Loan Default Prediction – Model Evaluation Report

The objective of this project was to build a machine learning model to predict whether a loan would be approved or not, based on various applicant features such as income, employment status, credit history, and property area. The dataset used for this analysis was the Loan Prediction dataset from Kaggle, which includes both categorical and numerical features. After preprocessing the data, including handling missing values and encoding categorical variables, three machine learning models were trained: Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier.

Each model was evaluated using standard classification metrics such as accuracy, precision, and recall. The Decision Tree model achieved an accuracy of 69.1%, a precision of 75.6%, and a recall of 77.5%. Gradient Boosting performed better, with an accuracy of 73.9%, precision of 75.0%, and recall of 90.0%. However, the Random Forest model stood out by achieving the highest accuracy of 75.6% and the highest recall of 93.8%, along with a precision of 75.0%. Random Forest also provides meaningful insights into feature importance. In our case, the most important features for predicting loan status were Credit_History, ApplicantIncome, LoanAmount, Education, and Married.

Based on the overall performance and robustness, the Random Forest Classifier is selected as the best model for this project. It provides a strong balance between precision and recall, handles overfitting effectively, and works well with mixed-type data. Future enhancements to this project could include hyperparameter tuning using GridSearchCV, handling class imbalance using techniques like SMOTE, and eventually deploying the model in a web or mobile application for real-time loan predictions.