Exploratory Data Analysis (EDA) and Feature Selection

Objective

The primary objective of this assignment is to provide you with hands-on experience in performing exploratory data analysis (EDA) on real-world datasets and implementing feature selection techniques using filtering and statistical methods.

Instructions

Select your dataset

Each person will pick 1 dataset to process from the following list:

1. Breast Cancer Wisconsin (Diagnostic) Dataset (UCI Machine Learning Repository) - link
2. German Credit Dataset (UCI Machine Learning Repository) - link
3. Census Income Dataset (UCI Machine Learning Repository) - link
4. Heart Disease UCI (UCI Machine Learning Repository) - link
5. Mushroom Classification Dataset (UCI Machine Learning Repository) - link
6. Statlog (Heart) Dataset (UCI Machine Learning Repository) - link
7. Hepatitis Dataset (UCI Machine Learning Repository) - link
8. Student Performance Dataset (UCI Machine Learning Repository) - link

Reserve your dataset here - link

If you want to use another dataset, you can propose it in the group chat. The use of other datasets is pending my approval.

Exploratory Data Analysis (EDA)

Using Google Colab or Jupyter Notebook, write Python code to perform EDA on the selected dataset. The EDA process should include the following steps:

1. **Data Loading:** Load the dataset into your notebook.

2. **Data Inspection:** Display the first few rows of the dataset, check for missing values, and provide summary statistics.
3. **Data Visualization:** Create relevant plots (e.g., histograms, box plots, scatter plots, correlation matrices) to visualize the data distribution and relationships between features.
4. **Data Preprocessing:** If necessary, perform data preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.

You can use any python libraries you want. Some good ones include:

- Data structure: pandas,
- Math: numpy, scipy
- Visualizations: matplotlib, sns
- ML Libraries: sklearn

Feature Selection (Filtering / Statistics)

After completing the EDA, implement feature selection techniques using filtering and statistical methods. The goal is to identify and select the most relevant features for machine learning models. Include the following steps:

- Feature Ranking: Use statistical methods (e.g., correlation coefficients, mutual information) to rank the features based on their importance.
- Feature Selection: Select a subset of the most important features based on the ranking obtained in the previous step.

Which features will you propose to use in the machine learning model, just based on statistical analysis? Why use these features?

Documentation and Analysis

Provide a written analysis of your findings. Discuss why certain features were selected or excluded based on the EDA and feature selection results. Explain how your choices may impact the performance of machine learning models in future tasks.

Submission Guidelines

- Submit your assignment as a Jupyter Notebook file or Google Colab Notebook link in EMAS2.
- Include clear and organized code, comments, and explanations.
- Clearly state which dataset you chose.
- Submit your analysis as part of the notebook file, there's no need to create a separate document.

Grading Criteria

The assignment will be graded based on the following criteria:

- Problem and dataset description (10%)
- Completeness and correctness of EDA process (40%)
- Implementation of feature selection techniques (30%)
- Documentation and clarity of code (10%)
- Quality of written analysis (10%)

Note

If you need any assistance, feel free to ask me. You can use any resource you need to work on this assignment (including GPT), but please make sure you understand what you are putting into the notebook. ***Be honest with yourself, your future self will be glad you did.***