**Project 1 Report: Web3gle Intership**

**Loan Approval Prediction**

**Introduction**

This project focuses on predicting loan approval using various machine learning models. We have explored and analyzed a dataset consisting of loan applications, aiming to identify the key features influencing loan approval and evaluate the performance of different models.

**Data Preprocessing**

1. **Null Values**: No null values were present in the dataset.
2. **Duplicates**: No duplicate entries were found.

**Exploratory Data Analysis (EDA)**

1. **Descriptive Statistics**:
   * Key observations included mean, standard deviation, min, and max values for features such as income\_annum, loan\_amount, cibil\_score, and various asset values.
   * High correlation (0.93) observed between loan\_amount and income\_annum, indicating a strong relationship.
   * Other significant correlations included luxury\_asset\_value with income\_annum (0.93) and loan\_amount (0.86).
2. **Outlier Detection**:
   * No outliers detected in income\_annum and loan\_amount using both box plots and the IQR method.

**Model Training and Evaluation**

1. **Manual Data Splitting**:
   * An 80/20 split was used to divide the data into training and testing sets.
   * Three models were trained: K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree.
2. **Model Performance (Manual Split)**:
   * **KNN**: Accuracy = 0.57, Precision = 0.67, Recall = 0.74, F1-score = 0.70, ROC-AUC = 0.55
   * **Logistic Regression**: Accuracy = 0.63
   * **Decision Tree**: Accuracy = 0.97
3. **Cross-Validation (KNN)**:
   * Cross-validation scores ranged from 0.55 to 0.58 with a mean score of 0.56.
   * Mean ROC-AUC score from cross-validation was 0.51.
4. **Feature Importance**:
   * Decision Tree model indicated that features such as income\_annum, loan\_amount, and cibil\_score were significant in predicting loan approval.

**Visualizations**

1. **Confusion Matrix**:
   * Provided a visual representation of the model’s performance on test data, highlighting true positives, true negatives, false positives, and false negatives.
2. **ROC Curve**:
   * Illustrated the trade-off between sensitivity and specificity for the KNN model.

**Summary**

* The Decision Tree model outperformed KNN and Logistic Regression with a significantly higher accuracy.
* Cross-validation provided a more robust evaluation of model performance, with consistent metrics across multiple folds.
* Key features influencing loan approval included income\_annum, loan\_amount, and cibil\_score.