Final Project

1. Project Kick-off

- **Group Formation**: Students will form groups of 4-5 members.
- **Dataset Selection and Problem Statement Formulation**:
    - Each group will select a dataset. This dataset can be chosen from public data repositories like Kaggle, UCI Machine Learning Repository, or any other credible source of datasets.
    - After selecting the dataset, the group must formulate a clear problem statement. This statement should outline what the group intends to solve or discover through their project. It could involve predicting an outcome (regression), classifying data points (classification), grouping data points (clustering), or any other machine learning task relevant to the dataset.

2. Data Preprocessing

- **Data Cleaning**: Identify and handle missing values, remove duplicates, and address outliers.
- **Feature Engineering**: Create new features from existing data to improve model performance.
- **Data Transformation**: Normalize or standardize data if necessary.

3. Exploratory Data Analysis (EDA)

- **Visual Analysis**: Use plots and charts to understand data distributions, relationships between features, and other patterns.
- **Statistical Analysis**: Apply statistical methods to gather insights and understand the dataset's characteristics.

4. Model Selection and Training

- **Model Selection**: Based on the problem statement, choose appropriate machine learning models. Consider using different models for comparison.
- **Training**: Train the models using the training subset of the dataset. Implement cross-validation to optimize model parameters and prevent overfitting.

5. Evaluation

- **Model Evaluation**: To evaluate the models, use appropriate metrics (e.g., accuracy, precision, recall, F1 score for classification tasks; MSE, RMSE for regression tasks).

- **Model Comparison**: Compare the performance of different models based on the evaluation metrics.

6. Reporting and Presentation

- **Detailed Report**: Prepare a detailed project report that includes (8-10 pages):
    - Introduction to the project and problem statement.
    - Provide a description of the dataset and any assumptions made.
    - Steps were taken in data preprocessing and EDA, with justification.
    - Overview of selected models, rationale for selection, and configuration details.
    - Evaluation results and interpretation.
    - Conclusions and possible improvements or future work.
- **Presentation**: Prepare a 10-minute presentation summarizing the project. Include key insights and findings and demonstrate a clear understanding of the project's machine-learning pipeline. Visual aids (e.g., slides and charts) should enhance the presentation.