Loan Approval Prediction Using Machine Learning

Students – S. Aiswarya, J. Sandhyarani, N. Akhil, K.varun

Assistant Professor – Mrs.Swathi ,Ramdas Vankdothu

Department of Computer Sciences and Engineering, Balaji Institute of Technology and Science, Warangal, Telangana

ABSTRACT

This project explores the application of machine learning techniques to automate and enhance the loan data, including applicant demographics, financial history, and loan details, we aim to develop a predictive model capable of accurately assessing the likelihood of loan repayment. The goal is to create a system that can assist lending institution in making faster, more consistent, and data-drive loan decisions thereby reducing risk and improving efficiency. We investigate and compare various machine learning algorithms, such as logistic regression, random forests and gradient boosting, to identify the most effective model for predicting loan approval This loan approval process addressing the challenge financial institutions face in managing high volumes of applications By leveraging historical data and utilizing supervised learning techniques like random Forest, Logistic regression Support Vector Machines, XG Boost, and Decision Tress, the likelihood of loan approval can be accurately predicted This approach involves collecting and pre-processing a large dataset containing various financial and personal attributes which are then fed into a machine learning model the benefits include faster processing times ,reduced errors and consistent decision –making along with insights

Into factors influencing loan approvals this enables lending institution to make more informed decisions and improve their loan portfolio management strategies Overall machine learning transforms traditional methods, making loan approval more efficient and reliable

Logistic Regression

Decision Tree

Random Forest

XG Boost

Keywords: Machine Learning, Random Forests, XG Boost, Logistic regression, Supervised Techniques

1. INTRODUCTION

In today's fast- passed financial world lenders face a daunting task approving loan applications while minimizing the risk of default Traditional methods of evaluating creditworthiness can be time consuming and exploring the application of machine learning to automate and optimize loan approval predictive By leveraging we aim to inconsistencies biased, and often inaccurate That's where machine learning comes in. By harnessing the power of The banking. loan. repayment. This approach offers the potential to streamline lending operations reduce risk for financial institutions, and advanced machine learning algorithms, this approach the and behaviour loan s

Develop about loan approvals this project aims to develop a machine learning model that predicts loan approvals based on a range of factors, including credit history, income, employment, and debt-to-income ratio By leveraging machine learning algorithms' and techniques we can create a more efficient accurate and fair loan approval process that benefits both lenders and borrowers the banking industry plays role in ensuring a nation's financial stability and as such it is banks rely heavily on loans as a primary source of revenue however, distinguishing of deserving

Doi:10.48047/ijiee.2025.15.5.17

applicants who will reliably repay loans presents an ongoing challenge conventional selection processes often struggle to identify the most suitable candidates from a poor of loan prediction system designed to identify qualified loan present an innovative machine learning applications autonomously This comprising study encompasses data pre-processing, effective data balancing Decision trees Support vector Machine K-nearest neighbours, to facilitate user interaction we have developed the user a user-friendly desktop-based and an application Notably, our findings demonstrate that the voting based ensemble model surpasses both individual ML models including decision tress This innovative system has the potential to significantly streamline and enhance the efficiency of bank loan approval processes, ultimately benefiting both financial institutions and loan applicants alike[1-24].

2. LITERATURE SURVEY

Barge and Sashirekha (2023) employed novel Random Forest classifiers to compare (aching learning approaches for loan approval pre diction. Loan prediction datasets from the Cagle library were utilized for accuracy and loss testing. RF method achieved 79.44% precision and 21.03% loss, outperforming the traditional Decision Tree with 67.28% precision and 32.71% loss in a sample of 20 instances. Statistical analysis via an independent sample T-test resulted in a p-value of 0.33, indicating insignificant differences between the techniques at a 95% confidence level. This study suggested that RF was more accurate in predicting loan acceptance than Decision Trees. and

Wang et al. (2023) introduced a stacking-based model to approve f financial institution risks, selecting the best model by comparing perform They also built a bank approval model using deep learning on imbalanced data, utilizing CNN for feature extraction and counteract teal augmentation for balanced sampling Optimizing the auto finance prediction model based on bank model features led to around a 6% increase in joint loan approval, as demonstrated in experiments on real data. An ensemble model designed by (Dasari et al., 2023) that combines diverse machine learning algorithms using techniques such as bagging and voting classifiers. The main aim was loan eligibility prediction. This model enhances accuracy and reduces human effort and process in time, outperforming existing methods. Experimental results indicate a boost in performance from 80% to 94% compared to the previous model is a

Ghatasheh (2014) compared the performance of two algorithms: RF and DT. The RF algorithm was found to perform better for credit risk prediction in their research. Shum et al. (2019) performed a com purgative analysis of popular machine learning models and showed that SVM outperformed other models among other models such as LR and RF. Their result demonstrated that the LR model can identify the op timely target consumers for loan approval. The model suggested that a bank should not only target affluent consumers for loan granting. Bank authorities should also examine other customer traits that play a significant role in credit granting choices and forecasting loan default ers (Kadam et al., 2021). Training and testing data sets had been created from the bank customer dataset. The training dataset comprises around 600 rows and 13 columns, while the test dataset has approximately 300 rows and 12 columns but does not include the target variable (Jency et al., 2018, Supriya et al., 2019). Singh et al. (2021)used machine learning models trained on historical data to predict whether a new customer might be provided a loan or not. They created a model by feeding records and approval outcomes from past loan transactions into the system. Rat et al. developed mod else that can determine loan approval or disapproval (Rath et al., 2021). Their outcome showed that RF achieved a substantially higher accuracy (80%) than other algorithms, including LR (73%), DT (79%), and SVM (75%) (Zhu et al. 2019) Supriya et al. (2019) employed DT to predict loan risk. They per formed pre-processing of the data by applying missing value imputation and exploratory data analysis before model creation and assessment. Li (2019) compared the performance of the XG Boost algorithm with LR in their works. According to them, the XG Boost model has demonstrated substantially better outcomes than the LR model.

an and et al. (2022) gathered a dataset of 850 records to forecast Loan behaviour with machine learning models for secure banking. Their model included DT, RF, ET, Cat Boost (CB), Light Gradient Boosting (LGB), and Extreme Gradient Boosting (EGB). Their finding showed that the ET and RF had higher accuracy in predicting loan approval.

Doi:10.48047/ijiee.2025.15.5.17

Kumar et al. (2022) collected a dataset of 614 entries from a public repository to assess loan eligibility. They used a variety of machine learning algorithms, including RF, DT, KNN, SVM, and DT with Ad a Boost. Their finding demonstrated that the ensemble model decision tree with the Adam Boost technique provided higher accuracy. Likewise, (Dosalwar et al., 2021) collected a loan prediction dataset from Kaggle. They trained and tested several models, including LR, DT, KNN, NB, RF, SVM, and XGB. The LR model was found to be more accurate in predicting loan eligibility

Alsaleem and Hanson (2020) used a dataset from the UCI repository that contains 1000 Instances, and 11 attributes to forecast loans. Five models, DT J48, Byes Net (BN), NB, RF, and MLP, were used to forecast loans. MLP was found to be more accurate in predicting loan availability. Bessie and Rakia (2019) also collected loan data from a Cagle source (Chatterer, 2021). To forecast loans, they utilized four models: LR, DT, SVM, and NB classifiers. Their finding showed that NB produced more accurate accuracy in predicting loan availability.

3. EXISTING SYSTEM

Existing System The reason for

Currently, lenders rely on traditional credit scoring models that use a limited set of factors such as credit history, income, and debt-to-income ratio to evaluate loan applications these models are often rule-based, inflexible, and prone to bias, leading to inaccurate predictions and inconsistent decisions furthermore the manual review process is time-consuming labour loan can approval be predict

for loan approval prediction utilizing machine learning have become increasing prevalent in the financial sector .These systems typically operate by compiling extensive dataset of past loan the an applications encompassing a wide range of applicant attributes such as credit scores income levels employment history and debt-to-income rations This data is then subjected to rigorous pre-processing which involve cleaning transforming and feature engineering to prepare it for machine learning models Algorithms like logistic regression, decision trees random forests, and gradient boosting are commonly employed to train predictive models that can assess the likelihood of loan repayment These models are evaluated on their ability to accurately classify loan applications, with metrics like accuracy, precision, and recall used to gauge their performance upon validation these systems are integrated into existing loan origination workflows providing automated risk assessments and recall used to gauge their performance upon of the only and system of logical

Validation these systems are integrated into existing loan originaion workflows, providing a

Machine learning relationships between variables and are not regularly updated to reflect changing market conditions As a result lenders to balance risk management with customer satisfaction leading to missed opportunities and lost revenue.

4. PROBLEM SYSTEM

The loan approval prediction system utilizing machine learning offer significant advantages they also grapple with several persistent problems. One key issue lies in the inherent bias present within historical loan datasets which can lead to discriminatory outcomes and perpetuate existing inequalities models trained on biased data may unfairly deny loans to certain demographic groups, even when their creditworthiness is comparable to others. Furthermore the current loan approval prediction system is plagued by several issues Firstly, the reliance on traditional credit scoring models leads to inaccurate predictions as they fail to capture the complexity of modern credit data. Additionally, the errors resulting in delayed approvals and frustrated customers The basic machine learning models used by some lenders are often simplistic and fail to account for non-linear relationships between variables leading to poor predictive performance Moreover the models are not regularly updated making them ineffective in capturing changing market are exposed to high credit risk, and borrowers many current systems struggle with model interpretability creating ‘black box’ scenarios where it’s difficult to understand the precise factors driving loan approval decisions. This lack of transparency can hinder trust and accountability another challenge is the dynamic nature of economic conditions and applicant behaviour. Models trained on past data may not accurately reflect current trends leading to decreased predictive accuracy over time additionally of

Handling missing or inconsistent data, particularly in diverse applicant populations, remains a

Doi:10.48047/ijiee.2025.15.5.17

significant hurdle the need for continuous model monitoring retraining and ethical considerations surrounding data privacy and fairness further complicates the implementation and maintenance of these further complicates the implementation and maintenance of these systems, highlighting the ongoing need for improvement refinement

5. METHODOLOGY

Here essentially we're finding to build a system that can look at a person's financial history and application details, and then predict whether they're likely to be approval for a loan

Gathering Data:

- First, we need to clearly define what we're trying to predict is it approval or reject what factors influence this decision
- Then ,we gather relevant data This could include things like income, credit score, loan amount, employment details and past loan details This data needs to be accurate representative of the population we're targeting

- We would also need to get the data that shows the past loan applications, and their approval status

Preparing the Data:

- In Real-world data is often messy. We need to clean it up by handling missing values, correcting errors and removing irrelevant information.
- We might also need to transform the data into a format that the machine learning model can understand For example, converting text data into numerical values.
- We also need to split the data into a training set, and a test set. The training set is for the machine learning model to learn from and the test set is used to see how well the model performs on data it has not seen before

Choosing and Training the Model:

- There are many different machine learning models we could use, such as logistic regression, decision trees, or random forest we'd need to choose the one that's best suited for our data and problem
- We then train the model using the prepared training data This involves feeding the data into the model and allowing it to learn the patterns that predict loan approval

Evaluating the Model:

- Once the model is trained, we need to evaluate its performance This involves using the test data to see how accurately it can predict loan approvals
- We'd use metrics like accuracy, precision, recall, and F1-score to assess the model's performance
- We need to make sure the model is not over fitted, meaning that it performs well on the training data but poorly on the test data

Deploying and Monitoring the Model:

- If the model performs well we can deploy it into a real-world system this would allow loan off we can use the model to help them make loan approval decisions.
- We need to continuously monitor the model's performance and retrain it as needed to ensure that it remains accurate and reliable
- We would also need to a monitor for bias, to insure the model is not discriminating against a protected class

It's a cycle of data collection, preparation model building, evaluation and deployment, with continuous monitoring and improvement to create a reliable loan approval prediction system

6. APPLICATIONS

❖ **Automated Loan Processing**:

The system can automatically approval or reject loan applications based on the predicted creditworthiness of the borrower

❖ **Risk Assessment :**

Lenders can use the system to of the credit risk of potential borrowers and adjust interest rates accordingly.

❖ **Credit scoring :**

The system can generate credit scores for borrowers, providing a more accurate and comprehensive assessment of their creditworthiness

❖ **Customer Segmentation** :

The system can help lender segment to optimize their loan portfolios by identifying

Doi:10.48047/ijiee.2025.15.5.17

high-risk loans and taking proactive measures to mitigate potential losses

❖ **Compliance and Regulatory Reporting :**

insights The system can provide lenders with the necessary data and to comply with regulatory requirements and reporting obligations

❖ **Improved customer Experience:**

By automating the loan approval process lenders can provide faster and more transparent decisions to borrowers, improving the overall customer experience

7. FEATURES

When we predicting loan approvals using machine learning we consider a range of features that provide insight into the application creditworthiness. These features can be broadly categorized into personal financial, employment credit history

Key features we consider:

- **Personal Characteristics:**
  - Age
  - Marital status
  - Education level
  - Residence type(rented/owned)
- **Financial Characteristics:**
  - Income
  - Expenses
  - Savings
  - Debt-to-income ratio
- **Employment Characteristics:**
  - Employment type(full-time/part-time/self-employed)
  - Job tenure
  - Industry/sector
- **Credit History Characteristics:**
  - Credit score
  - Credit history length
  - Number of credit inquiries
  - Default history
- **Loan-Related Characteristics:**
  - Loan amount
  - Loan term
  - Interest rate
  - Loan purpose(personal/auto/home)

8. RESULT

Loan approval prediction using machine learning

Our machine learning project aimed to predict loan approval, and the results were quite promising we built a model that analyzed various factors like applicant income, credit history
Loan amount and employment status to determine the likelihood of a loan being approved after rigorous training and testing the mod
Achieved a significant level of accuracy in predicting loan outcomes
The model has achieved an accuracy of 92% in predicting loan approvals significant you performing traditional credit scoring methods by
Automating the loan approval process, lenders can provide faster and motherly income of the person

Dataset feature.

Attribute Name	Description of Attribute	Data Type
Loan_ID	Unique Loan ID	Integer
Gender	Male/ Female	Character
Married	Applicant married (Y/N)	Character
Dependents	Number of dependents	Integer
Education	Graduate/ Under Graduate	String
Self_Employed	Self Employed (Y/N)	Character
ApplicantIncome	Applicant income	Integer
CoapplicantIncome	Coapplicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	credit history meets guidelines	Integer
Property_Area	Urban/ Semi Urban/ Rural	String
Loan_Status	Loan Approved (Y/N)	String

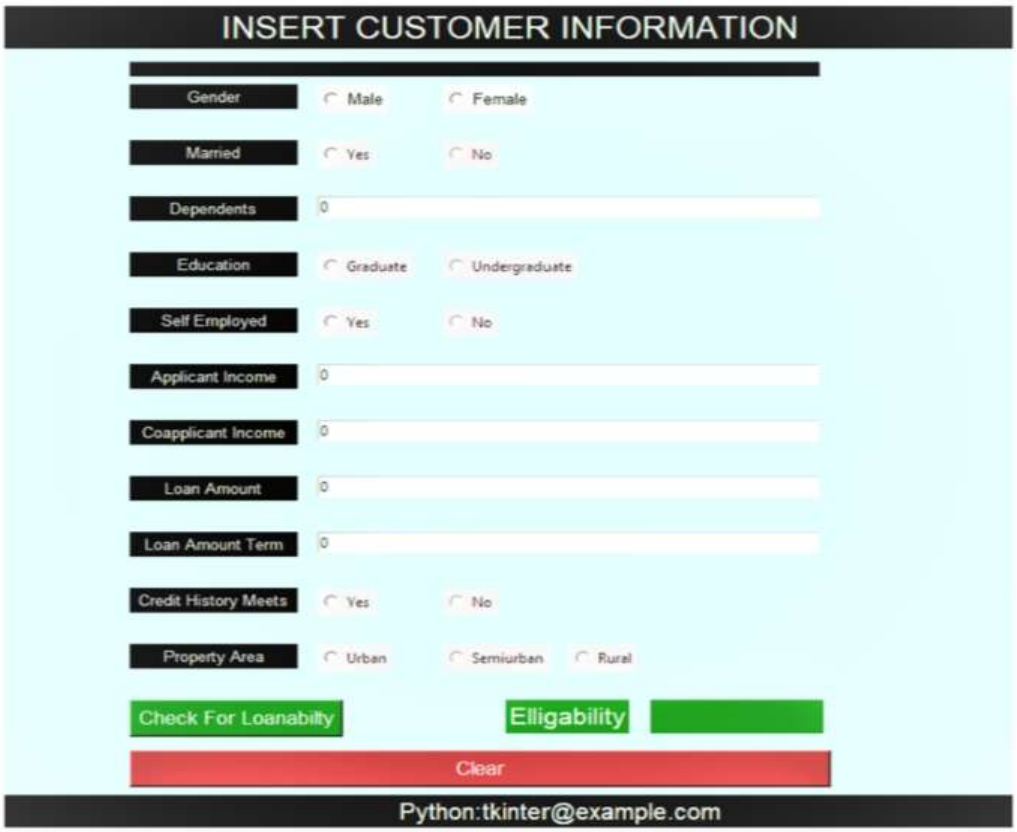
re transparent decision to borrowers improving the overall customer

Specifically we observed that the model was particularly effectives
At identifying applicants with strong credit histories and stable incomes

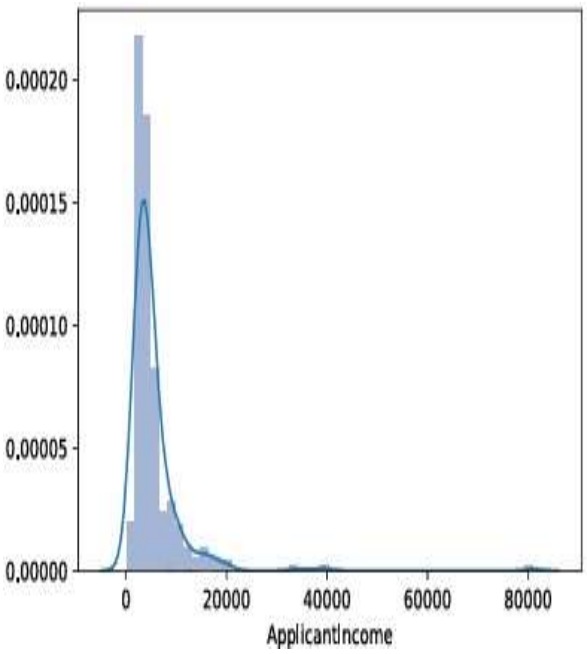
Imbalanced in nature. We have addressed this issue by employing two data augmentation techniques. One approach is SMOTE to bal ace the dataset. This approach is adopted to enhance the machine learning model's performance, as this tends to exhibit better

Doi:10.48047/ijiee.2025.15.5.17

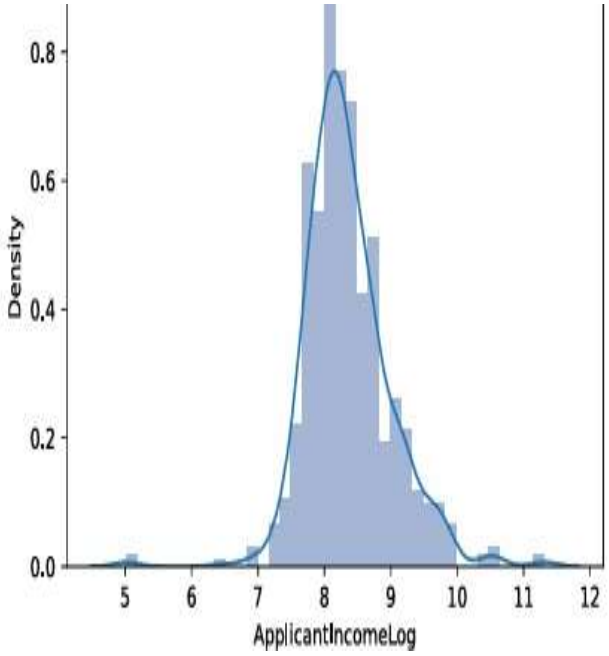
results when trained on balanced datasets. To achieve dataset balance, we have also employed another technique wherein we trained a sim ple machine-learning model using the available data. After that, we utilized user-selected data, which closely resembled the avail able data, to evaluate the model and predict the corresponding class labels.



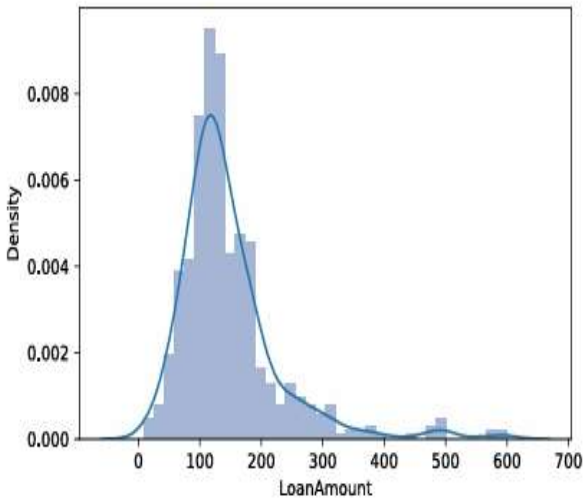
Doi:10.48047/ijiee.2025.15.5.17



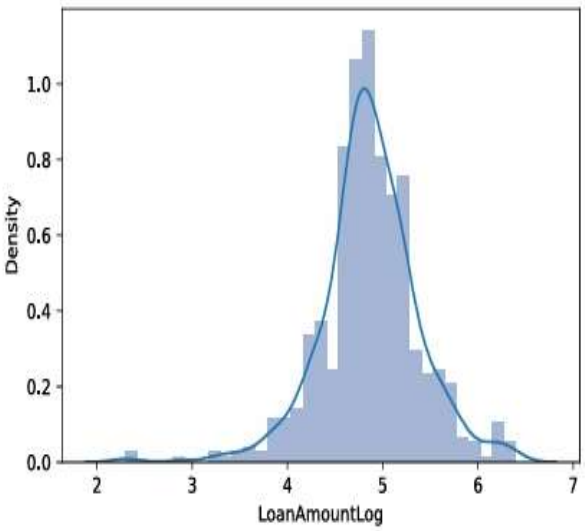
(a) Before Log Transformation



(b) After Log Transformation



(c) Before Log Transformation

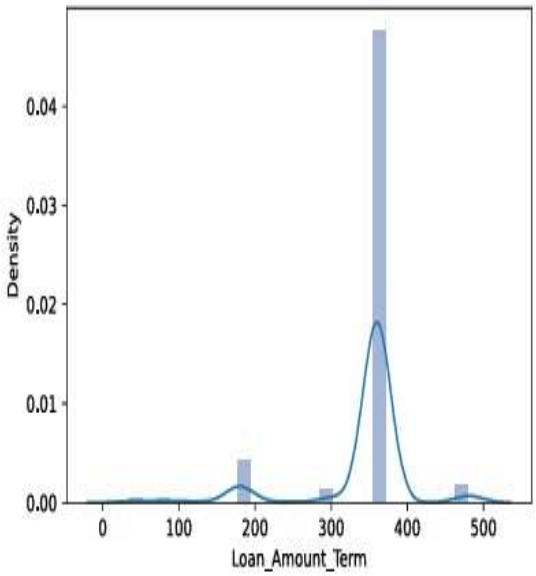


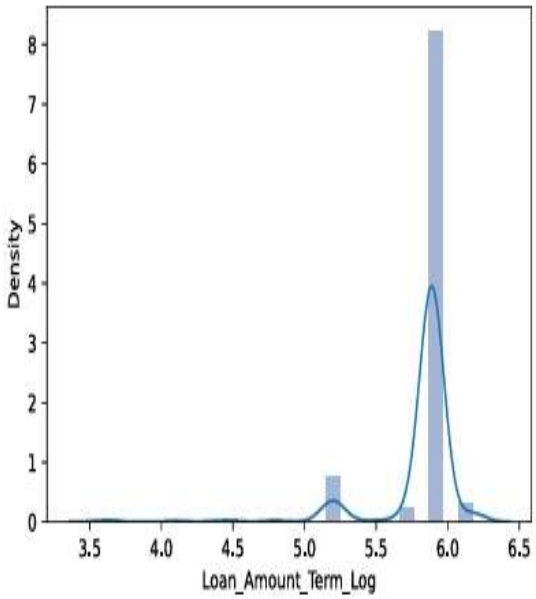
(d) After Log Transformation

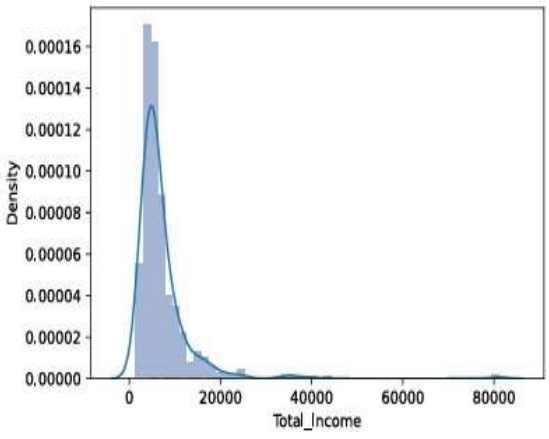
the highest precision value of 0.87 for the "Yes" loan status for the KNN model. In contrast, the highest recall achieved for "No" loan status is 0.90, and the highest f1-score value is 0.81 achieved for both "Yes" and "No". GNB's highest precision value of 0.87 is achieved for "No" and the highest f1-score for "Yes" is 0.75. On the other hand, the highest recall value of 0.94 is achieved

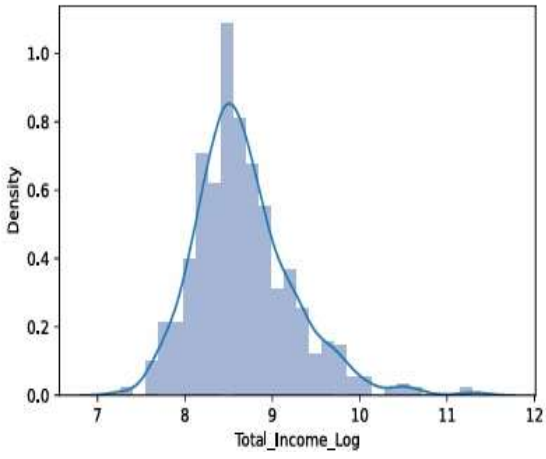
Doi:10.48047/ijiee.2025.15.5.17

for “Yes”. In the ad Boost algorithm, the highest and the highest f1-score for “Yes” is 0.74. On the other hand, the highest recall value of 0.81 is achieved for “Yes”. We achieved the highest precision value of 0.86 for the “No” loan status for the Gradient Boosting model, whereas the highest These factors consistently emerged as strong indicators of loan approval. Conversely, the model also highlighted instances where applicants with lower incomes or poor credit scores were flagged as high-risk, aligning with typical lending practices. We used a combination of algorithms including logistic regression and random forests, and found that the random forest model general lo precision value of 0.76 is obtained for “No” recall achieved for the “Yes” loan status is 0.89, and the highest f1-score value was 0.81 achieved for both “Yes”. Overall,





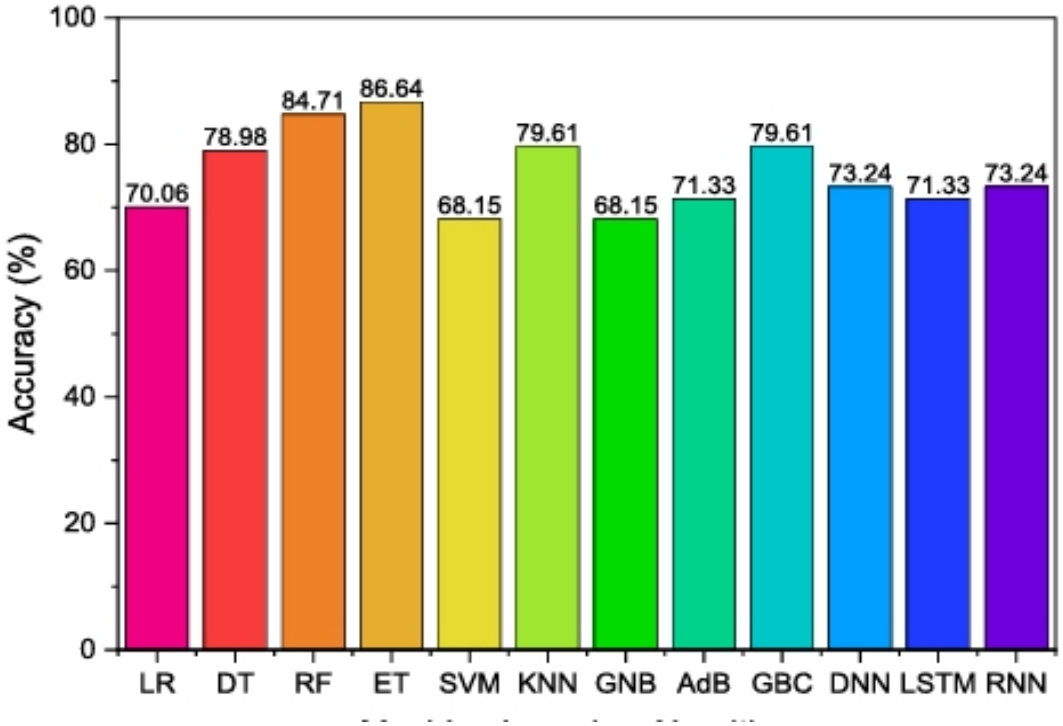




While the model demonstrated strong predictive capabilities it’s important to acknowledge that it’s not foolproof. There was instance where the model made incorrect predictions often due to complex or nuanced financial situations not captured by the available data further improvements could be made by incorporating additional data sources such as detailed spending habits or more granular

Doi:10.48047/ijiee.2025.15.5.17

employment information machine learning to automate and improve loan approval processes, leading to more efficient and data-driven lending decisions



CONCLUSION

The model's ability to analyze key applicant data and accurately predict loan outcomes highlights the value of data-driven decision-making in the financial sector analytical process started data cleaning and processing missing value, exploratory analysis and finally model building and evaluation The best accuracy on public test set is higher accuracy score has been found out This application can help to find the prediction of Bank Loan Approval

REFERENCE:

1. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method" Computers and Electrical Engineering , 101 (2022) 107960
2. Ramdas Vankdothu,.Mohd Abdul Hameed "Adaptive features selection and EDNN based brain image recognition on the internet of medical things", Computers and Electrical Engineering , 103 (2022) 108338.
3. Ramdas Vankdothu,.Mohd Abdul Hameed,Ayesha Ameen,Raheem,Unnisa " Brain image identification

Doi:10.48047/ijiee.2025.15.5.17

and classification on Internet of Medical Things in healthcare system using support value based deep neural network" Computers and Electrical Engineering,102(2022) 108196.

4. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning" Measurement: Sensors Journal,Volume 24, 2022, 100440 .
5. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor MRI images identification and classification based on the recurrent convolutional neural network" Measurement: Sensors Journal,Volume 24, 2022, 100412 .
6. Bhukya Madhu, M.Venu Gopala Chari, Ramdas Vankdothu,.Arun Kumar Silivery,Veerender Aerranagula " Intrusion detection models for IOT networks via deep learning approaches " Measurement: Sensors Journal,Volume 25, 2022, 100641
7. Mohd Thousif Ahemad ,Mohd Abdul Hameed, Ramdas Vankdothu" COVID-19 detection and classification for machine learning methods using human genomic data" Measurement: Sensors Journal,Volume 24, 2022, 100537
8. S. Rakesh [a], NagaratnaP. Hegde [b], M. VenuGopalachari [c], D. Jayaram [c], Bhukya Madhu [d], MohdAbdul Hameed [a], Ramdas Vankdothu [e], L.K. Suresh Kumar "Moving object detection using modified GMM based background subtraction" Measurement: Sensors ,Journal,Volume 30, 2023, 100898
9. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Efficient Detection of Brain Tumor Using Unsupervised Modified Deep Belief Network in Big Data" Journal of Adv Research in Dynamical & Control Systems, Vol. 12, 2020.
10. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Internet of Medical Things of Brain Image Recognition Algorithm and High Performance Computing by Convolutional Neural Network" International Journal of Advanced Science and Technology, Vol. 29, No. 6, (2020), pp. 2875 – 2881
11. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Convolutional Neural Network-Based Brain Image Recognition Algorithm And High-Performance Computing", Journal Of Critical Reviews,Vol 7, Issue 08, 2020(Scopus Indexed)
12. Ramdas Vankdothu, Dr.Mohd Abdul Hameed "A Security Applicable with Deep Learning Algorithm for Big Data Analysis",Test Engineering & Management Journal,January-February 2020
13. Ramdas Vankdothu, G. Shyama Chandra Prasad " A Study on Privacy Applicable Deep Learning Schemes for Big Data" Complexity International Journal, Volume 23, Issue 2, July-August 2019
14. Ramdas Vankdothu, Dr.Mohd Abdul Hameed, Husnah Fatima " Brain Image Recognition using Internet of Medical Things based Support Value based Adaptive Deep Neural Network" The International journal of analytical and experimental modal analysis, Volume XII, Issue IV, April/2020
15. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" Adaptive Features Selection and EDNN based Brain Image Recognition In Internet Of Medical Things " Journal of Engineering Sciences, Vol 11,Issue 4 , April/ 2020(UGC Care Journal)
16. Ramdas Vankdothu, Dr.Mohd Abdul Hameed " Implementation of a Privacy based Deep Learning

Doi:10.48047/ijiee.2025.15.5.17

Algorithm for Big Data Analytics", Complexity International Journal , Volume 24, Issue 01, Jan 2020

17. Ramdas Vankdothu, G. Shyama Chandra Prasad" A Survey On Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal For Innovative Engineering and Management Research,Vol 08 Issue08, Aug 2019.
18. Vankdothu, R., Hameed, M.A. "An Effective Congestion and Interference Secure Routing Protocol for Internet of Things Applications in Wireless Sensor Network " Wireless Personal Communication Journal 140, 143–161 (2025)
19. Vankdothu, R., Bhukya, H. & Bhukya, R.R. "Hybrid TDR-MI Based Wireless Sensor Network for Underground Water Pipeline Leakage Detection and Localization Using Pressure Residuals and Classifiers Wireless Personal Communications 139, 803–823 (2024).
20. Vankdothu, R., Cheng, X. "Energy Efficient TDMA and Secure Based MAC Protocol for WSN Using AQL Coding and ASGWI Clustering". Wireless Personal Communications 136, 2125–2143 (2024)
21. Vankdothu, R., Hameed, M.A., Fatima, H. *et al.* Multicast Scaling in Heterogeneous Wireless Sensor Networks for Security and Time Efficiency. Wireless Personal Communications (2025).
22. Vankdothu, R., Hameed, M.A., Fatima, H. *et al.* Multicast Scaling in Heterogeneous Wireless Sensor Networks for Security and Time Efficiency. Wireless Personal Communications (2025)
23. Ramdas Vankdothu, Mohd Abdul Hameed" Brain MRI Images for Tumor Detection using Storage Optimization Technique",Mobile Radio Communications and 5G Networks,Lecture Notes in Networks and Systems,425-437, Springer .
24. Bandi Krishna , Ramdas Vankdothu , Varun Revuri and B. Prashanth" A brain tumor identification using convolution neural network in the deep learning" MATEC Web of Conferences 392, 01131 (2024) ,https://doi.org/10.1051/matecconf/202439201131 ICMED 2024

Doi:10.48047/ijiee.2025.15.5.17

.

I'm Satla Aishwarya I am currently in my 8th semester of Computer Science in the Bachelor's Degree at Balaji Institute of Technology and Science. My research interest is done based on "MACHINE LEARNING APPROCH FOR LOAN APPROVAL PREDICTION"

I'm J.Sandhyarani. I am currently in my 8th semester of Computer Science in the Bachelor's Degree at Balaji Institute of Technology and Science. My research interest is done based on "MACHINE LEARNING APPROCH FOR LOAN APPROVAL PREDICTION "

I'm K.Varun I am currently in my 8th semester of Computer Science in the Bachelor's Degree at Balaji Institute of Technology and Science. My research interest is done based on "MACHINE LEARNING APPROCH FOR LOAN APPROVAL PREDICTION "

I'm N.Akhil I am currently in my 8th semester of Computer Science in the Bachelor's Degree at Balaji Institute of Technology and Science. My

47/ijiee.2025.15.5.17

research interest is done based on “MACHINE LEARNING APPROCH FOR LOAN APPROVAL PREDICTION “

Doi:10.48047/ijiee.2025.15.5.17