors from post contrast T1-weighted MR images using rough granular computing and random forest*. Applied Soft Computing. 2016 Apr 1;41:453-465.

Kshetri N. *The role of artificial intelligence in promoting financial inclusion in developing countries*. **Journal of Global Information Technology Management**. 2021 Jan 2;24(1):1-6

Kumar M, Goel V, Jain T, Singhal S, Goel L. *Neural network approach to loan default prediction*. **International Research Journal of Engineering and Technology (IRJET)**. 2018;5(4):4231-4.

Kumar Y, Saini S, Payal R. *Comparative analysis for loan detection using logistic regression, random forest and support vector machine*. Random Forest and Support Vector Machine (October 18, 2020). 2020 Oct 18.

Kwofie C, Owusu-Ansah C, Boadi C. *Predicting the probability of loan-default: An application of binary logistic regression*. **Research Journal of Mathematics and Statistics**. 2015 Nov 25;7(4):46-52.

Lee JM, Park N, Heo W. *Importance of subjective financial knowledge and perceived credit score in payday loan use*. **International Journal of Financial Studies**. 2019 Sep 17;7(3):53.

Leong WC, Bahadori A, Zhang J, Ahmad Z. *Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)*. **International Journal of River Basin Management.** 2021 Apr 3;19(2):149-56.

Li N, Shepperd M, Guo Y. *A systematic review of unsupervised learning techniques for software defect prediction*. Information and Software Technology. 2020 Jun 1;122:106287.

Liu LX, Liu S, Sathye M. *Predicting bank failures: A synthesis of literature and directions for future research.* **Journal of Risk and Financial Management**. Oct 8; 2021 14(10):474.

Lo YC, Rensi SE, Torng W, Altman RB. *Machine learning in chemoinformatics and drug discovery*. Drug discovery today. 2018 Aug 1;23(8):1538-46.

Maddox TM, Rumsfeld JS, Payne PR. *Questions for artificial intelligence in health care*. Jama. 2019 Jan 1;321(1):31-2.

Maharana K, Mondal S, Nemade B. A review: *Data pre-processing and data augmentation techniques*. Global Transitions Proceedings. 2022 Jun 1;3(1):91-9.

Mandala EP, Rianti E, Defit S. Classification of customer loans using hybrid data mining. **JUITA: Jurnal Informatika**. 2022 May 31;10(1):45-52.)

Maulud D, Abdulazeez AM. *A review on linear regression comprehensive in machine learning*. **Journal of Applied Science and Technology Trends**. 2020 Dec 31;1(4):140-7.

Mishra AK, Ramteke SV, Sen P & Kumar A, "random forest tree based approach for blast design in surface mine," Geotech. Geol. Eng., 2017, doi: 10.1007/s10706-017-0420-8.

Monarch RM. *Human-in-the-loop machine learning: Active learning and annotation for human-centered* AI. Simon and Schuster; 2021 Aug 17.

Moradi S, Mokhatab F, Rafiei. *A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks*. Financial Innovation. 2019 Dec;5(1):1-27.

Moradi S, Mokhatab Rafiei F. A *dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks*. Financial Innovation. 2019 Dec;5(1):1-27

Motwani A, Bajaj G, Mohane S. *Predictive modelling for credit risk detection using ensemble method*. **International Journal of Computer Sciences and Engineering**. 2018 Jun;6(6):863-7.

Nalawade S, Andhe S, Parab S, Sankhe A. *Loan approval prediction*. **International Research Journal of Engineering and Technology (IRJET)** e-ISSN: 2395-0056 Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Nie P, Roccotelli M, Fanti MP, Ming Z, Li Z. *Prediction of home energy consumption based on gradient boosting regression tree.* Energy Reports. 2021 Nov 1;7:1246-55.

Nielsen S. *Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions*. **Journal of Accounting & Organizational Change**. 2022 Oct 4;18(5):811-53.

Nisar TM, Prabhakar G, Torchia M. *Crowdfunding innovations in emerging economies: Risk and credit control in peer-to-peer lending network platforms*. Strategic Change. 2020 May;29(3):355-61.

Norton EC, Dowd BE. *Log odds and the interpretation of logit models*. Health services research. 2018 Apr;53(2):859-78.

Nozari H, Sadeghi ME. *Artificial intelligence and Machine Learning for Real-world problems (A survey)*. **International Journal of Innovation in Engineering**. 2021 Oct 7;1(3):38-47.

Nugraha ES, Sitepu GJ. A *Backpropagation Artificial Neural Network Approach for Loan Status Prediction*.**URI:** http://repository.president.ac.id/xmlui/handle/123456789/11222\

Nureni AA, Adekola OE. *Loan approval prediction based on machine learning approach*. **Fudma Journal of Sciences**. 2022 Jun 24;6(3):41-50.

Obare DM, Njoroge GG, Muraya MM. *Analysis of individual loan defaults using logit under supervised machine learning approach*. **Asian Journal of Probability and Statistics**. 2019 May 1;3(4):1-2.

Ogunfowora O, Najjaran H. *Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization.* **Journal of Manufacturing Systems**. 2023 Oct 1;70:244-63.

Otchere DA, Ganat TO, Gholami R, Ridha S. *Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models*. **Journal of Petroleum Science and Engineering.** 2021 May 1;200:108182.

Ouadah A, Zemmouchi-Ghomari L, Salhi N. *Selecting an appropriate supervised machine learning algorithm for predictive maintenance.* **The International Journal of Advanced Manufacturing Technology**. 2022 Apr;119(7-8):4277-301.

Paczkowski WR. *Data Visualization: The Basics. Business Analytics*: Data Science for Business Problems. 2021:85-126.

Padimi V, Venkata ST, Devarani DN. *Applying machine learning techniques to maximize the performance of loan default prediction*. **Journal of Neutrosophic and Fuzzy Systems (JNFS)**. 2022;2(2):44-56.

Pandey N, Gupta R, Uniyal S, Kumar V. *Loan approval prediction using machine learning algorithms approach.* **International Journal of Innovative Research in Technology**. 2021;8(1):898-902

Prathap KR, Bhavani R. *Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree)*. **Journal of Survey in Fisheries Sciences**. 2023 Mar 4;10(1S):2438-47

Putz D, Gumhalter M, Auer H. *A novel approach to multi-horizon wind power forecasting based on deep neural architecture*. Renewable Energy. 2021 Nov 1;178:494-505.

Qin C, Zhang Y, Bao F, Zhang C, Liu P, Liu P. *XGBoost optimized by adaptive particle swarm optimization for credit scoring*. Mathematical Problems in Engineering. 2021 Mar 23;2021:1-8.

Rani V, Nabi ST, Kumar M, Mittal A, Kumar K. *Self-supervised learning: A succinct review*. Archives of Computational Methods in Engineering. 2023 May;30(4):2761-75.

Reis I, Baron D, Shahaf S. *Probabilistic random forest: A machine learning algorithm for noisy data sets*. **The Astronomical Journal.** 2018 Dec 20;157(1):16

Roy PK, Shaw K. *A multicriteria credit scoring model for SMEs using hybrid BWM and TOPSIS.* Financial Innovation. 2021 Dec;7:1-27.

Sabbeh SF. *Machine-learning techniques for customer retention: A comparative study*. **International Journal of advanced computer Science and applications**. 2018;9(2).

Sadok H, Sakka F, El Maknouzi ME. *Artificial intelligence and bank credit analysis: A review.* Cogent Economics & Finance. 2022 Dec 31;10(1):2023262.

Sagi O, Rokach L. *Ensemble learning: A survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018 Jul;8(4):e1249.

Sah S. Machine learning working process, “machine learning: A review of learning types,” ResearchGate, no. July, 2020, doi: 10.20944/preprints202007.0230.v1.

Salazar A, Vergara L, Vidal E. *A proxy learning curve for the Bayes classifier*. Pattern Recognition. 2023 Apr 1;136:109240.

Sarker IH. *Machine learning: Algorithms, real-world applications and research directions*. SN computer science. 2021 May;2(3):160.

Sarker IH. *Machine learning: Algorithms, real-world applications and research directions.* SN computer science. 2021 May;2(3):160.

Sarker IH. *Machine learning: Algorithms, real-world applications and research directions.* SN computer science. 2021 May;2(3):160.

Sarker IH. Machine learning: *Algorithms, real-world applications and research directions.* SN computer science. 2021 May;2(3):160.

Sarker IH. *Machine learning: Algorithms, real-world applications and research directions. SN computer science*. 2021 May;2(3):160

Schober P, Vetter TR. *Logistic regression in medical research*. Anesthesia and analgesia. 2021 Feb;132(2):365.

Shakarami A, Ghobaei-Arani M, Shahidinejad A. *A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective.* Computer Networks. 2020 Dec 9;182:107496.

Sharma H, Tyagi I, Agarwal G, Gupta D. *An exhaustive investigation on loan prediction in banks using lrd*. **International Journal of Innovative Science and Research Technology, Volume 8, Issue 3**, March – 2023.ISSN No:-2456-2165

Sharma U, Saran S, Patil SM. *Fake news detection using machine learning algorithms*. **International Journal of Creative Research Thoughts (IJCRT)**. 2020 Jun 6;8(6):509-18.

Sharma V, Sharma R. *A systematic survey of automatic loan approval system based on machine learning*. **International Journal of Security and Privacy in Pervasive Computing (IJSPPC)**. 2022 Jan 1;14(1):1-25

Shi S, Tse R, Luo W, D'Addona S, Pau G. *Machine learning-driven credit risk: a systemic review*. Neural Computing and Applications. 2022 Sep;34(17):14327-39.

Sidey-Gibbons JA, Sidey-Gibbons CJ. *Machine learning in medicine: a practical introduction*. BMC medical research methodology. 2019 Dec;19:1-8.

Singh SK, Taylor RW, Pradhan B, Shirzadi A, Pham BT. *Predicting sustainable arsenic mitigation using machine learning techniques.* Ecotoxicology and Environmental Safety. 2022 Mar 1;232:113271.

Singh Y, Bhatia PK, Sangwan O. *A review of studies on machine learning techniques.* **International Journal of Computer Science and Security**. 2007 Jun;1(1):70-84.

Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL. Fixmatch: *Simplifying semi-supervised learning with consistency and confidence*. Advances in neural information processing systems. 2020;33:596-608.

Soreti Bekele Babo, Asrat Mulatu Beyene. Bank loan classification of imbalanced dataset using machine learning approach, 15 March 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2667057/v1]

Srinath H, Sharma AK, Akhil MR. *Reinforcement learning in real-world scenarios: Challenges, applications, and future directions.* I**nternational Journal of Research in Engineering, Science and Management**. 2023 Jul 31;6(7):40-5.

Srivastava, S.; Saranya, G.; Pratap, A.; Agrawal, R.; and Jain, A. *Loan default prediction using artificial neural networks.* **International Journal of Advanced Science and Technology**, 2020:29(6), 2761-2769. 21.

Sudarmaji E, Achsani NA, Arkeman Y, Fahmi I. *Credit-worthiness prediction in energy-saving finance using machine learning model*. Studies of Applied Economics. 2021 Oct 18;39(10).

Sujatha CN, Gudipalli A, Pushyami B, Karthik N, Sanjana BN. *Loan prediction using machine learning and its deployement on web application*. In2021 Innovations in Power and Advanced Computing Technologies (i-PACT) 2021 Nov 27 (pp. 1-7). IEEE.

Syahida Abdullah, Zakirah Othman, and Roshayu Mohamad. "*Predicting the risk of sme loan repayment using AI technology-machine learning techniques: A perspective of Malaysian financing institutions"*. **Journal of Advanced Research in Applied Sciences and Engineering Technology** 31, no. 2 (July 28, 2023): 320–326. http://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/article/view/2994.)

Tariq HI, Sohail A, Aslam U, Batcha NK. *Loan default prediction model using sample, explore, modify, model, and assess (SEMMA)*. **Journal of Computational and Theoretical Nanoscience**. 2019 Aug 1;16(8):3489-503.

Tejaswini J, Kavya TM, Ramya RD, Triveni PS, Maddumala VR. A*ccurate loan approval prediction based on machine learning approach*. **Journal of Engineering Science**. 2020;11(4):523-32.

Utkin, L.V., Kovalev, M.S., Coolen, F.P. I*mprecise weighted extensions of random forests for classification and regression*. Applied Soft Computing. 2020 Jul 1;92:106324.

Vardi N. *Creditworthiness assessment and other contractual duties as tools of 'responsible credit': The case of consumer loans.* InCreditworthiness and'Responsible Credit' 2022 Aug 4 (pp. 144-214). Brill Nijhoff.

Varshney KR. *Trustworthy machine learning and artificial intelligence*. XRDS: Crossroads, The ACM Magazine for Students. 2019 Apr 10;25(3):26-9.

Viswanatha V, Ramachandra A.C, Vishwas K N, and Adithya G. "*prediction of loan approval in banks using machine learning approach"*. **International Journal of Engineering and Management Research** 13, no. 4 (August 2, 2023): 7–19. https://ijemr.vandanapublications.com/index.php/ijemr/article/view/1318.

Von Borries GF, de Castro quadros av. *Roc app: An application to understand roc curves*. **Brazilian Journal of Biometrics**. 2022 Jun 10;40(2).

Vujović Ž. *Classification model evaluation metrics*. **International Journal of Advanced Computer Science and Applications**. 2021;12(6):599-606.

Wang J, Li P, Ran R, Che Y, Zhou Y. *A short-term photovoltaic power prediction model based on the gradient boost decision tree*. Applied Sciences. 2018 Apr 28;8(5):689.

Wang K, Lu J, Liu A, Zhang G, Xiong L. *Evolving gradient boost: A pruning scheme based on loss improvement ratio for learning under concept drift*. IEEE Transactions on Cybernetics. 2021 Oct 6.

Woldaregay AZ, Årsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, Hartvigsen G. *Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes.* Artificial intelligence in medicine. 2019 Jul 1;98:109-34.
Yan J, Wang X. *Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology.* **The Plant Journal.** 2022 Sep;111(6):1527-38.

Yeo GF, Akman D, Hudson I, Chan J. A s*tochastic approximation approach to fixed instance selection*. Information Sciences. 2023 May 1;628:558-79.

Yigin BO, Algin O, Saygili G. *Comparison of morphometric parameters in prediction of hydrocephalus using random forests*. Computers in Biology and Medicine. 2020 Jan 1;116:103547

Ying L. *Research on bank credit default prediction based on data mining algorithm*. **The International Journal of Social Sciences and Humanities Invention**. 2018;5(6):4820-3.

Yu S, Yang M, Wei L, Hu JS, Tseng HW, Meen TH. *Combination of self-organizing map and k-means methods of clustering for online games marketing*. Sensors & Materials. 2020 Aug 30;32.

Thesis

Asfaw W. *Addis Ababa Institute of Technology School of Electrical and Computer Engineering Telecommunication Engineering Graduate Program* (Doctoral dissertation, Addis Ababa University Addis Ababa).

Dissanayake T. *A machine learning approach to predict bank loan approval* (Doctoral dissertation).2022

Nilsson M, Shan Q. *Credit risk analysis with machine learning techniques in peer-to-peer lending market*. Stockholm Business School Master's Degree Thesis Master's Programme in Banking and Finance.2018

Sriranganathan K. *Bank loan approval prediction using machine learning approach: evidence from Sri Lanka* (Doctoral dissertation).2022)

Websites

Central Bank of Nigeria, Revised Guidelines for Primary Mortgage Banks in Nigeria. November, 2011 https://www.cbn.gov.ng/out/2013/ofisd/revised%20guidelines%20for%20primary%20mortgage%20banks%20in%20nigeria.pdf

Zendi. Data Science Nigeria Loan Default Prediction Challenge. 2024 https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data

Appendices

Appendix I: Design Source Code

```
In [69]: x = np.arange(0,len(df_timeline1),1)
fig, ax = plt.subplots(1,1,figsize=(20,5))
ax.plot(x,df_timeline1['num_of_transactions'])
ax.set_xticks(x)
ax.set_xticklabels(df_timeline1['year_month'])
ax.set_xlabel('Year Month')
ax.set_ylabel('Num of Transactions')
plt.show()
```

```
In [51]: #importing required packages
#modelues for EDA steps
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#modules for data cleaning and data analysis
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler
import scipy.stats as stats
#modules for model building
#algorithms for sampling
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler
from imblearn.over_sampling import SMOTE
#baseline linear model
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
```

```
from sklearn.ensemble import RandomForestClassifier
#modules for hyper parameter tuning
from sklearn.model_selection import GridSearchCV
#modules for model evaluation
from sklearn.model_selection import cross_val_score
from sklearn import metrics
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.metrics import precision_score, accuracy_score, f1_score, r2_score
from sklearn.metrics import precision_recall_curve, roc_curve
#modules for avoiding warnings
import warnings
warnings.filterwarnings('ignore')
#setting backend for matplotlib
%matplotlib inline
#setting formatting options
pd.options.display.max_columns = 100
pd.options.display.max_rows = 900
pd.set_option('float_format' , '{:f}'.format)
#setting plot style
plt.style.use('seaborn-darkgrid')
In [52]: #loading the dataset
df = pd.read_csv('data/loanTrain.csv')
Out[54]:
In [55]: df.columns
Out[55]:
In [56]: df.drop('Unnamed: 0', axis=1, inplace=True)
In [57]: # Plot the distribution of each variable
df.hist(figsize=(20,15))
plt.show()4/18/23, 5:51 PM
loan_detection
file:///C:/Users/owner/Downloads/loan_prediction.html
```

4/44

0 1289169

1 7506

Out[59]:

In [60]: *# Feature engineering*

 # Extract useful information from the "trans_date_trans_time" variable

#converting trans_date_trans_time into datetime

```
df['trans_date_trans_time'] = pd.to_datetime(df['trans_date_trans_time'])
df['trans_datetime'] = pd.to_datetime(df['trans_date_trans_time'])
df['day_of_week'] = df['trans_datetime'].dt.dayofweek
df['hour_of_day'] = df['trans_datetime'].dt.hour
df['trans_year_month'] = df['trans_date_trans_time'].dt.to_period('M')
df['time_since_last_trans'] = df.groupby(['cc_num'])['unix_time'].diff().fillna(0)
```

:

In [62]: *# Create a new variable that indicates the distance between the customer's location*

```
df['dist_customer_merchant'] = np.sqrt((df['lat'] - df['merch_lat'])**2 + (df['long
```

In [13]: *# Create a new variable that indicates the frequency of transactions made by each c*

```
df['freq_trans_customer_merchant'] = df.groupby(['cc_num', 'merchant'])['trans_num'
```

 # Create a new variable that indicates the time difference between the current tran

```
df['time_diff_customer_merchant'] = df.groupby(['cc_num', 'merchant'])['unix_time']
```

In [63]: *#finding age*

#converting 'dob' column to datetime

```
df['dob'] = pd.to_datetime(df['dob'])
df['age'] = np.round((df['trans_date_trans_time'] - df['dob'])/np.timedelta64(1, 'Y
df.age.head()
```

Out[63]:

In [64]: *#dropping variables*

```
df.drop(['trans_date_trans_time','first', 'last', 'dob'] , axis=1, inplace=True)
df.head()4/18/23, 5:51 PM
```

Out[66]:4/18/23, 5:51 PM

loan_detection

```
plot = [0,0,0]
#plotting the 'trans_hour' feature
plot[0] = sns.countplot(df.hour_of_day, ax = plt.subplot(221))
#plotting the 'trans_day_of_week' feature
plot[1] = sns.countplot(df.day_of_week, ax = plt.subplot(222))
#plotting the 'trans_year_month' feature
plot[2] = sns.countplot(df.trans_year_month, ax = plt.subplot(212))
for i in plot:
i.set_xticklabels(i.get_xticklabels(), rotation=30)
plt.show()
In [68]: #year_month vs number of transactions
df_timeline1 = df.groupby(df['trans_year_month'])[['trans_num','cc_num']].nunique()
df_timeline1.columns = ['year_month','num_of_transactions','customers']
df_timeline
Out[72]:
In [73]: x = np.arange(0,len(df_timeline2),1)
fig, ax = plt.subplots(1,1,figsize=(20,5))
ax.plot(x,df_timeline2['loan_customers'])
ax.set_xticks(x)
ax.set_xticklabels(df_timeline2['year_month'])
ax.set_xlabel('Year Month')
ax.set_ylabel('Number of loan customers')
plt.show()
In [74]: #creating the 'gender' distributed dataframe
df_gender = df[['gender','trans_num']].groupby(['gender']).count().reset_index()
df_gender.columns = ['Gender', 'gender_count']
#creating gender-loan distribution
In [76]: #let us first bin the age feature
for i in range(len(df.age)):
if df.age[i] <= 30:
```

```
df.age[i] = '< 30'
elif df.age[i] >30 and df.age[i] <= 45:
df.age[i] = '30-45'
elif df.age[i] >45 and df.age[i] <= 60:
df.age[i] = '46-60'
elif df.age[i] >60 and df.age[i] <= 75:
df.age[i] = '61-75'
else:
df.age[i] = '> 75'
df.age.head()4/18/23, 5:51 PM
In [77]: #constructing the age-transaction count distribution
df_age = df[['age','trans_num']].groupby(['age']).count().reset_index()
df_age.columns = ['age', 'age_count']
#creating the age-loan distribution
df_loan_age = df[['age', 'trans_num', 'is_loan']].groupby(['age','is_loan']).cou
df_loan_age.columns = ['age', 'is_loan', 'Transaction count']
df_loan_age = df_loan_age.merge(df_age[['age', 'age_count']], how='inner', on='ag
df_loan_age['Transaction percentage'] = (df_loan_age['Transaction count']/df_frau
df_loan_age
Out[77]:
In [78]: sns.barplot(data=df_loan_age, y='Transaction count', x='age', hue='is_loan')
plt.show()4/18/23, 5:51 PM
In [88]: #function to return highly correlated column above a threshold
def correlation(dataset, threshold):
col_corr = set() # This set stores the highly correlated columns
corr_matrix = dataset.corr() #correlation matrix
#traversing the correlation matrix
for i in range(len(corr_matrix.columns)):
for j in range(i):
if corr_matrix.iloc[i,j] >threshold:
colname = corr_matrix.columns[i] #selecting columns above threshold
```

col_corr.add(colname) *#adding columns to set*

return col_corr

In [89]: *#let us get the features with correlation above 85%*

corr_features = correlation(df,0.85)

corr_features4/18/23, 5:51 PM

44250.000000

1296675 rows × 8 columns

Out[89]:

In [90]: *#removing unnecessary variables*

df.drop(['zip', 'lat', 'long', 'city_pop', 'unix_time', 'merch_lat','merch_long','m

axis=1, inplace=**True**)

In [91]: df.head()

Out[91]:

In [124…

#split X and Y

X = df.drop(['is_loan'],axis=1)

y = df.is_loan

In [125…

X Out[125]:

In [93]: *#scaling*

scaler = StandardScaler()4/18/23, 5:51 PM

loan_prediction

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1296675 entries, 0 to 1296674

Data columns (total 9 columns):

Column Non-Null Count Dtype

--- ------ -------------- -----

0 category 1296675 non-null int32

1 amt 1296675 non-null float64

2 gender 1296675 non-null int32

3 is_loan 1296675 non-null int64

4 day_of_week 1296675 non-null int64

5 hour_of_day 1296675 non-null int64

6 time_since_last_trans 1296675 non-null float64

7 dist_customer_merchant 1296675 non-null float64

8 age 1296675 non-null int32

dtypes: float64(3), int32(3), int64(3)

memory usage: 74.2 MB

0 644585

1 3753

for train_index, test_index **in** skf.split(X,y):

X_train, X_test = X[train_index], X[test_index]

y_train, y_test = y[train_index], y[test_index]

y_train.value_counts()

Out[95]:

In [96]: lr = LogisticRegression(random_state=42)

model = lr.fit(X_train, y_train)

y_train_pred = model.predict(X_train)

y_test_pred = model.predict(X_test)

In [97]: *#evaluating the model*

model_name = 'Logistic Regression - imbalance class'

train_score = model.score(X_train,y_train)

test_score = model.score(X_test,y_test)

acc_score = accuracy_score(y_test,y_test_pred)

f_score = f1_score(y_test, y_test_pred, average='weighted')

precision = precision_score(y_test, y_test_pred)

recall = metrics.recall_score(y_test,y_test_pred)

#creating a dataframe to compare the performance of different models

model_eval_data = [[model_name, train_score, test_score, acc_score, f_score, precis

evaluate_df = pd.DataFrame(model_eval_data, columns=['Model Name', 'Training

Out[97]:

In [98]: *#random under sampling using imblearn*

```
rus = RandomUnderSampler()
X_rus, y_rus = rus.fit_resample(X_train,y_train)
y_rus.value_counts()
Out[98]:
In [99]: X_train, X_test, y_train, y_test = train_test_split(X_rus, y_rus, test_size=0.3, ra
In [100…
#evaluating the model
model_name = 'Logistic Regression - with Random Under Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)4/18/23, 5:51 PM
#adding calculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df
Out[103]:
In [104…
#oversampling with imblearn
ros = RandomOverSampler()
X_ros, y_ros = ros.fit_resample(X_train,y_train)
y_ros.value_counts()
Out[104]:
In [105…
#train Test split
X_train, X_test, y_train, y_test = train_test_split(X_ros,y_ros, test_size=0.3, str
y_train.value_counts()
Out[105]:
In [134…
```

```
#implementing logistic regression
lr = LogisticRegression(random_state=42)
#creating model
model = lr.fit(X_train, y_train)
y_train_pred = model.predict(X_train)
y_train_pred
Out[134]:
In [107…
test_pred = model.predict(X_test)
test_pred
Out[107]:
In [108…
#printing classification report
In [109…
#evaluating the model
model_name = 'Logistic Regression - Random Over Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df
Out[109]:
In [110…
#implementing logistic regression
lr = LogisticRegression(random_state=42)
```

```
#creating model
model = lr.fit(X_train, y_train)
y_train_pred = model.predict(X_train)
y_train_pred
Out[112]:
In [116…
model = model.predict(X_test)
model
Out[116]:
In [114…
#printing classification report
print(classification_report(y_test, test_pred))
In [115…
#evaluating the model
model_name = 'Logistic Regression - SMOTE'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
In [69]: #Decisiontree classifier
#train-test split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3, random_stat
dtree = DecisionTreeClassifier(max_depth=10)
model = dtree.fit(X_train,y_train)
```

```
y_test_pred = model.predict(X_test)
print(classification_report(y_test, y_test_pred))
#evaluating the model
model_name = 'Decision Tree - imbalance class'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,y_test_pred)
f_score = f1_score(y_test, y_test_pred, average='weighted')
precision = precision_score(y_test, y_test_pred)
recall = metrics.recall_score(y_test,y_test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
In [70]: #train-test split
X_train, X_test, y_train, y_test = train_test_split(X_rus,y_rus, test_size=0.3, ran
dtree = DecisionTreeClassifier(max_depth=10)
model = dtree.fit(X_train,y_train)
test_pred = model.predict(X_test)
print(classification_report(y_test, test_pred))
#evaluating the model
model_name = 'Decision Tree - Random Under Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
```

```
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
In [71]: #train-test split
X_train, X_test, y_train, y_test = train_test_split(X_ros,y_ros, test_size=0.3, ran
dtree = DecisionTreeClassifier(max_depth=10)
model = dtree.fit(X_train,y_train)
test_pred = model.predict(X_test)
print(classification_report(y_test, test_pred))
#evaluating the model
model_name = 'Decision Tree - Random Over Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
#evaluating the model
model_name = 'Decision Tree - SMOTE'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)4/18/23, 5:51 PM
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
```

```
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df
Out[72]:
In [73]: #train-test split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3, random_stat
rf = RandomForestClassifier(n_estimators=100, criterion='gini')
model = rf.fit(X_train,y_train)
y_test_pred = model.predict(X_test)
print(classification_report(y_test, y_test_pred))4/18/23, 5:51 PM
#evaluating the model
model_name = 'Random Forest
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,y_test_pred)
f_score = f1_score(y_test, y_test_pred, average='weighted')
precision = precision_score(y_test, y_test_pred)
recall = metrics.recall_score(y_test,y_test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
#evaluating the model
model_name = 'Random Forest - Random Under Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
```

```
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
#evaluating the model
model_name = 'Random Forest - Random Over Sampling'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
In [79]: #train-test split
X_train, X_test, y_train, y_test = train_test_split(X_sm,y_sm, test_size=0.3, rando
rf = RandomForestClassifier(n_estimators=100, criterion='gini')
model = rf.fit(X_train,y_train)
test_pred = model.predict(X_test)
print(classification_report(y_test, test_pred))
#evaluating the model
model_name = 'Random Forest - SMOTE'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)4/18/23, 5:51 PM
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
```

```
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
evaluate_df4/18/23, 5:51 PM
In [80]: #train-test split
model_name = 'Random Forest - SMOTE'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
In [81]: best_grid = RandomForestClassifier(max_features = 'sqrt', n_estimators=200,
random_
#train-test split
X_train, X_test, y_train, y_test = train_test_split(X_sm,y_sm, test_size=0.3, rando
model = best_grid.fit(X_train,y_train)
test_pred = model.predict(X_test)
print(classification_report(y_test, test_pred))
#evaluating the model
model_name = 'Random Forest - SMOTE [Hyperparameter Tuned]'
train_score = model.score(X_train,y_train)
test_score = model.score(X_test,y_test)
acc_score = accuracy_score(y_test,test_pred)
f_score = f1_score(y_test, test_pred, average='weighted')
precision = precision_score(y_test, test_pred)
recall = metrics.recall_score(y_test,test_pred)
#adding claculations to dataframe
model_eval_data = [model_name, train_score, test_score, acc_score, f_score, precisi
```

```
model_eval_dict = {evaluate_df.columns[i]:model_eval_data[i] for i in range(len(mod
evaluate_df = evaluate_df.append(model_eval_dict, ignore_index=True)
```

Bio data

A. Personal Data

1. **Full Name:** EFEKODO Kingsley Oghenekaro
2. **Date and Place of Birth:** 18th January 1981
3. **Nationality:** Nigerian
4. **Marital Status:** Married
5. **Place of Birth:** Ozoro
6. **Local Govt. Area:** Isoko-North
7. **State of Origin:** Delta
8. **Permanent Address:** 15 Oshidehin Street, Off Akilo Road, Ogba-Agege, Lagos
9. **Email:** keoghenekaro @yahoo.com
10. **Department:** Computer Science

B. Educational Background

Educational Institutions Attended with Dates and Qualification:

i.	Nasara Nursery/Primary School Chanchanga, Minna	1990 - 1994
ii.	Day Secondary School Tunga Minna, Niger State	1995 - 1997
iii.	Government Science College Izom , Niger State	1998 - 2000
iv.	Federal University of Technology Minna, Niger State	2001 - 2007
v.	Lead City University, Ibadan	2023-Till date

Educational Qualification with Dates:

i. First school leaving certificate	1994
ii. J.S.S.C.E Certificate	1997
iii. S.S.C.E Certificate	2000
iv. Bs.c Mathematics/Computer science	2007
v. Ms.c Computer Science	in-view

C. Work Experience: With Dates

Khamplus Multi-Concept Limited, Minna, Niger State	2009

Gucci Chis Nig. Ltd, Ikeja Lagos	2011
Pumoh Exclusive Online Service Ltd, Ikeja Lagos	2015
Platinum IT-Assets Consultancy Ltd	2017 till date

C. Names and Addresses of Referees

Prof. A. Akinola
Senior Lecturer
Lead City University, Ibadan
Department of Computer science
Solom202@yahoo.co.uk

Prof Akinola Akinlabi
Rector
Oyo State College of Agriculture and Technology, Igbo-ora
@gmail.com

The University Compliance Certification

This is to certify that this thesis by Kingsley Oghenekaro EFEKODO with Matriculation Number LCU/PG/003077 in the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan is in full compliance with the approval of the University's format and style.

…… ……..
Signature Date