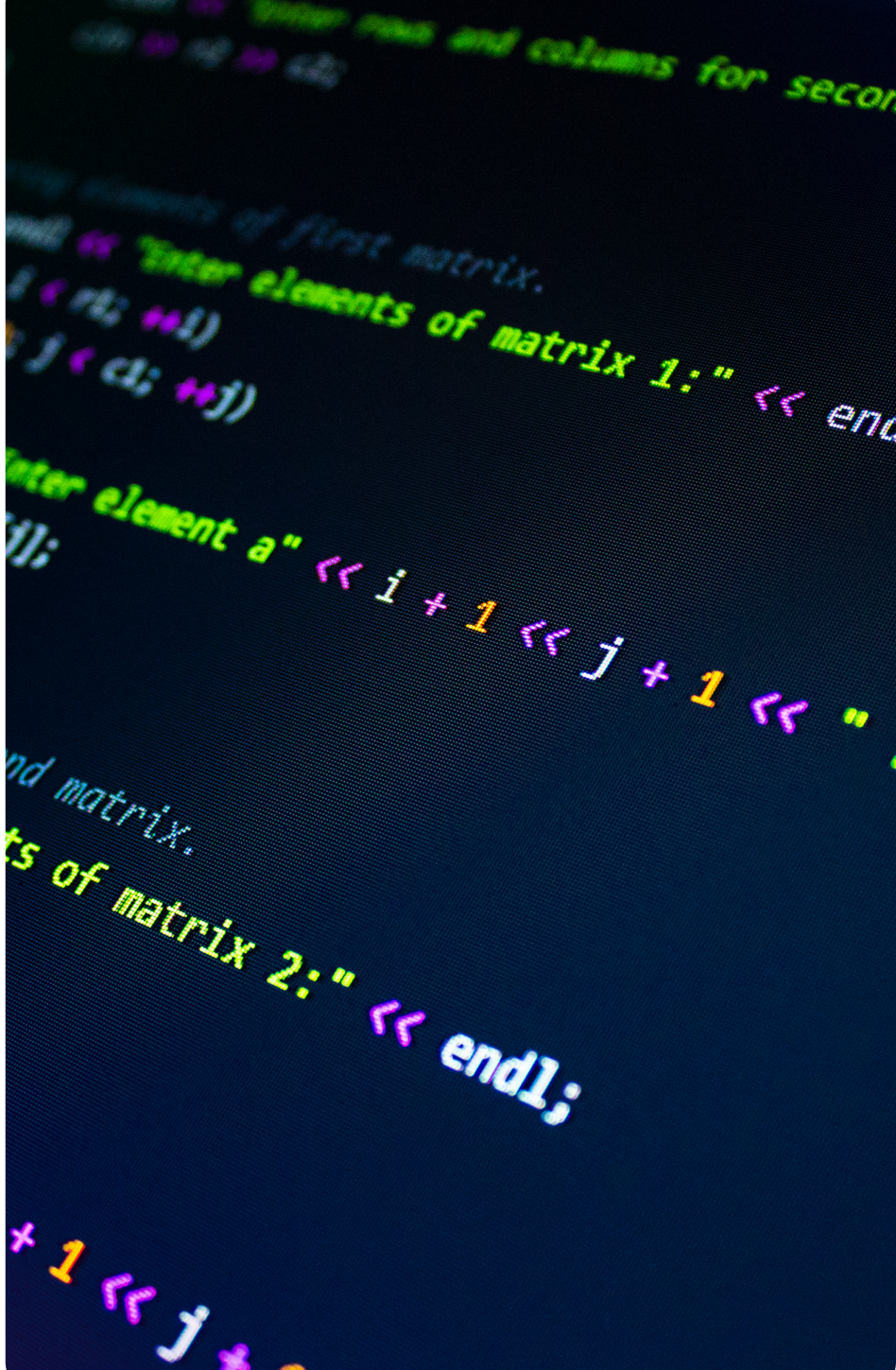
Machine Learning Final Project

Loan Approval Classification

SIS - 2203



Problem & Business Value

Problem

- Banks receive thousands of loan applications every day
- Manual review can take up to several days and requires experienced credit analysts
- Approval errors increase default risk, while rejection errors lead to customer loss

Business Value

- Automated decision-making within seconds
- Reduced operational costs and analyst workload
- More consistent and objective loan approval decisions

Project Goal

- Build a machine learning model that predicts loan approval outcomes
- Provide fast and reliable Approved / Rejected decisions to support banking operations

Machine Learning Approach

Problem Formulation:
Binary classification task

Target Variable

- Class 1: Loan Approved (Loan_Status = Y)
- Class 0: Loan Rejected (Loan_Status = N)

Evaluation Metrics

Accuracy — overall prediction correctness
F1-score — balance between Precision and Recall

Tools & Technologies

Python
scikit-learn
Random Forest classifier
Google Colab

Dataset

Kaggle dataset: Eligibility Prediction for Loan 2000 loan applications (e.g., income, education, credit history, etc.)

- License & Usage
- Public domain — free for educational purposes

Additional Info:

- Collected in 2019, India
- Target market: retail loans up to ₹300,000

Example of a Loan Application

Field	Value	Description
Gender	Male	Demographic information
Married	Yes	Marital status
Dependents	2	Number of dependents
Education	Graduate	Education level
Self_Employed	No	Self-employed or not
ApplicantIncome	$5,400	Applicant's monthly income
Credit_History	1	Any past defaults? (1 = No, 0 = Yes)
LoanAmount	$120,000	Requested loan amount
Target	Y	Bank's final decision (Approved / Rejected)

Key Insights from EDA

Class Distribution

40% of applications approved, 60% rejected

Slight class imbalance, not critical

Most Informative Feature

Credit_History = 40% approval

Credit_History = 60% denial

Income Analysis

Average income of approved applicants: $5,900

Average income of rejected applicants: $3,800

Location Effect

Urban applicants approved more often than Rural (12% difference)

Visualization

images (.png), (countplot of loan approvals)

Data Pipeline

Pipeline Steps

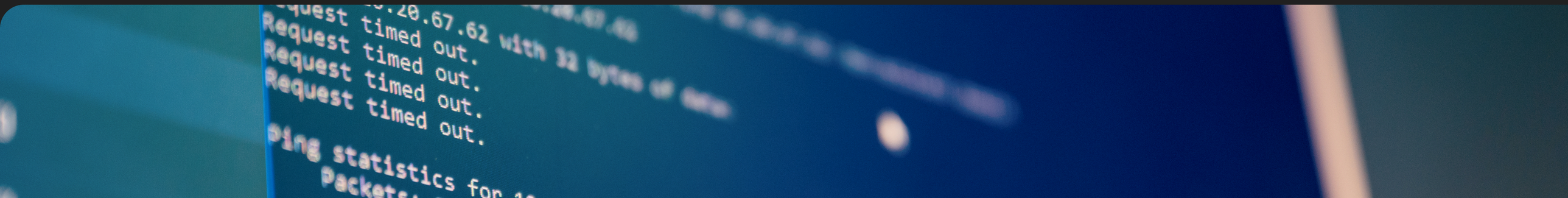
- Data acquisition
- Remove missing values (~5%)
- Encode target variable (Y/N → 1/0)
- Train/Test split (80/20, stratified)
- Model training
- Model evaluation

Tools & Environment

Entire pipeline implemented in Google Colab, no local Python required

Visualization

Represent pipeline with 5 rectangles and arrows in slides to show the workflow



Conclusion & Takeaways

Summary

- Built an end-to-end machine learning pipeline for loan approval prediction
- Automated decision-making reduces manual review time from days to seconds

Key Insights

- Credit history is the strongest predictor of loan approval
- Income level and applicant location also significantly influence decisions

Results

- The model demonstrates stable performance on a slightly imbalanced dataset
- Selected evaluation metrics ensure a balanced assessment of model quality

Impact

- Reduces operational costs and analyst workload
- Supports faster, more consistent, and data-driven credit decisions

Future Work

- Add more advanced models and hyperparameter tuning
- Include fairness and bias analysis
- Deploy the model as a web service or API

THANK YOU

SIS-2203

Madiyar Mustafin
Alisher Toleubay
Damir Izenbayev