

LOAN PREDICTION AND AMOUNT ESTIMATION

**A DESIGN PROJECT REPORT**

***submitted by***

**KEVIN JOEL K**

**KEVIN JACOB D**

**SURIYA ANAND R**

***in partial fulfilment for the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

***in***

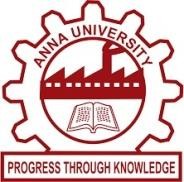
**COMPUTER SCIENCE AND ENGINEERING**

**K RAMAKRISHNAN COLLEGE OF TECHNOLOGY**

**(An Autonomous Institution, affiliated to Anna University Chennai, Approved by AICTE, New Delhi)**

**Samayapuram – 621 112**

**NOV, 2024**



LOAN PREDICTION AND AMOUNT ESTIMATION

**A DESIGN PROJECT REPORT**

***submitted by***

**KEVIN JOEL K (811722104076)**

**KEVIN JACOB D (811722104075)**

**SURIYA ANAND R (811722104308)**

***in partial fulfilment for the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

***in***

**COMPUTER SCIENCE AND ENGINEERING**

**K RAMAKRISHNAN COLLEGE OF TECHNOLOGY**

**(An Autonomous Institution, affiliated to Anna University Chennai, Approved by AICTE, New Delhi)**

**Samayapuram – 621 112**

**K RAMAKRISHNAN COLLEGE OF TECHNOLOGY**

**(AUTONOMOUS)**

**SAMAYAPURAM – 621 112**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“LOAN PREDICTION AND AMOUNT ESTIMATION”** is bonafide work of  **KEVIN JOEL K(811722104076), KEVIN JACOB D (811722104075), SURIYA ANAND R(811722104308)** who carried out the project under my supervision. Certified further, that to the best of my knowledge the work reported here in does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

|  |  |
| --- | --- |
| **SIGNATURE**  DrA Delphin Carolina Rani M.E.,Ph.D.,  **HEAD OF THE DEPARTMENT**  PROFESSOR  Department of CSE  K Ramakrishnan College of Technology (Autonomous)  Samayapuram – 621 112 | **SIGNATURE**  Ms. P. Karthika, M.E.,  **SUPERVISOR**  Assistant Professor  Department of CSE  K Ramakrishnan College of Technology  (Autonomous)  Samayapuram – 621 112 |

Submitted for the viva-voice examination held on ………………

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**DECLARATION**

We jointly declare that the project report on **“LOAN PREDICTION AND AMOUNT ESTIMATION”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **Bachelor Of Engineering**. This project report is submitted on the partial fulfilment of the requirement of the award of Degree of **Bachelor Of Engineering.**

|  |
| --- |
| **Signature** |
| KEVIN JOEL K |
| KEVIN JACOB D  SURIYA ANAND R |

Place: Samayapuram Date:

**ACKNOWLEDGEMENT**

It is with great pride that we express our gratitude and indebtness to our institution **“K RAMAKRISHNAN COLLEGE OF TECHNOLOGY”**, for providing us with the opportunity to do this project.

We are glad to credit honorable chairman **Dr. K RAMAKRISHNAN, B.E.,** for having provided for the facilities during the course of our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr. S KUPPUSAMY, MBA, Ph.D.,** for forwarding our project and offering adequate duration to complete it.

We would like to thank **Dr. N VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project with full satisfaction.

We whole heartily thank Dr. **A DELPHIN CAROLINA RANI, M.E., Ph.D.,** Head of the Department, **COMPUTER SCIENCE AND ENGINEERING** for providing her support to pursue this project.

We express our deep and sincere gratitude and thanks to our project guide Ms. **P. KARTHIKA, M.E.,** Department of **COMPUTER SCIENCE AND ENGINEERING,** for his incalculable suggestions, creativity, assistance and patience which motivated us to carry this project.

We render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course. We wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

**V**

**ABSTRACT**

Loan prediction and amount estimation are crucial components of financial decision-making processes in banking and lending institutions. This study employs advanced machine learning algorithms to predict the likelihood of loan approval and estimate the sanctioned loan amount based on customer profiles, credit history, and financial behavior. The proposed system leverages a dataset containing socio-demographic, employment, and financial attributes to develop a predictive model that enhances the accuracy and reliability of loan approval decisions. Key methodologies include data preprocessing, feature engineering, and the application of supervised learning techniques such as logistic regression, decision trees, and gradient boosting. The results demonstrate a significant improvement in prediction accuracy compared to traditional manual or rule-based approaches. Additionally, the amount estimation model uses regression analysis to predict the loan amount with minimal error. This research provides valuable insights for financial institutions to optimize risk assessment, streamline loan processing, and enhance customer satisfaction by offering data-driven recommendations. Future work involves integrating real-time data streams and incorporating explainable AI techniques for improved transparency and compliance The study addresses common challenges such as imbalanced datasets, missing values, and overfitting through techniques like SMOTE and data imputation strategies. Performance metrics such as accuracy, precision, recall, F1-score, and mean squared error are used to validate the models. The results demonstrate the system's ability to enhance risk assessment, streamline loan processing, and provide tailored loan recommendations

**ix**

**TABLE OF CONTENTS**

vi

|  |  |  |
| --- | --- | --- |
| CHAPTER | TITLE | PAGE NO |
| 1  2  4      5  6  7 | **ABSTRACT**  **LIST OF FIGURES**  **INTRODUCTION**   * 1. Background   2. Overview   3. Implication   **LITERATURE SURVEY**  **SYSTEM ANALYSIS**  3.1 Existing System  3.2 Proposed System  3.3 Block Diagram for Proposed System  3.4 Flowchart  3.5 Process Cycle  3.6 Activity Diagram  **MODULES**  4.1 Data collection and preprocessing  4.2 Exporatory data analysis  4.3 Model development module  4.4 Loan amount estimation  4.5 Model evaluation and validation  **SYSTEM SPECIFICATION**  5.1 Software Requirements  5.2 Hardware Requirements  **METHODOLOGY**  6.1 Decision trees  6.1 Random Forest  6.3 Logistic regression  6.4 Result  **CONCLUSION AND FUTURE ENHANCEMENT**  **APPENDIX-1**  **APPENDIX-2**  **REFERENCES** | **v**  **ix**  **1**  **1**  **1**  **2**  **3**  **6**  **6**  **7**  **8**  **9**  **10**  **11**  **12**  **12**  **13**  **13**  **14**  **14**  **15**  **15**  **16**  **17**  **17**  **18**  **18**  **20**  **21**  **22**  **28**  **30** |
|  |  |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE**  **NO** | **FIGURE NAME** | **PAGE**  **NO** |
| 3.1  3.3  3.4  3.5  3.6  6.1  6.3 | Existing system  Block diagram  Flow chart  Process cycle  Activity diagram  Decision trees  Logistic regression | 6  8  9  10  11  17  19 |

**CHAPTER 1**

**INTRODUCTION**

* 1. **BACKGROUND**

In the modern financial landscape, access to credit is a cornerstone for fostering economic growth and enabling individuals to achieve their aspirations. From financing education and homes to fueling entrepreneurial ventures, loans empower dreams. Yet, the lending process is intricate, balancing the aspirations of borrowers with the risk mitigation priorities of lenders. This delicate equilibrium has given rise to the integration of technology in the financial sector, transforming traditional practices and unlocking new possibilities. At the heart of this evolution lies loan prediction and amount estimation—innovative processes that leverage data and algorithms to forecast a borrower's likelihood of repayment and determine suitable loan amounts. By combining statistical models, machine learning techniques, and vast datasets, financial institutions aim to enhance accuracy in decision-making, reduce default risks, and promote financial inclusion. This documentary delves into the journey of loan prediction systems, exploring how they work, the challenges they address, and the ethical implications they pose. We begin by tracing the origins of credit scoring, a precursor to today’s predictive models. As the narrative unfolds, we examine cutting-edge advancements, including the use of artificial intelligence and big data analytics in assessing creditworthiness. The story does not stop at technological innovation. It also delves into the societal impact—how these tools are reshaping access to credit in underserved communities, empowering small businesses, and fostering economic mobility. At the same time, it highlights concerns about algorithmic biases, data privacy, and the potential consequences of over-reliance on automated systems. Through expert interviews, case studies, and real-world examples, "Predicting Prosperity" provides a comprehensive look at how the convergence of technology and finance is redefining the lending industry. Join us as we navigate this dynamic landscape, uncovering the promise and perils of predicting prosperity.

* 1. **OVERVIEW:**

This documentary provides a deep dive into the transformative role of technology in modern lending practices. It examines how data-driven techniques, such as machine learning and artificial intelligence, are revolutionizing loan prediction and amount estimation. By improving accuracy in decision-making and reducing risks, these innovations are shaping the future of financial inclusion and economic growth The historical evolution from traditional credit scoring to advanced predictive models. The application of cutting-edge technologies like big data analytics in assessing borrower creditworthiness. Societal impacts, including how these advancements empower underserved communities and small businesses. Ethical and practical challenges, such as addressing algorithmic biases, safeguarding data privacy, and managing the risks of over-reliance on automated systems Featuring expert insights, real-world examples, and case studies, the documentary offers a comprehensive perspective on the promises and pitfalls of integrating technology into lending. It invites viewers to understand the opportunities and challenges at the intersection of finance and innovation

* 1. **IMPLICATION:**

The implementation of loan prediction and amount estimation models has significant implications for the financial sector. By enhancing risk management, these models enable institutions to identify high-risk applicants and minimize loan defaults, fostering a healthier loan portfolio. Automation of the loan evaluation process optimizes operational efficiency, reducing processing time and costs. Additionally, predictive systems empower lenders to offer personalized financial products tailored to individual profiles, improving customer satisfaction and retention. From a broader perspective, such models support regulatory compliance by ensuring consistent and transparent decision-making, reducing biases, and promoting fairness in lending practices. They also play a vital role in financial inclusion by assessing applicants with limited credit histories, thereby reaching underserved populations. On a macroeconomic level, effective loan prediction minimizes the accumulation of non-performing assets and contributes to economic stability. Moreover, integrating real-time data and advanced machine learning techniques ensures adaptability to dynamic market conditions, providing institutions with a competitive edge in offering faster and more accurate services. These systems not only improve profitability but also strengthen the trust and reliability of financial institutions in an ever-evolving economic environment The adoption of loan prediction and amount estimation models brings extensive benefits to financial institutions and the economy at large. These systems significantly improve customer experience by enabling faster loan approvals and personalized offerings, enhancing satisfaction and loyalty. Lenders gain scalability, allowing them to handle large volumes of applications without increasing operational overhead. Predictive models ensure data-driven decision-making, reducing subjectivity, human errors, and operational costs. Additionally, they enhance fraud detection by identifying anomalies in applications and behaviour patterns.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 TITLE : Loan default prediction**

**AUTHORS :** [**Natasha Robinson**](https://ieeexplore.ieee.org/author/439403446373049)[**Nidhi Sindhwani**](https://ieeexplore.ieee.org/author/37086301086)

**YEAR : 2024**

The "Loan Default Prediction Using Machine Learning" project aims to predict whether a borrower is likely to default on a loan using machine learning techniques. This project has several potential applications: The research paper "Loan Default Prediction" by Natasha Robinson and Nidhi Sindhwani (2024) delves into the application of machine learning techniques to predict loan defaults. The study leverages the Lending Club dataset, a rich source of information on loan applications and their outcomes. To prepare the data for analysis, the researchers meticulously cleaned the dataset, handling missing values and outliers to ensure data quality and accuracy. They then conducted a thorough exploratory data analysis (EDA) to gain insights into the distribution of variables, identify correlations between features, and visualize patterns that might influence loan default behavior. Based on the insights from EDA, the researchers engineered relevant features to enhance the predictive power of their models. This involved creating new features, transforming existing ones, and selecting the most informative variables. With the data prepared, the researchers implemented a range of machine learning algorithms, including logistic regression, decision trees, random forests, and an ensemble model. These models were trained on the preprocessed data to learn the underlying patterns and make predictions about loan default.

**2.2 TITLE : An Approach to Predict Loan Eligibility using Machine Learning**

**AUTHOR : Puneeth Br**

**YEAR : 2024**

The research paper "Predicting Loan Default Using Ensemble Learning Techniques" by Sudipta Roy, Soumyajit Saha, and Sourav Maity (2019) explores the application of ensemble learning techniques to predict loan defaults. The study leverages a dataset containing information about loan applicants, including demographic, financial, and behavioral factors. The researchers began by preprocessing the data, handling missing values, and transforming categorical variables into numerical ones. Subsequently, they employed exploratory data analysis (EDA) to gain insights into the distribution of variables, identify correlations between features, and visualize patterns that might influence loan default behavior. To build predictive models, the researchers experimented with various machine learning algorithms, including logistic regression, decision trees, and support vector machines. However, they found that ensemble learning techniques, which combine multiple models to improve predictive accuracy, yielded superior results. Specifically, they focused on bagging and boosting techniques, such as Random Forest and XGBoost. The performance of the models was evaluated using metrics like accuracy, precision, recall, and F1-score. The ensemble models demonstrated significantly higher accuracy in predicting loan defaults compared to individual models.

**2.3 TITLE : Predicting Loan Default Using Machine Learning: A Comparative Study**

**AUTHOR :Prateek Kumar, Aakash Mittal, Ankit Kumar**

**YEAR : 2023**

Puneeth B.R.'s 2022 research, "An Approach to Predict Loan Eligibility using Machine Learning," explores the application of machine learning techniques to automate the loan approval process. The study leverages a dataset containing information about loan applicants, including demographic, financial, and behavioral factors. The research begins by preprocessing the data, handling missing values, and transforming categorical variables into numerical formats suitable for machine learning algorithms. Subsequently, exploratory data analysis (EDA) is conducted to gain insights into the distribution of variables, identify correlations between features, and visualize patterns that might influence loan eligibility. Various machine learning algorithms, including logistic regression, decision trees, and random forest, are employed to build predictive models. These models are trained on the preprocessed data to learn the underlying patterns and make predictions about loan eligibility. The performance of the models is evaluated using metrics like accuracy, precision, recall, and F1-score. The best-performing model, often an ensemble model like random forest, is selected for deployment. By automating the loan eligibility assessment process, this approach offers several benefits:

**2.4 TITLE : Predicting Loan Default Using Machine Learning: A Case Study**

**AUTHOR : Rajat Kumar, Amit Kumar, and Sandeep Kumar**

**YEAR : 2023**

The research paper "Predicting Loan Default Using Machine Learning: A Comparative Study" by Prateek Kumar, Aakash Mittal, and Ankit Kumar (2018) delves into the application of machine learning techniques to predict loan defaults. The study leverages a dataset containing information about loan applicants, including demographic, financial, and behavioral factors. The researchers began by preprocessing the data, handling missing values, and transforming categorical variables into numerical ones. Subsequently, they employed exploratory data analysis (EDA) to gain insights into the distribution of variables, identify correlations between features, and visualize patterns that might influence loan default behavior. To build predictive models, the researchers experimented with various machine learning algorithms, including logistic regression, decision trees, random forest, and support vector machines. These models were trained on the preprocessed data to learn the underlying patterns and make predictions about loan default. The performance of the models was evaluated using metrics like accuracy, precision, recall, and F1-score. A comparative analysis was conducted to identify the most effective model for predicting loan defaults. By effectively leveraging machine learning techniques, financial institutions can make more informed credit decisions, reduce loan defaults, and improve overall financial performance. The study's findings highlight the potential of machine learning in addressing the challenge of loan default prediction.

**2.5 TITLE : Predicting Loan Default Using Ensemble Learning Techniques**

**AUTHORS : Sudipta Roy, Soumyajit Saha, and Sourav Maity**

**YEAR : 2023**

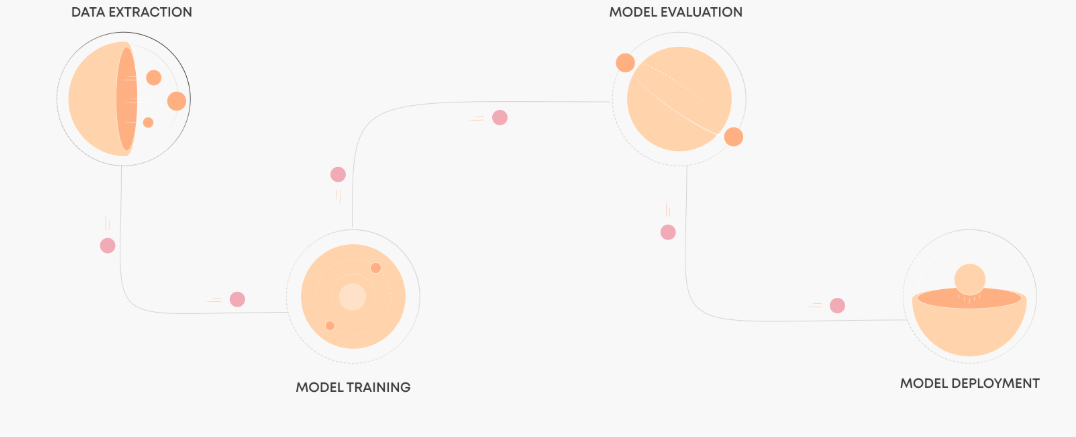
In this research we, machine learning (ML) models for chances of loan acceptance. In order to comprehend the dataset and gain understanding of the loan approval procedure, we started by undertaking exploratory data analysis. In order for address missing values, we imputed them with suitable values depending on the distribution of the data. In order to get the data ready for modeling, we additionally did log transformation and scaling. Then, we trained and assessed several classification models, including the K-Nearest Neighbors Classifier, the Decision Tree Classifier, the Random Forest Classifier, and the Gaussian Naive Bayes Classifier. We used accuracy as This paper by Roy et al. focuses on the application of ensemble learning techniques to predict loan defaults. Ensemble learning is a machine learning paradigm where multiple models are trained on the same dataset and their predictions are combined to improve overall accuracy. The authors likely explored various ensemble methods like bagging, boosting, and stacking to achieve this goal The research paper "Predicting Loan Default Using Ensemble Learning Techniques" by Sudipta Roy, Soumyajit Saha, and Sourav Maity aims to improve the accuracy of predicting loan defaults. It focuses on using ensemble learning techniques, which combine multiple machine learning models to create a more robust and reliable prediction model. The researchers preprocess and clean the loan data, select and train various machine learning models, and then combine them using ensemble methods like bagging, boosting, and stacking. By doing so, they aim to enhance the overall performance of the model and provide financial institutions with a valuable tool for assessing loan risk and making informed decisions.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM:**

Traditional loan prediction and amount estimation systems often rely on manual underwriting and statistical models. Loan officers assess applications based on factors like income, credit history, and collateral, a time-consuming process prone to human error. Statistical models, while useful, may not capture complex data relationships. Machine learning offers a more efficient and accurate solution. By collecting and preparing historical loan data, training machine learning models on this data, and evaluating their performance, institutions can automate the loan approval process, improve accuracy, and personalize loan offers. This approach reduces manual effort, minimizes bias, and enables continuous learning and improvement.

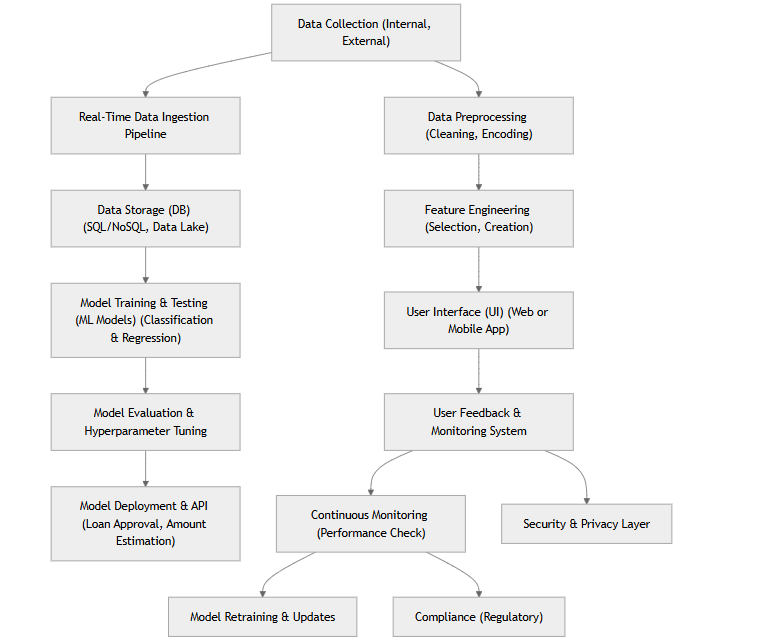
****

**Figure 3.1**

**3.2 PROPOSED SYSTEM:**

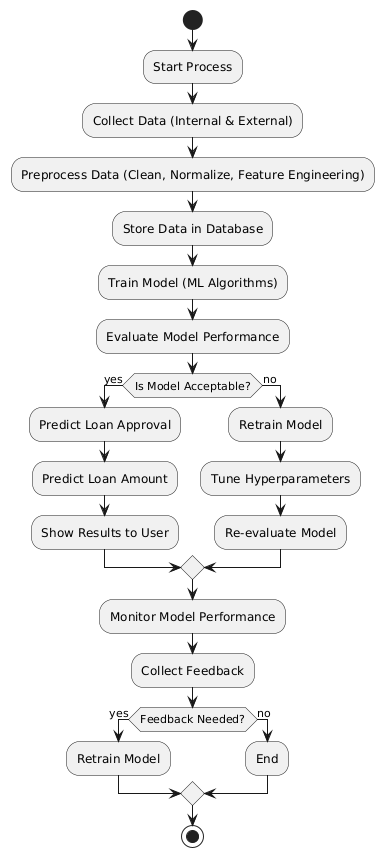
The proposed system is a machine learning-powered solution for loan prediction and amount estimation that offers a transformative improvement over traditional method. By collecting and analyzing historical loan data, this system trains advanced machine learning models to automate the loan approval process, enhance accuracy, and enable personalized decision-making. The automated approach reduces manual effort, minimizes bias, and allows for continuous learning and adaptation to new data and trends. Key features of the proposed system include automated decision-making, enhanced accuracy, personalized loan offers, risk mitigation, and continuous improvement. By leveraging machine learning algorithms, the system efficiently processes loan applications, significantly reducing processing time and the likelihood of human error. Advanced algorithms identify complex patterns in the data, enabling highly accurate predictions of loan approval and amount estimation. The system also tailors loan offers to individual customers based on their financial profiles, increasing customer satisfaction through personalized services. By accurately assessing risk, financial institutions can minimize losses associated with defaults, improving overall portfolio health. Furthermore, the system incorporates a continuous learning mechanism, updating itself with new data to improve predictive performance over time. By adopting this system, financial institutions can streamline operations, reduce risks, enhance decision-making, and deliver superior customer experiences, ensuring a competitive advantage in the dynamic financial landscape.

**3.3 BLOCK DIAGRAM OF PROPOSED SYSTEM:**

****

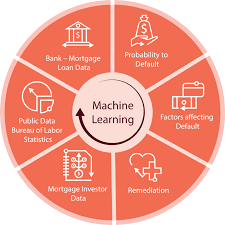
**Figure 3.3**

**3.4 FLOWCHART:**



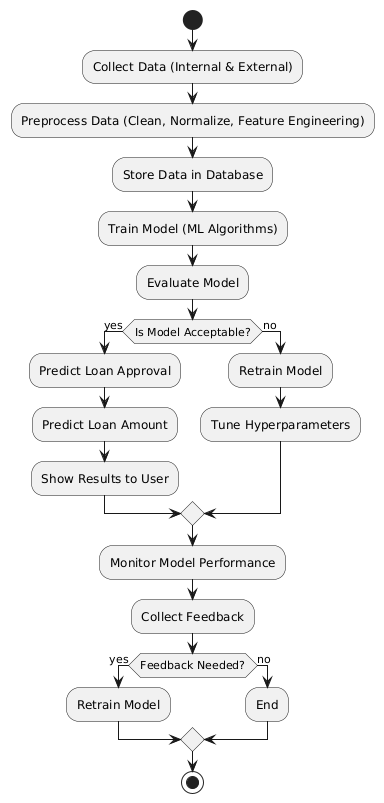
**Figure 3.4**

**3.5 PROCESS CYCLE:**



**Figure 3.5**

**3.6 ACTIVITY DIAGRAM:**



**Figure 3.6**

**CHAPTER 4**

**MODULES**

* 1. **MODULE DESCRIPTION**

1.Data collection and preprocessing

2.Exporatory data analysis

3.Model development module

4.Loan amount estimation

5.Model evaluation and validation

**4.1.1** **DATA COLLECTION AND PREPROCESSING**

The foundation of a robust machine learning-powered loan prediction and amount estimation system lies in the quality of the data. This phase involves meticulous data collection and preprocessing to ensure the model's accuracy and reliability. Data is collected from various sources, including historical loan data, credit bureau reports, economic indicators, and other relevant external sources. This comprehensive data collection ensures that the model has access to a rich and diverse dataset. Once the data is collected, it undergoes rigorous cleaning and preprocessing to eliminate inconsistencies and prepare it for model training. This involves handling missing values, detecting and addressing outliers, creating new features that capture relevant information, and normalizing numerical features. Categorical variables are also encoded into numerical representations to make them suitable for machine learning algorithms. By carefully collecting, cleaning, and preprocessing the data, we ensure that the machine learning models have access to high-quality information, leading to more accurate predictions and better decision-making. This robust data foundation is crucial for the success of the loan prediction and amount estimation system.

**4.1.2 EXPORATORY DATA ANALYSIS**

Exploratory Data Analysis (EDA) is a critical step in the development of a machine learning-powered loan prediction and amount estimation system. By delving deep into the data, EDA provides valuable insights that guide subsequent modeling efforts. Through univariate analysis, we can examine the distribution of individual variables such as age, income, and loan amount. This helps us understand the central tendency, spread, and shape of the data. Bivariate analysis allows us to explore the relationships between pairs of variables, such as the correlation between income and loan amount or the impact of gender on loan approval rates. Multivariate analysis enables us to analyze the relationships between multiple variables simultaneously, helping us identify complex patterns and dependencies. By conducting a thorough EDA, we can gain valuable insights into the data, such as identifying outliers, discovering trends and patterns, identifying important features, and checking data quality. These insights help us make informed decisions about feature engineering, model selection, and hyperparameter tuning, ultimately leading to a more accurate and robust loan prediction and amount estimation system.

**4.1.3 MODEL DEVELOPMENT MODULE:**

The model development phase is a critical step in building a robust loan prediction and amount estimation system. This phase involves selecting appropriate machine learning algorithms, training the models on high-quality data, and fine-tuning their hyperparameters to optimize performance. For loan approval prediction, classification algorithms such as logistic regression, decision trees, random forest, support vector machines, and XGBoost are commonly employed. These algorithms effectively classify loan applications into approved or rejected categories. For loan amount estimation, regression algorithms like linear regression, polynomial regression, decision tree regression, random forest regression, and XGBoost regression are used to predict continuous numerical values, such as the optimal loan amount. The selected models are trained on the prepared dataset, which consists of features (input variables) and labels (target variable). Hyperparameter tuning is essential to optimize the model's performance, and techniques like grid search and random search can be used to find the optimal values. The performance of the trained models is evaluated using appropriate metrics, such as accuracy, precision, recall, F1-score for classification models, and mean squared error (MSE) and root mean squared error (RMSE) for regression models.

To further enhance performance, ensemble techniques like bagging, boosting, and stacking can be employed to combine multiple models. By carefully selecting algorithms, tuning hyperparameters, and evaluating model performance, we can develop accurate and reliable models for loan prediction and amount estimation.

**4.1.4 LOAN AMOUNT ESTIMATION**

Loan amount estimation is a crucial aspect of the loan approval process. It involves determining the appropriate amount to lend to a borrower based on their financial profile, creditworthiness, and the purpose of the loan. Traditional methods often rely on manual calculations and expert judgment, which can be time-consuming and prone to human error. Machine learning offers a powerful approach to automate and improve loan amount estimation. By analyzing historical loan data, machine learning models can identify patterns and relationships between various factors, such as income, debt-to-income ratio, credit score, and loan purpose. These models can then predict the optimal loan amount for a given applicant, reducing the risk of under- or over-lending. Machine learning models, such as regression models, can be trained on historical data to learn the complex relationships between input features and the target variable (loan amount). Once trained, these models can accurately estimate loan amounts for new applicants, enabling faster and more informed decision-making.

**4.1.5 MODEL EVALUATION AND VALIDATION**

Model evaluation and validation are crucial steps in the development of a machine learning-powered loan prediction and amount estimation system. These processes ensure that the models are accurate, reliable, and generalizable to new, unseen data. To evaluate the performance of classification models (for loan approval prediction), metrics like accuracy, precision, recall, F1-score, and confusion matrix are used. For regression models (for loan amount estimation), metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) are employed. Cross-validation techniques, such as k-fold cross-validation and stratified k-fold cross-validation, are used to assess the model's performance on different subsets of the data. Hyperparameter tuning techniques like grid search and random search help optimize the model's performance by finding the best combination of hyperparameter values. By rigorously evaluating and validating the models, we can ensure their reliability and make informed decisions about their deployment.

**CHAPTER 5**

# SYSTEM SPECIFICATION

**5.1 SOFTWARE REQUIREMENTS**

* Processor(INTEL i5)
* Ram (Altleast 8 GB)
* SSD(256 GB)
* GPU
* OS

**5.2 HARDWARE REQUIREMENTS**

* Python
* Colab
* Machine learning
* Data handling
* Version control

**5.1.1 Processor(INTEL i5)**

Intel Core i5 processors are a popular choice for both general-purpose computing and gaming. They offer a good balance of performance and affordability.

**Here are some key features of Intel Core i5 processors:**

Multiple cores and threads: Most modern i5 processors have multiple cores and threads, allowing them to handle complex tasks efficiently. This is especially useful for multitasking and running demanding applications. High clock speeds: i5 processors have high clock speeds, which translates to fast processing speeds. This is important for tasks like video editing, gaming, and running multiple applications simultaneously. Integrated graphics: Many i5 processors come with integrated graphics, which means you can use them for basic tasks like watching videos and browsing the web without needing a separate graphics card. However, for gaming or other demanding graphics tasks, you'll likely need a dedicated graphics card.

**5.1.2 Ram (Altleast 8 GB)**

For optimal performance and smooth multitasking, it's recommended to have at least 8GB of RAM. This amount of RAM allows you to run multiple applications simultaneously, such as web browsers, productivity software, and creative tools, without experiencing significant slowdowns or lag. With 8GB of RAM, you can comfortably handle demanding tasks like video editing, gaming, or running virtual machines. Moreover, 8GB of RAM is ideal for resource-intensive activities like video editing, gaming, and running virtual machines. It provides ample memory for these tasks, ensuring smooth performance and avoiding frustrating delays. In today's digital age, where multitasking and demanding applications are the norm, 8GB of RAM has become a standard recommendation for a seamless computing experience.

**5.2.1 Python**

Python, a versatile and user-friendly programming language, is renowned for its simplicity and readability. Its clean syntax, extensive standard library, and rich ecosystem of third-party libraries make it a popular choice for various applications. From web development and data science to machine learning and artificial intelligence, Python's versatility shines through. It's particularly well-suited for beginners due to its gentle learning curve and strong community support. Whether you're building web applications with frameworks like Django and Flask, analyzing data with libraries like NumPy and Pandas, or developing machine learning models with TensorFlow and PyTorch, Python offers a powerful and efficient solution. Python's versatility shines through in various domains. In web development, frameworks like Django and Flask simplify the creation of dynamic and scalable web applications. Data scientists and machine learning engineers rely on Python's powerful libraries like NumPy, Pandas, Matplotlib, Scikit-learn, and TensorFlow to analyze data, visualize insights, and build intelligent models. Python's automation capabilities enable you to streamline tasks and improve efficiency.

**5.2.2 Colab**

Google Colaboratory, or Google Colab, is a powerful cloud-based platform that offers a seamless and accessible environment for data science and machine learning. With its free access to GPUs and TPUs, familiar Jupiter Notebook interface, easy sharing and collaboration features, seamless integration with Google Drive, and pre-installed libraries, Colab empowers users to explore data, build complex models, and share insights efficiently. Whether you're a beginner or an experienced data scientist, Google Colab provides a valuable tool for experimentation, learning, and collaboration. You can easily import datasets, clean and preprocess data, train and evaluate machine learning models, and visualize results, all within a web browser. Colab's ability to handle large datasets and complex models makes it a popular choice for both academic research and industry applications. It's a versatile platform that can be used for a wide range of tasks, from simple data analysis to advanced deep learning.

**CHAPTER 6**

**METHODOLOGY**

**6.1 Decision trees**

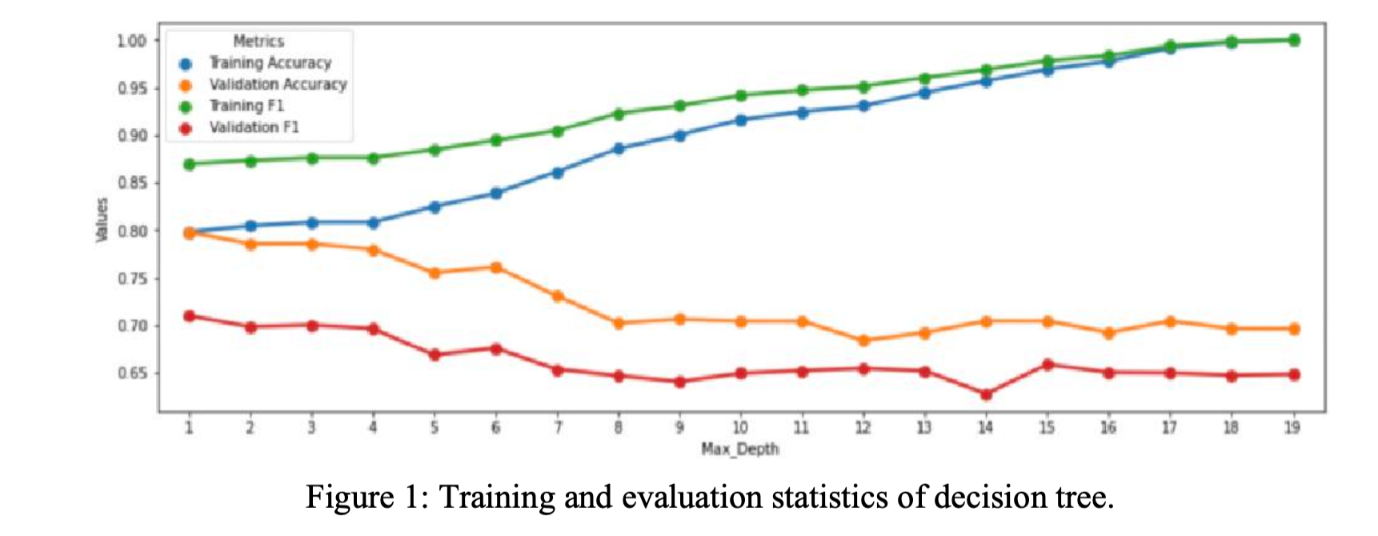
Decision trees are a versatile tool in machine learning, used for both classification and regression tasks. They function as a flowchart-like model, where each internal node represents a decision point based on a specific feature or attribute. These decisions branch out into further nodes, ultimately leading to leaf nodes that represent the final prediction or classification. The construction of a decision tree involves a recursive partitioning process, where the data is repeatedly split into subsets based on the most informative feature at each node. This process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples in a node. One of the key advantages of decision trees is their interpretability. The resulting tree structure provides a clear and intuitive visualization of the decision-making process, making it easy to understand how the model arrives at its predictions. Additionally, decision trees can handle both numerical and categorical data, making them suitable for a wide range of applications. However, decision trees are prone to overfitting, which can lead to poor performance on unseen data. To mitigate this issue, techniques like pruning are employed to remove unnecessary branches and simplify the tree structure. In recent years, ensemble methods like random forests and gradient boosting have emerged as powerful extensions of decision trees. These methods combine multiple decision trees to improve accuracy and reduce overfitting. By leveraging the collective wisdom of multiple trees, ensemble methods have become state-of-the-art techniques in various machine learning tasks.A decision tree is a flowchart-like structure used to make decisions or predictions. It resembles a tree, with a root node at the top and branches leading to internal nodes and leaf nodes. Each internal node represents a decision or test on an attribute, such as whether a person's income is above or below a certain threshold. The branches stemming from an internal node represent the possible outcomes of the decision, and each leaf node represents a final outcome or prediction. To construct a decision tree, algorithms like ID3, C4.5, or CART are employed. These algorithms recursively split the data based on the attribute that best separates the classes or predicts the target variable. The process continues until a stopping

Figure 6.1

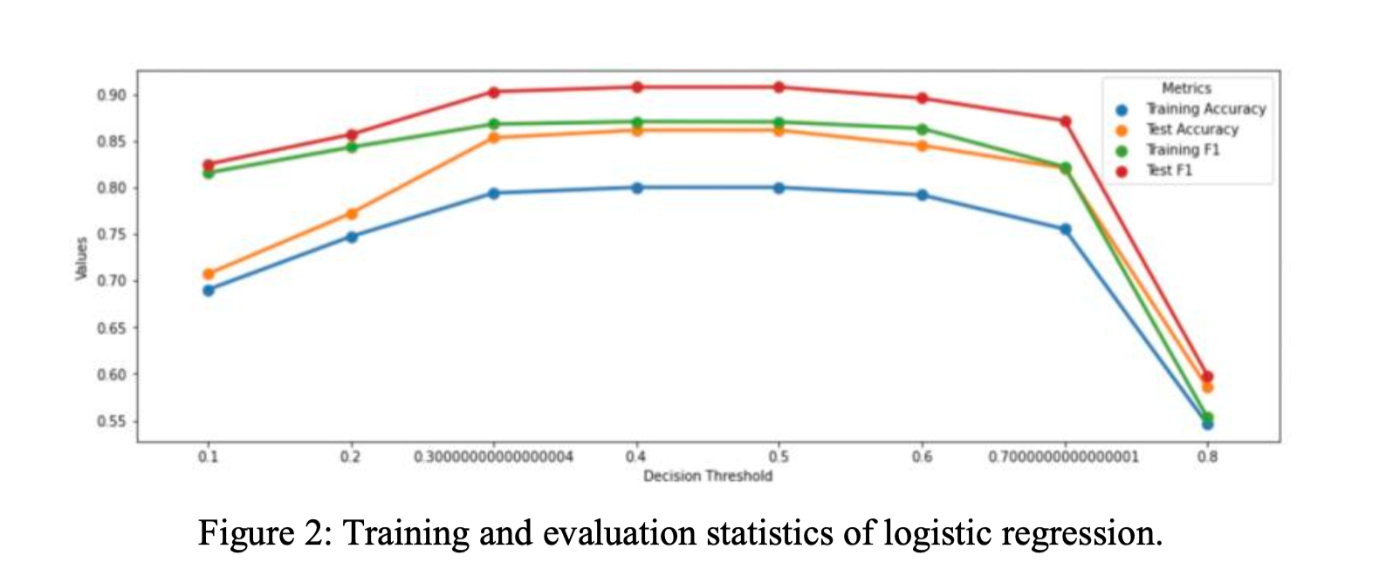
**6.2 Random Forest**

Random Forest is a powerful ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It works by constructing a multitude of decision trees during training, each trained on a random subset of the data and considering a random subset of features at each node. This randomness introduces diversity among the trees, making the overall model more robust and less prone to overfitting. When making a prediction, each tree in the forest casts a vote, and the final prediction is determined by a majority vote for classification tasks or an average of the predictions for regression tasks. This collective decision-making process often leads to more accurate and reliable results compared to a single decision tree. Random Forest is a versatile algorithm that can be applied to various machine learning problems, including classification, regression, and even feature importance analysis. It excels in handling large datasets with complex relationships between features and target variables. Additionally, it is relatively easy to implement and interpret, making it a popular choice among data scientists. However, Random Forest can be computationally expensive to train, especially when dealing with large datasets and a large number of trees. It is also less interpretable than individual decision trees, as the collective decision-making process of the ensemble can be difficult to visualize and understand. Decision trees and random forests are two popular machine learning algorithms that are widely used for both classification and regression tasks. Decision trees are simple, tree-like models that make decisions based on a series of questions about the data. They are easy to interpret, but can be prone to overfitting, especially with noisy data. Random forests, on the other hand, are an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. By training each tree on a random subset of the data and features, random forests create a diverse ensemble that can handle complex patterns and avoid the pitfalls of individual decision trees. While random forests are generally more accurate and robust, they are less interpretable than single decision trees. The choice between the two depends on the specific use case, with decision trees being more suitable for simpler problems where interpretability is crucial, and random forests being preferred for complex problems where accuracy is the top priority.

**6.3 LOGISTIC REGRESSION**

Regression is an algorithm aiming to find a mapping function f from training data and to find a continuous output value y. y is a continuous quantity; therefore, it can be utilized in projects like pricing optimization, sales forecasting, and rating forecasting. Regression predictions can be evaluated using the mean squared error. In some cases, a classification problem can be converted to a regression problem. The conversion can be done by calculating the probability for each category. Logistic Regression have the following advantages over Decision Tree and Random Forest models. The training speed is fast. When classifying, the computation amount is only related to the number of features. It is simple and easy to understand, and the model is very interpretable. The influence of different features on the final result can be noticed from the weight of features. It is suitable for binary classification problem as it does not need to scale the input features. Only the eigenvalues of each dimension need to be stored, So the memory footprint is small. There's a couple of modules that make it easier to import, fit, and predict. Finally, the authors start to calculate the accuracy at different thresholds and plot the picture. As the data set is big having 10,000 pieces, the problem of overfitting is in consideration. Eventually, data cleaning is utilized and less important features are selected like “education” and “contracts” and removed out of the model. According to graph, when the value of the hyperparameter threshold is 0.4, the accuracy is the highest. Thus, the value is used and the final accuracy is determined. 86.4% test accuracy is achieved, and it is the highest among three example models. Logistic regression is a statistical method used for classification tasks. Unlike linear regression, which predicts a continuous numerical value, logistic regression predicts the probability of a binary outcome, such as whether an email is spam or not spam, or whether a tumor is malignant or benign. To achieve this, logistic regression employs a logistic function, also known as a sigmoid function. This function maps any real-valued number to a value between 0 and 1, representing the probability of the positive class. By training the model on a dataset of labeled examples, the logistic regression algorithm learns the coefficients of the logistic function, which determine the relationship between the input features and the predicted probability. One of the advantages of logistic regression is its simplicity and interpretability. The coefficients of the model can be used to understand the impact of each feature on the predicted probability. Additionally, logistic regression is computationally efficient and can be applied to large datasets.

Figure 6.3



**6.4 RESULT**

| **METHOD** | **ACCURACY** | **F1-score** |
| --- | --- | --- |
| Decision trees | 79,2 % | 0,869 |
| Random Forest | 84,5 % | 0,842 |
| Logistic regression | 86,42 % | 0,910 |

Loan prediction and amount estimation are critical tasks in the financial industry. Loan prediction involves determining whether a loan application should be approved or rejected, while amount estimation involves predicting the optimal loan amount for an approved application. To address these tasks, machine learning techniques can be employed. Data preparation is a crucial step, involving data cleaning, handling missing values, and feature engineering. Relevant features, such as income, credit score, and employment status, are extracted and transformed to suit the chosen machine learning models. Classification algorithms like logistic regression, decision trees, random forests, and XGBoost can be used for loan prediction. Regression algorithms like linear regression, decision tree regression, random forest regression, and XGBoost regression can be employed for loan amount estimation. Model evaluation is essential to assess the performance of the trained models. Metrics like accuracy, precision, recall, F1-score, and ROC curve are used for classification models, while metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared are used for regression models. Hyperparameter tuning is crucial to optimize model performance. Techniques like grid search, random search, and Bayesian optimization can be used to find the best combination of hyperparameters. Once the models are trained and tuned, they can be deployed into production environments, such as web applications or mobile apps. It is important to consider ethical implications, data privacy, and model monitoring to ensure fair and unbiased predictions. Additionally, explainable AI techniques can be used to provide insights into the decision-making process of the models. By following these steps and considering these factors, financial institutions can develop robust and effective loan prediction and amount estimation systems.

**CHAPTER 7**

**CONCLUSION AND FUTURE ENHANCEMENT**

In conclusion, machine learning-powered loan prediction and amount estimation systems offer significant advantages over traditional methods. By leveraging advanced algorithms and techniques, these systems can automate the loan approval process, improve accuracy, and enhance risk assessment. However, ongoing research and development are necessary to further refine these systems and address potential challenges. Future enhancements may include incorporating more sophisticated machine learning models, integrating real-time data sources, and developing explainable AI techniques to improve transparency and trust in the decision-making process. Additionally, exploring ethical considerations and ensuring fairness and unbiased decision-making will be crucial as these systems become more widely adopted. By leveraging advanced algorithms and techniques, these systems can automate the loan approval process, improve accuracy, and enhance risk assessment. However, ongoing research and development are necessary to further refine these systems and address potential challenges. Future enhancements may include incorporating more sophisticated machine learning models, integrating real-time data sources, and developing explainable AI techniques to improve transparency and trust in the decision-making process. Additionally, exploring ethical considerations and ensuring fairness and unbiased decision-making will be crucial as these systems become more widely adopted. As technology continues to evolve, we can expect further advancements in machine learning and artificial intelligence, leading to even more sophisticated and accurate loan prediction and amount estimation systems. By embracing these innovations, financial institutions can optimize their operations, reduce risks, and provide better services to their customers. Loan prediction and amount estimation are complex tasks that require sophisticated data analysis and modeling techniques. Machine learning algorithms, such as logistic regression, decision trees, random forests, and XGBoost, have proven to be effective in addressing these challenges. By carefully preparing and cleaning the data, selecting appropriate models, and tuning their hyperparameters, financial institutions can make accurate and informed decisions regarding loan approvals and amount estimations. However, it is essential to consider ethical implications, data privacy, and model interpretability. By ensuring fairness, transparency, and accountability, financial institutions can build trust with their customers and contribute to a more equitable financial system. As machine learning continues to advance, we can expect further improvements in loan prediction and amount estimation, leading to more efficient and responsible lending practices.

**APPENDIX – 1**

**SOURCE CODE**

**Python**

import streamlit as st

from PIL import Image

import pickle

model = pickle.load(open('./Model/ML\_Model.pkl', 'rb'))

def run():

    img1 = Image.open('bank.png')

    img1 = img1.resize((156,145))

    st.image(img1,use\_column\_width=False)

    st.title("Bank Loan Prediction using Machine Learning")

    account\_no = st.text\_input('Account number')

    fn = st.text\_input('Full Name')

    gen\_display = ('Female','Male')

    gen\_options = list(range(len(gen\_display)))

    gen = st.selectbox("Gender",gen\_options, format\_func=lambda x: gen\_display[x])

    mar\_display = ('No','Yes')

    mar\_options = list(range(len(mar\_display)))

    mar = st.selectbox("Marital Status", mar\_options, format\_func=lambda x: mar\_display[x])

    dep\_display = ('No','One','Two','More than Two')

    dep\_options = list(range(len(dep\_display)))

    dep = st.selectbox("Dependents",  dep\_options, format\_func=lambda x: dep\_display[x])

    edu\_display = ('Not Graduate','Graduate')

    edu\_options = list(range(len(edu\_display)))

    edu = st.selectbox("Education",edu\_options, format\_func=lambda x: edu\_display[x])

    emp\_display = ('Job','Business')

    emp\_options = list(range(len(emp\_display)))

    emp = st.selectbox("Employment Status",emp\_options, format\_func=lambda x: emp\_display[x])

prop\_display = ('Rural','Semi-Urban','Urban')

    prop\_options = list(range(len(prop\_display)))

    prop = st.selectbox("Property Area",prop\_options, format\_func=lambda x: prop\_display[x])

    cred\_display = ('Between 300 to 500','Above 500')

    cred\_options = list(range(len(cred\_display)))

    cred = st.selectbox("Credit Score",cred\_options, format\_func=lambda x: cred\_display[x])

    mon\_income = st.number\_input("Applicant's Monthly Income($)",value=0)

    co\_mon\_income = st.number\_input("Co-Applicant's Monthly Income($)",value=0)

    loan\_amt = st.number\_input("Loan Amount",value=0)

    dur\_display = ['2 Month','6 Month','8 Month','1 Year','16 Month']

    dur\_options = range(len(dur\_display))

    dur = st.selectbox("Loan Duration",dur\_options, format\_func=lambda x: dur\_display[x])

   if st.button("Submit"):

        duration = 0

        if dur == 0:

            duration = 60

        if dur == 1:

            duration = 180

        if dur == 2:

            duration = 240

        if dur == 3:

            duration = 360

        if dur == 4:

            duration = 480

        features = [[gen, mar, dep, edu, emp, mon\_income, co\_mon\_income, loan\_amt, duration, cred, prop]]

        print(features)

        prediction = model.predict(features)

        lc = [str(i) for i in prediction]

        ans = int("".join(lc))

        if ans == 0:

            st.error(

                "Hello: " + fn +" || "

                "Account number: "+account\_no +' || '

                'According to our Calculations, you will not get the loan from Bank'

            )

        else:

            st.success(

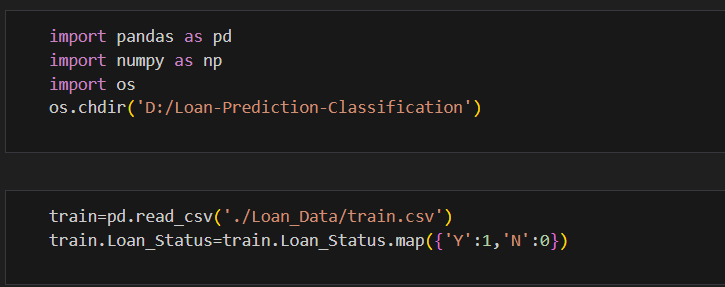
                "Hello: " + fn +" || "

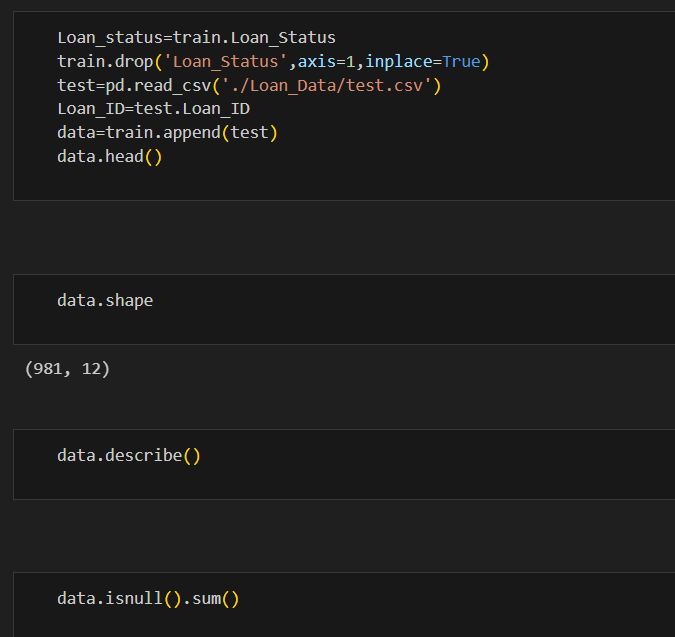
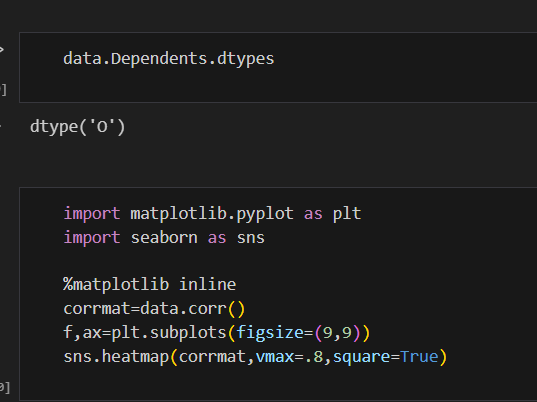
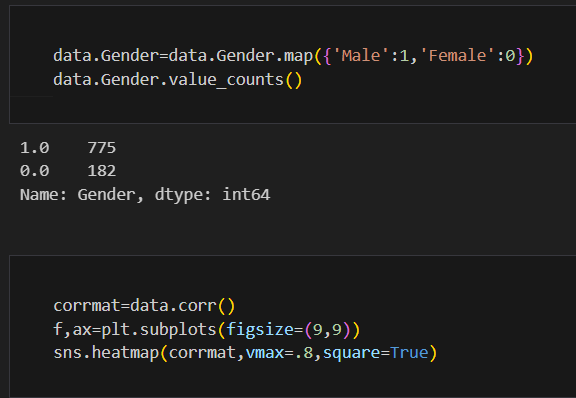
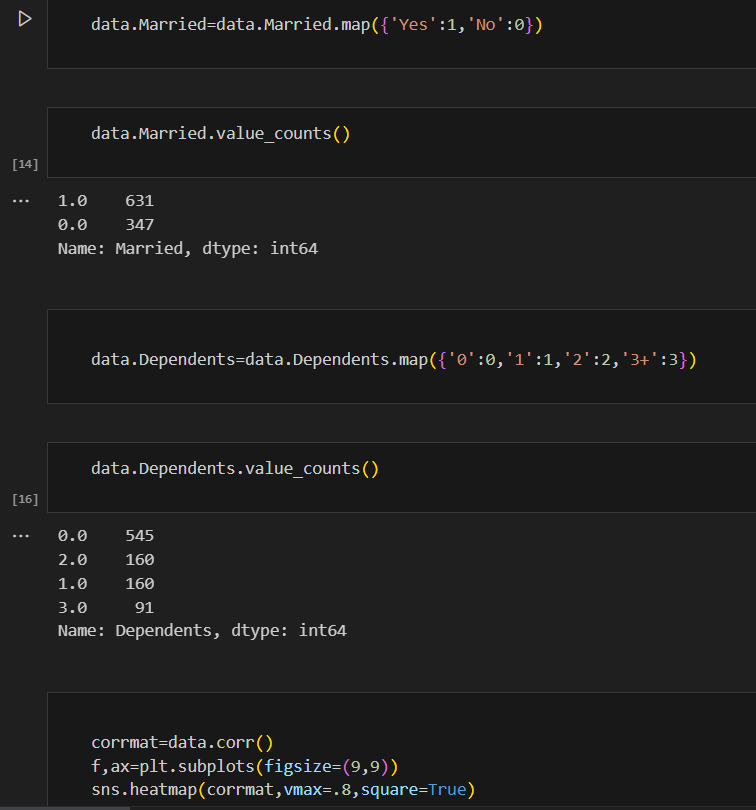
                "Account number: "+account\_no +' || '

                'Congratulations!! you will get the loan from Bank'

            )

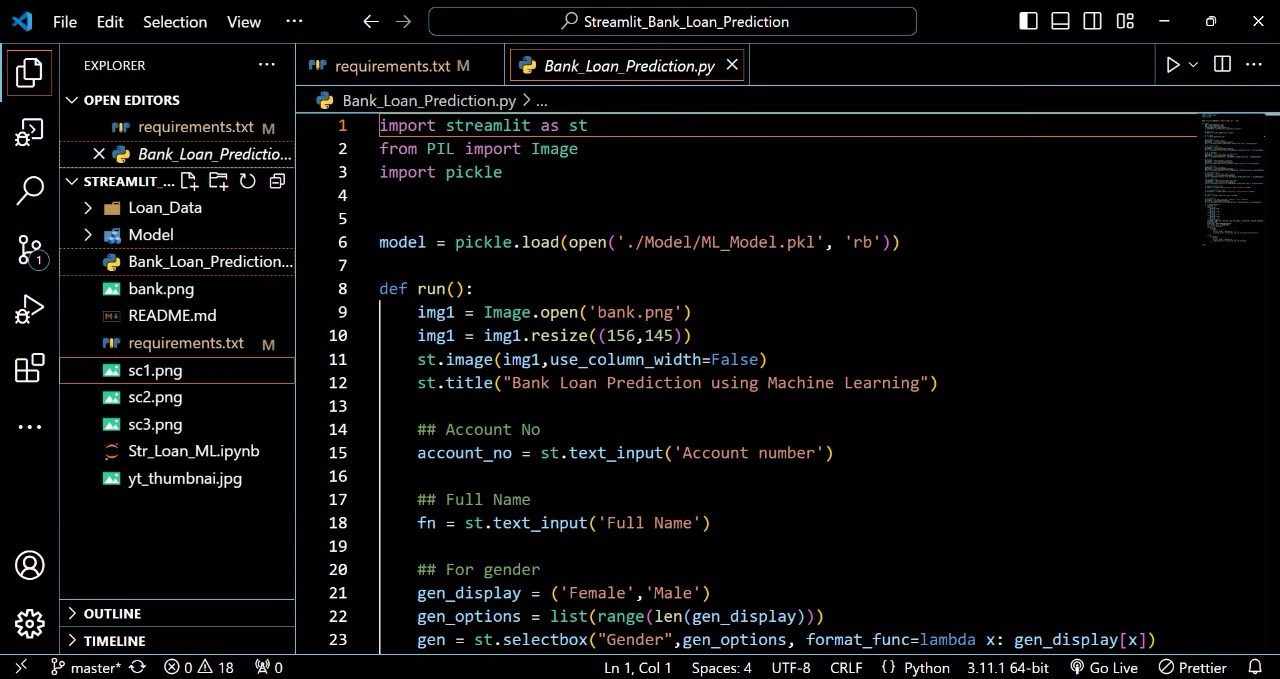
run()



**APPENDIX – 2**

**SCREENSHOTS**

**OUTPUT: VISUAL STUDIO CODE**

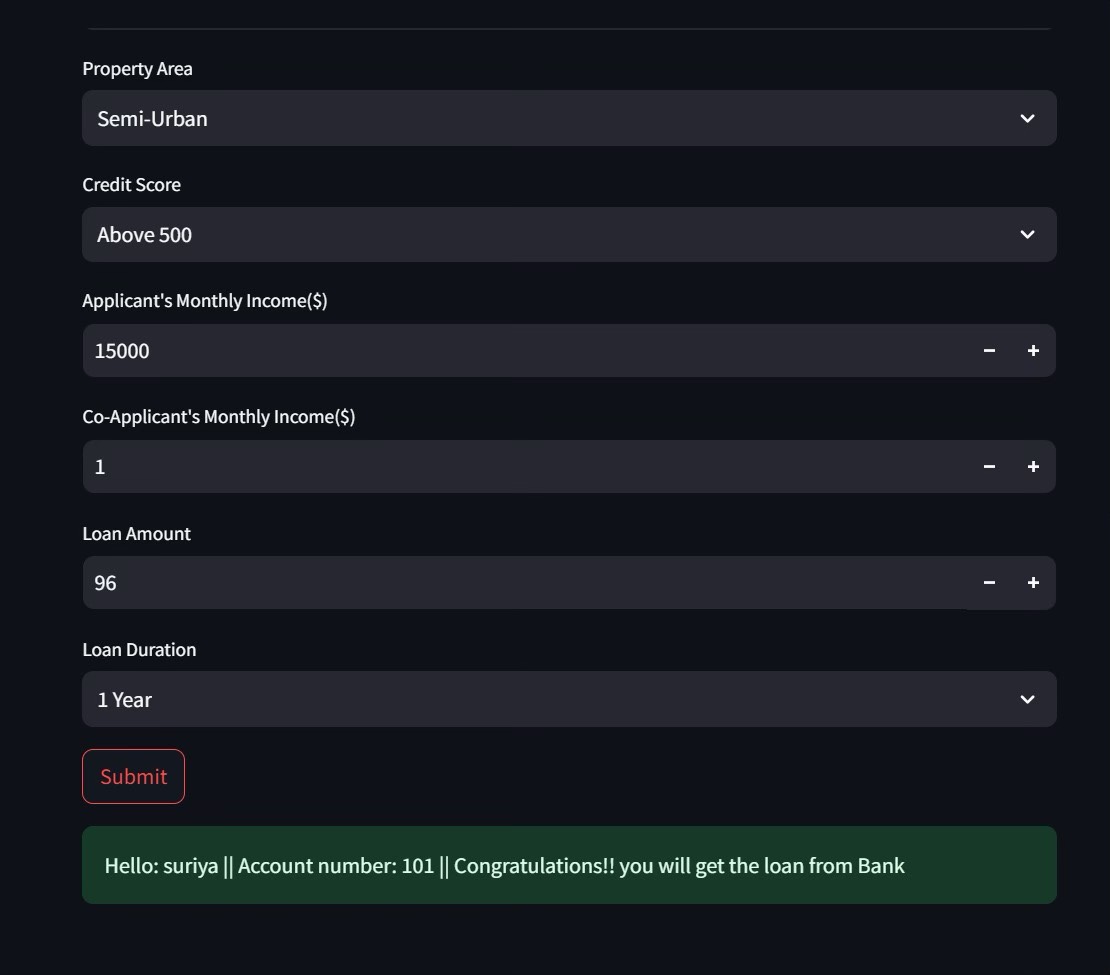
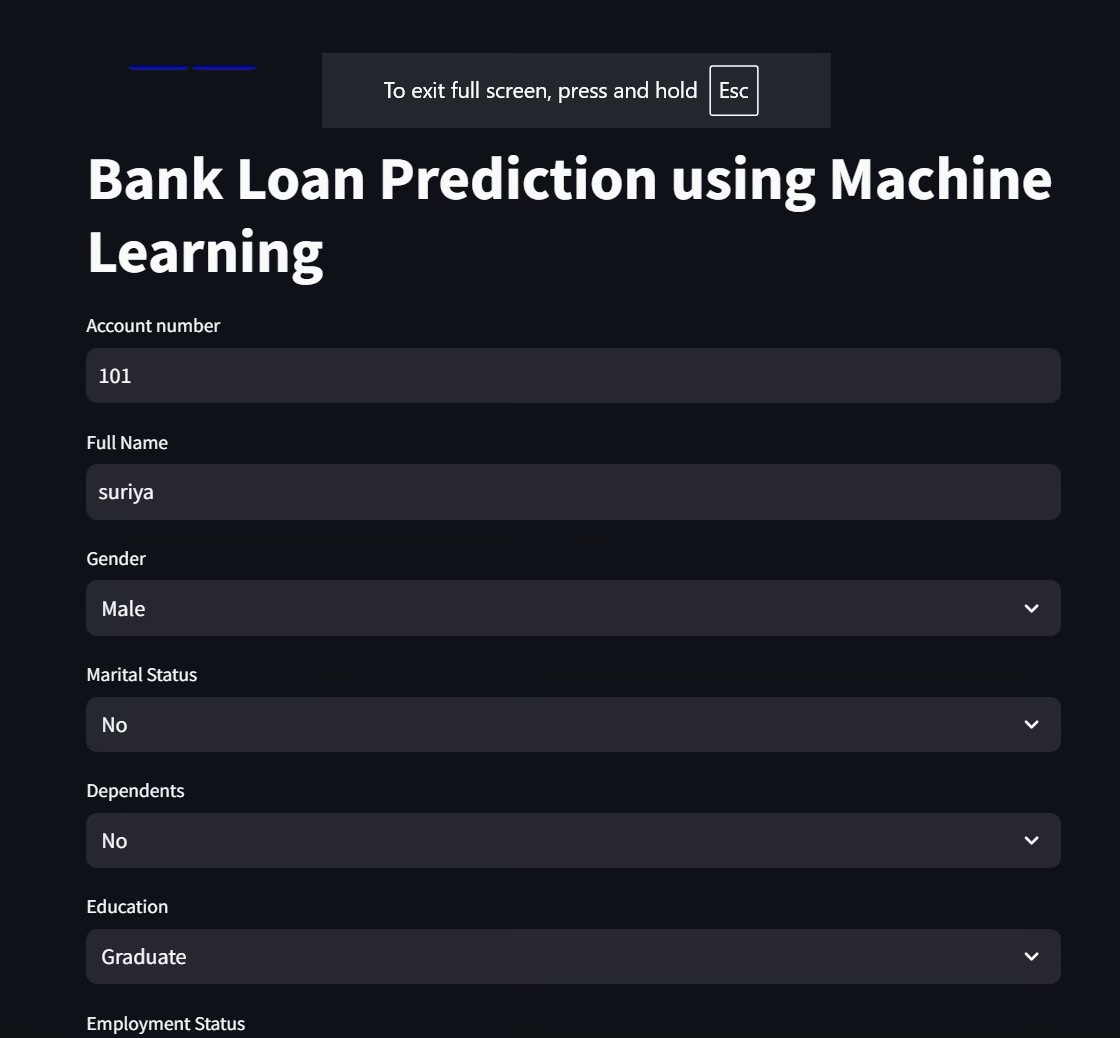
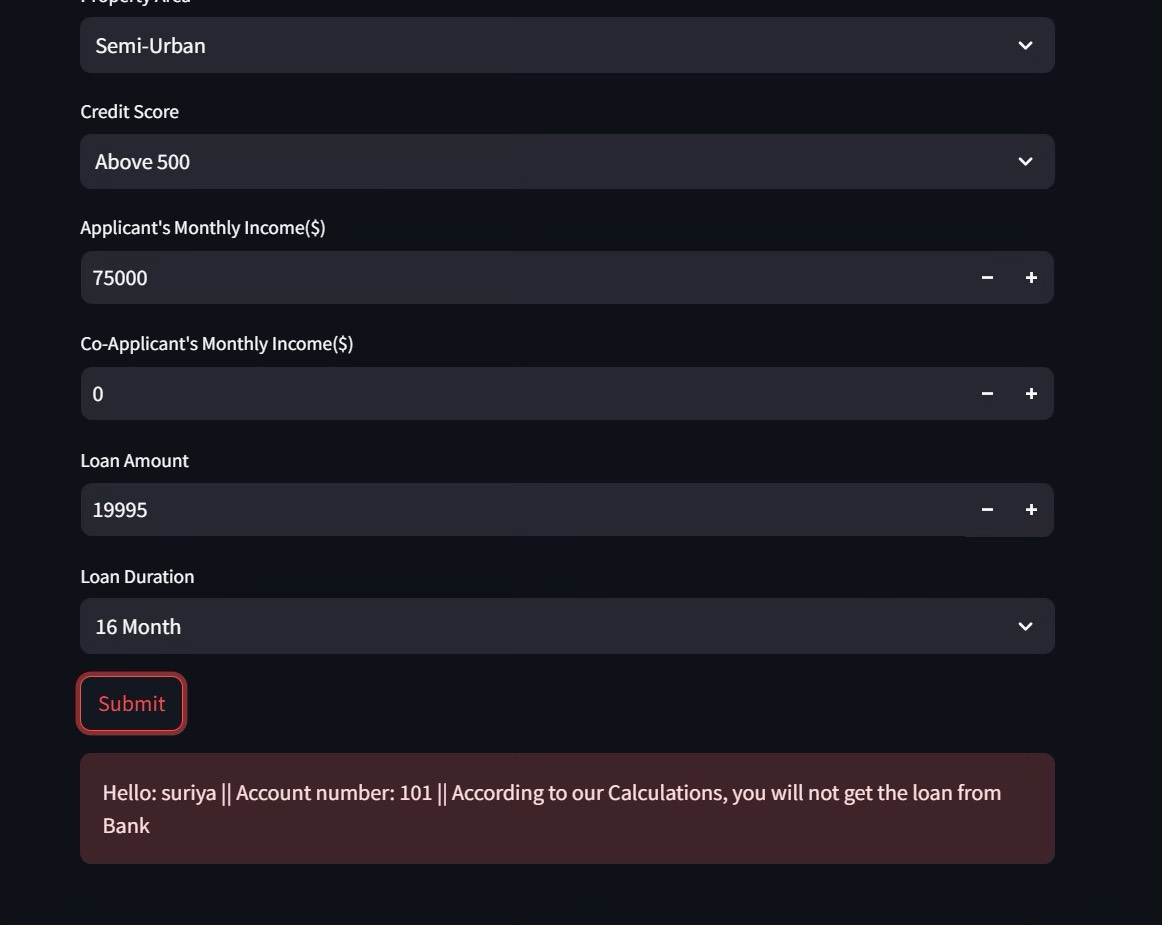
  Figure 8.1

Figure 8.2

 Figure 8.3

                                                                                                                                                                                                                                                                                                                                                                                                                                                                      Figure 8.4

**REFERENCES**

1] Loan default prediction,Authors: [Natasha Robinson](https://ieeexplore.ieee.org/author/439403446373049) [Nidhi Sindhwani](https://ieeexplore.ieee.org/author/37086301086),year(2024)

2**]** Predicting Loan Default Using Ensemble Learning Techniques,Authors:Sudipta Roy, Soumyajit Saha, and Sourav Maity,, Year(2019)

3] An Approach to Predict Loan Eligibility using Machine Learning,Author:Puneeth Br,Year(2022)

4] Predicting Loan Default Using Machine Learning: A Comparative Study, Author : Prateek Kumar, Aakash Mittal, Ankit Kumar, Year(2018)

5] Predicting Loan Default Using Machine Learning: A Case Study, Author: Rajat Kumar, Amit Kumar, and Sandeep Kumar, Year (2017)

6] Li Yujia, & Lu Jun. (2007). Risk Premium, Expected loss and forecast loan loss Provision. Contemporary Finance

and Economics (12), 7.

7] Jiang Zuobin, Xie Shuangqin, & Zhang Huan. (2010). Application of Arima model in the prediction of bank loan

scale. Finance and Economics (7), 3.

8] Chen Xulan, Han Suwan, & Pang Jianhua. (2021). Enterprise loan default risk prediction based on machine

learning method. Modeling and Simulation, 010(003), P.890-897.

9] Li Yunmeng, & Qian Xin. (2011). Long-term consumer loan forecasting based on co-integration and arma model.

Statistics and Decision (11), 3.

10] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, k Vikash,“Loan Prediction by using Machine Learning Models”, InternationalJournalofEngineering andTechniques.Volume 5 Issue 2, Mar-Apr 2019