**TO PREDICT LOAN APPROVAL USING APPLICANT INCOME, LOAN AMOUNT, CREDIT HISTORY, AND OTHER PERSONAL FACTORS**

Table of Contents

[Introduction 1](#_Toc184791389)

[Background 1](#_Toc184791390)

[Motivation 2](#_Toc184791391)

[Data Summary: 3](#_Toc184791392)

[Statistical Summary and Correlation Matrix 5](#_Toc184791393)

[Methodology 6](#_Toc184791394)

[Results 8](#_Toc184791395)

[Logistic Regression 8](#_Toc184791396)

[Decision Tree 8](#_Toc184791397)

[Random Forest 9](#_Toc184791398)

[K-Nearest Neighbors (KNN) 9](#_Toc184791399)

[Support Vector Machine (SVM) 9](#_Toc184791400)

[Accuracy Scores 10](#_Toc184791401)

[Cross-Validation Results for Model Accuracy 11](#_Toc184791402)

[Model Selection 11](#_Toc184791403)

[Conclusion 11](#_Toc184791404)

[Findings 11](#_Toc184791405)

[Future Works 12](#_Toc184791406)

# 

# Introduction

The process of loan approval is a critical decision-making step in the financial industry. Traditionally, lenders assess a variety of factors such as an applicant's income, loan amount, credit history, and other personal information to determine whether a loan application will be approved. Given the increasing volume of loan applications and the need for quick, accurate decision-making, leveraging machine learning techniques for automating this process has gained significant importance.

This report outlines the development of a Loan Approval Prediction Model using machine learning algorithms in Jupyter Notebook. The goal is to predict whether a loan application will be approved or not based on various factors, including the applicant's income, loan amount, credit history, and other relevant attributes. The model aims to help financial institutions streamline the loan approval process, reduce human error, and improve decision-making efficiency.

Several machine learning techniques, including Random Forest, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN), were evaluated for their performance in predicting loan approval. The evaluation was based on metrics such as accuracy, confusion matrix, and classification report, along with cross-validation to ensure model robustness.

# Background

Loan approval is a fundamental aspect of the financial industry, where lenders assess an applicant's ability to repay a loan based on various factors such as income, credit history, loan amount, and personal details. Traditionally, this process has been manual, relying on human judgment to interpret these factors and make decisions. However, as the volume of loan applications increases and borrower profiles become more complex, the traditional approach becomes less efficient and prone to inconsistencies.

In recent years, machine learning has emerged as a powerful tool to automate and enhance decision-making in various industries, including finance. By training algorithms on historical data, machine learning models can predict the likelihood of loan approval more accurately and quickly than manual processes. These models analyze numerous features, such as applicant income, loan amount, and credit history, to classify loan applications into approved or denied categories based on patterns learned from past decisions.

This shift towards data-driven decision-making allows financial institutions to streamline the loan approval process, reduce human error, and improve overall efficiency. In this project, a Loan Approval Prediction Model was built using machine learning algorithms, with the goal of predicting whether a loan application will be approved based on several key factors. The implementation of such predictive models promises to support more informed and reliable lending decisions in the future.

# Motivation

1. **Improve Loan Approval Efficiency:**

Automating the loan approval process using machine learning models ensures quicker and more consistent decision-making. With the ability to process vast amounts of data in real-time, financial institutions can reduce the time taken for each loan application, providing applicants with faster responses and improving overall operational efficiency.

1. **Enhance Decision Making:**

By replacing traditional methods with data-driven algorithms, the prediction model allows financial institutions to make more informed and objective decisions. Machine learning models, trained on historical data, minimize human biases and errors, making the lending process more rational and improving the accuracy of loan approvals, which is crucial for reducing bad debt risks.

1. **Reduce Operational Costs:**

Through automation, the model reduces the need for extensive manual intervention, thus lowering operational costs. Financial institutions can save time and resources spent on manual evaluations, while also avoiding the costly errors associated with human judgment. The streamlined process allows them to allocate resources more effectively and improve profitability.

1. **Adapt to Economic Changes:**

The financial landscape is constantly evolving, with changes in the economy, regulations, and technology. Machine learning models can be updated with new data, enabling financial institutions to remain competitive and adaptable. The ability to incorporate new information into predictive models ensures that lenders can better anticipate market shifts and adjust their strategies accordingly.

1. **Improve Risk Management:**

By predicting the likelihood of loan approvals based on factors such as applicant income and credit history, financial institutions can identify potential risks early in the process. The model helps mitigate risks by flagging high-risk applicants, allowing institutions to tailor their loan offerings and minimize exposure to defaults, thus maintaining a healthy loan portfolio.

# Data Summary:

**Data Source**

The data used in this analysis is sourced from a Google Drive link. It consists of loan application data, containing various features such as gender, marital status, education, income, loan amount, credit history, and other factors relevant to loan approval. This dataset serves as a basis for building a predictive model for loan approval prediction, leveraging machine learning techniques.

Data source: <https://drive.google.com/file/d/1X_A4WaAfhtK5-icyix_TRuxZtzeMtetg/view>

**Data Details**

|  |  |
| --- | --- |
| Total Number of Rows | 614 |
| Total Number of Columns | 13 |

**Target Variable: Attrition (YES/NO)**

|  |  |
| --- | --- |
| Loan Not Approved | 0 |
| Loan Approved | 1 |

**FEATURE VARIABLE DEFINITION**

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **DEFINITION** | **NOTES** |
| **Gender** | The gender of the applicant (0 = Female, 1 = Male) | Binary variable representing the gender of the applicant. |
| **Married** | The marital status of the applicant (0 = No, 1 = Yes) | Binary variable indicating whether the applicant is married. |
| **Dependents** | Number of dependents (0, 1, 2, 3+, where 3+ represents more than three dependents) | Discrete variable representing the number of dependents. |
| **Education** | The education level of the applicant (0 = Not Graduate, 1 = Graduate) | Binary variable indicating whether the applicant is a graduate or not. |
| **Self Employed** | Whether the applicant is self-employed (0 = No, 1 = Yes) | Binary variable indicating if the applicant is self-employed. |
| **Applicant Income** | The income of the applicant (in currency units, e.g., USD) | Continuous variable representing the income of the applicant. |
| **Co-applicant Income** | The income of the coapplicant (in currency units, e.g., USD) | Continuous variable representing the income of the coapplicant. |
| **Loan Amount** | The loan amount applied for (in currency units, e.g., USD) | Continuous variable representing the loan amount requested. |
| **Loan Amount Term** | The term of the loan in months (e.g., 360 months or 12 months) | Continuous variable representing the length of the loan term. |
| **Credit History** | The credit history of the applicant (1 = No Dues, 0 = Dues) | Binary variable indicating whether the applicant has cleared their previous loans. |
| **Property Area** | The area where the property is located (0 = Urban, 1 = Semiurban, 2 = Rural) | Categorical variable representing the type of area (Urban/Semiurban/Rural). |
| **Loan Status** | The target variable indicating if the loan was approved (0 = Not Approved, 1 = Approved) | Binary target variable indicating whether the loan was approved or not. |

**Data Cleaning**

The data cleaning process involved several key steps to ensure the dataset’s quality. First, missing values were handled by imputing numerical columns like Applicant Income and Loan Amount with the median, and categorical columns such as Self Employed and Education with the mode. Duplicates were checked and removed to maintain data integrity. Categorical variables like Gender, Married, and Property Area were encoded using binary or one-hot encoding. Outliers were detected using the Interquartile Range (IQR) method and were either capped or removed to minimize their impact on the analysis. Additionally, numerical features were scaled using StandardScaler to ensure consistency in data ranges, especially for models sensitive to feature scale. After these cleaning steps, the dataset was thoroughly checked for completeness and consistency, making it ready for analysis and model development.

# Statistical Summary and Correlation Matrix

**Descriptive Statistics of the Cleaned Dataset**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Gender** | 614 | 0.82 | 0.39 | 0 | 1 | 1 | 1 | 1 |
| **Married** | 614 | 0.65 | 0.48 | 0 | 0 | 1 | 1 | 1 |
| **Dependents** | 614 | 0.74 | 1.01 | 0 | 0 | 0 | 1 | 3 |
| **Education** | 614 | 0.22 | 0.41 | 0 | 0 | 0 | 0 | 1 |
| **Self Employed** | 614 | 0.13 | 0.34 | 0 | 0 | 0 | 0 | 1 |
| **Applicant Income** | 614 | 5403.46 | 6109.04 | 150 | 2877 | 3813 | 5795 | 81000 |
| **Co-applicant Income** | 614 | 1621.25 | 2926.25 | 0 | 0 | 1189 | 2297 | 41667 |
| **Loan Amount** | 614 | 145.75 | 84.11 | 9 | 100.25 | 128 | 164.75 | 700 |
| **Loan Amount Term** | 614 | 342.41 | 64.43 | 12 | 360 | 360 | 360 | 480 |
| **Credit History** | 614 | 0.86 | 0.35 | 0 | 1 | 1 | 1 | 1 |
| **Property Area** | 614 | 1.04 | 0.79 | 0 | 0 | 1 | 2 | 2 |
| **Loan Status** | 614 | 0.69 | 0.46 | 0 | 0 | 1 | 1 | 1 |

**Correlation Matrix**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Gender** | **Married** | **Dependents** | **Education** | **Self Employed** | **Applicant Income** | **Co-applicant Income** | **Loan Amount** | **Loan Amount Term** | **Credit History** | **Property Area** | **Loan Status** |
| **Gender** | 1.000 | 0.365 | 0.173 | 0.045 | -0.001 | 0.059 | 0.083 | 0.107 | -0.074 | 0.009 | -0.026 | 0.018 |
| **Married** | 0.365 | 1.000 | 0.334 | 0.012 | 0.004 | 0.052 | 0.076 | 0.147 | -0.101 | 0.011 | 0.004 | 0.091 |
| **Dependents** | 0.173 | 0.334 | 1.000 | 0.056 | 0.057 | 0.118 | 0.030 | 0.163 | -0.104 | -0.040 | -0.000 | 0.010 |
| **Education** | 0.045 | 0.012 | 0.056 | 1.000 | -0.010 | -0.141 | -0.062 | -0.169 | -0.074 | -0.074 | -0.065 | -0.086 |
| **Self Employed** | -0.001 | 0.004 | 0.057 | -0.010 | 1.000 | 0.127 | -0.016 | 0.115 | -0.034 | -0.002 | -0.031 | -0.004 |
| **Applicant Income** | 0.059 | 0.052 | 0.118 | -0.142 | 0.127 | 1.000 | -0.117 | 0.565 | -0.047 | -0.019 | -0.010 | -0.005 |
| **Co-applicant Income** | 0.083 | 0.076 | 0.030 | -0.062 | -0.016 | -0.117 | 1.000 | 0.189 | -0.059 | 0.011 | 0.011 | -0.059 |
| **Loan Amount** | 0.107 | 0.147 | 0.163 | -0.170 | 0.115 | 0.565 | 0.189 | 1.000 | 0.037 | -0.000 | -0.047 | -0.033 |
| **Loan Amount Term** | -0.074 | -0.101 | -0.104 | -0.074 | -0.034 | -0.047 | -0.059 | 0.037 | 1.000 | -0.005 | -0.076 | -0.023 |
| **Credit History** | 0.009 | 0.011 | -0.040 | -0.074 | -0.002 | -0.019 | 0.011 | -0.000 | -0.005 | 1.000 | 0.002 | 0.541 |
| **Property Area** | -0.026 | 0.004 | -0.000 | -0.065 | -0.031 | -0.010 | 0.011 | -0.047 | -0.076 | 0.002 | 1.000 | 0.032 |
| **Loan Status** | 0.018 | 0.091 | 0.010 | -0.086 | -0.004 | -0.005 | -0.059 | -0.033 | -0.023 | 0.541 | 0.032 | 1.000 |

# Methodology

**Predictive Model**

This analysis focuses on building a model to predict whether a loan application will be approved or rejected. The model utilizes machine learning algorithms to analyze relationships between input features such as applicant income, loan amount, credit history, and other factors to determine the likelihood of loan approval. A cleaned dataset containing independent variables (features) and a single dependent variable (target) was used for model training and evaluation.

**Classification Model**

As the target variable, Loan Status, was binary in nature (1 = Loan Approved, 0 = Loan Not Approved), classification techniques were applied in this study. Five classification models were tested to evaluate and compare their performance:

1. **Logistic regression**

Logistic regression, a statistical method for modeling binary outcomes, was used to predict whether a loan was approved (value of 1) or not approved (value of 0). Features such as Applicant Income, Co-applicant Income, Loan Amount, Credit History, and Property Area were utilized as input variables to forecast loan approval probabilities. Logistic regression estimated the likelihood of loan approval based on these features.

1. **Decision Tree**

The decision tree, a classification technique that organizes the dataset into distinct subcategories using feature variables, was implemented to predict the binary outcome of loan approval status. The decision tree split the data at each branch based on feature variables like Credit History, Loan Amount, and Applicant Income to arrive at predictions. This model used hierarchical splitting to maximize accuracy in predicting loan approval outcomes.

1. **Random Forest**

Random forest, an ensemble method that uses multiple decision trees to improve classification performance, was employed in this analysis. Each tree in the forest predicted loan approval, and the majority vote from all trees determined the final output. By aggregating predictions, random forest reduced overfitting and enhanced accuracy. Features such as Loan Amount, Applicant Income, Credit History, and Loan Amount Term were analyzed across multiple decision trees to determine loan approval status.

1. **K-Nearest Neighbors**

KNN, a distance-based algorithm, was used to classify loan approval outcomes by finding the closest data points (neighbors) in the feature space. The model evaluated the similarity between a new data point and its k-nearest neighbors to predict Loan Status. Features such as Credit History, Loan Amount, and Applicant Income were used to measure the distance and identify patterns in the dataset for classification.

1. **Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a powerful classification algorithm that separates classes by finding the optimal hyperplane. In this analysis, SVM was used to predict loan approval status by classifying data points into two categories: approved (1) or not approved (0). Features such as Loan Amount, Applicant Income, Credit History, and Property Area were employed in the model. SVM maximized the margin between the classes to improve classification accuracy and generalizability, especially for complex decision boundaries.

**Model Evaluation**

To assess the performance of the classification models, accuracy, precision, recall, F1 score, and ROC-AUC metrics were utilized as measures for evaluating loan approval prediction. These metrics allowed for a comprehensive comparison of the Logistic Regression, Decision Tree, Random Forest, support vector machine (SVM), and K-Nearest Neighbors (KNN) models in correctly classifying loan approvals and rejections.

# Results

## Logistic Regression

**Logistic Regression Model Evaluation Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0 (Loan Rejected)** | **1 (Loan Approved)** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| Precision | 0.95 | 0.76 | 0.79 | 0.85 | 0.83 |
| Recall | 0.42 | 0.99 |  | 0.70 | 0.79 |
| F1-Score | 0.58 | 0.86 |  | 0.72 | 0.76 |
| Support | 43 | 80 | 123 |  |  |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0 (Rejected)** | **Predicted 1 (Approved)** |
| Actual 0 | 18 | 25 |
| Actual 1 | 1 | 79 |

## Decision Tree

**Decision Tree Model Evaluation Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0 (Loan Rejected)** | **1 (Loan Approved)** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| Precision | 0.56 | 0.76 | 0.69 | 0.66 | 0.69 |
| Recall | 0.53 | 0.78 |  | 0.65 | 0.69 |
| F1-Score | 0.55 | 0.77 |  | 0.66 | 0.69 |
| Support | 43 | 80 | 123 |  |  |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0 (Rejected)** | **Predicted 1 (Approved)** |
| Actual 0 | 23 | 20 |
| Actual 1 | 18 | 62 |

## Random Forest

**Random Forest Model Evaluation Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0 (Loan Rejected)** | **1 (Loan Approved)** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| Precision | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 |
| Recall | 0.42 | 0.93 |  | 0.67 | 0.75 |
| F1-Score | 0.54 | 0.83 |  | 0.68 | 0.73 |
| Support | 43 | 80 | 123 |  |  |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0 (Rejected)** | **Predicted 1 (Approved)** |
| Actual 0 | 18 | 25 |
| Actual 1 | 6 | 74 |

## K-Nearest Neighbors (KNN)

**K-Nearest Neighbors Model Evaluation Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0 (Loan Rejected)** | **1 (Loan Approved)** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| Precision | 0.81 | 0.75 | 0.76 | 0.78 | 0.77 |
| Recall | 0.40 | 0.95 |  | 0.67 | 0.76 |
| F1-Score | 0.53 | 0.84 |  | 0.68 | 0.73 |
| Support | 43 | 80 | 123 |  |  |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0 (Rejected)** | **Predicted 1 (Approved)** |
| Actual 0 | 17 | 26 |
| Actual 1 | 4 | 76 |

## Support Vector Machine (SVM)

**Support Vector Machine Model Evaluation Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0 (Loan Rejected)** | **1 (Loan Approved)** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| Precision | 0.95 | 0.76 | 0.79 | 0.85 | 0.83 |
| Recall | 0.42 | 0.99 |  | 0.70 | 0.79 |
| F1-Score | 0.58 | 0.86 |  | 0.72 | 0.76 |
| Support | 43 | 80 | 123 |  |  |

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | **Predicted 0 (Rejected)** | **Predicted 1 (Approved)** |
| Actual 0 | 18 | 25 |
| Actual 1 | 1 | 79 |

## Accuracy Scores

**Model Accuracy Scores**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy Score** | **Comment** |
| Logistic Regression | 0.79 | High accuracy, but struggles with predicting non-repayments. |
| Decision Tree | 0.69 | Moderate accuracy, with issues in classifying the positive class. |
| Random Forest | 0.75 | Moderate accuracy, performs better in predicting loan approvals. |
| K-Nearest Neighbors (KNN) | 0.76 | Good accuracy, but less effective in predicting non-repayments. |
| Support Vector Machine (SVM) | 0.79 | High accuracy, performs well in predicting loan repayments. |

The accuracy scores of the models vary, with Logistic Regression and Support Vector Machine (SVM) achieving the highest accuracy of 0.79, indicating good performance in predicting loan repayment outcomes. However, both models struggle with predicting non-repayments effectively, as evidenced by the lower recall for class 0 (not fully paid). The Decision Tree classifier performed the weakest, with an accuracy of 0.69, and it had difficulty in classifying the positive class (borrowers who did not fully repay their loans). Random Forest (0.75) and K-Nearest Neighbors (0.76) also showed moderate accuracy, with Random Forest performing better in predicting loan approvals, while KNN struggled more with non-repayments. Overall, Logistic Regression and SVM appear to be the most reliable models for this dataset.

## Cross-Validation Results for Model Accuracy

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Accuracy** | **Standard Deviation** |
| Logistic Regression | 0.81 | 0.03 |
| Decision Tree Classifier | 0.71 | 0.04 |
| Random Forest Classifier | 0.79 | 0.03 |
| K-Nearest Neighbors | 0.61 | 0.02 |
| Support Vector Machine | 0.69 | 0.00 |

## Model Selection

**Logistic Regression**: Has the highest mean accuracy of 0.81 with a small standard deviation of 0.03, indicating stability across folds. However, it struggles with predicting non-repayment cases, as seen in the confusion matrix.

**Support Vector Machine (SVM)**: Shows a mean accuracy of 0.69 with no variation (standard deviation = 0.00). While it performs consistently, it fails to predict non-repayments, affecting its overall effectiveness.

**Decision Tree Classifier**: Exhibits a mean accuracy of 0.71 with a standard deviation of 0.04. It performs moderately, but accuracy is slightly lower compared to other models.

**K-Nearest Neighbors (KNN)**: Has the lowest mean accuracy of 0.61 and a standard deviation of 0.02. This model performs the worst in terms of overall accuracy and stability.

**Random Forest Classifier**: Shows a mean accuracy of 0.79 with a standard deviation of 0.03. It balances high accuracy and reasonable stability, making it a strong contender among the models tested.

# Conclusion

## Findings

* The Logistic Regression model showed an accuracy of 79%, making it a reliable predictor of loan repayment outcomes. Key predictors for loan approval included factors such as applicant income, loan amount, credit score, and employment status. These factors were strongly correlated with the likelihood of loan repayment.
* The Support Vector Machine (SVM) model also demonstrated an accuracy of 79%, similar to Logistic Regression, but with a higher tendency to misclassify the non-repayment cases, leading to false negatives.
* The Decision Tree model had an accuracy of 69%, showing some promise but struggled with classifying non-repayment loans, as it had a higher number of false positives compared to other models.
* The Random Forest model performed well with an accuracy of **75%** and demonstrated robust performance across various loan approval cases. It outperformed Decision Tree but was slightly less accurate than Logistic Regression.
* The K-Nearest Neighbors (KNN) model showed a lower accuracy of **76%**, indicating moderate performance in predicting loan approval outcomes but still not as reliable as Logistic Regression or SVM.
* **Cross-validation results** reinforced these findings, with Logistic Regression consistently outperforming other models, achieving a mean accuracy of **81%** and demonstrating less fluctuation across different folds compared to other models.
* These results suggest that Logistic Regression is the most suitable model for predicting loan approval, with high accuracy and stable performance across validation tests.

## Future Works

* **Incorporate More Features**: Include additional features such as the applicant's credit history, employment duration, and marital status to improve the model's predictive power.
* **Model Tuning**: Explore hyperparameter tuning for models like Logistic Regression, Decision Tree, and Random Forest to optimize performance and further enhance accuracy.
* **Implement Advanced Algorithms**: Experiment with advanced machine learning techniques like Gradient Boosting or Neural Networks to compare their performance with traditional models.
* **Real-Time Prediction System**: Develop a real-time loan approval system that can predict loan approval status for new applicants instantly.
* **Address Class Imbalance**: Use techniques like oversampling or undersampling to handle class imbalances in the dataset, improving prediction accuracy for less frequent classes.
* **Explore Other Evaluation Metrics**: Besides accuracy, include evaluation metrics such as AUC-ROC, precision-recall curves, and confusion matrices to assess model performance comprehensively.
* **Integrate External Data**: Integrate external datasets like economic indicators or demographic data to potentially improve the model's predictions.
* **Model Interpretability**: Implement model interpretability techniques such as SHAP or LIME to provide insights into the factors influencing loan approval decisions.