Task 4 Report – Loan Default Prediction

1. Title:

Predicting Loan Default Risk Using German Credit Dataset

2. Objective

This project aims to build a predictive machine learning model that classifies whether a loan applicant is likely to **default** (fail to repay) or **repay** the loan in full. The model will assist lenders in assessing loan risks and taking preventive decisions.

3. Dataset Overview

Feature	Description
Dataset Source	UCI Machine Learning Repository
Dataset Name	Statlog (German Credit Data)
Number of Records	1,000
Number of Features	20 (categorical + numerical)
Target Variable	Target: 0 = Fully Paid, 1 = Defaulted

Each record represents a loan applicant with features like credit history, loan amount, employment status, housing, etc.

4. Data Preprocessing

- **Column headers** were assigned manually from documentation.
- **Categorical features** (e.g., employment, housing) were encoded using Label Encoding.
- **Target Variable Mapping**:
  - 1 → Default (High Risk)
  - 0 → No Default (Low Risk)
- **Missing Values**: No missing values were present in the dataset.
- **Feature Scaling**: StandardScaler was used to normalize numeric columns.

5. Handling Class Imbalance

The dataset is slightly imbalanced (70% paid, 30% defaulted).
To address this:

- Applied **SMOTE (Synthetic Minority Over-sampling Technique)** on training data
- Balanced both classes (defaults = non-defaults)

6. Model Selection and Training

Two models were tested:

Model	Reason for Choice
LightGBM	Fast, interpretable, high performance
SVM (RBF Kernel)	Good for classification with scaled data

Final training was done using the **LightGBM Classifier**, as it offered the best results.

7. Model Evaluation

- Evaluation done on unseen 20% test set
- Metrics Used: Precision, Recall, F1 Score, AUC-ROC
- **Best Model: LightGBM**

Classification Report Summary:

Metric	Score
Precision	~0.77
Recall	~0.78
F1-Score	~0.77
AUC-ROC	~0.80

ROC Curve was plotted to visualize model performance.

8. Key Insights

1. **Credit History** and **Credit Amount** are top predictors of default.
2. Short employment duration and low savings increase risk.
3. SMOTE significantly improved the model's ability to detect defaults.
4. LightGBM showed strong generalization and is deployment-ready.
5. This model can be used by banks to auto-flag high-risk loan applications.

9. Conclusion

The project successfully built a reliable and interpretable classification model for loan default prediction using the German Credit dataset. By identifying risky customers early, the model can assist financial institutions in making data-driven lending decisions and reducing default-related losses.

10. Future Recommendations

- Consider using a larger and more recent loan dataset for higher real-world relevance.
- Explore explainability tools like **SHAP** to show why the model predicts default.
- Integrate model into lending systems for real-time risk scoring.