Assignment 2: Loan Data Analysis (95%)

Objective

This assignment focuses on two key machine learning techniques: clustering analysis and classification using logistic regression. You will analyze the provided dataset (`credit_risk_dataset.csv`) and draw meaningful insights from your models.

Data Explanation

1. person_age: Age
2. person_income: Annual Income
3. personhomeownership: Home ownership
4. personemplength: Employment length (in years)
5. loan_intent: Loan intent
6. loan_grade: Loan grade
7. loan_amnt: Loan amount
8. loanintrate: Interest rate
9. loan_status: Loan status (0 is non default 1 is default)
10. loanpercentincome: Percent income
11. cbpersondefaultonfile: Historical default
12. cbpresoncredhistlength: Credit history length

Problem 1: Clustering Analysis

Task

Perform a clustering analysis on the dataset to identify groups of loan applicants based on numerical variables.

Steps

1. Load the dataset and preprocess it:
    - Handle missing values appropriately.
    - Standardize numerical variables.
2. Select relevant numerical columns for clustering (for example, personage, personincome, personemplength, cbpresoncredhistlength, etc.).
3. Use the **K-Means algorithm** to perform clustering.
    - Determine the optimal number of clusters using the **Elbow Method**.
    - Fit the K-Means model and assign cluster labels.

4. Interpret the clusters:
    - What patterns do you observe in the clusters?
    - How do different clusters compare in terms of loan characteristics (e.g., loan amount, income, loan status)?

Problem 2: Classification Using Logistic Regression

Task

Choose one meaningful categorical variable from the dataset and build a logistic regression model to classify **loan status** (default or non-default).

Steps

1. **Select a categorical variable** (e.g., `cb_person_default_on_file`, `loan_grade`, or `loan_intent`).
2. **Preprocess data:**
    - Convert categorical variables into numerical form (e.g., one-hot encoding or label encoding).
    - Handle missing values if applicable.
    - Split the data into training and testing sets.
3. **Train a logistic regression model** to predict `loan_status` (0 or 1).
4. **Evaluate model performance** using:
    - **Confusion matrix**
    - **Accuracy, Precision, Recall, and F1-score**
5. **Interpret results:**
    - What insights can you gain from the logistic regression model?
    - What features are most important for predicting loan defaults?
    - How reliable is the model for making predictions?

Group Survey (5%)

Each group (2 or 4 students) will complete a short survey providing feedback on the course. The survey should cover the following aspects:

Questions

1. **Course Rating (1-5 scale):** How would you rate this course overall? (1: Poor, 2: Fair, 3: Good, 4: Very Good, 5: Excellent)
2. **Difficulty Level (1-5 scale):** How difficult was this course? (1: Very Easy, 2: Easy, 3: Moderate, 4: Hard, 5: Very Hard)
3. **Pros:** What aspects of the course did you find beneficial?
4. **Cons:** What challenges did you face in this course?
5. **Suggestions:** How can this course be improved in future iterations?

Each group should submit one collective response. The feedback will be considered confidential and used to improve the course in the future.
