HEART ATTACK PREDICTION MODEL

Data preprocessing marked the inception of our data science project, initiated by utilizing the pandas package to read the 'heart.csv' dataset and inspecting its structure through its tail. To enhance clarity, we opted to rename columns for improved readability. Addressing data quality, we dealt with missing values and outliers, specifically substituting 0 values in the 'thall' column. Additionally, we increased interpretability by mapping category values to more descriptive labels in columns like 'Chest Pain type,' 'slp,' 'thall,' 'resting electrocardiographic findings,' and 'Gender_of_the_person.' Our exploration extended to visualizations, including a pie chart representing gender distribution and a boxplot investigating the link between chest pain types and resting blood pressure across genders.

Subsequently, we transitioned into data visualization, generating a histogram to illustrate the distribution of chest discomfort types and a correlation heatmap to explore links between different variables. These visualizations played a pivotal role in acquiring insights about the dataset's features.

In the machine learning models phase, we employed various classifiers such as Logistic Regression, Random Forest, KNeighbors, Decision Tree, Ada Boost, Bagging, Gradient Boosting, and XGBoost for classification. Utilizing 5-fold cross-validation, we evaluated model performance, measuring accuracy on both training and test sets.

Feature engineering emerged as a critical aspect of our project, involving the definition of functions based on specific conditions. These functions were instrumental in generating new columns, including 'new_heart_rate,' 'new_blood_pres,' 'new_age_grp,' 'new_chest_pain,' 'heart_chest_pain,' and 'thall_new.' This process significantly enhanced the dataset by integrating meaningful derived characteristics.

The implementation of the Random Forest Classifier marked a watershed moment in our project. We divided the data into training and testing sets, trained the classifier on the training data, and assessed its accuracy on the testing data. Further enhancing model robustness, GridSearchCV was employed to discover the best hyperparameters for the Random Forest model.

In the concluding stages of the project, during model evaluation, we scrutinized the Random Forest model's performance on the test set. This assessment was measured using accuracy metrics and visually represented using a confusion matrix and heatmap, offering an intuitive depiction of the model's predictive capabilities. Overall, these processes contributed to a comprehensive and systematic data science initiative focusing on heart disease classification.

As a notable conclusion, the model's accuracy showed a significant improvement throughout the project lifecycle. Starting at 80%, it was enhanced to 84% after feature engineering, and fine-tuning the parameters further elevated the accuracy to an impressive 88%. This progression underscores the effectiveness of the applied methodologies in refining the predictive capabilities of the model.