CO544 Machine Learning and Data Mining
Take Home Assignment - E/20

Objective:

The purpose of this assignment is to give you hands-on experience with a Machine Learning pipeline from dataset selection to model evaluation and improvement. You will select a real-world dataset, preprocess it, apply feature engineering, build and evaluate models, and explore ensemble methods.

Select a dataset from the UCI Machine Learning Repository. The dataset must have at least 300 records, and more than 10 features (before preprocessing/ feature engineering).

Tasks:

You must complete the following steps using Python (preferably Jupyter/ Colab Notebook) and necessary libraries.

Your notebook should be well-structured, clearly divided into sections, and contain both code and explanations. Use headings, comments, and visualizations where appropriate.

Expected Structure:

1. Introduction and Dataset Description
2. Data Preprocessing
3. Exploratory Data Analysis: Use basic statistics, distributions, correlations to obtain insights about data
4. Feature Engineering - Feature Selection and Feature Extraction
5. Modeling: Train at least two classifiers, Show performance results with appropriate evaluation metrics, compare models
6. Model Improvements
7. Ensemble Learning: Implement at least one ensemble method (e.g., bagging, Boosting) and compare performance with individual models

Submission Guidelines:

- Submission 1: Python Notebook
    - Ensure the notebook is clean, structured, commented, and runs without errors
- Submission 2: A 2-5 pages report (excluding references, if any)
    - A concise summary of your work is expected. It should explain the reasoning behind your choices, not just the results. Include the following.
        - Introduction to the selected dataset (source, domain, target variable, goal of your analysis etc.)
        - Discussion of the results of the above and justification for your choices at each level (preprocessing, modeling, evaluation, improvements, ensemble methods)
        - Discussion and Conclusion: What worked best, what didn't, limitations of your approaches, possible further improvements.