Project presentation

Loan Approval Prediction

What is the problem we are attempting to solve?

1. Loans are vital in today's world, driving bank profits and enabling various needs like education and major purchases.
2. Assessing loan eligibility involves factors like Marital Status, Education, Income, and Credit History.
3. We'll use Python-based Machine Learning to predict profile relevance, enhancing decision-making efficiency.

Ideation process

01

Data Preprocessing and visualisation

02

Different model training and testing

03

Model selection

04

Final model evaluation

What have we accomplished so far?

1. We have performed dataset preprocessing.
2. We've carried out exploratory data analysis (EDA) on the dataset.
3. As a result of our analysis, we've identified the features with the highest correlation with the loan approval status.

Data Preprocessing

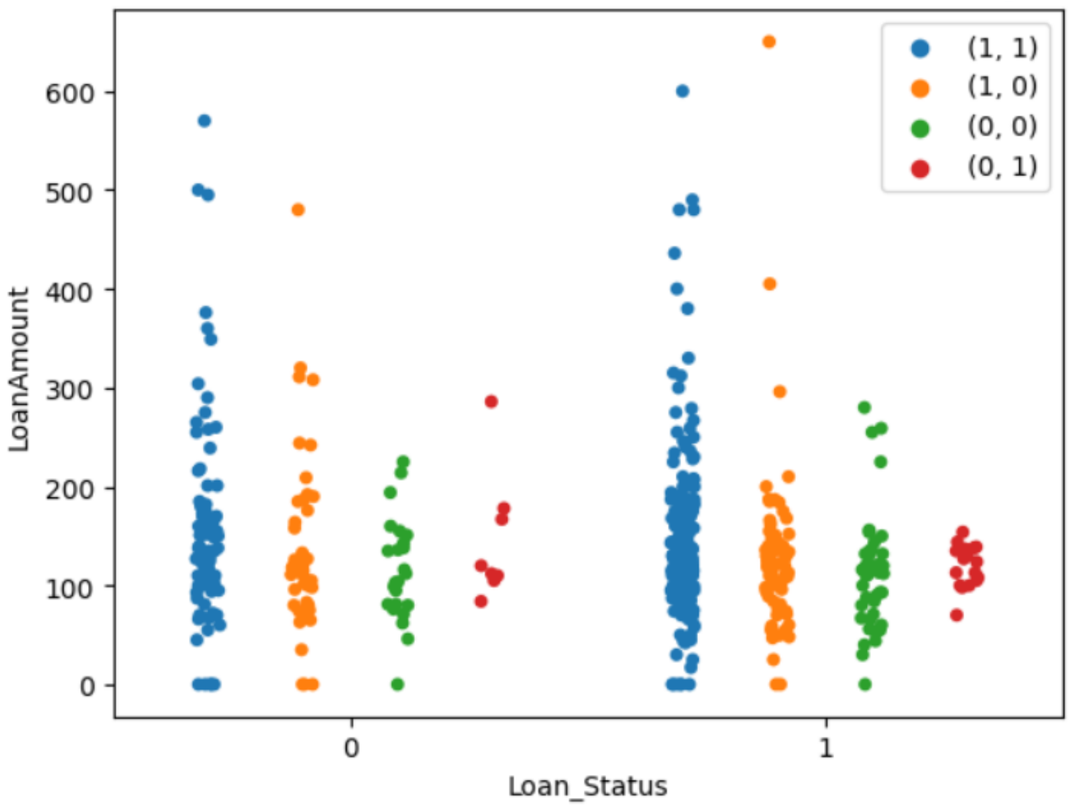
In data preprocessing, we've undertaken the following tasks:

1. **Importing necessary libraries**: Import NumPy, Pandas, Seaborn, and Matplotlib for data manipulation and visualization.
2. **Loading the dataset**: Read a CSV file ("LoanApprovalPrediction.csv") into a Pandas DataFrame and store it in the variable "data."
3. **Displaying the dataset**: Output the contents of the DataFrame to view the data.
4. **Handling Missing Values:** Dealing with incomplete data.
5. **Encoding the Dataset:** Converting categorical data into numbers.
6. **Normalizing:** Scaling numerical features to a common range.
7. **Adding Relevant Graphs:** Incorporating visualizations for insights.
8. **Correlation:** Analyzing relationships between variables.
9. **Splitting the Dataset:** Dividing data for model training and testing.

Data Visualization

Strip plot

- The strip plot visualizes the distribution of 'LoanAmount' values.
- It categorizes the data based on 'Loan_Status' into different groups.
- The 'Gender' and 'Marital Status' (Married/Not Married) of applicants are considered for differentiation.
- The 'dodge' parameter separates data points for each 'Loan_Status' category.
- This separation enables a straightforward comparison of 'LoanAmount' variations within each category, considering both 'Gender' and 'Marital Status' as contributing factors.



The following observations were made:

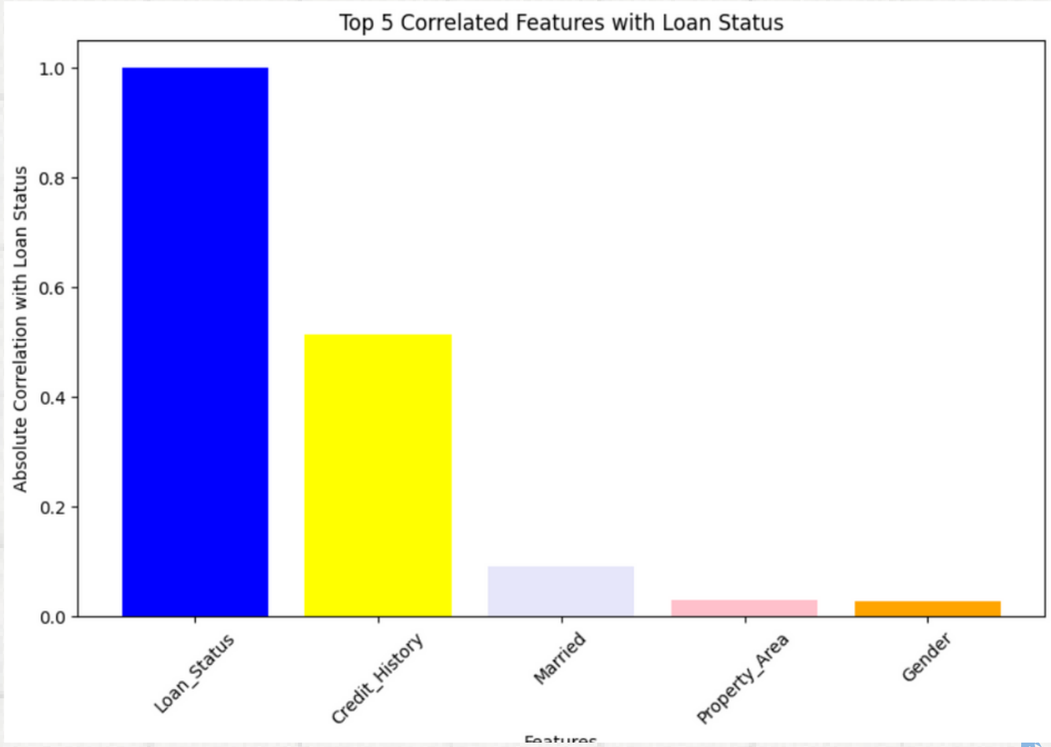
- Being a married male with a loan request below 200 increases the likelihood of loan approval.
- Being a married female with a loan request exceeding 150 decreases the likelihood of loan approval.

Data Visualization

Correlation

A correlation plot is used to:

- Visualize relationships between variables.
- Identify patterns and associations.
- Aid in feature selection by revealing correlations between variables.



Model exploration and comparative analysis

For Loan Approval Prediction

For the models we have mainly used 5 different models with parameter tuning using Grid Search. The models were :

- Linear SVM
- RBF SVM
- Polynomial SVM
- Logistic Regression
- Decision Tree

Finally we also used K-fold cross validation technique to assess the performance and generalizability of the chosen machine learning model, which majorly helped us for:

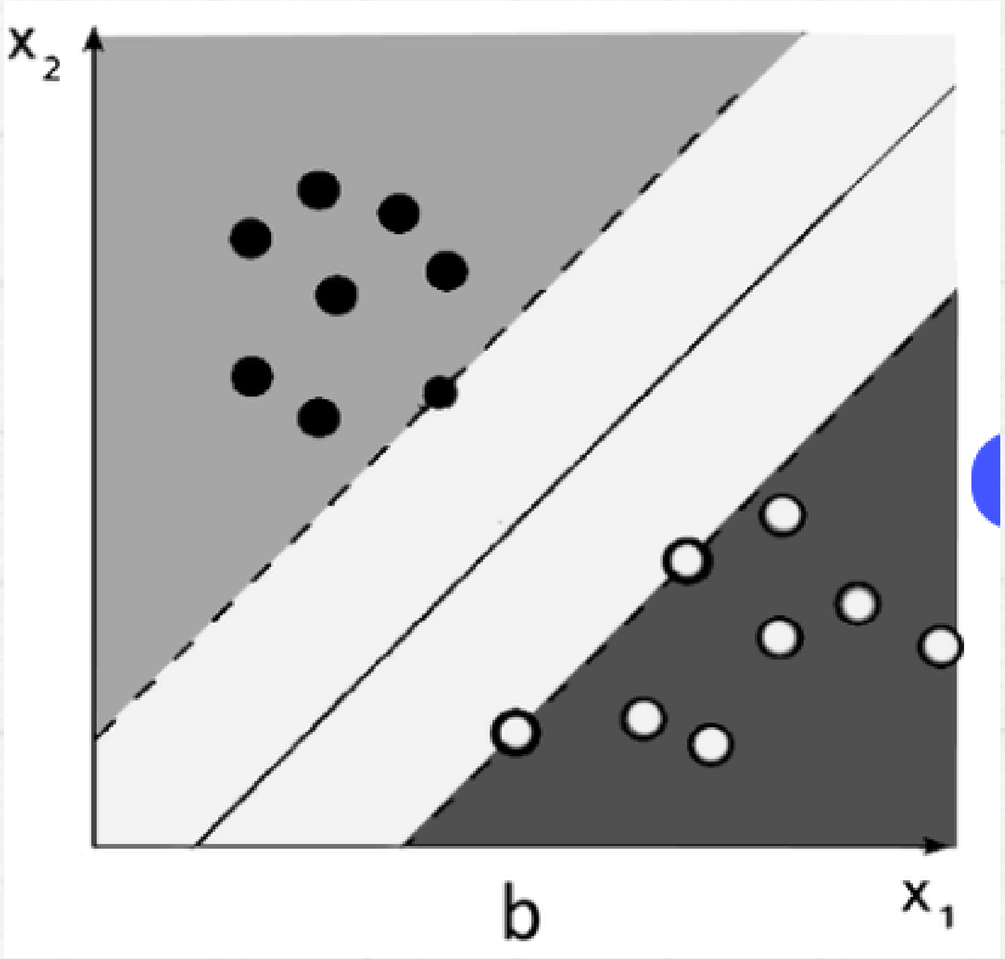
- Reducing variance in performance estimates
- Avoiding overfitting
- Assessing Model Stability

Linear SVM

For starters we used a linear kernel SVM, with parameter C = 10, chosen Randomly.

Results obtained were as follows:

- Accuracy on Dev Set = 81%
- Precision = 0.80
- Recall = 0.72
- F1 Score = 0.74

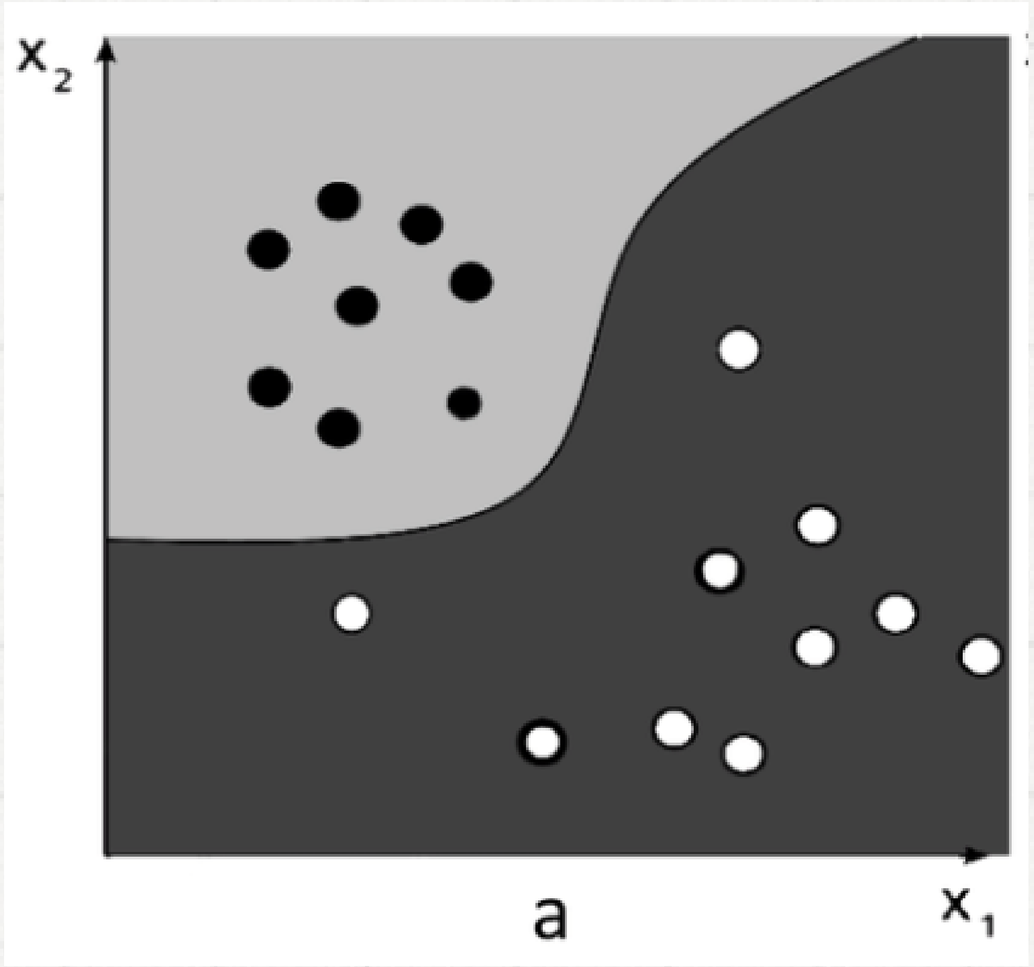


RBF SVM

Then we used a RBF kernel SVM, with random state parameter set as 42.

Results obtained were as follows:

- Accuracy on Dev Set = 81%
- Precision = 0.80
- Recall = 0.72
- F1 Score = 0.74

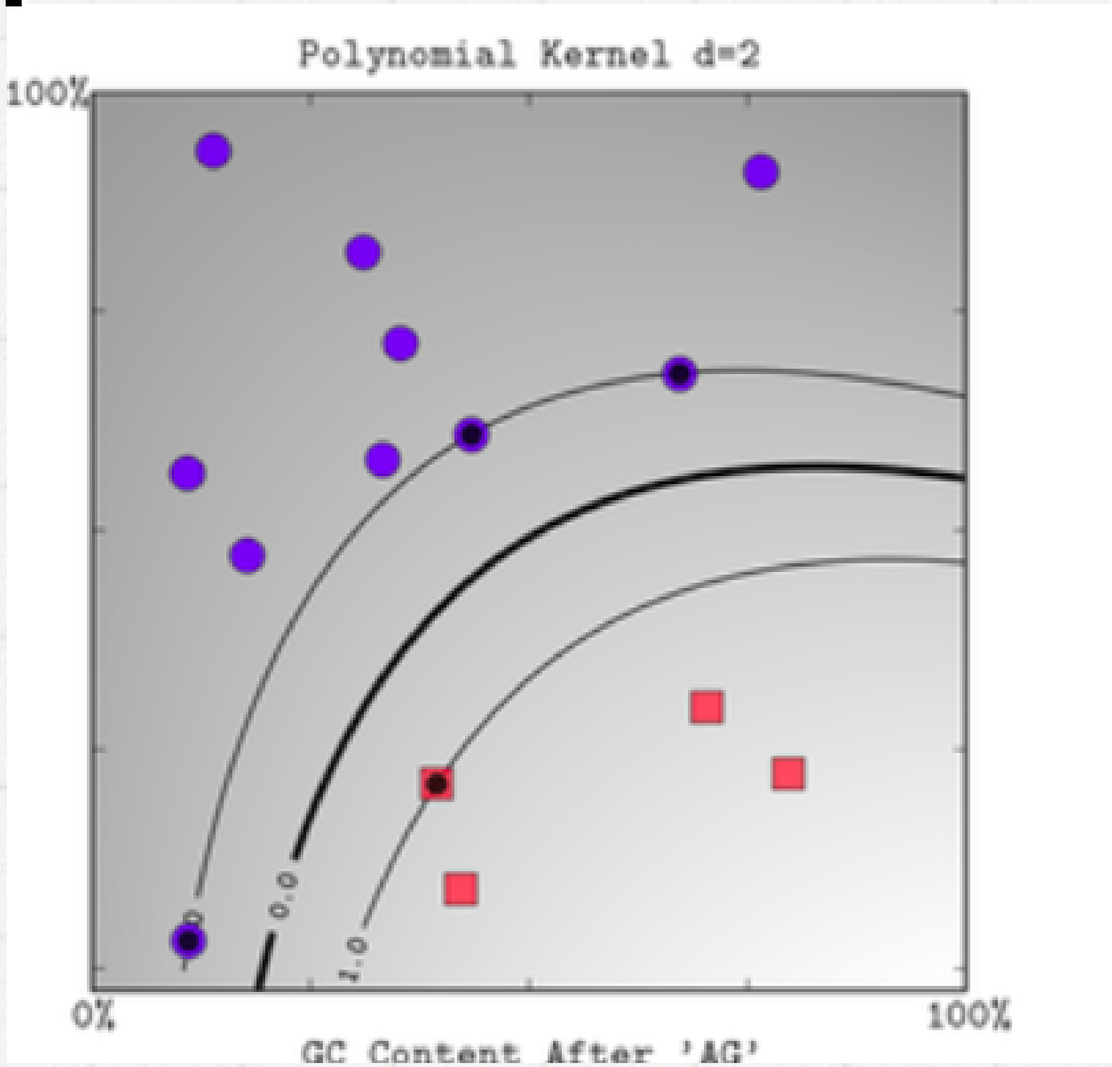


Polynomial kernel SVM

Then we used a polynomial with degree 3 kernel SVM, with random state parameter set as 42.

Results obtained were as follows:

- Accuracy on Dev Set = 75%
- Precision = 0.73
- Recall = 0.63
- F1 Score = 0.64



GridSearch on SVM's

First we do Grid Search CV for Linear Kernel SVM:

- with value of C = 0.1, 1, 10, 100
- We obtained the best accuracy of 81% for C = 10

Then, we do GridSearch CV for RBF kernel SVM:

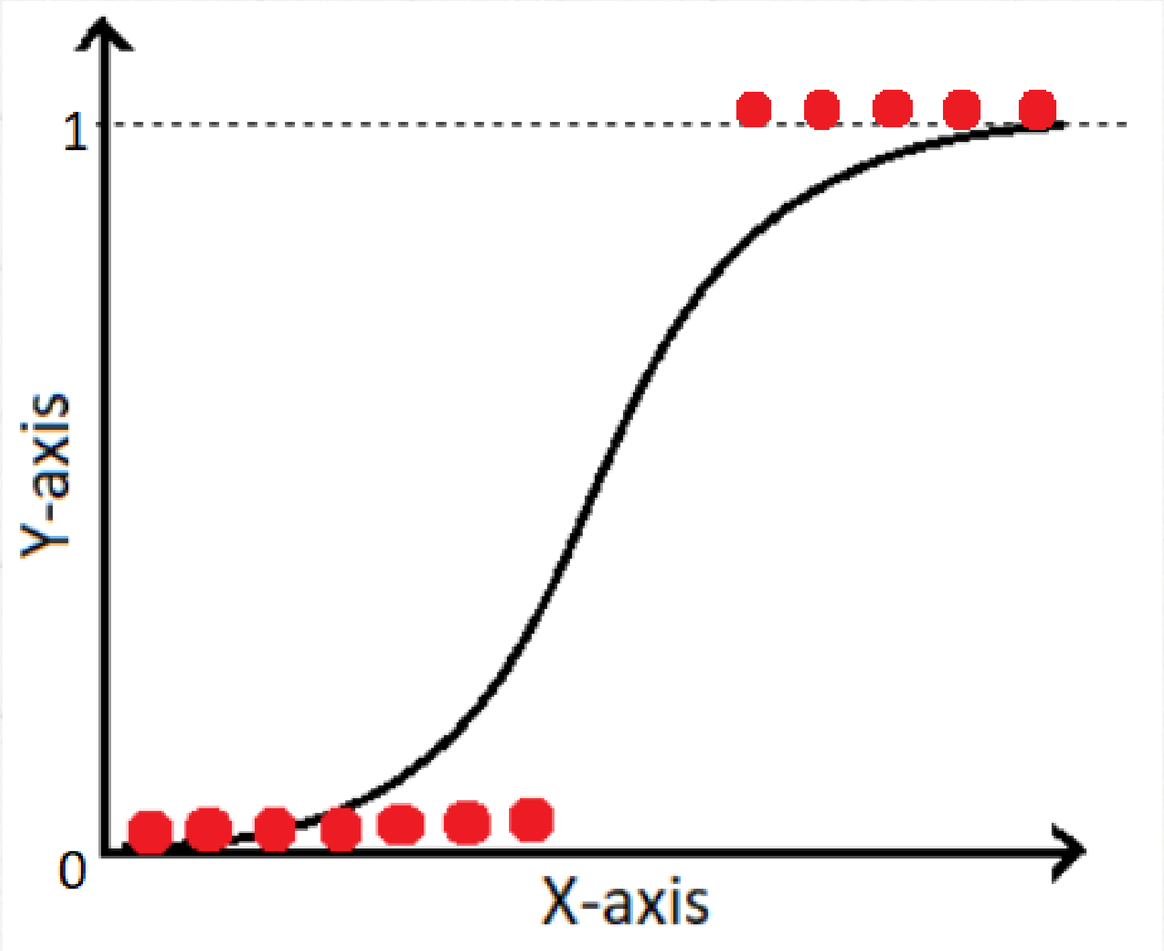
- with value of C = 0.1, 1, 10, 100 and Gamma as 0.001, 0.01, 0.1, 1
- We obtained the best accuracy of 74.35% for C = 10 and Gamma = 0.001

Logistic Regression

Then we used logistic Regression, with random state parameter set as 42.

Results obtained were as follows:

- Accuracy on Dev Set = 80.77%
- Precision = 0.80
- Recall = 0.72
- F1 Score = 0.74

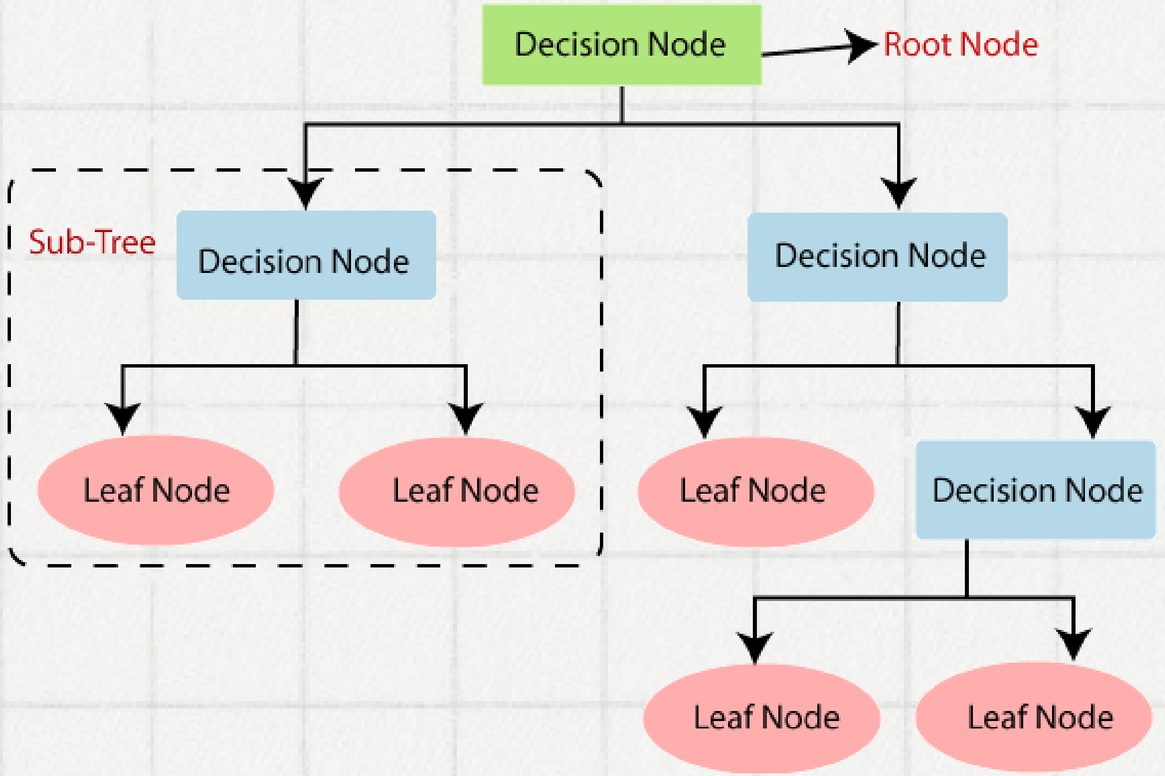


Decision Trees

Then we used Decision Tree Classifier, with random state parameter set as 42.

Results obtained were as follows:

- Accuracy on Dev Set = 71.15%
- Precision = 0.65
- Recall = 0.64
- F1 Score = 0.64



Final Model Selection

For the final model or the production model we would like to choose the Linear SVM with C = 10 as it had the best accuracy, precision, recall and F1 Score.

Results obtained were as follows:

- Accuracy on Dev Set = 85%
- Precision = 0.85
- Recall = 0.79
- F1 Score = 0.81

K-Fold Cross Validation

- We had performed K-fold cross validation on the final model.
- We used 4 different values of K (3,5,7,10) and calculated the accuracy for each of them.
- From the results we concluded that for K=10 we have the best accuracy.

Conclusion

In this project for Predicting Loan approval using Classical Machine Learning Techniques and other performance enhancing techniques. We have concluded that Linear SVM with C = 10 is the best model for this. It gives us an accuracy of 85% and it can be improved further by using the methods of bagging, boosting and Component analysis.

CONTRIBUTION

1. **Samridhi Jain(B22ES008) - SVM model + Report**
2. **Kashvi Pandya (B22BB045)- Logistics Regressionn Model + PPT**
3. **Palak Bhawsar (B22CH018)- Decision Tree + PPT**
4. **Shagun Suryavanshi (B22ME059)- Data- preprocessing**
5. **Shubham Kumar (B22CI038)- SVM model+ Final model**