IDC4022C Module 6 Project Part 1 K-Nearest Neighbors Analysis on Loan Default Prediction

This assignment will help you understand how to prepare data for modeling, specify features, and evaluate a KNN classifier.

Instructions:
Using the provided dataset 'module6data.csv', perform a K-Nearest Neighbors analysis using the "Credit Score" and "Debt-to-Income Ratio" as features to predict loan defaults.

Your analysis will be completed using a new Jupyter Notebook on Google Colaboratory.

Verify the necessary libraries (pandas, scikit-learn) are installed by running `!pip list` in a cell. Install any missing libraries by running `!pip install library-name` in the first cell in your notebook as necessary.

Complete the following steps in separate cells of the Jupyter notebook, **one cell per step** (your completed submission should contain 8 cells).

1. Upload the dataset to Colab and create a Pandas dataframe
   - Use the copy path menu option in Colab to determine the file's location and use the pandas read_csv() function to load the data into a dataframe.
2. Preliminary Data Analysis:
   - Use Pandas dataframe info() and describe() methods to understand the dataset structure and statistics.
3. Select Features and Target Variable:
   - Extract 'Credit Score' and 'Debt-to-Income Ratio' from the dataframe as features.
   - Select 'Loan Default' as the target variable.
4. Split the Dataset:
   - Use train_test_split() to divide the data into training and testing sets.
   - Use a test size of 20% and a random state of 42 for reproducibility.
5. Standardize the Features:
   - Use StandardScaler to normalize the feature data since KNN is distance-based.
6. Initialize and Train the KNN Classifier:
   - Initialize the KNN classifier with 5 neighbors.
   - Train the classifier on the scaled training data.
7. Make Predictions:
   - Use the trained classifier to predict the target values for the test set.
8. Evaluate the Model:
   - Generate a classification report and confusion matrix to evaluate the model performance.
   - Print out the classification report and confusion matrix.

Submitting Your Work:

- Be sure your work contains a 4-line ID header containing the file name, date, and one-line summary.
- Ensure your script runs without errors and produces expected output.

- Download your notebook as a .ipynb (Jupyter notebook) file.
- Create a Word document containing a report with the following content:
  - A brief description of the KNN algorithm.
  - Your observations from the classification report and confusion matrix.
  - Any insights or patterns noticed in the prediction results.
- Commit your .ipynb and report documents to the GitHub classroom repository.

Evaluation Criteria:

- ID header present
- Correct implementation of the KNN algorithm.
- Code readability and comments.

Demonstration of understanding of the model evaluation via report document.