**Project Report: Loan Prediction Analysis Using Machine Learning**

**Submitted By: Surbhi Chauhan, Menpal Dhundhara**

**1. Introduction**

This project focuses on predicting loan approval status using machine learning techniques. The dataset used contains demographic, financial, and loan-related information. The analysis includes data preprocessing, feature engineering, visualization, and model training using Logistic Regression and XGBoost.

**2. Objectives**

1. Preprocess and clean the dataset.
2. Engineer features to improve model performance.
3. Visualize data insights using plots such as heatmaps, violin plots, Sankey diagrams, and radar charts.
4. Train and evaluate machine learning models.
5. Explain model predictions using SHAP values for interpretability.

**3. Methodology**

**3.1 Data Preprocessing**

* **Missing Values Handling:** Missing values were imputed using mode or median depending on the feature type.
* **Categorical Encoding:** One-hot encoding was applied to categorical variables.
* **Outlier Handling:** Winsorization was used to cap extreme values at the 5th and 95th percentiles.

**3.2 Feature Engineering**

* Created new features such as:
  + *TotalIncome*: Sum of applicant and co-applicant income.
  + *Debt-to-Income Ratio (DTI)*: Loan amount divided by total income.
  + *LoanTermYears*: Loan term converted into years.

**3.3 Visualization**

* Generated insights through:
  + Correlation heatmaps to identify relationships between features.
  + Violin plots to visualize DTI distribution by loan status.
  + Sankey diagrams to show flow between gender and education categories.
  + Radar charts to compare normalized financial indicators.

**3.4 Model Development**

Two machine learning models were trained:

1. Logistic Regression
2. XGBoost (with hyperparameter tuning via GridSearchCV)

**3.5 Model Evaluation**

Evaluation metrics included:

* Classification report (precision, recall, F1-score).
* AUC-ROC scores for model comparison.

**3.6 Explainability**

SHAP (SHapley Additive exPlanations) was used for:

* Global feature importance.
* Local explanations for individual predictions.

**4. Results**

**4.1 Visualization Outputs**

1. Correlation Heatmap: Displays relationships between numerical features.
2. Violin Plot: Shows DTI distribution across loan statuses.
3. Sankey Diagram: Highlights flow between gender and education categories.
4. Radar Chart: Compares normalized financial indicators.

**4.2 SHAP Insights**

SHAP visualizations provided interpretability for:

* Global feature importance (Credit History was the most impactful feature).
* Individual predictions (e.g., explanation for Applicant #10).

**5. Conclusion**

The project successfully demonstrated the prediction of loan approval status using machine learning models with robust preprocessing and visualization techniques. XGBoost outperformed Logistic Regression in terms of AUC-ROC scores.