HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEVELOPING LOAN DEFAULT PREDICTION MODEL USING MACHINE LEARNING TECHNIQUES

By
Tewodros Teshale Alemu

Major Advisor: Efrem Yohannes (PhD)
Co-Advisor: Solomon Tsegaye (MSc)

MAY, 2024

DEVELOPING LOAN DEFAULT PREDICTION MODEL USING MACHINE LEARNING TECHNIQUES

By

Tewodros Teshale

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF HAWASSA UNIVERSITY INSTITUTE OF TECHNOLOGY

IN PARTIAL FULFILLMENTN OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

HAWASSA, ETHIOPIA

Program: MSc. Computer Science

Major Advisor: Efrem Yohannes (PhD)

Co-Advisor: Solomon Tsegaye (MSc)

MAY, 2024

HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES FACULTY OF INFORMATICS

THESIS APPROVAL SHEET-I

This is to certify that the thesis entitled "**Developing Loan Default Prediction Model Using Machine Learning Technique**" submitted in partial fulfilment of the requirements for the degree of Master's with specialization in Computer Science, the graduate program of the school of informatics, and has been carried out by Tewodros Teshale. Therefore, we recommend that the student has fulfilled the requirements and hence herby can submit the thesis to the department

Name of Major Advisor: Efrem Yohannes (PhD)

Signature _____

Date of Submissioı 05/17/2024

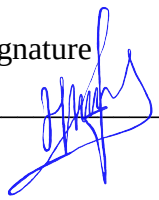
Name of Co-advisor: Solomon Tsegaye (MSc)

Signature: _____

Date of Submission: _____

THESIS APPROVAL SHEET –II

We, the undersigned members of the Board of Examiners of the final open defence by Tewodros Teshale have read and evaluated his thesis entitled "**Developing Loan Default Prediction Model Using Machine Learning Technique**", and examined the candidate. This is, therefore. To certify that the thesis has been accepted in partial fulfilment of the requirements for the degree of Master of Science in "Computer Science".

Name of Major Advisor	Signature	Date
Efrem Yohannes (PhD)		05/17/2024
Name of Internal Examiner-I	Signature	Date
Name of Internal Examiner-II	Signature	Date
Name of External Examiner	Signature	Date
Dr. Hussien Seid		July 3, 2024
SGS Approval	Signature	Date

ACKNOWLEDGMENT

First and foremost, I would like to thank the almighty God including his mother St. Mary's who helped me to live and unlimited support in every aspect of my life.

This thesis would not have been possible without the help, support, and guidance of my advisor, Dr. Efrem Yohannes for his encouragement right from the beginning to the completion of the work also my co-advisor Mr. Solomon Tsegaye for your guidance and support.

I also would like to thank Wegagen Bank Sc. Directors, Manager and Credit department staffs for their full cooperation for the provision of the required data for my study. Finally, I need to express my deep thanks to my family and my friends.

STATEMENT OF THE AUTHOR

I hereby declare that this MSc thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name of Student:	**Tewodros Teshale**
Signature:	_____
Date of Submission:	_____
Place:	**Institute of Technology, Hawassa University, Hawassa**

ABSTRACT

Loan defaults pose a significant risk to financial institution, leading to substantial financial losses and impacting their stability and profitability. Existing predictive models often overlook key borrower characteristics, resulting in less accurate predictions. This study aims to improve loan default prediction by integrating borrower-specific features and loan characteristics using a blending ensemble model. Specifically, we focus on borrower characteristics such as business location, loan product type, yearly business income, location of collateral, total years of experience, and educational status, which are used by some Ethiopian banks for risk assessment but have been underexplored in previous studies.

We employ three base models: logistic regression, multilayer perceptron, and random forest. These models are combined using a weighted average blending ensemble approach to enhance predictive performance. The dataset, consisting of 18,184 records from a single bank, was split using an 70/30 ratio for training and testing.

Our findings demonstrate that the blending ensemble model outperform individual base models in predicting loan defaults, achieving higher accuracy (98.62%), precision, recall, and F1-score. The most significance predictors identified includes sex, collected total, educational status, employment status, and age, while gender and marital status shower lesser impact. This study contributes to the field by providing a more robust predictive model that incorporates underexplored borrower characteristics, offering financial institutions a more accurate tool for risk assessment and decision-making.

Keyword: Loan default, machine learning, loan status, normal loan, special mention, substandard loans, doubtful loans, and loss loan, blending ensemble, multilayer perceptron, random forest, logistic regression.

Table of Contents Page

LIST OF FIGURE

LIST OF TABLE

ACRONYMS/ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial Neural Network
CIC	Credit Information Center
DNN	Deep Neural Network
DOUB	Doubtful
EAD	Exposure at Default
EL	Expected Loss
EMRF	Ensemble Mixture Random Forest Model
ISHOPA	Imperial Savings & Home Ownership Public Association
LGD	Loan Given Default
ML	Machine Learning
MLP	Multilayer perceptron
NPL	Non-Performing Loan
NBE	National Bank of Ethiopia
NORM	Normal
PAR	Portfolio at Risk
ReLU	Rectified Neural Network
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-Sampling Technique
SOM	Self-organizing Map
SPME	Special Mention
SUBS	Substandard
SVM	Support Vector Machine

CHAPTER ONE

INTRODUCTION

1.1. Background of Study

The number of banks in Ethiopia is increasing after the new policy proclamations. These Banks are owned by the government or by shareholders. Both private and governmental banks are established with the goal of serving the financial needs of the society. These banks play significant role in Ethiopia's economic growth, as they are source of finance for every sector. The agricultural sector, healthcare, education, trade, marketing, etc., cannot fully function without bank involvement in their daily operations. A bank is the bridge between customers who have huge capital and a customer who has innovative ideas, and this combination creates profitability [1].

Similarly, commercial and developmental banks loans are their main source of revenue. Interest on loans is a major source of profit for banks. Every nation's economic development depends on its banks' ability to turn a profit; on the other hand, banks with poor management practices impede national economic growth. Due to the increasing demand for financial institutions, banks' roles are becoming more and more important in developing nations. While the majority of loans are returned according to their schedules, some do not get repaid. Non-performing loans (NPL) are the term used to describe these types of defaulting loans.

Default is the failure to make the required interest or principal repayment on a debt, whether that debt is a loan or security. Individuals, Businesses, and even countries can default on debt obligations.

For commercial and developmental banks, assessing and predicting customers' ability to repay debt has been one of the most difficult problems in recent years. People from all over the most difficult problems that commercial and development banks face. In one way or another, banks serve people all over the world by lending them money to help them get over financial challenges and fulfill some of their dreams. The procedure of obtaining a loan has grown essential due to the economy's continual fluctuations and the intensifying rivalry among financial institutions.

Furthermore, lending is a vital source of income for banks of all sizes, helping them to maintain operations and make ends meet even in lean economic times. Loan default means that a borrower

does not make the required payments or does not comply with the loan terms. The profit or loss of the financial lender largely depends on loan repayments, that is, whether customers are playing back the loans(defaulting). Therefore, when loans default, financial institutions will loss money, which might even lead to bankruptcy and the collapse of institutions. By predicting loan default, financial institutions(lenders) can reduce credit risk, prevent loan default and increase profit by evaluating the ability of the borrower to deliver on their obligation of loan repayment i.e., loan default prediction [2].

Taking a loan has been inevitable for people because individuals around the world depend on loans to overcome financial constraints to achieve their personal goals, and organizations rely on loans to expand their production [3]. In most cases, loan lending is beneficial to both borrowers and lenders. However, loan default is still unavoidable, which carries great risk and may even end up in a financial crisis. Therefore, it is important to identify whether a candidate is eligible to receive a loan. In the past, evaluation primarily depended on manual review, which was time-consuming and labour-intensive [4].

Artificial intelligence (AI) and machine learning (ML) techniques are transforming and revolutionizing loan default risk management. Currently, monitoring risk becomes possible after the growth of AI-driven solutions, such as deciding on the amount of money a bank should lend to its customers, providing warning signals during financial market risks detecting customer potential, better compliance, and improved default loan prediction [5].

Loan default prediction is the most serious activity of banks in reducing their probability of default. Default is the delay of loan repayment or totally unable to pay for the bank. This has different causes, including incorrect estimations of loan customers. Traditionally, banks profit from their loan customers through demographic measurement. The aim of this thesis is to use machine learning methods to create a loan default prediction model for Wegagen Bank. This research focuses on collecting and analyzing loan application data to accurately assess risk and manage portfolios at risk ratios precisely.

1.2. Bank in Ethiopia

Modern banking in Ethiopia was introduced in 1905 after the agreement that was reached between Emperor Minilik II and Mr. Ma Gillivary representative of the British-owned National Bank of Egypt.

The first bank called Bank of Abyssinia was inaugurated in Feb.16, 1906, by the Emperor. The Bank was managed by the Egyptian National Bank and the following rights and concessions were agreed upon the establishment of the Bank of Abyssinia. [1] The society at that time being new to the banking service, Bank of Abyssinia had faced the difficulty of familiarizing the public with it. Despite its monopolistic position, the bank earned no profit until 194, a reflection of the challenges faced in establishing a modern banking system in Ethiopia.

The bank had faced many pressures as a result of its inefficiency and was purely profit-oriented. Thus, an agreement was reached to abandon its operation and be liquidated so as to disengage banking from foreign control and to make the institution responsible to Ethiopian's credit need. By shortly after Haile Sellassie came to power, the Bank of Ethiopia was a purely Ethiopian institution and was the commercial activities of the Bank of Abyssinia and was authorized to issue notes and coins.

During the invasion (1935) the Italians established branches of their main banks namely Bankca d' Italia, Banco diRoma, Banko diNapoli, and Banca Nazionale delavoro [2].

In 1941, another foreign bank, Barclays Bank came to service in Addis Ababa till it withdraws in 1943, shortly before the commencement of the full operation of the state Bank of Ethiopia.

The state Bank of Ethiopia acted as the central Bank of Ethiopia and as the principal commercial bank in the country and engaged in all commercial banking activities until ceased to exist by a bank proclamation issued in December 1963. [2]**.**

The National Bank of Ethiopia and the Commercial Bank of Ethiopia were established as a result of the Ethiopian Monetary and Banking Proclamation, which divided the role of commercial and central banking. While a commercial bank of Ethiopia took over the commercial banking operations of the old state bank of Ethiopia, the National Bank of Ethiopia still handles central banking functions. In addition to the country's first private bank, Addis Ababa Bank Share Company, which opened for business in 1964, other financial institutions included Imperial Savings and Home Ownership Public Associations (ISHOPA) and Savings and Mortgage Corporation of Ethiopia. These organizations' objectives were to accept savings and trust accounts, provide loans for home improvement, repair, and construction, and accept deposits.The Agricultural Bank was founded in 1945 and succeeded by the Investment Bank in 1951. The Agricultural Bank offered financing for projects related to agriculture and other similar fields [2]**.**

1.3. Statement of the Problem

Lending a loan benefits both lenders and borrowers, it carries inherent risks, particularly the risk of the borrower failing to repay the loan as agreed, known as credit risk. Accurately assessing a client's creditworthiness before granting a loan is essential for minimizing these risks.

Traditional lending processes often rely on the 5C principle- Character, Capital, Capacity, Collateral, and Conditions to evaluate borrowers. This method, which largely depends on the personal judgment and experience of the bank CRM (credit relationship manager), has notable limitation. Even with thorough verification and validation procedures, there is no certainty that an approved applicant will repay the loan on time.

Credit scoring systems, which assign a numerical score based on applicants' credit histories and backgrounds, are widely used to assess creditworthiness. These scores classify applicants into good payers (likely to repay on time) and bad payers (likely to default). Although useful, traditional credit scoring methods frequently overlook important borrower characteristics that can affect loan default risk.

This study aims to enhance loan default prediction by incorporating borrowers-specific characteristics that are crucial in real-world loan risk assessments but have been underexplored in previous research. It focuses on factors used by some Ethiopian banks for risk assessment, such as sex, marital status, educational background, business income, collateral location, total year of experience of the borrowers. By integrating these features, the study seeks to improve the accuracy of loan default predictions, offering financial institutions a more reliable assessment tool.

1.4. Research Question

To achieve the objectives listed below, this study will answer the following research question.

- Are ensemble method more effective than individual machine learning techniques in loan default prediction?
- What are the most important features or variables in predicting loan default?

1.5. Objective of the Study

1.5.1. General Objective

The general objective of the study is to develop an accurate loan default prediction model using machine learning techniques and identify the most important features influencing loan defaults.

1.5.2. Specific Objective

- Evaluate the effectiveness of ensemble methods
- Identify key predictive features
- Integrate diverse borrower and loan characteristics
- Compare model performance

1.6. Significance of Study

This study makes several contributions to both knowledge-building and practice improvement in loan default prediction. From this study, Bank industries in Ethiopia are mainly beneficiaries:

- For identifying and detecting possible defaulter and active loan applicants before granting loans to the borrowers.
- To limit or to completely end granting loans default loan
- To instantaneously approve low-risk customers and secure PAR (Portfolio at risk) at an internationally accepted level below 5%.

1.7. Scope of the Study

The scope of this thesis focuses on investigating and identifying patterns that aid in predicting the probability of default for loan applicants. It aims to develop a model that can classify loan applications into accepted or rejected categories based on historical bank data of borrowers. The study involves various stages, starting from pre-processing the loan and borrower's historical data and culminating in accurate predictions and classifications of loan applicants.

1.8. Limitation of the Study

This study has a limitation that must be acknowledged. It is the dataset used was sourced from a single financial institution, Wegagen Bank because other institutions were unwilling to provide data. This introduces potential bias as the data may not fully represent the diverse borrowing patterns and defaulter characteristics found in other banks. As a result, the model's generalizability to other financial institutions may be limited.

1.9. Organization of Study

This study is organized into five chapters. The first chapter discusses the introduction of the thesis to highlight background information, the problem statement, the study's objective, significance, and scope and limitations. Chapter Two is the literature review part that focuses on credit risk management, credit process, machine learning concepts, machine learning algorithms, model evaluation concepts, and machine learning-related research papers on loan default prediction. Chapter Three covers data understanding, Business understanding, and Data preprocessing including data cleaning, feature selection, feature engineering, data normalization, and data splitting. Chapter Four presents the proposed architecture and how the data is prepared before the experiment by explaining Data Preprocessing tasks. In addition, it presents how the training and testing dataset split, the predictive models experimented using the proposed algorithms, a comparison of the results of the models, identification of major risk factors, and finally a discussion of the result. Finally, Chapter Five presents the conclusion and recommendations for future works presented.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Credit Risk

The risk of a borrower not fulfilling their commitment to service debt and not making payments is known as credit risk. They claim that non-payment on account of failure by the other party to make payments on time or otherwise may be the cause of the non-payment [6]. Furthermore, they stated that a bank's credit risk is by nature its most obvious risk. Furthermore, the author uses three factors to describe credit risk. The first is default risk, which is the chance that money won't be sent out for at least three months. Due to a higher default chance, sector knowledge, and management quality, counterparts with bad financial standing, huge debt loads, low and uncertain revenue will experience this delay. The second factor is exposure risk, which is the uncertainty surrounding the precise amount at risk at the exact moment of a potential default, and loss risk, also known as loss given default (LGD), which is a portion of exposure in the event of non-payment. Similarly, [7] says that default risk is another name for credit risk and that it is the bank's risk of loss if a counterparty fails to fulfil their payment obligations. In the banking industry, credit risk is the primary and oldest source of risk, as he also makes clear. Credit risk includes the potential for both an obligor's credit quality to decline and result in an economic loss, as well as the probability that a borrower will default by neglecting to repay principal and interest on time [8].

Credit risk emerges and can result in the bankruptcy of one of the counter parties to a transaction if they fail to properly clear up while the funds are outstanding or later on [9].

Credit risk, as defined by [10], is the inability to fulfil a stipulated contractual financial obligation before the deadline for payment by the other party. Credit risk is the potential for debtors or borrowers to be unable to fulfil their responsibilities in accordance with predetermined contractual agreements made during the credit approval process, which would negatively impact the lender's operating environment.

Credit risk is characterized as the likelihood that a counterparty or bank borrower may not fulfil its contractual commitments on time [11]. Credit risk occurs when a lender faces potential loss as a result of a counterparty, borrower, or obligor not fulfilling their contractual obligations.

Credit risk is the possibility that borrowers will not make their loan payments in full or in part for various reasons [12] According to this statement, the bank's exposure to higher credit risk might result in unforeseen financial crises, while lower credit risk can reduce the likelihood of such crises because this particular department of the bank will create significant profits.

2.2 Credit Risk Management

All management tasks, such as identifying, measuring, monitoring, and controlling credit risk exposure, are included in credit risk management, according to [7] Additionally, the author stated that in the current business climate, strong credit risk management practices are essential for the banking sector's long-term success and that inadequate credit risk management policies will seriously contribute to a banking industry disaster.

The process of analyzing and updating credit risk management documents, as well as applying them consistently to the actual credit granting, credit administration, monitoring, and risk-controlling processes, is referred to as credit risk management practice [13]. This definition includes understanding and identifying risk in order to minimize the unfavorable effects of taking on credit, and the effectiveness of the credit risk management process depends on a number of factors, including staff quality, credit culture, devoted top management bodies, adequate training programs, appropriate organizational structures, ample levels of internal control, and performance of the intermediation function.

Credit risk management encompasses various aspects, including the creation and execution of an appropriate credit risk strategy, policy, and procedure, precise risk assessment, optimal credit granting procedures, credit administration, monitoring and reporting procedures, regulating the frequency and techniques of credit policy and procedure reviews, and assigning authority and responsibility. Moreover, he discussed how to improve credit performance by creating an appropriate credit risk environment, a reasonable credit limit, the best credit granting procedure, appropriate monitoring and control of credit risk, and optimizing risk-return of a bank credit risk management.

The impact of loan arrangements on capital structure, lending, earnings, and risk of banks is empirically examined by the author of [14]. Banks involved in the loan sales market are found to have lower capitalization ratios and to create riskier loan arrangements than other banks. According to their findings, rather than lowering risk in the banking system, developments in credit

risk management have improved credit accessibility. Globally, all financial organizations now have credit risk management as one of their main goals. Credit risk management seeks to maximize return on investment (ROI) for banks while keeping exposure to credit risk within reasonable bounds [15]. According to [10] we can create an effective credit risk management system by creating and implementing a credit risk framework, conducting a credit risk assessment, developing credit risk scoring models and credit risk reporting control panels, forecasting loan losses, and so on. He also thinks that the best credit risk management systems concentrate on people, culture, processes, and organizations because those are the things we work with. Credit risk management encompasses both proactive and reactive strategies. Preventive measures include risk assessment, measurement, pricing, and early warning systems to identify prospective default signals ahead of time and better diversify credit portfolios. Through actions like securitization, derivative trading, risk sharing, and legal enforcement, the curative remedy seeks to reduce post-sanction loan losses [16].

2.2.1 Credit Risk Management Life Cycle

The degree of credit risk management strategy and appropriate credit process implementation in each credit risk management life cycle determine whether the institution is profitable or insolvent. According to [17] states that the credit risk management life cycle is divided into four levels:

1. **Collect loan and borrower information:** Parties required her loan, external data, and borrower. Interpreting financial data, system integration, obtaining borrower and loan data, data quality, and external rating data are the main tasks and problems of this stage.
2. **Calculate credit risk:** At this point, credit risk will be determined using a risk assessment system that takes into account significant variations in exposure and risk across various companies. The primary tasks and obstacles her faces include creating a rating model, estimating the likelihood of a default, rating methodology, data comprehensiveness, model comprehensiveness, calculating loss given default (LGD), exposure at default (EAD), and expected loss (EL).
3. **Risk rating monitoring and management:** This phase involves keeping an eye on and overseeing the risk rating system. The interface with the internal collection, back testing ratings, reducing manual dependency on feedback and alerts, creating a procedure to

manage rating approval, and making sure that external rating changes are communicated to her are her primary tasks and problems.

4. **Managing Portfolio Allocation of Capital:** In today's financial environment, this credit risk management life cycle phase is the most difficult. The most significant tasks and difficulties at this point are calculating and tracking portfolio risk, permitting risk transfer, reporting on risk, and doing scenario analysis and stress testing in addition to the back testing difficulty.

2.3 Credit Process

Making profitable loans with the least amount of risk is the main goal of consumer and commercial lending. Targeting particular areas or sectors where lending officers excel is a good idea for management. The procedures and controls of each bank are what enable management and credit officers to assess the trade-offs between risk and return in the credit process.

2.3.1 Credit Application

The first step in the credit management procedure is the credit application. A loan application is necessary regardless of the loan's amount and purpose. Even though the candidates may think these are straightforward questions, they should recognize the significance of the document. The application materials include comprehensive personal data for the candidate. The applicant's name, residential address, age, phone number, marital status, number of dependents, educational background, hometown, business type, location, years in business, reasons for loan, required amount, repayment period, security pledge, if any, and guarantors are among the other details included in the information [18]. The loan application form includes acceptance and rejection signatures from the loan committee members. This is the most important document in the loan arrangement. This contract specifies that if a borrower fails, the credit union may seek any applicable legal remedies. Given that this is the beginning of the credit management process, any errors committed here will have a big negative influence on the entire process. If the loan application form is not well-designed, a defaulting borrower may be able to avoid legal action. As a result, it is critical to evaluate the current loan application forms to ensure their acceptable structure for protecting credit unions (Ibid).

2.3.2 Credit Information

To make wise lending decisions, one must have access to timely and sufficient information that allows one to evaluate applicants' creditworthiness for bank loans. One of the main elements assuring banks' financial stability is prudent lending decisions based on sufficient information about borrowers' creditworthiness. Nevertheless, it has proven extremely difficult to obtain timely and reliable information about potential borrowers in Ethiopia, which makes it difficult to make such cautious lending decisions. The creation of a Credit Information Center (CIC), where pertinent data about borrowers is presumed to be gathered and made available to lending banks, is one way to address the challenge of obtaining reliable and timely information on potential borrowers.

These instructions to build a Credit Information Center (CIC) have been issued by the National Bank Ethiopia (NBE) in accordance with article 36 of the Licensing and Supervision of Banking Business Proclamation No. 84/1994. The following is a summary of the directive, notwithstanding the fact that there are still significant limits in the accuracy of the credit information extracted:

- Banks must use the online system to furnish, modify, and update credit information about each of their borrowers.
- In response to written requests from banks, the NBE's Supervision Department will give written access to all credit data on potential borrowers found in the Central Database within three business days of the request being received;
- The user group will be the only ones with access to the Central Database;
- The NBE's responsibilities will be limited to managing the Credit Information Sharing system, giving banks written access to credit information about borrowers that is available at the Credit Information Centre, making sure that only authorized individuals have access to the online system to update or modify credit information, and making sure the system is reliable and efficient;

When borrowers provide false, deceptive, or insufficient credit information, the NBE will not be held liable for any losses, demands, or liabilities that may ensue.

- Banks through the Credit Information Centre and exchanged with other banks via the NBE.
- Every bank is required to electronically provide to the Credit Information Centre all of its borrowers' initial credit information as well as any additional pertinent data;

- The Credit Information Centre must receive accurate, comprehensive, and timely credit information from each bank, and this responsibility rests entirely with them. When mistakes occur, the relevant bank must fix them right away;
- Any bank that provides the Credit Information Centre with false, deceptive, or incomplete credit information, or fails to provide the Centre with information that should have been provided in accordance with these instructions, will be solely liable for any resulting losses, demands, or liabilities;
- When determining whether to grant a loan, each bank must consider the credit data about borrowers that it has collected from the Credit Information Centre Central Database. These data must be handled with the highest confidentiality and must not be shared with outside parties or utilized for any other reason;
- Any bank that uses borrower credit information acquired from the Credit Information Centre for reasons other than lending decisions shall be solely liable for any losses, liabilities, or claims that may come from such use or disclosure to third parties.

2.3.3 Credit Analysis

The process of determining whether or not to give credit to a specific consumer is known as credit analysis. Bank officials evaluate all relevant data when a customer requests a loan to ascertain whether the loan satisfies the bank's risk-return goals. In essence, credit analysis is default risk analysis, when a loan officer assesses a borrower's capacity and willingness to repay [19]. In his book, Koch elucidated how Eric Compton distinguished three district areas of commercial risk analysis associated with the following inquiries:

- Which hazards are inherent in the way the firm operates?
- What steps have managers taken to reduce those risks, or not taken?
- How may a lender allocate its risks when providing credit?

In order to answer the first question, the credit analyst must come up with a list of things that could negatively impact a borrower's capacity to repay. The latter acknowledges that a borrower's decisions play a major role in repayment. The final query pushes the analyst to elaborate on risk management strategies so the bank can create a loan arrangement that works. When assigning credit, Pandey (1990) and [19] mentioned in their book that their primary considerations are taken

into account. Character, capital, capacity, condition, and collateral are the five Cs of credit, according to Koch, while Pandey lists the three Cs of credit as character, capacity, and capital.

Character: describes the borrower's reliability, honesty, and capacity for fulfilling responsibilities. The honesty and future repayment intent of the borrower must be evaluated by an analyst. The loan should be denied if there are any substantial uncertainties.

Capital: is used to describe the borrower's wealth position as determined by their market standing and financial stability. Losses are lessened and the chance of bankruptcy is decreased.

Capacity: includes the borrower's legal status as well as management know-how to keep the business running so the company or person can pay back the loan. To pay off debt, a company or an individual must have a recognizable cash flow or income.

Condition: refers to the financial climate or supply, manufacturing, and distribution elements unique to a given sector that affect a firm's operations. Cash-repayment sources frequently fluctuate based on customer demand or the business cycle.

Collateral: Collateral serves as a backup source of funding or security for the lender in the event of default. While having an asset the bank can take possession of and sell off in the event of a default lessens loss, it does not excuse lending proceeds at the time of credit decision-making

2.4 Functional Terminology Definition

- **Non-Performing Loan (NPL):** refer to loans where the borrower has failed to make interest or principal payments for a specified period, typically 90 days or more.
- **Provision:** refers to the amount set aside by financial institution cover potential losses on loans and other assets.
- **Loan Pricing:** is the process of determining the interest rate and other terms that alender will offer on a loan. This involves assessing various factors to ensure that the loan is both attractive to the borrower and profitable for the lender.
- **Credit processing Pool:** refers to a centralized system or a collaborative group of financial institutions or lenders that collectively handle the processing and assessment of credit application.
- **Credit portfolio:** is the collection of all loans and credit exposures held by a financial institution or lender

- **Credit Risk:** is the possibility that a borrower will fail to meet their obligations in accordance with agreed terms, leading to a loss for the lender.
- **Credit Risk Management:** refers to a critical function for financial institutions, involving a range of strategies and practices to minimize the potential for borrower default and mitigate the impact of defaults when they occur.
- **Discretional Lending Limit:** refers to the maximum amount of credit that a bank or financial institution allows its staff to approve without needing higher-level authorization. This limit is set to manage risk and ensure that large lending decisions undergo appropriate scrutiny and approval processes.

2.4.1 Classification of Loan Status

A. **Normal (Pass):** This category of loans and advances is completely protected by the borrower's present financial and payment ability and is not open to dispute. This type of loan is fully secured, meaning that principal and interest payments are covered by cash or cash equivalents, regardless of past due amounts or other negative credit characteristics.
B. **Special Mention:** Any loan or advance that is past due by at least thirty days but not more than ninety.
C. **Substandard:** Fulfilling loans or advances that are above 90 days but under 180 days.
D. **Doubtful:** Non-performing loans are those that are 180 days or more past due but less than 360 days.
E. **Loss:** Non-performing loans or advances that have been past due for 360 days are classified as losses.

2.5 Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on the use of data and algorithms to mimic how humans perceive things and improve their accuracy over time. ML is the science of providing data and information to computers without explicitly programming them to learn and act like humans [20] Depending on the behavior of the problem and data, machine learning uses different types of algorithms. The datasets are fed to the algorithm and the system learns from each pattern of the data, then the algorithm can predict when new data is fed to the system. This indicates that, first machine learning can learn using historical data and then make decisions for new input data [21].

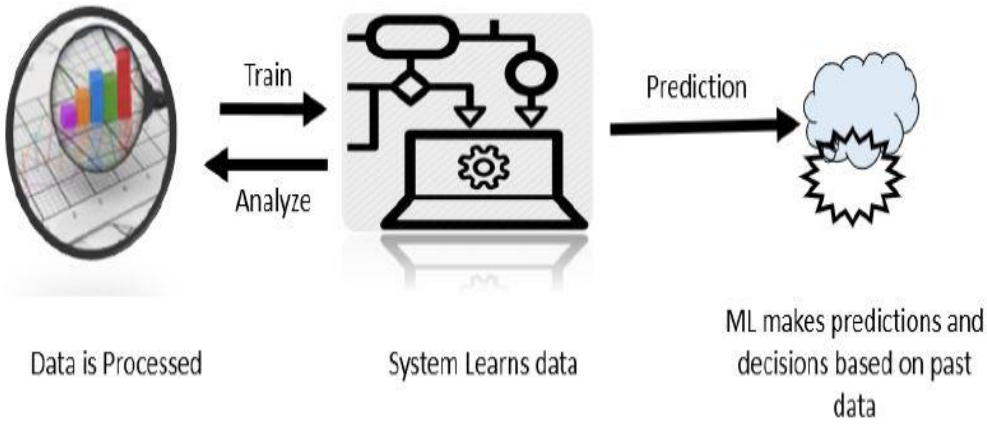


Figure 2. 1 Machine learning working process

Machine learning algorithms allow the systems to make decisions autonomously without any external support. Such decisions are made by finding valuable underlying patterns within complex data [20] Generally machine learning is specified into descriptive and predictive analytics. Descriptive analytics are used to describe the general property and aspects of the data, whereas, predictive analytics are used when you need to know anything about the future.

Based on the learning capability, type of input and output data, and the behavior of the problem machine learning algorithms are categorized into supervised, unsupervised, reinforcement, and semi-supervised learning [21]

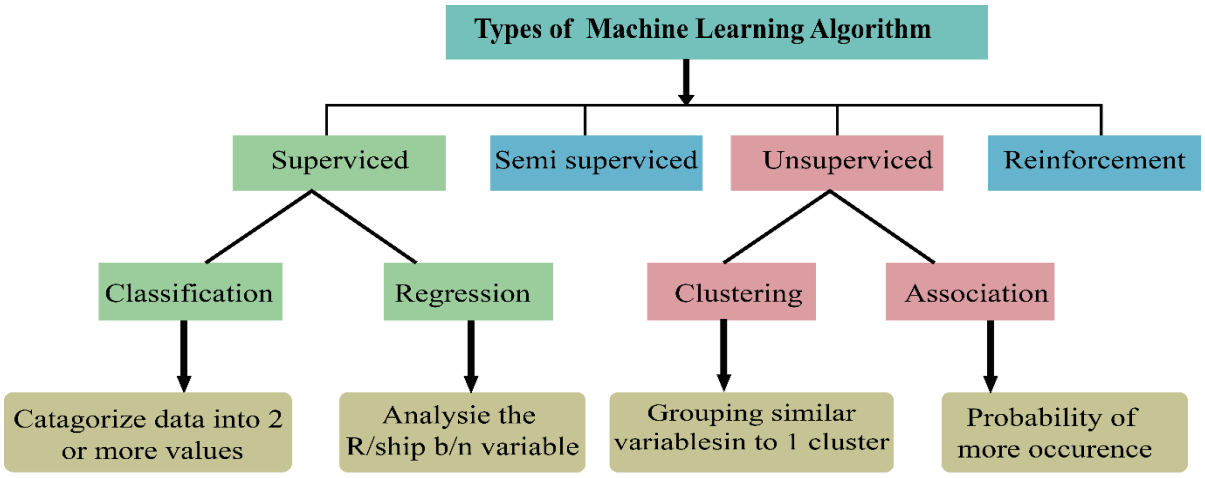


Figure 2. 2 Types of machine learning algorithms

2.5.1 Supervised Learning

Supervised learning is a sort of machine learning that is used when data has an input variable or attributes and produces a target or output value. The algorithm does this by learning from the input data and predicting the output value. Data pre-processing in supervised learning includes operations such as data cleaning, normalization, transformation, feature extraction, selection, and so on. The final training set is the result of data pre-processing [22]. In general, there are two types of supervised learning: (1) classification, and (2) regression [23].

Classification

Classification is a machine learning function that uses labeled data to train and classify new data into two or more categories. The main goal of the classification is to classify the target class from the input data accurately. The simplest type of classification is the binary classification.

This thesis uses a multiclass classifier on data that has five classes: NORM, SPME, SUBS, DOUB, and LOSS. Nondefaulters are customers who repay their loan in the time given to settle the loan, while defaulters are those who cannot repay their loan in the time given to settle the loan.

Some common classification algorithms are logistic regression, J48 decision tree types, gradient boosting machines, Naïve Bayes, random forest, SVM (support vector machine), Multi-layer perceptron, and K-Nearest Neighbor. From these algorithms, we selected to use random forest, Multi-layer perceptron, and logistic regression in this thesis.

2.5.2 Unsupervised Learning

Unsupervised learning, in contrast to supervised learning, is a type of machine learning that can find new patterns in unlabeled data and is used when there is only input data and no output. To take advantage of the vast amount of unlabeled data, unsupervised learning techniques are employed to learn complicated, highly non-linear models with millions of parameters [20].

The goal of unsupervised learning algorithms is to learn and cluster unlabeled datasets. These algorithms realize hidden patterns or data groupings without the interaction of human intervention. Unsupervised ML algorithms are classified into clustering and association as shown in Figure 2.2.

A. Clustering

Clustering is the process of grouping comparable things into a single cluster. The algorithms learn the patterns in the input data to group the comparable things into a single cluster. Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects based on similarities and similar traits [[20]. The most common machine-learning clustering algorithms are K-means clustering, Mean-Shift Clustering, Hierarchical Clustering, and Spectral clustering.

B. Association

Associative learning is a type of unsupervised rule-based machine learning that identifies significant relationships between features in a dataset. If someone burns themselves on a hot stove, they may learn to link hot stoves with pain and avoid touching them. K-means clustering, hierarchical clustering, and Self-Organizing Map(SOM) are the most prevalent unsupervised Machine learning techniques.

2.5.3 Semi-supervised Learning

Semi-supervised learning is a type of machine learning that combines labeled and unlabeled data sets. Semi-supervised learning is a term that refers to a combination of supervised and unsupervised learning. There is a limited amount of labeled data and a huge number of unlabeled data [24]. The semi-supervised approach aims to solve the scarcity of labeled data by first creating clusters of the unlabeled data using the unsupervised learning technique, and then labeling the clusters using the labeled dataset and supervised learning.

2.5.4 Reinforcement Learning

Reinforcement learning is a sort of machine learning that creates a training sequence based on rewarding selected behaviors and punishing those that aren't. An artificial agent receives either rewards or penalties for the behaviors it does during the learning process. The objective is to maximize the total reward [24]. Figure 2.2 shows the reinforcement information exchange.

2.6 Description of Machine Learning Method Used

Machine learning uses algorithms to turn a data set into a model that can identify patterns or make predictions from new data. Which algorithm works best depends on the problem [25]

Machine learning algorithms are the engines of machine learning, meaning it is the algorithms that turn a data set into a model. Which kind of algorithm works best (supervised, unsupervised, classification, regression, etc.) depends on the kind of problem you're solving, the computing resources available, and the nature of the data [25]. In this study three machine learning algorithms namely Logistic Regression, Multi-Layer Perceptron, and Random Forest have been used.

2.6.1 Logistic Regression

Logistic regression estimates the probability of an event occurring in the feature based on previously provided data [26]. When the dependent or target variable has two values (0/1, yes/no, true/false) depending on the provided independent variables, the classification algorithms known as logistic regression is employed. Artificial neural networks and logistic regression are the models of choice for many medical data classification problems. [27].

There are two different types of logistic regression models: multinomial logistic regression and binary logistic regression. When the dependent variable is categorical or continuous and the independent variables are either continuous or discrete, binary logistic regression is utilized. Multinomial regression explains the relationship between one nominal dependent variable and one or more independent variables. When the dependent variable is dichotomous and has more than two categories, a multinomial logistic regression can be used [27].

$$Y = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

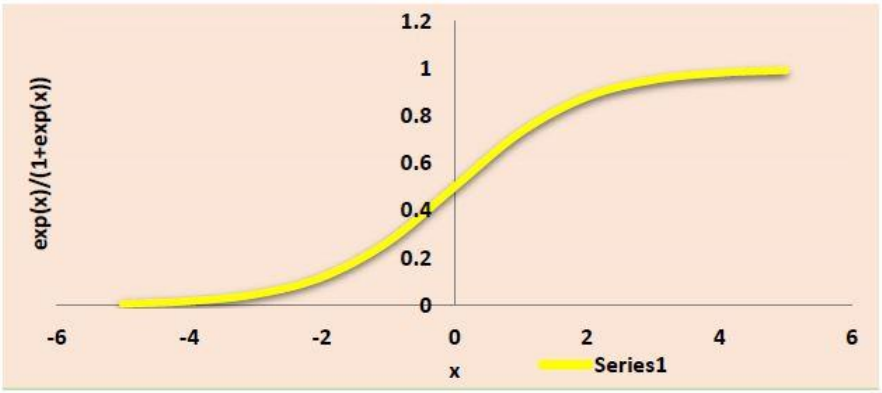


Figure 2. 3 Graph of Logistic Regression Curve

Where **α** and **β** determine the logistic intercept and slope.

Numerous benefits come with logistic regression, including its easy extension to several classes (multinomial regression), speedy training, lightning-fast classification of unknown records, resistance to overfitting, and high accuracy for a wide range of basic datasets. The most popular models in biomedicine right now are artificial neural networks and logistic regression, with 28,500, 1100, and 1300 references to journal articles for decision trees, neural networks, and k-nearest neighbors, respectively, according to several publications indexed in MEDLINE, the National Library of Medicine's premier bibliographic database that has more than 29 million references to journal articles [27].

2.6.2 Multilayer Perceptron

Multi-layer Perceptron (MLP), a supervised learning approach, is a type of feedforward artificial neural network (ANN). It is also known as the foundation architecture of deep neural networks (DNN) or deep learning. [28] A typical MLP is a fully connected network that consists of an input layer that receives input data, an output layer that makes a decision or prediction about the input signal, and one or more hidden layers between these two that are considered the network's computational engine [29]. The output of an MLP network is determined using a variety of activation functions, also known as transfer functions, such as ReLU (Rectified Linear Unit), Tanh, Sigmoid, and Softmax. [28], [30]. To train MLP employs the most extensively used algorithm "Backpropagation" a supervised learning technique, which is also known as the most basic building block of a neural network. During the training process, various optimization approaches such as Stochastic Gradient Descent (SGD), Limited Memory BFGS (L-BFGS), and Adaptive Moment Estimation (Adam) are applied. MLP requires tuning of several hyperparameter such as the number of hidden layers, neurons, and iterations, which could make solving a complicated model computationally expensive. However, through partial fit, MLP offers the advantage of learning non-linear models in real-time or online [28].

MLP consists of multiple layers of nodes or neurons, organized into three main types of layers: an input layer, one or more hidden layers, and an output layer. Each connection between nodes has an associated weight, and the network learns by adjusting these weights based on training data.

A multiclass MLP (Multi-Layer Perceptron) is a type of neural network designed for solving classification problems with more than two classes. In a multiclass classification task, the goal is to assign each input instance to one of several possible classes.

- **Output Layer Activation Function:** the activation function of the output layer is typically chosen based on the nature of the problem.
    - The activation function of the output layer is typically chosen based on the nature of the problem.
    - For binary classification, a sigmoid activation is often used.
    - For multiclass classification, a softmax activation function is commonly employed. Softmax normalizes the output into a probability distribution across multiple classes.
- **Output Layer Neurons:**
    - The number of neurons in the output layer should match the number of classes in your problem.
    - Each neuron in the output layer corresponds to a class, and the softmax activation ensures that the outputs sum to 1, forming a probability distribution.
- **Loss Function:**
    - The choice of loss function depends on the problem. For multiclass classification, categorical cross entropy is commonly used.
    - Categorical cross-entropy is suitable for problems where each instance belongs to exactly one class.
- **Encoding of Target Labels:**
    - The target labels need to be one-hot encoded for multiclass classification.
    - Each target label is represented as a binary vector where only the index corresponding to the true class is set to 1, and the others are 0.
- **Number of Hidden Layers and Neurons:**
    - The architecture of the hidden layers can vary based on the complexity of the problem and the size of the dataset.
    - Experiment with different numbers of hidden layers and neurons to find a suitable architecture. Techniques such as cross-validation can help in model selection.

- **Activation Functions in Hidden Layers:**
    - Common activation functions for hidden layers include Relu (Rectified Linear Unit) and its variants. Experiment with different activation functions to find what works best for your specific problem.
- **Regularizations:**
    - Consider using regularization techniques such as dropout to prevent overfitting, especially if you have a limited amount of training data. [31]

2.6.3 Random Forest

Random forest is a collection of decision trees and there are a lot of differences in their behavior. Random Forest has a significant performance improvement when it is compared to a single tree classifier like C4.5 [32]. Bagging is a principle used by Random Forest. Bagging also known as bootstrap aggregation is an ensemble technique used by random forest, which chooses a random sample from the data set. Hence each model is generated from the samples (bootstrap samples) provided by the original data with replacement known as row sampling, it is a step of row sampling with replacement called bootstrap. Each model is trained independently and generates results. Aggregation is the process that involves combining all results and generating output based on majority voting.

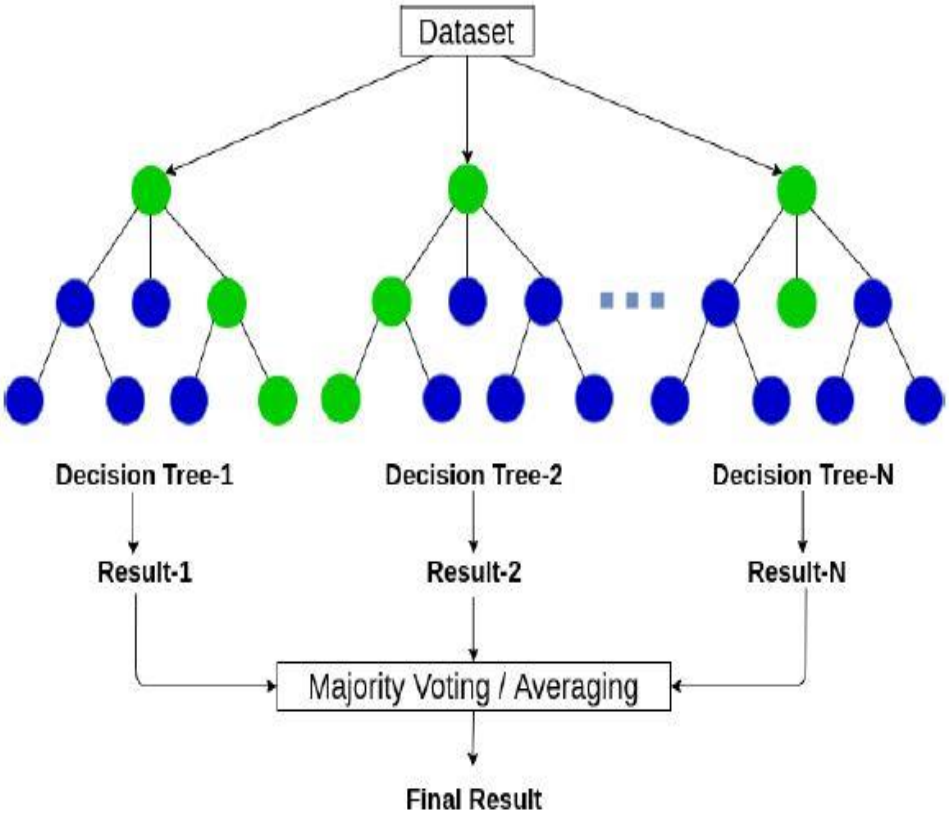


Figure 2. 4 Random forest classifier

The RF Decision Tree is a forest of decision trees that are generated at random, with each node using a random selection of features to determine the output. Using majority voting, the random forest then combines the results of each individual decision tree to produce the ultimate result.

Important Features of Random Forest

- **Diversity-** Each tree uses different attributes while making an individual tree not all attributes are used in a single tree.
- **Immune to the curse of dimensionality-** Each tree does not consider all the features so the feature space is reduced.
- **Parallelization-** We can make full use of the processor to build random forests by creating each tree independently out of different data and attributes.
- **Train-Test split-** In a random forest we don't have to segregate the data for training and testing as there will always be 30% of the data which is not seen by the decision tree.
- **Stability-** The result is based on majority voting or averaging so stability arises.

Hype parameters are used in random forests to either speed up the model or improve its performance and prediction ability. Random forests are far more successful than choice trees if the trees are acceptable and diversified. When compared to a single decision tree, Random Forest produces more accurate and exact answers and performs better in categorization when there are more occurrences. Additionally, it resolves the over-fitting issue and missing values issue brought on by missing values in the datasets. Consequently, Random Forest is advised for a range of classification issues with confidence if one must select among the tree-based classifiers provided for those tasks [33].

2.6.4 Ensemble Model

Ensemble methods are prevalent in machine learning and statistics. Ensemble methods offers techniques to merge multiple single classifiers or predictors to form a committee, to achieve amassed decision for better and accurate results than any of the single or base predictors [34] [35]. Thus, the ensemble method highlights the strong point and watered-down the feebleness of the single classifiers [34] [35]. Two types of ensemble methods are defined by Opitz and Maclin [35], namely: cooperative and competitive ensemble classifiers. Ensemble involves training diverse single classifiers independently with the same or different dataset, but not with the same parameters. Then, the final prediction (expected output) is obtained by finding an average of all individual single or base classifier outputs (or other similarities). Whiles the cooperative ensemble is a divide and conquer-based approach. The prediction task is subdivided into two or more tasks, where each subtask is sent to the appropriate single classifier based on the characteristics and nature of the subtasks, and the final prediction output is obtained by the sum of all distinct single or base classifiers. In the creation of ensemble classifier and regresses models, three factors need careful consideration. (1) The availability of numerous classification and regression methods makes it difficult to identify which one of them is suitable for the application domain. (2) The number of single classifiers or repressors to assembled for better and higher accuracy. (3) The amalgamation techniques are suitable for combining the outcomes (outputs) of the various single classifiers and repressor to obtain the final prediction or output. We present a brief discussion of some basic and advanced combination techniques for EL in the subsequent section.

Advanced ensemble techniques

The following section discusses three advanced combination techniques in brief.

Stacking

Stacking is an ensemble learning technique that makes use of predictions from several models (m1, m2,...,mn) to construct a new model, where the new model is employed for making predictions on the test dataset. Stacking seeks to increase the predictive power of a classifier [36]. The basic idea of stacking is to "stack" the predictions of (m1,m2,...,mn) by a linear combination of weights aj,,(i = 1,...,n) as expressed in Eq. (10)

Blending

The blending ensemble approach is like the stacking technique. The only difference is that, while stacking uses test dataset for prediction blending uses a holdout (validation) dataset from the training dataset to make predictions. That is predictions take place on only the validation dataset from the training dataset. The outcome of the predicted dataset and validation dataset is used for building the final model for predictions on the test dataset.

Bagging

Bagging also called bootstrap aggregating involves combining the outcome of several models (for instance, N number of K-NNs) to acquire a generalised outcome. Bagging employs bootstrapping-sampling techniques to create numerous subsets (bags) of the original train dataset with replacement. The bags created by the bagging techniques severs as an avenue for the bagging technique to obtain a non-discriminatory idea of the sharing (complete set) [48]. The bags' sizes are lesser than the original dataset. Some machine learning algorithms that use the bagging techniques are bagging meta estimator and random forest. BAG seeks to decrease the variance of models.

Boosting

Boosting also called "meta-algorithm" is a chronological or sequential process, where each successive model tries to remedy or correct the errors of the preceding model. Here, every successive model depends on the preceding model [57]. A boosting algorithm seeks to decrease the model's bias. Hence, the boosting techniques lump together several weak-learners to form a strong leaner. However, the single models might not achieve better accuracy of the entire dataset;

they perform well for some fragment of the dataset. Therefore, each of the single models substantially improves (boosts) the performance of the ensemble. Some commonly boosting algorithms are AdaBoost, GBM, XGBM, Light GBM and CatBoost.

2.7 Model Evaluation

The methodology for assessing the model loan default prediction's performance is provided in this section. The confusion matrix is used to evaluate the effectiveness of a classification-based machine learning model [37]. This table is frequently used to show how well a classification model performs when applied to a set of test data for which the actual values are known [37]. A confusion matrix is a table used to summarize the number of correct and incorrect predictions made by a classifier (or classification model) in multiclass classification tasks. This study employed the confusion matrix as an evaluation tool to assess the model's performance [37].

2.7.1 Confusion Matrix

A confusion matrix is a technique for calculating the performance of ML classification models. It's just a matrix comparing actual categorical values to expected categorical values .By comparing the actual and predicted classifications that visualizes a classifier's accuracy. When we accurately anticipate actual values, we call them a true positive, and when we predict incorrect values as incorrect, we call them a true negative. A false positive occurs when we anticipate a value that does not occur, and a false negative occurs when we do not predict a value that does occur [38].

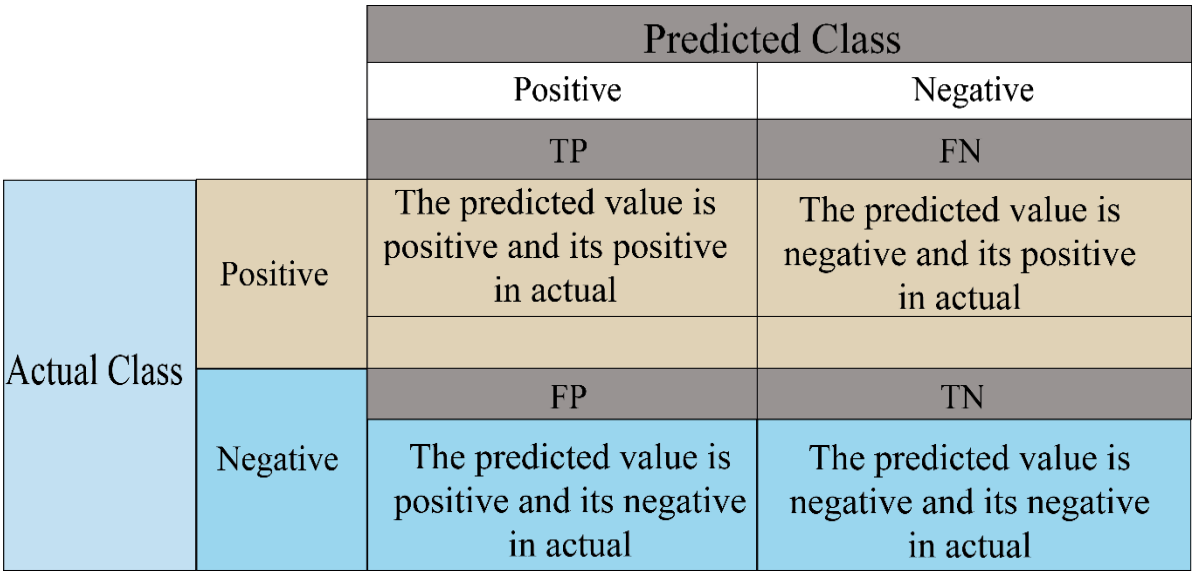


Figure 2. 5 Confusion Matrix

Confusion Matrix has four dimensions

- **True Positives (TP)** − the number of instances that were correctly predicted as belonging to a particular class.
- **True Negatives (TN)** − the number of instances that were correctly predicted as not belonging to a particular class.
- **False Positives (FP)** − the number of instances that were incorrectly predicted as belonging to a particular class when they belong to a different class.
- **False Negatives (FN)** − the number of instances that were incorrectly predicted as not belonging to a particular class when they belong to that class.

Accuracy

Accuracy is a measures the proportion of correctly classified instances among all the instances. [35]. is the ratio of the number of correct predictions to the total number of prediction.

The accuracy metric is shown in equation 1

$$Accuracy = \frac{Tp + TN}{TP + FN + FP + TN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.. (1)$$

Misclassification:

The Misclassification refers to instances where a classification model incorrectly predicts the class label of an instance. It occurs when a predicted label does not match the true label of the instance in the dataset.

$$Mis\ classification = FP + FNTP + TN + FP + FN \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.. (2)$$

Precision

Precision measures the accuracy of positive predictions This determines whether the model is reliable or not.

Precision focuses on the relevance of the model's predictions for a specific class and is calculated using the formula [35].

$$Precision = \frac{TP}{TN + Fp} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Recall (sensitivity)

Recall measures the ability of the model to correctly identify instances of a class. It is particularly concerned with the model's ability to correctly identify all positive instances, including those that are falsely predicted as negative.

$$Recall = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

A higher recall will result in the majority of positive instances (TP+FN) being classified as positive (TP), which will increase the number of FP measurements and decrease overall accuracy [35]. A low recall corresponds to a high FN, indicating that the data should be positive but is being classified as negative. This implies that in the event that a positive example is discovered, the model will be more convinced that it is a true positive.

F1-Score

F1-Score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall.

$$F1 - score = 2x \frac{Precision\ x\ Recall}{Precision + Recall} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Since the F1-Score is difficult to interpret, we must use it in conjunction with other assessment metrics to obtain a whole picture. Without these, we are unable to determine if the classifier is optimizing recall or precision.

False Positive rate

The false positives rate refers to instances where the model incorrectly predicts a class as positive (or present) when it should not have [60].

FPR=$FPTN+FP$...(5)

2.8 Related Work

The research work [39], compared a decision tree model, the Cox model, and logistic regression in their study. The purpose of their study on loan defaults was to identify loans that would fail within a year of being awarded, in addition to predicting which loans would default overall. Furthermore, the most crucial factors for forecasting defaults on short-term loans were presented. They said that loans that default soon after being granted cause the lender to suffer significant

losses. In their investigation, sensitivity and precision were employed as evaluation indicators. The accuracy with which the models predicted short-term defaults was used to compare the results.

Their decision tree model reached a specificity of 81.9% and a precision of 83.3%. Logistic regression had a sensitivity of 46% and a precision of 64.2%, whereas the Cox model achieved 45.5% sensitivity and 45.5% precision in predicting short-term loan defaults. The most important variables in predicting short-term loan defaults were variables "background", "macro", "liability_ratio", "fixed_ratio", "DSR", "competitor evaluation" and "law".

Study [40], conducted a default probability study on a peer-to-peer lending data set. They compared the ensemble mixture random forest model (EMRF) with a standard mixture cure model, the Cox proportional hazards model, and the logistic regression model. They stated that their EMRF model outperformed all the other models based on the mean area under the ROC curve.

According to [41], studied default risk with logistic regression and discriminant analysis. Tunisian commercial bank's data set consisting of 603 loans was used in the study. The default rate was higher compared to other loan default studies, reaching 56.55%. Their logistic regression model had 99.41% sensitivity and 98.47% specificity. The discriminant analysis provided a sensitivity of 75.36% and a specificity of 59.54%, thus they declared logistic regression to be more powerful. Variables "loan amount", "outstanding loan" and "socio-professional category" were relevant in the logistic regression model.

The study [42], studied default risk with logistic regression on a Portuguese credit data set. The data set consisted of 3221 individuals with a 10% default rate. In the result analysis, they provided relevant variables and logistic regression's accuracy metric was compared to other studies. Variables "Spread", "Term", "Age", "Credit cards", "Salary" and "Tax echelon" were found relevant and the percentage of correctly classified cases was 89.79%. A sensitivity of 0.94% and a specificity of 99.55% can be counted from the presented confusion matrix.

The research work [43], performed their loan default predictions on a data set from a credit assessment company. The data set had 50 000 observations, 350 columns and according to them, more than 70% of the observations were non-defaulted loans. They compared seven machine learning algorithms, including logistic regression, decision tree, random forest, and gradient boosting tree. Evaluation metrics used were accuracy, f1 score, and the AUC. Their logistic

regression model reached an AUC of 0.84, with 74.43% accuracy. The decision tree had an AUC of 0.85, with 84.68% accuracy. Random forest had an AUC of 0.96 and an accuracy of 88.96%, whereas gradient-boosting trees reached an AUC of 0.97 and an accuracy of 90.99%.

According to [44], studied credit card default prediction with logistic regression, decision tree, ad boosting, and random forest. In addition, he created weighted models for each model to overcome data imbalance. The results were compared with each other based on accuracy. Confusion matrices were presented but they were only used to calculate accuracy.

Table 2. 1 Summary of Related work

Research Title	**Authors**	**Algorithms**	**Limitations**
Default Probability Study on Peer-to-Peer Lending Data	Xu, J., Lu, Z. & Xie, Y. (2021)	Random forest(RF), extreme gradient boosting tree(XGBT), gradient boosting model (GBM), neural network (NN)	Data source, sample size, time period, model selection, model interpretability, causal inference, external validity
Default Risk Analysis with Logistic Regression and Discriminant	Abid, L., Masmoudi, A. & Zouari-Ghorbel, S. (2018).	Logistic Regression, Discriminant Analysis	Small dataset size, limited analysis of model performance
Predicting Loan Default Risk: A Portuguese Credit Data Set	Silva, E.C., Lopes, I.C., Correia, A. & Faria, S. (2020).	Logistic Regression	Did not explore other machine learning algorithms or feature selection
Loan Default Prediction Using Seven Machine Learning Algorithms	Tian, Z., Xiao, J., Feng, H. & Wei, Y. (2020)	Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Tree	Focused on algorithm comparison, did not provide in-depth analysis of model performance or feature importance
Credit Card Default Prediction with Weighted Models	Yu, Y. (2020, August)	Logistic Regression, Decision Tree, Random Forest	Limited evaluation metrics(only used accuracy), did not consider data balancing techniques

CHAPTER THREE

METHODOLOGY

3.1 Overview

This section involves the building, formatting, and rigorous cleansing of the data to make sure its quality satisfies the requirements for the chosen analytic techniques, after thorough identification and careful selection of data sources. To get the dataset ready for later modelling techniques, actions like data coding, extensive data cleaning, attribute selection, and data transformation are performed.

This thesis main goal is to create an accurate loan default prediction model by thoroughly studying loan data.

3.2 Business Understanding

Lending is the amount of money provided by a lender and taken by a borrower, payable at some future date on specific terms and conditions that are governed by a legal contract. It can also be defined as an offer of capital to a person or entry with the expectation that repayment would be made with interest either by instalments or in one amount by a specified date, otherwise, where necessary a lender will protect himself by asking the customer to provide collateral [45]. Banks and financial institutions like Wegagen Bank are involved in credit facilities to serve society as a source of credit and to obtain profit from the credit they disburse. Each bank and financial institution in Ethiopia is governed under the rules and regulations of the national bank. The National Bank of Ethiopia controls credit activities in every financial institution including determining interest rates, collateral types and conditions, loan types, rights and obligations of the creditor and the borrower, etc. If a loan customer does not make a loan repayment according to the agreed schedule creditors use legal actions including asking the customer into the court. This is usually time-consuming, expensive, and may not always be successful. As a result, proper customer identification is very useful. It is necessary to estimate the loan status of a customer by learning from the previous loan history of other customers using machine learning techniques. By using the loan history of customers this method predicts whether a customer will repay his loan in a normal way or will default. There are many reasons for a loan to default [45]. These include poor management, unwillingness to repay the loan, willful negligence and improper appraisal by credit officers, domestic products decline, exchange rate depreciation, economic shocks, occupation

type, marital status, scale of operation, education level of the borrower, age of the borrower and the like. These attributes determine the healthiness or creditworthiness of a given loan. Nowadays the use of bank services is increasing in Ethiopia and the rest of the world. Bank customers deposit their money in banks, transact by the account they have, etc. As customers deposit their money they need interest on the money they deposited depending on the bank policy. To pay interest for the deposited money a bank should invest the deposited money as a loan to other customers and to get better interest. The success of a bank highly depends on the level of loan status. In this age of digitization, much of the money a country has is found in banks. These allow financial institutions to lend a lot of money to their customers. The loan investment that banks lend plays a great contribution to the economic growth of the country. A loan that is lent to the right customer will contribute to the country's economy positively and a loan disbursed to the wrong customer will be an obstacle to the growth of a country's economy. As a result, financial institutions need to identify loan customers whether the customer is a good, moderate customer, or bad customer in loan repayment. Good loan customers are those who repay their loan on the given schedule. For good loan customers, no default event will occur. Moderate customers repay their loans with minimum dalliance of the agreed schedule. For moderate customers, a substandard default type occurs. Bad customers are those customers whose loan is considered as almost loan due to much dalliance. A loan disbursed to the bad customers would more probably go to default. A default is a condition where a loan repayment would be delayed from the agreed schedule. Banks have different mechanisms to identify their customers. Before they disburse a loan they use some measurements like demographic information and the customer history in the bank. They collect demographic information like the customer's age, gender, location, marital status, employment conditions, income level, loan purpose, loan amount, loan guarantee, and the like. In countries like Ethiopia, this happened without the assistance of a computer. I.e. They do this by manpower. Credit activity in financial institutions is done by credit officers found at the district level. Managers and Credit officers are responsible for selecting appropriate customers for a loan who will repay the loan in the given time by using different parameters like collecting some information about the customer or from the previous experience of a customer in the bank to prevent wrong loan for wrong customer. They make manual estimations of customers for loan candidates.

3.3 Data Understanding

After understanding the business where the problem is located, the next step is to understand and analyze the data that is available for analysis and model building.

3.3.1 Data source Description

For this study, all the data was collected from Wegagen Bank core banking system. The core banking system encompasses numerous tables that detail various credit activities and other operational modules within the bank. From this extensive database, tables relevant to customer and loan contract records were extracted.

The total data comprises 18,743 records with 20 features, including one class features categorized into five classes Normal, Special Mention, Substandard, Doubtful, and Loss.

The decision to use data solely from Wegagen Bank was influenced by several critical considerations. Firstly, gaining access to banking data from multiple institutions would require navigating a complex and lengthy process of obtaining permissions and approvals. Banks are generally highly protective of their data due to stringent privacy and confidentiality regulations.

Furthermore, the dataset from Wegagen Bank was deemed comprehensive and representative enough for the purposes of this study. The selected features encompass various borrower characteristics and loan details that are pivotal in assessing credit risk. These features include, but are not limited to, business location, loan product type, yearly business income, location of collateral, total years of experience, and educational status of borrowers. These attributes are critical in the real-world loan risk assessment practices employed by financial institutions.

The chosen dataset provides a robust foundation for the research, offering sufficient variability and detail to build and validate predictive models effectively. By using this dataset, the study aims to improve the accuracy of loan default predictions, thus providing a more reliable and effective risk assessment tool for financial institutions. This focus on a single, well-documented dataset also ensures that the analysis remains consistent and manageable, facilitating more precise and actionable insights.

The features description and categorical values explored for the study are described in the below table:

Table 3. 1 Feature and description

S.N	Features	Description
1.	Loan Product	The specific type of loan offered by a lending operation that is certified by a regulating authority.
2.	Business Income	Income received from the sale of products or services.
3.	Business Experience	Work experience in business.
4.	Business Location	Place is concerned with how goods and services reach customers.
5.	Sex	The characteristics of women and men.
6.	Marital Status	The state of being married or not married.
7.	Educational Background	Educational attainment individuals.
8.	Approved Amount	The maximum amount the Lender is willing to fund the Loan.
9.	Disbursed Date	The dates on which the Bank funds the Loans.
10.	Loan Exp. Date	The date on which the term of the loan expires and the outstanding principal balance of the loan must be repaid to the lender.
11.	Interest Rate	The cost of debt for the borrower and the rate of return for the lender.
12.	Collected Total	the amount shown in the monthly Report as the sum of Branch Account Collections,
13.	Collaterals Value	The amount of assets that have been put up to secure a loan.
14.	Status	The indicator where your loan is in the process.

3.4 Data Preprocessing

Data available for mining is raw data and may be in different formats as it comes from different sources, it may consist of noisy data, irrelevant attributes, missing data, etc. Data needs to be pre-processed before applying any kind of machine learning algorithm [46]. We have applied pre-processing tasks to the data.

3.4.1 Data Cleaning

Among the total dataset extracted, i.e., 18,743 records, 559 of them have missing values for features such as educational background and business income. Instead of imputing these missing values with statistical methods like mean, median, or mode, a row deletion approach was

employed. This decision was justified based on several considerations. Firstly, the proportion of missing values constituted a relatively small fraction of the entire dataset (approximately 3.1%). In such cases, removing rows with missing values is often preferred to avoid introducing potential biases or inaccuracies that can result from imputation techniques. Imputing missing values might distort the natural distribution of the data, especially for features like business income, which could vary widely among different borrowers. Secondly, the integrity of the dataset is critical for ensuring accurate model training and prediction. Features like educational background and business income are essential for predicting loan defaults, and any imputation method could compromise the quality of the data. By deleting rows with missing values, we maintain the dataset's integrity, ensuring that the machine learning models are trained on accurate and reliable data.

3.4.2 Feature Selection

Feature selection from the data set was done based on the objective of the study at hand. Hence the account number, customer’s name, spouse name, maker name, checker name and branch code features are removed in order to piracy customers or bank and reduce the data to only most important ones. Some of the feature before collected the data deleted by the bank; this would minimize the effort required for further processing.

3.4.3 Feature Encoding

Feature encoding is a crucial step in the data pre-processing pipeline, especially when dealing with categorical data that needs to be converted into a numerical format for machine learning models to process effectively. In this study, the dataset included several categorical features such as business location, loan product type, and educational status of borrowers. To encode these categorical variables, the One-Hot Encoding technique was employed.

In this study, features such as 'business location', 'loan product type', and 'educational status' were encoded using One-Hot Encoding. This process was implemented using the pd.get_dummies() function in Python, which efficiently handles categorical data and converts them into the required binary format. By applying One-Hot Encoding, the dataset was transformed into a numerical format, making it ready for the subsequent machine learning modelling steps. This ensured that all features were represented appropriately, thereby enhancing the model's ability to learn from the data and make accurate predictions.

3.4.4 Data Normalization

Data normalization is one of pre-processing step in machine learning that ensures each feature contributes equally to the model. Normalizing the data helps improve the efficiency and performance of many machine learning algorithms, particularly those that are sensitive to the scale of the input data, such as logistic regression and neural networks. In this study, data normalization was applied to the numerical features to bring all the data into a common scale without distorting differences in the ranges of values. This was achieved using the Min-Max Scaling technique, which transforms the features to a fixed range, typically [0, 1].

3.4.5 Data Splitting

To ensure the development of robust and reliable predictive models, the collected dataset was split into training and testing sets. This splitting is a crucial step in the machine learning pipeline, as it allows for the evaluation of model performance on unseen data, thereby mitigating the risk of overfitting and ensuring the generalizability of the models. In this study, a 70/30 split ratio was used. This means that 70% of the dataset (12,729 records) was allocated to the training set, while the remaining 30% (5,455 records) was reserved for testing. This split ratio is chosen to ensure a significant portion of the data is used for training while also providing a substantial amount of data for evaluation.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 3. 1 Data splitting

3.5 Training and Test Dataset

For the purpose of developing and evaluating the loan default prediction models, a well-structured dataset was utilized. This dataset, extracted from a core banking system, comprises 18,184 records encompassing various attributes pertinent to loan activities, customer profiles, and credit information.

3.5.1 Training Dataset

The training dataset consists of 12,728 records, which represent approximately 70% of the total dataset. This set was used to train the machine learning models, enabling them to learn the intricate patterns and relationships within the data. The attributes in the training dataset include:

Customer Information: Demographic details such as age, sex, marital status, and educational background.

- Loan Details: Information regarding loan products, approved amounts, interest rates, collateral values, business income, and loan expiry dates.
- Business Information: Business location, years of business experience, and business income.
- Loan Status: The target variable indicating the loan status, categorized into NORM (normal), SPME (special mention), SUBS (substandard), DOUB (doubtful), and LOSS.

The training dataset was preprocessed to handle missing values, encode categorical variables, and normalize numerical features.

3.5.2 Test Dataset

The test dataset comprises 5,456 records, making up approximately 30% of the total dataset. This set was reserved for evaluating the performance of the trained models. The attributes in the test dataset are identical to those in the training dataset, ensuring consistency in the evaluation process.

Table 3. 2 Summary of training and testing dataset

Class Total	No of Training Sample	No of testing sample	Total sample
NORM	10,137	4344	**14,481**
SPME	1025	440	**1465**
SUBS	374	160	**534**
DOUB	640	275	**915**
LOSS	552	237	**789**

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 Overview

In this chapter, discuss how to develop an optimal predictive model for loan default using machine learning techniques. The dataset utilized comprised 18,184 records from a financial institution's core banking system. To ensure robust model training and evaluation, the dataset was divided into a training set of 12,728 samples and a testing set of 5,456 samples, maintaining a standard 70/30 split ratio. Three individual base models were employed: Logistic Regression, Multilayer Perceptron (MLP), and Random Forest. These models were chosen for their distinct methodologies and strengths in handling classification tasks. Additionally, a blending ensemble model was constructed to combine predictions from these base models, aiming to leverage their complementary strengths and improve overall predictive performance. Performance evaluation was conducted using standard metrics: accuracy, precision, recall, and F1-score. These metrics provided a comprehensive assessment of each model's ability to classify loan default across five class categories: DOUB (doubtful), LOSS (loss), NORM (normal), SPME (special mention), and SUBS (substandard). The experiments were conducted on a laptop equipped with an Intel Core i7 processor operating at 2.7 GHz, 8 GB of RAM, and a 1 TB hard disk drive. This configuration ensured efficient execution of machine learning algorithms, facilitating rapid model training and evaluation.

4.2 Dataset for Training and Testing

A comprehensive dataset containing 18,184 records was utilized, source from Wegagen Bank core banking system. The data set includes 13 features and one class features with five categories. To develop robust predictive models, the dataset was accurately prepared and split into 70/30 ration: a training set and a testing set

Table 4. 1 Training and testing dataset data distribution

Class Name	Training Sample	Testing Sample	Sample Total
NORM	10,137	4,344	14,484
SPME	1,025	440	1,465
SUBS	374	160	534

DOUB	640	275	915
LOSS	552	237	789

4.3 Deployed Model Architecture

The deployed architecture in this research is illustrated in Figure 4.1. The proposed architecture shows the steps followed in the Prediction of loan default of customers from the dataset obtained. In the proposed architecture, borrower's data is taken as input and stored in the dataset and this dataset contains borrowers with credit status such as NORM, SPME, SUBS, DAUT, and LOSS. To validate the proposed model, we use the borrower dataset. The architecture passes through different stages after business understanding which is described in Chapter Three. After understanding the business from the business perspective, the first task is obtaining ethical clearance from Wegagen Bank, data collection takes place. The data was collected from the Wegagen Bank credit department then the dataset (borrower's data) went through pre-processing tasks to clean attributes in the dataset. Then train each base model (Logistic regression, Multilayer perceptron, Random forest) independently on the training data. After training the base models a holdout set (validation set) is used to generate prediction for next stage: creating the meta-model. These predictions, which reflect the likelihood of each class for the validation samples, are sacked to form a new feature matrix. This meta-model learns to combine the base model predictions in a way that enhances overall predictive accuracy. Then using the meta model for making final prediction on the test set. Finally evaluate the performance of blending ensemble model involves using appropriate multiclass classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix

analysis. The final output of proposed ensemble model predicts the loan status of applicant as NORM, SUPS, DAUT, and LOSS.

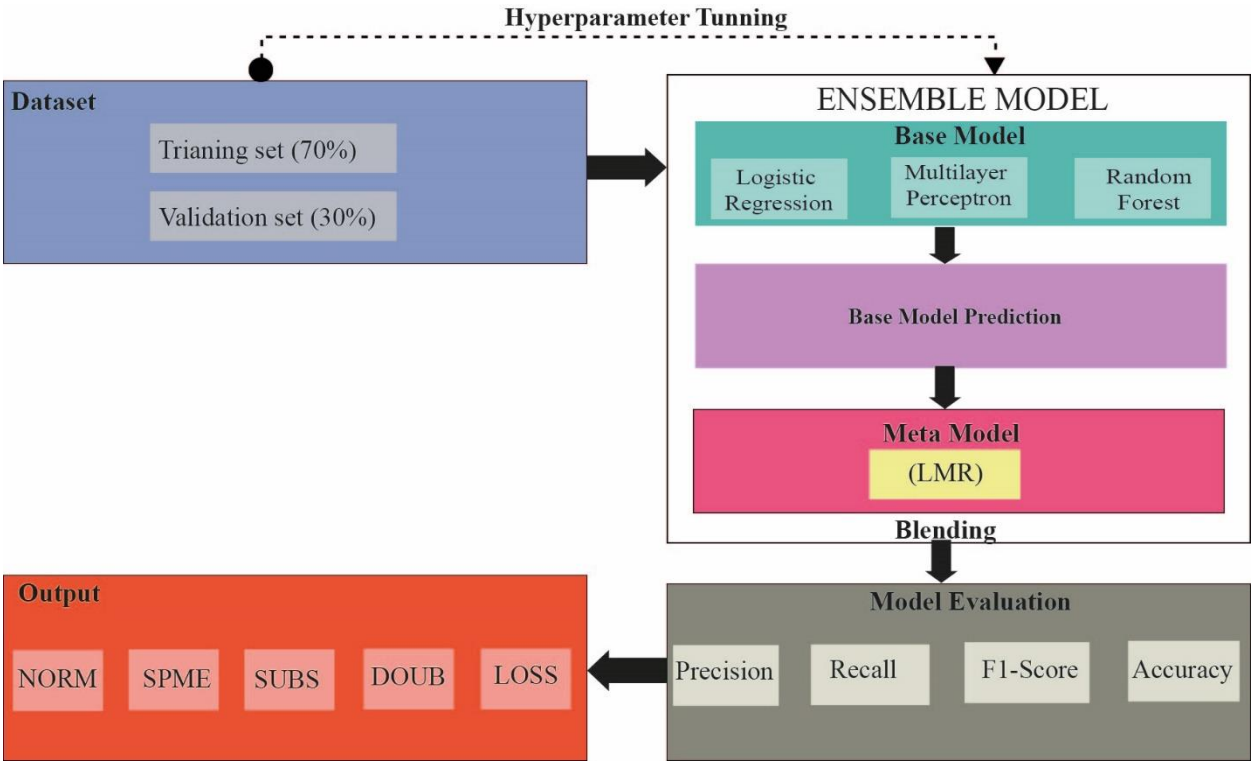


Figure 4. 1 Overview of Deployed Model Architecture

4.4 Dataset for Experiment

The dataset used in this study is crucial for developing and validating the loan default prediction models. The data was carefully selected and processed to ensure it accurately represents the factors influencing loan repayment behaviors.

The data for this study was collected from Wegagen bank core banking system. The core banking system contains several tables related to loan operations and other modules of the bank. Regarding the privacy of financial information, unable to get data from multiple banks. Therefore, the data obtained from Wegagen bank was considered sufficient and representative for this study

The dataset consists of 18,184 records with 13 features, including one class features with five categories Normal, Special Mention, Substandard, Doubtful, and Loss. The feature includes Sex, Martial status, Educational background, Business location, Business experience, Lon product type, Yearly Business income, and Location of collateral.

To prepare the data for experiments, several preprocessing steps were undertaken. Among the total dataset, 559 records had missing values for features such as educational background and business

income. Instead of imputing these values, rows with missing data were deleted to ensure the integrity of the analysis. Categorical features such as loan product type and business location were encoded using one-hot encoding to convert them into numerical format suitable for machine learning models. Feature such as yearly business income was normalized using Min-Max scaling to ensure they fall within a specific range, improving model performance.

The dataset was split into training and testing sets using a 70/30 ratio. This means 70% of the data (12,729 records) was used for training the models, and 30% (5,455 records) was used for testing. The split ratio was chosen to ensure a sufficient amount of data for both training and evaluation, minimizing the risk of overfitting while providing reliable performance metrics.

4.5 Selected Model Experiment

Three machine learning models were selected for predicting loan defaults: Logistic Regression, Multilayer Perceptron, and Random Forest. These models were chosen for their complementary strengths and the ability to provide robust and accurate predictions when used in a blending ensemble approach.

4.5.1 Logistic regression Model

Logistic Regression is simple yet powerful linear model widely used for binary and multiclass classification problems. It is easy to implement, interpretable, and efficient, making it suitable for initial baseline comparisons. It provides probabilistic outputs that are useful for understanding the likelihood of a borrower defaulting on a loan. The hyperparameter used for this experiment are: -

Table 4. 2 Hyperparameter Used for Logistic Regression Model

Hyper parameter	**Values**
C (regularization strength)	1.0
solver	liblinear
penalty	12
Class_weight	None

4.5.1.1 Evaluating Model Performance

Accuracy

The logistic regression model achieved an accuracy of 88.5% on the testing dataset. An accuracy of 88.5% indicates that the model correctly predicted the loan default status for 88.5% of the instance in the test set.

```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.8856304985337243
```

Figure 4. 2 Accuracy evaluation of logistic regression model

Confusion Matrix

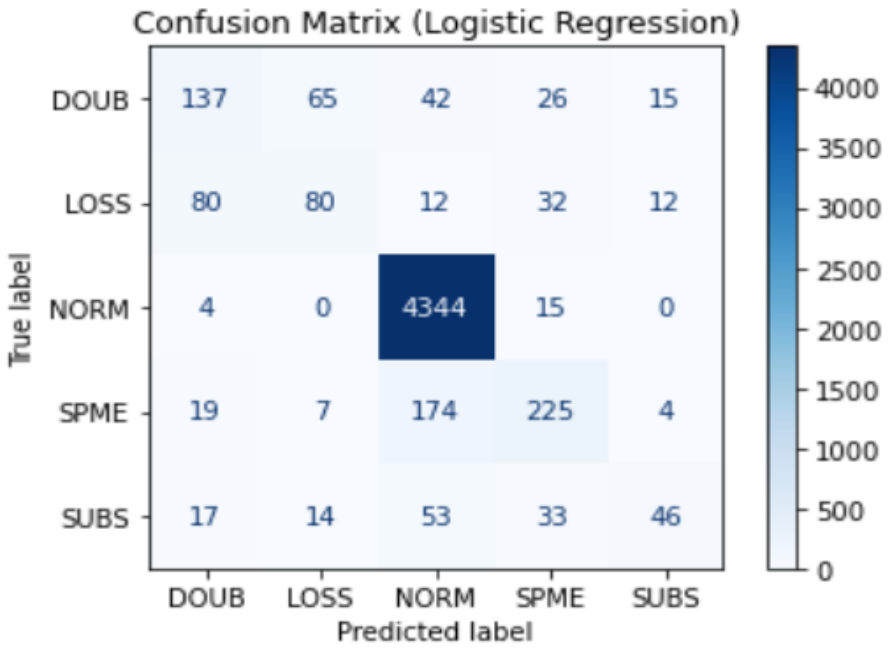


Figure 4. 3 Confusion matrix evaluation of logistic regression model

Confusion Matrix Interpretation

DOUB CLASS:

- Correctly classified 137 instances as Class Doub.
- Misclassified 65 instances as Loss Class, 42 instances as Norm Class, 26 instance as Spme class, and 15 instance as Subs Class.

LOSS CLASS:

- Correctly classified 80 instances as Class Loss.
- Misclassified 80 instances as Doub Class, 12 instances as Class Norm, 32 instance as class Spme, and 12 instance as Class Subs.

NORM CLASS:

- Correctly classified 4344 instances as Class Norm.
- Misclassified 4 instances as Doub Class, and 15 instance as Class Subs.

SPME CLASS:

- Correctly classified 225 instances as Class Spme.
- Misclassified 19 instances as Doub Class, 7 instances as Class Loss, 174 instance as Class Norm, and 4 instance as Class Subs.

SUBS CLASS:

- Correctly classified 46 instances as Class Subs.
- Misclassified 17 instances as Doub Class, 14 instances as Class Loss, 53 instance as class Norm, and 33 instance as Class Spme.

Classification Report

```
# Display detailed classification report
print('Classification Report of Logistic Regression classification:')
print(classification_report(y_test, y_pred))
```

```
Classification Report of Logistic Regression classification:
              precision    recall  f1-score   support

        DOUB       0.53      0.48      0.51       285
        LOSS       0.48      0.37      0.42       216
        NORM       0.94      1.00      0.97      4363
        SPME       0.68      0.52      0.59       429
        SUBS       0.60      0.28      0.38       163

    accuracy                           0.89      5456
   macro avg       0.65      0.53      0.57      5456
weighted avg       0.87      0.89      0.87      5456
```

Figure 4. 4 Classification report Logistic regression Model

4.5.2 Multilayer Perceptron Model

Multilayer Perceptron is a type of artificial neural network capable of capturing complex patterns in the data. It is effective for tasks where the relationship between input features and target variable is non-linear. MLP can model interactions between features that simpler models might miss.

Table 4. 3 Hyperparameter Used for Multilayer Perceptron Model

Hyper Parameter	**Values**
hidden layer Sizes	100,50
activation Function	relu
Solver	adam
Alpha	0.0001
Maximum Iteration:	500
Batch_size	32
Learning Rate	0.001
Momentum	0.9
Early_stopping	False
Validation_fraction	0.1
N_iter_no_change	10
Tolerance (tol)	le_4

4.5.2.1 Evaluating Model Performance

Accuracy

Multilayer perceptron (MLP) model achieved an accuracy of 79.7%. this indicate that 79.7% of the total predictions made by the model were correct. Accuracy measures the overall correctness of the model.

```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.7971041055718475
```

Figure 4. 5 Accuracy evaluation of Multilayer Perceptron Model

Confusion Matrix

```
# Plot the confusion matrix
class_labels = sorted(y.unique())  # Assuming your target variable is categorical
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=class_labels)
disp.plot(cmap='Blues', values_format='d')
plt.title('Confusion Matrix (MLP)')
plt.show()
```

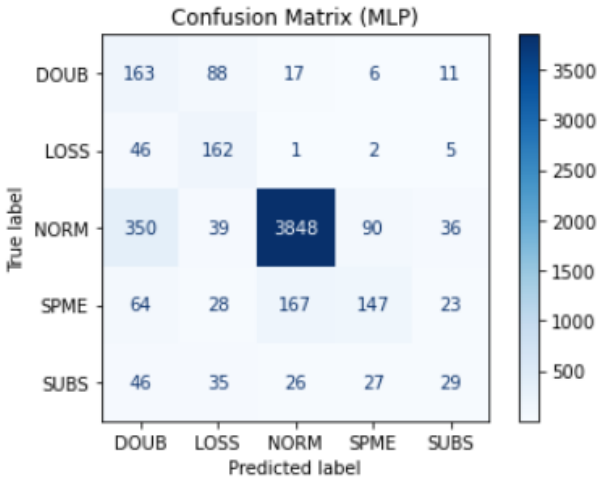


Figure 4. 6 Confusion matrix evaluation Multilayer Preceptron Model

Confusion Matrix Interpretation

DOUB CLASS:

- Correctly classified 163 instances as Doub Class.
- Misclassified 88 instances as Loss Class, 17 instances as Norm Class, 6 instance as Spme class, and 11 instance as Subs Class.

LOSS CLASS:

- Correctly classified 162 instances as Class Loss.
- Misclassified 46 instances as Doub Class, 1 instances as Class Norm, 2 instance as class Spme, and 5 instance as Class Subs.

NORM CLASS:

- Correctly classified 3848 instances as Class Norm.
- Misclassified 350 instances as Doub Class 39 instance as Loss Class, 90 instances as Spme Class, and 36 instance as Class Subs.

SPME CLASS:

- Correctly classified 147 instances as Class Spme.
- Misclassified 64 instances as Doub Class, 28 instances as Class Loss, 167 instance as Class Norm, and 23 instance as Class Subs.

SUBS CLASS:

- Correctly classified 29 instances as Class Spme.
- Misclassified 46 instances as Doub Class, 35 instances as Class Loss, 26 instance as Class Norm, and 27 instance as Class Subs.

Classification report

```
# Display detailed classification report
print('Classification Report of Mulilayer Perceptron(MLP)classification:')
print(classification_report(y_test, y_pred))
```

```
Classification Report of Mulilayer Perceptron(MLP)classification:
              precision    recall  f1-score   support

        DOUB       0.24      0.57      0.34       285
        LOSS       0.46      0.75      0.57       216
        NORM       0.95      0.88      0.91      4363
        SPME       0.54      0.34      0.42       429
        SUBS       0.28      0.18      0.22       163

    accuracy                           0.80      5456
   macro avg       0.49      0.54      0.49      5456
weighted avg       0.84      0.80      0.81      5456
```

Figure 4. 7 Classification report of Multilayer Perceptron

4.5.3 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. It is robust to noise and can handle large datasets with high dimensionality, making it a good choice for loan default prediction.

Table 4. 4 Hyperparameter Used for Random Forest Model

Hyperparameter	Values
n_estimators	100
Criterion	gini
Minimum Samples Split	2
Maximum Depth	None
Minimum Samples Leaf	1
Random state	42

4.5.3.1 Evaluating Model Performance

Accuracy

The Random Forest model achieved an accuracy 97.1%. this metrics indicates that 97.1% of the predictions made by the model were correct

```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
```

```
Accuracy: 0.9710410557184751
```

Figure 4. 8 Accuracy evaluation of Random Forest model

Confusion Matrix

```
# Plot the confusion matrix
class_labels = sorted(y.unique())  # Assuming your target variable is categorical
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=class_labels)
disp.plot(cmap='Blues', values_format='d')
plt.title('Confusion Matrix (Random Forest)')
plt.show()
```

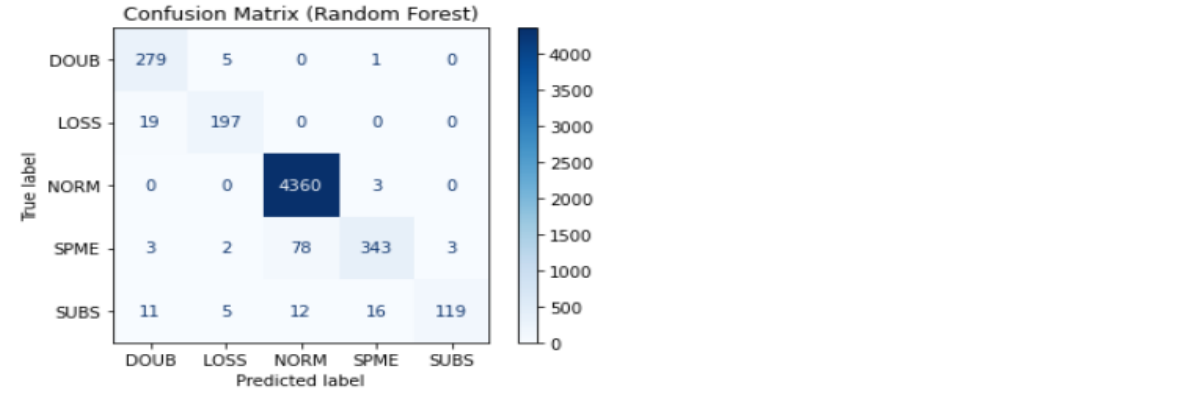


Figure 4. 9 Confusion Matrix of Random Forest Model

Confusion Matrix Interpretation of Random forest

DOUB CLASS:

- Correctly classified 279 instances as Doub Class.
- Misclassified 5 instances as Loss Class, and 1 instances as Spme Class.

LOSS CLASS:

- Correctly classified 197 instances as Loss Class.
- Misclassified 19 instances as Loss Class.

NORM CLASS:

- Correctly classified 4360 instances as Norm Class.
- Misclassified 3 instances as Loss Class.

SPME CLASS:

- Correctly classified 343 instances as Spme Class.
- Misclassified 3 instances as Doub Class, 2 instances as Class Loss, 78 instance as Class Norm, and 3 instance as Class Subs.

SUBS CLASS:

- Correctly classified 119 instances as Class Subs.
- Misclassified 11 instances as Doub Class, 5 instances as Class Loss, 12 instance as Class Norm, and 16 instance as Class Spme.

Classification Report

```
# Display detailed classification report
print('Classification Report:')
print(classification_report(y_test, y_pred))
```

```
Classification Report:
              precision    recall  f1-score   support

        DOUB       0.89      0.98      0.93       285
        LOSS       0.94      0.91      0.93       216
        NORM       0.98      1.00      0.99      4363
        SPME       0.94      0.80      0.87       429
        SUBS       0.98      0.73      0.84       163

    accuracy                           0.97      5456
   macro avg       0.95      0.88      0.91      5456
weighted avg       0.97      0.97      0.97      5456
```

Figure 4. 10 Classification report of Random Forest Model

4.5.4 Blending Ensemble Model

Blending Ensemble to leverage the strengths of each base model, a blending ensemble approach was used. This approach involves training each base model separately and then combining their predictions using a meta-model, which in this case, was a weighted average method.

4.5.4.1 Evaluating Model Performance

Accuracy

The Blending Ensemble model achieved an accuracy 98.6%. this metrics indicates that 98.6% of the predictions made by the model were correct.

```
# Evaluate predictions
accuracy = accuracy_score(y_val, y_val_pred)
print("Blending Ensemble Accuracy:", accuracy)
```

```
Blending Ensemble Accuracy: 0.9862536656891495
```

Figure 4. 11 Accuracy evaluation of Blending Ensemble Model

Confusion Matrix

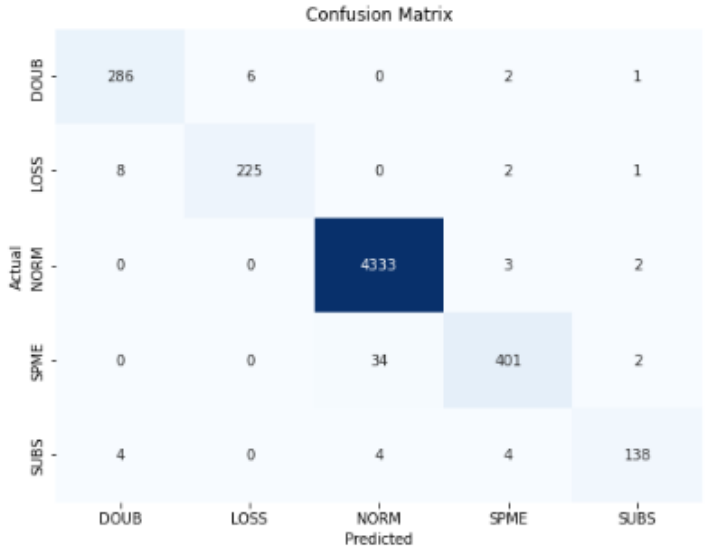


Figure 4. 12 Confusion Matrix of Blending Ensemble Model

Confusion Matrix Interpretation

DOUB CLASS:

- Correctly classified 286 instances as Doub Class.
- Misclassified 6 instances as Loss Class, 2 instances as Spme Class, and 1 instances as Subs Class.

LOSS CLASS:

- Correctly classified 225 instances as Loss Class.
- Misclassified 8 instances as Loss Class, 2 instances as Spme Class, and 1 instances as Subs Class.

NORM CLASS:

- Correctly classified 4333 instances as Norm Class.
- Misclassified 3 instances as Spme Class, and 2 instances as Subs Class.

SPME CLASS:

- Correctly classified 401 instances as Spme Class
- Misclassified 34 instances as Norm Class, and 2 instances as Subs Class.

SUBS CLASS:

- Correctly classified 138 instances as Subs Class
- Misclassified 4 instances as Doub Class, 4 instances as Norm Class, and 4 instances as Spme Class.

Classification Report

Classification Report:

	precision	recall	f1-score	support
DOUB	0.96	0.98	0.97	295
LOSS	0.98	0.96	0.97	236
NORM	0.99	1.00	0.99	4338
SPME	0.96	0.92	0.94	437
SUBS	0.96	0.92	0.94	150
accuracy			0.99	5456
macro avg	0.97	0.95	0.96	5456
weighted avg	0.99	0.99	0.99	5456

Figure 4. 13 Classification report of Blending Ensemble Model

4.6 Comparison of Model Performance

A comparative performance analysis was conducted between the three individual base models and the blending ensemble model. This analysis used evaluation metrics such as accuracy, precision, recall, and F1 score. The result showed that blending ensemble model outperformed the individual base models (logistic regression, multilayer perceptron, and random forest) in all these metrics, indicating its superior performance in predicting loan defaults as shown in Figures 4.14 – 4.16.

The accuracy of the loan default prediction model, developed using a blending ensemble approach that combines logistic regression, multilayer perceptron, and random forest, is shown in Figure 4.14. as illustrated, the blending ensemble model achieves higher accuracy compared to each of the individual base models (logistic regression, multilayer perceptron, and random forest).

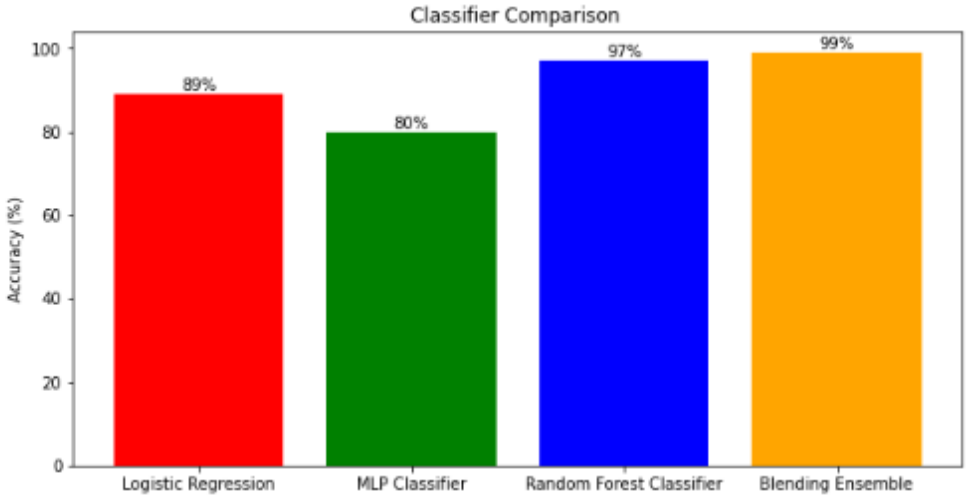


Figure 4. 14 Comparison of accuracy of model employed

The precision results for each class in the classifier comparison are shown in Figure 4.15. The blending ensemble model, which combines predictions from Logistic Regression, Multilayer Perceptron and Random Forest, further enhance precision across all classes, outperforming each individual model. This demonstrates the blending ensemble effectiveness in leveraging the strengths of multiple models to achieve superior predictive accuracy.

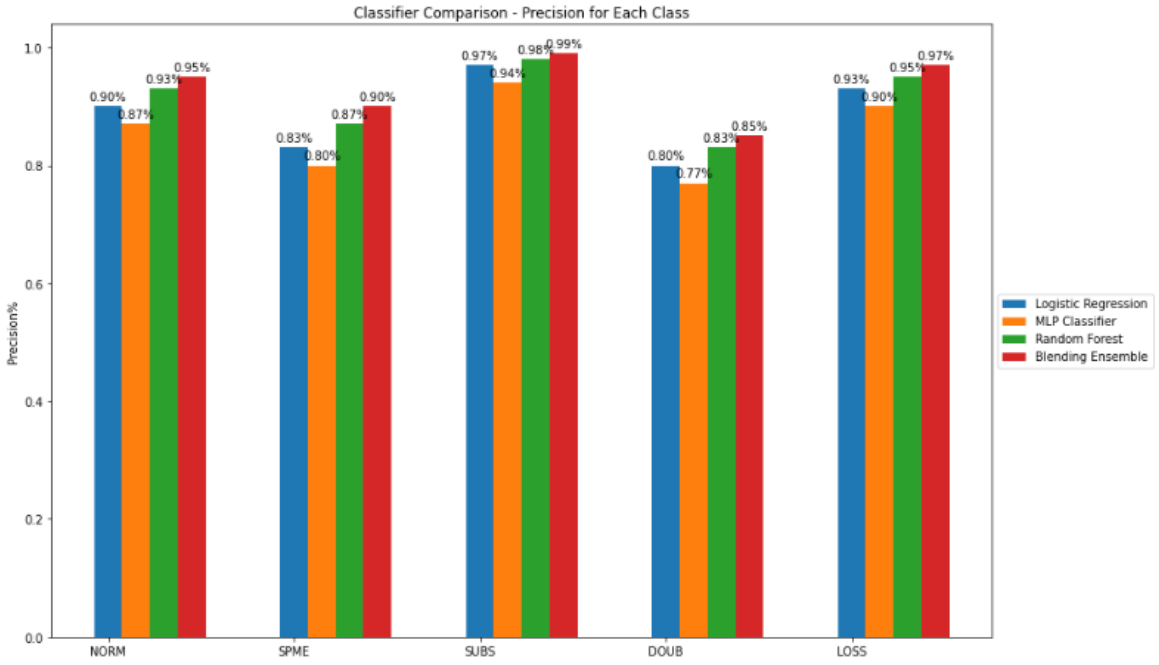


Figure 4. 15 Comparison of Precision of the four classifiers in five classes

Figure 4.16 shows the recall performance of three base models (Random forest, Multilayer Perceptron, Logistic Regression) and blending ensemble model using the test dataset. It is evident that the Blending Ensemble model exhibits superior recall across all five classes compared to individual Base models. Specifically, blending ensemble model, which combines predictions from RF, MLP, and Logistic Regression, further enhances recall performance by leveraging the strengths of these individual models.

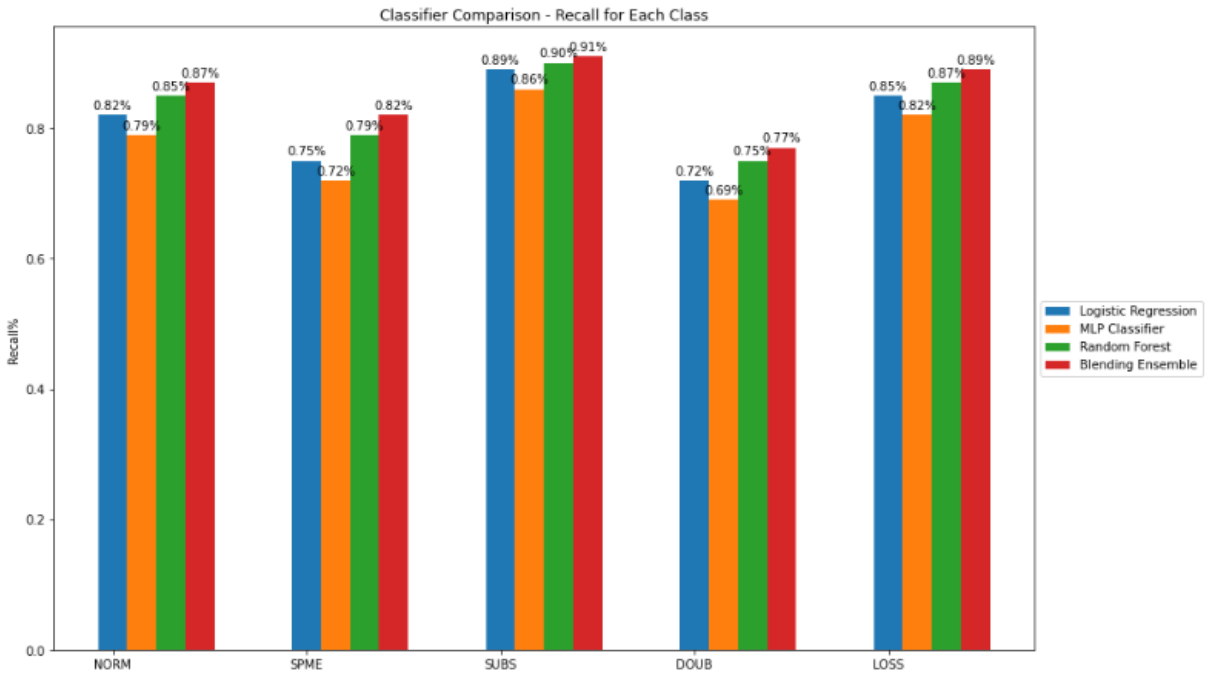


Figure 4. 16 Comparison of Recall of the four classifiers in five classes

4.6.1.1 Summary of Comparison of Models Performance Metrics

To develop an optimal loan default prediction model, four different machine learning approaches were evaluated: logistic regression, Multilayer perceptron, random forest and Blending ensemble model. The performance of each model was assessed using key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The table below summarizes the performance metrics for each model.

Table 4. 5 Summary of model comparative performance metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89%	0.65	0.53	0.57
Multilayer Perceptron	80%	0.49	0.54	0.49
Random Forest	97%	0.95	0.88	0.91
Blending Ensemble	99%	0.97	0.95	0.96

4.7 Feature Importance

Feature importance is a technique used in machine learning to identify and rank the features (or variable) in a dataset based on their influence on the predictive model's output. This analysis helps in understanding which factors are most significant in determine the model's output. This analysis helps in understanding which factors are most significant in determining the model's predictions, providing valuable insights for decision-making and improving model interpretability. In this study, feature importance was assessed to determine the key factors affecting loan default, using a blending ensemble method.

To identify the most important factors influencing loan default, we employed a feature importance analysis within our blending ensemble model. This method evaluates the contribution of each feature to the model's predictive power. The blending ensemble model utilized the Gini impurity approach to rank features based on their importance scores. This approach effectively identifies the features that contribute most significantly to predicting whether borrowers fall into one of the loan status: NORM (normal), SPME (special mention), SUBS (substandard), DOUB (doubtful), and LOSS.

The feature importance scores revealed that "Sex" was the most important factor, with a score of 0.284507, indicating its substantial influence on loan default prediction. This finding suggests that gender-related socio- economic factors play a critical role in determining loan repayment behavior. Other notable features included "Marital Status" (0.128325), "Educational Background" (0.068495), and "Interest Rate" (0.067427. Interestingly, "Interest Rate" traditionally considered a crucial financial factor, was less influential than demographic factors such as gender and marital status. This underscores the significance of demographic characteristics in understanding and predicting loan defaults.

```
In [22]: feature_scores = pd.Series(clf.feature_importances_, index=X_train.columns).sort_values(ascending=False)

         feature_scores

Out[22]: SEX                    0.284507
         MARTIAL STATUS         0.128325
         EDUCATION BACKGROUND   0.068495
         INTREST RATE           0.067427
         COLLECTED TOTAL        0.061462
         DISBURSED DATE         0.059477
         LOAN EXP DATE          0.057238
         BUSINESS EXPERIENCE    0.051758
         APPROVED AMOUNT        0.050784
         COLLATORAL VALUE       0.050245
         LOAN PRODUCT           0.044172
         BUSINESS INCOME        0.041917
         BUSINESS LOCATION      0.034193
         dtype: float64
```

Figure 4. 17 Feature importance

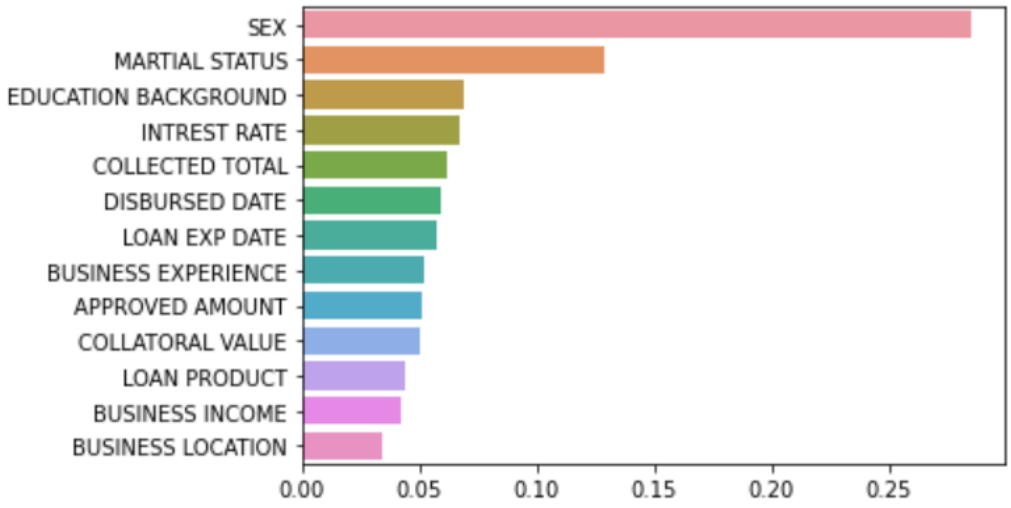


Figure 4. 18 Feature importance plot

4.8 Discussion of Result

The main objective of this study was to develop an optimal machine learning model for predicting loan defaults. To achieve this, three base models—Logistic Regression, Multilayer Perceptron, and Random Forest—were evaluated, along with a blending ensemble model. The dataset used for training comprised 12,728 samples, while the test set included 5,456 samples.

Among the base models, the Random Forest model achieved the highest accuracy on the test set, with a notable 97.10%. This was followed by Logistic Regression with an accuracy of 88.56%, and Multilayer Perceptron, which achieved an accuracy of 72.65%. The Random Forest model not only outperformed the other base models in terms of accuracy but also demonstrated superior recall

across all five loan default classes: DOUB (doubtful), LOSS, NORM (normal), SPME (special mention), and SUBS (substandard).

The blending ensemble model, which combined the predictions of the three base models, further improved the predictive performance. This model achieved an accuracy of 98.7% on the test set, demonstrating its superiority over the individual base models. This result indicates that the blending ensemble approach can effectively leverage the strengths of multiple models to enhance prediction accuracy. Based on these findings, the blending ensemble model is recommended for loan default prediction due to its better performance.

To address the first research question, "Are ensemble methods more effective than individual machine learning techniques in loan default prediction?", the results clearly demonstrate that ensemble methods, specifically the blending ensemble model, are more effective. The blending ensemble model outperformed each of the individual base models in terms of accuracy and recall, thus providing a more reliable prediction tool for loan defaults.

Feature importance analysis provided additional insights into the factors influencing loan default predictions. The analysis revealed that "Sex" was the most important factor, with a feature importance score of 0.284507. This suggests that gender-related socio-economic factors significantly impact loan repayment behavior. Other important factors included "Marital Status" (0.128325), "Educational Background" (0.068495), and "Interest Rate" (0.067427). Interestingly, while "Interest Rate" is traditionally considered a crucial financial factor, it was found to be less influential than demographic factors such as gender and marital status. This highlights the critical role of demographic characteristics in predicting loan defaults.

Regarding the second research question, "What are the most important features or variables in predicting loan default?", the feature importance analysis identified "Sex" as the most significant factor, followed by "Marital Status," "Educational Background," and "Interest Rate." These findings underscore the importance of considering demographic characteristics alongside traditional financial metrics when predicting loan defaults.

Overall, the results of this study underscore the importance of using a robust and comprehensive approach to loan default prediction. The blending ensemble model, in particular, has demonstrated

its effectiveness in accurately predicting loan defaults, providing valuable insights for financial institutions in assessing and managing credit risk.

This study aimed to develop an optimal machine learning model for predicting loan defaults, leveraging both individual base models logistic regression, multilayer perceptron, random forest and blending ensemble model. The findings demonstrated that the blending ensemble model, which combined the strengths of the three base models, achieved superior performance with an accuracy of 98.62%, outperforming the best performing base model, Random Forest, which had an accuracy of 97.10%. This result underscores the efficacy of ensemble methods in enhancing predictive accuracy and reliability in loan default prediction.

Moreover, the feature importance analysis revealed significant insights into the factors influencing loan default. “Sex” emerged as the most critical factors, followed by “Martial Status”, “Educational Background”, and “Interest Rate.” These findings highlight the consideration impact of demographic characteristics on loan repayment behavior, suggesting that financial institutions should incorporate these variables alongside traditional financial metrics for more accurate risk assessments.

In conclusion, the study successfully demonstrated that ensemble methods are more effective than individual machine learning techniques in predicting loan defaults. The blending ensemble model not only improved prediction accuracy but also provided a comprehensive understanding of the key factor influencing loan default. These insights can aid financial institutions in developing more robust credit risk assessment strategies, ultimately enhancing their decision making processes and mitigating potential loan default risks. The research underscores the importance of integrating diverse data source and leveraging advanced machine learning techniques to achieve superior predictive performance and deeper insights into borrower behavior.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study aimed to develop an optimal machine learning model for predicting loan defaults, leveraging both individual base models logistic regression, multilayer perceptron, random forest and blending ensemble model. The findings demonstrated that the blending ensemble model, which combined the strengths of the three base models, achieved superior performance with an accuracy of 98.62%, outperforming the best performing base model, Random Forest, which had an accuracy of 97.10%. This result underscores the efficacy of ensemble methods in enhancing predictive accuracy and reliability in loan default prediction.

Moreover, the feature importance analysis revealed significant insights into the factors influencing loan default. “Sex” emerged as the most critical factors, followed by “Martial Status”, “Educational Background”, and “Interest Rate.” These findings highlight the consideration impact of demographic characteristics on loan repayment behavior, suggesting that financial institutions should incorporate these variables alongside traditional financial metrics for more accurate risk assessments.

In conclusion, the study successfully demonstrated that ensemble methods are more effective than individual machine learning techniques in predicting loan defaults. The blending ensemble model not only improved prediction accuracy but also provided a comprehensive understanding of the key factor influencing loan default. These insights can aid financial institutions in developing more robust credit risk assessment strategies, ultimately enhancing their decision making processes and mitigating potential loan default risks. The research underscores the importance of integrating diverse data source and leveraging advanced machine learning techniques to achieve superior predictive performance and deeper insights into borrower behavior.

5.2 Future Work

Future research could address the limitations identified in this study and explore several ways to enhance loan default prediction models further. First, expanding the dataset to include data from multiple financial institutions would improve the model's generalizability and robustness. This would help ensure that the model can accurately predict loan defaults across diverse economic environments and borrower characteristics.

Second, incorporating additional features and data source could provide a more comprehensive view of the factors influencing loan defaults. integrating macroeconomic indicators, detailed credit histories, and real-time transaction data can provide a richer insight into borrower characteristics and improve prediction accuracy

Finally, employing more advanced machine learning techniques and deep learning models could enhance the predictive power of the model.

REFERENCE

[1] A. Mauri, "The Short Life of the Bank of Ethiopia," no. issue 4(4), pp. 104-116, December, 2010.

[2] NBE, "National Bank of Ethiopia," [Online]. Available: https://nbe.gov.et/about-us/our-history/. [Accessed 2023].

[3] Aslam, Uzair; Aziz, Hafiz Ilyas Tariq; Sohail, Asim; Batcha, Nowshath Kadhar, "An Empirical Study on Loan Default Prediction Model," vol. Volume 16, August 2019.

[4] L. Lai, "Loan Default Prediction with Machine Learning Techniques," 2020.

[5] M. S. Irfan Ahmed, P. Ramila Rajaleximi, "An Empirical Study on Credit Scoring and Credit," vol. Volume 8, no. Issue 7, 2019.

[6] Beasens and Gestel, " Credit Risk Management Basic Concepts.," *New York: Oxford University Press Inc*, 2009.

[7] Singh.A, " Credit Risk Management Practices in Ghana Commercial Banks. I," *nternational Journal of Marketing and Management Research*, pp. , 2, 47-51. , 2013.

[8] Li.Ong, "The Credit Risk Transfer Market and Stability Implications for U.K. Financial Institutions," *International Monetary Fund (IMF)*, p. 27, 2006.

[9] Baker, J.C. , "Risk Management in Global Finance Vol. 21 No. 4 https://doi.org/10.1108/eb018508," *Managerial Finance*, pp. 1-4, 1995.

[10] Arora, Anju, "The impact of size on credit risk management strategies in commercial banks : empirical evidence from India," *The IUP journal of financial risk management* , 2012.

[11] Committee,Basel, "Best Practice for Credit Risk Disclosure," *Committee on Banking*, 2000.

[12] Committee,Basel, "Sound Practice for The Management of Operational Risk," 2001.

[13] ATAKELT, P. VENI, "DETERMINANTS OF CREDIT RISK IN ETHIOPIAN PRIVATE COMMERCIAL BANKS," *International Journal of Accounting and*, 2015.

[14] Cebenoyan, A.S. and Strahan, P.E. , "Risk Management, Capital Structure and Lending at Banks., 28, . http://dx.doi.org/10.1016/S0378-4266(02)00391-6," *Journal of Banking & Finance*, pp. 19-43, 2004.

[15] B. Committee, "Principles for the Management of Credit risk," *Basel Committee on Banking Supervision*, 1999.

[16] P. J.B, "Credit Risk Management in Indian Banks,," *International Journal of Advance Research in Computer Science and Management Studies*, *2(1)*, *309-313.*, 2014.

[17] S. a. M. George, "Credit Risk Management Life Cycle," (2008).

[18] Michael, K. O., "Risk Management –A Modern Perspective, Elsevier Inc, London, UK.," 2015.

[19] Koch, W.T, "Bank management (3rd Ed.) USA: The Dryden press, see Harbor Drive.," 1995.

[20] S.Sah, "Machine Learning: A Review of Learning Types," *ResearchGate,* 2020.

[21] T.O.Ayaodele, "Types of Machine Learning," 2010.

[22] S. B. Kotsiantis,D. Kanellopoulos and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," 2014.

[23] V.Nasteski, "An overview of the supervised machine learning methods," pp. 51-62, 2017.

[24] B. Dushimimana, Y. Wambui, T. Lubega, and P. E. Mcsharry, "Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans," 2017,2020.

[25] M. Heller, "The engines of AI: Machine learning algorithms explained," *InfoWorld,* 2023.

[26] K. Wakefield, "A guide to machine learning algorithms and their applications," 2019.

[27] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models," pp. 352-359, 2002.

[28] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, "Scikit-learn: machine learning in python. J Mach Learn Res.," p. 2825–30, 2011.

[29] Han J, Pei J, Kamber M. , "Data mining: concepts and techniques. Amsterdam: Elsevier; 2011., Sarker IH, Furhad MH, Nowrozy R. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer. Science.," p. 1–18, 2021;2(3).

[30] S. IH., "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. SN Computer. Science. :," p. 1–16, 2021;2(3).

[31] Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.

[32] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," pp. 272-278, 2012.

[33] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," pp. 272-278, 2012.

[34] Guzman E, El-halaby M, Bruegge B., "Ensemble methods for app review classification : an approach for software evoluti," 2015.

[35] Ren Y, Suganthan PN, Srikanth N., "Ensemble methods for wind and solar power forecasting—a state-of-the-art review. Renew Sustain Energy Rev," no. https ://doi.org/10.1016/j.rser.2015.04.081., 2015.

[36] Khairalla MA, Ning X, AL-Jallad NT, El-Faroug MO. , "Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. Energies. ;1," no. https ://doi.org/10.3390/en110 61605 ., p. 1:1–21. , 2018.

[37] H. Dalianis, ""Evaluation Metrics and Evaluation," in Clinical Text Mining," *springer,* 15 May 2018. .

[38] Mirka Saarela , Susanne Jauhiainen,, "''Comparison of feature importance measures as explanations," *no. SN Applied Sciences ,* no. https://doi.org/10.1007/s42452-021-04148-9,, pp. pp. 1-12, 3 , February 2021..

[39] Chang, Y.C., Chang, K.H., Chu, H.H. & Tong, L.I. (, "Establishing decision tree based short-term default credit risk assessment models. Communications in Statistics – Theory and Methods," no. https://doi.org/10.1080/03610926.2014.96, pp. 45(23), 6803-6815., 2016.

[40] Xu, J., Lu, Z. & Xie, Y. , " Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11.," https://doi.org/10.1038/s41598-021- 98361-6, (2021).

[41] Abid, L., Masmoudi, A. & Zouari-Ghorbel, S. . , "The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank.," *Journal of the Knowledge Economy,* pp. 9(3), 948- 9, (2018).

[42] Silva, E.C., Lopes, I.C., Correia, A. & Faria, S. , "A logistic regression model for consumer default risk.," *Journal of Applied Statistics,,* no. https://doi.org/10.1080/02664763.2020.1759030, pp. 47(13-15), 2879-2894., (2020). .

[43] Tian, Z., Xiao, J., Feng, H. & Wei, Y. (2020)., "Credit Risk Assessment based on Gradient Boosting Decision Tree," *Procedia Computer Science,,* no. https://doi.org/10.1016/j.procs.2020.06.070, pp. 174, 150-160.

[44] Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction.," *2020 International Conference on Computing and Data Science (CDS), Stanford, CA, USA.,* no. https://doi.org/10.1109/CDS49703.2020.00050, (2020, August)..

[45] Ebisa Deribie, Getachew Nigussie and Fikadu Mitiku, "Filling the breach: Microfinance," *Journal of Business and Economic Management 1(1): 010-017,* January 2013.

[46] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar , Introduction to Data mining, Pearson Education Limited, 2014.

[47] S. Birla, K. Kohli, and A. Dutta, "“Machine Learning on imbalanced data in Credit Risk,”," *IEEE,* 2016.

[48] P. Verhagen, "Predictive Modeling," 2018.

[49] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network".

[50] Iqbal H. Sarker, Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions, under exclusive licence to Springe, 18 August 2021.

[51] S. William, "CREDIT CONTROL ATTITUDE AND COMPLIANCE WITH CENTRAL BANK OF NIGERIA PRUDENTIAL GUIDELINES: EVIDENCE FROM A DEPOSIT MONEY BANK IN NIGERIA," *International Journal of Economics, Business and Management Research,* 2020.

[52] J. Murray, Government Gazette, Cape Town, 15 March 2006.

[53] M. Madaan, "Loan default prediction using decision trees and," 2021.

[54] Saqib Aziz & Michael Dowling , "Machine Learning and AI for Risk Management: FinTech and Strategy in the 21st Century," 2020.

[55] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., "Learning from class-imbalanced data: Review of methods and applications," 2017.

APPENDIX

Appendix1. Importing necessary libraries and data preprocessing

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, plot_confusion_matrix
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestClassifier
rom sklearn.metrics import accuracy_score
from numpy import hstack
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import pandas as pd



#Importing Dataset
df = pd.read_csv("FD.csv")


# Assuming your data is in a pandas DataFrame df
# Extract features and target variable
X = df[['LOAN PRODUCT', 'BUSINESS INCOME', 'BUSINESS EXPERIENCE', 'BUSINESS LOCATION', 'SEX', 'MARTIAL STATUS', 'EDUCATION BACKGR
y = df['STATUS']


# Convert categorical variables to one-hot encoding
X = pd.get_dummies(X, columns=['LOAN PRODUCT', 'BUSINESS INCOME', 'BUSINESS EXPERIENCE', 'BUSINESS LOCATION', 'SEX', 'MARTIAL STA



# One-hot encode categorical variables
categorical_cols = ['LOAN PRODUCT', 'BUSINESS INCOME', 'BUSINESS EXPERIENCE', 'BUSINESS LOCATION', 'SEX', 'MARTIAL STATUS', 'EDUC
encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')
X_encoded = pd.DataFrame(encoder.fit_transform(X[categorical_cols]))
X.drop(categorical_cols, axis=1, inplace=True)
```

Appendix. 2 Define base model

```
# Define base models
def get_models():
    models = []
    models.append(('lrr', LogisticRegression()))
    models.append(('mlp', MLPClassifier()))
    models.append(('rf', RandomForestClassifier()))
    return models

# Fit the ensembles
def fit_ensemble(models, X_train, X_val, y_train, y_val):
    meta_X = list()
    for name, model in models:
        model.fit(X_train, y_train)
        yhat = model.predict(X_val)
        yhat = yhat.reshape(len(yhat), 1)
        meta_X.append(yhat)
    meta_X = hstack(meta_X)
    blender = LogisticRegression()
    blender.fit(meta_X, y_val)
    return blender
```