Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 07 July-2024, Page No.- 4532-4549

DOI: 10.47191/etj/v9i07.24, I.F. – 8.227

Enhancing Bank Loan Approval Efficiency Using Machine Learning: An Ensemble Model Approach

Md. Rezaul Islam[1], Md. Shariful Islam[1], Sharmin Shama[2], Aniruddha Islam Chowdhury[2], Md. Masudul Hasan Lamyea[2]

[1]Lecturer, Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh

[2]Student, Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh

ABSTRACT: Lending is a major source of income for banks, but identifying worthy borrowers who will consistently repay loans is a constant problem. From a pool of loan applicants, conventional selection procedures frequently fail to find the most qualified individuals. To make loan applications faster, we created a new system that uses machine learning to automatically find people who qualify for loans. This comprehensive analysis involves data preprocessing, effective data balancing using SMOTE, and the application of various machine learning models, including Decision Trees, Support Vector Machines, K-Nearest Neighbors, Gaussian Naive Bayes, AdaBoost, Gradient Boosting, Logistic Regression, and advanced deep learning models like recurrent neural networks, deep neural networks, and long short-term memory models. We thoroughly evaluate the models based on accuracy, recall, and F1 score. Our experimental results demonstrate that the Extra Trees model outperforms its counterparts. Furthermore, we achieve a significant 0.62% increase in accuracy over the Extra Trees model by using an ensemble voting model that combines the top three machine learning models to predict bank loan defaulters. An intuitive desktop application has been developed to enhance user engagement. Remarkably, our findings indicate that the voting-based ensemble model surpasses both current state-of-the-art methods and individual ML models, including Extra Trees, with an impressive accuracy of 87.26%. Ultimately, this innovative system promises substantial improvements and efficiency in bank loan approval processes, benefiting both financial institutions and loan applicants.

KEYWORDS: Loan Prediction, EDA analysis, Machine learning algorithms, Confusion matrices, etc.

1. INTRODUCTION

The banking sector is crucial for maintaining a country's financial stability, prompting most nations to establish extensive regulatory frameworks. Lending is a primary operation for banks, generating significant assets from interest revenue [1]. However, the current loan approval process has significant shortcomings in terms of accuracy and efficiency due to its heavy reliance on manual procedures. These procedures place the responsibility of determining an applicant's eligibility and potential default risk on specific bank managers. Such manual processes can have severe ramifications, from financial losses for banks to catastrophic economic disruptions. Historically, selecting creditworthy customers from a large pool of applicants has been a significant challenge in the loan approval process. Accurate prediction of loan defaults has become increasingly crucial in today's financial systems, with inaccurate forecasts leading to wide-ranging repercussions, including banking crises [2]. The manual identification of loan defaulters is particularly challenging given the complexity of the modern banking environment and the rising demand for credit [3]. Machine learning (ML) algorithms, which enable computers to interpret patterns and predict outcomes based on data, have emerged as a viable method for assessing loan default risk. These algorithms analyze clients' transaction histories and social profiles, identifying common traits and activity patterns that provide insights into future repayment tendencies. Additionally, deep learning methods, such as deep neural networks, have gained popularity due to their superior ability to capture complex, non-linear relationships in the data, enhancing the assessment of a customer's likelihood to default on a loan [4]. As a result, ML-based methodologies have been extensively researched to provide a comprehensive evaluation of loan default risk [5]. The importance of banks in assessing credit risk cannot be overstated. Lenders must evaluate applicants' credit histories to distinguish probable defaulters from non-defaulters [6]. Due to the complexity of this task and the growing loan demand, manually creating a reliable prediction system is a formidable challenge. Previous research has shown that a single classifier is inadequate for real-world deployment. Consequently, some studies have explored the use of machine learning algorithms for forecasting loan defaults. To improve loan approval systems, we propose an ensemble strategy that leverages the best machine learning techniques. Loan approval is a critical process for financial organizations,

involving a thorough assessment of various factors such as citizenship, income, and social status to determine loan eligibility. Prompt loan repayment is essential for banks to operate profitably, making an accurate evaluation of borrowers' repayment ability crucial [7]. Several projects have investigated the use of machine learning to enhance loan approval prediction, finding that Random Forest classifiers outperform traditional Decision Trees [8]. For financial organizations, loan approval is an essential process that involves a thorough assessment of several variables, such as citizenship, income, and social position, in order to establish loan eligibility. For banks to operate profitably, prompt loan payback is essential, which makes an accurate evaluation of borrowers' repayment ability necessary. Machine learning has been investigated in a number of projects to improve loan approval prediction, and it was discovered that Random Forest classifiers performed better than conventional Decision Trees [9]. A stacking-based methodology that greatly enhanced the acceptance of joint loans [10]. Created an ensemble model that increased the 80% to 94% prediction accuracy [11]. Utilized cutting-edge machine learning methods, in particular Random Forest, to forecast nonperforming loans with an emphasis on developing economies. Previous studies in this field have put out a number of models, but there hasn't been a thorough examination that takes into account both machine learning and deep learning techniques. Furthermore, the majority of currently available solutions have not shown to be accurate enough to be used in real-world circumstances [12]. The following are the contributions made by this paper:

- Using a combination of ensemble machine learning techniques, we have developed a predictive model for loan approval that surpasses the performance of individual machine learning methods. The model we propose incorporates nine distinct machine learning algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), Gaussian Bayesian (GB), Random Forest (RF), AdaBoost (AdB), Extra Trees (ET), K- Nearest Neighbors (KNN), and Gradient Boosting (GB).

1. Related Works

Machine learning algorithms have demonstrated efficacy in the realm of bank loan prediction, aligning with their success across various domains such as data mining, cybersecurity, game theory optimization, natural language processing, traffic management, and medical applications like brain tumors, breast cancer, leaf disease, and human behavior analysis (Alzubi et al., 2022). Bhargav and Sashirekha (2023) employed state-of-the-art Random Forest classifiers to assess multiple machine learning strategies for loan acceptance prediction, utilizing loan prediction datasets from the Kaggle library for accuracy and loss testing. In a sample of 20 cases, the RF approach surpassed the conventional Decision Tree, achieving 67.28% precision and 32.71% loss, as opposed to 79.44% precision and 21.03% loss. Statistical analysis utilizing an independent sample T-test yielded a p-value of 0.33, indicating no significant variances between the methodologies at a 95% confidence level. The study concluded that RF exhibited superior loan approval prediction performance compared to Decision Trees. Dasari et al. (2023) developed an ensemble model that integrates multiple machine learning techniques with voting classifiers and bagging methods, focusing on predicting loan eligibility. This model outperformed existing techniques by enhancing accuracy and reducing human labor and processing time requirements. A performance boost from 80% to 94% was observed when comparing experimental results to the previous model. Abdullah et al. (2023) employed diverse machine learning approaches to predict nonperforming loans in financial institutions in developing countries. Advanced machine learning models, specifically random forest, outperformed linear techniques with 76.10% accuracy while analyzing data from 322 banks across 15 countries. When it came to predicting nonperforming loans, bank diversity prevailed over macroeconomic conditions as the primary predictor. In order to approve financial institution risks, In their study, Wang et al. (2023) introduced a model based on the stacking technique. The optimal model is chosen by comparing performance. Additionally, they used deep learning to create a bank clearance model using unbalanced data, extracting features with CNN and balancing samples with counterfactual augmentation. Experiments on actual data show that optimizing the car financing prediction model based on bank model attributes increased joint loan approval by around 6%. The objective of evaluating bank loan risks more accurately through the use of traditional methods led to an analysis of the evaluation of machine learning models. Alsaleem and Hasoon (2020) [14] discovered that in the realm of bank loan risk classification, Multilayer Perceptron (MLP) exhibited superior accuracy compared to RF, BayesNet, Naive Bayes (NB), and DTJ48 algorithms. The evaluation of the model's efficacy was based on a dataset comprising 1000 loans and their repayment statuses, utilizing standard metrics. Wang et al. (2019) [15] employed deep learning methodologies to assess consumer credit risk in scenarios where e-commerce platforms extend unsecured loans to facilitate customer transactions. Supriya et al. (2019) [16] utilized LR, DT, and GB techniques for the purposes of predicting loan risks. Sun & Vasarhalyi (2021) [17] utilized a neural network that underwent training via a backpropagation learning algorithm. The primary objective was to aid a credit card issuer in predicting the likelihood of credit card defaults through the development of a predictive system. Additionally, he looked at how well deep learning might forecast credit card risks. Two machine learning models were presented by Madaan et al. (2021) [18] that examined specific characteristics that are more helpful in deciding whether to approve a loan for a person. Their approach might help

banking authorities choose the right candidates from a list of loan applications. Anand et al. (2022) [19] collected 850 records in order to use machine learning models for safe banking to predict loan behavior. Decision Trees, Random Forest, Extra Trees, CatBoost (CB), Light Gradient Boosting (LGB), and Extreme Gradient Boosting (EGB) were all incorporated into their model. The outcomes they presented illustrated that RF and ET could predict loan approval with enhanced precision. In order to ascertain loan eligibility, Kumar et al. (2022) [20] collected a dataset of 614 entries from a public repository. A variety of machine learning methods, such as RF, DT, KNN, SVM, and DT with AdaBoost, were put into practice. Their findings indicated that the decision tree ensemble model utilizing the AdaBoost technique provided superior accuracy. Similarly, (Dosalwar et al., 2021) [21] acquired a Kaggle dataset for loan prediction. Several models, LR, DT, KNN, NB, RF, SVM, and XGB, were trained and evaluated. It was observed that the LR model accurately predicted loan eligibility. Alsaleem and Hasoon (2020) [14] employed a dataset comprising 1000 instances and 11 attributes from the UCI repository to predict loans. Five models were utilized for loan prediction: DT J48, Bayes Net (BN), NB, RF, and MLP. The findings indicated that MLP offered more precise predictions regarding loan availability. Furthermore, Blessie and Rekha (2019) [22] collected loan data from a Kaggle dataset [23]. They utilized four classifiers (LR, DT, SVM, and NB) for loan prediction. Their investigation demonstrated that NB produced more precise forecasts concerning credit availability. In Table 1, a comparison of different existing works' results is presented.

Table 1: Comparison of Different Existing Work's Results.

Model for Loan Prediction used in existing work (Accuracy %)														**Paper**	**Year**
RF	**ET**	**CB**	**LGB**	**EGB**	**DT**	**KNN**	**SVM**	**DT+AB**	**LR**	**NB**	**XGB**	**BN**	**MP**		
79.4														(Bhargav & Sashirekha,2023)	2023
76.1														(Abdullah et al.,2023)	2023
85.6	**86.2**	84.9	84	83.9										(Anand et al.,2022)	2022
72					69	59	70	**84**						(Kumar et al.,2022)	2022
77.3					66.2	61.9	65		**78.5**	77.9	77.3			(Dosalwar et al.,2021)	2021
78.5					73.5					77.5		75	**80**	(Alsaleem & Hasoon,2020)	2020
					71.9		78.9			**80.4**				(Blessie & Rekha,2019)	2019

2. METHODOLOGY

In the depicted Fig. 1, the proposed model's overarching design is presented. The objective of this study is to identify the optimal model and seamlessly integrate it into a user interface for predicting loan eligibility. A web application has been developed for assessing whether a customer qualifies for a loan. The illustration in the figure outlines the preprocessing steps applied to the loan data, encompassing the division of data into training and testing sets. Following this, nine machine learning algorithms undergo training, leading to the creation of an ensemble model comprising the top three performing algorithms. The model's efficacy is then assessed in terms of accuracy, precision, recall, and F1 score. Notably, various features, such as Gender, marital status, dependents, education, self-employed status, applicant income, co-applicant income, loan amount, loan amount term, credit history, and property, are deemed crucial in determining loan approval.

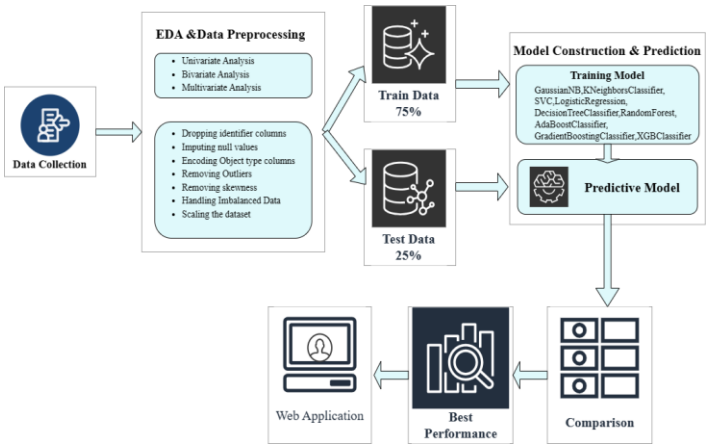


Fig. 1. Workflow diagram of the proposed system.

3.1 Understanding the Problem Statement

Loans represent a significant revenue stream for banks, constituting the primary source of their profits. Despite thorough verification and assessment processes, there's no assurance that selected candidates are the most suitable. Manual operations in this regard are time-consuming. However, leveraging machine learning, we can automate the entire assessment process, predicting the safety of potential borrowers. This automation benefits both the bank and applicants. The bank aims to streamline the loan eligibility process in real-time, utilizing customer details provided in online application forms such as gender, marital status, education, number of dependents, income, loan amount, credit history, among others. To achieve this, they've supplied a dataset to pinpoint eligible customer segments, enabling targeted marketing efforts. This task involves binary classification, where the objective is to predict the "Loan Status" label. "Loan Status" can be either "Yes" if the loan is approved or "No" if it's not. By training our model on the provided dataset, we aim to predict the "Loan Status" for the test dataset.

3.2 Dataset Features

A dataset for loan prediction was gathered from Kaggle. Table 2 presents the details of this loan dataset, sourced from Kaggle, including a range of features outlined within the same table. A company aims to automate the loan eligibility process in real-time by leveraging customer details provided during the online application process. To do this, it has published this dataset, which includes 614 rows and 13 columns.

Table 2: Introducing Features

Attribute Name	Details of Attribute	Data Type
Loan ID	Loan reference number	Integer
Gender	Applicant gender	Character
Married	Applicant marital status	Character
Dependents	Number of family members	Integer
Education	Graduate/ Not Graduate	String
Self Employed	Applicant employment status	Character
Applicant Income	Applicant's monthly salary/income	Integer
Co-applicant Income	Additional applicant's monthly	Integer
Loan Amount	Loan amounts in thousands	Integer
Loan Amount Term	The loan's repayment period (in days)	Integer
Credit History	Records of previous credit history	Integer
Property Area	The location of property	String
Loan Status	Status of loan	String

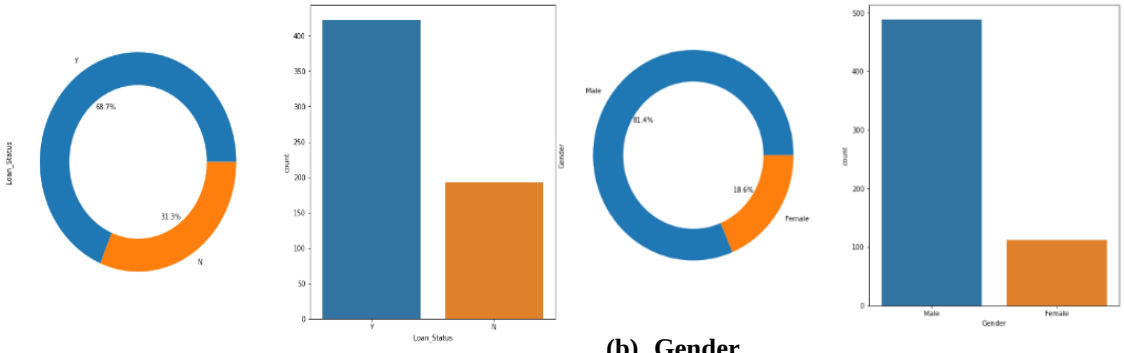
3.3 Dataset Visualization

The dataset under consideration comprises 614 rows and 13 columns, featuring an identifier column labeled "loan id" that necessitates removal for analytical purposes. Among the columns, 10 contain object-type data, while the remaining columns consist of numerical values. Notably, an inherent skewness is observed in the dataset, signaling a potential need for normalization or transformation. Moreover, an imbalance in the dataset is evident, raising concerns about the potential impact on model training and predictive accuracy. A closer examination reveals that the "Loan amount" and "Credit history" columns exhibit counts below the dataset's total,

indicating the presence of missing values that warrant attention during preprocessing. Further exploration into the numerical columns reveals interesting insights. The mean values of "ApplicantIncome", "CoapplicantIncome", and "LoanAmount" exceed their respective medians, indicative of right skewness. Conversely, the mean values in the last two columns are lower than the medians, signaling left skewness. This distribution asymmetry suggests a non-uniform spread of data. Examining the descriptive statistics, disparities in the minimum and maximum values, as well as the interquartile ranges, hint at the presence of outliers. The variability in these measures across different columns implies a need for robust outlier detection and handling techniques during data preprocessing. In summary, a comprehensive analysis of the dataset highlights the importance of addressing missing values, imbalances, skewness, and outliers. These considerations are crucial for ensuring the reliability and accuracy of any subsequent machine learning models applied to the dataset.

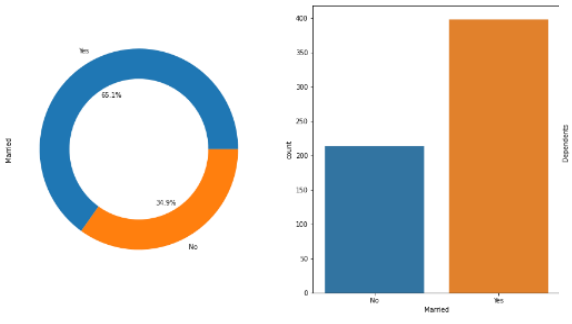
3.3.1 (a) Univariate Analysis

In Fig. 2, the dataset reveals crucial insights into various aspects of loan applications. Firstly, Fig. 2(a) highlights a significant imbalance in label classes, with unequal counts across categories. Moving forward, Fig. 2(b) shows that the number of male applicants exceeds that of female applicants by approximately fourfold, indicating a gender disparity in loan applications. Similarly, Fig. 2(c) demonstrates a higher proportion of married applicants compared to their unmarried counterparts. Further analysis in Fig. 2(d) provides insights into the dependency status of applicants, revealing that the majority have zero dependents, while those with three or more dependents are notably scarce. This distribution suggests a concentration of loan applicants with fewer dependents. The influence of educational background on loan applications is depicted in Fig. 2(e), with a noticeable prevalence of graduate applicants over non-graduates. Additionally, Fig. 2(f) indicates a lower participation of self-employed individuals in loan applications, suggesting a potential trend or simply a smaller population of self-employed applicants. Examining loan terms in Fig. 2(g) reveals that the majority of applicants opt for a 360-month loan term, with a notable scarcity of those choosing shorter terms. Fig. 2(h) indicates a significant number of loan applicants with credit history. Geographically, Fig. 2(i) suggests that a considerable number of loan applicants reside in semi-urban areas, while rural areas have the least representation. This spatial distribution offers valuable insights into the regional dynamics of loan applications, aiding in understanding the diverse nature of loan applicants.

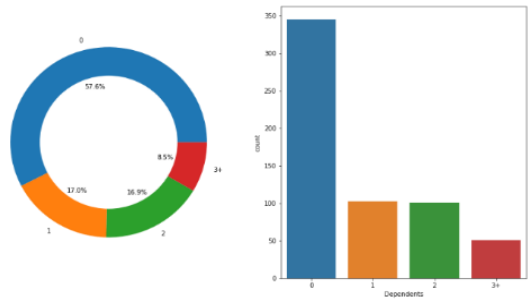


(a) Loan Status

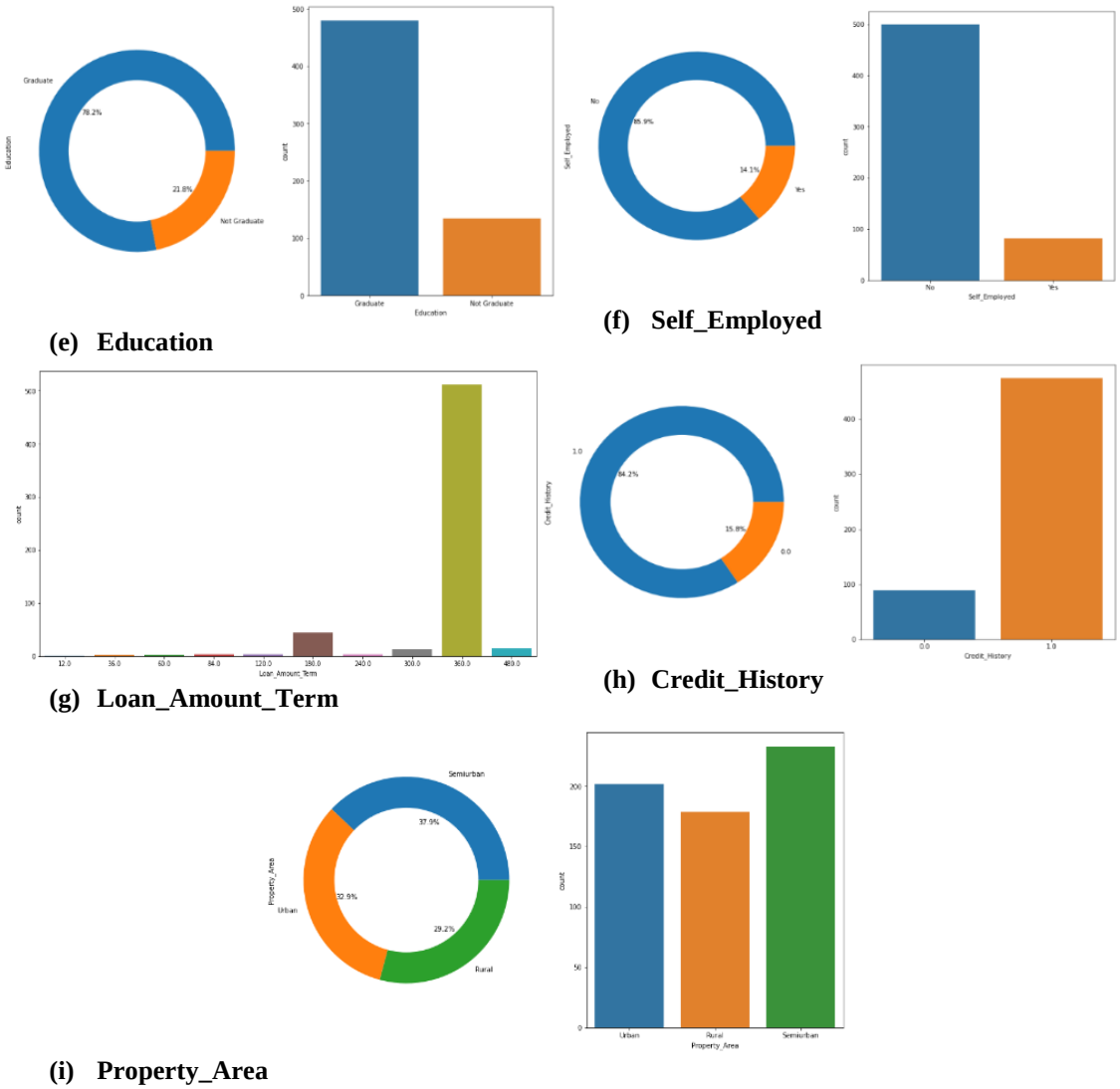
(b) Gender



(c) Married



(d) Dependent



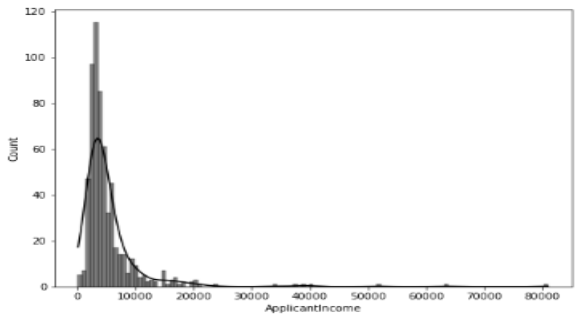
(e) Education **(f) Self_Employed**

(g) Loan_Amount_Term **(h) Credit_History**

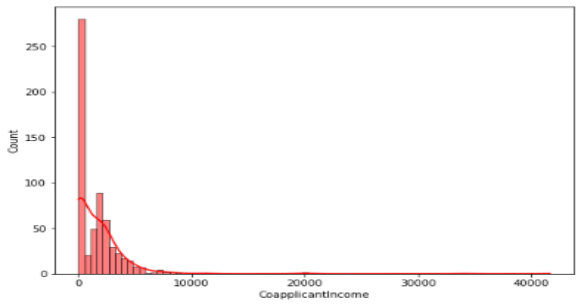
(i) Property_Area

Fig. 2. Univariate Analysis

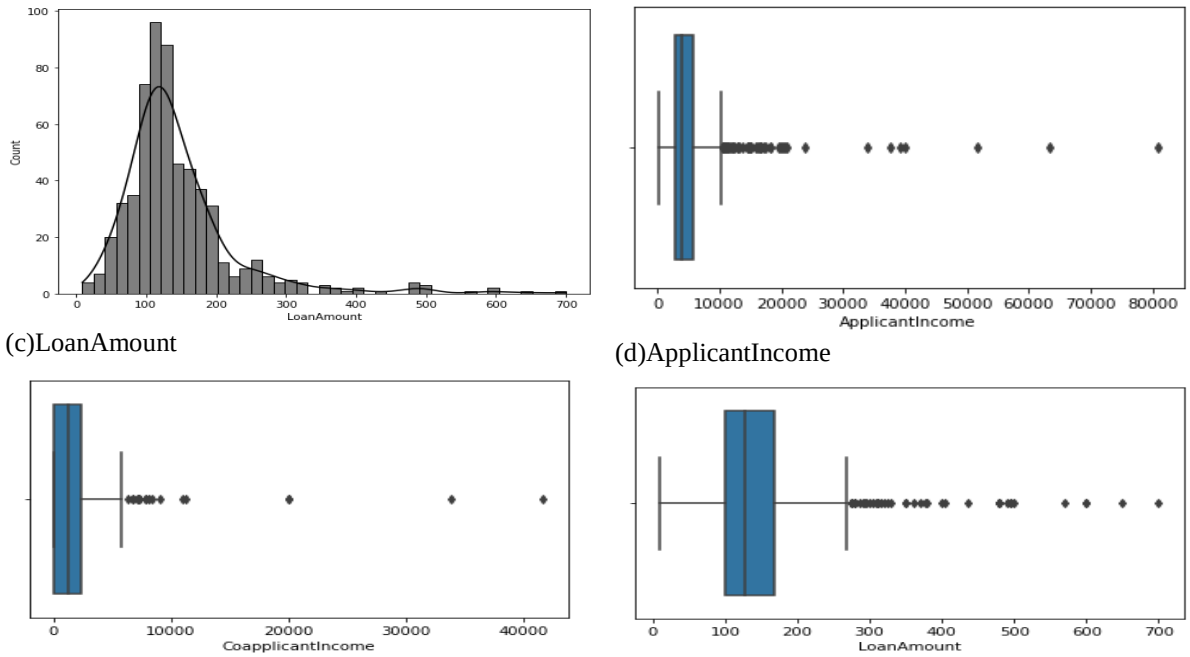
Fig. 3. (a) Reveals a skewed income distribution, with most applicants falling within the income range of 1000 to 10000. However, outliers exist, with some reporting incomes as low as 150 and as high as 81000, contributing to the right skewness in the income data. Similarly, in Fig. 3. (b), co-applicant income displays a skewed distribution to the right, indicating lower incomes compared to primary applicants. Loan amounts vary widely, as depicted in Fig. 3. (c), with a concentration in the range of 80 to 180. The skewed Distribution, with a mean greater than the median, suggests a rightward bias in the loan amount data. In Figures 3 (d), (e), and (f), it is notable that all three columns—Applicant Income, Coapplicant Income, and Loan Amount—exhibit outliers, with the Loan Amount column showing the highest concentration of such outliers. These outliers require meticulous attention during the data preprocessing phase to ensure the resilience of subsequent analyses and models.



(a)ApplicantIncome



(b)CoapplicantIncome



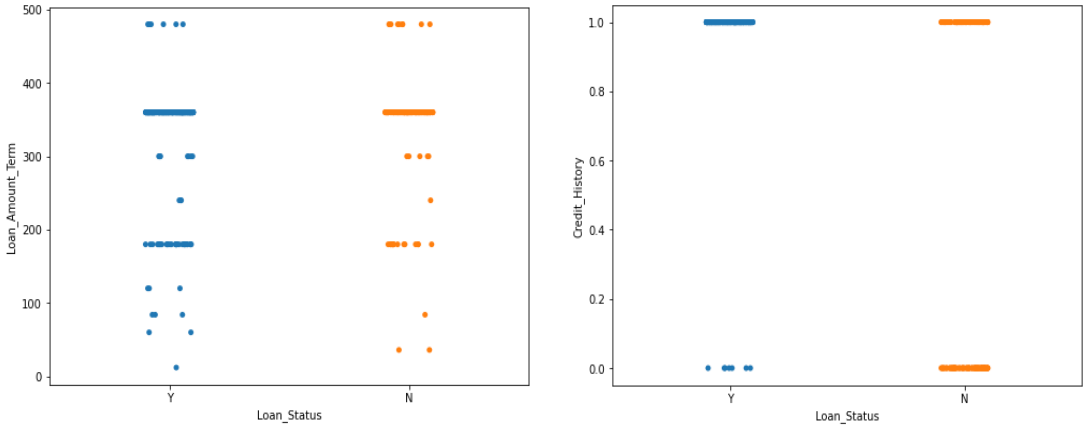
(c)LoanAmount (d)ApplicantIncome

(e)CoapplicantIncome (f)LoanAmount

Fig. 3. Univariate Analysis

3.3.2 (b) Bivariate Analysis

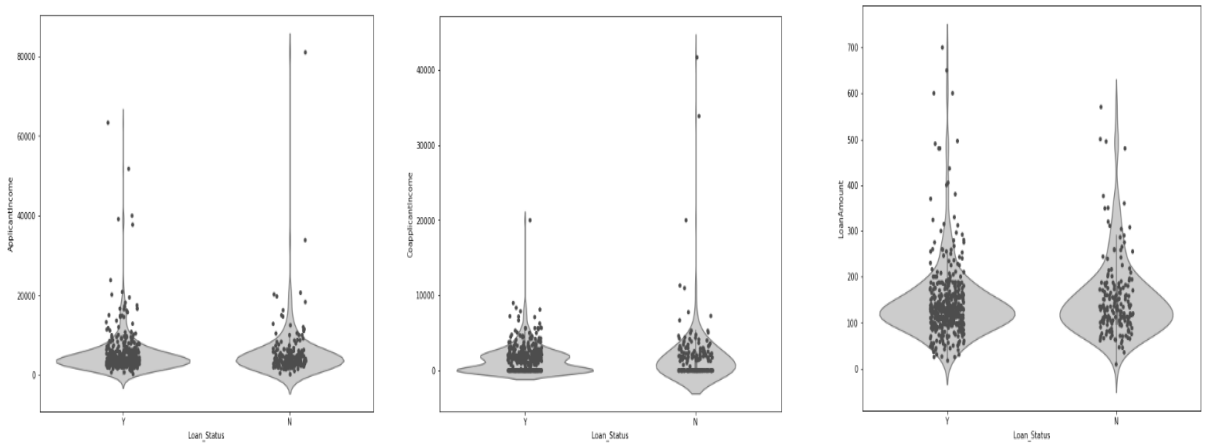
In Fig. 4, intriguing patterns in loan applications and approvals emerge. Particularly in Fig. 4. (a), there is a higher acceptance rate for loan amounts less than 150, despite the majority of applications being for higher amounts. Surprisingly, as shown in Fig. 4. (b), some applicants with a bad credit history still secure loan approvals, indicating that credit history alone does not guarantee rejection.



(a)Loan Status **(b)Loan Status**

Fig. 4. Bivariate Analysis

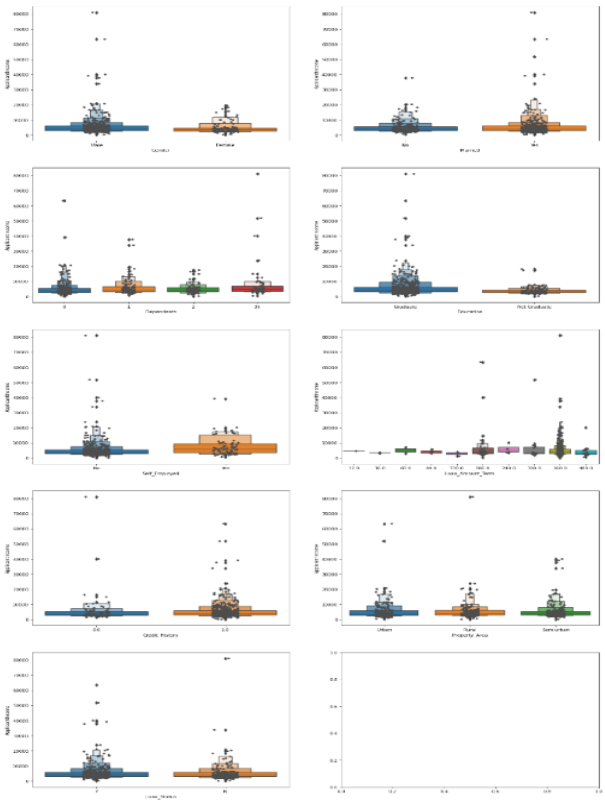
In contrast, Fig. 5. (a) reveals that having a good credit history does not guarantee loan approval. The analysis further challenges the notion that loan approval is solely based on the applicant's income, as rejections occur even with high incomes. Similarly, in Fig. 5. (b), co-applicant income alone does not determine loan approval, with rejected cases observed even for high co-applicant incomes. In Fig. 5. (c), a concentration of loan applications is observed for amounts below 200. Interestingly, all loans above 600 are accepted, emphasizing that loan amount alone does not conclusively determine approval.



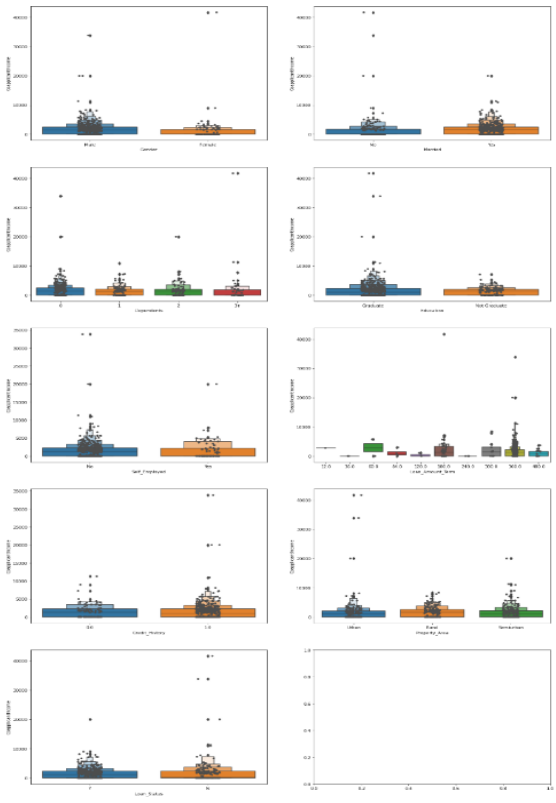
(a)Loan Status (b)Loan Status (c)Loan Status

Fig. 5. Bivariate Analysis

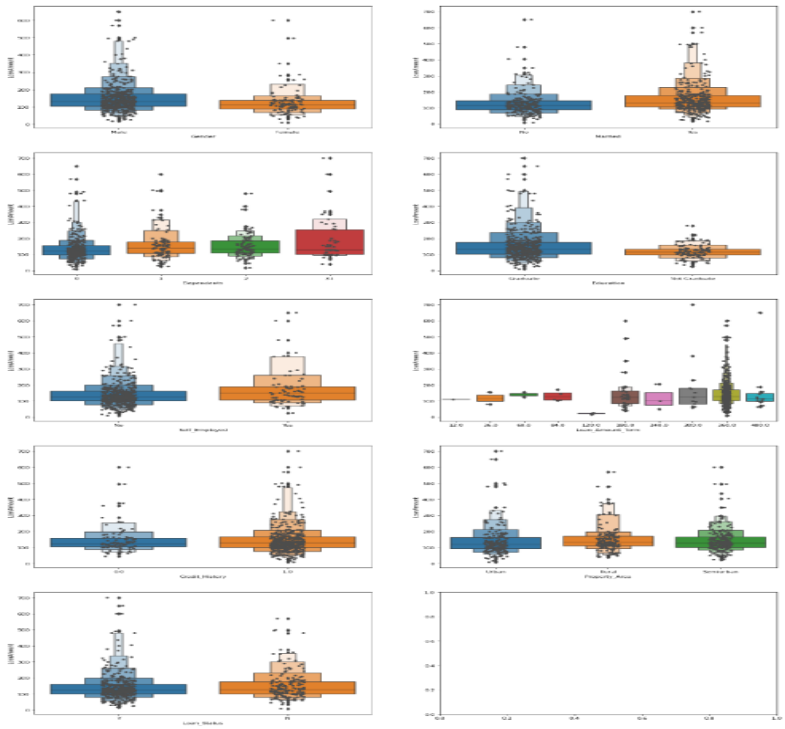
In Fig. 6. (a) gender-based insights indicate that male applicants generally have higher incomes than females, and there are more male applicants overall. Unmarried applicants tend not to take loans exceeding 40,000, while the majority of applicants are married. The number of dependents correlates with a decrease in the applicant's income, with numerous outliers present. Graduates exhibit higher incomes compared to non-graduates, and self-employed individuals, despite having higher mean incomes, represent a smaller proportion of applicants. In Fig. 6. (b) analysis of co-applicants reveals gender disparities, with males generally having higher incomes. Unmarried co-applicants tend to have higher salaries than their married counterparts. The income of the co-applicant appears to be unrelated to the number of dependents of the applicant. Graduates among co-applicants also display higher incomes. Interestingly, non-self-employed co-applicants have higher mean incomes than their self-employed counterparts. In Fig. 6. (c) examining loan amounts and terms, applicants with higher incomes tend to take longer loan terms. Good credit history among both applicants and co-applicants correlates with higher incomes, though outliers are present. Regionally, people from rural areas tend to have the highest mean income, while urban areas exhibit the highest loan amounts. Notably, loan status appears independent of the loan amount, as both high and low loan amounts experience approvals and rejections.



(a) ApplicantIncome vs Categorical features



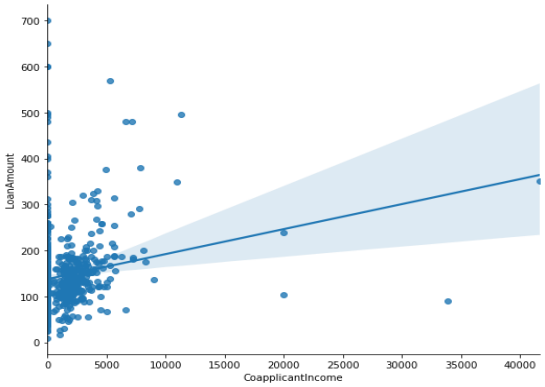
(b) CoapplicantIncome vs Categorical features



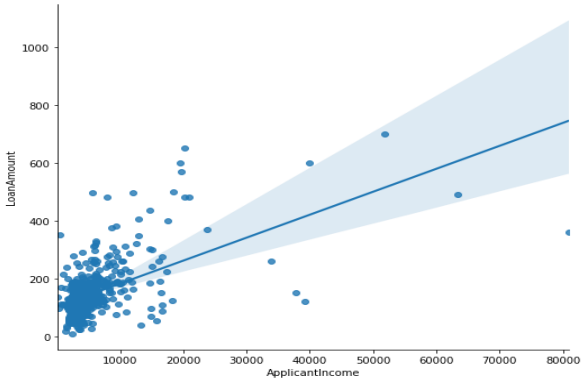
(c) LoanAmount vs Categorical features

Fig. 6. Bivariate Analysis

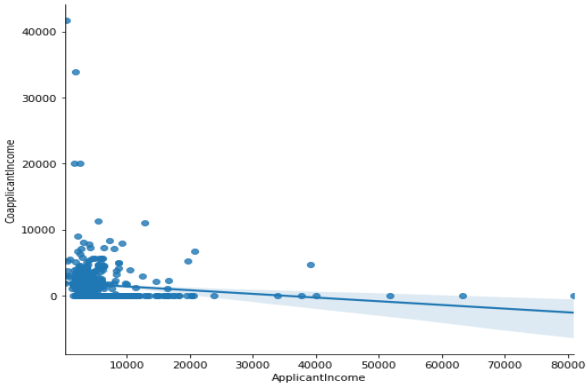
In Figure 7, denoted as (a), (b), and (c), the correlation between loan amount and co-applicant income, as well as loan amount and applicant income, indicates a positive relationship. However, there is a slight decrease in co-applicant income as loan applicant income increases. These intricate insights underscore the multifaceted factors influencing loan applications, approvals, and their interdependencies within the dataset.



(a) CoapplicantIncome



(b) ApplicantIncome

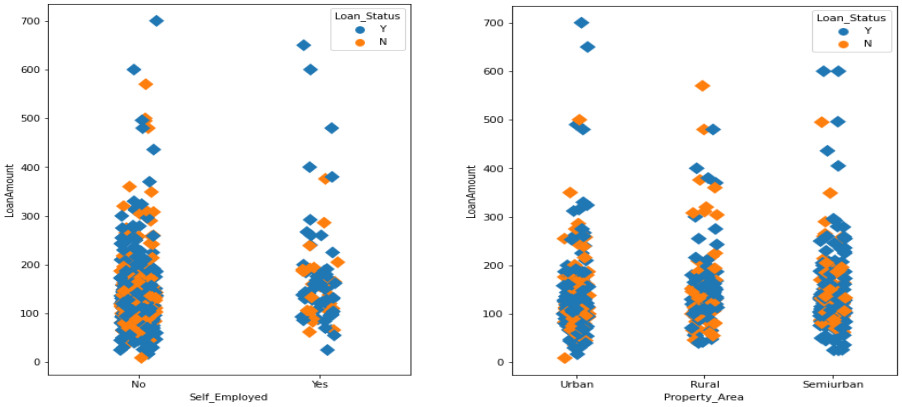


(c) ApplicantIncome

Fig. 7. Bivariate Analysis

3.3.3 (c) Multivariate Analysis

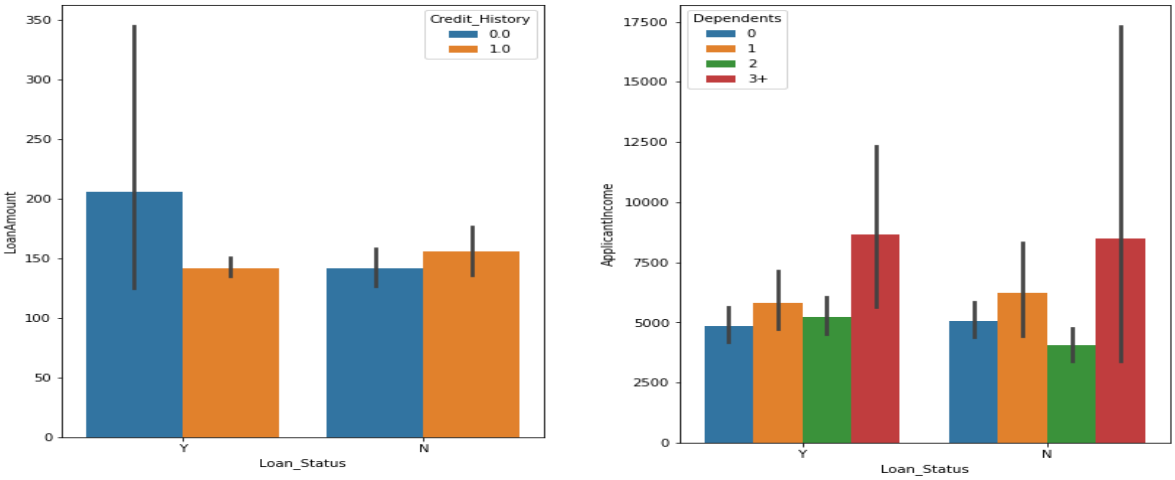
The analysis of the dataset in Figure 8 offers detailed insights into the complexities of loan applications and approvals, uncovering patterns that extend beyond conventional expectations. One notable trend is the tendency for self-employed individuals to seek higher loan amounts compared to their non-self-employed counterparts. However, in Fig. 8. (a), a nuanced observation reveals that loans for amounts exceeding 400 face rejection, irrespective of the applicant's self-employed status.



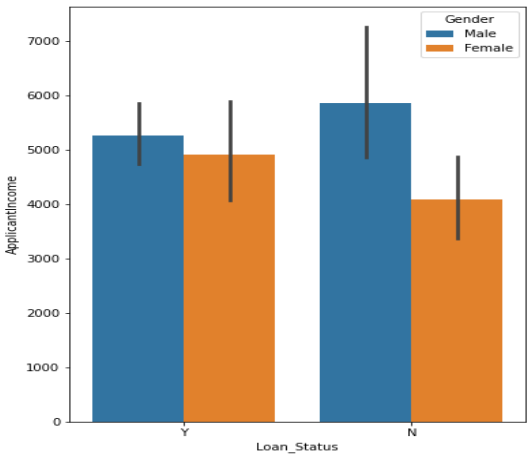
(a) LoanAmount vs Self_Employed **(b) LoanAmount vs Property_Area**

Fig. 8. Multivariate Analysis

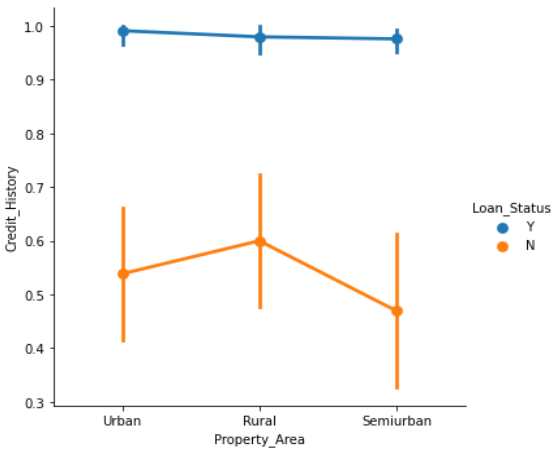
In Fig. 8. (b), loan approval is significantly influenced by regional disparities, with rural areas securing higher loan amounts more frequently than urban areas. This trend persists despite the higher number of applicants from semi-urban areas. Interestingly, in Fig. 9. (a), the dataset suggests that higher loan amounts may be approved, while lower amounts face rejection, even with favorable credit histories. In Fig. 9. (b), the complex relationship between the number of dependents and loan approval is evident. Despite applicants with three dependents having the highest incomes, loan approval depends on various other factors beyond these metrics. In Fig. 9. (c), the interplay between gender and income adds another layer of complexity. Surprisingly, females with higher incomes are often favored for loan approval, contrary to male applicants who, despite comparable earnings, frequently face rejection. Fig. 9. (d) reveals a discernible pattern regarding the influence of credit history on loan approval. Candidates with poor credit histories consistently encounter denial, regardless of their demographic profiles. This underscores the crucial significance credit history holds in the loan approval process. Finally, in Fig. 9. (e), analyzing the connection between loan size and applicant income uncovers an expected positive correlation: as loan amounts rise, so does applicant income. However, the dataset defies expectations by showcasing that loan approval isn't solely based on this correlation. Instances where lower incomes secure higher loan amounts and vice versa underscore the intricate factors at play in loan approval decisions.



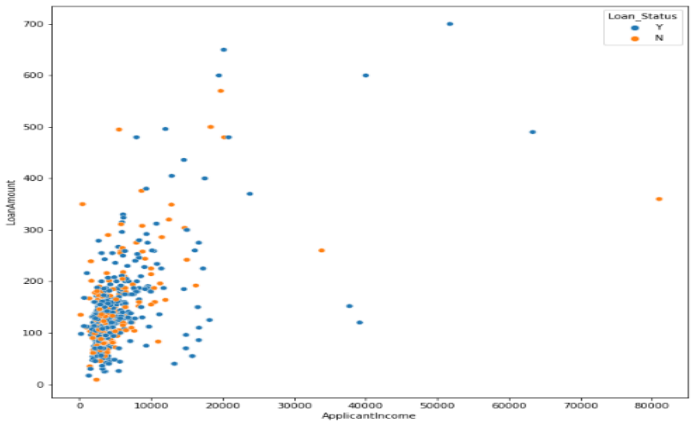
(a) LoanAmount vs Loan_Status **(b) ApplicantIncome vs Loan_Status**



(c) ApplicantIncome vs Loan_Status



(d) Credit_History vs Property_Area



(e) LoanAmount vs ApplicantIncome

Fig. 9. Multivariate Analysis

Essentially, the dataset provides a rich and intricate depiction of the factors influencing loan approvals, highlighting the necessity for a comprehensive understanding beyond conventional assumptions. These findings emphasize the importance of considering multiple variables and their intricate interdependencies when assessing the likelihood of loan approval. In Fig. 10, there isn't much correlation between independent variables except for loan amount, which shows a positive correlation with applicants' and co-applicants' inco

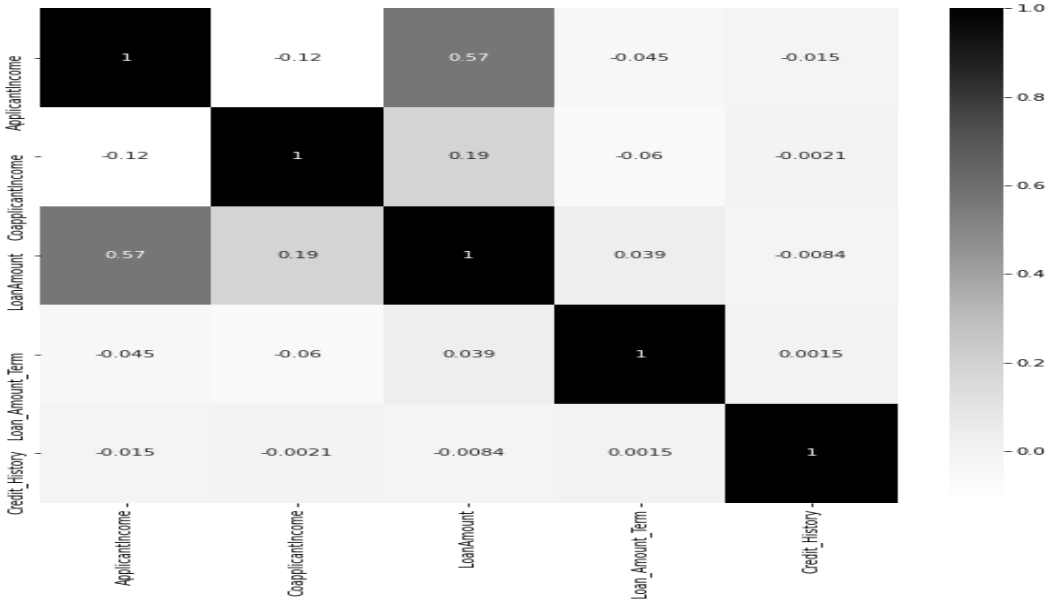


Fig. 10. Multivariate Analysis (Correlation Matrix)

3.4 Data Preprocessing

In the endeavor to create a comprehensive and well-structured dataset, several crucial steps were undertaken to enhance its quality and effectiveness. Initially, null values in categorical features were imputed using the mode, ensuring completeness in the dataset. For the continuous feature "LoanAmount," which exhibited high skewness, null values were filled with the median to alleviate the impact of skewed data. Fig. 11 depicts that the dataset does not contain any null values.

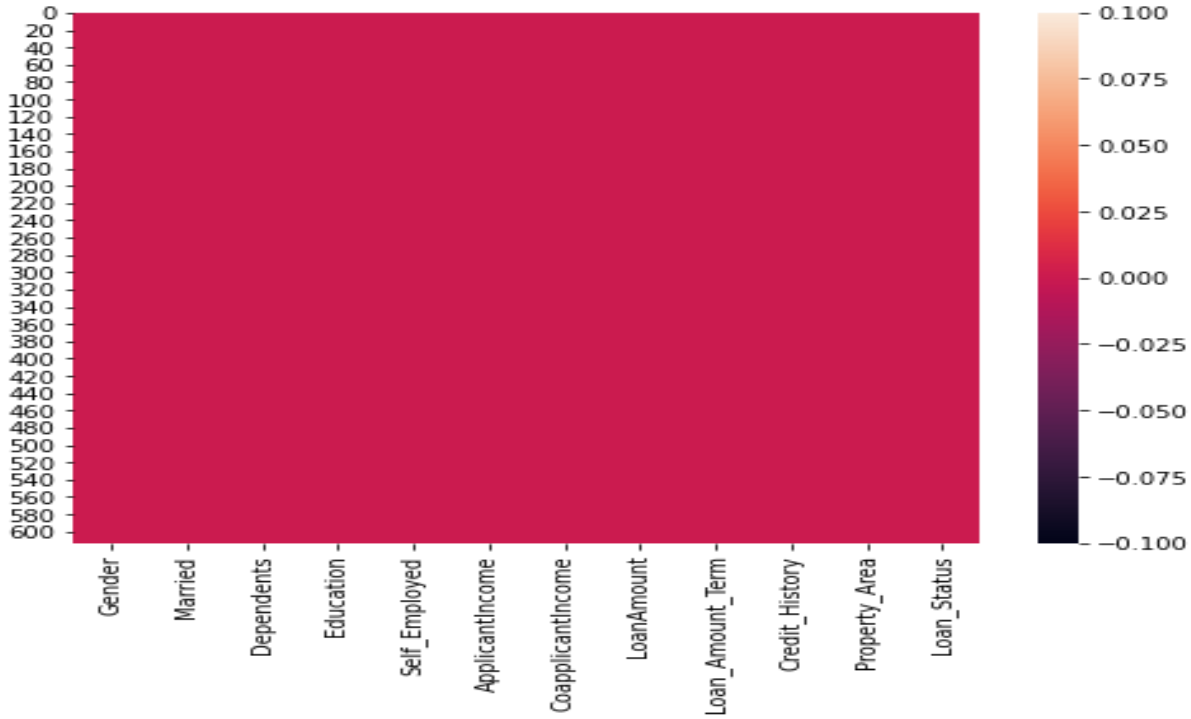


Fig. 11. Null values in the dataset.

Following these imputations, the dataset underwent a thorough examination, and no further null values were identified, indicating successful data preprocessing. To facilitate effective machine learning model training, object-type columns were encoded. Independent features underwent ordinal encoding, while label encoding was applied to the target variable. A systematic approach using the z-score method was employed to address the presence of outliers in the data. A threshold of 3.6 was determined, considering the nature of the data and its limited quantity. Outliers exceeding this threshold were subsequently removed, ensuring the overall integrity of the dataset. A critical aspect of data preprocessing involved addressing skewness. In Fig. 12, through careful measures, skewness was significantly reduced, as evidenced by distribution plots illustrating the before-and-after scenarios. Continuing with dataset refinement, the dependent and independent features were separated, laying the groundwork for effective model training. Acknowledging the imbalance in the data, a strategic approach was adopted to address this issue. The Synthetic Minority Oversampling Technique (SMOTE) was employed to achieve a balanced distribution, resulting in equal representation (50%) for each income category, as demonstrated in Fig. 13. The final step entailed scaling the data to ensure uniformity and enhance model performance. The Min-Max scaler was employed to standardize the numerical features, ensuring consistency across the dataset. This comprehensive data preprocessing pipeline establishes the foundation for developing robust machine learning models, facilitating accurate predictions and valuable insights.

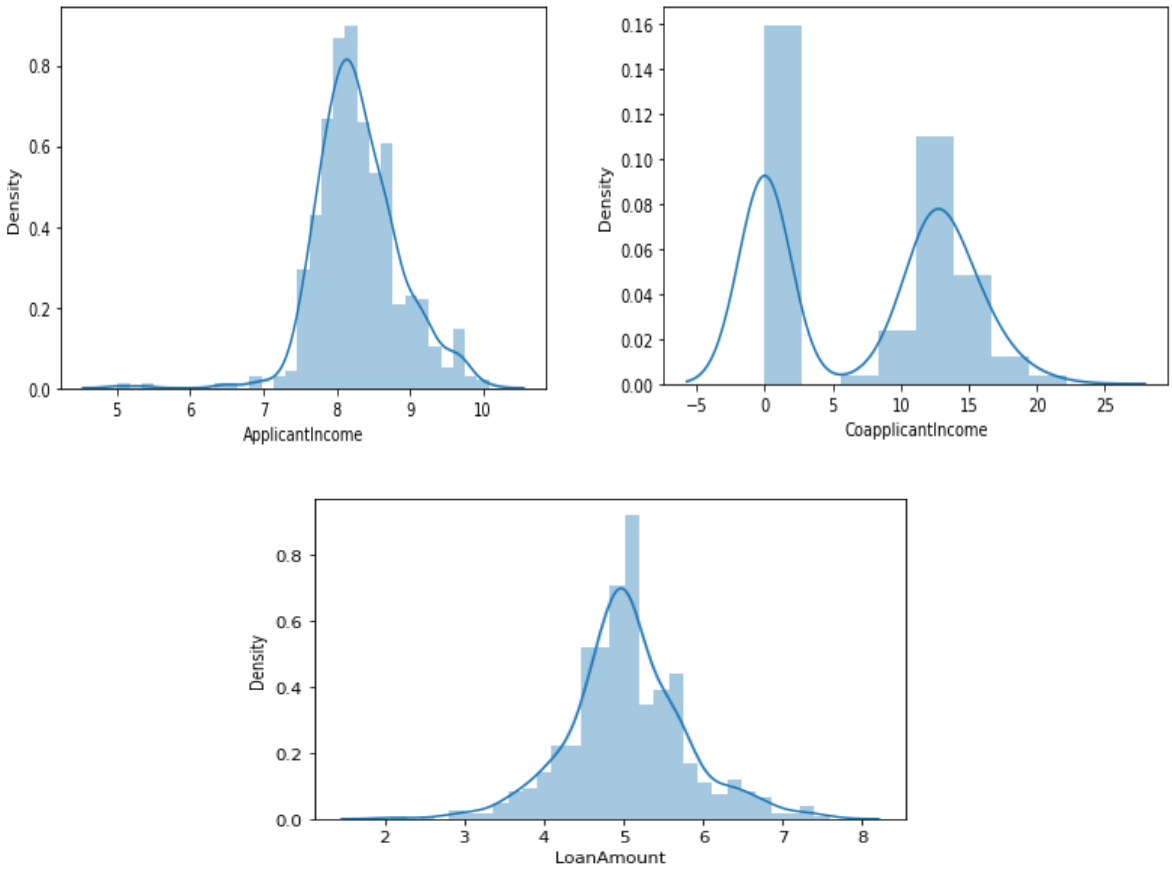


Fig. 12. Skewness in distribution plot

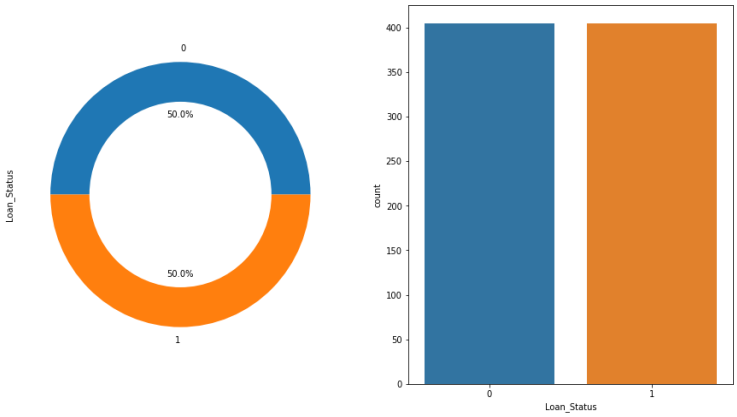


Fig. 13. Balanced distribution using SMOTE.

3.5 Hyperparameter Tuning

In Table 3, Random Forest, Gradient Boost, and Xtreme Gradient Boosting demonstrated the highest performance, prompting us to apply hyperparameter tuning to these models. Ultimately, Random Forest emerged as the optimal choice for our problem, showcasing a robust cross-validation score with minimal variance between its cross-validation and accuracy scores. In the pursuit of optimizing predictive models for loan approval, a comprehensive evaluation of various classifiers has been conducted. The performance metrics of each model shed light on their respective strengths and areas for potential improvement. The Gaussian Naive Bayes (GNB) model demonstrates a commendable accuracy of 79.31%, with a mean cross-validation score of 75.43%.

Table 3: Comparison of Cross Validation Score and Accuracy Score for Top Three Models.

Model Name	Accuracy (%)	Cross Validation Score (%)
RandomForest	**89.66%**	**84.44%**
GradientBoosting	85.22%	82.35%
XGBClassifier	92.12%	83.83%

3.6 Performance matrices

Machine learning models can exhibit a wide range of characteristics and behaviors, making it challenging to identify the optimal model for a specific task. Therefore, it is crucial to have a toolkit for effectively assessing model performance. Common quality control measures in machine learning include accuracy, precision, recall, and F1-score. These metrics, calculated using the confusion matrix (shown in Table-4), are essential for evaluating model performance.

Next, each classifier is analyzed using the (1-5) metrics formula. The confusion matrix categorizes outcomes into:

- True Positives (TP): Both the prediction and the actual output are YES.
- True Negatives (TN): Both the prediction and the actual output are NO.
- False Positives (FP): The prediction is YES, but the actual output is NO.
- False Negatives (FN): The prediction is NO, but the actual output is YES.

Table 4: Confusion Matrices.

		Predicted Class	
Actual Class		**True**	**False**
	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

$$Precision = \frac{TP}{(TP + TN)} \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

$$F1\_score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} \quad (5)$$

4. RESULT AND ANALYSIS

Table 5 provides the performance metrics including Accuracy, Precision, Recall, F1-Score, and Support for all models. The precision and recall metrics reveal balanced performance, particularly in distinguishing between approved and rejected loans. The K-Nearest Neighbors (KNN) classifier achieves an accuracy of 80.3%, outperforming GNB. Notably, an AUC-ROC score of 0.8886 indicates a strong capability to distinguish between positive and negative instances. The confusion matrix and classification report emphasize the model's proficiency in handling loan approval classifications. The Support Vector Classifier (SVC) with probability estimation achieves an accuracy of 80.79%, with an AUC-ROC score of 0.9127. Its effectiveness in identifying true positives and negatives is reflected in its precision, recall, and F1-score metrics. Logistic Regression achieves the highest accuracy among the evaluated models, with a score of 82.27%. Additionally, an AUC-ROC score of 0.8909 confirms its proficiency in distinguishing between loan approvals outcomes. The confusion matrix and classification report underscore its reliability in handling diverse loan scenarios. While the Decision Tree Classifier achieves an accuracy of 79.31%, it demonstrates balanced performance with a mean cross-validation score of 78.64%. The AUC-ROC score of 0.7971 highlights its ability to discriminate between loan approval categories. The Random Forest Classifier emerges as a robust performer, attaining an impressive accuracy of 89.66%. With a high AUC-ROC score of 0.9546, this model excels in both precision and recall metrics, showcasing its proficiency in handling loan approval predictions. AdaBoost and Gradient Boosting classifiers deliver accuracies of 84.73% and 89.16%, respectively. Fig. 14 presents the confusion matrices for the following classifiers: (a) GaussianNB, (b) KNN, (c) SVC, (d) Logistic Regression, (e) Decision Tree, (f) Random Forest, (g) AdaBoost, (h) Gradient Boosting, and (i) XGB. Both models exhibit strong AUC-ROC scores (showed in Fig. 15), highlighting their ability to differentiate between positive and negative instances. The XGBoost classifier, with an accuracy of 87.68%, demonstrates balanced performance and robust

discriminatory power, as evidenced by the AUC-ROC score of 0.9338. In summary, the evaluation of diverse classifiers provides a nuanced understanding of their strengths and weaknesses in predicting loan approval outcomes. Selecting the optimal model depends on the lending institution's specific priorities and requirements, including precision, recall, and overall accuracy. This thorough analysis offers stakeholders valuable guidance for implementing an effective loan approval system.

Table 5: Accuracy, Precision, Recall, F1-Score, Support performance for all classification model.

Parameter	Accuracy (%)	Precision		Recall		F1-Score		Support	
		Yes	No	Yes	No	Yes	No	Yes	No
GaussianNB	79.31%	0.94	0.63	0.74	0.91	0.83	0.74	136	67
KNeighbors Classifier	80.3%	0.84	0.76	0.79	0.81	0.82	0.79	112	91
SVC	80.79%	0.94	066	0.75	0.91	0.84	0.77	133	70
Logistic Regression	82.27%	0.99	0.64	0.75	0.98	0.85	0.77	140	63
Decision Tree	79.31%	0.71	0.89	0.87	0.74	0.78	0.80	86	117
Random Forest	**89.66%**	0.92	0.87	0.88	0.91	0.90	0.89	111	92
AdaBoost Classifier	84.73%	0.90	0.79	0.83	0.88	0.86	0.83	115	88
Gradient Boosting	**89.16%**	0.94	0.84	0.86	0.93	0.90	0.88	116	87
XGBClassifier	**87.68%**	0.87	0.89	0.89	0.86	0.88	0.87	103	100

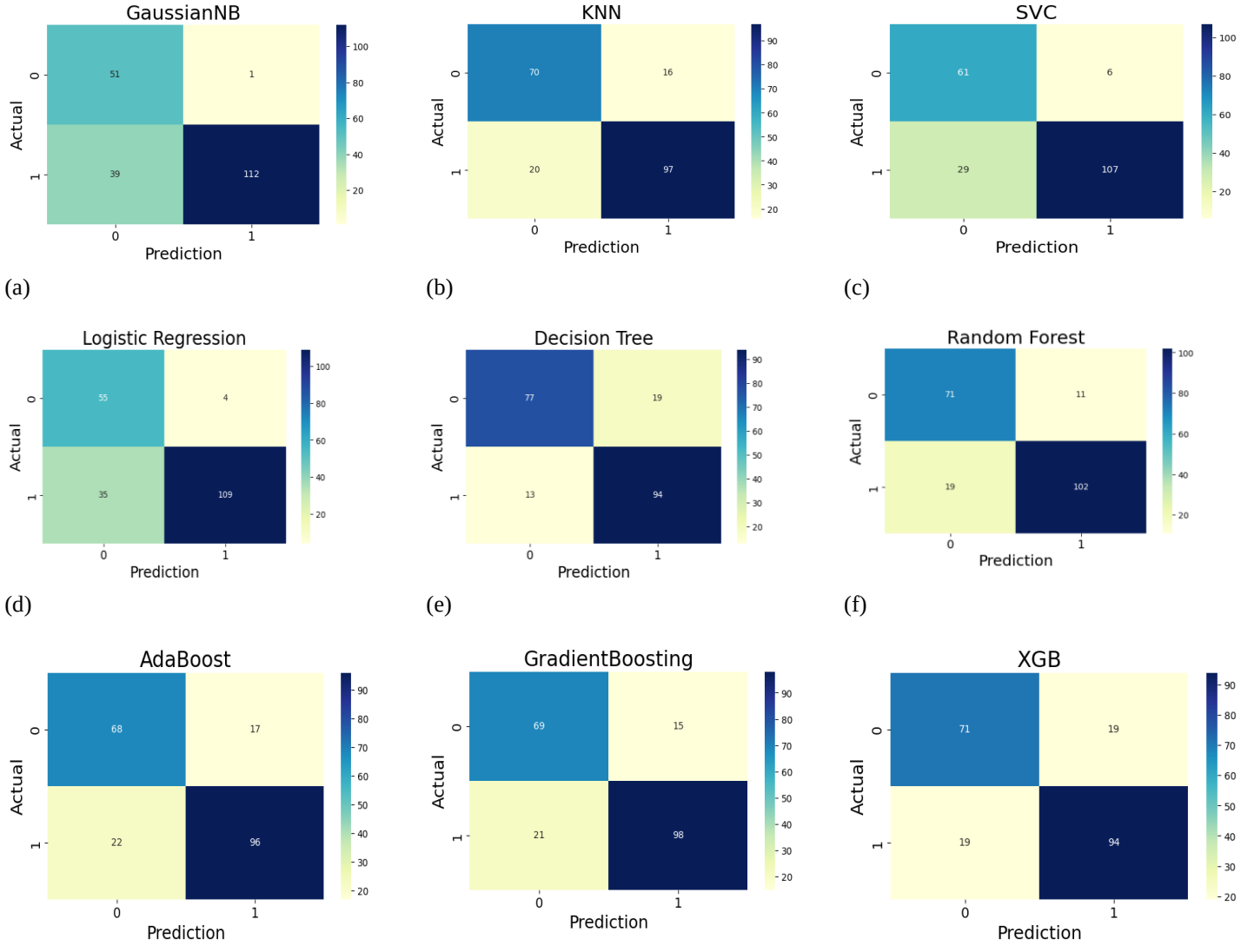


Fig. 14: Confusion matrix for (a) GaussianNB (b) KNN (c) SVC (d) Logistic Regression (e) Decision Tree (f) Random Forest (g) AdaBoost (h) Gradient Boosting (i) XGB.

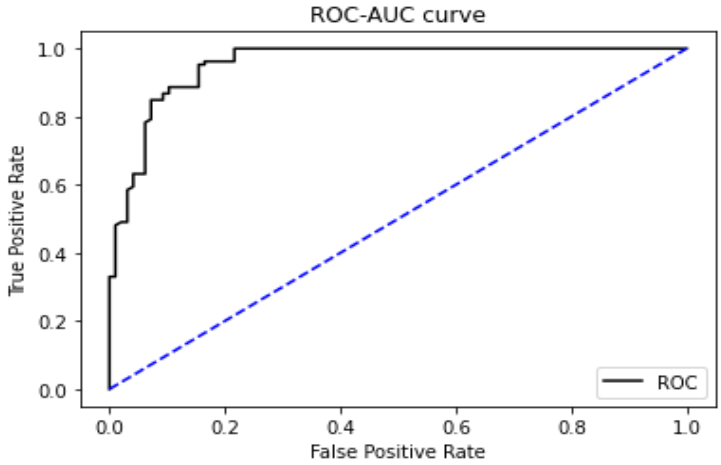


Fig. 15. ROC-AUC Curve

5. USER INTERFACE PERFORMANCE

Fig. 17 illustrates the graphical representation of the implemented user interface. Our implementation utilizes a loan dataset comprising eleven independent attributes and one dependent attribute, referred to as the Target attribute. To check loan availability in the user interface, all eleven independent attributes must be provided. The dependent attribute, Loan Status, relies on the following independent attributes: Gender, Married, Dependents, Education, Self-Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, and Property Area. Fig. 16 showcases the API created using the Flask Python library, enabling users to input values and check loan eligibility easily. In Fig. 16, the mandatory data input validation check is depicted, ensuring all data fields are completed before performing the loan status check for a specific customer. When all input data requirements are met, the user interface displays 'Loan eligibility: yes'. If the conditions are not met, 'Loan eligibility: no' is shown, as depicted in Fig. 16. Table 4 compares the highest model accuracy with existing work, highlighting that our proposed work, focusing on the Random Forest model, achieves an accuracy of 89.66%, outperforming other models in the literature. The performance is further enhanced using hyperparameter tuning, confirming Random Forest as the best model among the three evaluated. The effectiveness of machine learning techniques depends on the dataset used for training and evaluation. This study utilizes publicly available secondary loan data from the Kaggle repository. While various built-in machine learning models are used, the goal is to train these models with real-time data in the future. The study employs hyperparameter tuning to assess and optimize model performance.

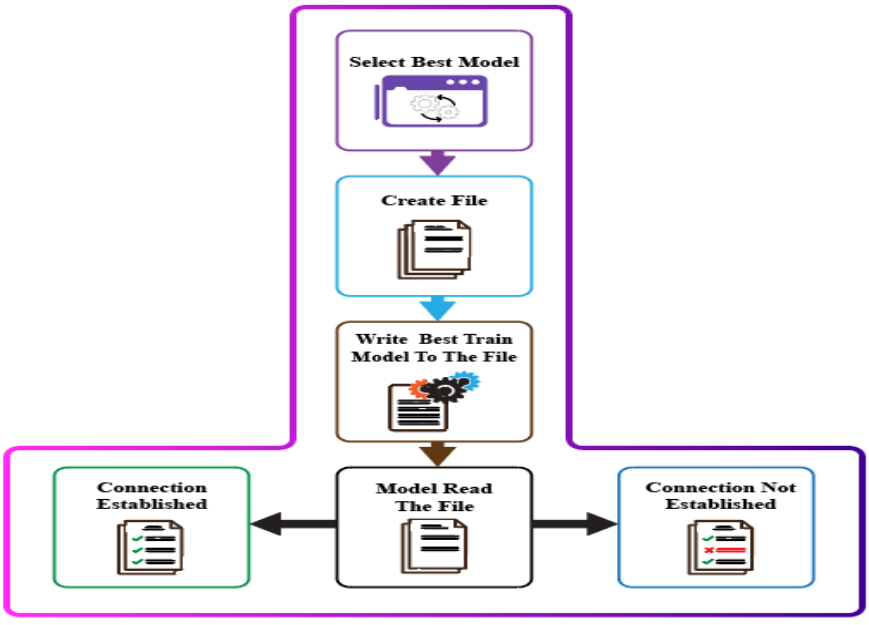


Fig. 16: Best Model Connectivity Model.

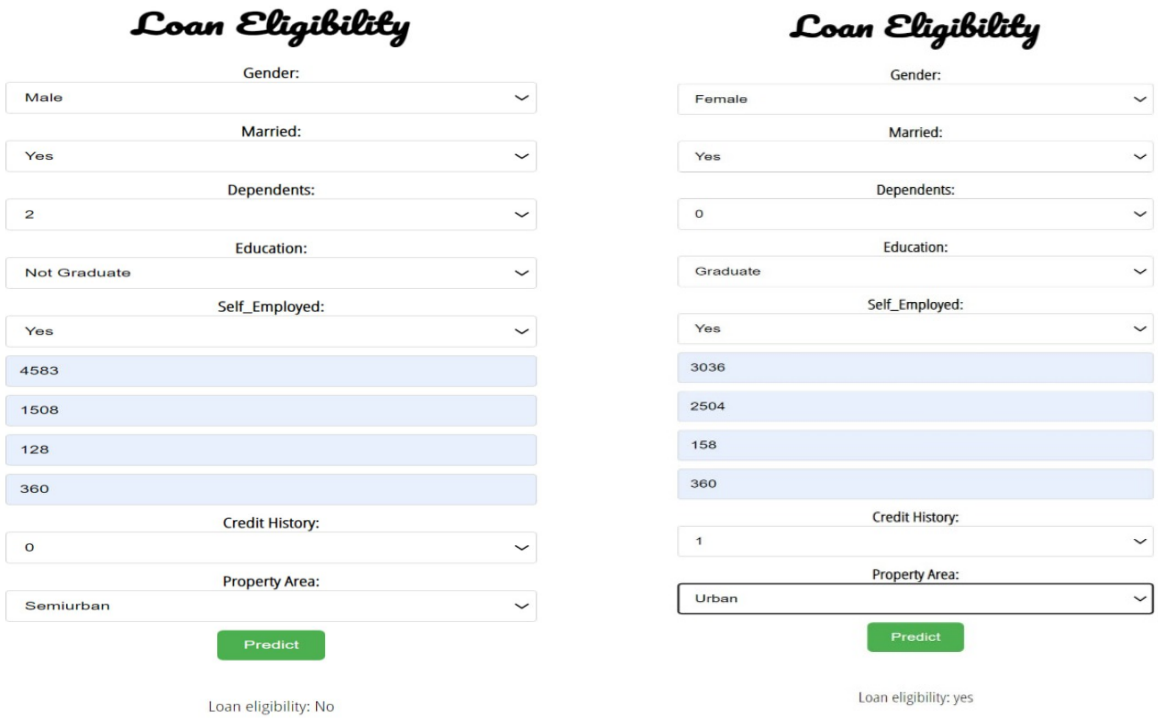


Fig. 17: Graphical View of User Interface.

6. DISCUSSION

We determined the best model through two distinct approaches. Firstly, we trained individual models using random state data splitting. Secondly, we employed Hyperparameter Tuning to identify the optimal model. In the random data splitting approach, we divided the dataset randomly into two segments: the training dataset, which comprised 75% of the data, and the testing dataset, which constituted the remaining 25%. This division enabled us to evaluate the performance of the model effectively.

7. CONCLUSION

The study primarily focused on evaluating default risks associated with loans in the banking sector, utilizing both machine learning and deep learning algorithms. The selection of the appropriate algorithm played a crucial role in managing loan decisions by aiding in determining the probability of clients defaulting on their loans. Accurate prediction of loan default probabilities enables financial institutions to effectively mitigate the risk of incurring financial losses during the loan approval process. The initial phase involved cleansing the dataset, which included eliminating variables with a significant amount of missing data. Subsequently, efforts were made to address issues related to unbalanced data and outliers before inputting the data into machine learning algorithms. Various algorithms were examined in the study, including LR, DT, RF, ET, SVM, KNN, GNB, AdaBoost, and GB models. Among these, three high-performing models were identified, with an ensemble classifier being suggested for predicting bank loan defaults. A comprehensive assessment verified that the approach notably improved the precision and dependability of loan default forecasts. Additionally, a user-friendly desktop-based interface was developed to evaluate the eligibility of loans for specific customers, thereby assisting banks in streamlining the loan approval process and enhancing operational efficiency. The discoveries made in this study have the potential to advance risk management methodologies within the banking sector and provide valuable insights for future research endeavors in this field.

REFERENCES

1. Dansana, D., Patro, S.G.K., Mishra, B.K., Prasad, V., Razak, A. and Wodajo, A.W., 2024. Analyzing the impact of loan features on bank loan prediction using R andom F orest algorithm. *Engineering Reports*, *6*(2), p.e12707.Khairi et al., 2021
2. Khairi, A., Bahri, B. and Artha, B., 2021. A literature review of non-performing loan. *Journal of Business and Management Review*, *2*(5), pp.366-373.Musdholifah et al., 2020
3. Musdholifah, M., Hartono, U. and Wulandari, Y., 2020. Banking crisis prediction: emerging crisis determinants in Indonesian banks. *International Journal of Economics and Financial Issues*, *10*(2), p.124.Ma et al., 2023
4. Ma, Y., Yu, C., Yan, M., Sangaiah, A.K. and Wu, Y., 2023. Dark-side avoidance of mobile applications with data biases elimination in socio-

cyber world. *IEEE Transactions on Computational Social Systems*.Koulouridi et al., 2021

5. Koulouridi, E., Kumar, S., Nario, L., Pepanides, T. and Vettori, M., 2020. Managing and monitoring credit risk after the COVID-19 pandemic. *McKinsey & Company*.
6. Alzubi, O.A., Alzubi, J.A., Al-Zoubi, A.M., Hassonah, M.A. and Kose, U., 2022. An efficient malware detection approach with feature weighting based on Harris Hawks optimization. *Cluster Computing*, pp.1-19.
7. Koulouridi, E., Kumar, S., Nario, L., Pepanides, T. and Vettori, M., 2020. Managing and monitoring credit risk after the COVID-19 pandemic. *McKinsey & Company*.
8. Rawate, K.R. and Tijare, P.A., 2017. Review on prediction system for bank loan credibility. *International Journal of Advance Engineering and Research Development*, *4*(12), pp.860-867.
9. Bhargav, P. and Sashirekha, K., 2023. A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms. *Journal of Survey in Fisheries Sciences*, *10*(1S), pp.1803-1813.
10. Dasari, Y., Rishitha, K. and Gandhi, O., 2023. Prediction of bank loan status using machine learning algorithms. *International Journal of Computing and Digital Systems*, *14*(1), pp.1-1.
11. Abdullah, M., Chowdhury, M.A.F., Uddin, A. and Moudud-Ul-Huq, S., 2023. Forecasting nonperforming loans using machine learning. *Journal of Forecasting*, *42*(7), pp.1664-1689.
12. Kavitha, M.N., Saranya, S.S., Dhinesh, E., Sabarish, L. and Gokulkrishnan, A., 2023, March. Hybrid ML classifier for loan prediction system. In *2023 international conference on sustainable computing and data communication systems (ICSCDS)* (pp. 1543-1548). IEEE.
13. Wang, Y., Wang, M., Pan, Y. and Chen, J., 2023. Joint loan risk prediction based on deep learning-optimized stacking model. *Engineering Reports*, p.e12748.
14. Alsaleem, M.Y. and Hasoon, S.O., 2020. Predicting bank loan risks using machine learning algorithms. *AL-Rafidain Journal of Computer Sciences and Mathematics*, *14*(1), pp.149-158.
15. Wang, D., Wu, Q. and Zhang, W., 2019. Neural learning of online consumer credit risk. *arXiv preprint arXiv:1906.01923*.
16. Supriya, P., Pavani, M., Saisushma, N., Kumari, N.V. and Vikas, K., 2019. Loan prediction by using machine learning models. *International Journal of Engineering and Techniques*, *5*(2), pp.144-147.
17. Sun, T. and Vasarhalyi, M.A., 2021. Predicting credit card delinquencies: An application of deep neural networks. Handbook of Financial Econometrics, Mathematics. *Statistics, and Machine Learning*, pp.4349-4381.
18. Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P., 2021. Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012042). IOP Publishing.
19. Anand, M., Velu, A. and Whig, P., 2022. Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering (JCSE)*, *3*(1), pp.1-13.
20. Kumar, C.N., Keerthana, D., Kavitha, M. and Kalyani, M., 2022, June. Customer loan eligibility prediction using machine learning algorithms in banking sector. In *2022 7th international conference on communication and electronics systems (ICCES)* (pp. 1007-1012). IEEE.
21. Dosalwar, S., Kinkar, K., Sannat, R. and Pise, N., 2021. Analysis of loan availability using machine learning techniques. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, *9*(1), pp.15-20.
22. Blessie, E.C. and Rekha, R., 2019. Exploring the machine learning algorithm for prediction the loan sanctioning process. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(1), pp.2714-2719.
23. Loan prediction problem dataset. Loan Prediction Problem Dataset (kaggle.com)