Student Name: Carlos Nakagomi

Instructor: Quynh Nguyen

Langara College

Cyber Security Case Study

Explaining the Baseline and Process of Optimization

The process began with data loading, where the dataset was imported using pandas.read\_excel(). An initial inspection using .head() provided an overview of the dataset's structure, offering insight into feature names, data types, and potential inconsistencies.

To improve data clarity, the column names were translated to English using a dictionary mapping strategy. The .rename() method was applied to ensure column names were clear and interpretable. Additionally, the "Duongvao" column, identified as irrelevant to the analysis, was dropped using .drop().

A detailed univariate analysis followed, aimed at exploring the distribution of individual features. Visualizations such as histograms, boxplots, and violin plots were generated using matplotlib and seaborn to identify data trends, skewed distributions, and outliers. A Q-Q plot and the Shapiro-Wilk test were also incorporated to assess the normality of numerical features, providing both visual and statistical insights into feature behavior.

The notebook then moved on to bivariate analysis, where scatterplots were created to explore relationships between numerical features and the target variables. Bar plots were employed to evaluate the impact of categorical variables on the target outcomes. Multivariate analysis was conducted using a correlation matrix visualized with a heatmap, which highlighted strong linear relationships between key features. A pairplot further visualized the interactions between features with notable correlations.

For data cleaning, the notebook included steps to identify and address duplicated rows using .duplicated(). Missing values were examined using .isnull().sum() and subsequently addressed using appropriate imputation strategies. Numerical features were imputed with their median values, while categorical features were filled with their most frequent values (mode). Outliers were managed through the IQR method, with extreme values capped at calculated lower and upper limits.

Feature selection was performed to identify the most relevant predictors for the AKI Outcome Prediction and Dialysis Treatment Prediction tasks. The selected features for AKI Outcome Prediction included:

* Thomay
* Glasgow
* Mach
* HATB
* SOFA
* APACHEII
* pH
* HCO3
* Lactac0
* Urea\_Level
* Creatinine\_Level
* PCT0
* BiLIrubin

For the Dialysis Treatment Prediction model, the selected features included the same features listed above, with the addition of:

* Albumin\_Level
* BC0

By refining the dataset to include only these selected features, the model's efficiency improved by reducing noise and focusing on the most impactful variables.

The dataset was then split into training and testing sets using train\_test\_split() to facilitate fair model evaluation. Random Forest models were constructed for both AKI Outcome Prediction and Dialysis Treatment Prediction using RandomForestClassifier. These initial models, trained with default hyperparameters, served as baseline models to establish performance benchmarks.

To improve model performance, hyperparameter tuning was conducted using GridSearchCV, where parameters such as the number of estimators, maximum depth, and minimum samples per split were optimized. This step resulted in improved evaluation metrics, particularly in precision and recall for predicting rare events.

Using shap.TreeExplainer(), SHAP values were calculated for both Random Forest models. SHAP summary plots revealed that Thomay was the most influential feature in predicting AKI outcomes, while Glasgow played a secondary yet meaningful role. SHAP force plots provided deeper insights into individual misclassified observations, revealing how key features influenced prediction outcomes.

By combining data cleaning, feature selection, hyperparameter tuning, and model interpretability through SHAP analysis, the optimized models achieved improved predictive performance and delivered meaningful insights into the factors driving model decisions.