Project Description:

This project was started with 72 raw (.edf) files. The project involved extracting and analyzing features from 72 .edf files, each containing EEG data for Rest and Load conditions for 36 subjects.

Data Preprocessing

- **Input Files**: loaded 72 .edf files using the pyedflib library.
- **Iteration**: A for loop iterated over all .edf files in the specified directory.
- **Combining Data**:
  - Data from each file was converted into a pandas DataFrame.
  - Each DataFrame was appended to a list and finally concatenated into a single DataFrame (final_df) containing all data from 72 files.
- **Columns**: The dataset includes EEG channel data, an "ECG ECG" column, and a "Subject" column for identifying individual files.

Data manipulation:

After creating the dataset, I analyzed dataset with some basic EDA, then I dropped duplicates. Following this I selected columns containing EEG and converted it into numpay array. The next task was to assign 4 different labels (A, B, C, D). I used KMeans clustering technique here, and then mapped the clusters with these variables. Now the final form of our dataset is "new_df" and final shape of our dataset is

[4267195 rows x 22 columns]

Feature Extraction

We extracted four key features for each microstate (A, B, C, D) from the EEG data:

1. **Global Explained Variance (GEV)**:
   - Quantified how much of the overall variance in the EEG signal each microstate explained.
   - Calculated by dividing the variance within a microstate by the total variance of the dataset.
2. **Time Coverage (timecov)**:
   - Calculated the proportion of time spent in each microstate.
   - Provided insight into the dominance of each microstate in Rest and Load conditions.
3. **Mean Duration (meandurs)**:
   - Measured the average temporal duration of segments spent in each microstate.
   - Gave information about the stability of each microstate before transitioning to another.
4. **Occurrence per Second (occurrence)**:
   - Counted the number of transitions into each microstate per second.
   - Represented the switching behavior between states.

Classification Target

- The target labels for classification (Rest and Load) were assigned based on the subject and file name patterns (_1 for Rest, _2 for Load).
- Each subject's features were computed separately for Rest and Load conditions.