Lab5 Assignment: High School Students Data Mining

Overview

You have been tasked by a company to analyze and process data related to High School students (https://archive.ics.uci.edu/dataset/320/student+performance). Your role involves extracting key statistical insights from the dataset and visualizing the data to uncover patterns and trends. This project will test your ability to work with data mining, apply statistical methods, and create informative visualizations.

Objectives

- Extract and calculate important statistical measures from the provided dataset.
- Practice constructing data mining models with Python coding.
- Create visual representations of the data to highlight key findings.

Statistic Task

1. Load student-mat.csv and student-por.csv from UCI.
    a. How many rows and columns does each file contain?
    b. List categorical vs numeric features.
    c. Use a **Markdown** cell to explain your results.
2. For two CSVs respectively, identify the three features (attributes) that have the greatest impact on students' grade G1, G2, and G3. You may use heatmap to visualize the correlation matrix from Provided dataset. Use a **Markdown** cell to explain your results.
3. For two CSVs respectively, select any four attributes and appropriate statistical results, and use **bar** plots to present them. Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.
4. For two CSVs respectively, select any four attributes and appropriate statistical results, and use **line** plots to present them. Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.

Data Mining Tasks

You are required to analyze the data using different data mining techniques:

1. **Decision Tree for** student-mat.csv:
    - Please use the decision tree model to extract a Rule related to the student's grade classification (You can define your mining goals) from the provided dataset. (You may just use a few features not all).
    - You need to visualize the extracted rules and set the maximum depth to 4.
    - Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.

2. **Naive Bayes for** student-por.csv:
    - Create a binary target variable passed (1 if G3 $\geq$ 12, 0 otherwise). Show the class distribution (% passed vs failed). Explain your results a **Markdown** cell.
    - Given the features *studytime*, *absences*, and *G1* (first period grade), can we predict whether a student will *pass* or *fail (same rules 1 if G1 $\geqslant$ 12, 0 otherwise.)* the course using a Naive Bayes classifier? Explain your results a **Markdown** cell.

You need to visualize your results with scatter (True vs prediction) and Confusion Matrix.

- Using the features *studytime, failures, schoolsup, famsup*, and *activities*, can we predict whether a student will achieve a high, medium, or low final grade G3 (You can define these level range) using a Multinomial Naive Bayes classifier? Explain your results a **Markdown** cell. You need to visualize your results with Confusion Matrix.

3. Please add necessary comments for each key function in your code.

Submission

- Save your visualization results and upload them to GitHub.
- Code: Include the entire code used for the analysis and upload them to GitHub.
- Submit your GitHub link to Canvas.