Final Project:

Overview

The final lab project requires students to select one dataset from real-world to conduct comprehensive data analysis using data mining techniques. Each team, consisting of multiple members, will be responsible for customizing a specific number of data mining-related questions based on the dataset chosen. The number of **unique** questions to be customized will be equal to the number of group members x 3. These questions will encompass a variety of statistical and data mining methods, with up to three questions allowed to be of statistical nature, while the remaining questions will require the application of analytical models.

Key Components of the Final Lab Project

1. **Selection of Dataset**: Students will choose one dataset from a provided list, ensuring it aligns with their interests and objectives for analysis.

2. **Customized Questions**: Each team will formulate a set of data mining-related questions based on the dataset chosen. These questions should cover various aspects of data analysis, including descriptive statistics, predictive modeling, and pattern recognition.

    A. Statistical Questions: Up to **three questions (**for each group**)** can focus on statistical analysis, such as max/min/mean value testing or correlation analysis.

    B. Data Mining Techniques: The remaining questions will require the application of data mining techniques, such as classification, clustering, or Association rules.

3. **Visualization**: All problems in the final lab project must include visualization of the results with detailed decoration such as font size, color, text, and x-y labels. Visual representations, such as charts, or graphs, should be utilized to enhance understanding and interpretation of the data analysis outcomes.

4. **Final Report**: Teams will prepare a final report documenting their analysis process for each question. The report should include problem definitions, descriptions of analysis methods employed, visualizations of results, and highlights/takeaways from the analysis.

5. **Speech Presentation**: Each team member will present their assigned problem and research results during a speech presentation. Additional points will be awarded for effective communication and presentation skills.

6. **Submission of Report and Code**: Teams will upload the final report along with the original code used for analysis.

- Gain Report: Compile your findings, analysis, and visualizations into a comprehensive report. The report should be submitted as a PDF.
- Code: Include the entire code used for the analysis and upload them to GitHub.
- Submit your GitHub link to Canvas.

Note:

Data mining involves uncovering *hidden patterns, trends, and relationships* within datasets that are not immediately obvious. Your project should demonstrate an ability to explore data beyond simple descriptive statistics.

To receive full credit, your questions must reflect **depth and analytical insight**.
Overly simplistic or surface-level questions (e.g., "What is the average value of X?") will result in a **50% deduction**.

A simple example using the **Iris dataset** is provided for reference, but you are expected to design **original, creative, and meaningful data mining questions** relevant to your chosen dataset.

Ideas for Meaningful Data Mining Questions

Here are some question types that go beyond surface-level analysis:

1. Correlation and Pattern Discovery

- *Example:* "Which combinations of student habits (study time, internet usage, absences) most strongly correlate with final grades?"
- *Idea:* Look for **hidden interactions** between multiple variables instead of single-variable summaries.

2. Classification and Prediction

- *Example:* "Can we predict whether a student will pass or fail based on social activities and family support?"
- *Idea:* Use **Decision Tree** or **Naive Bayes** models to classify outcomes and analyze feature importance.

3. Clustering and Segmentation

- *Example:* "Can we segment customers/students/users into meaningful clusters based on behavior or performance metrics?"
- *Idea:* Apply **K-Means** to find groups and then interpret what characterizes each cluster.

4. Association Rules

- *Example:* "Which products or behaviors commonly occur together?"
  (e.g., "If a student spends more than 2 hours online and goes out frequently, are they likely to have lower grades?")
- *Idea:* Use **Apriori or FP-Growth** to find frequent patterns or co-occurrences.

5. Anomaly or Outlier Detection

- *Example:* "Which data points represent unusual or extreme behaviors, and what might explain them?"
- *Idea:* Use clustering distances or z-scores to detect anomalies.

6. Temporal or Sequential Patterns

- *Example:* "How do student performance or user activities change over time?"
- *Idea:* Compare patterns across semesters, months, or sessions to find **temporal trends**.