Homework – Week 4: Decision Trees and Boosting

Course: Introduction to Machine Learning
Andres Mendez-Vazquez

Objective

Apply decision tree models to a real-world dataset of your choice. You will explore tree-based classification or regression, and optionally experiment with AdaBoost or XGBoost.

Instructions

1. Go to **https://www.kaggle.com/datasets** and choose a dataset that interests you.
   - It can be about sports, health, finance, education, etc.
   - The dataset must have at least one target column and multiple features.
2. Load the dataset into a Jupyter Notebook and perform basic preprocessing:
   - Clean missing or invalid data
   - Encode categorical variables if needed
   - Split the data into training and test sets
3. Choose one of the following tasks:
   - **Classification** using a Decision Tree (e.g., predict a label or category)
   - **Regression** using a Regression Tree (e.g., predict a number)
4. Train a Decision Tree model and evaluate its performance. Include:
   - Accuracy (for classification) or MSE (for regression)
   - A plot of the tree or feature importances
5. (Optional Bonus) Train an AdaBoost or XGBoost model and compare the results with the decision tree.
6. Submit:
   - A PDF or notebook (.ipynb) with your code, outputs, and short explanations.

- Include a short summary of what you learned.
- Link to the dataset used (URL on Kaggle).