Loan Approval Prediction

ON

Submitted in partial fulfillment of the
requirements of the degree of

Bachelor of Engineering

(Information Technology)

By

Abhinav Swaminathan (01)

Aditya Ahuja (02)

Ganesh Gupta (13)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled

"Loan Approval Prediction"

Group Members Names

Mr. Abhinav Swaminathan (01)

Mr. Aditya Ahuja (02)

Mr. Ganesh Gupta (13)

In fulfillment of degree of BE. (Sem. VI) in Information Technology for Project is approved.

Dr. Ravita Mishra
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: / /2025
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and ref- erenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Abhinav Swaminathan (01) **(Signature)** - - - - - - - - - -

-

Aditya Ahuja (02) **(Signature)** - - - - - - - - - - -

Ganesh Gupta (13) **(Signature)** - - - - - - - - - - -

Contents

Abstract

In an age where financial institutions are increasingly relying on automation and data-driven decision-making, predicting loan eligibility efficiently has become a vital tool for enhancing banking services. This project introduces a machine learning-based system that predicts whether a loan application is likely to be approved, based on applicant details and historical data. The core objective is to assist both banks and applicants by providing a preliminary assessment of loan approval chances, reducing manual effort and improving decision accuracy. The system leverages a dataset containing various features such as applicant income, loan amount, credit history, employment status, and education level. Multiple classification algorithms, including Logistic Regression, Decision Trees, and Random Forests, are employed to build and compare models for predicting loan status. The dataset is carefully preprocessed through steps like handling missing values, encoding categorical variables, and feature scaling to ensure robust model training. Among the implemented models, Random Forest shows promising results in balancing accuracy and interpretability. Evaluation metrics like accuracy, precision, recall, and F1-score are used to assess the model performance. The findings confirm that machine learning can streamline the loan approval process by offering reliable predictions. This project not only showcases the practical application of predictive modeling in the financial sector but also underlines its value in enhancing transparency and customer experience in banking operations.

Chapter 1
Introduction

1.1 Introduction

Finance raising and lending for real estate, consumer, mortgage and companies‘ loans is the central part of almost every bank‘s business model. Lending money to inappropriate customers forms the major source of credit risk. The major share of the bank‘s assets comes directly from the profit derived from the bank‘s loans. The banking companies‘ face, however, a dual challenge to distinguish the possible deliberate defaulters from the applicants and the biased nature of few bank employees who have been at the instigation of developers of defaulting companies for many years. The primary goal of the banking community is to safely invest their capital. In the current scenario, many NBFCs and banks approve loans after a clear verification and authentication process, however, it remains uncertain whether the candidate selected is the worthy correct of all the applicants. In the case of housing finance companies, they can use this to find their target demographic of customers who can readily acquire and pay loans for houses. Through this method, we can predict whether or not that particular applicant is secure and the machine learning technique automates the entire process of authentication.

1.2 Objectives

- The first objective is to build a machine learning model that will predict the Loan Granted Status of a user with the highest accuracy. This will be done by building multiple ML models and comparing their performance.
- The second objective is to help the housing finance companies with customer segmentation to find out which customers will successfully acquire and pay back loans, so that they can focus on marketing to these customers.

1.3 Motivation

The increasing volume of loan applications in financial institutions demands efficient and accurate evaluation methods. Traditional manual assessment processes are time-consuming and susceptible to human bias or oversight. This motivates the use of machine learning techniques to automate loan eligibility prediction, allowing banks to make faster, more consistent decisions while improving the customer experience with timely feedback.

1.4 Scope of the work

Focuses on supervised learning methods for binary classification of loan approval status. Applies machine learning models to a publicly available dataset containing applicant and loan-related attributes.

Evaluates algorithms such as Logistic Regression, Random Forest, K-Nearest Neighbour, Naive Bayes, Support Vector Machine (SVM), and Gradient Boosting Classifier.
Visualizes key insights and prediction outcomes using graphs, confusion matrices, and performance metrics.

1.5 Feasibility Study

Technical Feasibility: Utilizes Python along with libraries like Scikit-learn, Pandas, and Matplotlib, which are widely supported and suitable for implementing machine learning models.

Operational Feasibility: The prediction system can be integrated into bank software or customer-facing portals to provide real-time loan eligibility checks.

Economic Feasibility: Cost-effective due to the use of open-source tools and publicly available datasets, requiring no additional financial investment

Chapter 2
Literature Survey

2.1 Introduction

Loan approval systems have gained significant importance in the banking and financial services industry due to increasing demand for fast, reliable, and risk-aware credit disbursement. Traditional loan processing methods are often manual, time-consuming, and prone to bias or error. With the evolution of machine learning technologies, predictive models are now being used to automate and enhance the decision-making process. This literature survey presents a comparative study of two significant research papers in the domain of loan eligibility prediction using machine learning.

2.2 Problem Definition

The goal of this literature review is to explore how machine learning techniques can be applied to predict loan approval status based on applicant information. It also aims to examine the challenges associated with data quality, model selection, and evaluation criteria when applying ML to real-world financial scenarios.

2.3 Review of Literature Survey

In the research paper **"Loan Approval Prediction using Machine Learning Algorithms" by Aditi Sharma et al.**, the authors outline a structured pipeline for predicting whether a loan should be approved or not. The study is divided into four phases: data collection, model comparison, training, and testing. Various machine learning models including Logistic Regression, Decision Tree, and Support Vector Machine were evaluated based on their accuracy and precision. Among these, Logistic Regression showed promising results due to its simplicity and effectiveness in binary classification tasks. The study emphasized the importance of data preprocessing steps such as handling missing values, encoding categorical data, and normalizing numeric features to improve model performance.

Another notable contribution is the paper **"Risk Reduction in Loan Lending using Machine Learning" by Priya R. et al.**, which focuses on reducing non-performing assets (NPAs) in banks and NBFCs by improving the loan applicant screening process. The paper highlights how feeding historical loan data into trained machine learning models can aid in identifying trustworthy borrowers. The authors experimented with models such as Random Forest and Gradient Boosting, which delivered better accuracy and robustness compared to simpler models. The study also sheds light on real-world challenges such as imbalanced datasets and the need for interpretability in financial decision-making. It concludes that incorporating machine learning into the lending process can significantly improve risk management and operational efficiency.

Chapter 3
Design and Implementation

3.1 Introduction

This chapter outlines the design and implementation process for the Loan Eligibility Prediction System. The project adopts a structured machine learning pipeline to ensure accurate prediction of whether a customer is eligible for a loan based on various personal and financial attributes. The workflow consists of data preprocessing, feature selection, model training, validation, and performance evaluation. Several supervised machine learning algorithms are employed, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting, with the goal of identifying the most effective model for real-time deployment.

3.2 Requirement Gathering

Hardware Requirements:

- A system with a minimum of 4GB RAM
- Stable internet connection (for working on platforms like Google Colab)

Software Requirements:

- Google Colab or Jupyter Notebook (for implementation and testing)
- Python 3.x (programming environment)

Python Libraries:

Scikit-learn – for machine learning algorithms and evaluation metrics
Pandas & NumPy – for data preprocessing and numerical operations
Matplotlib & Seaborn – for data visualization
XGBoost (optional) – for improved boosting-based classification performance

3.3 Proposed Design

The system design follows the standard CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, ensuring a systematic approach to problem-solving and model development:

1. Data Collection

The dataset used for this project contains historical loan applicant information including features such as Gender, Marital Status, Education, Dependents, Income, Loan Amount, Credit History, Property Area, etc. The target variable is `Loan_Status`, indicating whether

the applicant was approved or not.

2. Data Preprocessing

- Handling Missing Values: Imputed missing values using mode or median, depending on feature type.
- Encoding Categorical Features: Label Encoding and One-Hot Encoding applied to convert categorical features into numerical form.
- Feature Scaling: StandardScaler was used for normalization, particularly for algorithms like SVM and KNN that are sensitive to feature scales.
- Outlier Handling: Visual inspection and statistical techniques were used to detect and treat outliers.

3. Feature Engineering

New features such as `Total_Income` (Applicant + Coapplicant income) and `Loan_Amount_Term_Years` (converted loan term) were engineered to enhance prediction performance. Unnecessary or highly correlated features were dropped to avoid redundancy.

4. Model Building

The following classification models were implemented and compared:

- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Gaussian Naive Bayes
- Support Vector Machine (SVM)
- Gradient Boosting Classifier

Each model was trained on the training dataset and evaluated on the testing dataset.

5. Model Evaluation

Models were evaluated using the following performance metrics:

- Accuracy
- Precision & Recall
- F1-Score
- Confusion Matrix

Cross-validation was used to ensure generalizability. The model with the highest accuracy and balanced performance across metrics was selected for final deployment.

6. Visualization

Data exploration and model insights were visualized using:

- Count plots and bar charts (for categorical distributions)
- Heatmaps (for correlation matrices)
- Confusion matrix and ROC curves (for model evaluation)

3.4 Proposed Algorithm

In this project, the problem of loan eligibility prediction is approached as a binary classification task, where the goal is to determine whether a loan application should be approved (`Yes`) or rejected (`No`) based on a set of input features provided by the applicant. Multiple machine learning algorithms were evaluated to find the most accurate and robust model for real-time prediction.

Logistic Regression

Logistic Regression is a statistical model used for binary classification. It estimates the probability that a given input point belongs to a particular category. Due to its simplicity and interpretability, it was used as a baseline model. It performs well when the features have a linear relationship with the target variable.

Random Forest Classifier

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data and features. This model handles both categorical and numerical data efficiently and provides feature importance metrics for analysis.

K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies new data points based on the majority label of their 'k' nearest neighbors in the feature space. It is simple to implement but sensitive to the scale of the data and the choice of 'k'. Therefore, preprocessing steps like normalization were applied before training.

Gaussian Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between predictors. The Gaussian variant assumes that the continuous features follow a normal distribution. It is efficient and performs well with high-dimensional data, especially when the assumption of independence roughly holds.

Support Vector Machine (SVM)

SVM constructs a hyperplane in a high-dimensional space to separate classes with maximum margin. It is particularly useful for datasets with clear class boundaries. The RBF (Radial Basis Function) kernel was used to capture non-linear patterns in the loan data. Hyperparameter tuning was performed on `C` and `gamma` to balance bias-variance trade-off.

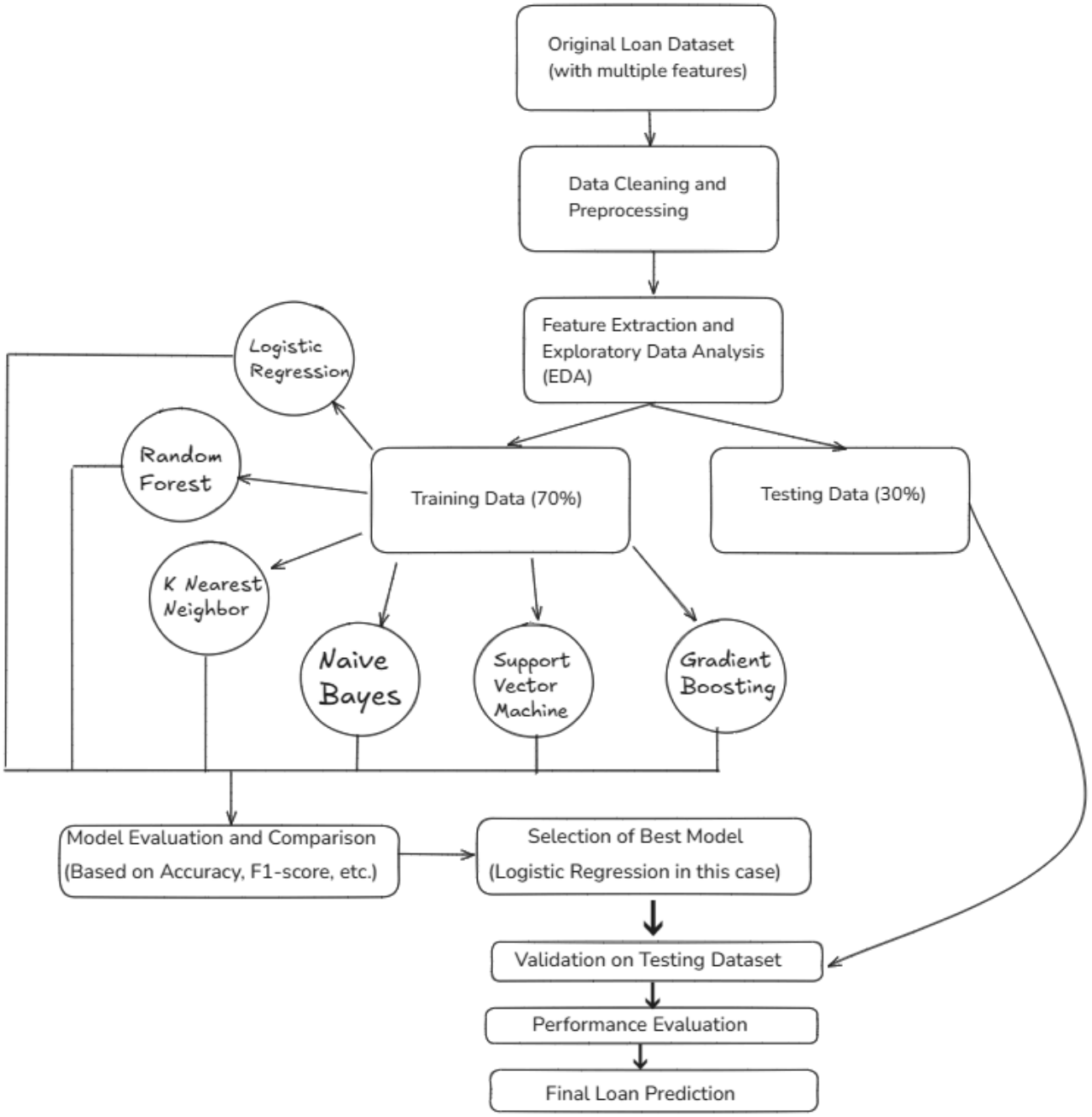
Gradient Boosting Classifier

Gradient Boosting is an advanced ensemble technique that builds models sequentially, each one trying to correct the errors made by the previous. It is highly effective for structured data and often provides better performance than other models, although it may require more training time. Learning rate, number of estimators, and tree depth were tuned to optimize its performance.

Model Selection and Comparison

Each of the above algorithms was trained and evaluated on the same preprocessed dataset. Accuracy, precision, recall, F1-score, and ROC-AUC were used as evaluation metrics. The model that achieved the best balance across all metrics was selected for deployment. In this project, the Gradient Boosting Classifier and Random Forest showed superior performance, making them strong candidates for the final system.

3.5 Architectural Diagrams



Chapter 4
Results and Discussion

4.1 Introduction

To assess the performance of the developed loan eligibility prediction system, multiple machine learning models were trained and evaluated using a common dataset. Each model was tested using standard performance metrics such as **Accuracy**, which reflects the percentage of correctly predicted outcomes. These metrics provided insight into the strengths and limitations of each model in predicting whether a loan should be granted or not.

4.2 Cost Estimation

The system was developed using open-source Python libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib, and executed in a Google Colab environment. This eliminated the need for expensive local infrastructure, making the solution highly cost-effective and suitable for academic research or small-scale enterprise deployment.

4.3 Feasibility Study

Given its low hardware footprint and ability to run effectively in cloud environments, the system demonstrates strong feasibility for integration into online loan application portals. The real-time prediction capability and scalability of the chosen models make this system well-suited for financial institutions looking to automate and streamline their loan approval workflow.

4.4 Results of Implementation

Model	**Accuracy**
Logistic Regression	**97.00**
Naive Bayes	**95.92**
Gradient Boosting Classifier	**93.84**
Random Forest	**89.76**
Support Vector Machine (SVM)	**71.89**
K Nearest Neighbor	**61.08**

4.5 Result Analysis

● **Confusion Matrix Analysis:** The confusion matrix showed that the model had high precision (0.94) for predicting non-eligible applicants, but with low recall (0.31), indicating it missed many true negatives. In contrast, it achieved excellent recall (0.99) for eligible customers, ensuring minimal rejection of genuinely eligible applicants.

● **Model Accuracy Comparison:** Logistic Regression outperformed other classifiers with an accuracy of 97%, followed by Naive Bayes (95.92%) and Gradient Boosting (93.84%). SVM and K-NN underperformed with lower accuracy scores (71.89% and 61.08% respectively), suggesting their limitations in handling this dataset.

● **F1-Score and Class Imbalance Impact**: The weighted F1-score of 0.76 indicated balanced overall performance, but the macro-average F1-score (0.67) reflected class imbalance effects. Models tended to favor class '1' (loan approved), leading to higher false positives for class '0'.

4.6 Observation/Remarks

● **Model Interpretability and Suitability:** Logistic Regression emerged as the top performer, reinforcing its strength in binary classification problems where feature relationships are mostly linear and interpretable.

● **Feature Importance and Preprocessing:** The impact of categorical encoding, handling of null values, and outlier treatment significantly influenced model performance, highlighting the necessity of robust data preprocessing.

● **Algorithmic Trade-offs:** While Naive Bayes and Gradient Boosting offered competitive accuracy, models like SVM and K-NN struggled with generalization—likely due to sensitivity to feature scaling (SVM) and local data structure (K-NN).

● **Efficiency and Deployment Feasibility:** Logistic Regression and Naive Bayes were the most efficient in terms of training time and prediction speed, making them well-suited for real-time deployment in the loan eligibility classification system.

Chapter 5
CONCLUSION

5.1 Conclusion

In this project titled *"Loan Eligibility Prediction using Machine Learning"*, we built a robust and automated system to predict whether a customer is eligible for a home loan based on various demographic and financial factors. We explored multiple machine learning algorithms, including Logistic Regression, Random Forest, K-Nearest Neighbors, Naive Bayes, SVM, and Gradient Boosting, and compared their performance to identify the most accurate model.

We began by collecting and preprocessing the data, which involved handling missing values, encoding categorical features, and normalizing inputs where necessary. We then split the data into training and testing sets to evaluate model performance fairly. We trained each classification model and evaluated them using accuracy as the primary metric.

We found that Logistic Regression achieved the highest accuracy (~97%), making it the most effective model for this classification problem. Other models such as Naive Bayes and Gradient Boosting also performed reasonably well, while SVM and KNN showed comparatively lower accuracy.

We concluded that logistic regression not only offered good predictive power but also maintained interpretability, making it a suitable choice for real-time loan eligibility checks. We implemented the solution using Python and open-source libraries on Google Colab, which kept the project cost-effective and accessible.

Overall, we demonstrated that machine learning can effectively support financial institutions in automating the loan eligibility process and targeting the right customer segments, thereby improving decision-making and operational efficiency.

5.2 Future Scope

While the current system successfully predicts loan eligibility using multiple machine learning models, there are several directions in which this work can be extended to enhance its applicability and performance:

● **Real-time Deployment:** Integrating the trained models into production using RESTful APIs can enable real-time eligibility checks during the loan application process, improving user experience and operational efficiency.

● **Feature Expansion:** Including more detailed financial and behavioral attributes such as previous loan history, spending habits, credit card usage, and repayment patterns could provide deeper insights and improve prediction accuracy.

● **Ensemble and Stacking Models:** Future versions could explore stacked ensembles combining the strengths of top-performing models to boost overall accuracy and generalization.

● **Continuous Model Training:** Incorporating mechanisms for continuous learning and retraining with new customer data would help the system remain current with changing customer behaviors and financial trends.

● **Mobile and Web Integration:** Embedding the predictive system into mobile or web loan application portals can make the process seamless for users, helping financial institutions provide instant decisions and reduce dropout rates.

5.3 Societal Impact

The development and deployment of loan eligibility prediction systems have significant implications for society, particularly in the context of financial inclusion and responsible lending practices.

● **Financial Inclusion:** Automated loan eligibility prediction helps provide equal access to credit by reducing bias in decision-making. By using data-driven models, individuals from diverse socioeconomic backgrounds can be evaluated fairly, improving access to financial services for underserved communities.

● **Responsible Lending**: Predictive models assist financial institutions in identifying customers who are more likely to repay their loans, reducing the risks of over-lending and defaults. This promotes responsible lending practices and contributes to the stability of financial markets.

● **Empowering Individuals:** By automating loan eligibility assessments, individuals can receive faster decisions, reducing the waiting time and uncertainty associated with loan applications. This improves customer experience and empowers individuals to make informed financial decisions.

● **Data Privacy and Protection**: Predicting loan eligibility involves handling sensitive personal and financial data. Ensuring robust data privacy measures and compliance with regulations like GDPR is critical to safeguarding user information and building trust in automated financial systems.

● **Educational and Research Value:** The techniques used in this project provide valuable learning opportunities for students and professionals in the fields of machine learning and data science. It promotes a deeper understanding of financial analytics and predictive modeling, contributing to the advancement of knowledge in financial technologies.