Project Description

The project is to explore and evaluate various machine learning algorithms to predict loan approval for customers. Accurate loan approval predictions are essential for financial institutions to minimize risks and make informed decisions. By leveraging the power of machine learning, this project will assess the performance of different algorithms and determine the most effective model for loan approval predictions.

The technical process of the project includes the following steps:

1) **Data Collection:** The dataset will be collected from kaggle.

2) **Data Preprocessing:**The dataset will be cleaned and preprocessed to ensure its suitability for the machine learning algorithms.This step involves handling missing values, removing duplicates, normalizing numeric features and encoding categorical variables.

3) **Feature Selection / Engineering**: In this step the most informative features will be identified that have a significant impact on Loan Approval. Statistical tests, Correlation Analysis or domain knowledge will be used to select relevant features.

4) **Data Split:** Divide the preprocessed dataset into training and testing sets. The training set is used to train the machine learning models, while the testing set evaluates their performance on unseen data.

5) **Model Selection:** In this project the machine learning algorithms that will be used for loan approval prediction includes:

A. **Logistic Regression**: It is a generalized form of Linear Regression and is used whenever the outcome of a problem is dichotomous and is dependent on some other variables.
B. **Decision Tree** : A decision tree is a supervised machine learning algorithm that is used for both classification and regression tasks. It is a flowchart-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or predicted value.
C. **Random Forest:** Random Forest is a popular ML method that is part of the supervised learning technique. The machine learning algorithm is utilized to edit as well as retrieve issues. It is based on the integrated learning approach, which employs multiple disciplines to tackle complex issues and improve model

performance.It separates the tree into many decision trees for a variety of databases and metrics to improve prediction data accuracy. Rather than relying on a single tree for judgment, a random forest takes a forecast from each tree and backs it up with several predictable votes to predict the final outcome.

D. **Naive Bayes:** Naive Bayes theorem is a probabilistic algorithm used in machine learning and statistics. It is based on Bayes' theorem with an assumption of independence among the features.Naive Bayes is particularly useful for classification problems, where the goal is to assign an input to one of several predefined categories or classes.

E. **Support Vector Machine (SVM)**: The Support Vector Machine (SVM) algorithm is a powerful supervised machine learning algorithm used for both classification and regression tasks. It is particularly effective in solving complex classification problems where the decision boundary between classes is not linear.The main idea behind SVM is to find an optimal hyperplane that separates the data points of different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points of each class. This optimal hyperplane is chosen to have the largest margin to improve the generalization ability of the classifier.

F. **K-Nearest Neighbor**:The k-nearest neighbors (K-NN) algorithm is a simple and widely used machine learning algorithm for both classification and regression tasks.

6) **Model Evaluation:** All of these models will be implemented one by one and every model’s performance will be evaluated using evaluation metrics such as accuracy,precision,recall and F-1 Score. The testing set will be used to evaluate how well the models generalize to unseen data.

7) **Hyperparameter Tuning:** In this step the hyperparameters of the selected models will be fine-tuned to improve their performance. It involves techniques like grid search or random search to find the best combinations of hyperparameters.

8) **Model Selection:** Performance of all the models will be compared and the model that provides the best results based on the chosen evaluation metrics will be selected.

9) **Model Deployment:** Once the best performing model is selected it will be deployed into a production environment where it can be used for loan approval prediction.

Aims and Objectives

The aim of the project is to develop a reliable and accurate loan approval prediction model by exploring and evaluating various machine learning algorithms. The project aims to assist financial institutions in making data-driven decisions to minimize risks and enhance the loan approval process.

Objectives:

1. Identify the most suitable machine learning algorithms for the loan approval prediction task.
2. Assess and compare the performance of selected algorithms to determine their accuracy, precision, recall, and F1-score.
3. Develop an efficient and reliable loan approval prediction model to assist financial institutions in making data-driven decisions.
4. Investigate feature importance to understand which factors influence the loan approval process significantly.

Output Results

After comparing the performance of six different machine learning models on a given dataset using the accuracy metric. Here's a breakdown of the results:

	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	90.4255 %	0.89	1	0.94
K-Nearest Neighbor (KNN)	85.1064 %	0.87	0.93	0.9
Decision Tree Classifier	78.7234 %	0.91	0.81	0.85
Random Forest	90.4255 %	0.88	0.94	0.91
Gaussian Naive Bayes	84.0426 %	0.88	0.92	0.9
Support Vector Machine (SVM)	90.4255%	0.89	1	0.94

From the results, we can observe the following:

Accuracy: Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. It gives an overall measure of how well the model is performing. The models with the highest accuracy are Logistic Regression, Random Forest, and Support Vector Machine (SVM), all with an accuracy of around 90.43%.

Precision: Precision is the ratio of true positive predictions to the total positive predictions made by the model. It indicates how many of the positive predictions made by the model were actually correct. Models like Logistic Regression, K-Nearest Neighbor (KNN), Gaussian Naive Bayes, and Support Vector Machine (SVM) have relatively high precision values, ranging from 0.87 to 0.91.

Recall: Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total actual positive instances in the dataset. It indicates the ability of the model to correctly identify positive instances. In this case, Decision Tree Classifier and SVM have the highest recall of 1, which suggests that they are able to identify all positive instances.

F-1 Score: The F-1 Score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance by considering both false positives and false negatives. Among the models listed, the Decision Tree Classifier has the lowest F-1 score of 0.85, which indicates a trade-off between precision and recall.

Logistic Regression: This model shows good accuracy, precision, recall, and F-1 score values, making it a well-rounded performer in this scenario.

K-Nearest Neighbor (KNN): While KNN has a relatively lower accuracy and F-1 score compared to some other models, its precision and recall are still quite competitive.

Decision Tree Classifier: The Decision Tree Classifier has lower accuracy, precision, and F-1 score values compared to other models, but it has a high recall of 1, meaning it's very good at identifying positive instances.

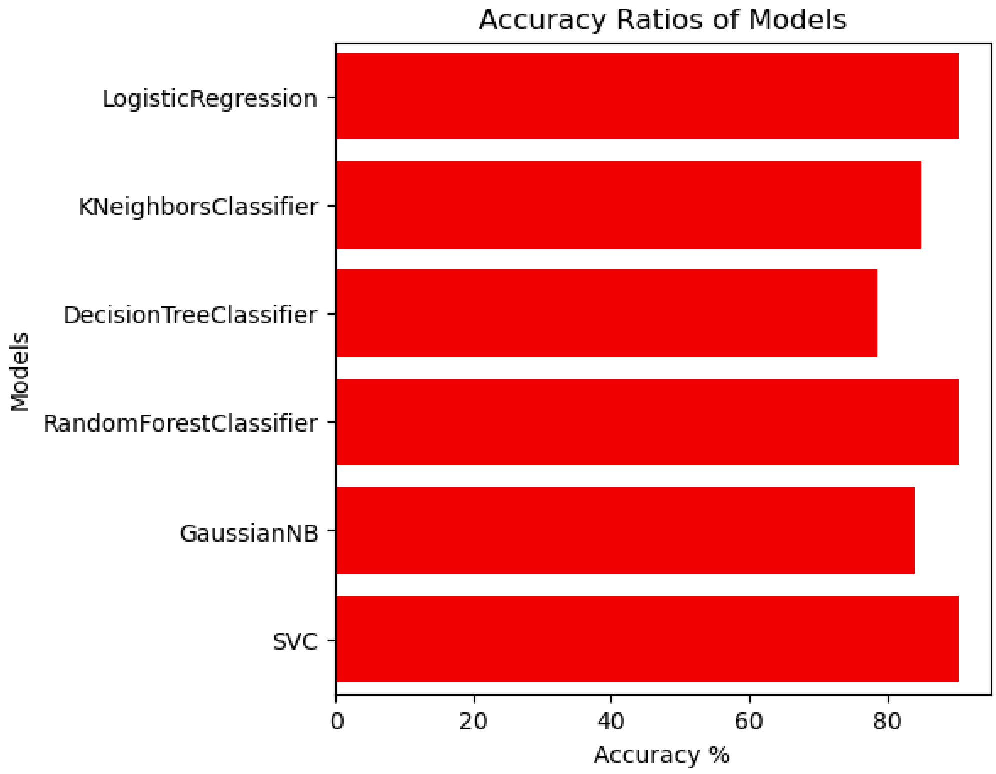
Random Forest: Random Forest demonstrates high accuracy and competitive precision, recall, and F-1 score values. It's a robust ensemble model that combines multiple decision trees to make predictions.

Gaussian Naive Bayes: This model has good precision and F-1 score values, but its recall is slightly lower, indicating that it might miss some positive instances.

Support Vector Machine (SVM): Like Logistic Regression, SVM shows consistently high performance across all metrics, with a recall of 1 indicating its ability to identify all positive instances.

Model Comparison

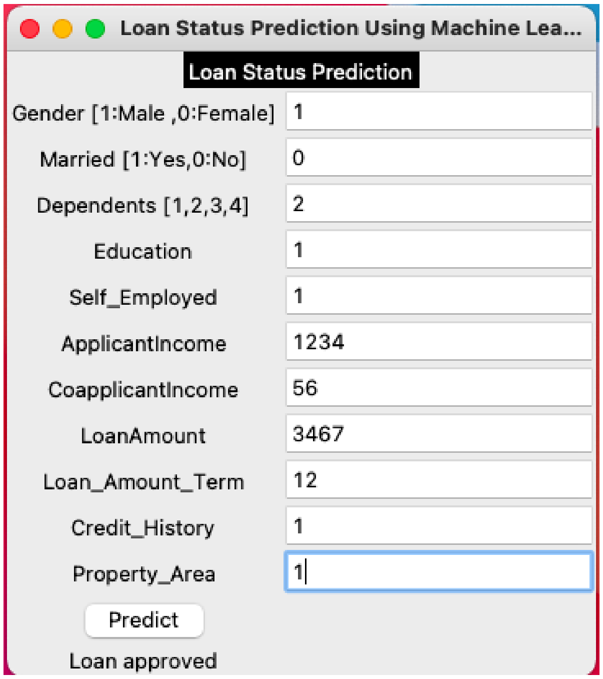
The figure shows the comparison of accuracy for all the six machine learning models. From the results we can see that the accuracy of the Logistic Regression, Random Forest and Support Vector Machine (SVM) model is the same.



Evaluation

After the comparative analysis of the machine learning models the best models that have high accuracy include Logistic Regression, Random Forest and Support Vector Machine. For the deployment Random Forest model has been selected and a Graphic User Interface has been developed to predict the customer loan approval.

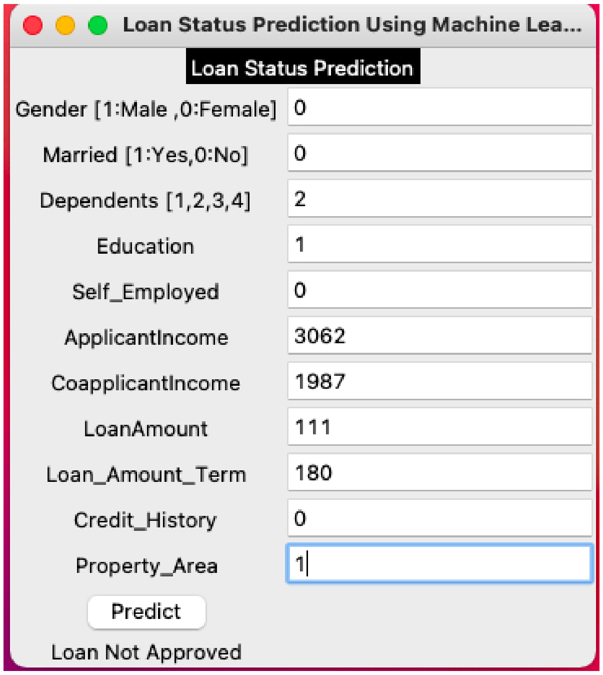
Graphic User Interface (GUI)



Loan Status Prediction

Gender [1:Male ,0:Female]	1
Married [1:Yes,0:No]	0
Dependents [1,2,3,4]	2
Education	1
Self_Employed	1
ApplicantIncome	1234
CoapplicantIncome	56
LoanAmount	3467
Loan_Amount_Term	12
Credit_History	1
Property_Area	1

Loan approved



Loan Status Prediction

Gender [1:Male ,0:Female]	0
Married [1:Yes,0:No]	0
Dependents [1,2,3,4]	2
Education	1
Self_Employed	0
ApplicantIncome	3062
CoapplicantIncome	1987
LoanAmount	111
Loan_Amount_Term	180
Credit_History	0
Property_Area	1

Loan Not Approved

References

Dataset : https://www.kaggle.com/code/vikasukani/loan-eligibility-prediction-machine-learning

Link for video :
https://drive.google.com/file/d/1Wj4eRrR6UGHtPX3NgwyzLMfDrdTIyhTZ/view?usp=sharing

Link for Presentation Slides :
https://docs.google.com/presentation/d/1j6JoWhjIQg54vqdHql95HYOW614hv3T-Mv-5KjOeTzw/edit?usp=sharing