INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

www.ijraset.com

Call: ◎08813907089 | E-mail ID: ijraset@gmail.com

A Comparison of Various Machine Learning Algorithms and Deep Learning Algorithms for Prediction of Loan Eligibility

Archana S.[1], Divyalakshmi K. S.[2]

Department of Mathematics, Amrita Vishwa Vidyapeetham

Abstract: Banking serves as a link between people who have money in excess and those who lack it for any reason, making it a crucial component of the entire financial system. People depend on banks to borrow money to meet their needs Banks provide money as loans with an interest rate that has to be repaid. The loans are approved by banks based on the borrower's numerous characteristics. The development of artificial intelligence gave financial organizations a method to broaden their lending practices without taking on excessive financial risk. The process can be sped up, made more effective, and less error-prone by using machine learning and deep learning models for loan approval. In this study, we compare ten Machine Learning models, including Decision Tree, Logistic Regression, K Nearest Neighbour (KNN), Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes, XGBoost, Gradient Boosting, Adaboost, and Deep Learning models, including Deep Neural Network (DNN) and Long Short Term Memory network (LSTM), to predict loan applicants who deserve the money. Both Logistic Regression and Linear Discriminant Analysis had the best accuracy of 82.43% after examining all of these models.

Keywords: Machine Learning, Deep Learning, Loan Prediction, Ensemble Methods, Artificial Intelligence

I. INTRODUCTION

In the lives of people, money is quite important. In the modern world, it has evolved beyond a basic necessity. Without the assistance of money, nothing is possible. Every person can't have enough money to fulfill all of their desires. In order to obtain the needed amount of money, he borrows money, which he then repays. Lending money is a crucial aspect of the bank's operations. On loans that are approved for their customers, banks charge interest. Banks ensure that the money they lend will be repaid along with the interest by the customer. Therefore, determining the customers' creditworthiness is crucial before approving the loan. There is no guarantee that an applicant will get his loan approved. Banks and financial institutions undergo a rigorous process before approving loans. The customer must wait weeks to find out if they merit it. Artificial intelligence can speed up the process and cut down on time and labor requirements. The loan sanctioning procedure is accelerated by artificial intelligence, which also improves accuracy and reduces costs. Through automation, we can predict a customer's loan eligibility using machine learning and deep learning algorithms. The loan forecast will be advantageous for both bank personnel and applicants if the best algorithms are used.

The purpose of this study is to evaluate the effectiveness of various machine learning and deep learning models for predicting loan eligibility. We utilized a Kaggle dataset with 614 samples and 13 attributes for this. To discover a model with greater accuracy, training and testing of the model are carried out after collecting the data, addressing missing values, and standardizing the independent variables. The results are then analyzed. To achieve this, we employ well-known Machine Learning techniques like Decision Tree, Logistic Regression, K-Nearest Neighbour (KNN), Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis, Gaussian Naive Bayes, XGBoost, Gradient Boosting, and AdaBoost, as well as Deep Learning techniques like Deep Neural Network (DNN) and Long Short Term Memory network (LSTM). A variety of criteria are used to evaluate the models, including F1 score, accuracy, and precision. The literature review, methodology, and a brief discussion of the analyzed results are covered in the next section.

II. LITERATURE SURVEY

A prediction is something one may think that can happen in the future. In Artificial Intelligence, Prediction or forecasting refers to a statistical algorithm that feeds the model with historical data and predicts the probability of an outcome. There are many models in Machine Learning which is efficient in loan prediction. This paper helps find the best classifier for forecasting loan approval.

A study comparing machine learning methodologies was suggested in S.Sobana et al.'s publication [1]. This project uses data from a financial institution that includes attributes like Gender, Married, Education, Dependents, Credit History, Loan Amount, and others. After preprocessing the data, two classifiers—logistic regression and decision tree—were included. After evaluating the classifiers, the ideal model is chosen. This model's implementation was done using the R programming language. Following evaluation, it was discovered that Decision Tree had the highest accuracy, at 85.4%.

The goal of Philippe et al.'s work [2] is to compare six different machine learning models to find the one that best predicts the loan sanction status. The project's dataset, which had eleven attributes, was taken from the Kaggle website. The Radial Basic Function model, which was tested, exhibits the best prediction with an accuracy of 86%.

The study [3] by Ndayisenga et al. used information from the Bank of Kigali, a commercial bank, to forecast applicant behavior by creating and evaluating the efficacy of various Machine Learning models, including ensemble methods. Gradient Boosting fits the best (with an accuracy of 80.40%), while other models fared badly.

Using data gathered from multiple sources, preprocessed, and put into the three machine learning models of Logistic Regression, Decision Tree, and Random Forest, [4] by A. Sarkar developed a statistical model. Logistic Regression was the most accurate model for predicting loan acceptance, with an accuracy rate of 80.78%.

Prateek Dutta offered a study where three machine learning algorithms were explored for the improvement of loan prediction in his article [5]. Logistic Regression, Decision Tree, and Random Forest were the three models that were used. Pandas, Numpy, Matplotlib, Seaborn, and Sklearn were utilized in the model along with the data that was obtained from Kaggle. The most accurate model compared to the other two was determined to be logistic regression, which had an accuracy rate of 89.70%.

A strong and efficient software method that classifies based on 13 properties of the data from Kaggle was developed in the paper [6] by Sharayu Dosalwar et al. The following models were examined: SVM, Decision Tree, Random Forest, Logistic Regression, XGBoost, KNN, and Naive Bayes. Given the lucidity and adaptability of its mathematics, it is discovered that Logistic Regression provides greater accuracy, yielding an accuracy of 78.5%.

The information of previous customers from various banks was used in the work [7] by Nitesh Pandey et al. to examine loan prediction by the machine learning technique. They employed the Support Vector Machine (SVM), which has higher accuracy, among four machine learning methods, along with Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. Confusion matrix, precision, recall, accuracy, and F1 score are employed for evaluation.

Machine learning is used in [8] by A. Gupta et al. to determine who is qualified for a loan based on the history of the person to whom the loan amount has previously been accredited. 13 attributes from the Kaggle data were preprocessed before being fed into machine learning models like Random Forest and Logistic Regression. Django, HTML, and CSS are used to deploy the findings on a local server. The major objective of the project was to determine the loan status, or whether it is safe or dangerous, by entering factors like gender, marital status, education, employment status, application income, etc.

P. Tumuluru et al.'s article [9] offered a study in which Machine Learning techniques are used to extract patterns in predicting future loan defaulters from a typical loan-approved dataset. The investigation was conducted using data on past customers, including their age, income, loan amount, and employment information. The most relevant features or those that had the most impact on the prediction outcome were found using a variety of Machine Learning approaches, such as Random Forest, Support Vector Machine, K-Nearest Neighbour, and Logistic Regression. These algorithms are compared and evaluated using accepted measures. The best model is found to be Random Forest, which has an accuracy rate of 81%.

Anand Mayank et al. predicted whether or not a loan applicant will miss payments in their paper [10]. They employed several machine learning algorithms, including Linear Discriminant Analysis, Support Vector Machine, Naive Bayes, Extra Trees Classifier, Random Forest Classifier, CatBoost Classifier, Light Gradient Boosting, Extreme Gradient Boosting, Gradient Boost, Decision Tree, K Nearest Neighbour, and Logistic Regression. The most accurate classifier was Extra Trees Classifier.

Two machine learning methods, Random Forest and Decision Tree, are suggested in the publication [11] by Madaan et al. They made use of the Kaggle Lending Club dataset. They concluded that the Random Forest classifier, which has an accuracy of 80%, is the best.

The performance evaluation of several machine learning algorithms for forecasting the acceptance of client loans is the main topic of the study [12] by R Karthiban et al. Naive Bayes, Generalised Linear Model, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting are some of the algorithms that were taken into consideration.

The goal of the study [13] by Bhattad S et al. is to swiftly and effectively identify loan applicants who are deserving of them. They employed machine learning techniques like Logistic Regression, Random Forest, and Decision Tree. They came to the conclusion that Random Forest, which had an accuracy of 80.20%, is the best.

By using a Logistic Regression method, the [14] by Shinde A et al. presents a system that determines whether or not to grant a loan to a consumer. The dataset's origin was not disclosed. After contrasting Logistic Regression with Random Forest, they found that it was the most accurate. The accuracy of logistic regression was roughly 82%.

An artificial intelligence (AI) model is used in the study [15] by Diwate Y et al. to describe a system that predicts customer loan acceptance. In this system, features are retrieved once the customer submits the details, which were pre-processed. For the purpose of predicting loan eligibility, the Support Vector Machine method is employed.

The data is gathered from prior clients of different banks who received loans, and the ML models are trained on it for accurate results in the study [16] by Kadam A et al. Loan safety is predicted using machine learning methods like Support Vector Machine and Naive Bayes. It is discovered that the Naive Bayes model is effective since it accurately meets all of the demands of bankers.

A system for automatic loan approval is suggested in [17] by Tejaswini J et al. They used three machine learning algorithms to predict client loan acceptance, including Logistic Regression, Decision Tree, and Random Forest. Due to Decision Tree's high accuracy they came to the conclusion that it was the better model and hence used in their system.

In contrast to using solely CIBIL codes, Ramya S. et al. in [18] developed a model to verify the customer and examine their background based on numerous variables like Gender, Income, Employment Status, etc. to assure the client's creditworthiness so that the loan can be sanctioned or not. Information about applicants is gathered into a dataset. To forecast the required outcome, the data is analyzed, prepared, and input into the machine learning model. Using the Logistic Regression model, precise predictions are made for numerous independent entries. To prevent the problem of overfitting, not all the variables are taken into consideration; only those that directly affect the applicant's eligibility are taken into account.

Al Mamun Miraz et al. conducted a comparative analysis of various machine learning algorithms, including Random Forest, XGBoost, Adaboost, Lightgbm, Decision Tree, and K Nearest Neighbour, in their paper [19]. The dataset for the analysis is sourced from Kaggle. Given that LightGBM achieved the highest accuracy of 92 percent, the article recommends it as the ideal model.

III. METHODOLOGY

Data collection, data preprocessing, model training, model testing, and outcome prediction are all included in the methodology (Figure 1). We used ten machine learning algorithms to train and test our data, including Deep Learning algorithms like Deep Neural Network (DNN) and Long Short Term Memory (LSTM), as well as Decision Tree, Gaussian Naive Bayes, Logistic Regression, KNN, SVM, Random Forest, Linear Discriminant Analysis, XGBoost, Gradient Boosting, and AdaBoost. All of these models are examined to show which one is the best for predicting bank loans. To implement these models, we used Python.
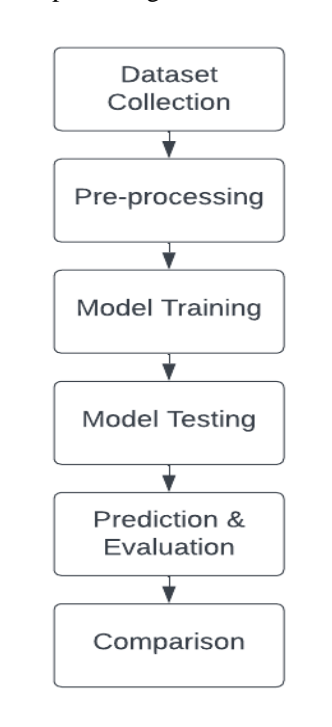

Fig. 1. Procedure

A. Data Collection

In this paper, we used the dataset that is available on Kaggle online website (https://www.kaggle.com/datasets/loan-eligibility-prediction). This dataset consists of 614 samples. There are 13 attributes such as Gender, Married, Dependents, Education, Applicant income, Coapplicant Income, Loan amount, Loan Amount Term, Credit history, Property Area, and Loan status. The description of the dataset is as in Table 1.

TABLE I
DATASET DESCRIPTION

Attribute	Description	Type
Loan_ID	Unique Loan ID	Object
Gender	Male/Female	Object
Married	Applicant Married(Y/N)	Object
Dependents	Number of dependents	Object
Education	Applicant Education (Graduate/Undergraduate)	Object
Self_Employed	Self-employed (Y/N)	Object
ApplicantIncome	Applicant Income	Integer
CoapplicantIncome	Co-Applicant Income	Float
LoanAmount	Loan Amount (in Thousands)	Float
Loan_Amount_Term	Term of a loan in months	Float
Credit_History	Credit history meets guidelines	Float
Property_Area	Urban/Semi-Urban/Rural	Object
Loan_Status	Loan Approved	Object

B. Pre-processing

Data pre-processing is an important step that enhances the quality of data to extract meaningful insights from the data. This will make it suitable for building and training the model. First, we identified and replaced the missing values in the data. The numerical attributes are replaced with the mean value of each attribute and categorical attributes with the mode value of each attribute. Then the data is split into train and test using train_test_split() from the Sklearn library. The data is then encoded using Label Encoder. Finally, we standardize the data using Standard Scalar to make all values lie in same range.

C. Model training

The dataset was split into two parts, with 70% as train data and 30% as test data. Once the data was prepared, different algorithms were applied to the train data, and prediction is done on test data. The resulting confusion matrix was then obtained. Using this confusion matrix, the accuracy, precision, and F1 score of each algorithm are calculated and compared to evaluate the performance of different algorithms.

D. Algorithms Used

1) *Decision Tree*: An effective supervised learning approach for classification issues is the decision tree algorithm. Using the dataset features to create yes/no questions, the Decision Tree method aims to repeatedly split the dataset until all the data points that belong to each class are isolated. Because of how similar its structure is to a tree, which grows from the root node and branches outward, it is known as a decision tree [17] [12].

2) *Gaussian Naïve Bayes:* An algorithm for probabilistic machine learning, known as Naive Bayes, is based on the Bayes Theorem. In addition to the Naive Bayes formula and Gaussian distribution, Gaussian Naive Bayes is a variation of the original Naive Bayes theory. The training data's mean and standard deviation must be determined before using any other functions to estimate the distribution of the data, hence Gaussian or Normal distributions are the easiest to implement. The continuous numerical attributes of Gaussian Naive Bayes are assumed to be regularly distributed. To determine the variance and mean of the characteristic for each output class, the attribute is first segmented based on the output class [12] [13].

3) *Logistic Regression:* The goal of the supervised learning technique known as logistic regression, which is frequently used for classification purposes, is to estimate the likelihood that a given instance will belong to a certain class or not. The categorical dependent variable is predicted using a set of independent variables. The logistic regression model transforms the continuous value output of the linear regression function into a categorical value output by using a sigmoid function, which transfers any real-valued collection of independent variables input into a value between 0 and 1 [12] [8].

4) *K-Nearest Neighbor:* The KNN method categorizes new variables based on a similarity metric after storing all of the existing variables. They don't base their conclusions on the underlying data in any way. It operates on the premise that observations nearest to a specific data point in the data set are those that are most "similar" to it, allowing one to classify unanticipated points based on the values of the nearby existing points. To determine how many nearby observations will be used in the method, the user can select K [6].

5) *Support Vector Machine:* We depict raw data as points in n-dimensional space using the SVM algorithm, where n is the number of features. The fact that each feature value is linked to a particular coordinate makes it simple to classify the data. A hyperplane is an ideal boundary for making decisions. Numerous acceptable hyperplanes can be used to divide two groups of data points. Utilizing the plane with the largest margin yields the best results [7].

6) *Random Forest:* A collection of decision trees is called Random Forest. It is based on ensemble learning and employs several classifiers to tackle a challenging issue and enhance performance. To boost the dataset's predicted accuracy, numerous decision trees are used on various input subsets, and the results are averaged. The accuracy increases with the number of trees used, and overfitting is avoided [12] [8].

7) *Linear Discriminant Analysis:* When it comes to efficiently separating two or more classes with many features, linear discriminant analysis is regarded as the standard technique. By projecting the dataset onto a modestly dimensional space with a true class of separable features, it lowers overfitting and computational costs. Finding a group of linear discriminants that maximize the proportion of between-class variation to within-class variance is how this is accomplished [10].

8) *Gradient Boosting:* One of the ensemble technique variations that combine several weak models to improve performance is gradient boosting. Each subsequent model in this process is trained using gradient descent to reduce the loss function of the preceding model. In each iteration, the algorithm calculates the gradient of the loss function relating to the predictions made by the current ensemble, and then it trains a new weak model to try to minimize this gradient. The ensemble is then updated with the predictions from the new model, and the process is repeated until a stopping condition is met [12] [3].

9) *XGBoost:* Extreme Gradient Boosting, also known as XGBoost, is a machine learning technique that is popular due to its versatility and capacity for handling big datasets. Decision trees are generated sequentially in this approach. Before being fed into the decision tree that forecasts the outcomes, each independent variable is given a weight. Before being fed into the second decision tree, variables that the first one predicted erroneously are given additional weight. Each such predictor is then combined to produce a powerful, accurate model [3].

10) *Adaboost:* Adaboost is also called Adaptive Boosting this combines multiple weak classifiers into a single strong classifier. In Adaboost, multiple weak learners are allowed to learn on training data and then combine these models to generate a meta-model that resolves the errors as performed by each weak learner. This meta-model has high accuracy as compared to their corresponding counterparts in prediction [19].

11) *Deep Neural Network:* Between the input and output layers of an artificial neural network are several hidden layers, which is what is known as a DNN. DNNs are typically feed-forward networks, which means that data propagates forward from the input layer to the output layer without stopping. The DNN begins by building a map of virtual neurons and connecting them with random weights or numerical values. An output between 0 and 1 is generated after multiplying the weights and inputs. If the network is unable to recognize a certain pattern with accuracy, the weights would change. By doing so, it can decide which mathematical operation will allow for the most thorough analysis of the data while also increasing the influence of specific parameters [12].

12) *Long Short Term Memory:* Recurrent Neural Networks (RNNs) can be used to learn long-term dependencies in data, and one such RNN is the LSTM. The RNN algorithm evaluates the input while taking into consideration the results of earlier events (feedback), and then briefly stores the processed information in the user's memory (short-term memory). However, RNNs have problems with vanishing gradient, which makes learning lengthy data sequences challenging, and exploding gradient, where the slope tends to grow exponentially rather than degrade while the neural network is being trained. To address this, LSTM was created by including a memory cell with a large storage capacity. Three gates make up the memory cell. The input gate regulates the amount of data that is added to the memory cell by point-wise multiplying the 'sigmoid' and 'tanh' functions. The

forget gate controls what information should be removed from the memory cell by using a 'sigmoid' function and the output gate controls what information is to be the output from the memory cell.

IV. RESULT ANALYSIS

On test data, we applied the suggested models, and a confusion matrix was produced. The F1 score, Accuracy, and Precision are then determined. Table 2 lists the effectiveness of various deep learning and machine learning models. Analyzing the data, it was discovered that, with an accuracy of 82.43%, Logistic Regression and Linear Discriminant Analysis offer the most precise predictions among these Machine Learning Models and the Deep Learning models. With an accuracy of 81.95%, the Support Vector Machine comes in second. The Decision Tree model has the lowest accuracy, at 68.29%, as we have seen. A bar plot is used to compare the accuracy of various algorithms, as seen in Figure 2.

TABLE III
MEASURES FOR DIFFERENT ALGORITHMS IN %

| Proposed Models	Accuracy	Precision	F1 Score
Decision Tree	68.29	74.48	76.86
Gaussian Naive Bayes	80.97	96.55	87.77
Logistic Regression	82.43	98.62	88.81
KNN	79.02	91.72	86.08
SVM	81.95	98.62	88.54
Random Forest	78.53	90.34	85.62
Gradient Boosting	77.07	91.03	84.88
Linear Discriminant Analysis	82.43	98.62	88.81
XGBoost	80.97	95.17	87.61
AdaBoost	75.6	87.58	83.55
Deep Neural Network	77.23	75.49	84.61
LSTM	78.38	75.64	85.51
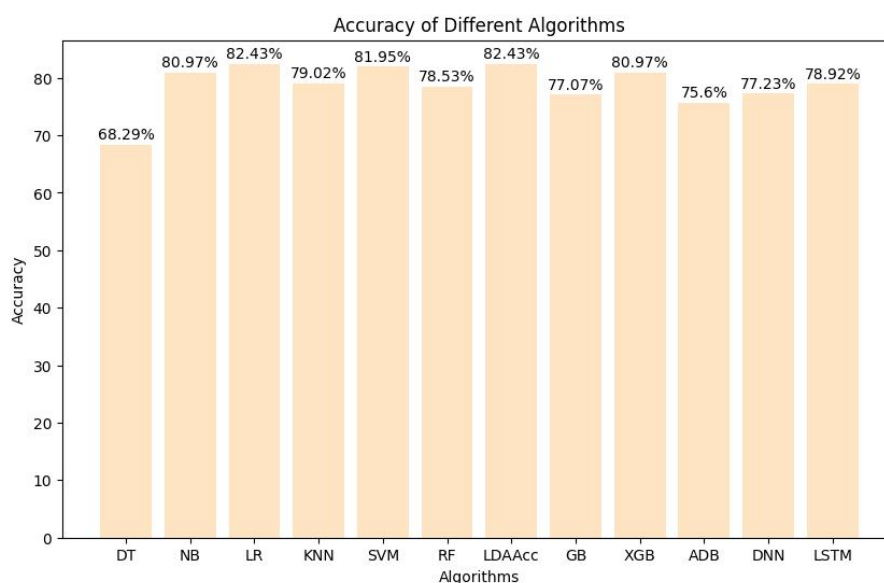
Fig. 2. Comparison of Accuracy for Different Algorithms

V. CONCLUSION

Using machine learning and deep learning models, this research aims to predict loan eligibility. It then compares the results to identify the most effective model. Due to the time savings in loan operations and the decrease in fraud during the loan approval process, these models are advantageous for both the bank and the clients. The steps in this prediction process include data collection, preprocessing, model development, data training, and data testing. After analysis, we discovered that Logistic Regression (82.43%) and Linear Discriminant Analysis (82.43%) are closely followed by Support Vector Machine (81.95%) in terms of accuracy. The future scope includes creating superior deep learning models that will surpass current machine learning algorithms for predicting loan eligibility.

REFERENCES

[1] S. Sobana and P. J. L. Ebenezer, A comparative study on machine learning algorithms for loan approval prediction analysis, IRJMETS, vol. 4, no. 12, pp. [565-569]], Dec. 2022.

[2] P. Kabenla, Automation of Loan Eligibility Process with Deep Machine Learning, Ph.D. dissertation, Utica University, 2022.

[3] Ndayisenga, Theoneste, Bank Loan Approval Prediction Using Machine Learning Techniques, Diss. 2021.

[4] A. Sarkar, Machine Learning Techniques for Recognizing the Loan Eligibility, in IRJMETS, vol. 03, no. 12, 2021.

[5] P. Dutta, A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction, in International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), vol. 3, no. 1, 2021.

[6] S. Dosalwar, K. Kinkar, R. Sannat, and N. Pise, Analysis of loan availability using machine learning techniques, in International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 9, no. 1, pp. 15-20, 2021.

[7] N. Pandey, R. Gupta, S. Uniyal, and V. Kumar, Loan Approval Prediction using Machine Learning Algorithms Approach, in International Journal of Innovative Research in Technology (IJIRT), vol. 8, no. 1, 2021.

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

[9] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba and N. Sunanda, Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms, 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 349-353.

[10] A. Mayank, A. Velu, and P. Whig, Prediction of loan behaviour with machine learning models for secure banking, Journal of Computer Science and Engineering (JCSE), vol. 3, no. 1, pp. 1-13, 2022.

[11] M. Mehul, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, 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, 2021.

[12] R. Karthiban, M. Ambika, and K. E. Kannammal, A Review on Machine Learning Classification Technique for Bank Loan Approval, in 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6.

[13] Bhattad S, Sumit Bawane, Shweta Agrawal, Unnati Ramteke, and P. B. Ambhore, Loan Prediction using Machine Learning Algorithms. International Journal of Computer Science Trends and Technology 9, no. 3 : 143-146, 2021.

[14] A. Shinde, Y. Patil, I. Kotian, A. Shinde, and R. Gulwani, Loan Prediction System Using Machine Learning, in ITM Web of Conferences, vol. 44, p. 03019, EDP Sciences, 2022.

[15] Y. Diwate, P. Rana, and P. Chavan, Loan Approval Prediction Using Machine Learning, International Research Journal of Engineering and Technology, vol. 8, no. 5, pp. 1741-1745, 2021.

[16] A. Kadam, S. Nikam, A. Aher, G. Shelke, and A. Chandgude, Prediction for Loan Approval using Machine Learning Algorithm, International Research Journal of Engineering and Technology, vol. 8, no. 4, pp. 4089-4092, 2021.

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

[18] Ramya S , Priyesh Shekhar Jha , Ilaa Raghupathi Vasishtha, Shashank H, Neha Zafar, Monetary Loan Eligibility Prediction using Machine Learning, IJESC,Vol.11,Issue.07, 2021.

[19] A. M. Miraz, A. Farjana, and M. Mamun, Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis, International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 05, 2021.

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ◎ (24*7 Support on Whatsapp)