Checkpoint 2: Data Collection, Feature Selection, and Model Plan

Group 7: Ayala Wang, Shashank Bhagwani, Shiyuan Wang, Nandhan Natarajan

Link to GitHub Repo: GitHub Repository

1. Update on Data Collection

- **Status**:
  - We collected **120 packet traces** using the NetUnicorn platform.
  - Each trace includes:
    - **Download/upload speed**
    - **Latency**
    - **Jitter**
    - **Packet loss**
  - Data was collected from three key campus locations: **library**, **lecture halls**, and **outdoor plazas**.
  - Traces were gathered during **peak** and **off-peak** hours for variability.
- **Challenges**:
  - Minor disruptions occurred during outdoor data collection due to power and Wi-Fi instability but were quickly resolved.
  - Sequential data collection extended the process slightly due to limited device availability.
- **Scaling Plan**:
  - **No further scaling is planned** as the current dataset is sufficient for our proof-of-concept model.

2. Planned Features

- **Extracted Metrics**:
  - **Download/upload speed**
  - **Latency**
  - **Jitter**
  - **Packet loss**
- **Justification**:
  - These metrics are directly tied to evaluating network performance and align with the project goal of assessing UCSB Wi-Fi quality.

3. Model Plan

- **Model Type**:
  - A **Random Forest Classifier** will categorize network performance into three levels: **Good**, **Moderate**, and **Poor**.
- **High-Level Explanation**:
  - Random Forest is ideal for handling small datasets and mixed feature types (e.g., continuous and categorical).
  - It is robust, interpretable, and provides feature importance metrics to prioritize key network issues.
- **Scikit-learn Implementation**:
  - Random Forest Classifier Documentation

---

4. Next Steps

- **Feature Engineering**:
  - Extract the listed metrics from the packet traces.
  - Preprocess the data for model input (e.g., normalize values as needed).
- **Model Training**:
  - Train the Random Forest Classifier on the labeled dataset.
  - Evaluate the model using metrics such as **accuracy** and **F1-score**.
- **Proof of Concept**:
  - Validate the approach by categorizing network quality across sampled locations.
