Credit Risk Prediction by using Ensemble Machine Learning Algorithms

Er. Sarita Chhetri[a], Ramesh Parajuli[b], Assoc. Prof. Dr. Gajendra Sharma[c]

[a] *sarita13mca21@kcc.edu.np*

Kantipur City College, Kathmandu, 44600 , Nepal

Abstract

In the financial sector, credit lines are the main source of income. The main income source is investing in the amount and collecting interest using the principal amount. However, they are not able to collect all investments because defaulters are not ready to pay the amount. Loan prediction played a vital role in this scenario. By predicting loan defaulters, the institution can reduce the number of faulty account holders.
The most popular method is the scientific method, where Machine Learning is used. Machine Learning has statistical models that can perform specific task to predict the credit data of upcoming future. Prediction can be performed using supervised learning techniques such as decision tree, random forests, Gaussian Naïve Bayes, AdaBoost, Support vector machines, and logistic regression. This study aims to build credit scoring by adopting five of them for further study. The primary data are collected from a cooperative financial institution where 13600 data points on credit was collected, 70% of the data was trained, and 30% was separated for testing for test 1. Additionally, same data set is trained and test in 8:2 ratio for test 2.

Keywords: Financial sector; credit risk prediction; machine learning

1. INTRODUCTION

A) BACKGROUND

Credit lending is a crucial task for income generation in financial institutions. The bank is collecting the money in different headings and investing it as a loan to the clients. Clients are collecting it for different purposes like business, properties like land, gold, hire purpose, education, travels, treatments and for marriage and ceremonies. While clients are asking for money, they need to pay certain additional amount to bank which is known as interest. The interest is the monetary income for financial institutions. For generating the more interest, a huge amount of deposits is invested in the market as loans. The investment which is done on market is not returning on schedule period and some of them are fell in the category of non- performing loan. Before supplying the loan, the authorized personalities of financial institutions should do prediction whether the money will return to organization or not.

Financial institutions use traditional methods to decide whether to lend credit to their clients. Traditional way means executives and tactical level managers are doing manual prediction for bank loan. Traditionally, they provide insights into clients' habits, deposits, applicant income, co-applicant income, loan-duration, dependency, credit-history, marital status, qualification, gender, age and property to make decisions on credit transfer. There are various loop holes in the traditional method of loan distribution. Most of the seven characteristics–saving account, occupation, work duration, home, ownership, annual income, and income ratio– influence loan default [1]. The nature, perception and nurture of managers are playing a vital role for manual way of prediction. It means that the loan request accept or rejection may differ person wise. However, for loan supply there is the team of loan committee to do the decision. Credit ratings can be predicted scientifically as well as by using the machine learning algorithms. Machine learning can be used in loan prediction in less presence of human interruption. Machine learning can be used to predict exact information from a previous data set [2]. In the machine learning algorithm, advanced algorithms and statistical models can be used. The future credit risk prediction can be done on historical data set. In past days, what types of cases are being defaulter and what are the issues behind loan defaulters can be used to set the models. The previous data can be used to trained and test the model. The varieties of models are available in machine learning for credit risk

prediction. The crucial part for it is the trained and test the model in huge number of data set in the range of thousands to millions. Machine learning algorithms like supervised learning, unsupervised learning, reinforcement learning as well as ensemble learning can be used for credit risk analysis. However, in institutional loan prediction, machine learning can be adopted using logical regression as well [3].

B) STATEMENT OF PROBLEM

Financial institutions, such as cooperatives, finances, and banks, deal with various types of loans, such as housing credit, personal credit, and business loans. As the banking sector is growing daily, a large number of clients are applying for loans. In today's era, financial institutions providing loans have become a general phenomenon. At the same time, the rise in loan applications and consumption has resulted in worse credit losses.

The major problem in the financial sector is that it faces an increasing rate of loan defaults, and executive and tactical managers are experiencing difficulty in accessing the correct loan request. It is stiff to analyze how risky the borrower is and should the loan be supplied.

C) RESEARCH OBJECTIVES

The objectives of the proposed paper are as follows: -

- To perform comparative analysis of ensemble Machine learning among Logistic Regression, Random Forest, Naïve Bayes, Ada Boost and XG Boost.
- To find out the significant parameters for bank loan prediction.
- To design and implement hybrid model for loan prediction.

D) RESEARCH QUESTIONS

The study aims to answer the following research question:

- RQ1: What is the best performing Ensemble algorithm for loan prediction in terms of accuracy, precision, recall and F1 score among different machine learning algorithms: Logistic Regression, Random Forest, Naïve Bayes, Ada Boost and XG Boost?
- RQ2: What are the significant parameters for bank loan prediction?
- RQ3: Which combination of machine learning can be used to design hybrid classifier to optimize the performance and accuracy?

E) RATIONAL OF THE STUDY

As in the financial sector, the main source of income is credit measure out, and the huge risk of this process is that loans will be converted into non-performing loan. Varieties of traditional loan prediction ways are used by tactical level and executive level managers which are not scientific in nature.
Scientific methods, such as machine learning algorithms, help to perform loan prediction in a better way. It helps to overcome issues such as the complexity of data, automated decision making with high accuracy, and risk minimization. In addition, it discards the concepts of bias and personalization.

2. LITERATURE REVIEW

In the financial sector, institutions have several products and activities, but core earnings concepts are centralized on credit distribution. The profit of financial institutions depends on loan recovery. Loan recovery is a crucial task, as many loans fall into the category of non-performing loan. A very important approach is the correct prediction of loans using appropriate approach of predictive analytics. Increasing loan applications and defaulting applicants creates credit losses. Credit loans are issued by people for several purposes, such as education, medicine, travel, business, and hire purchases. Financial institutions must use the blueprint of effectual models to know about facts and figures about clients' habits, monetary usage patterns, and other relevant characteristics. According to this study, there are seven characteristics: occupation, work duration, home ownership, annual income of clients, oldness, and age. In addition, location, debt-income ratio, loan term, and house ownership also play a vital role.

A) THEORETICAL FOUNDATION

It was found that the banks relied on credit scores earned by loan investment. All the loan investments are not returned back to the bank as credit pay by clients. The main focus of banking institutions centralized on client's income, assets, credit history, educational qualification, employment types and loan term [2]. However, these parameters may not minimize the defaulters' number in banks. In some context, bankers fail to judge whether to invest loan or not. The bad investment creates the non- performing loan. The non-performing loan refers to a specific amount of credit taken by a borrower but the debtor has declined in making agreed instalment paybacks in 90 days for commercial banking loans and 180 days for consumer loans [3]. In these scenarios, machine learning algorithms may play a vital role. Machine learning is the rapidly evolving branch of artificial intelligence which is widely used in modern computer-based technology [4].

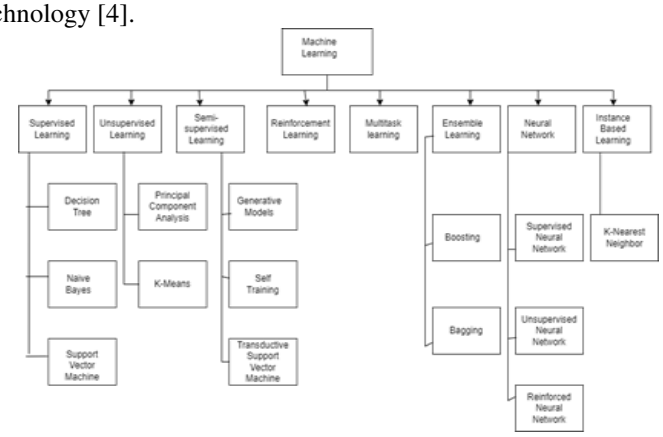


Fig. 1. Categories of Machine Learning Algorithms [4]

B) TRADITIONAL WAY OF BANK LOAN PREDICTION

Loans are a crucial source of income for the financial sectors. However, investing loans on clients have more financial risks as many loans are going to be non- performing loans. Everyday a large number of people are applying for loan for various purposes and some of them are being approved on basis of certain criteria [2]. The most of the financial organizations are doing the process of loan approval manually, which is the slow process [32]. Banking authorities are trying to mitigate the risk [15]. However, conventional selection processes often struggle to identify the most suitable candidates from a pool of loan applicants. In response to this challenge, machine learning algorithm are best options [18].

C) BANK LOAN PREDICTION BY USING MACHINE LEARNING ALGORITHMS

There have been several traditional methods for predicting loan information for decision-making, and scientific methods using machine learning algorithms are also available [5]. The machine learning is a fleetly growing field that enables overkill of innovative approaches for solving real-world problems. It endorses machines to learn without human involvement from data and is used in a variety of applications [6]. The machine learning algorithm is defined as the scientific study of algorithms and statistical models in which computer systems are used. In machine learning, once an algorithm learns the pattern and strategy of data, it can further work automatically by learning and doing. Machine learning (ML) is used to teach machines how to handle data effectively and efficiently. Machine learning helps to deal with big data, as data are increasing daily. Machine learning algorithms can be further categorized using different algorithms.

D) COMPARATIVE ANALYSIS OF ENSEMBLE ALGORITHM FOR BANK LOAN PREDICTION

By using the Jupiter notebook, platform for python, the algorithms performance was evaluated in terms of accuracy, precision, recall and F1-score. This study highlighted that the several supervised learning can be formed as hybrid to show the better performance and that is considered as ensemble learning.

The performance of loan prediction model differs on the basis of parameters what they used and performance metrics what they preferred. In this paper, the parameters age, purpose, credit history, credit duration, sex, Number of dependents, qualification, annual income were taken and the logistic regression was taken as an algorithm where the accuracy was 0.811 [27].

E) MACHINE LEARNING FOR BANK LOAN PREDICTION: REVIEW, APPROACHES AND OPEN RESEARCH PROBLEMS

It summarizes the previous works as paper with author reference, data set description, parameters, algorithms, performance metrics and merits and demerits of the papers. This paper focused on the finding of best machine learning algorithm among selected ensemble learning technique. The previous summarized papers in table 1 helps to find out the best ensemble learning for research work for this paper. The evaluation of these papers helps to find out the previous dataset used by the papers, parameters for evaluations and algorithms for comparisons.

Table 1: Summary of ML-Based Bank Loan Prediction Paper

S. No	Author Refere n ce	Datase t Descri p tion	Parameters	Compa red algorit hm	Perfor mance Metrics	Merits of Paper	Limitations
1	M. A Sheikh, A. K Goel, and T. Kumar [27]	Dataset collecte d from Kaggle	Age, purpose, credit history, credit duration, sex, No of dependents, qualification , annual income	Logisti c regress ion	Sensitiv ity, Specifi city	The paper clearly talks about model evaluat ive.	Gender and marital status seem not to taken in consider ation.
2	M. Anand,		Age, education,	Logisti c Regres sion,	Confusi on matrix, accurac	Predict ive Modeli ng	

	A. Velu, and P. Whig [35]		employment status, year of experience, address, income, debt income, credit to debit ratio, Other debt	Decisi on Tree, KNN, SVM, RF	y, recall, F1- score analysis		
3	A. Gupta, V. Pant, S. Kumar, and P. K. Bansal [22]		Gender, married, dependents, education, self employeed, applicant income, coapplicant income, loan amount, loan amount term, credit history, property area, loan status	Logisti c Regres sion, Rando m Forest			
4	C.K. Gomat hy [7]		Loan_id, gender, Married,dep endents, education, applicant income, coapplicant income, loan amount, loan amount term, credit history area, loan status	Decisi on Tree		COB techniq ue used. It can be fixed with automated prophecy system	It is compro mised with noise and outlier data of classification
5	V. Singh, A. Yadav, R.		Income, marital status, loan amount, loan duration.	Decisi on Tree, Rando m Forest, XG Boost		Eligibi lity criteria are set to predict loan pay.	This paper is unanswe rable when the client face some disaster conditio ns.

	Awasthi, and G. N. Partheeban [33]						
6	K. Gautam, A.P. Singh, K. Tyagi, and S. Kumar [36]	The dataset is collected from the banking sector. AREF format	Loan_id, gender, marital status, dependents, education, self employeed, applicant income, co_applicant income, loan amount, loan amount term, credit history, property area, loan status	Decision Tree, Random Forest	accuracy	Working with different confidence to increase accuracy.	For lower confidence factor, more pruning is done.
7	A. Goyal, R. Kaur [17]					Ensemble model enhance the accuracy.	The modal is not suitable for less number of data.
8	S. Sreesouthry, A. Ayubkhan, M. M. Rizwan, D. Lokes	Dataset is collected from Kaggle.	Loan amount, marital status, gender, dependent, graduation, selfemployed, income, co_applicant income, loan term, credit history, location	Logistic Regression	Accuracy	Gender and marital status doesnot make sense.	The paper does not explain about heat map.

	h, and K. P. Raj [28]						
9	A.S. Aphale and D. S. R. Shinde [8]	The dataset is collected from cooperative bank.			Accuracy, precision, recall, specificity, F1-score.	The model helps to formulate the bank risk automated system.	
10	M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath [10]	The dataset is publicly available Lending Club dataset from Kaggle.	Loan amount, term, interest rate, installment, grade, sub grade, home ownership, purpose, loan status, zip code, revol_balance	Decision tree, Random Forest	Accuracy	The model can help to identified the default er type scenario for loan investment.	The algorithm puts some of the nondefaulters in the defaulter class.
11	V. Moscato, A. Picariello, and G. Sperlí [37]			Random Forest, Logistic Regression, Multilayer perceptron	Sensitivity, specificity, precision, FP-value, GMean	It is able to manage unbalanced data set.	It devotes to improve the evaluation in upcoming works.
12	I.O. Eweoya, A. A. Adebiyi, A. A.		Age, sex, income, employment status, the track of the last three payments, balance of	Decision tree,	Accuracy, confusion matrix, ROC curve	Casebased, analogy- based reasoning and statistical approaches have been employed	-Fraudulent attempts cannot be discovered by these approac

	Azeta, and A. E. Azeta [3]		loan taken				hes. -Dat set descripti on is not mention ed.
13	P. Chotwa ni, A. Tiwari, and M. Hooda [21]	The data set is collecte d from bankin g sector which is in ARFF format.	Loan id, gender, marital status, dependents, education, self employed, applicant income, co_applicant income, loan amount, loan amount term, credit history, property area, loan status.	Logisti c regress ion, Rando m Forest, Decisi on Tree	Accura cy	It can be plugge d with other system s as well.	It has some cases of compute r glitches, errors in content and weight of features is fixed.
14	Lili Lai [12]	The data set is of actual busines s occasio n in Xiame n Internat ional Bank.	Id, target, credit id number, gender, age, region, education, job, ethic, credit start date, credit valid date	RF, Ada Boost, XG Boost, KNN, MLP	AUC, Accura cy	Ada Boost is showin g 100% accura cy.	RF, KNN and MLP are overfit the training data and AUC results are weaker for them.
15	J. Tejasw i		Loan id, gender, married, dependents,	Decisi on tree, LR,	Precisio n, recall,	It can be plugge d with other system s as well.	It has some cases of compute r glitches, errors

	ni, T.M. Kavya, R.D. Naga Ramya, P. S. Triveni, V.R Madddu mala [9].		education, self_employed, applicant income, co_applicant income, loan amount, loan amount term, credit history, property area, loan status.	Random Forest,			in content and weight of features is fixed.
16	D. Dansana, S. G. K. Patro, B. K Mishra, V. Prasad, A. Razak, and A. W. Wodajo [2]		Gender, education, employment type, business type, loan term, marital status	Random Forest	No. of customer in different category	The marital status parame ter is highly monito red.	This paper is not able to work on deep learning. The data set is not much larger.

3. RESEARCH METHODOLOGY

A) Research Framework

Fig. 2. Theoretical Framework

The theoretical framework is designed as shown in figure 2 which is described below.

i) Raw Data Collection:

Firstly, the raw data was collected from banking institutions. The total number of data is 13600. The data has features like loan amount, applicant income, co-applicant income, dependency, credit history, marital status, age, interest rate, duration and remarks.

ii) Data Cleansing:

The next step taken for data was cleansing. In this step, the missing data were handling with deletion method. Deletion method is the way of deleting rows and columns where it has not any data or having null value. Out of the total number of 13600 data, the 145 data were removed as they have null cells. The remaining 13455 were taken for further processing. After that, the second step of cleansing is finding outliers. In this paper, the outliers are detected by using visualization box plot method in python. In some cases of feature selection log transformation has been done to reduce the impact of outlier. The data formatting has not been done as there is not presence of any duplicate data.

iii) Feature Engineering:

The second step followed by cleansing is feature engineering. In this step, firstly, categorical variables are handled in numeric form according to the requirement. Among above mentioned features, some features like marital status and remarks are converted into numerical label. The feature marital status is converted into numeric form like 0,1 and 2 for married, unmarried and divorced by using label encoding method.

Likewise, the feature remark is converted into binary values 0 and 1 by using one-hot encoding method. In this method, paid accounts are categorized as 1 and default accounts are categorized as 0.

The next step in feature engineering is creating interaction terms. For this, the new column total income is generated by summing up the applicant income and co-applicant income.

iv) Data Splitting:

The next step after feature engineering is data splitting. The total data set is separated into training data set and testing data set in the ratio of 7:3 for test 1 and the 8:2 ratio is selected for test 2.

v) Choose a model and train it:

The training data set is used for choosing a model. In this paper, the Logistic Regression, Random Forest, AdaBoost, XGBoost and Guassian Naïve Bayes have been chosen as model for analysis. After that, model has been trained and their performance has been measured on the basis of accuracy, F1-score, precision value, recall and confusion matrix.

vi) Evaluate on test data:

The all selected model will be evaluated on test data.

vii) Interpret Results and Deploy the model:

After getting the test results, results will be interpreted. As the testing data will be succeed the model will deploy.

a. Dataset Description

For this dissertation, 13600 datasets were collected from co-operative banking institution which is the provided data set from institution of their customers. It consists of 14 features including dependent variable remarks.

The figure 2 represents the working of model of this dissertation. It provides a rough idea how the loan prediction system works.

b. Data Cleaning

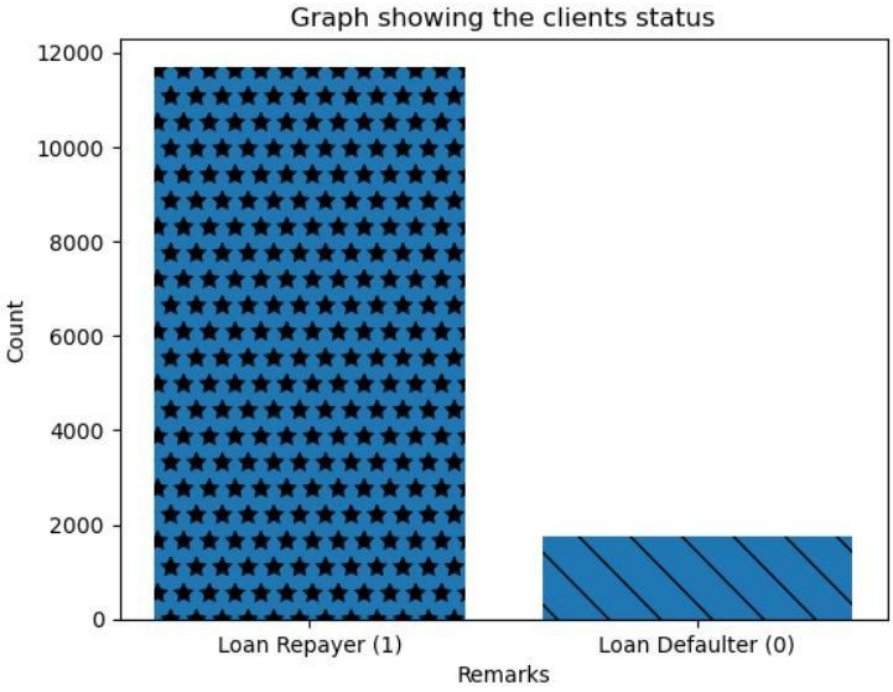
After collecting the data, it is most important to clean it as it may contain null values and empty cells. The dataset may contain unnecessary feature cell as well. Therefore, the data cleaning is most required phase before further processing to handle missing values.

The dataset is in CSV (Comma-Separated value) format that is accepted by jupiter notebook. There are 14 attributes using for this paper. The depth description of data is shown in table 1[36][38]. Secondly, out of many features, features like account number, loan amount, applicant income, credit history, marital status, co-applicant income, qualification, dependency, gender, age, interest rate, duration was selected as feature. After that categorical values are converted into numerical values.

Table 2: Data set variables along with description and type

Variable Name	Description	Type
Accountno	Unique ID	Integer

Loanamount	Loan amount in thousands	Integer
Applicantincome	Applicant income	Integer
Credithistory	Credit history meets the guidelines	Integer
maritalstatus	Applicant married(M/U/D)	Character
Coapplicantincome	Co-applicant income	Integer
Qualification	Qualification of applicant as literate or not? (Y/N)	Character
dependency	Number of Dependents	Integer
Gender	Male/Female	Character
Age	Applicant age	Integer



mo
in

fig

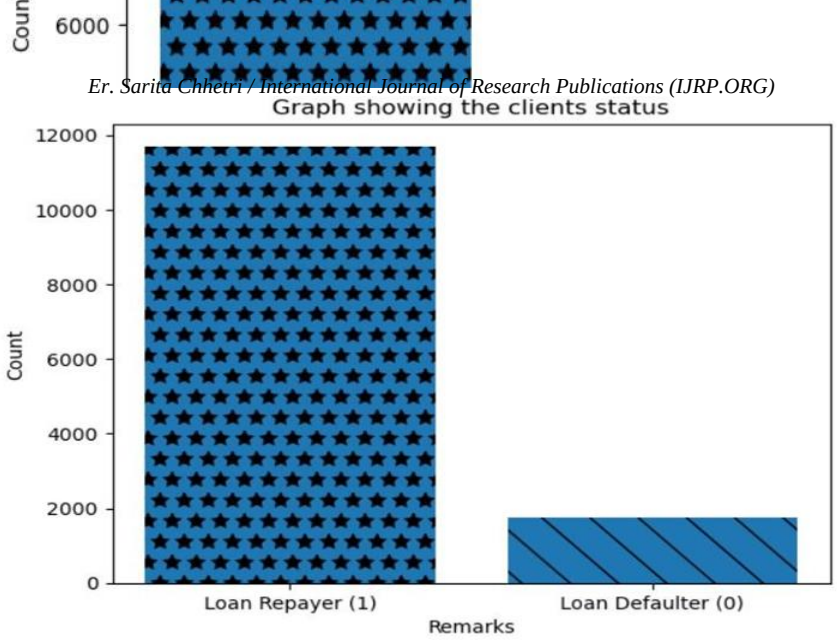


Fig -3: Graph showing Classification of 0 and 1

3.5 Modeling

The whole data set was split into two subsets known as training data set and testing data set. 70% of the whole data set is used to train the machine learning model and known as training data set. The remaining 30% data is used for testing set to evaluate its performance and called testing data set. Now in machine learning model, we first apply the training data set, in this data set the model is trained with known examples. The entries of new applicants will act a test data which are to be filled at the moment of submitting the application model. After performing such tests, model can be evaluated whether the credit approved to the person is safe or not basically about the credit approval on the basis of various training data sets. The chronology of the data is showing in figure 4[21].

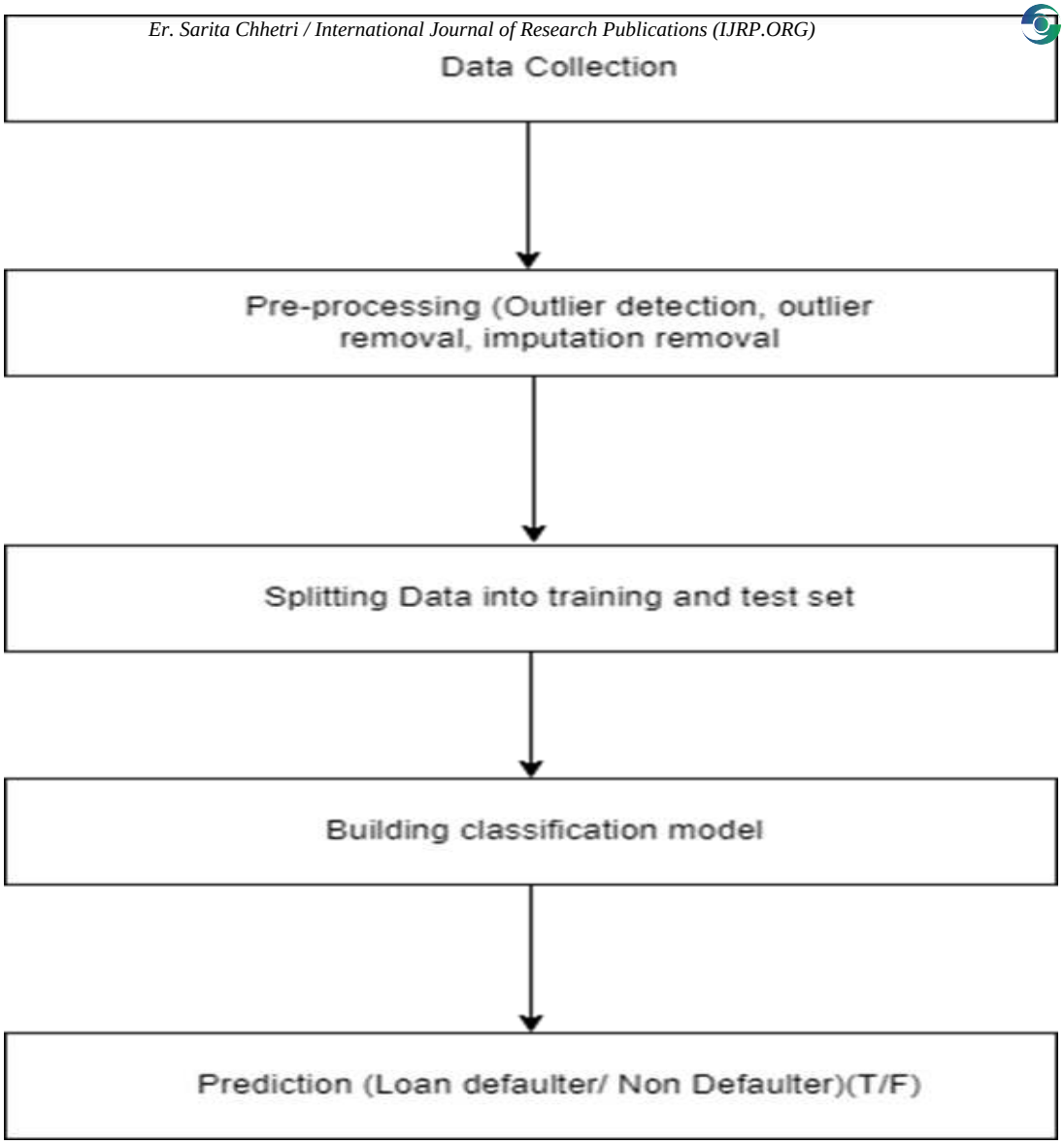


Fig 4: Chronology of Data [21]

3.5.1 Dataset analysis

For exhibiting this research work, a Jupiter notebook is used along with a laptop 11th Gen Intel® Core™ i7-1165G7 @2.80GHz processor, 8GB RAM with 64-bit operating system. To analyze the performance of collected data and train the model, python is used where Sci-kit-learn open source machine learning library is used. Python with version 3.11.4 which is packaged by Anaconda is used for the prediction purpose. Other libraries like Numerical python, matplot library and seaborn were used. Numerical python or Numpy is a fundamental cornerstone library of python which is used for scientific computation and data analysis. Numpy provides multidimensional arrays and matrices. Numpy is required to run other libraries like Matplotlib, pandas, XGBoost and scipy. Matplot library is a widely used library in python for creating static, animated and interactive visualizations and plots. Matplot lib helps to get high quality graphs, charts and other

representations of data. Along with matplotlib, seaborn library is used for visualizing statistical relationships and exploring datasets. All the libraries can be call by using the import keyword.

4. RESULT ANALYSIS AND COMPARISON

A) Result of the Selected Model

Machine Learning is the crucial for bank loan prediction rather than manual prediction way. Five machine learning models are used with different features like loan amount, total income, credit history, dependency, age, interest rate, duration and marital status. Using a dataset of 13600 was collected and 145 null data set were removed. This dissertation was working with 13455 from banking institutions as a primary data. In this paper, five most suited algorithms namely random forest, ada boost, logistic regression, XG Boost and Naïve Bayes were used for classification learning. These algorithms were imported from Scikit learn library and used for classification. They were implemented to perform the comparative analysis of their performance considering for different evaluation matrix like accuracy, precision, f1-score and recall.

The total data set 13455 was train with two trained and test ratios. Firstly, the collected data set was trained with 70% data and test with 30% dataset. The collected result was summarized in table 3.

Table 3: Evaluation metrics on 7:3 trained and test dataset

Algorithm	**0/1**	**Random Forest**	**Logistic Regression**	**AdaBoost**	**XG Boost**	**Naïve Bayes**
Precision	0	0.17	0	0.47	-	0.29
	1	0.87	0.87	0.87	0.89	0.87
Recall	0	0.09	0	0.02	-	0
	1	0.94	1	1	0.96	1
F1-Score	0	0.11	0	0.03	-	0.01
	1	0.90	0.93	0.93	0.931	0.93
Accuracy		82.83%	87.19%	87%	87.61%	87.11%

From table 3, it can be clearly observed that the XG Boost is showing better performance as compared to the other machine learning algorithm.
The total data set 13455 was train with two trained and test ratios. Firstly, the collected data set was trained with 80% data and test with 20% dataset. The collected result was summarized in table 4.

Table -4: Evaluation metrics on 8:2 trained and test dataset

Algorithm	**0/1**	**Random Forest**	**Logistic Regression**	**AdaBoost**	**XG Boost**	**Naïve Bayes**
Precision	0	0.21	0	0.61	-	0.33
	1	0.88	0.87	0.88	0.90	0.87
Recall	0	0.11	0	0.03	-	0
	1	0.94	1	1	0.97	1
F1-Score	0	0.14	0	0.06	-	0.01
	1	0.91	0.93	0.93	0.935	0.93
Accuracy		83.65%	87.73%	87.51%	88.37%	87.32%

From table 4, it can be clearly observed that the XG Boost is showing better performance of accuracy with 88.37% as compared to the other machine learning algorithms. The logistic regression is getting better performance after XG Boost that is 87.73%.

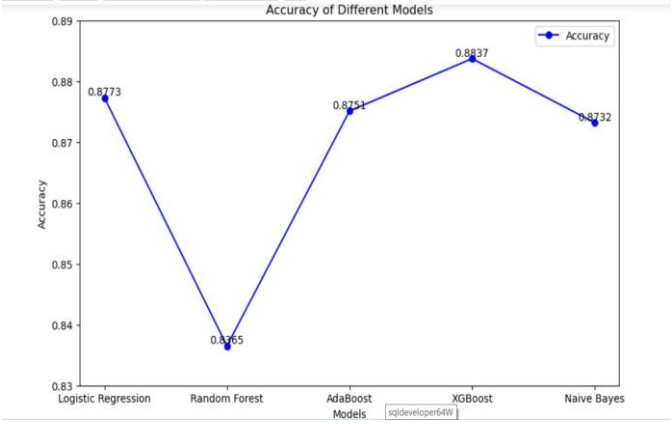


Fig 5: Comparative Analysis on the basis of Accuracy

Figure 5 represents the comparative analysis of accuracy of selected model namely logistic regression, random forest, adaboost, xgboost and naïve bayes. The comparative line graph shows that the accuracy of xg boost is higher with 88.37% accuracy whereas the lower accuracy is 83.65% is of random forest regression. This plot is taken for the dataset of total data 13455 and trained in ratio of 8:2.

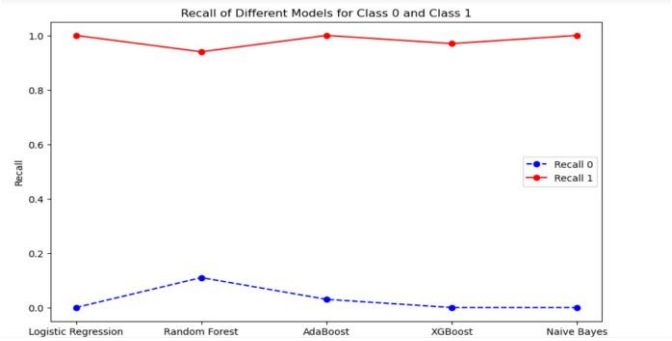


Fig 6: Comparative Analysis on the basis of Recall

Figure 6 represents the comparative analysis of recall value of selected model namely logistic regression, random forest, Adaboost, xgboost and naïve bayes. The comparative line graph shows that the recall value of logistic regression, adaboost and naïve bayes is higher with 1 whereas the 0-precision value is higher in random forest that is 0.11.

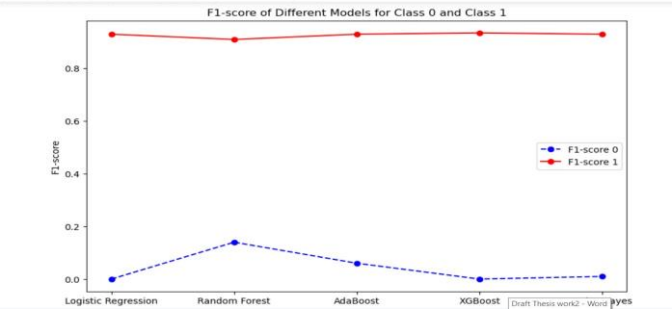


Fig 7: Comparative Analysis on the basis of F1-Score

Figure 7 clearly illustrates that the Random Forest shows the better value with 0.14 in class 0. And XG Boost shows the better performance with value 0.935 inclass 1. The comparative value shows the quite progressive result than others models.

B) Result of the Proposed Hybrid model

The hybrid model is generated by adopting the best two models based on the performance metrics. Among all selected models like logistic regression, random forest, xg boost, ada boost and naïve bayes, xg boost and logistic regression is showing best performance in most of the performance metrics like precision, recall, f1-score and accuracy. Therefore, the xg boost and logistic regression is selected to build hybrid model. While modeling the hybrid, the hyper parameter tuning is done to show the best performance.

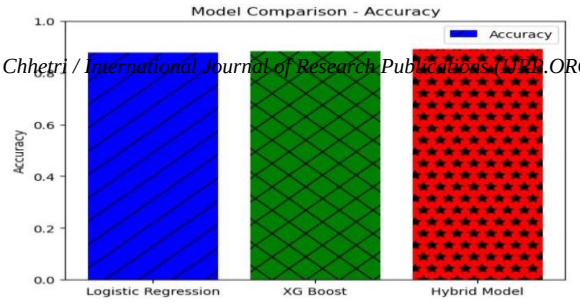


Fig 8: Comparative Analysis of Accuracy of Purposed Hybrid model along with Logistic regression and XG Boost

Table -5: Comparative metrics for proposed hybrid model

Performance Metrics	**0/1**	**Hybrid Model**
Precision	0	0.65
	1	0.90
Recall	0	0.06
	1	1
F1-Score	0	0.13
	1	0.94
Accuracy		88.79%

In the proposed hybrid model, the model is created by using the hybrid of logistic regression and XG Boost. First of all, the two models are combined and hyper parameter tuning is done on parameters. As in previous selected ensemble models, the measurement metrics are accuracy, recall, precision and F1-Score. The accuracy of purposed model is 88.79% whereas the precision value at class 0 is 65% and in class 1 is 90%. Likewise, the recall value in class 0 is 6% and in class 1 is 100%. In case of F1-score in class 0 the value is 13% and in class 1, the value is 94%.

The measure metrics are plot in bars in figure 9 below.

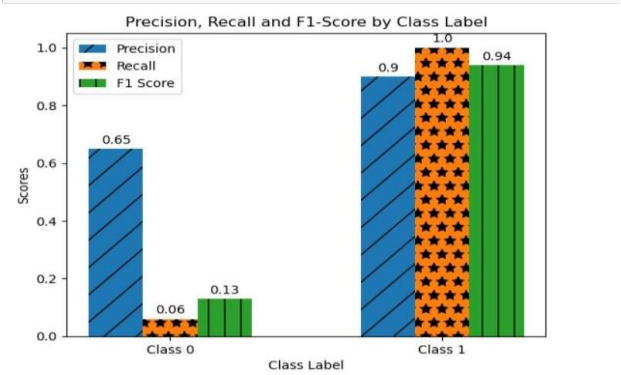


Fig 9: Comparative Analysis on the basis of performance metrics

C) Comparative analysis of Proposed Hybrid model along with Logistic Regression and XG Boost

Figure 10 is the graphical representation of comparative analysis of accuracy of purposed hybrid model along with logistic regression and XG Boost. From the diagram, it can be clearly seen that the accuracy of hybrid model is greater than logistic regression and XG Boost. The accuracy of hybrid model is 89% whereas the accuracy of LR and XG Boost is 87.73% and 88.79% respectively.

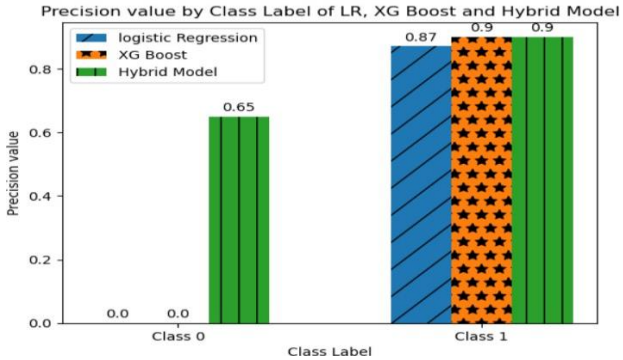


Fig 10: Comparative Analysis of Precision value of Purposed Hybrid model along with Logistic regression and XG Boost

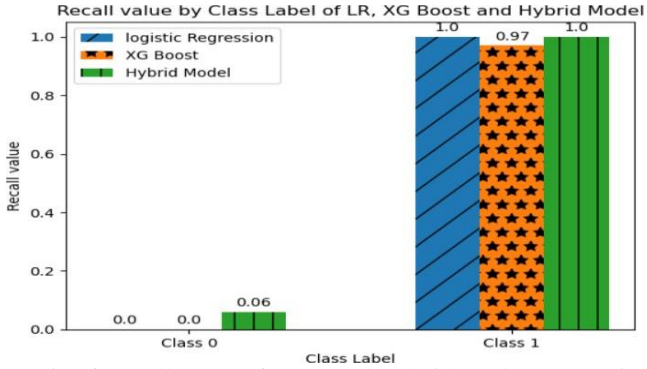


Fig 11: Comparative Analysis of Recall value of Purposed Hybrid model along with Logistic regression and XG Boost

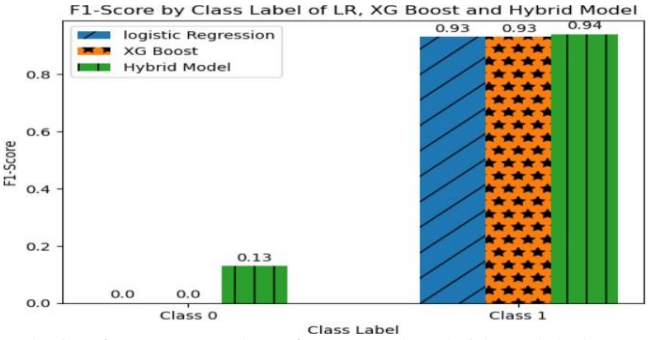


Fig 12: Comparative Analysis of F1-score value of Purposed Hybrid model along with Logistic regression and XG Boost

5. CONTRIBUTIONS

- This paper introduces a novel credit scoring model with comparing the five ensemble learning models with their performance accuracy, f1score, precision, confusion matrix and recall.
- This model plays a crucial role in resource optimization as it replaces manual work.
- As the data set is collected as primary data, it can be further use as secondary data for further research work by anyone.

6. LIMITATIONS

- Less number of datasets.
- Working with only supervised algorithms.

7. CONCLUSIONS

This study may have demonstrated the viability of machine learning algorithm for bank loan prediction. In addition, this study also tries to highlight the traditional way of bank loan prediction. In machine learning, this study focused on five ensemble learning namely random forest, logistic regression, ada boost, XG Boost and Naïve Bayes. The best two ensemble learning model is chosen based on accuracy, precision, recall and F1-score that is logistic regression and XG Boost. The XG Boost is showing high accuracy of 88.37% followed by LR with 87.73%. From these two ensemble models, proposed hybrid model is generated with accuracy of 88.79%. This proposed hybrid model can be fruitful in banking sector for loan prediction. This paper highlights the future enhancement and contributions in the field of banking sector.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude the most enlighten personality for this paper, my supervisor, Asst. Prof. Ramesh Parajuli who has consistently encouraged, inspired and provided wealthy knowledge for fulfilling this task. I am pleased to express my gratitude to respected personality Assoc. Prof. Dr. Gajendra Sharma who assisted me and provided the fruitful guidelines for this paper. And, the final sincerity goes to, my respected husband and lovable son who always done the immense of support and never-ending encouragement during my study period and research time.

REFERENCES

[1] M. Jul and L. Nrb, “List of Banks and Financial Institutions,” vol. 2020, pp. 1–3, 2020.
[2] D. Dansana, S. G. K. Patro, B. K. Mishra, V. Prasad, A. Razak, and A. W. Wodajo, “Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm,” Eng. Reports, no. May, pp. 1–17, 2023, doi: 10.1002/eng2.12707.
[3] I. O. Eweoya, A. A. Adebiyi, A. A. Azeta, and A. E. Azeta, “Fraud prediction in bank loan

administration using decision tree," J. Phys. Conf. Ser., vol. 1299, no. 1, 2019, doi: 10.1088/1742-6596/1299/1/012037.
[4] H. Bhatt, V. Shah, K. Shah, R. Shah, and M. Shah, "State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review," Intell. Med., vol. 3, no. 3, pp. 180–190, 2023, doi: 10.1016/j.imed.2022.08.004.
[5] B. Mahesh, "Machine Learning Algorithms-A Review," Int. J. Sci. Res., 2018, doi: 10.21275/ART20203995.
[6] S. Dridi, V. Machine, D. Tree, R. Forest, and L. Regression, "S l - a s l r," 2021.
[7] C. K. Gomathy, "(Pdf) the Loan Prediction Using Machine Learning," no. December, 2021, [Online]. Available: https://www.researchgate.net/publication/357449126_THE_LOAN_PREDICTION_USING_MACHINE_LEARNING
[8] A. S. Aphale and D. S. R. Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval," Int. J. Eng. Res. Technol., vol. 9, no. 8, pp. 991–995, 2020, [Online]. Available: www.ijert.org
[9] J. Kajornrit, W. Inchamnam, and W. Jirapanthong, "An Investigation of Machine Learning Techniques for Loan Default Payments Prediction," vol. 13, no. 1, pp. 38–44, 2023.
[10] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021, doi: 10.1088/1757- 899X/1022/1/012042.
[11] A. A. Nureni and O. E. Adekola, "Loan Approval Prediction Based on Machine Learning Approach," Fudma J. Sci., vol. 6, no. 3, pp. 41–50, 2022, doi: 10.33003/fjs-2022-0603-830.
[12] L. Lai, "Loan Default Prediction with Machine Learning Techniques," Proc.- 2020 Int. Conf. Comput. Commun. Netw. Secur. CCNS 2020, pp. 5–9, 2020, doi: 10.1109/CCNS50731.2020.00009.
[13] K. Bogelly, C. R. Rao, S. Uppari, and K. Shilpa, "Hybrid Classification Using Ensemble Model to Predict Cardiovascular Diseases," Int. J. Res. Appl. Sci. Eng. Technol., vol. 11, no. 4, pp. 697–705, 2023, doi: 10.22214/ijraset.2023.50111.
[14] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting," Ann. Data Sci., vol. 10, no. 1, pp. 183–208, 2023, doi: 10.1007/s40745-021-00344-x.
[15] Ankit Karmakar, "Machine Learning Approach to Credit Risk Prediction: A Comparative Study Using Decision Tree, Random Forest, Support Vector Machine and Logistic Regression," FinancialMathematicsTermPaperofAnkitKarmakar, no. March, pp. 0–14, 2023, doi: 10.13140/RG.2.2.31652.14725.
[16] F. Aryanto, A. Fauzi, A. Fitri Nur Masruriyah, A. Lia Hananto, and Darmansyah, "Sentiment Analysis Of Vaccination Using The K-Nearest Neighbor Algorithm," Edutran Comput. Sci. Inf. Technol., vol. 1, no. 1, pp. 34–41, 2023, doi: 10.59805/ecsit.v1i1.6.
[17] A. Goyal, R. Kaur, and] Research Scolar, "A survey on Ensemble Model for Loan Prediction," 2013. [Online]. Available: www.ijetajournal.org
[18] N. Uddin, M. K. Uddin Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder, and S. Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," Int. J. Cogn. Comput. Eng., vol. 4, no. September, pp. 327–339, 2023, doi: 10.1016/j.ijcce.2023.09.001.
[19] Z. Liu, Z. Zhang, H. Yang, G. Wang, and Z. Xu, "An innovative model fusion algorithm to improve the recall rate of peer-to-peer lending default customers," Intell. Syst. with Appl., vol. 20, no. March, p. 200272, 2023, doi: 10.1016/j.iswa.2023.200272.
[20] P. Bhargav and P. Rama Parvathy, "Comparing Random Forest with the Naive Bayes Algorithm with Improved Accuracy: An Effective Machine Learning Method for Loan Prediction," J. Surv. Fish. Sci., vol. 10, no. 1S, pp. 2018–2029, 2023, [Online]. Available:

http://sifisheriessciences.com/journal/index.php/journal/article/view/436
[21] P. Chotwani, A. Tiwari, and M. Hooda, "Fraudulent loan prediction using machine learning algorithms," Indian J. Public Heal. Res. Dev., vol. 10, no. 5, pp. 845–850, 2019, doi: 10.5958/0976-5506.2019.01187.2.
[22] A. Gupta, R. Biwal, S. Joshi, and P. Singh, "a Comparative Study on Liver Disease Prediction Using Support Vector Machine Algorithm," Int. Res. J.Mod. Eng. Technol.Sci., no. 01, pp. 1369–1382, 2023, doi: 10.56726/irjmets33175.
[23] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Comput. Sci., vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.
[24] M. M. Aziz, M. D. Purbalaksono, and A. Adiwijaya, "Method comparison of Naïve Bayes, Logistic Regression, and SVM for Analyzing Movie Reviews," Build. Informatics, Technol. Sci., vol. 4, no. 4, pp. 1714–1720, 2023, doi: 10.47065/bits.v4i4.2644.
[25] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, "Bank loan prediction system using machine learning," Proc. 2020 9th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2020, pp. 423–426, 2020, doi: 10.1109/SMART50582.2020.9336801.
[26] E. E. Hussein, M. Y. Jat Baloch, A. Nigar, H. F. Abualkhair, F. K. Aldawood, and E. Tageldin, "Machine Learning Algorithms for Predicting the Water Quality Index," Water (Switzerland), vol. 15, no. 20, pp. 3677–3685, 2023, doi: 10.3390/w15203540.
[27] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 490–494, 2020, doi: 10.1109/ICESC48915.2020.9155614.
[28] S. Sreesouthry, A. Ayubkhan, M. M. Rizwan, D. Lokesh, and K. P. Raj, "Loan Prediction Using Logistic Regression in Machine Learning," vol. 25, no. 4, pp. 2790–2794, 2021, [Online]. Available: http://annalsofrscb.ro
[29] Y. Zhou, "Loan Default Prediction Based on Machine Learning Methods," no. Ml, 2023, doi: 10.4108/eai.2-12-2022.2328740.
[30] K. R. Prathap and R. Bhavani, "Study comparing classification algorithms for loan approval predictability (Logistic Regression , XG boost , Random Forest , Decision Tree)," vol. 10, pp. 2438–2447, 2023.
[31] A. Shobana, N. Kokilavani, R. Menaga, and M. Ramya, "Bank Loan Prediction Using Knn Algorithm," Int. Res. J. Mod. Eng. Technol. Sci., no. 03, pp. 2944–2948, 2023, doi: 10.56726/irjmets34927.
[32] T. Zhang and B. Li, "Loan Prediction Model Based on AdaBoost and PSO- SVM," vol. 147, no. Ncce, pp. 733–739, 2018, doi: 10.2991/ncce- 18.2018.120.
[33] V. Singh, A. Yadav, R. Awasthi, and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," 2021 Int. Conf. Intell. Technol. CONIT 2021, pp. 21–24, 2021, doi: 10.1109/CONIT51480.2021.9498475.
[34] Z. Nabavi, M. Mirzehi, H. Dehghani, and P. Ashtari, "A Hybrid Model for Back-Break Prediction using XGBoost Machine learning and Metaheuristic Algorithms in Chadormalu Iron Mine," J. Min. Environ., vol. 14, no. 2, pp. 689–712, 2023, doi: 10.22044/jme.2023.12796.2323.
[35] M. Anand, A. Velu, and P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," J. Comput. Sci. Eng., vol. 3, no. 1, pp. 1–13, 2022, doi: 10.36596/jcse.v3i1.237.
[36] K. Gautam, A. P. Singh, K. Tyagi, and S. Kumar, "Loan Prediction using Decision Tree and Random Forest," pp. 853–856, 2020.
[37] V. Moscato, A. Picariello, and G. Sperlí, "A benchmark of machine learning approaches for credit score prediction," Expert Syst. Appl., vol. 165, no. September 2020, p. 113986, 2021, doi: J. Matuszewski and A. Rajkowski, "The use of machine learning algorithms for image recognition," vol. 13, no. 3, p. 48, 2020, doi: 10.1117/12.2565546.10.1016/j.eswa.2020.113986.
[38] J. Matuszewski and A. Rajkowski, "The use of machine learning algorithms for image recognition,"

vol. 13, no. 3, p. 48, 2020, doi: 10.1117/12.2565546.
[39] S. A. Fard and S. Finance, "Risk Prediction for Loan Applications By Machine Learning Algorithms," 2023.
[40] Y. Dasari, K. Rishitha, and O. Gandhi, "Prediction of Bank Loan Status Using Machine Learning Algorithms," Int. J. Comput. Digit. Syst., vol. 14, no. 1, pp. 139–146, 2023, doi: 10.12785/ijcds/140113.

BIBLIOGRAPHY

[41] H. K. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9028580.
[42] NRB, "Unofficial Translation Monetary Policy for 2022/23," Nepal Rastra Bank, no. 7, pp. 1–67, 2022, [Online]. Available: www.nrb.org.np
[43] J. C. Alejandrino, J. P. Bolacoy, and J. V. B. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction," Int. J. Electr. Comput. Eng., vol. 13, no. 2, pp. 1837–1847, 2023, doi: 10.11591/ijece.v13i2.pp1837-1847.
[44] A. Huang, R. Xu, Y. Chen, and M. Guo, "Research on multi-label user classification of social media based on ML-KNN algorithm," Technol. Forecast. Soc. Change, vol. 188, no. May 2022, p. 122271, 2023, doi: 10.1016/j.techfore.2022.122271.