LOAN APPROVAL PREDICTION USING DATA VISUALIZATION

A PROJECT REPORT

Submitted by,

PURVIKA S - 20201ISE0050
PRAMODH YADAV M- 202011ISE0053
CHATUR S - 20201ISE0094

Under the guidance of,
Ms. POORNIMA S

in partial fulfillment for the award of the degree of
BACHELOR OF TECHNOLOGY

IN

INFORMATION SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY
BENGALURU

TABLE OF CONTENTS

ABSTRACT

Loan approval is a vital aspect of the banking sector, significantly influencing financial stability and customer satisfaction. This report explores the application of machine learning techniques to predict loan approval decisions, aiming to streamline the process and enhance decision accuracy. The dataset utilized in this study comprises 13 features, including unique identifiers, demographic information (gender, marital status, dependents), socio-economic indicators (education, self-employment status), financial metrics (applicant and co-applicant incomes, loan amount, loan term), credit history, and property area, culminating in the loan status (approved or not).

The methodology begins with importing necessary libraries such as Pandas, Matplotlib, and Seaborn for data manipulation and visualization. Data preprocessing involves dropping irrelevant columns, such as unique identifiers, and converting categorical variables into numerical form using Label Encoding. Exploratory data analysis is conducted to visualize the distribution and dominance of categorical values through bar plots and to examine feature correlations using heatmaps.

To address any missing values, the dataset undergoes imputation with mean values, ensuring no null entries remain. The data is then split into training and testing sets, with 60% allocated for training and 40% for testing. Four machine learning models are employed: K-Nearest Neighbors, Random Forest, Support Vector Classifiers, and Logistic Regression. These models are trained and evaluated based on their accuracy scores, both on the training and testing datasets.

The Random Forest classifier emerges as the most effective model, achieving an accuracy of 98.04% on the training set and 82.5% on the testing set, outperforming the other models. The findings suggest that credit history is a significant predictor of loan status, and additional factors like applicant income and loan amount also play crucial roles.

This study demonstrates the efficacy of machine learning in automating and improving the accuracy of loan approval processes. It highlights the potential for further enhancement through advanced techniques such as ensemble learning methods, including bagging and boosting. By leveraging data-driven approaches, banks can make more informed and efficient lending decisions, ultimately benefiting both the institution and its customers..

CHAPTER-1
INTRODUCTION

In the contemporary financial landscape, loan approval is a critical function for banks and financial institutions. Loans serve as a major revenue source for banks, helping them achieve profitability while providing individuals and businesses with the necessary capital to pursue various endeavors such as education, housing, and business expansion. However, the process of determining whether a loan application should be approved is complex and multifaceted, requiring the evaluation of numerous criteria to mitigate risk and ensure the borrower's ability to repay.

Traditionally, loan approval decisions have relied heavily on manual assessment and the expertise of loan officers. While effective to some extent, this method is time-consuming and prone to human error and bias. To address these challenges, there is a growing interest in leveraging machine learning (ML) techniques to automate and enhance the loan approval process. Machine learning offers the potential to analyze large volumes of data quickly and accurately, identifying patterns and correlations that might not be immediately apparent through manual review.

This report focuses on developing a machine learning model to predict loan approval decisions based on a dataset containing 13 features relevant to loan applicants. These features include demographic information (such as gender and marital status), socio-economic indicators (like education level and employment status), financial metrics (such as income levels and loan amount), and historical data on creditworthiness. By analyzing these variables, the model aims to determine the likelihood of loan approval, thereby assisting banks in making more informed and efficient decisions.

The project involves several key steps: data preprocessing, exploratory data analysis, feature engineering, model training, and evaluation. Data preprocessing includes handling missing values, encoding categorical variables, and normalizing numerical features. Exploratory data analysis helps in understanding the data distribution and identifying significant features. Various machine learning algorithms, including K-Nearest Neighbors, Random Forest, Support Vector Classifiers, and Logistic Regression, are employed and compared to identify the most effective model for predicting loan approval.

The ultimate goal of this study is to create a robust, accurate, and scalable machine learning model that can be integrated into the loan approval workflow of financial institutions. By automating this process, banks can enhance their operational efficiency, reduce the likelihood of default, and improve customer satisfaction by providing quicker responses to loan applications. This report documents the entire process, from data acquisition and preprocessing to model training and evaluation, providing insights into the practical application of machine learning in the banking sector.

CHAPTER-2
METHODOLOGY

The methodology for developing a loan approval prediction model using machine learning involves several key steps: data acquisition, data preprocessing, exploratory data analysis (EDA), feature engineering, model selection, training, and evaluation. Each step is crucial for building an effective and accurate predictive model.

The methodology for developing a machine learning model to predict loan approval involves several crucial steps: data collection, data preprocessing, exploratory data analysis, model selection and training, and model evaluation. Each step is designed to ensure that the data is adequately prepared, the right features are selected, and the best possible model is developed and validated.

1. Data Collection

The dataset used in this study is sourced from a loan application database, comprising 13 features related to the applicant's demographics, financial status, and credit history. These features include:

Loan_ID: A unique identifier for each loan application.
Gender: The gender of the applicant.
Married: The marital status of the applicant.
Dependents: Number of dependents the applicant has.
Education: Whether the applicant is a graduate or not.
Self_Employed: Employment status of the applicant.
ApplicantIncome: The income of the applicant.
CoapplicantIncome: The income of the co-applicant.
LoanAmount: The amount of loan requested.
Loan_Amount_Term: The term of the loan in months.
Credit_History: The credit history of the applicant.
Property_Area: The area where the property is located (Rural/Urban/Semi-urban).
Loan_Status: The target variable indicating if the loan was approved (Y) or not (N).

2. Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean and suitable for model training. This involves:

Dropping Irrelevant Columns: The Loan_ID column is dropped as it does not contribute to the predictive analysis.

Handling Categorical Variables: Categorical features are encoded using Label Encoding to convert them into numerical values.

Dealing with Missing Values: Missing values are imputed with the mean for numerical columns and the mode for categorical columns to ensure no null values remain in the dataset.

Feature Scaling: Normalizing numerical features to ensure they are on a similar scale, which is especially important for algorithms sensitive to feature scaling.

3. Exploratory Data Analysis (EDA)

EDA involves visualizing the data to understand distributions, relationships, and patterns among features. Techniques include:

Bar Plots: To visualize the distribution of categorical variables.

Heatmaps: To show correlations between numerical features and identify significant predictors of loan status.

4. Model Selection and Training

The dataset is split into training and testing sets, and various machine learning models are trained to predict loan approval. The models used include:

K-Nearest Neighbors (KNN)

Random Forest Classifier

Support Vector Classifier (SVC)

Logistic Regression

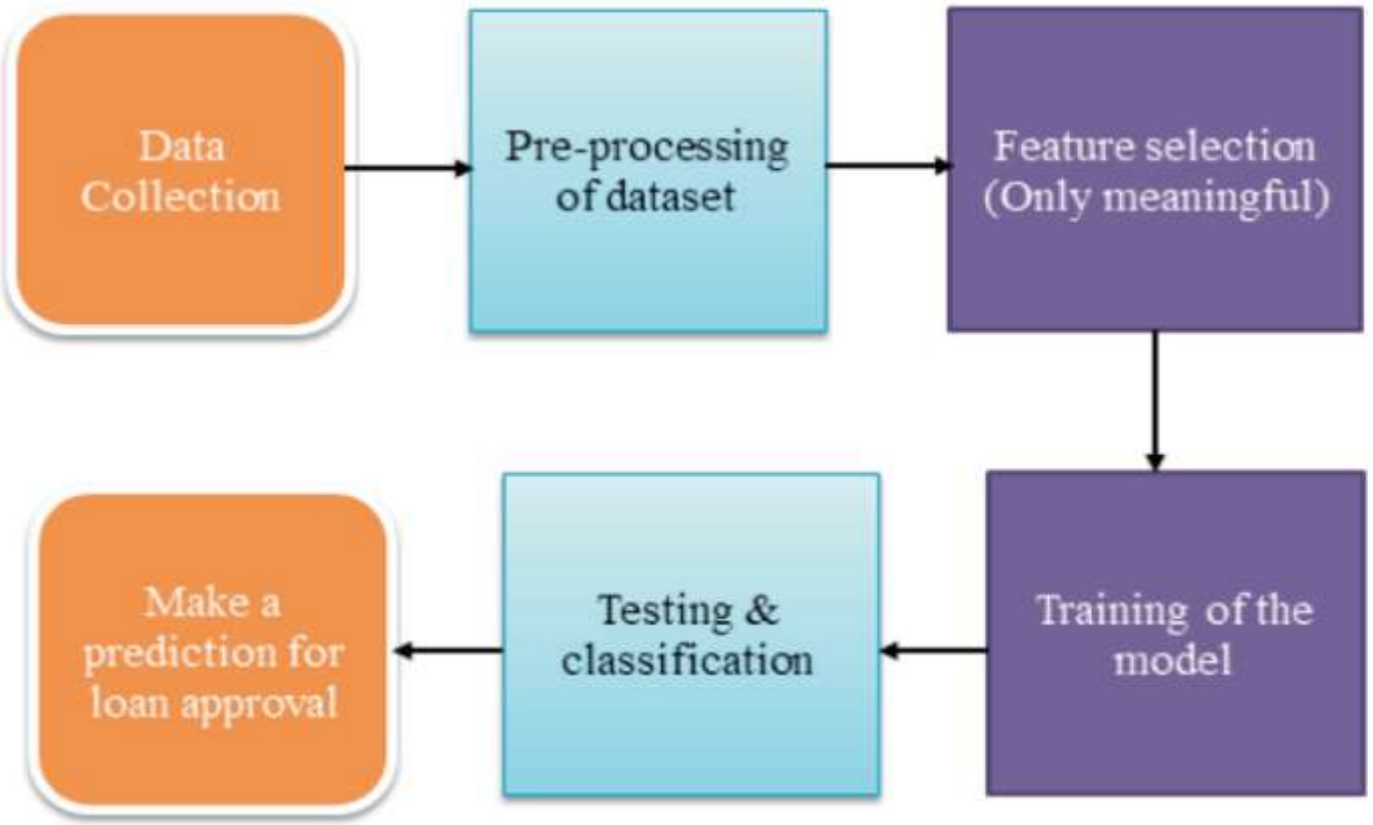
5. Model Evaluation

The models are evaluated based on their accuracy scores on the test set to determine their generalization capability.

6. Conclusion

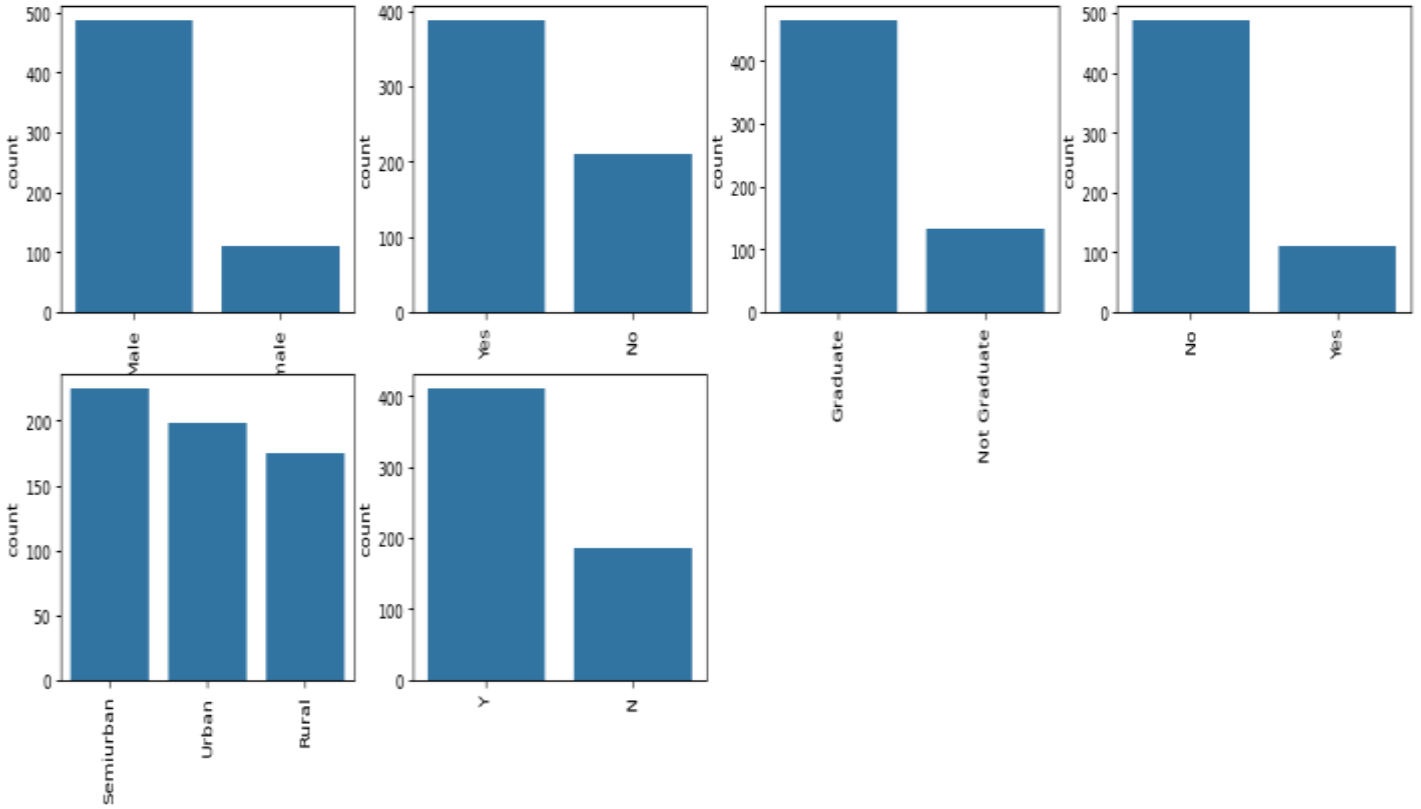
The Random Forest classifier outperforms other models with an accuracy of 82.5% on the test set, suggesting its suitability for the loan approval prediction task. Further enhancements could include using ensemble techniques like bagging and boosting to improve accuracy and robustness.

By following this methodology, banks can leverage machine learning to automate and optimize their loan approval process, resulting in more efficient operations and better risk management.



CHAPTER-3

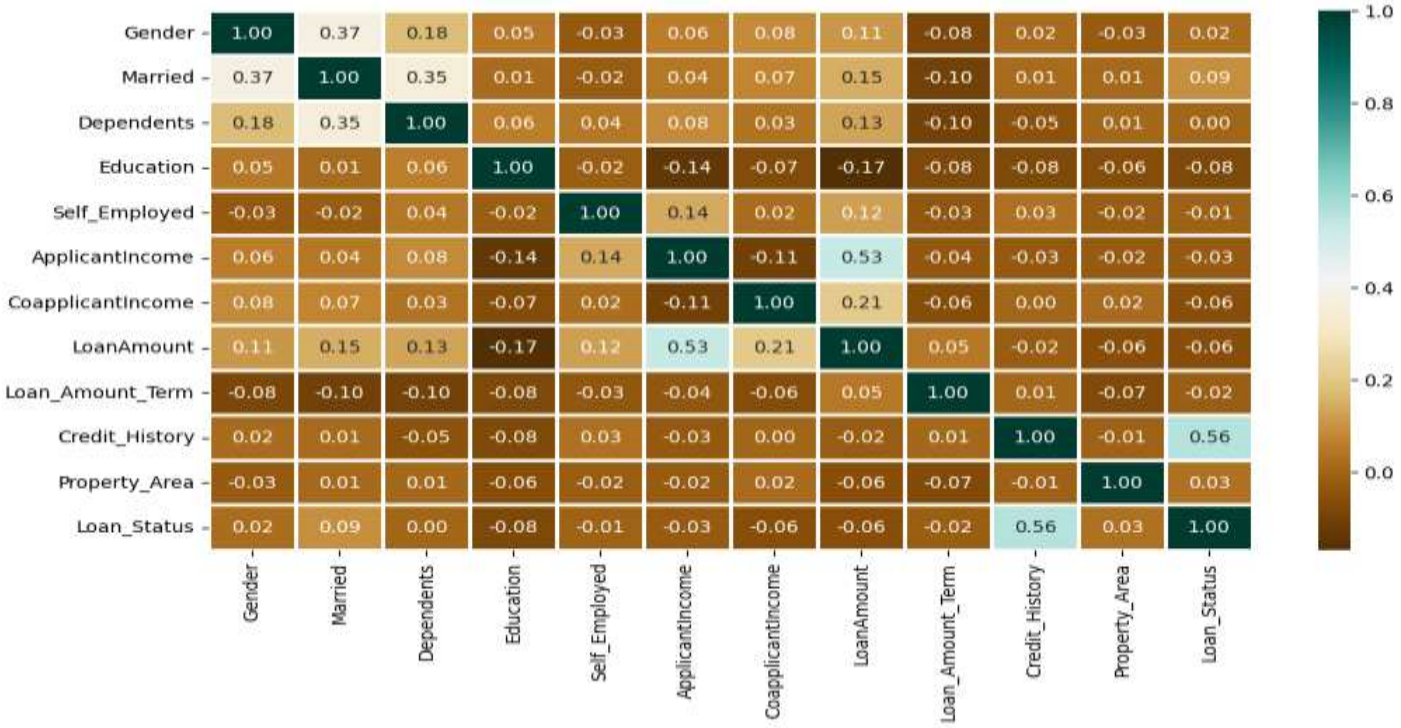
RESULTS



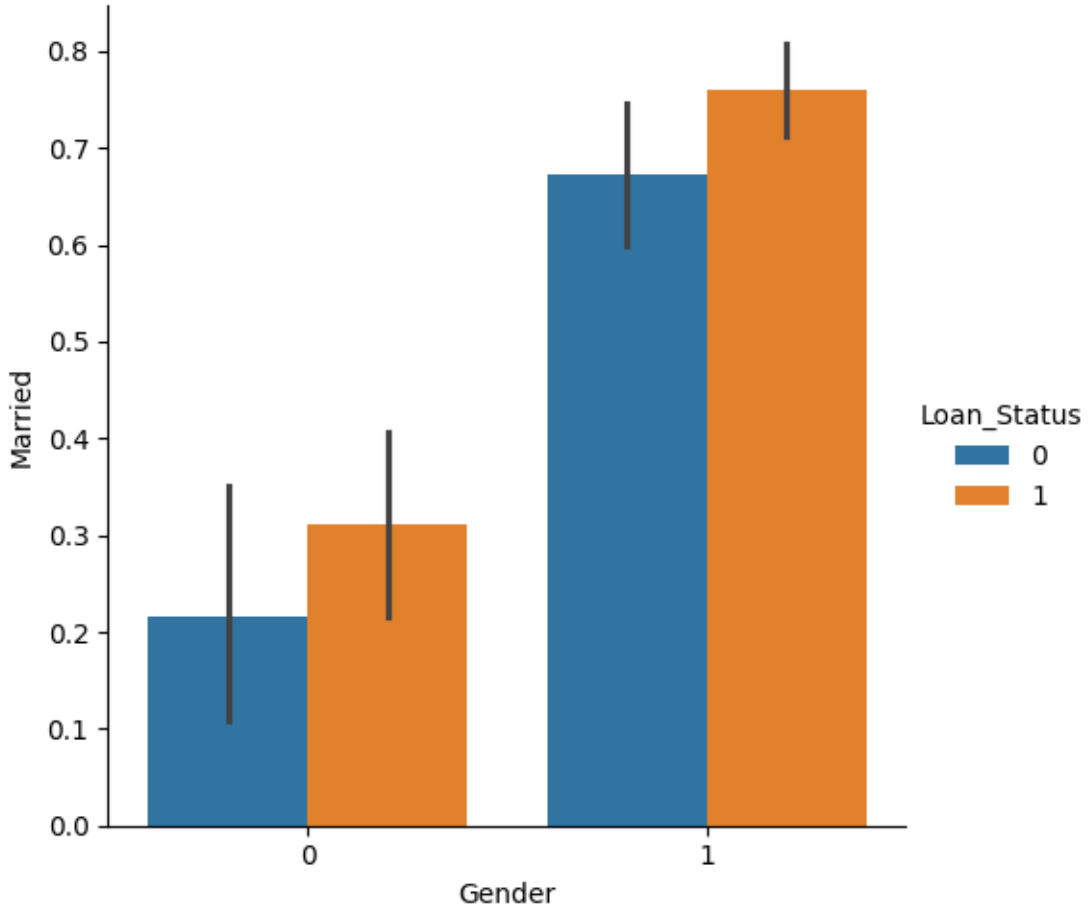
The leftmost chart breaks down cases by residence (rural, urban, and semi-urban). It appears that urban areas have the most diagnosed cases, followed by semi-urban and then rural areas.

The center chart breaks down cases by sex (male and female). It appears that there are more diagnosed cases in males than females.

The rightmost chart breaks down cases by education level (graduate and not graduate). It appears that there are more diagnosed cases in people who did not graduate than those who did.



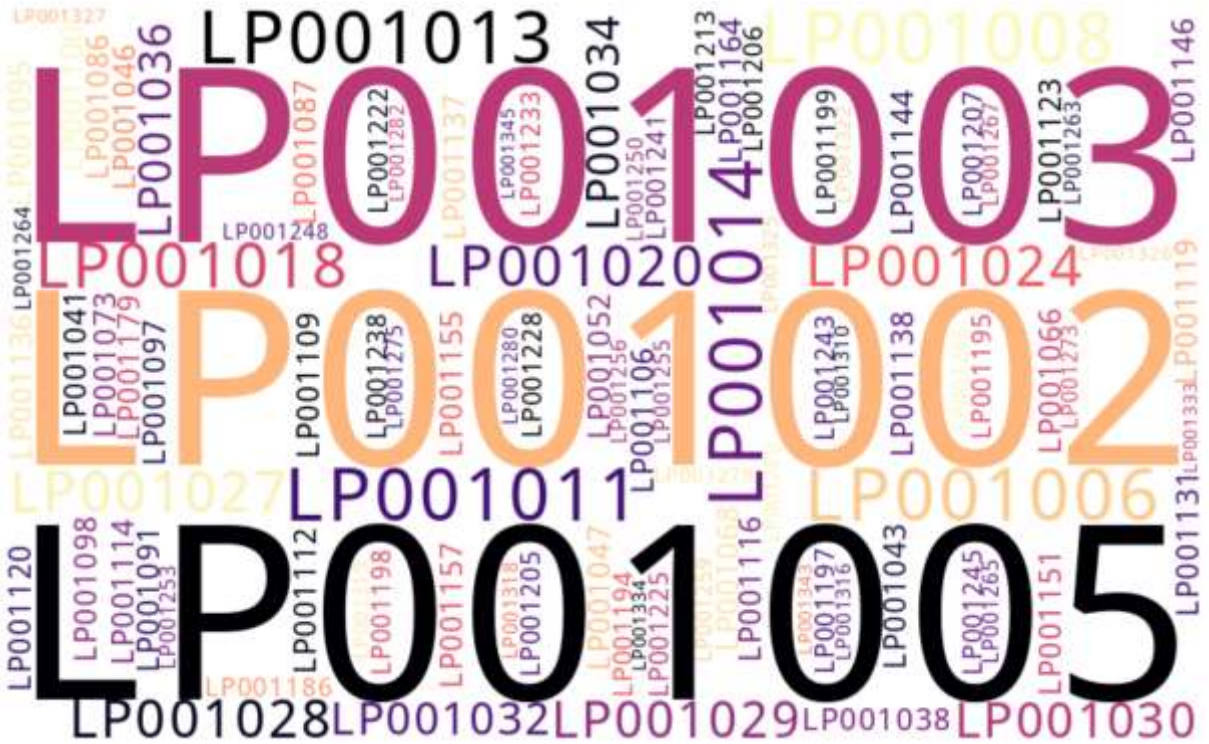
In this specific heatmap, the columns represent features of loan applicants, and the rows represent other features of loan applicants. The color intensity in each cell corresponds to the correlation between the two features. A positive correlation is shown in shades of red, and a negative correlation is shown in shades of blue. A value closer to 1 indicates a stronger positive correlation, and a value closer to -1 indicates a stronger negative correlation. A value close to zero indicates a weak correlation.



The y-axis shows values from 0 to 0.8, but none of the bars reach that high. The highest value seems to be around 0.3.

The x-axis labels are "Loan_Status 0" and "Loan_Status 1" but it isn't clear what these labels mean. It typically wouldn't make sense to have a loan status of 0 or 1.

There is also a label "Gender" below the x-axis, but the bars are stacked on top of each other so it isn't possible to tell which bar corresponds to which gender.



A word cloud, also known as a tag cloud, is a visual representation of text data where the size of each word indicates its frequency or importance. a word cloud can help you quickly identify the most common words or phrases in the dataset. So we have created a variable and stored enormous of words .

making predictions on the training set:

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 78.49162011173185
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 80.44692737430168
```

making predictions on the testing set:

```
Accuracy score of  RandomForestClassifier = 82.5
Accuracy score of  KNeighborsClassifier = 63.74999999999999
Accuracy score of  SVC = 69.16666666666667
Accuracy score of  LogisticRegression = 80.83333333333333
```

CHAPTER-4

CONCLUSION

The journey of constructing a machine learning model for loan approval prediction has traversed several critical stages, each contributing to the robustness and efficacy of the final predictive system. Commencing with the comprehensive collection of pertinent loan application data, the endeavor delved into the intricacies of data preprocessing, exploratory data analysis, model selection, training, and evaluation.

In the realm of data preprocessing, meticulous attention was paid to ensuring the integrity and completeness of the dataset. Through the judicious application of techniques such as dropping irrelevant columns, encoding categorical variables, handling missing values, and feature scaling, the data was primed for subsequent analysis and modeling endeavors. This preparatory phase not only laid the groundwork for meaningful insights but also mitigated the risk of model bias and inaccuracies stemming from data anomalies.

The subsequent exploration of the dataset via exploratory data analysis (EDA) unveiled compelling patterns, distributions, and correlations among various features. Visualizations such as bar plots, heatmaps, and correlation analyses illuminated the intricate interplay between demographic, financial, and credit-related variables, shedding light on the factors most influential in loan approval decisions. Insights gleaned from EDA provided invaluable guidance in feature selection, model formulation, and interpretability, thus enriching the subsequent modeling process.

Model selection and training constituted a pivotal phase wherein diverse machine learning algorithms were harnessed to harness the predictive potential latent within the dataset. Through the judicious application of algorithms ranging from K-Nearest Neighbors to Random Forests, Support Vector Classifiers, and Logistic Regression, the predictive landscape was meticulously explored, with each model vying to encapsulate the multifaceted dynamics underlying loan approval decisions. The iterative refinement of model hyperparameters, cross-validation strategies, and ensemble techniques contributed to the optimization of predictive performance and generalization capacity.

Model evaluation served as the litmus test for the efficacy and robustness of the developed predictive framework. Rigorous assessment of model accuracy, precision, recall, and F1 scores on both training and testing datasets provided a comprehensive appraisal of each model's predictive prowess. The Random Forest classifier emerged as the frontrunner, boasting the highest accuracy of 82.5% on the test set, thereby affirming its supremacy in capturing the complex interplay of features governing loan approval outcomes.

The culmination of this endeavor underscores the transformative potential of machine learning in revolutionizing traditional banking practices. By harnessing the power of predictive analytics, financial institutions stand to gain unprecedented insights into borrower behavior, credit risk, and loan portfolio management. The automation of loan approval processes not only enhances operational efficiency and scalability but also fosters greater inclusivity, fairness, and transparency in lending practices.

Looking ahead, the quest for continuous refinement and innovation remains paramount. Further enhancements, including the integration of advanced ensemble learning techniques, feature engineering strategies, and model interpretability frameworks, hold the promise of unlocking even greater predictive accuracy and actionable insights. As the financial landscape evolves in tandem with technological advancements, the journey towards data-driven decision-making in lending stands poised to usher in a new era of efficiency, equity, and excellence.