Data Analysis and Data Mining Assignment

Objective:
The purpose of this assignment is to perform data collection, cleaning, exploratory data analysis, hypothesis testing, and classification modeling on a dataset sourced from the internet.

Task instructions:

1. Dataset Collection and Cleaning
- Search the internet to identify and download a relevant dataset.
- Perform data cleaning to prepare the dataset for analysis. Document the steps taken in the data cleaning process.

2. Univariate and Bivariate Analysis
- Conduct a univariate analysis on the dataset to understand the distribution of individual variables.
- Perform bivariate analysis to examine relationships between two variables.

3. Hypothesis Formulation and Testing
- Formulate a hypothesis related to the dataset.
- Test this hypothesis using appropriate statistical methods and interpret the results.

4. Classification Modeling
- Build two classification models based on the cleaned dataset using machine learning algorithms in Python.
- Compare the performance of the two models and discuss the results.

5. Submission and presentation requirements

Submit the following items:
- Python source code used for data cleaning, analysis, hypothesis testing, and modeling.
- The dataset used in the analysis.
- A PDF report containing:
  - Source and description of the dataset.
  - Detailed data cleaning steps.
  - Summary of univariate and bivariate analyses.
  - Hypothesis statement and results of the hypothesis testing.
  - Description of the machine learning algorithms used for classification.
  - Comparison of the classification models' performance.
  - Forecast or prediction based on the model outcomes.

Presentation: The completed assignment must be presented in person, summarizing key findings, methodology, and conclusions.

Deadline for assignment submission in MS Teams: 1st December, 2024

Presentation dates: 4th and 11th of December, 2024