**Title: Loan Approval Prediction Analysis   
&   
KNN Model**

1. **Introduction**   
   This project aims to analyze the Loan Approval Prediction Dataset from Kaggle and build a K-Nearest Neighbors (KNN) model to predict loan approval status. The analysis includes data cleaning, exploratory data analysis (EDA), outlier detection and treatment, and model training and evaluation.  
     
   **About the Dataset**

* The Loan Approval Prediction Dataset, sourced from Kaggle, contains information about loan applicants and their loan approval status. Key features include:
* Demographic information: Number of dependents, education level, self-employment status
* Financial information: Annual income, loan amount requested, loan term
* Credit information: CIBIL score
* Asset information: Residential, commercial, and luxury asset values, bank asset value
* Target variable: Loan status (Approved/Rejected)
* The dataset consists of 4,269 entries with 13 features, providing a comprehensive view of factors potentially influencing loan approval decisions. This dataset is suitable for binary classification tasks, where the goal is to predict whether a loan application will be approved or rejected based on the given features.

1. **Data Cleaning**

* Loaded the dataset into a Pandas DataFrame
* Checked for missing values: No missing values found
* Removed unnecessary spaces in column names and values
* Dropped the 'loan\_id' column as it was not needed for analysis

1. **Exploratory Data Analysis (EDA) Key findings:**

* The dataset contains more approved loans than rejected ones
* Strong positive correlation between income and loan amount
* High correlation between CIBIL score and loan approval
* Applicants with higher income tend to apply for larger loan amounts
* Applicants with higher incomes generally have higher bank balances
* No clear relationship between income and loan approval status
* CIBIL score is a crucial factor in loan approval, with higher scores more likely to be approved

1. **Outlier Detection and Treatment**

* Used Interquartile Range (IQR) method to detect outliers
* Detected outliers in 'residential\_assets\_value', 'commercial\_assets\_value', and 'bank\_asset\_value'
* Chose to cap outliers instead of removing them to preserve data
* Applied capping to all numerical columns using the IQR method

1. **Data Splitting and Model Training**

* Manually split the data into 80% training and 20% testing sets
* Performed feature scaling using StandardScaler
* Trained a K-Nearest Neighbors (KNN) model with n\_neighbors=16
* Implemented 5-fold cross-validation to assess model performance

1. **Model Evaluation Metrics on test set:  
     
     
   Detailed Model Performance Insights:**

* **Accuracy: 0.9016 -** The model correctly predicts loan approval status for 90.16% of cases.
* **Precision: 0.9330 -** When the model predicts loan approval, it's correct 93.30% of the time.
* **Recall: 0.9086** - The model correctly identifies 90.86% of all actual approved loans.
* **F1-score: 0.9206 -** Indicates a good balance between precision and recall.
* **ROC-AUC: 0.8993 -** Shows strong discriminative ability between approved and rejected loans.

**Confusion Matrix:**

* True Positives: 283
* False Positives: 35
* False Negatives: 49
* True Negatives: 487

**Cross-validation results:**

* Mean CV score: 0.9078
* Individual CV scores: [0.88726208, 0.89751098, 0.91508053, 0.91800878, 0.92093704]  
    
    
  **The consistent and high cross-validation scores indicate model stability and good generalization ability. The ROC Curve with an AUC of 0.90 further confirms the model's strong ability to distinguish between classes.**

1. **Conclusion**

* **Consistent results between manual split and cross-validation indicate model stability**
* **CIBIL score is identified as a crucial factor in loan approval decisions**
* **The model shows good balance between precision and recall, as evidenced by the high F1-score**
* **The K-Nearest Neighbors (KNN) model developed for loan approval prediction demonstrates strong performance and reliability:**
* **With an accuracy of 90.16%, the model shows excellent overall predictive capability.**
* **High precision (93.30%) and recall (90.86%) indicate the model's effectiveness in both approving worthy candidates and identifying potential defaults.**
* **An ROC-AUC score of 0.8993 suggests the model's strong ability to distinguish between approved and rejected loan applications.**
* **Cross-validation scores (mean 0.9078) demonstrate the model's stability across different data subsets, indicating good generalizability.**
* **The analysis highlighted CIBIL score, income, and loan amount as crucial factors in loan approval decisions.**
* **The model's performance suggests it could be a valuable tool in assisting loan approval decisions in real-world scenarios.**