Bayesian Network Group Activity

Overall Learning Objectives:

- Practice data exploration techniques relevant to probabilistic modeling.
- Construct a plausible Bayesian network structure representing conditional dependencies for a given scenario.
- Identify and articulate conditional independencies implied by the network structure.
- Estimate joint and conditional probability distributions from data.
- Perform probabilistic inference (e.g., calculating the probability of an event given evidence) using the constructed network.
- Interpret and communicate the results of the Bayesian network analysis.

General Guidance for All Groups:

- Tools: You can use Python libraries like pgmpy, bnlearn, or even spreadsheet software for calculations and visualization, depending on your comfort level. Whiteboards are great for sketching initial network structures.
- Assumptions: Clearly state any assumptions you make when constructing your network (e.g., assumed causal directions not explicitly evident in the data).
- Data Limitations: Acknowledge that the provided datasets are small and simplified. Real-world applications would require larger, more nuanced data.
- Collaboration: Discuss interpretations and justifications for your network structure and findings within your group.

Group 1
Data Center Outage

(Technical Systems)

- **Context:** Investigate the potential causes of server crashes within a data center environment by analyzing relationships between system metrics.
- **Objective:** Build a Bayesian network to model the dependencies between server metrics and crashes. Use the model to determine the most likely cause(s) of a server crash given observations of other metrics (e.g., high CPU load, network spikes).
- **Variables:**
    - `CPU_Load` (Categorical: Low, Medium, High)
    - `Power_Supply` (Categorical: Stable, Fluctuating, Failed)
    - `Network_Traffic` (Categorical: Low, Medium, High)
    - `Disk_Usage` (Categorical: Normal, High, Critical)
    - `Malware_Detected` (Binary: True, False)
    - `Server_Crash` (Binary: True, False)
- **Tasks:**

1. **Explore Data:** Analyze the provided dataset (datacenter_logs.csv) to understand the distribution of each variable and identify potential correlations. Visualize relationships.
2. **Construct Network:** Propose a directed acyclic graph (DAG) representing the causal relationships (or strong correlations) between the variables. Justify your chosen structure (e.g., Does high CPU load cause high disk usage, or vice versa? Does malware detection influence network traffic?).
3. **Estimate Probabilities:** Calculate the Conditional Probability Tables (CPTs) for each node in your network based on the data.
4. **Perform Inference:** Use your network and CPTs to answer questions like:
    - What is the probability of a `Server_Crash` if `CPU_Load` is High and `Power_Supply` is Fluctuating?

        - If a `Server_Crash` occurred, what is the most likely state of `Malware_Detected`?
        - How does observing `Network_Traffic` = High change the probability of `Disk_Usage` being Critical?
- **Present Findings:** Summarize your network structure, key probability findings, and inference results.
- **Dataset:** datacenter_logs.csv (Sample Data)