Project Report: End-to-End Machine Learning

# Introduction

This project applies machine learning techniques to a synthetic dataset with multiple numerical and categorical features. The main objective is to predict the likelihood of default risk using a systematic pipeline of EDA, preprocessing, model building, and hyperparameter tuning.

# Task 1: Exploratory Data Analysis (EDA)

- Basic Checks: Verified dataset shape, data types, summary statistics.  
- Missing Values: Identified and quantified missing data across numerical and categorical variables.  
- Target Variable Analysis: Checked class balance for `default\_risk`.  
- Visualizations:  
 - Histograms and boxplots for distribution and outlier detection.  
 - Bar charts for categorical features.  
 - Heatmap and pair plot for feature correlation and relationships.

# Task 2: Data Preprocessing

- Categorical Encoding: Used ColumnTransformer with OneHotEncoding/OrdinalEncoding.  
- Missing Value Imputation:  
 - Numerical columns filled with mean/median.  
 - Categorical columns filled with mode.  
- Scaling:  
 - MinMaxScaler and StandardScaler applied to numerical features for model compatibility.

# Task 3: Model Building

Trained multiple models on the processed dataset using different scaling strategies:  
- Logistic Regression  
- Support Vector Machine (SVM)  
- Decision Tree  
- Random Forest  
- XGBoost  
- Perceptron  
- Multi-Layer Perceptron (MLP Classifier)

# Task 4: Hyperparameter Tuning

- Applied GridSearchCV to optimize hyperparameters for each model.  
- Compared model performance across different scalers.  
- Evaluated models using accuracy, precision, recall, and F1-score.

# Conclusion

This project demonstrates a complete machine learning workflow:  
1. EDA for data understanding.  
2. Preprocessing for clean and consistent data.  
3. Training and comparing multiple models.  
4. Hyperparameter tuning for optimization.  
  
The pipeline ensures robust preparation and evaluation, making it suitable for real-world applications in credit risk prediction.