AN ENSEMBLE METHOD FOR PREDICTING LOAN ELIGIBILITY IN COMMERCIAL BANK USING MACHINE LEARNING

BY

JAHIRUL ISLAM
ID: 191-15-2752

AND

KH. SHAKIL
ID: 191-15-2759

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

FATEMA TUJ JOHORA
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

AL AMIN BISWAS
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Thesis titled “**AN ENSEMBLE METHOD FOR PREDICTING LOAN ELIGIBILITY IN COMMERCIAL BANK USING MACHINE LEARNING**”, submitted by **Jahirul Islam,** ID No: **191-15-2752** and **Kh. Shakil,** ID No: **191-15-2759** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 23-01-2023.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Tarek Habib — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Tapasy Rabeya — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

23-01-23

Dr. Dewan Md Farid — **External Examiner**
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Fatema Tuj Johora, Senior Lecturer, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Fatema Tuj Johora
Senior Lecturer
Department of Computer Science and Engineering

Daffodil International University

Co-Supervised by:

Al Amin Biswas
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Jahirul

Jahirul Islam
ID: 191-15-2752
Department of Computer Science and Engineering
Daffodil International University

KH. SHAKIL

Kh. Shakil
ID: 191-15-2759
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Fatema Tuj Johora**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In recent days in Bangladesh the number of loan applicants for loans in commercial banks are gradually increasing every year. Banking sectors always need a more accurate system for handling many issues. In order to select the right applicant who can return the loan amount within given time, the bank employees do a lot of analysis on the information provided by the applicant and based on the analysis give a prediction. But it is very difficult and time-consuming process for bank employees. To deal with this particular problem of predicting the right applicant for loan request we use the EDA (Exploratory Data Analysis) technique. A variety of machine learning models are used to aid in the task of loan prediction. A dataset made up of loan projections is used to assess the study. Data cleaning procedures were applied, such as deleting null columns, using the mean mode approach to fill in missing values, and converting categorical values to numeric format. We employ two different methods to provide the greatest outcomes from the feature selection process. Traditional machine learning models employ distinct training and testing processes for both features, which are derived from various feature selections. Bagging Classifier, out of all the models, has attained the highest level of accuracy (88.00%), as well as a high recall and F1 score. The method of univariate feature selection was used to achieve this. As a result, the results suggest that Bagging Classifier might do very well when it comes to the task of predicting loan defaults.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-4**
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Project Management and Finance	4
1.7 Report Layout	4
CHAPTER 2: BACKGROUND	**5-11**
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	10
2.4 Scope of the Problem	11
2.5 Challenges	11
CHAPTER 3: RESEARCH METHODOLOGY	**12-26**

3.1 Research Subject and Instrumentation 12

3.2 Data Collection Procedure/Dataset Utilized 14

3.3 Statistical Analysis 14

3.4 Proposed Methodology/Applied 15

3.5 Implementation Requirements 25

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **27-32**

4.1 Experimental Setup 27

4.2 Experimental Results & Analysis 27

4.3 Discussion 32

CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **33-35**

5.1 Impact on Society 33

5.2 Impact on Environment 33

5.3 Ethical Aspects 34

5.4 Sustainability Plan 35

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **36-37**

6.1 Summary of the Study 36

6.2 Conclusions 36

6.3 Implication for Further Study 37

REFERENCES **38-40**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Dataset in CSV format	14
Figure 3.2: Stages of Loan Prediction Analysis	15
Figure 3.3: Ensemble method	20
Figure 3.4: Bagging Classifier	21
Figure 3.5: Adaboost Classifier	24
Figure 4.1: Accuracy comparison of various machine learning models without feature selection	31
Figure 4.2: Accuracy Comparison of various machine learning models with feature importance selection	31
Figure 4.3: Accuracy Comparison of various machine learning models with univariate feature selection	32

LIST OF TABLES

Tables	**PAGE NO**
Table 2.1: Comparative analysis with previous work	10
Table 4.1: Classification Report of the machine learning models without any feature selection	28
Table 4.2: Classification Report using features of Feature importance selection	29
Table 4.3: Classification Report using features of Univariate feature selection	30
Table 4.4: Accuracies of the models using voting classifier	30

CHAPTER 1

Introduction

1.1 Introduction

In order to overcome financial obstacles and realize some of their personal goals, people all over the world rely in some way on loans provided by banks. The ever-changing economy and the growing level of financial competition have necessitated the loan application process. Additionally, loan distribution is nearly every bank's primary business. The majority of a bank's assets are primarily linked to profits from loans issued by the bank authority. Banks only give these loans to applicants who have the capability to repay them in order to continue in business and reduce the danger of loan defaults. Therefore, after a lengthy process of validation and verification, many financial institutions and banks accept loans. However, there is no assurance that the selected candidate is the most meritorious of all applicants.[1]

One of the most significant issues facing the banking industry is the rating of borrowers' creditworthiness when making commercial loan lending decisions. The danger of borrowers failing to repay their current debt is characterized as credit risk [2].

Using the credit scoring method, one may estimate credit risk and cut down on illicit activity [3]. This credit rating system is utilized to make decisions based on information regarding the borrowers. Bankers seek the lowest default risk and highest return when deciding to give a loan. [4]. In order to mitigate the risk of default, it is required to build and improve a credit lending management system and to anticipate loan default problems.

In the past, the main technique for predicting loan default was human screening. Because there is so much data, it always takes a lot of time and labor and causes a lot of problems. Utilizing machine learning to estimate the possibility of a default risk saves time and reduces the amount of work required, hence enhancing the effectiveness and precision significantly.

Machine learning techniques are used in financial applications for two different reasons. With the growth of internet and mobile banking, as well as the emergence of third-party online and mobile payment platforms, banks are gathering more information on their clients from both internal and external sources. This information is then used to anticipate loan default using big data. The idea of implementing machine and deep learning models in a number of financial applications that were previously dominated by manual processes was inspired by the rapid development and success of doing so in areas like image classification, stock market prediction, and traffic forecasting.

In this study, we forecast whether loan default will occur based on real-world data provided by a bank and assess the effectiveness of five machine learning models from various families, such as RG, DT, Adaboost and Gradient Boost etc. In addition to analyzing the prediction result, this study also outlines our prepossessing procedure. The outcome shows that Bagging Classifier achieves an 88.00% accuracy rate in this data set, demonstrating a promising use of Bagging Classifier in the real world.

The main contribution of this study is:

1. We address the problem of predicting loan default using a publicly available set of real-world data.
2. We undertake trials with ten sophisticated models of machine learning.
3. We compared the outcomes of different machine learning algorithms.
4. A voting classifier is used to improve the performance of every model.
5. With our feature extraction approaches, we were able to attain a high accuracy.

1.2 Motivation

Almost every bank in the banking industry's core business is to provide loan. The main assets of bank’s mainly came from the profit of loan transactions in a bank. Every bank's primary goal is to put their valuable assets in a safe hand. After a regress process of validation and verification banks and other financial companies approves loan. Despite of this process of validation and verification there is no guaranty whether the chosen applicants are eligible or not. We are able to generate the prediction report which shows whether a current loan applicant is safe or not using this system. Machine Learning applications generally automate the whole process of verifying and validating features.

1.3 Rationale of the Study

We survey the loan applicant's history about their previous banking transaction. Most of the time in banking sector, we are manually selecting the applicants. It is quite challenging to predict the eligible applicants to apply for a loan. Consequently, we develop a model that automatically selects the best candidate who are eligible to getting loan. To choose the right candidate, we must therefore check previous customers banking history. We begin a study on the existing dataset of banks and make evaluations as a result. Among the Predictive Analysis classifiers, we employ are AdaBoost Classifier, Gradient Boosting Classifier (GB), Bagging Classifier, Decision Tree (DT).

1.4 Research Question

These are the research questions:

- What is loan prediction?
- How does we can predict the eligibility of loan applicants?
- Develop a model that can determine whether an applicant is defaulter?
- What issues does have in predicting the eligibility of applicants in the banking sector?
- How can the results be more specific?

1.5 Expected output

The results of our study are crucial for forecasting the loan defaulters. In this research, we tried to use some machine learning models to categorize the loan defaulters for predicting the eligibility of those who have applied for the loan. The goal of the project is to classify loan defaulters in quickly and precisely.

1.6 Project Management and Finance

There are a couple restrictions through this project that we hope to get past in the future. We employed feature selection approaches and fewer classifiers that we would like to work in the future.

Finance: Our research project finance by Daffodil International University.

1.7 Report Layout

- The project's objective, justification for the study, research question, meaning of the study, and study limitations are all outlined in Chapter 1.
- Introduction, Related Work, and Research Summary in Chapter 2.
- Contains a data set by means of associated data collecting and processing procedures in the Chapter 3.
- System Design and Analysis in Chapter 4.
- Impact on Society and Environment, Ethical Aspects and Sustainability Plan in Chapter 5.
- Summary, Assumption, and Future work in Chapter 6.

CHAPTER 2

Background study

2.1 Preliminaries

Loan defaulting is becoming a major problem in developing country as we discussed in this chapter. As a result, the banking sector of a country face a horrible loss in business, because the main profit of banks depends on their distribution of loans. In chapter, we discuss some relevant work on this particular subject. Finally, there is a comparative study has been provided.

2.2 Related Works

The loan defaulters were foreseen by Kadam et al. [5] collected and sanitized the data. After evaluating performance, it can be determined with certainty that the Nave Bayes model is highly efficient and produces superior results compared to other models. By examining the past records of the suitable loan applicant. The model is capable of making the most precise forecasts. This study is separated into the following sections: They compared various machine learning models, trained the system using the models, and performed testing. This system forecasts loan information using models. For example, SVM and the Naive Bayes (NB) Model. Kumar et al. [6] minimized the risk associated with picking the qualified individual in order to save the bank a substantial amount of time and resources. It employs many machine learning algorithms. Examples include Decision Trees, Random Forest, Support Vector Machine, Linear Models, Neural Network, and Adaboost. These six classification methods for machine learning have been used to predict the predicted outcome. Comparing datasets across several machine learning models, supplying the system with employed models, and performing tests. Supriya et al. [7] predicted whether or not it is safe to grant the loan to a specific individual. After cleaning and processing the collected dataset, the next section is presented. After examining the system's performance, we can deduce from the obtained predictable outcomes that the decision tree model provides the highest degree of precision. It is relatively simple to tune.

This model of machine learning was used to train the system on these records, which were then used to test the system. This system's primary function is to reduce the risk associated with selecting the best candidate. Using classification, logic regression, decision tree, and gradient boosting, this system forecasts the loan data. Gautam et al. [8] predicted the kind or history or believability of the loan applicant's client. They employ exploratory data analysis to solve the problem of approving or rejecting loan requests, or loan prediction, in short. Two classification models based on machine learning have been employed to predict the expected outcome. Examples include Decision Tree and Random Forest. This paper's primary objective is to determine if a certain individual or organization should be granted a loan. Gupta et al. [9] have deployed many machine learning models on the Kaggle dataset to predict whether a loan should be granted or not. There are a total of 981 applicant records with the values of their categories and numeric attributes. In the experiment and result analysis portion, the logistic regression, random forest machine learning technique is utilized. Bhattad et al. [10] decide if a loan to a certain individual or organization will be granted. They utilized three classification techniques, including Decision Tree, Logistic Regression, and Random Forest. The best result obtained in trials without attribute selection was 80.20 percent. This system's primary objective is to categorize and evaluate the character of loan applications. Consequently, they developed this model. Gomathy et al. [11] predict the loan data using Decision Tree machine learning methods. After cleaning and processing the collected dataset, the next section is presented. After analyzing the system's performance. Implementing this system allows us to conclude that the Decision Tree algorithm can analyze and comprehend the prediction process. The primary objective of this approach is to identify qualified people who can afford it. Mehul et al. [12] studied the use of various machine learning models to the loan-lending process in order to determine the optimal strategy. Random Forest and Decision Trees were the classifiers utilized to construct the model. By utilizing these algorithms, they independently analyze the dataset and uncover patterns within it. They anticipate if a new applicant is likely to default on a loan based on their analysis. They employ exploratory data analysis (EDA) to solve this particular difficulty of selecting the best candidate. Using Decision Tree, they achieved 73% accuracy, whereas Random Forest yielded 80% accuracy. The Random Forest Classifier yielded an 80% accuracy rate. Decision Tree Classifier, on the other hand, provided an accuracy of 73%. The Random Forest appears to be the superior choice between these two algorithms. Sheikh et al. [13] anticipated loan security. The logistic regression algorithm is utilized to forecast the safety of a loan. This algorithm is used to analyze the dataset and create predictions based on this dataset. They anticipate if a new loan applicant is likely to default on a loan based on their analysis. Using Random Forest, they achieved an 81% accuracy rate. There are numerous approaches for evaluating the performance of models in model evaluation. Including Confusion metrics, Accuracy, Precision, Recall, and the F1 score, among others. Each method yields the same performance result. Before performing the algorithm, this system employs approaches for data cleansing and processing. Including Imputation, Outlier Handling, Binning, and Log Transformation. When Regina et al. [14] predict the safety of a loan, they employ the logistic regression approach. This algorithm is used to analyze the dataset and create predictions based on this dataset. They anticipate if a new loan applicant is likely to default on a loan based on their analysis. Using Random Forest, they achieved an 81% accuracy rate. There are numerous approaches for evaluating the performance of models in model evaluation. Including Confusion metrics, Accuracy, Precision, Recall, and the F1 score, among others. Each method yields the same performance result. Before performing the algorithm, this system employs approaches for data cleansing and processing. Imputation, Handling Outliers, Binning, Log Transformation are examples. The proposed model achieved a classification accuracy of 75.08% when utilizing the R package. Shinde et al. [15] created a system to give an immediate, rapid, and simple method for selecting the best applicant for commercial bank loan lending. The credit analysis system enables the prioritization of the individual candidate who can afford this loan. Business understanding, data convention, data processing, and modeling are the four different aspects of the proposed framework. This system employs the Logistic Regression with stratified k-folds cross-validation and Random Forest methods. The Logistic Regression Model's mean validation accuracy and mean validation f1 score are 72% and 83%, respectively. The mean validation accuracy of Random Forest Model hyperparameters is 79%. Priya et al. [16] determined a customer's creditworthiness for loan approval. This process is contingent on a number of characteristics, including credit history, Installment, etc. After gathering data, they preprocessed it, constructed a classification model, and made predictions. In this system, the Random Forest classification technique was implemented. By assessing this system, it is possible to determine that Random Forest provides excellent accuracy. This method is a technique for supervised learning. 81.10 percent was the experimental result, which is relatively straightforward to modify. Applicants are applying for employment-related bank loans. Due to the bank's limited resources, only a limited number of applicants can qualify. By analyzing all of the available data, our model concludes that applicants with high income and smaller loan amounts are more likely to be approved, which makes sense given that they can repay the loan on time. The analyzing study indicates that some fundamental characteristics, such as gender and marital status, do not appear to be taken into account. The primary goal of this system is to divide the groups of customers who apply for loans into those who are in default and those who are not, so that credit lenders can use the information. It employs the Min-Max normalization and K-Nearest Neighbor (K-NN) classifier in combination. In The proposed model in this paper predicts whether a loan applicant will be a valid customer or a default customer. The classifier essentially provides the dataset's credit scoring results. These credit scores are used to determine which classifier produces the most accurate results. The credit scoring model that is being proposed is based on K-NN and offers greater accuracy than other classifiers. These suggested models have an accuracy rate of 75.08% when classifying credit applicants with the R package. This accuracy was achieved by splitting the dataset into two equal pieces. Data from training and testing, for instance [1]. Sujatha et al. [17] created a web-based application to do extensive and much more accurate prediction utilizing logistic regression, which was integrated in the Python programming language, after realizing the significance of loan forecasting in the modern banking system. The system can produce results with a high degree of accuracy and a minimal loss of training and validation data. However, it should be emphasized that the system's functionality is constrained by a number of characteristics and cannot help customers outside of those features. In order to lower the risk involved in choosing the safe individual and conserve a lot of bank resources and labor, Foster et al. [18] surely made a contribution to this research by using loan defaults and risk models to forecast bankruptcy. When loan default condition and/or audit opinion factors are left out of hazard bankruptcy prediction models, the authors looked at how the findings change. In order to find variables for practical bankruptcy prediction models and validate hypotheses, the research used logistic regression. The findings enhance the hazard model's accuracy for samples with limited financial resources. Data mining was used by Kruppa et al. [19] to discover the industries that frequently commit financial statement fraud. On a dataset of 202 Chinese enterprises, six algorithms were applied, evaluated, and analyzed with and without feature selections. Probabilistic Neural Network (PNN) performed the best when there was no feature selection out of all six data mining techniques. In terms of feature extraction, GP and PNN fared better than others with accuracy levels that were almost equal. Both traditional algorithms and manual loan approval processes have been characterized by poor performance and low recognition rates. In light of this, Zhang et al. [20] suggested an integrated training classification model that made use of support vector machines that used particle swarm optimization (PSO) (SVM). SVM was optimized using PSO, and an established prediction model was integrated with SVM's weak classifier using AdaBoost. It was found that the AdaBoost-PSO-SVM method may successfully raise the accuracy level. The tiny sample size employed for the categorization is the main challenge. In an effort to determine the credit worthiness of potential customers, Luczak et al. [21] tried to compare the performance of seven classifiers. Two datasets were used. The difficulty with the work is that the ensemble cannot be led for improved performance and evaluation. According to Arushi Jain et al. [22], financial fraud detection can have an impact on a variety of industries, including banking, insurance, government organizations, and law enforcement. Every day, there are numerous currency swaps and transactions, and the number of fraud incidents is increasing. Utilized in this manner Although there are more complex expansions, assets retrieval can also be modeled mathematically as a structural function simulating binary dynamics. Asset order evaluates the order model's parameters in multivariate analysis. The joint conditional probability density function is specified. Belief networks, Bayesian networks, and probabilistic networks are other names for them. It is a skeleton that resembles a flowchart, with each internal node standing for a "test" on the attribute, each branch for an extension of the test, and each leaf node for a class label.

2.3 Comparative Analysis and Summary

Table 2.1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Kadam et al. [5]	SVM and Naïve Bayes	Naïve Bayes gives better accuracy
2.	Kumar et al. [6]	Decision Tree, Support Vector Machine, Random Forest, Linear Models, Adaboost, Neural network	———
3.	Supriya et al. [7]	Decision Tree, Random Forest, Logistic Regression	81%
4.	Gautam et al. [8]	Decision Tree, Random Forest	Random Forest=85.75%
5.	Gupta et al. [9]	Logistic regression, Random Forest	———
6.	Aruthiothi et al. [1]	KNN, Min-Max Normalization	KNN=88.63%
7.	Gomathy et al. [11]	Logistic Regression, KNN, Random Forest, Decision Tree, SVM, Neural Network	———
8.	Mehul et al. [12]	Decision Tree, Random Forest	Random Forest=80%
9.	Sheikh et al. [13]	Logistic Regression, KNN, SVM	81.10%
10.	Regina et al. [14]	Decision Tree, Linear regression, KNN, Naive Bayes, Neural Network, Ensemble Method	Naïve Bayes=98%
11.	Shinede A. et al. [15]	Random Forest, Logistic Regression	Random Forest=79.47%
12	Kruppa et al. [19]	Random Forest, KNN, bNN	Random Forest performed best

2.4 Scope of the Problem

Using data analysis, machine learning algorithms, and ensemble techniques, this project is contributing to the development of a model that can predict loan defaulters. We can use this model to predict the reasons why people default on loans, which will help society deal with those problems.

As a result, the sole focus of this model is the identification of the potential causes of banking issues. We are developing a model using machine learning and artificial intelligence to identify the riskiest loan applicants. As a result, we considered developing a model that could anticipate loan defaulters.

2.5 Challenges

We face quite a challenge with this project. When carrying out the project in accordance with our plans, we encountered numerous challenges. It takes a long time to finish anything, especially when it comes to training. We need to give this a lot of time in order to find people who default on loans.

CHAPTER 3

Research methodology

3.1 Research Subject and Instrumentation

Investigation of loan prediction assessments via machine learning algorithm and ensemble methods are the focus included in our study.

Google Colab

Colab is essentially a totally mist based, free Jupyter sketchbook atmosphere. Most significantly, Colab doesn't essential to be established, and the sketchpads you generate can have multiple team associates editing that one at once, alike to how you run documents in Google Docs dictionary. The fact that Colab cares the most broadly castoff machine learning collections and that they are modest to weight onto your notebook is its greatest advantage.

It's a held Jupyter notebook with a countless allowed kind that delivers free admission to Google dispensation incomes like GPUs and TPUs and necessitates no arrangement.

NumPy

The collection acknowledged as NumPy, or "Arithmetical Python," has multidimensional collection substances and methods for processing those collections. It also offers purposes for employed in the parts of media, the Fourier convert, and line algebra. In the year 2005, Travis Oliphant established NumPy. You can practice it for it is an exposed basis scheme. A universal- drive set for treatment collections is named NumPy. Its proposals a multidimensional collection thing with unresolved rapidity as well as competences for interrelating with these collections. It is the keystone Python unit for scientific calculating. The programmed is exposed basis. The goalmouth of NumPy is to offer collection substances that are up to 50 times earlier than conservative Python slopes.

Pandas

The most well-known Python programming language package for data operation and examination is called Pandas. Pandas provides information structures and processes for strong, versatile in order to user-friendly data analysis and manipulation as an exposed basis software library developed on highest of Python exactly for information operation and analysis. Pandas enhances Python by enabling it to interact with data similar to spreadsheets, facilitating quick loading, aligning, manipulating, and merging in addition to other crucial operations. Data Frames, which are two-dimensional array-like data tables with one variable's values in each column and a set of those values in each row for each row, are part of the Pandas open-source package. A Data Frame may contain data of the character, factor, or numeric types. Data frames created with Pandas can alternatively be compared to a dictionary or a grouping of series objects.

Matplotlib

The large information arithmetical organization tool NumPy comprises the graphing set Matplotlib for the Python programming linguistic. Conspiracies are entrenched in Python programmed by means of Matplotlib's thing concerned with API. Python's Matplotlib toolkit delivers a whole instrument for construction still, lively, and collaborating imaginings. Matplotlib makes problematic things conceivable and modest belongings informal.

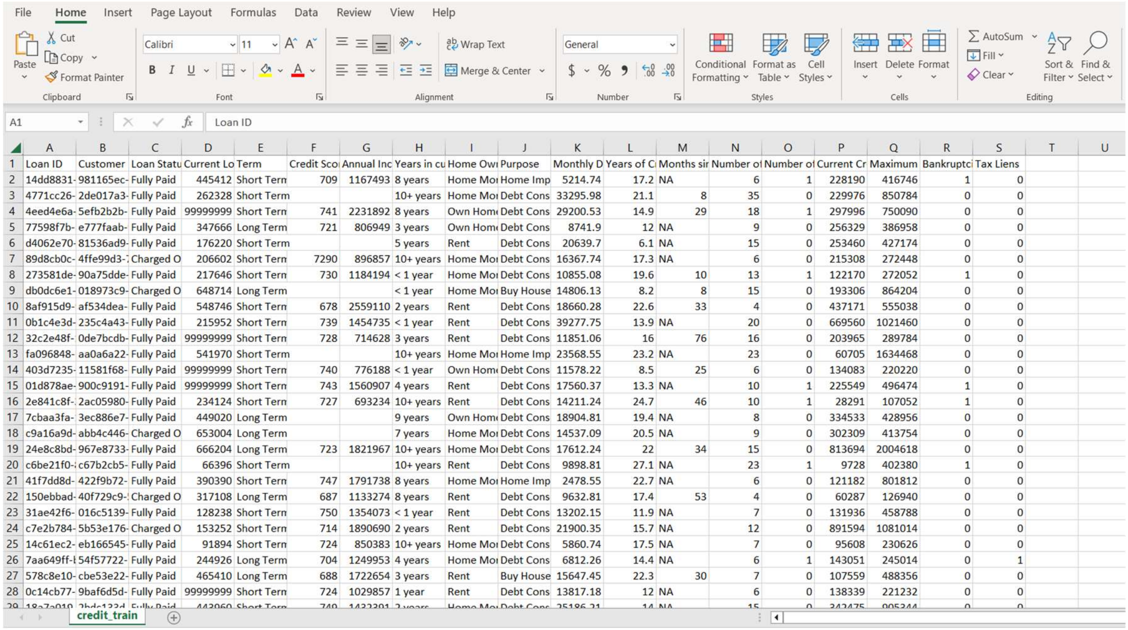
Seaborn

Seaborn basically a python data visualization library which is based on Matplotlib. we use seaborn to creating attractive and informative statistical graph. It is more comfortable to handling pandas data frames.

3.2 Data Collection Procedure/Dataset Utilized

In information accumulation step we are gathering information from Kaggle. Our dataset is downloaded in CSV format and use for research purpose. Total of 100514 entries is used in our test project.

In figure-3.1, it shows the dataset in CSV format-



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Loan ID	Customer	Loan Statu	Current Lo	Term	Credit Sco	Annual Inc	Years in cu	Home Ow	Purpose	Monthly D	Years of C	Months sir	Number of	Number of	Current Cr	Maximum	Bankruptc	Tax Liens
2	14dd8831-	981165ec-	Fully Paid	445412	Short Term	709	1167493	8 years	Home Mo	Home Imp	5214.74	17.2	NA	6	1	228190	416746	1	0
3	4771cc26-	2de017a3-	Fully Paid	262328	Short Term			10+ years	Home Mo	Debt Cons	33295.98	21.1	8	35	0	229976	850784	0	0
4	4eed4e6a-	5efb2b2b-	Fully Paid	99999999	Short Term	741	2231892	8 years	Own Hom	Debt Cons	29200.53	14.9	29	18	1	297996	750090	0	0
5	77598f7b-	e777faab-	Fully Paid	347666	Long Term	721	806949	3 years	Own Hom	Debt Cons	8741.9	12	NA	9	0	256329	386958	0	0
6	d4062e70-	81536ad9-	Fully Paid	176220	Short Term			5 years	Rent	Debt Cons	20639.7	6.1	NA	15	0	253460	427174	0	0
7	89d8cb0c-	4ffe99d3-	Charged O	206602	Short Term	7290	896857	10+ years	Home Mo	Debt Cons	16367.74	17.3	NA	6	0	215308	272448	0	0
8	273581de-	90a75dde-	Fully Paid	217646	Short Term	730	1184194	< 1 year	Home Mo	Debt Cons	10855.08	19.6	10	13	1	122170	272052	1	0
9	db0dc6e1-	018973c9-	Charged O	648714	Long Term			< 1 year	Home Mo	Buy House	14806.13	8.2	8	15	0	193306	864204	0	0
10	8af915d9-	af534dea-	Fully Paid	548746	Short Term	678	2559110	2 years	Rent	Debt Cons	18660.28	22.6	33	4	0	437171	555038	0	0
11	0b1c4e3d-	235c4a43-	Fully Paid	215952	Short Term	739	1454735	< 1 year	Rent	Debt Cons	39277.75	13.9	NA	20	0	669560	1021460	0	0
12	32c2e48f-	0de7bcdb-	Fully Paid	99999999	Short Term	728	714628	3 years	Rent	Debt Cons	11851.06	16	76	16	0	203965	289784	0	0
13	fa096848-	aa0a6a22-	Fully Paid	541970	Short Term			10+ years	Home Mo	Home Imp	23568.55	23.2	NA	23	0	60705	1634468	0	0
14	403d7235-	11581f68-	Fully Paid	99999999	Short Term	740	776188	< 1 year	Own Hom	Debt Cons	11578.22	8.5	25	6	0	134083	220220	0	0
15	01d878ae-	900c9191-	Fully Paid	99999999	Short Term	743	1560907	4 years	Rent	Debt Cons	17560.37	13.3	NA	10	1	225549	496474	1	0
16	2e841c8f-	2ac05980-	Fully Paid	234124	Short Term	727	693234	10+ years	Rent	Debt Cons	14211.24	24.7	46	10	1	28291	107052	1	0
17	7cbaa3fa-	3ec886e7-	Fully Paid	449020	Long Term			9 years	Own Hom	Debt Cons	18904.81	19.4	NA	8	0	334533	428956	0	0
18	c9a16a9d-	abb4c446-	Charged O	653004	Long Term			7 years	Home Mo	Debt Cons	14537.09	20.5	NA	9	0	302309	413754	0	0
19	24e8c8bd-	967e8733-	Fully Paid	666204	Long Term	723	1821967	10+ years	Home Mo	Debt Cons	17612.24	22	34	15	0	813694	2004618	0	0
20	c6be21f0-	c67b2cb5-	Fully Paid	66396	Short Term			10+ years	Rent	Debt Cons	9898.81	27.1	NA	23	1	9728	402380	1	0
21	41f7dd8d-	422f9b72-	Fully Paid	390390	Short Term	747	1791738	8 years	Home Mo	Home Imp	2478.55	22.7	NA	6	0	121182	801812	0	0
22	150ebbad-	40f729c9-	Charged O	317108	Long Term	687	1133274	8 years	Rent	Debt Cons	9632.81	17.4	53	4	0	60287	126940	0	0
23	31ae42f6-	016c5139-	Fully Paid	128238	Short Term	750	1354073	< 1 year	Rent	Debt Cons	13202.15	11.9	NA	7	0	131936	458788	0	0
24	c7e2b784-	5b53e176-	Charged O	153252	Short Term	714	1890690	2 years	Rent	Debt Cons	21900.35	15.7	NA	12	0	891594	1081014	0	0
25	14c61ec2-	eb166545-	Fully Paid	91894	Short Term	724	850383	10+ years	Home Mo	Debt Cons	5860.74	17.5	NA	7	0	95608	230626	0	0
26	7aa649ff-	54f57722-	Fully Paid	244926	Long Term	704	1249953	4 years	Home Mo	Debt Cons	6812.26	14.4	NA	6	1	143051	245014	0	1
27	578c8e10-	cbe53e22-	Fully Paid	465410	Long Term	688	1722654	3 years	Rent	Buy House	15647.45	22.3	30	7	0	107559	488356	0	0
28	0c14cb77-	9baf6d5d-	Fully Paid	99999999	Short Term	724	1029857	1 year	Rent	Debt Cons	13817.18	12	NA	6	0	138339	221232	0	0

Figure 3.1: Dataset in CSV format

3.3 Statistical Analysis

- Our dataset has two types of data such as Categorial and Numerical.
- Categorical data are Loan Status, Years in current job, Home Ownership, Purpose, Term.
- Unnecessary attributes are Loan ID and Customer ID which are categorical in number type.
- Numerical data are Credit Score, Annual Income, Monthly Debt, Years of Credit History, Months since last delinquent, Number of Open Accounts, Number of

Credit Problems, Current Credit Balance, Maximum Open Credit, Bankruptcies, Tas Liens

- Dataset is saved in Microsoft excel which extension is CSV.

3.4 Proposed Methodology/Applied

In figure-3.2, it shows the entire working process-

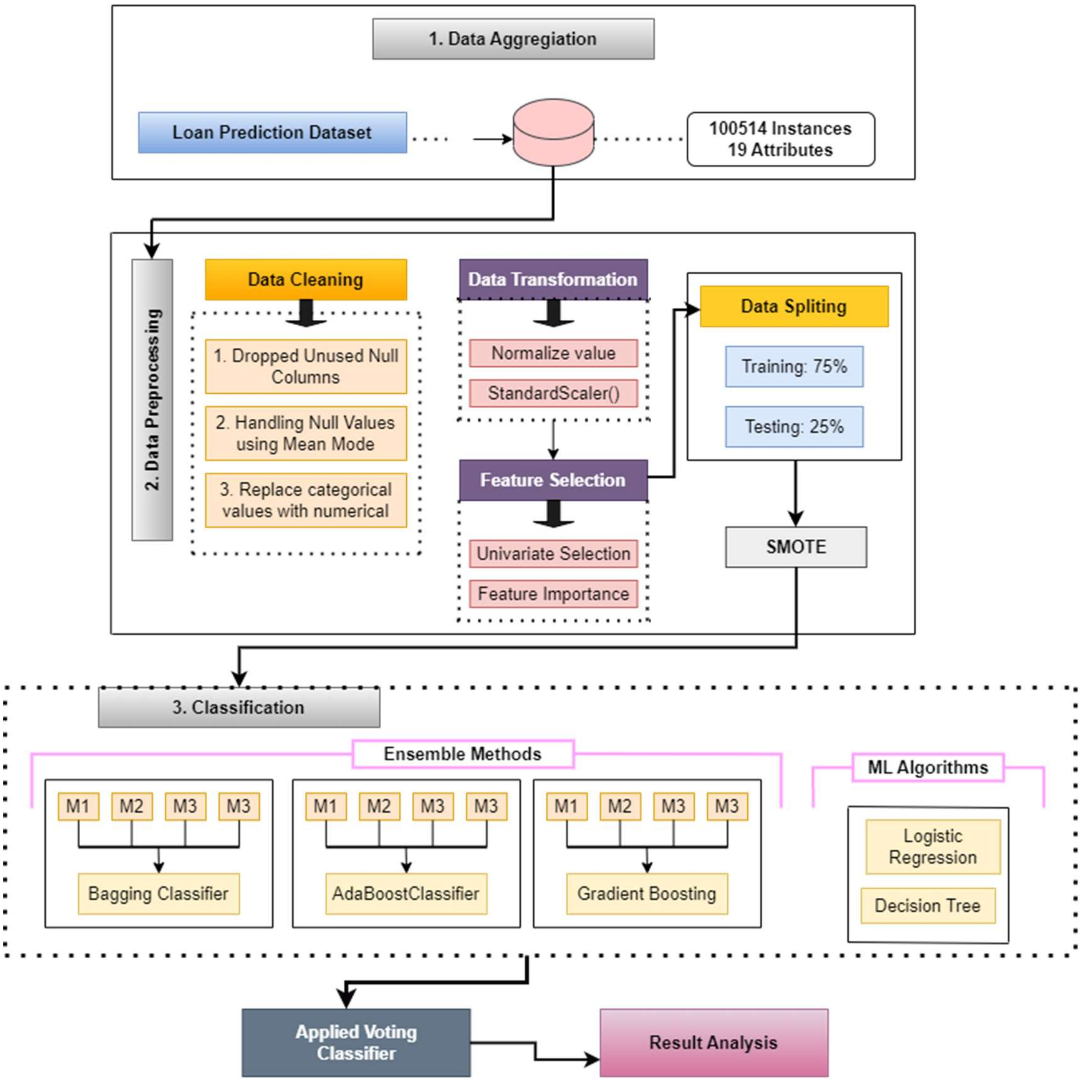


Figure 3.2: Stages of Loan Prediction Analysis

Kaggle is used to obtain a publicly available loan prediction dataset. Several types of data preprocessing are performed to help the model make a better prediction. Unused null columns are taken out of the features, several null values are filled in using the mean mode method, and categorical values are changed to numeric values in the data preprocessing method. In data transformation, StandardScaler is performed and the values are normalized. For feature selection, univariate selection and feature importance is performed. Top features from each feature selection are splitted by 70% for training and 30% for testing. Then SMOTE is performed in the dataset to equalize the data. The equalized data is fed to the machine learning models. To boost the result of the models, voting classifier is applied. Finally, the results from the model are compared with each other and the best model is proposed to predict loan defaults.

Data Pre-processing

Data preprocessing edits, rewrites, and organizes data for analysis. Classification often ignores some facts. This study's data cleaning and preparation are described below.

a) **Drop unnecessary columns:** The dataset contains some irrelevant information that would not contribute to classification. In our dataset, such columns include 'Credit Score', 'Current Loan Amount', 'Annual Income', 'Years of Credit History', 'Current Credit Balance', 'Months since last delinquent', 'Maximum Open Credit', 'Monthly Debt' and 'Number of Open Accounts'. They were eliminated because they serve no classification purpose.

b) **Handling Null Values:** There are several null values in the columns. The values are replaced using mean-mode method [23].

The arithmetic mean is obtained by adding the numbers and dividing the result by the total number of numbers in the list. The term "average" is typically used to refer to this.

The equation for the mean is:

$$\text{Mean Formula} = \frac{\text{Sum of Observations}}{\text{Total Numbers of Observations}} \tag{1}$$

Mode is the value that appears most often in a list. The equation for the mode is:

$$\text{Mode formula} = \mathrm{M} + \mathrm{i}\,\frac{(x_m - x_1)}{(x_m - x_1) + (x_m - x_2)} \tag{2}$$

Here,

- 'M' is modal class's lower limit.
- 'i' is Class interval size.
- 'x_m' is modal class frequency.
- 'x_1' is the class before the modal class's frequency.
- 'x_2' is the class following the modal class's frequency.

c) **Replaced categorical values with numerical:** Replaced the categorical values of the features to numerical value to train the data into the model. The converted values are: "Home Ownership":{'Home Mortgage':0,'Rent':1,'Own Home':2,'HaveMortgage':3}, "Term":{'Short Term':0,'Long Term':1} and so on.

Data Transformation

The data values are standardized using the StandardScaler [24] to create a standard format. StandardScaler rescales the dataset to have a mean of zero and a standard deviation of one. Changes the dataset's look. Subtracting the mean from the original number and dividing by the standard deviation yields the converted value.

Feature Selection

Important features help the model to perform the best prediction. To extract the most important features from the dataset two different feature selection techniques are applied. They are described below.

- **Univariate feature selection:** In univariate feature selection, univariate statistical tests are used to select the best features [25]. We compare the two to see if there is a statistically significant connection between each feature and the target variable. Moreover, known as examination of fluctuation (ANOVA). We don't consider different highlights while inspecting the association between an element and the objective variable. This is why it is referred to as "univariate." A test result exists for each feature. The highlights of the high.

- **Feature importance:** Another method for selecting features that makes use of the Extra Trees Classifier is feature importance [26]. The Extremely Randomized Trees Classifier, also known as the Extra Trees Classifier, is a type of ensemble learning method that produces its classification result by combining the results of numerous de-correlated decision trees gathered in a "forest." The construction of the decision trees in the forest is the only conceptual distinction between it and a Random Forest Classifier.

  The underlying preparation test is utilized to fabricate every choice tree in the Additional Trees Timberland. After that, at each test node, a random sample of k features from the feature set is given to each tree. From this sample, the tree must select the most suitable feature to divide the data in accordance with a particular mathematical criterion. This random sampling of features results in the production of numerous de-correlated decision trees.

  During the creation of the forest, the normalized total drop with in mathematical criteria utilized in the decision of characteristic of split is computed for each feature in order to perform feature extraction using the aforementioned forest structure. The Gini Value of the feature is the name given to this value. Each feature is ranked according to its Gini Importance in descending order to perform feature extraction, and the user then chooses the top k features that appeal to them.

From each feature selection technique, we get top fourteen features. The top fourteen features of feature importance selection are: Term, Years in current job, Home Ownership, Purpose, Number of Credit Problems, Number of Open AccountsLog, Monthly DebtLog, Maximum Open CreditLog, Months since last delinquentLog, Current Credit BalanceLog, Years of Credit HistoryLog, Annual IncomeLog, Current Loan AmountLog and Credit ScoreLog. The top fourteen features of Univariate feature selection are: Credit ScoreLog, Current Loan AmountLog, Term, Annual IncomeLog, Maximum Open CreditLog,Home Ownership, Years of Credit HistoryLog, Monthly DebtLog, Number of Open

AccountsLog, Months since last delinquentLog, Tax Liens, Current Credit BalanceLog, Bankruptcies and Purpose. These features are split and trained differently to the models. The results of different feature selection are presented in the result section.

Data Split

Data splitting, the model is trained, and the loan prediction dataset is tested or evaluated using a two-part split. The dataset is divided into train and test datasets at a ratio of 75:25.

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is carried out to address the loan dataset's data imbalance issue [27]. Synthetic Minority Oversampling Technique, or SMOTE for short, is a potent remedy for data imbalances. SMOTE, an algorithm for data augmentation, generates fictitious data points based on the real ones. SMOTE can be thought of as an enhanced form of oversampling or as a particular method for enhancing data. With SMOTE, we produce artificial data points that are only slightly different from the actual data points rather than producing duplicate data points.

The following is how the SMOTE algorithm works. Our sample comes from the minority class at random. The observation in this dataset will be used to calculate the k closest neighbors. After that, one of those neighbors will be chosen to determine the vector that runs between the current data point and the chosen neighbor. The vector is given a random number from 0 to 1. The synthetic data point is made by combining this with the existing data point.

In point of fact, this procedure is comparable to a data point that has been slightly shifted in the direction of a neighbor. By doing this, we make sure that our made-up data point is not exactly the same as an actual data point and that it is also not significantly different from the known data for our minority class.

Models

Three ensemble classifiers and two machine learning based classifier is used to evaluate this study. The details of the models below [28–30]:

Ensemble method

An ensemble method is a machine learning technique that combines the predictions of multiple individual models to make a final prediction. Ensemble methods are used to improve the stability and predictive power of models by combining the strengths of multiple models. This is often done by training multiple models on the same data and then averaging their predictions, although there are many other ways to combine the predictions of multiple models. Ensemble methods are widely used in practice because they often lead to improved performance over a single model.

In figure-3.3, it shows the entire working process of ensemble method-

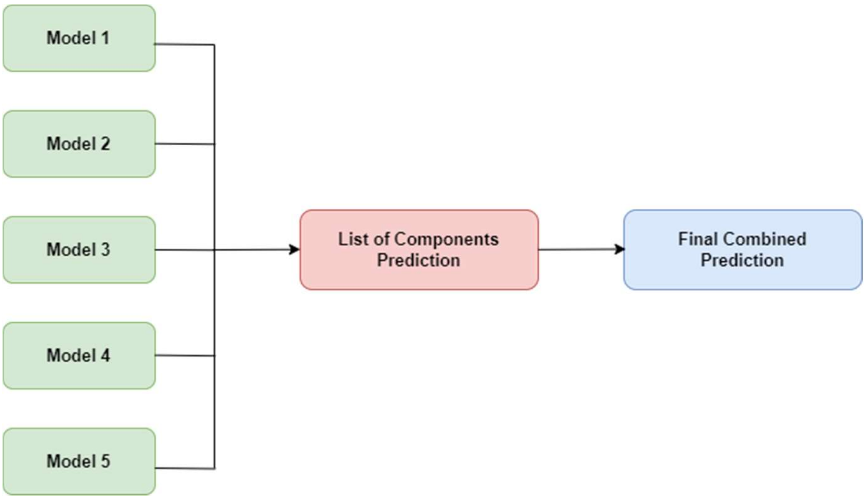


Figure 3.3: Ensemble method

Among many types of ensemble method, Bagging, Gradient Boosting and Adaboost are prominent algorithm. In our paper we used these ensemble methods, the details about the models are below:

- **Bagging Classifier:** Bagging (short for bootstrapped aggregation) is an ensemble machine learning algorithm that can be used to improve the stability and accuracy of machine learning models. It works by training multiple models on different random subsets of the training data and then averaging the predictions of these models.

  Here's a brief overview of how bagging works:

  - The training data is split into a number of random subsets.
  - A model is trained on each of these subsets.
  - The predictions of the individual models are combined to make a final prediction.

  The final prediction is often made by averaging the predictions of the individual models, although other methods such as weighted averaging can also be used.

  In figure-3.4, it shows the entire working process of Bagging Classifier-

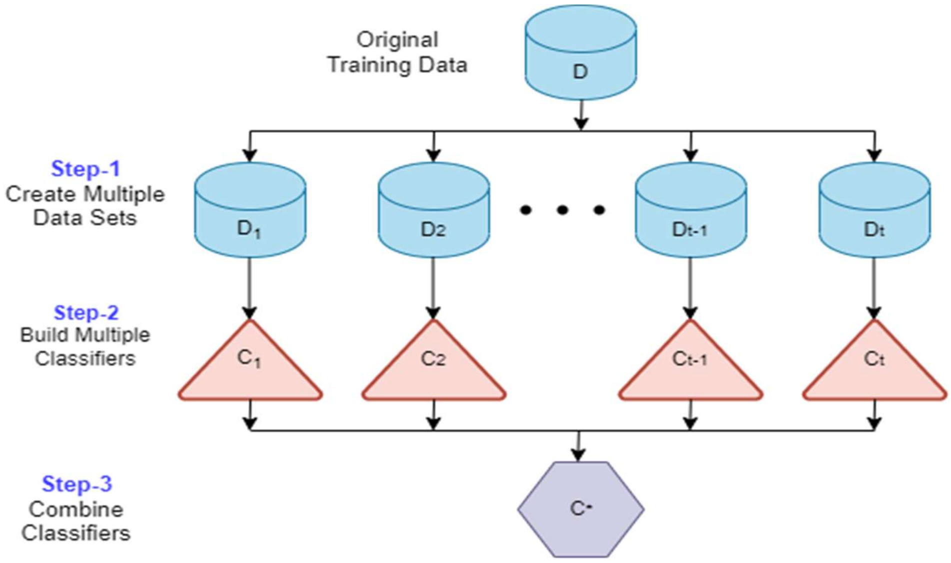


Figure 3.4. Bagging Classifier

  Bagging can be used with any machine learning algorithm, but it is particularly effective with models that have high variance, such as decision trees. By training multiple models on different subsets of the data, bagging can help to reduce the variance of the overall model and improve its generalization performance.

- **Gradient Boosting Classifier (GB):** Gradient Boosting is an ensemble machine learning algorithm that can be used to improve the accuracy of a model by combining the predictions of multiple individual models. It works by training a series of models in a sequential manner, with each model trying to correct the mistakes of the previous model.

  Here's a brief overview of how gradient boosting works

  - A model is trained on the training data.
  - The model makes predictions on the test data.
  - The predictions of the model are compared to the true values, and the error is calculated.
  - A new model is trained to predict the error of the previous model.
  - Steps 2-4 are repeated until a predetermined number of models have been trained.
  - The predictions of the individual models are combined to make the final prediction.

  Gradient Boosting is often used with decision trees, but it can be used with any machine learning algorithm. It is particularly effective for tasks where the model needs to make a lot of decisions based on complex data. By training a series of models to correct the mistakes of the previous model, gradient boosting can produce highly accurate models.

- **AdaBoost Classifier:** AdaBoost (short for Adaptive Boosting) is an ensemble machine learning algorithm that can be used to improve the accuracy of a model by combining the predictions of multiple individual models. It works by training a series of models in a sequential manner, with each model trying to correct the mistakes of the previous model.

Here's a brief overview of how AdaBoost works:

- A model is trained on the training data.
- The model makes predictions on the test data.
- The predictions of the model are compared to the true values, and the error is calculated.
- The weights of the incorrectly classified data points are increased, so that the next model pays more attention to them.
- A new model is trained using the updated weights.
- Steps 2-5 are repeated until a predetermined number of models have been trained.
- The predictions of the individual models are combined to make the final prediction.

AdaBoost is often used with decision trees, but it can be used with any machine learning algorithm. It is particularly effective for tasks where the model needs to make a lot of decisions based on complex data. By training a series of models to correct the mistakes of the previous model, AdaBoost can produce highly accurate models.

The equation of Adaboost is:

$$\boldsymbol{n_i \epsilon \mathbb{R}^x, z_i \epsilon \{-1, 1\}}. \qquad (1)$$

Our dataset has x characteristics, or real number dimensions. Data points are n. The goal variable, z, is either -1 or 1, indicating the first or second class in a binary classification issue.

In figure-3.5, it shows the entire working process of Adaboost Classifier-

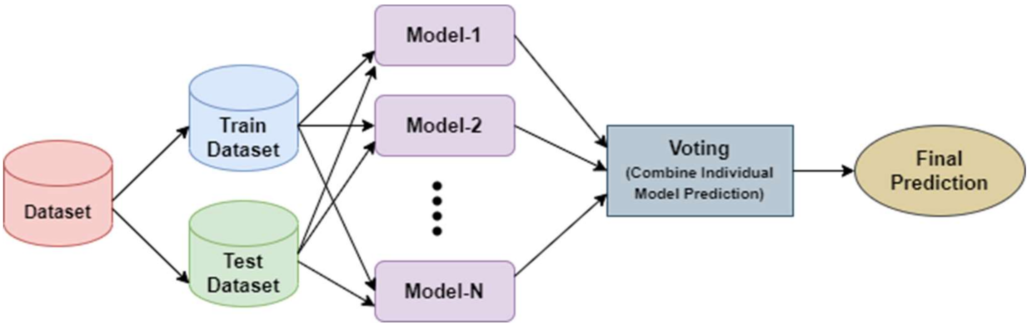


Figure 3.5: Adaboost Classifier

Machine Learning Models

Among several machine learning models, Decision Tree and Logistic Regression are used in this experiment.

- **Decision Tree Classifier (DT):** A decision tree is a flowchart-like tree structure used to make a decision or prediction. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, and a leaf node represents a decision. The topmost node in a decision tree is the root node, and the nodes that do not have any children are called leaf nodes.

  Decision trees are used in many areas, including business, economics, and medicine. They are a useful tool for predicting the likelihood of an event based on certain conditions. For example, a decision tree could be used to predict whether an individual is likely to have a certain disease based on their age, sex, and other factors.

  The decision tree algorithm is a supervised learning algorithm, which means that it requires a training data set to learn from. The algorithm starts with the root node and then splits the data based on the most important feature (a feature is a specific characteristic of the data). The resulting sub-nodes are then split based on the next most important feature, and so on, until the leaf nodes are reached. The resulting tree can be used to make predictions on new data by following the same path through the tree as was taken by the training data.

- **Logistic Regression:** Logistic regression is a statistical method for predicting binary outcomes. It is used to model the probability of a certain class or event occurring. The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable (the thing we want to predict) and one or more independent variables (the things that we are using to predict the dependent variable).

  In logistic regression, the dependent variable is binary, meaning that it can only take on two values, such as "win" or "lose," "success" or "failure," or "dead" or "alive." The independent variables can be continuous or categorical.

  The logistic regression model is a mathematical equation that takes the form:

$$z = \frac{e^{(a_0+a_1y)}}{1+ e^{(a_0+a_1y)}} \qquad (2)$$

  here z is the probability of the event occurring, y is the independent variable, and a_0 and a_1 are the model coefficients.

  Logistic regression is used in a variety of fields, including finance, psychology, and biology. It is a widely used tool in data analysis and predictive modeling.

3.5 Implementation Requirements

We have explored previous banking history of banking sector and we analyzing all statistical and theoretical concepts and approaches related to this research project.

Software/Hardware:

- Operating System

- Hard disk (minimum 500GB)
- RAM (minimum 4GB)

Developing tools:

- Colab environment
- Google drive
- Good internet connection
- Any browser (Mozilla Firefox or Google Chrome)

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

At first, we collected our dataset from Kaggle which have 100514 rows and 19 columns. Secondly, we preprocessed the dataset and split the dataset into two parts such as train data and test data. After that we trained the machine using the train data. The test data is then taken for analysis.

4.2 Experimental Results & Analysis

The evaluation is using the confusion matrix like, accuracy, precision, recall, and F1 score. True positive (TP) values are true in reality. False positives (FP) occur when false results are mislabeled. The third form, false negative (FN), occurs when a correct value is misinterpreted as negative. TN and FN are the fourth and fifth choices. A true negative (TN) is a positive value misidentified as negative. Fourth is true negative (TN).

Precision: The correct optimistic to total true positive in addition false positive ratio is known as precision. Precision checks to see how many false positives were included in the sample. If there aren't any false positives (FPs), the model's precision was 100 percent. The precision will appear uglier when additional FPs are extra to the mix.

$$\text{Precision} = TP/(TP+FP)$$

Recall: Recall takes a dissimilar path. Recall examines the number of mistaken rejections that were comprised in the prediction process somewhat than the quantity of untrue positives the model predicted. Each period a forecast false negative happens, the recall rate is punished. The equations themselves are opposites because of the penalties for precision and recollection. The yin and yang of evaluating the confusion matrix are precision and recall.

$$\text{Recall} = TP/ (TP+FN)$$

F-measure: Individual once exactness and recall are together 1 fixes the F1 Score turn out to be 1. Individual at what time exactness and recall are both strong container the F1 score rise. An additional valuable metric than accurateness is the F1 score, which is the vocal mean of recall and precision.

F-score = 2*((Precision*Recall) / (Precision + Recall))

Accuracy: The exactness is strongminded by in-between the entire number of correct guesses by the total amount of explanations in the dataset. The precision ranges from 0.0 to 1.0, with 1.0 existence the finest. It can also be strongminded by dividing by the ERR.

This section will go through the results of this paper. As mentioned earlier, we have done two different feature selections and top features from both feature selection are ran them separately to the eleven machine learning models.

Table 4.1 shows the accuracy, precision, recall and F1 score of eleven machine learning models without any feature selection.

Table 4.1: Classification Report of the machine learning models without any feature selection.

Model	**Accuracy (%**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Decision Tree (DT)	73.97	84.00	82.00	83.00
Gradient Boosting Classifier (GB)	70.79	85.00	76.00	80.0
AdaBoost Classifier	76.00	83.00	86.00	85.00
Bagging Classifier	**80.37**	83.00	**94.00**	**88.00**
Logistic Regression (LR)	66.00	**86.00**	67.00	75.00

From Table 4.1, we can see that, in terms of loan prediction, the highest accuracy of 80.37% is achieved by Bagging Classifier. It also achieved the highest recall and F1-score of 94.00% and 88.00%. The lowest accuracy is achieved by Logistic Regression, which is 66.00%. The lowest precision of 83.00% is achieved by AdaBoost and Bagging classifier.

Logistic Regression has the lowest recall and F1 score of 67.00% and 75.00%, respectively. The other models' accuracy ranges between 71% and 76%. So, without any feature selection, Bagging may be the best fit.

Table 4.2 shows the accuracy, precision, recall and F1 score of eleven machine learning models using the features of feature importance selection.

Table 4.2: Classification report using features of the Feature importance selection.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree (DT)	73.41	84.00	81.00	83.00
Gradient Boosting Classifier (GB)	70.45	85.00	75.00	80.00
AdaBoost Classifier	74.85	84.00	84.00	84.00
Bagging Classifier	**80.63**	83.00	**94.00**	**88.00**
Logistic Regression (LR)	66.02	**86.00**	67.00	75.00

From Table 4.2, we can see that, in terms of loan prediction, the highest accuracy of 80.63% is achieved by Bagging Classifier. It also achieved the highest recall and F1-score of 94.00% and 88.00%. The lowest accuracy is achieved by Logistic Regression which is 66.02%. The lowest precision of 83.00% is achieved by Gradient Boosting Classifier. Logistic Regression has the lowest recall and F1 score of 67.00% and 75.00%, respectively. The other models' accuracy ranges between 70% and 75%. So, with the features of feature importance's features, Bagging Classifier may be the best fit.

Table 4.3 shows the accuracy, precision, recall and F1 score of eleven machine learning models using the features of univariate feature selection.

Table 4.3: Classificatin Report using features of the Univariate feature selection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree (DT)	73.80	84.00	82.00	83.00
Gradient Boosting Classifier	63.21	**87.00**	62.00	72.00
AdaBoost Classifier	74.23	84.00	82.00	83.00
Bagging Classifier	**80.00**	83.00	**93.00**	**88.00**
Logistic Regression (LR)	66.50	86.00	68.00	76.00

According to Table 4.3, the Bagging classifier has the highest loan prediction accuracy, with 80.00%. It also achieved the highest recall and F1-score of 93.00% and 88.00%. The lowest accuracy is achieved by Gradient Boosting Classifier which is 63.21%. The lowest precision of 83.00% is achieved by Bagging Classifier. Gradient Boosting Classifier has the lowest recall and F1 score of 62.00% and 72.00%, respectively. The other models' accuracy ranges between 66% and 75%. So, with the features of Univariate feature selection, Bagging Classifier may be the best fit.

Table 4.4 shows the accuracy of different machine learning models after using the voting classifier on them.

Table 4.4: Accuracies of the models using voting classifier

Model	Without Feature Selection (%)	Feature Importance (%)	Univariate Feature Selection (%)
Decision Tree	80.42	80.27	81.00
Gradient Boosting Classifier	68.42	68.04	67.00
AdaBoost Classifier	77.30	77.21	76.00
Bagging Classifier	**86.86**	**87.00**	**88.00**
Logistic Regression	64.83	65.00	65.00

Without feature selection technique, Bagging Classifier achieves the highest accuracy of 86.86 %, and with Feature Importance and Univariate Feature Selection, Bagging classifier achieves the highest accuracy of 87.00% and 88.00%.

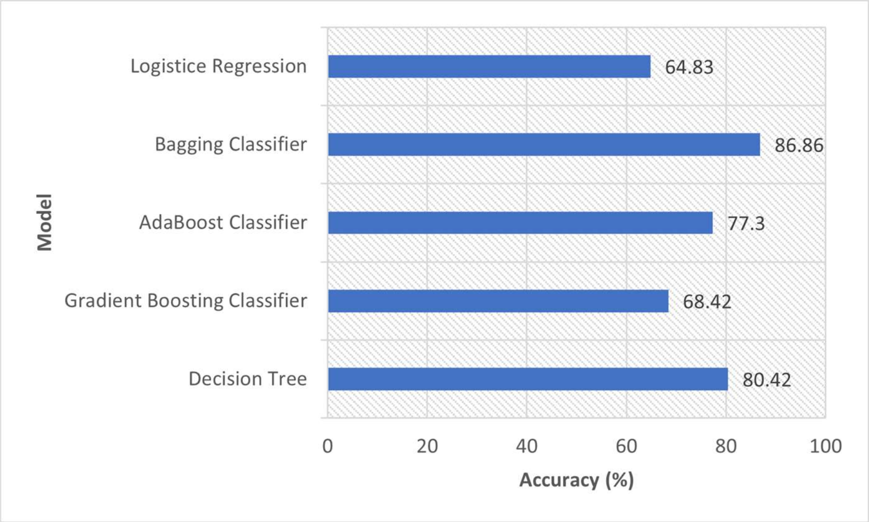


Figure 4.1: Accuracy comparison of various machine learning models without feature selection

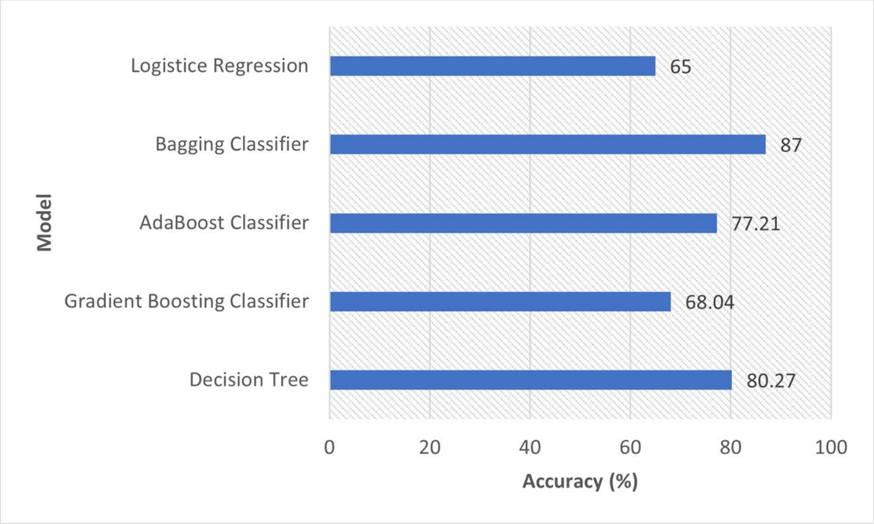


Figure 4.2: Accuracy comparison of various machine learning models with feature importance selection

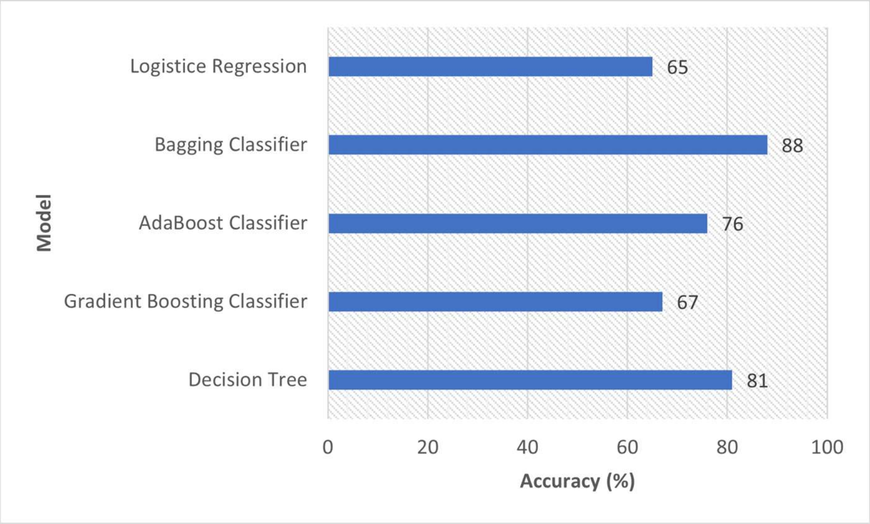


Figure 4.3: Accuracy comparison of various machine learning models with univariate feature selection

From the figure, Figure 4.1 is showing the results without any feature selection technique. Figure 4.2 is showing the results with feature's of feature importance technique. Figure 4.3 is showing the results with feature's of univariate feature selection. It is clearly visible without any feature selection technique, the features of feature importance technique and univariate feature selection technique, features of univariate feature selection is helping the machine learning models to achieve a good result for loan prediction. Among the models with the feature's of univariate feature selection, Bagging Classifier performed the best.

4.3 Discussion

After using voting classifier on the ensemble methods and machine learning algorithms, we got the best accuracy which was 88.00% with the help of Bagging Classifier. It is safe to say that the Bagging classifier is quite efficient and outperforms other models in terms of results. It works as intended and meets all bankers' requirements. The result is accurately and precisely calculated by this technology. It accurately predicts whether a customer or loan application will be accepted or rejected.

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

A loan prediction system is a tool that uses data and machine learning algorithms to predict the likelihood of a loan applicant being approved for a loan or credit. The use of a loan prediction system can potentially have a number of impacts on society.

One potential impact is that a loan prediction system can help to improve the efficiency and accuracy of the loan application process. By automating certain aspects of the process, a loan prediction system can help to reduce the time and resources required to evaluate loan applications, which can benefit both lenders and borrowers.

Another potential impact of a loan prediction system is that it can help to reduce the risk of discrimination in the loan application process. By basing loan decisions on objective data rather than subjective factors, a loan prediction system can help to reduce the risk of biased or unfair lending practices.

On the other hand, there are also potential concerns about the use of a loan prediction system, such as the possibility of errors or biases in the data used to train the system, which could lead to incorrect loan decisions. There is also the potential for the system to be used to exclude certain groups of people from accessing credit, which could have negative consequences for those individuals and for society as a whole.

Overall, the impact of a loan prediction system on society will depend on how it is designed and implemented, and it is important to carefully consider the potential impacts and take steps to mitigate any negative effects.

5.2 Impact on Environment

The impact of a loan prediction system on the environment will depend on the specific methods used to develop and implement the system. In general, the use of electronic or digital technologies, such as machine learning algorithms and computer systems, for the development and operation of a loan prediction system could potentially have a negative impact on the environment due to the energy and resources required to power and maintain these systems.

However, there are also ways in which a loan prediction system could potentially have a positive impact on the environment. For example, if the system is able to streamline and automate the loan application process, it could potentially reduce the need for certain types of paper-based documentation and communication, which could help to reduce the environmental impact of paper production and disposal.

Overall, the environmental impact of a loan prediction system will depend on the balance between the potential positive and negative impacts of the technology. It is important to carefully consider these potential impacts and take steps to minimize any negative effects.

5.3 Ethical Aspects

There are several ethical aspects to consider in the development and use of a loan prediction system.

One ethical concern is the potential for bias in the system. If the data used to train the system is biased or unrepresentative, it could lead to incorrect or unfair loan decisions. For example, if the data used to train the system is predominantly from a certain demographic group, the system may be more likely to approve loans for members of that group and less likely to approve loans for members of other groups. It is important to ensure that the data used to train the system is diverse and representative in order to avoid bias.

Another ethical concern is the potential for the system to be used to exclude certain groups of people from accessing credit. If the system is designed in a way that disproportionately denies loans to certain groups, it could have negative consequences for those individuals and for society as a whole. It is important to design the system in a way that is fair and unbiased, and to consider the potential impacts on disadvantaged or marginalized groups.

Another ethical aspect to consider is the transparency of the system. It is important to be transparent about how the system works and how loan decisions are made, so that borrowers understand the basis for the decisions and can challenge any errors or biases.

Overall, it is important to carefully consider the ethical implications of a loan prediction system and to take steps to ensure that it is fair, unbiased, and transparent.

5.4 Sustainability Plan

Here is a potential sustainability plan for a loan prediction system:

a) **Continuous evaluation:** The loan prediction system should be continuously evaluated for its performance and accuracy. This can be done through regularly scheduled evaluations or by using a monitoring system that tracks the system's performance in real-time.

b) **Regular updates:** The system should be regularly updated with new data and algorithms to improve its performance and accuracy. This can be done through updates to the training data or by incorporating new machine learning algorithms.

c) **Performance monitoring:** The system should be monitored for its performance and any issues should be promptly addressed. This can be done through automated monitoring systems or by a team of data scientists responsible for maintaining the system.

d) **User feedback:** The system should solicit and incorporate user feedback to improve its performance and usability. This can be done through surveys, focus groups, or other methods of gathering user feedback.

CHAPTER 6

Summary, conclusion, recommendation and implication for future research

6.1 Summary of the Study

We looked at a variety of Machine Learning classification models. There were primary components to our investigations. With the use of data pre-processing methods including null value handling using mean and mode, categorical data visualization, numerical data visualization, level encoding, replacing value, group by attributes, co-relation matrix, drop unnecessary columns, creating new attribute using the Kaggle dataset. Following that, we applied a variety of classifiers, including Decision Tree (DT), Gradient Boosting Classifier (GB), AdaBoost Classifier, Bagging Classifier and Logistic Regression. After that we split the dataset into two parts such as train and test. Then, we employed "Feature selection technique" to improve model accuracy while decreasing model complexity. Finally, we use SMOTE to balance the ratio of target class. Among the models, Bagging Classifier with its univariate feature selection method and high recall and F1 score achieved the highest accuracy with 88.00%.

6.2 Conclusions

Various machine learning models are utilized to perform loan prediction. A loan prediction data set is utilized for the evaluation of the study. Several data cleansing operations, including removing null columns, filling null values using the mean mode method, and converting categorical to numeric values. For optimal feature selection, two distinct feature selection strategies are employed. In conventional machine learning models, both features from different feature selections are trained and tested in a distinct manner. Among the models, Bagging Classifier with its univariate feature selection method and high recall and F1 score achieved the highest accuracy with 88.00%. Consequently, the results indicate that Bagging Classifier may be suited to the task of loan default prediction. In the future, it will be our responsibility to increase the volume of data in the dataset, and we will use data preprocessing and omission to improve the accuracy of this work.

6.3 Implication for Further Study

Our aim will predict loan eligibility of applicants in banking sector using more deciding parameters like- credit score, life style, career. Using deciding parameters precisely predict the right candidate within a shorter time. If we can able to do that it will helpful for our society to identify the eligible loan applicant to getting loan from banks.

Reference:

[1] Arutjothi, G. and Senthamarai, C., 2017, December. Prediction of loan status in commercial bank using machine learning classifier. In 2017 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 416-419). IEEE.

[2] Abdelmoula, A.K., 2015. Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. Accounting and Management Information Systems, 14(1), p.79.

[3] Devi, C.D. and Chezian, R.M., 2016, October. A relative evaluation of the performance of ensemble learning in credit scoring. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA) (pp. 161-165). IEEE.

[4] Goyal, A. and Kaur, R., 2016. A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA), 3(1), pp.32-37.

[5] Kadam, A., Nikam, S., Aher, A., Shelke, G. and Chandgude, A., 2021. Prediction for Loan Approval using Machine Learning Algorithm. International Research Journal of Engineering and Technology, 8(04), pp.4089-4092.

[6] Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), pp.18-21.

[7] Supriya, P., Pavani, M., Saisushma, N., Kumari, N.V. and Vikas, K., 2019. Loan prediction by using machine learning models. International Journal of Engineering and Techniques, 5(22), pp.144-148.

[8] Gautam, K., Singh, A.P., Tyagi, K. and Kumar, M.S., 2020. Loan Prediction using Decision Tree and Random Forest. International Research Journal of Engineering and Technology (IRJET), 7(08).

[9] Gupta, A., Pant, V., Kumar, S. and Bansal, P.K., 2020, December. Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[10] Bhattad, S., Bawane, S., Agrawal, S., Ramteke, U. and Ambhore, P.B., 2021. Loan Prediction using Machine Learning Algorithms. International Journal of Computer Science Trends and Technology, 9(3), pp.143-146.

[11] Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G. and Düştegör, D., 2021. Employability prediction: a survey of current approaches, research challenges and applications. Journal of Ambient Intelligence and Humanized Computing, pp.1-17.

[12] Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P., 2021. Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012042). IOP Publishing.

[13] Sheikh, M.A., Goel, A.K. and Kumar, T., 2020, July. An approach for prediction of loan approval using machine learning algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 490-494). IEEE.

[14] Turkson, R.E., Baagyere, E.Y. and Wenya, G.E., 2016, September. A machine learning approach for predicting bank credit worthiness. In 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR) (pp. 1-7). IEEE

[15] Shinde, A., Patil, Y., Kotian, I., Shinde, A. and Gulwani, R., 2022. Loan Prediction System Using Machine Learning. In ITM Web of Conferences (Vol. 44, p. 03019). EDP Sciences.

[16] Priya, K.U., Pushpa, S., Kalaivani, K. and Sartiha, A., 2018. Exploratory analysis on prediction of loan privilege for customers using random forest. International Journal of Engineering Technology, 7(2.21), pp.339-341.

[17] Sujatha, C.N., Gudipalli, A., Pushyami, B., Karthik, N. and Sanjana, B.N., 2021, November. Loan Prediction Using Machine Learning and Its Deployement On Web Application. In 2021 Innovations in Power and Advanced Computing Technologies (i-PACT) (pp. 1-7). IEEE.

[18] Foster, B.P. and Zurada, J., 2013. Loan defaults and hazard models for bankruptcy prediction. Managerial Auditing Journal.

[19] Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A., 2013. Consumer credit risk: Individual probability estimates using machine learning. Expert systems with applications, 40(13), pp.5125-5131.

[20] Zhang, T. and Li, B., 2018, May. Loan Prediction Model Based on AdaBoost and PSO-SVM. In 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018) (pp. 733-739). Atlantis Press.

[21] Łuczak, A., Ganzha, M. and Paprzycki, M., 2021. Probability of Loan Default—Applying Data Analytics to Financial Credit Risk Prediction. In Intelligent Systems, Technologies and Applications (pp. 1-16). Springer, Singapore.

[22] Jain, A. and Shinde, S., 2019, March. A Comprehensive Study of Data Mining-based Financial Fraud Detection Research. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.

[23] Zelaya, C.V.G., 2019, April. Towards explaining the effects of data preprocessing on machine learning. In 2019 IEEE 35th international conference on data engineering (ICDE) (pp. 2086-2090). IEEE.

[24] Raju, V.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A. and Padma, V., 2020, August. Study the influence of normalization/transformation process on the accuracy of supervised classification. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 729-735). IEEE.

[25] Guru, D.S., Suhil, M., Raju, L.N. and Kumar, N.V., 2018. An alternative framework for univariate filter based feature selection for text categorization. Pattern Recognition Letters, 103, pp.23-31.

[26] Sharaff, A. and Gupta, H., 2019. Extra-tree classifier with metaheuristics approach for email classification. In Advances in computer communication and computational sciences (pp. 189-197). Springer, Singapore.

[27] Chawla, N.V., Bowyer, K.W. and Lawrence, O., Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique, pp.321-357.

[28] Chen, R.C., Dewi, C., Huang, S.W. and Caraka, R.E., 2020. Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), pp.1-26.

[29] Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), pp.128-138.

[30] Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E., 2006. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), pp.159-190.

[31] Pandey, P. and Prabhakar, R., 2016, August. An analysis of machine learning techniques (J48 & AdaBoost)-for classification. In 2016 1st India International Conference on Information Processing (IICIP) (pp. 1-6). IEEE.

Final Plagiarism

ORIGINALITY REPORT

26%	18%	10%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to University of Sunderland Student Paper	1%
4	repository.seeu.edu.mk Internet Source	1%
5	Submitted to University of Westminster Student Paper	1%
6	www.coursehero.com Internet Source	1%
7	fjs.fudutsinma.edu.ng Internet Source	1%
8	Submitted to National College of Ireland Student Paper	1%
9	Submitted to Liverpool John Moores University Student Paper	<1%