Loan Default Risk Prediction

Objective

The objective of this project is to develop a model that accurately predicts **loan default risk** using the provided dataset. By analyzing key financial indicators, the model aims to assist lenders in making informed loan approval decisions.

Methodology

Data Loading and Preprocessing

- The loan default risk dataset was **loaded and examined** for inconsistencies.
- Missing values were **imputed using the mean** for 'Debt_Amount' and 'Monthly_Savings'.
- Duplicate entries were checked and removed to ensure data integrity.

Exploratory Data Analysis (EDA)

- **Histograms** were used to visualize the distributions of 'Retirement_Age', 'Debt_Amount', and 'Monthly_Savings'.
- A **correlation heatmap** was generated to identify relationships between variables.
- **Boxplots** were utilized to detect outliers in 'Retirement_Age' and 'Debt_Amount'.
- The relationship between features and **loan default risk** was explored using boxplots.
- A **log transformation** was applied to 'Debt_Amount' to correct for skewness.

Feature Scaling

- 'Debt_Amount', 'Monthly_Savings', and 'Retirement_Age' were **standardized** using StandardScaler to ensure uniform feature scaling.

Data Splitting and Balancing

- The dataset was split into **80% training and 20% testing sets**.
- **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to address class imbalance by increasing the minority class samples.

Model Training and Evaluation

- A **Support Vector Machine (SVM) classifier with an RBF kernel** was trained on the dataset.

- The model's performance was evaluated using **accuracy, precision, recall, F1-score, and a confusion matrix**.

Results

- **Accuracy:** The SVM model achieved an accuracy of **0.95 (95%)** on the test dataset, indicating strong predictive performance.
- **Classification Report:**
  - The model provided high **precision and recall** scores, ensuring balanced performance across both default and non-default cases.
- **Confusion Matrix:**
  - The confusion matrix was analyzed to understand the distribution of **true positives, true negatives, false positives, and false negatives**.
  - Minimal misclassifications were observed, reinforcing the model's reliability in distinguishing between defaulters and non-defaulters.

Recommendations for Lenders

1. **Utilize the Developed Model**
   - Integrate the trained **SVM model** into the loan approval process to assess the risk of default before approving loans.
2. **Focus on Key Features**
   - Based on EDA findings, **'Debt_Amount' and 'Retirement_Age'** significantly influence loan default risk.
   - Lenders should carefully evaluate applicants with **high debt levels or early retirement ages**.
3. **Set Appropriate Decision Thresholds**
   - Depending on the lender's **risk tolerance**, the probability threshold for classifying applicants as **high risk** can be adjusted.
4. **Monitor and Update the Model**
   - The model should be **regularly retrained** on new loan applications to ensure continued accuracy.

   - Economic conditions and borrower behaviors change over time, requiring periodic updates to maintain predictive performance.

5. **Consider Additional Features**
   - Future iterations of the model could include additional factors such as **credit history, income levels, employment stability, and past loan repayment behavior** to further enhance prediction accuracy.