Project 3 – German Credit Dataset

Context: The original dataset contains 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. It is difficult to understand the original dataset due to its complicated system of categories and symbols. Thus, several columns are simply ignored, because they are not important, or their descriptions are obscure. The selected attributes are:

- Age (numeric)
- Sex (text: male, female)
- Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- Housing (text: own, rent, or free)
- Saving accounts (text - little, moderate, quite rich, rich)
- Checking account (numeric, in DM - Deutsch Mark)
- Credit amount (numeric, in DM)
- Duration (numeric, in month)
- Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
- Risk (Target value - Good or Bad Risk)

Dataset Source: Click here

Task: Goal of this project is to cluster the customers and classify whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants

Implementation:

- Perform EDA and any data cleaning if necessary.
- Perform one hot encoding for categorical variables
- Visualize the histograms of numerical features. Do you observe skewness in the data? If yes apply the log transformation. Check the histograms again to see if data has been normalized.
- Apply Feature Scaling
- Choose only the numerical features for clustering
- Apply elbow method to find best number of clusters. Plot the graph.
- Choose optimum number of clusters and visualize it using PCA
- Implement KFOLD CV and use any classifier of your choosing and report the evaluation metrics

Submission Instructions: Please just submit one jupyter notebook containing all the code and make use of markdown cells to include the comments, answers, reasoning, analysis, etc.

Note: Name of your file should be your “Project3-id_Firstname_Lastname.ipynb”