# Loan Approval Prediction using Machine Learning

Dr. Rama Krishna Eluri[1], Battu Venkata Siva[2], Shaik Aariz Ahmed[3], Shaik Mastan Vali[4]
[1] Professor, [2, 3 & 4] Student
[1]drrkeluri@gmail.com, [2] venkatasivabattu2002@gmail.com, [3] aarizahmed2002@gmail.com, [4] skmasthanvali032@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

***ABSTRACT-Loans are a crucial part of the modern world, and banks receive a significant portion of their profits from them. However, deciding whether to grant a loan to an applicant is a complex process that requires banks to consider many factors.***

***In this study, we suggest a machine learning-based method to streamline the loan acceptance prediction process. To determine whether or not a loan applicant's profile is relevant for approval, we employ effective machine learning algorithms. We base our predictions on important features. Additionally, we present a comparison study of various categorization methods to demonstrate how machine learning algorithms might enhance the loan approval procedure. Our results show that machine learning algorithms can significantly reduce the risk of loan defaults and improve the loan approval process. Moreover, to enhance prediction accuracy, we incorporate a voting ensemble technique into our methodology. This additional layer of analysis further refines our predictions, contributing to more reliable loan approval decisions.***

***KEYWORDS--Loan approval prediction, Loan default risk, Predictive modeling, Feature selection, SMOTE, Decision Tree,, Gradient Boost, Random Forest, Extra Tree, Comparative analysis, Voting Ensemble.***

## 1. INTRODUCTION

One of the main functions of the banking sector is lending, which is crucial to maintaining the financial stability of a country. Banks receive a large amount of their interest revenue from loans, thus the loan approval procedure is quite important. However, because it mostly relies on manual methods, this approach has accuracy and efficiency issues. Individual bank managers evaluate the risk of loan default and application eligibility, which might cause systemic disruptions that could affect the economy as a whole as well as possible financial losses for banks.

It has always been difficult for banks to pick creditworthy borrowers from a large application pool. Effective loan default prediction is crucial to risk mitigation in contemporary banking systems. Miscalculations in anticipating loan defaults may result in financial crises with extensive ramifications.

In this research, we use data-driven strategies to streamline the loan approval process. We start with the removal of null values, duplicates, and outliers from the data, cleaning it up, and examining any relationships between the variables. The purpose of feature selection is to find the most pertinent qualities. To address the disparity in class sizes, we employ the SMOTE. We then use machine learning methods to forecast the results of loan approval, such as gradient boost, decision tree, random forest, and other techniques. Each algorithm's performance is assessed using various metrics. We also examine how each algorithm contributes to enhancing the loan approval procedure.

Since the primary objective of this study is to employ data-driven methodologies to enhance the precision and effectiveness of the loan approval procedure.We do this with methodically addressing the main issues with conventional loan approval procedures. The following succinctly describes our paper's primary contributions:

### **A.** *Selection of Datasets*

To start, we take great care in choosing datasets relevant to the loan approval procedure. These files provide vital details about candidates, such as work status, credit ratings, financial histories, and other pertinent characteristics. We ensure that our research is well-founded by using datasets that are both extensive and indicative of real-world circumstances.

### *B. Data Preprocessing*

To clean and prepare the data, we go through a number of preprocessing stages before using machine learning algorithms. Managing missing numbers, eliminating duplicates, and spotting outliers that can bias our analysis are all part of this process. Furthermore, we utilise feature selection methods to determine which attributes are most pertinent for forecasting loan approval results. In order to guarantee the durability of our classification models, we also address class imbalance issues utilising methods like the Synthetic Minority Over-sampling Technique (SMOTE).

### *C. Data Splitting*

We divided the preprocessed datasets into training and testing sets in order to efficiently assess the performance of our machine learning models. This guarantees that our models are tested on a different subset of data after being trained on a different one, which helps us determine how well they can generalise.

By allocating 70% of the data for training and 30% for testing, we are able to strike a balance between the two processes of model evaluation and training.

### *D. Application of Machine Learning methods*

We use a variety of machine learning methods, such as Random Forest, Decision Tree, Extra Tree, and Gradient Boost, to predict the results of loan acceptance. These approaches all have different benefits in terms of interpretability and prediction accuracy. Our goal is to use these algorithms to find useful patterns and trends in our datasets so that decision-makers can

make well-informed decisions when approving loans. In addition, we employ a voting ensemble technique, generating the final forecast by aggregating the predictions from the top three performing models.

## 2. LITERATURE SURVEY

Manjeet Kumar et al. [3] evaluated a number of classifiers, such as Light Gradient Boosting Machine (LGBM), Extra Trees, Random Forest, and Extreme Gradient Boosting (XGB) for the purpose of predicting bank loan default. Their study provides insightful information for financial institutions by highlighting the significance of debt income and work history in forecasting defaults. The comparative study of classifiers by Kumar et al. offers a thorough grasp of performance criteria including key metrics.

In the research, Mehul Madaan et al. [4] obtained 73% and 80% accuracy, respectively, in loan default prediction using Decision Trees and Random Forest algorithms. For financial organisations looking to enhance loan approval procedures and reduce credit risks, their study provides insightful information. Through the analysis of these algorithms' performance on a shared dataset, the research adds to the continuing investigation of machine learning applications in the banking industry.

Decision Trees and other machine learning models were used by Supriya et al. [2] to predict loan defaulters with an accuracy of 81.1%. Their research focused on data preparation methods, such as managing outliers and missing information, and resulted in a thorough examination of the characteristics that affect loan acceptance. The writers provided insightful information for improving credit risk assessment in the banking industry by highlighting the importance of variables like income level and credit history in loan sanctioning decisions.

Mahankali et al. [1] forecast loan approvals with an accuracy rate of 80.945% by using logistic regression. Their all-inclusive strategy comprises testing, model creation, and data pretreatment, offering a solid foundation for automated loan approval systems. This study provides useful information about the use of machine learning algorithms in banking settings and sets the standard for further research in the area.

When comparing machine learning algorithms for forecasting bank loan risks, Alsaleem et al. [5] discovered that Multilayer Perceptron has the best accuracy (80%). With an emphasis on useful applications and decision support systems, this study provides a baseline for comparable research in the field and offers insightful information on using neural networks for loan classification.

In order to predict loan acceptance, Ramachandra et al.'s [6] study used machine learning techniques like Random Forest, Decision Tree, Logistic Regression and it was 86% accurate. Their study provides insights into data pretreatment, algorithm selection, and result interpretation, and it shows that cloud-based platforms can be a viable solution for implementing loan prediction models.

A modernised loan approval system based on the XGBoost, Random Forest, and Decision Tree algorithms was created by Singh et al. [7] to achieve accurate loan prediction. By reliably determining loan eligibility and boosting lending volume, their study helps banks reduce losses. The architecture diagram of the system demonstrates how well it can forecast the results of loan approval, improving banking efficiency and risk management.

## 3. THE PROPOSED SYSTEM FOR LOAN APPROVAL PREDICTION

Our proposed model for loan approval prediction is structured around distinct phases and steps as shown in Fig.1 , each tailored to maximize accuracy, efficiency, interpretability, robustness, and fairness. The criteria as follows:

- Analysis of Dataset
- Visualization of Data
- Preprocessing Techniques
- Model Development

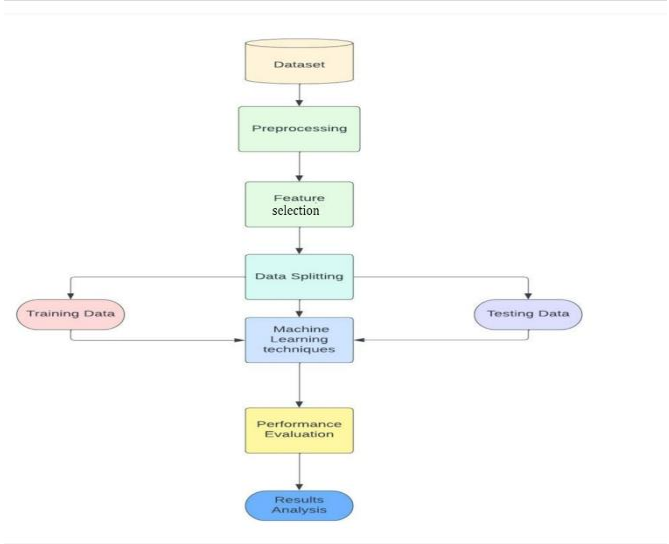


Fig.1 Flow Chart

*A. Analysis of Dataset*

We used the well-known Kaggle website, which hosts a variety of datasets, to find the datasets we needed for our predictive study. The dataset that we used is accessible [6] For our research to be successful, having access to these datasets shown in Fig.2 is essential for performing in-depth analysis and forecasts. The dataset consists of 13 columns and 614 entries. The information is as follows:

| Column Name | Meaning |
|---|---|
| Loan_ID | Unique identifier for each loan application |
| Gender | Gender of the applicant (Male/Female) |
| Married | Marital status of the applicant (Yes/No) |
| Dependents | Number of dependents the applicant has |
| Education | Educational qualification of the applicant (Graduate/Not Graduate) |
| Self_Employed | Indicates whether the applicant is self-employed (Yes/No) |
| ApplicantIncome | Gross income of the applicant |
| CoapplicantIncome | Gross income of the co-applicant |
| LoanAmount | Amount of loan applied for |
| Loan_Amount_Term | Term of the loan (in months) |
| Credit_History | Credit history of the applicant (1: Good, 0: Bad) |
| Property_Area | Location of the property associated with the loan application (Urban/Semiurban/Rural) |
| Loan_Status | Status of the loan application (Approved/Rejected) |

Fig.2: Data Set

*B. Visualization of Data*

As the target class to be predicted, the "Loan_Status" column shows a class imbalance in the Fig.3, with roughly 68.7% of the entries labelled as "Y" (meaning loan approval) and the remaining 31.3% labelled as "N" (showing loan denial). The performance of several models, especially those that are sensitive to class distribution, may be impacted by this imbalance, necessitating the use of procedures like resampling or model assessment metrics that are specifically designed for imbalanced datasets

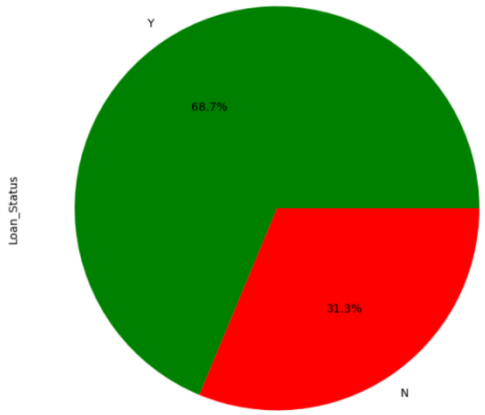


Fig3. Loan_Status

*1.Comparing the features with respect to the target variable*

We carried out a thorough investigation to look at the connection between different characteristics and the target variable, concentrating on those that have a big impact on loan approval choices. We sought to determine important variables and comprehend their influence on the loan approval procedure through our comparison analysis.

- *Gender vs Loan_Status*

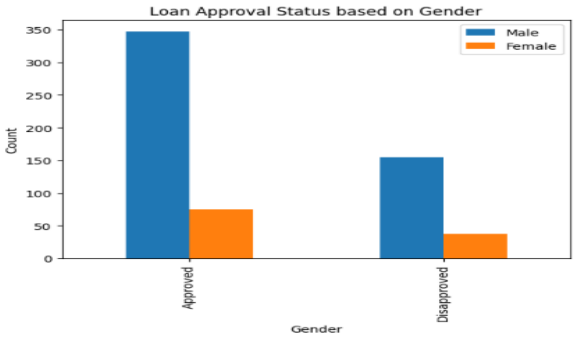


Fig 4: Gender vs Loan_Status

The data in Fig.4 displays loan approval status by gender, highlighting differences in approval rates. In the case of men, 340 loans were authorised vs 130 denied, and in the case of women, 80 loans were accepted and 40 rejected. There may be gender-based differences in the outcomes of loan approvals, since this data points to a greater approval rate for men than for women.

- *Marrital status vs Loan_Status:*

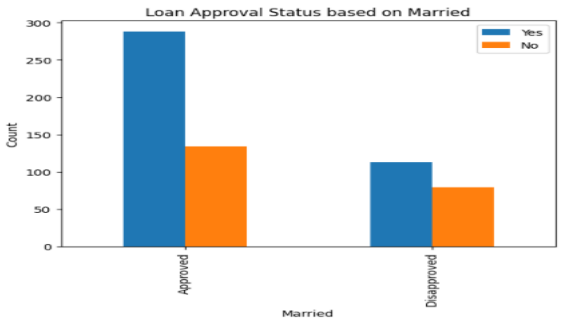


Fig 5: Marrital status vs Loan_Status

The loan approval status with respect to marital status behaved as 280 married individuals had 280 approved loans compared to 110 denied, while 140 unmarried individuals had 140 approvals compared to 70 disapprovals in Fig.5. The data indicates that applicants who are married had a higher approval rate, suggesting that marital status may have an impact on loan acceptance results.

- *Education vs Loan_Status:*

The education status has a significant impact on loan acceptance rates; graduates have a significantly higher approval rate than non-graduates which is clearly represented in Fig.6. On the other hand, a larger

percentage of loan denials was encountered non-graduates, suggesting a possible relationship between educational attainment and loan approval result.

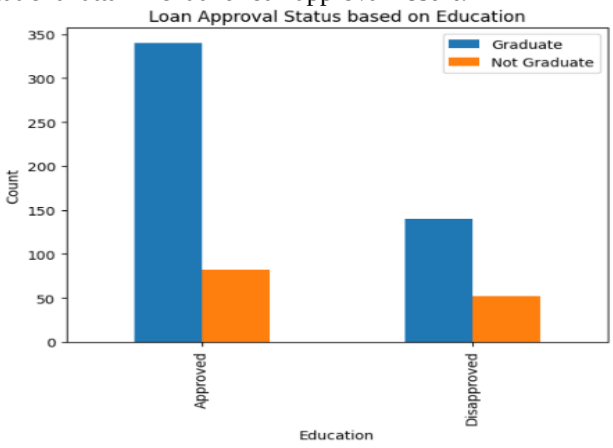


Fig6: *Education vs Loan_Status*

- *Self-Employment vs Loan_Status:*

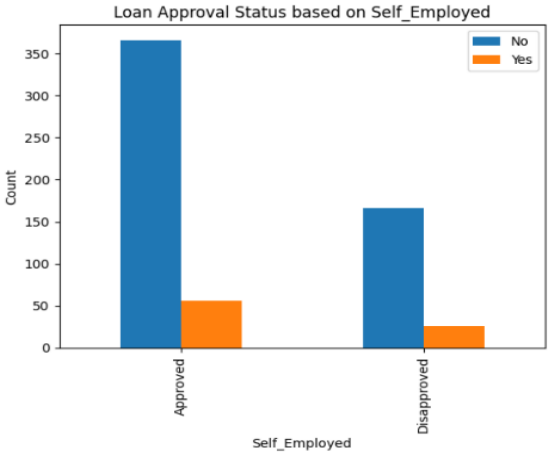


Fig7: Self-Employment vs Loan_Status

From Fig.7, Self-employment status has a substantial impact on loan approval status; those who are self-employed are approved at a higher rate than those who are not. On the other hand, applicants who are not self-employed exhibit a comparatively lower percentage of loan denial.

- *Credit History vs Loan_Status*:

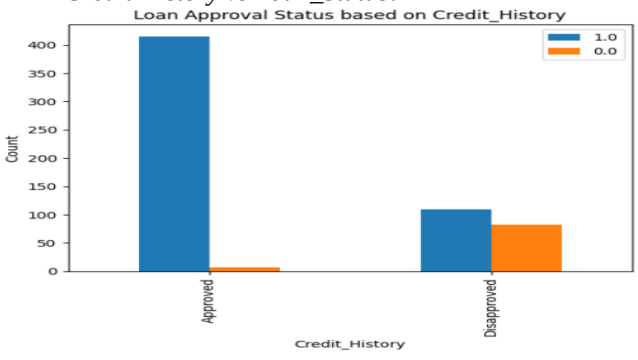


Fig8: *Credit History vs Loan_Status*

Credit history has a significant impact on loan acceptance status; Fig.8 shows that candidates with a favourable credit history are far more likely to be approved than those without one. On the other hand, those without a credit history had a comparatively greater rate of loan denial, suggesting that credit history has a substantial influence on loan approval results.

- *Property_Area vs Loan_Status:*


Fig 9: Property_Area vs Loan_Status

The status of loan approval varies depending on the type of property; with Fig.9 we can state that semi-urban areas have the greatest approval rates, followed by rural and urban areas. On the other hand, loan denials occur more frequently in metropolitan areas than in semi-urban and rural areas. These differences highlight how crucial it is to take locality into account when approving loans. Financial areas can use this information to improve risk management procedures that are suited to particular types of property and to optimise lending strategies.

*C. Preprocessing Techniques*

The next stage after data visualisation is preprocessing, which gets the data ready for model training. The following are the steps.

- *Dealing Null Values:*

Reliability of the model and data quality depend on addressing null values. By managing null values using methods like imputation or elimination, machine learning models become more accurate and the dataset's integrity is preserved. Potential biases are reduced by efficient null value management, allowing for more precise forecasts and perceptive analysis The fig 10(a), fig 10(b) illustrates the frequency of missing values in different columns. It exposes deficiencies in important characteristics such as Gender, Dependents, Self_Employed, and Credit_History, for example. Handling these blank values is crucial to ensuring the integrity and dependability of the data in subsequent research. In

order to address this problem, we imputed missing values using the fillna approach, which reduced the possibility of biases and improved the overall quality of the dataset.

```
df.isnull().sum()
```

```
Loan_ID               0
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```

Fig.10 (a) Null Values in the Data set

```
df.isnull().sum()
```

```
Loan_ID               0
Gender                0
Married               0
Dependents            0
Education             0
Self_Employed         0
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount            0
Loan_Amount_Term      0
Credit_History        0
Property_Area         0
Loan_Status           0
dtype: int64
```

Fig 10 (b) After Removal of Null Values

- *One-hot Encoding:*

  To ensure compatibility with numerical computations, one-hot encoding is essential for transforming categorical variables into a format that is appropriate for machine learning methods. Algorithms can successfully read and learn from categorical features by expressing categorical data as binary vectors. By doing this, the feature space is increased, maintaining the distinctive qualities of every category and improving. The prediction ability of the model. This process can be carried out as shown in fig 11 (a).

```
df = pd.get_dummies(df)
df = df.drop(['Gender_Female', 'Married_No', 'Education_Not Graduate',
              'Self_Employed_No', 'Loan_Status_N'], axis = 1)
new = {'Gender_Male': 'Gender', 'Married_Yes': 'Married',
       'Education_Graduate': 'Education', 'Self_Employed_Yes': 'Self_Employed',
       'Loan_Status_Y': 'Loan_Status'}

df.rename(columns=new, inplace=True)
print(df.head())
df.shape
```

Fig11 (a): One-Hot Encoding Process

```
df.columns
```

```
Index(['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount',
       'Loan_Amount_Term', 'Credit_History', 'Gender', 'Married',
       'Dependents_0', 'Dependents_1', 'Dependents_2', 'Dependents_3+',
       'Education', 'Self_Employed', 'Property_Area_Rural',
       'Property_Area_Semiurban', 'Property_Area_Urban', 'Loan_Status'],
      dtype='object')
```

Fig 11 (b) Columns after One-hot encoding

- *Eliminating Data Outliers*

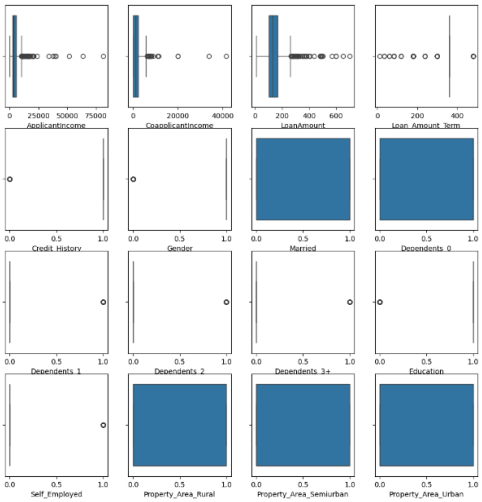


Fig 12: Visualizing data outliers

Handling outliers is essential to preserving the accuracy and integrity of machine learning models and statistical analysis. Results can be severely skewed by outliers, which can also influence interpretations and provide false findings. Researchers can guarantee robust and dependable data analysis, enhancing the calibre and validity of their findings, by recognising and managing outliers. Fig.12 Illustrating the presence of Outliers in the data. Since, the dataset had outliers that could have distorted the results of our investigation and the performance of the model. We utilised the quantile approach to solve this problem, identifying and eliminating extreme values by establishing thresholds based on quantiles. By using this strategy, we were able to lessen the impact of outliers on our findings and maintain the validity and precision of our study.

- *Square Root Transformation:*

  Because it lessens the impact of extreme values, square root transformation is essential for stabilising variance, particularly in datasets with skewed distributions. It works by improving the symmetry and conformance of the data to normalcy assumptions, which can guarantee reliable inference and enhance the performance of specific statistical models. The influence of SRT on applicant income, coapplicant income, and loan amount is seen in the following fig 13.

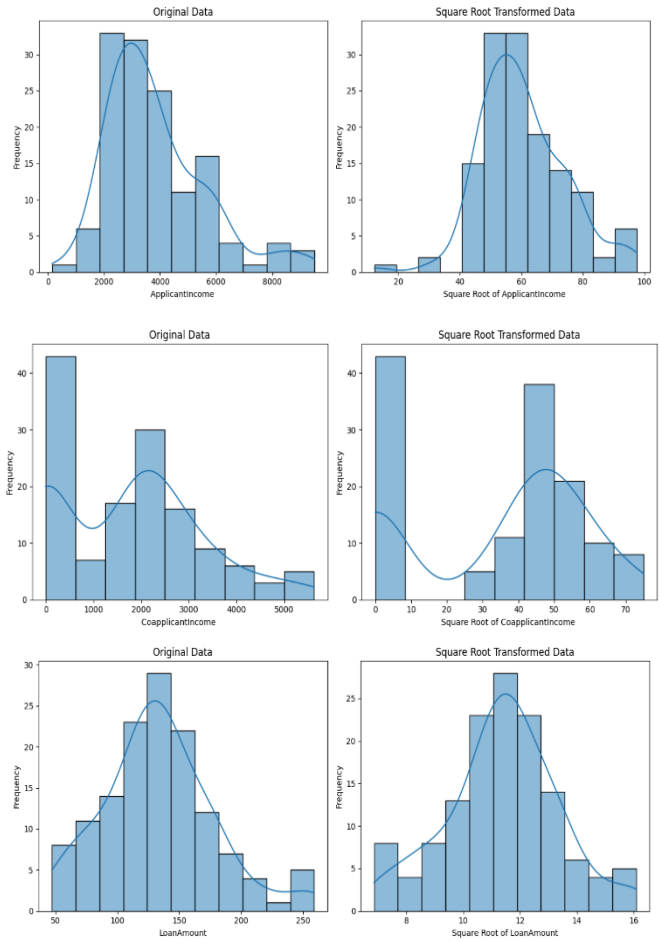


Fig 13: Square Root Transformation

- *Correlation:*

Correlation analysis is used to quantify the linear relationship between two variables, indicating both its direction and intensity. There in fig(14), doesn't seem to be any discernible relationship between ApplicantIncome, Coapplicant_Income, and Loanamount in the illustration. This implies that there is not always a linear relationship between changes in one variable and changes in the other variables.

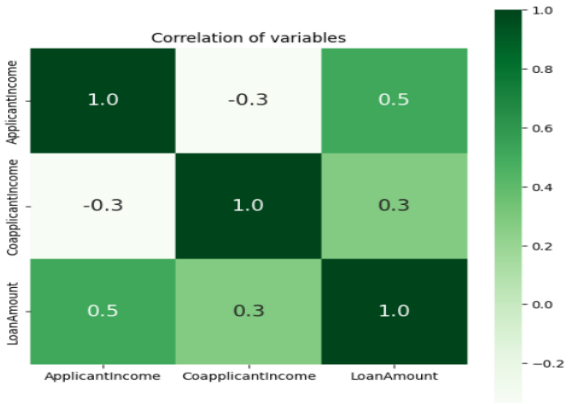


Fig.14 Correlation

- *Feature Selection:*

Feature selection is an essential process that helps machine learning models operate at their best by locating the most significant predictors for the target variable. We employ Recursive Feature Elimination (RFE) in our method, which is an iterative process of selecting features according to their importance ratings. The impact of each feature to the loan approval prediction job is carefully evaluated by integrating RFE with a logistic regression estimator. We are able to determine a subset of features shown in Fig 15, that minimises model complexity and maximises predicted accuracy through this iterative method. A more comprehensible and reliable predictive model is made possible by the chosen features, which offer insightful information about the underlying patterns and relationships within the dataset. Our goal with this feature selection method is to increase the model's capacity for generalisation and boost its effectiveness in actual loan acceptance situations.

```
Selected Features: Index(['ApplicantIncome', 'LoanAmount', 'Married', 'Property_Area_Rural',
       'Property_Area_Semiurban', 'Property_Area_Urban'],
      dtype='object')
```

Fig 15:Feature selection

- *SMOTE:*

When one class predominates over another in a situation like predicting loan acceptance, SMOTE (Synthetic Minority Over-sampling Technique) is an essential technique for resolving class imbalance in datasets. SMOTE helps balance the dataset, reducing bias towards the majority class and enhancing model performance by creating synthetic examples for the minority class. In Fig 16, the class imbalance bias is successfully reduced by using SMOTE, resulting in a more representative and trustworthy dataset that we can use to train our machine learning models. By using this method, the model's capacity to generalise across both groups is improved, which eventually leads to forecasts of loan acceptance that are more fair and accurate.

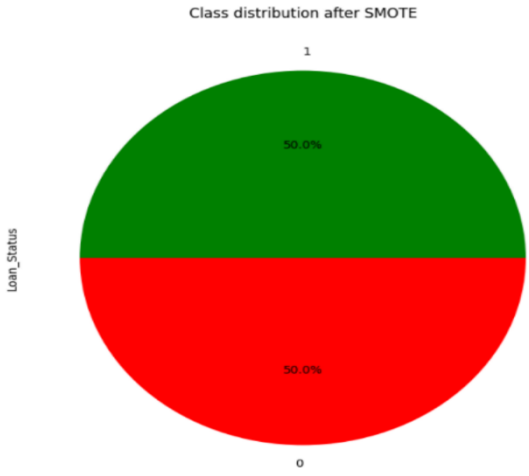


Fig 16:Loan_Status after SMOTE

*D. Model Development*

In order to ensure robustness and generalisation, machine learning models are developed and tested using training data. Their performance is then assessed, and their efficacy is evaluated on test data that hasn't been seen before. Here, the following algorithms were employed:

- Random Forest: Random Forest is an ensemble learning technique that builds several decision trees during training and produces the mean prediction (regression) or mode of the classes (classification) for each individual tree. The key formula involves aggregating predictions from individual trees:

$$\hat{y}\_RF(x) = 1/N\_trees \sum (i=1)^{(N\_trees)} f\_i(x) \quad (1)$$

- Extra Trees: Extremely Randomized Trees or Extra Trees is to Random Forest but with more randomness. Instead of selecting the best split, Extra Trees randomly select splits at each node, leading to faster training and reduced variance. The key formula for prediction is the same as Random Forest.
- Decision Tree: A straightforward, interpretable model called a decision tree recursively divides the feature space according to the values of the input features with the goal of minimising impurity in each node.
  The key formula for classification involves calculating impurity measures like Gini impurity or entropy:

$$IG(t) = 1 - \sum (i=1)^{c} p(i|t)^2 \quad (2)$$

- Logistic Regression:
  A linear model used for binary classification is called logistic regression. It uses the logistic function to represent the likelihood that a binary event will occur. The key formula for logistic regression is the sigmoid function:

$$p(y=1|x) = 1 / (1 + e^{-(\beta\_0 + \beta\_1 x\_1 + ... + \beta\_n x\_n)}) \quad (3)$$

- K-Nearest Neighbors (KNN): Data points are categorised by KNN, a non-parametric, lazy learning technique, according to the majority class of their k nearest neighbours. Finding the distance between data points and choosing the majority class among the k nearest neighbours constitute the prediction formula.

- Gradient Boosting: Gradient Boosting is an ensemble learning method in which models are built one after the other, with each model fixing the mistakes of the one before it. It makes use of gradient descent to minimise a loss function. The gradient of the loss function is used to update the model's predictions, and this is the crucial formula:

$$F\_m(x) = F\_(m-1)(x) + \gamma \sum (i=1)^{N} \nabla L(y\_i, F\_(m-1)(x\_i)) \quad (4)$$

- Voting Ensemble Method:
  A machine learning method called voting ensemble combines several models to provide predictions. A majority vote or an average is used to decide the final forecast, which is based on the weighted predictions of each model. When compared to separate models, it frequently results in increased accuracy and resilience.

$$y\hat{}=argmax_{ji=1} \sum n1(y\hat{}i=j) \quad (5)$$

## 4. RESULTS & DISCUSSION

We found that the models that were used are Decision Trees, Random Forest, Logistic Regression, Extra Trees, Gradient Boosting and K-Nearest Neighbors performed differently. Every algorithm performed differently, as shown below:

### A. Random Forest:

With an accuracy of 95.56%, the Random Forest model proved to be a reliable classifier of data points. The model successfully learned from the training data, as seen by its testing accuracy of 95.556%, but with a training accuracy of 98.325%.

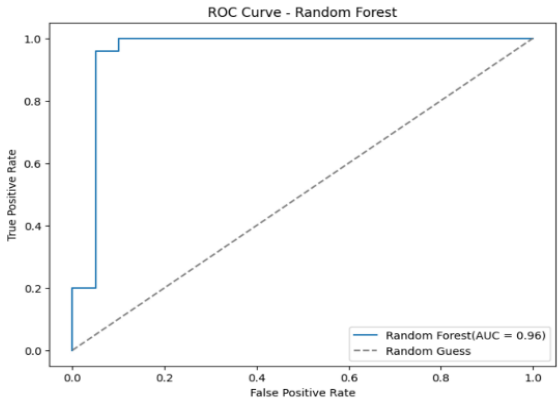


Fig.17(a) ROC Curve-Random Forest

The confusion matrix shows the results of 90 loan applications using the model. It correctly identified 3 applications that were denied (True Negatives) and 78 accepted applications (True Positives). But four rejected applications were mistakenly marked as approved (False Positives). With a precision of 95.3%, recall of 93.9%, and accuracy of 87.8%.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.95 | 0.95 | 20 |
| 1 | 0.96 | 0.96 | 0.96 | 25 |
| accuracy | | | 0.96 | 45 |
| macro avg | 0.95 | 0.95 | 0.96 | 45 |
| weighted avg | 0.96 | 0.96 | 0.96 | 45 |

Fig.17(b) Metrics for Random Forest

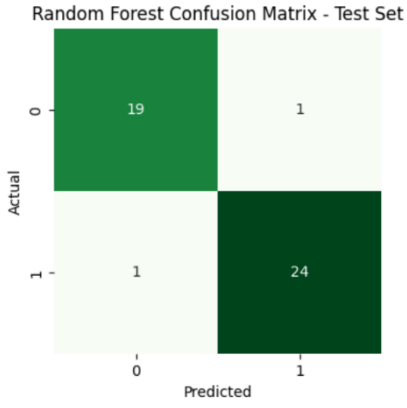


Fig.17(c) Confusion matrix of Random Forest

*B. Extra Tree Classifier:*

Possesing a respectable accuracy of 91.11%, the Extra Trees Classifier demonstrated its efficacy in classifying tests. Using the training data, the model showed strong learning skills with a training accuracy of 92.1%. The marginally reduced testing accuracy of 91.11% indicates a modest degree of overfitting and validates the model's ability.

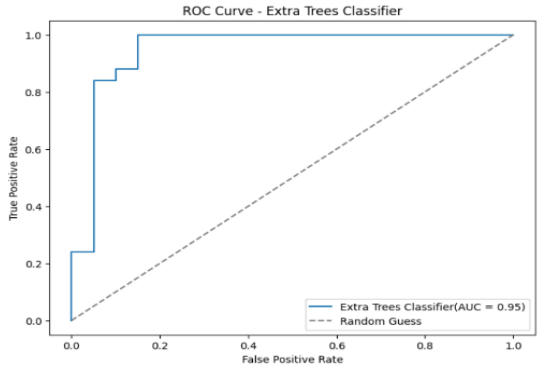


Fig.18(a) ROC Curve-Extra Tree Classifier

Nineteen loan applications were accurately categorised as accepted (True Positives) in the confusion matrix, while twenty-four were correctly labelled as refused (True Negatives). One false positive, on the other hand, occurred when an application that was denied was mistakenly marked as granted. With a flawless recall and a precision of 95%, the overall accuracy is 97.7%. This indicates that the model classifies applications as accepted or refused with minimum misclassifications.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.85 | 0.89 | 20 |
| 1 | 0.89 | 0.96 | 0.92 | 25 |
| accuracy | | | 0.91 | 45 |
| macro avg | 0.92 | 0.91 | 0.91 | 45 |
| weighted avg | 0.91 | 0.91 | 0.91 | 45 |

Fig 18(b) Metrics for Extra Tree Classifier

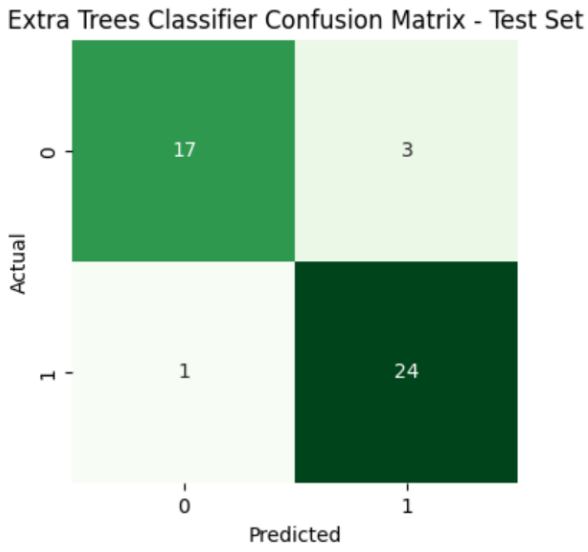


Fig.18 (c) Confusion matrix of Extra Tree Classifier

*C. Logistic Regression:*

Its 82.22% accuracy suggests that the Logistic Regression model is a good fit for jobs involving binary categorization. In training, the model performed reasonably well, with a training accuracy of 79.33%. Reliability of the model for prediction tasks is reinforced by its testing accuracy of 82.223%, which indicates high generalisation to unknown data.

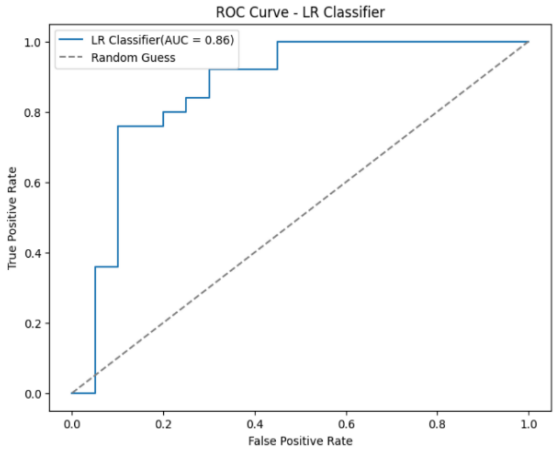


Fig.19 (a) ROC Curve- LR Classifier

The accuracy, recall, and F1-score table's highlighted values show excellent performance for both classes. The accuracy and recall of Class 1 (presumably accepted) are 0.95 and 0.90, respectively, and Class 0 (likely refused) has 0.90 and 0.76, respectively. Overall, the model shows good recall and accuracy, indicating that it can classify loan applications well. Nonetheless, the somewhat reduced recall for Class 0 could suggest a greater probability of false positives.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.90 | 0.82 | 20 |
| 1 | 0.90 | 0.76 | 0.83 | 25 |
| accuracy | | | 0.82 | 45 |
| macro avg | 0.83 | 0.83 | 0.82 | 45 |
| weighted avg | 0.84 | 0.82 | 0.82 | 45 |

Fig.19(b) Metrics for LR Classifier

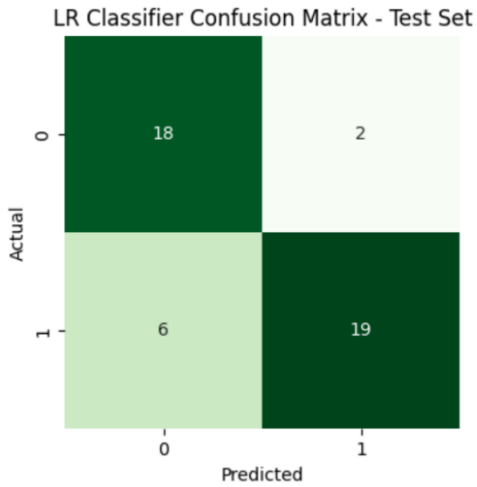


Fig.19 (c) Confusion matrix of LR Classifier

*D. KNN Classifier:*

The K-Nearest Neighbours (KNN) model is good at identifying data items according to their closest neighbours; it reached its maximum accuracy of 80.00%. The model performed somewhat worse on unknown data, with a testing accuracy of 60.0%, despite a training accuracy of 70.95%. This shows that although the model performs reasonably on the test set, it may have overfit the training set by a little amount.

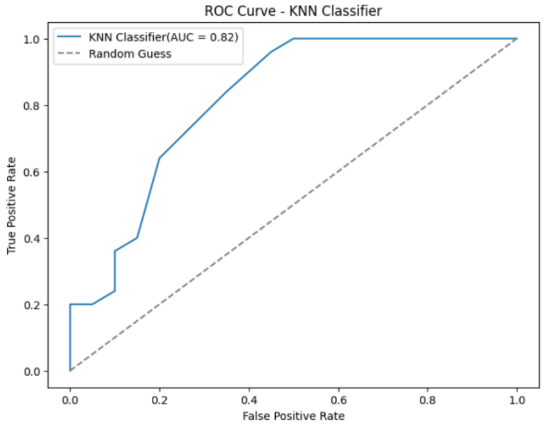


Fig.20 (a)ROC Curve-KNN Classifier

Based on the highlighted values in the confusion matrix. It accurately identifies 34 loan applications that are denied (True Negatives) and 61 loan applications that are granted (True Positives). Nevertheless, it incorrectly flags 8 accepted applications as rejected (False Negatives) and 7 rejected applications as approved (False Positives). The model performs fairly in categorising loan approvals.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.53 | 0.85 | 0.65 | 20 |
| 1 | 0.77 | 0.40 | 0.53 | 25 |
| accuracy | | | 0.60 | 45 |
| macro avg | 0.65 | 0.62 | 0.59 | 45 |
| weighted avg | 0.66 | 0.60 | 0.58 | 45 |

Fig.20(b) Metrics for KNN Classifier

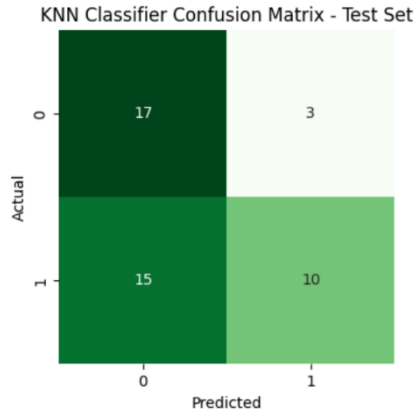


Fig.20 (c) Confusion matrix of KNN Classifier

*E. DecisionTree Classifier:*

The Decision Tree model demonstrated its efficacy in classifying tests with an accuracy of 88.89%. A testing accuracy of 88.887% and a training accuracy of 93.85% show that the model performs well in generalisation, showing that it can make predictions on data that hasn't been seen before.

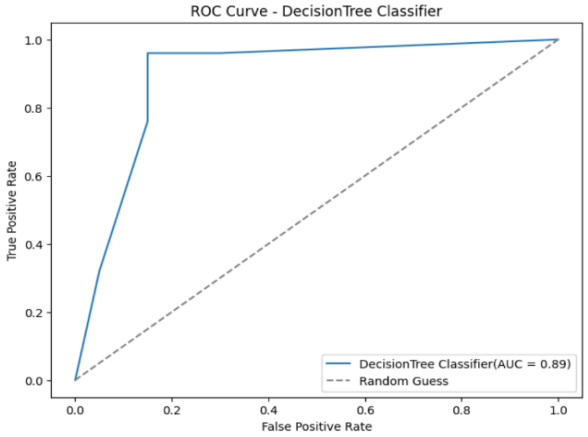


Fig.21 (a) ROC Curve-Decision Tree

The table with precision, recall, and F1-scores demonstrates both classes' excellent performance. Precision for Class 1 (presumably accepted) is 0.89, recall is 0.85, and recall is 0.92. Class 0 (likely rejected) has precision of 0.88 and recall of 0.92. Class 0 shows somewhat higher recall, indicating better identification of rejected applications. These values reflect effective classification of loan applications.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.85 | 0.87 | 20 |
| 1 | 0.88 | 0.92 | 0.90 | 25 |
| accuracy | | | 0.89 | 45 |
| macro avg | 0.89 | 0.89 | 0.89 | 45 |
| weighted avg | 0.89 | 0.89 | 0.89 | 45 |

Fig.21 (b) Metrics for Decision Tree

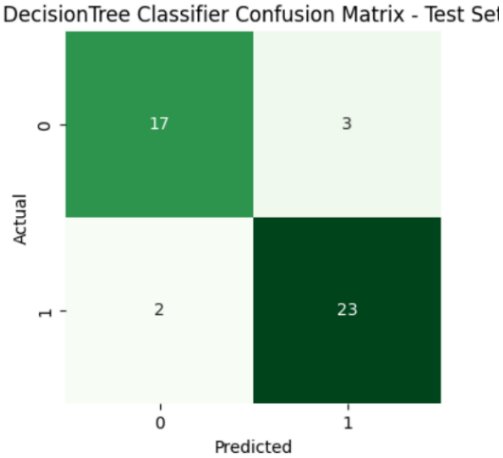


Fig.21 (c) Confusion matrix of Decision Tree

*F. GradientBoosting Classifier:*

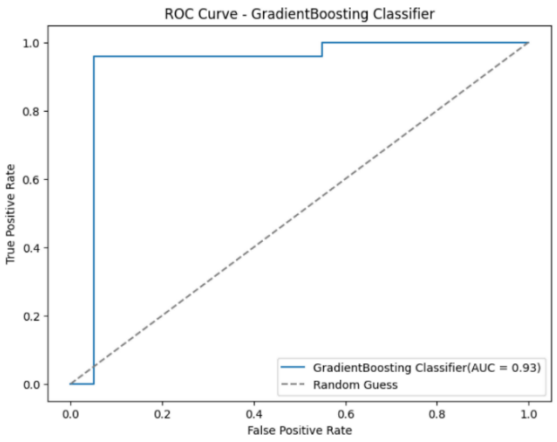


Fig.22(a) ROC Curve-GradientBoost Classifier

The Gradient Boosting model demonstrated its efficacy in loan approval prediction with an accuracy of 93.33%. As evidenced by its testing accuracy of 93.332% and training accuracy of 100.0%, respectively, the model performs well in both training and generalisation, indicating its resilience to complex datasets.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.95 | 0.93 | 20 |
| 1 | 0.96 | 0.92 | 0.94 | 25 |
| accuracy | | | 0.93 | 45 |
| macro avg | 0.93 | 0.94 | 0.93 | 45 |
| weighted avg | 0.93 | 0.93 | 0.93 | 45 |

Fig.22(b) Metrics of GradientBoost Classifier

The effectiveness of the model for predicting loan acceptance is examined based on the values that are emphasised in the confusion matrix. Only one rejected loan was mistakenly classified as approved, out of the 19 approved and 24 rejected loans that were accurately classified. For authorised loans, the model achieves excellent accuracy (97.7%), precision (95%), and potentially perfect recall (1) assuming no false negatives.

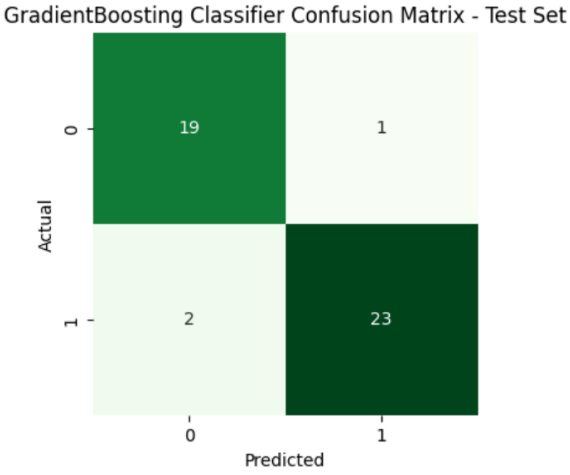


Fig.22(c) Confusion matrix of GradientBoost Classifier

In order to highlight, With an accuracy of 95.56%, Random Forest stood out among the group of models employed to forecast loan acceptance. It demonstrated resilience during both the training (98.325%) and testing (95.556%) stages of the process. Closely behind, Extra Trees Classifier showed great predictive strength with an accuracy of 91.11%, and Gradient Boosting, though doing effective, showed its effectiveness with an accuracy of 93.33%. On the other hand, the accuracies of the Decision Tree, K-Nearest Neighbours, and Logistic Regression models were much lower at 88.89%, 80.00%, and 82.22%, respectively. As shown in fig 23.

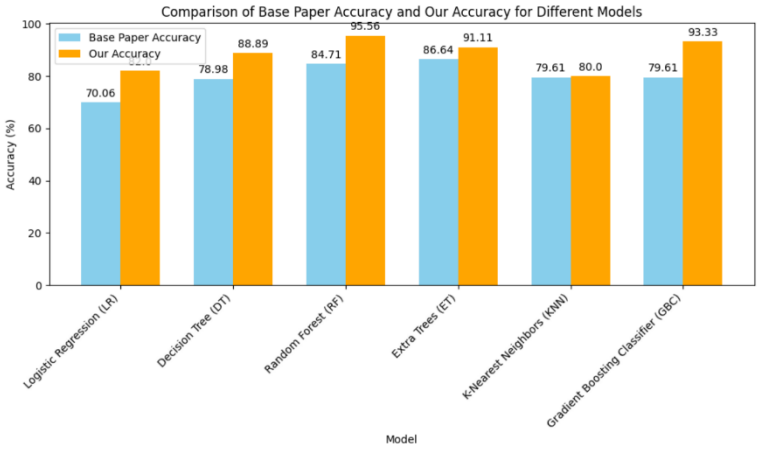


Fig.23 Comparing Accuracies

*G. Voting Ensemble Method:*

In order to increase the overall accuracy, we will integrate the predictions of several machine learning models in our next stage, which is the implementation of a voting ensemble approach.

After careful examination, Random Forest was clearly the best performer, with an astounding accuracy of 95.56%. In training and testing, this model demonstrated exceptional robustness, with accuracy values of 98.325% and 95.556%, respectively. With an accuracy of 91.11%, the Extra Trees Classifier demonstrated noteworthy predictive capability, trailing closely behind. Likewise, Gradient Boosting demonstrated its efficacy with a 93.33% accuracy rate.

By combining Random Forest, Extra Trees Classifier, and Gradient Boosting, we will use ensemble learning to take advantage of the advantages of these high-performing models. In order to improve the accuracy of loan approval forecasts, this strategy seeks to use the combined predictive potential of models. Using the best-performing models and the ensemble approach, we were able to obtain an astounding accuracy of 95.55%. This suggests a notable improvement in the accuracy of loan approval forecasts.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.95 | 0.95 | 20 |
| 1 | 0.96 | 0.96 | 0.96 | 25 |
| accuracy | | | 0.96 | 45 |
| macro avg | 0.95 | 0.95 | 0.96 | 45 |
| weighted avg | 0.96 | 0.96 | 0.96 | 45 |

Fig.24 (a) Metrics for Voting Ensemble Model

The confusion matrix's highlighted values suggest that the model is performing satisfactorily in classifying loan applications. The accuracy is 94% with 78 true positives and 24 genuine negatives. With a precision of 95% for loans that have been granted, there is a high percentage of accurate forecasts among loans with this approval status. Furthermore, for loans that are authorised, recall is 100% perfect, providing there are no false negatives. Four false positives, on the other hand, show that some loans were incorrectly identified as approved despite being refused.

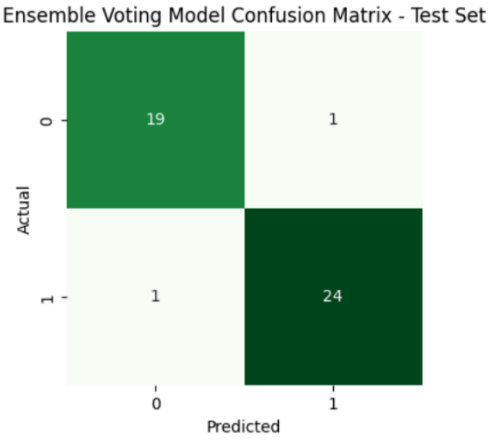


Fig.24(b) Confusion Matrix of Voting Ensemble Model

USER INTERFACE:

In order to provide smooth communication between users and our prediction models, the user interface (UI) is essential. A user-friendly and intuitive interface facilitates easy access to and interpretation of the model predictions by all stakeholders, including bank management and loan applicants. Our user interface (UI) provides clear feedback mechanisms and interactive visualisations to enable users to make educated decisions about loan acceptance.as follow:

## Home Page:


Fig.25 Home Page

## Prediction Form:

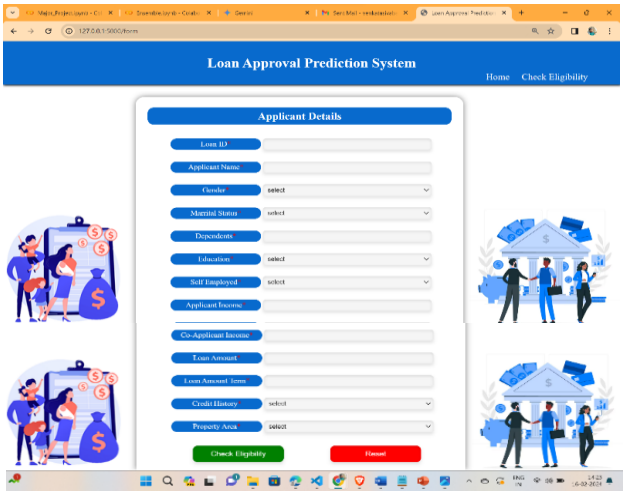


Fig.26 Prediction Form

## OutPut Pages:

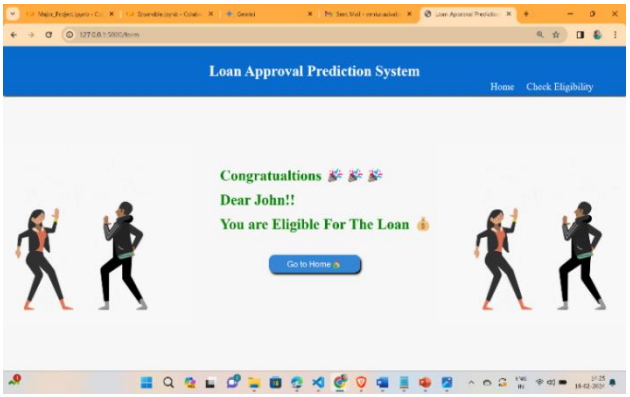


Fig.27 Approval Status

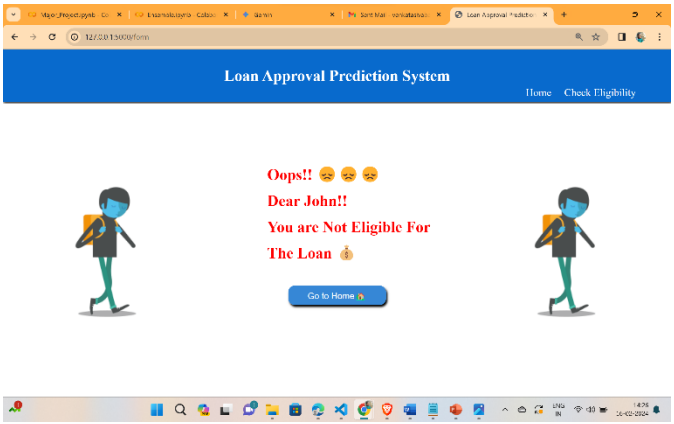


Fig.28 Disapproval status

## 5. CONCLUSION & FUTURE SCOPE

In summary, our initiative has shed light on the significant contribution that machine learning (ML) approaches have made to transforming the process of predicting loan approval. We have established a solid basis for our predictive models by carefully examining our dataset and utilising sophisticated preprocessing methods, such as managing null values, outliers, square root transformation, correlation analysis, feauture selection and SMOTE.

Going forward, our model achieved an amazing accuracy of 95.56%—serve as a testament to the ability of ML to provide incredibly precise predictions. This represents a notable improvement above the baseline accuracy of 87.26% from the previous base study and highlights the enormous progress made possible by our rigorous methodology.

Table1: Comparing Accuracies

| | Accuracy(%) |
|---|---|
| Base Accuracy | 87.26 |
| Our Accuracy | 95.56 |

Our research opens up a world of opportunities for improving loan approval prediction models in the future. Subsequent investigations may explore further into the field of deep learning, utilising neural networks capacity to address complex patterns in loan data. In addition, to guarantee flexibility in response to changing patterns and dynamics in loan application and repayment behaviour, ongoing model monitoring systems will be necessary. We are steadfast in our research to advance ML innovation and bring about revolutionary change in the field of financial decision-making processes as we move forward.

## 6. REFERENCES


[1] Gopinath, Mahankali, K. Srinivas Shankar Maheep, and R. Sethuraman. 2021. “Customer Loan Approval Prediction Using Logistic Regression.” Advances in Parallel Computing. https://doi.org/10.3233/apc210103.

[2] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,“Loan Prediction by using Machine Learning Models”, International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019.

[3] A. Uzair, T. Aziz, H. Ilyas, S. Asim, B. N. Kadhar, "An Empirical Study on Loan Default Prediction Models" Journal of Computational and Theoretical Nanoscience, Volume 16, Number 8, August 2019, pp. 3483-3488(6). DOI: https://doi.org/10.1166/jctn.2019.8312

[4] M. Madaan et al. "Loan default prediction using decision trees and random forest: A comparative study" IOP Conf. Ser.: Mater. Sci. Eng. 2014. doi: 10.1088/1757-899X/1022/1/012042.

[5] Alsaleem, M. Y., & Hasoon, S. O. (2020). Predicting bank loan risks using machine learning algorithms. AL-Rafidain J. Comput. Sci. Math., 14(1), 149–158.

[6] Ramachandra, H. V., G. Balaraju, R. Divyashree, and Harish Patil. 2021. “Design and Simulation of Loan Approval Prediction Model Using AWS Platform.” 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/esci50559.20219397049. Data set used https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset/.%20.

[7] Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban. 2021. “Prediction ofModernized Loan Approval System Based on MachineLearning Approach.” 2021 International Conference on IntelligentTechnologies(CONIT).https://doi.org/10.1109/conit51480.2021.9498475.

[8] S.m., Karthikeyan, S. M. Karthikeyan, and Pushpa Ravikumar. 2021. “A Comparative Analysis of Feature Selection for Loan Prediction Model.” InternationalJournalofComputerApplications.https://doi.org/10.5120/ijca2021920992.

[9] Hassan, Amira Kamil Ibrahim and Ajith Abraham. "Modeling consumer loan default prediction using ensemble neural networks" , 2013 International Conference On Computing, Electrical And Electronic Engineering (ICCEEE). IEEE, 2013.

[10] .Nitesh Pandey et al. (2022). "Loan Approval Prediction using Machine Learning Algorithms Approach.", IRJMETS, this paper achieved an accuracy of 78.3% using Logistic Regression, Decision Trees, and KNN. It discusses limitations in feature engineering and imbalanced datasets.

[11] J. Tejaswini (2022). "Accurate Loan Approval Prediction Based on Machine Learning Approach.", IRJMETS, this paper achieved an accuracy of 88.1% using XGBoost. It emphasizes the importance of careful data pre-processing and feature engineering for optimal performance.

[12] Anant Shinde et al. (2022). "Loan Prediction System Using Machine Learning." Published in the field of Finance by IEEE, this paper compares Logistic Regression, Decision Tree, and Random Forest models. It concludes that Decision Tree and Random Forest outperform Logistic Regression due to their non-linearity handling capabilities.

[13] Dharavath Sai Kiran. Avula et al. (2023). "Loan Approval Prediction using Adversarial Training and Data Science." Within the domain of Finance, this paper discusses a model that produced low accuracy values and could only handle minimum-sized data.

[14] Zhu L, Qiu D, Ergu D, Ying C, Liu K (2019) A study on predicting loan default based on the random forest algorithm. Procedia Comput Sci 162:503–513.

[15] Nigmonov A, Shams S, Alam K (2022) Research Macroeconomic determinants of loan defaults: evidence from the U.S. peer-to-peer lending market. Res Int Bus Finan 59:101516.

[16] Lee JW, Lee WK, Sohn SY (2021) Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Syst Appl 168:114411.

[17] Lim S-J, Thiel C, Sehm B, Deserno L, Lepsien J, Obleser J (2022) Distributed networks for auditory memory differentially contribute to recall precision. NeuroImage 256:119227.

[18] Fontem B, Smith J (2019) Analysis of a chance-constrained new product risk model with multiple customer classes. Eur J Oper Res 272(3):999–1016 23. Bianco S, Mazzini D, Napoletano P, Schettin R (2019) Multitask painting categorization by deep multibranch neural network. Expert Syst Appl 135:90–101.

[19] Wang L, Chen Y, Jiang H, Yao J (2020) Imbalanced credit risk evaluation based on multiple sampling, multiple kernel fuzzy self-organizing map and local accuracy ensemble. Appl Soft Comput 91:106262

[20] Vuttipittayamongkol P, Elyan E, Petrovski A (2021) On the class overlap problem in imbalanced data classification. Knowl-Based Syst 212:106631 26. Papouskova M, Hajek P (2019) Two stage consumer credit risk modelling using heterogeneous ensemble learning. Decis Support Syst 118:33–45.

[21] Ashofteh A, Bravo JM (2021) A conservative approach for online credit scoring. Expert Syst Appl 176:114835.