Loan Default Prediction

Project Summary

Project Goal

To build a machine learning model that predicts whether a loan applicant is likely to repay or default on a personal loan, using historical applicant data. This helps financial institutions make data-driven decisions and reduce the risk of approving loans to high-risk applicants.

Chosen Model

After evaluating multiple algorithms including Logistic Regression, Decision Tree, and Random Forest, the final selected model is:

Logistic Regression

It provided the best balance between accuracy, interpretability, and performance on this dataset.

Final Accuracy

95.3% accuracy on the test data using Logistic Regression.

Evaluation Metrics:

- High Precision, Recall, and F1-Score for both classes (repay and default).
- The Confusion Matrix showed minimal misclassifications.

Model	Accuracy	Precision (1)	Recall (1)	F1-Score(1)
Random Forest	0.991	0.99	0.92	0.96
Decision Tree	0.988	0.95	0.93	0.94
Logistic Regression	0.95	0.86	0.66	0.75

Insights from Exploratory Data Analysis (EDA)

- Applicants with higher **Income** and **Education Level 3** were more likely to repay loans.
- The **CCAvg** (average monthly credit card spending) was a strong predictor of default risk.
- Applicants using **online banking** and those with a **CreditCard** showed slightly higher approval rates.
- Irrelevant columns such as **ID** and **ZIPCode** were removed from the dataset.