**K-Nearest Neighbors (KNN) Classifier Exercise with Loan Prediction Dataset**

**Objective**

The goal of this exercise is to apply the **K-Nearest Neighbors (KNN)** Classifier to predict the **loan status** of an individual based on their personal and financial characteristics.

**1. Data Dictionary**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Feature Name** | **Description** | **Data Type** | | --- | --- | --- | | person\_age | Age of the individual. | Integer | | person\_gender | Gender of the individual. | Categorical | | person\_education | Educational qualification of the individual. | Categorical | | person\_income | Annual income of the individual. | Float | | person\_emp\_exp | Work experience in years. | Float | | person\_home\_ownership | Type of home ownership (e.g., RENT, OWN). | Categorical | | loan\_amnt | Loan amount requested by the individual. | Float | | loan\_intent | Purpose for which the loan is requested (e.g., PERSONAL, EDUCATION). | Categorical | | loan\_int\_rate | Interest rate for the loan. | Float | | loan\_percent\_income | Percentage of income allocated to loan repayment. | Float | | cb\_person\_cred\_hist\_length | Length of the individual's credit history. | Integer | | credit\_score | Credit score of the individual. | Integer | | previous\_loan\_defaults\_on\_file | Whether the individual has defaulted on previous loans (Yes/No). | Categorical | | loan\_status | Target variable indicating if the loan was approved (1) or denied (0). | Integer | |

**2. Problem Statement**

Using the provided dataset, the task is to build a **KNN Classifier** model that can accurately predict the loan approval status of individuals based on their characteristics.

**3. Data Pre-Processing**

**3.1 Data Cleaning and Preparation**

* **Missing Values:**
  + Identify and handle any missing values appropriately.
* **Encoding Categorical Variables:**
  + Convert all categorical variables (e.g., person\_gender, person\_home\_ownership) into numerical format for compatibility with the KNN algorithm.

**3.2 Feature Scaling**

* Normalize or standardize numerical variables to ensure all features contribute equally to the distance metric used by KNN.

**4. Exploratory Data Analysis (EDA)**

**4.1 Univariate Analysis**

Visualize individual variables to understand their distribution (e.g., histograms for numerical data, bar charts for categorical data). Make notes on any unique features or trends you observe.

### 4. Bivariate and Multivariate Analysis

Explore relationships between variables by visualizing pairs of variables or groups of variables (e.g., scatter plots, heatmaps). Document any significant patterns or correlations that emerge from your analysis.

**5. Model Building**

**5.1 Training and Testing Split**

* Divide the dataset into training and testing sets (e.g., 80% training, 20% testing).

**5.2 Model Training**

* Train the **KNN Classifier** using the training data, experimenting with different values of **K** to find the optimal number of neighbors.