Chronic Kidney Disease Prediction Using Boosting Algorithms

Background:

Boosting is a machine learning technique that builds a strong classifier by combining multiple weaker models in a sequential manner. Each model learns from the errors of the previous one, gradually improving overall performance. This approach reduces bias and variance and is particularly useful in complex classification tasks. Common boosting algorithms include AdaBoost, XGBoost, and Histogram-Based Gradient Boosting, which have been widely adopted in clinical prediction problems.

Objective:

The goal of this project was to develop predictive models to identify chronic kidney disease (CKD) using a dataset with a wide range of demographic, clinical, and lifestyle variables. The project compared the performance of three boosting methods: XGBoost, AdaBoost, and HistGradientBoosting.

Data Overview:

The dataset included 1,659 patient records and 54 variables, such as:

- Demographics: age, gender, socioeconomic status, education
- Clinical indicators: blood pressure, GFR, serum creatinine, HbA1c
- Lifestyle: diet quality, physical activity, alcohol consumption
- Medication use, health literacy, and medical history

The target variable was a diagnosis label indicating whether a patient had CKD (1) or not (0).

Methodology:

- Data preprocessing included imputing missing values, scaling numeric features, and encoding categorical variables using a ColumnTransformer.
- Separate pipelines were built for each model: XGBoost, AdaBoost, and HistGradientBoosting.
- Hyperparameter tuning was performed using GridSearchCV with ROC AUC as the scoring metric.
- The best models were evaluated using ROC and precision-recall curves, confusion matrices, and feature importance analysis.

Results:

- HistGradientBoosting achieved the highest AUC of 0.853.
- XGBoost performed similarly with an AUC of 0.851.
- AdaBoost had a lower AUC of 0.791.
- The most important features across models included GFR, serum creatinine, BUN levels, HbA1c, and medication adherence.

Conclusion:

Both HistGradientBoosting and XGBoost demonstrated strong performance in predicting CKD. Boosting algorithms proved effective for capturing patterns in complex clinical data and can support early identification of patients at risk for kidney disease.