Cardiovascular Risk Prediction

Introduction:
This is a classification project where the objective is to predict whether a patient has a 10-year risk of future coronary heart disease (CHD) or not. The dependent variable, also known as the output variable, is named "TenYearCHD," and it takes two values: 0 (indicating no risk of CHD) and 1 (indicating a risk of CHD within the next 10 years). The goal is to build a model that can accurately classify patients into these two categories based on their health data and risk factors.

Data Preprocessing:
The dataset contains three types of features: demographic (sex, age), behavioral attributes (smoking status, average cigarettes per day), and medical history (blood pressure meds, stroke, hypertension, diabetes). The target variable is the 10-year risk of coronary heart disease (CHD), represented as 1 for "Yes" and 0 for "No." The goal is to build a classification model to predict this CHD risk based on the provided features.

Missing Data Handling:
The dataset contains 3,390 rows and 17 columns. There are no duplicate values in the dataset. However, several columns have missing values. The dataset has less than 10% of missing data, which is not considered a major issue and can be easily handled. To address the missing values, we have replaced them with the median values of the respective columns. Using the median helps preserve the data's original skewness and is less sensitive to extreme values. Additionally, the data type format of some variables was incorrect, and it has been corrected during the data preprocessing step.

Handling Skewed Data:
Then we checked the distribution of each feature by creating histogram plots. We saw 'cigsPerDay' is a highly skewed and did not follow a normal pattern.

To investigate further, we employed the Interquartile Range (IQR) method to identify outliers within the 'cigsPerDay' data and we replaced the identified outliers with the median value of the 'cigsPerDay' data. The IQR method is effective in detecting outliers, especially in skewed distributions, as it is less sensitive to extreme values compared to other methods like z-scores or standard deviations. Replacing outliers with the median helps to maintain the central tendency of the data while reducing the impact of extreme values on the distribution.

Encoding Categorical Variables:
Then we had two variables where encoding was required,sex and is_smoking variable. Both had binary categories either male, female or yes, no, so we simply used binary encoding instead of using one-hot and label encoding. As Binary encoding can effectively represent each category as a binary digit.

Feature Manipulation and Selection:
Then we went for feature manipulation and selection to identify the most important features for our analysis. During this process, we observed that the 'smoking' feature and 'cigsPerDay' feature were highly linearly dependent. Since 'smoking' could be easily derived from 'cigsPerDay', we decided to remove the 'smoking' feature to avoid redundancy. Additionally, we noticed a high correlation between the 'sysBP' (systolic blood pressure) and 'diaBP' (diastolic blood pressure) features. To address this correlation and reduce multicollinearity among independent variables, we combined the two features to create a new column named 'hypertension'. Moreover, to ensure the efficiency and accuracy of our analysis, we

removed any independent variables that were either redundant, not useful for our predictive model, or showing high correlation with other features.

Addressing Data Imbalance:
Then we checked for data imbalance in the dataset, and it was evident that the data was highly skewed, with only around 15% of patients diagnosed with coronary heart disease. To address this imbalance, we used the SMOTE (Synthetic Minority Oversampling Technique) to oversample the training dataset. SMOTE generates synthetic samples of the minority class (patients with heart disease) to balance the class distribution. This approach ensures that the model is trained equally on all outcomes and prevents bias towards the majority class, leading to a more reliable and accurate prediction.

Data Splitting and Scaling:
Then we split the data into train and test datasets in a ratio of 80:20, where 80% of the data is used for training and the remaining 20% is used for testing and evaluating the model's performance. We used the StandardScaler to bring all variables onto the same scale. Scaling the data ensures that each feature contributes equally to the model during training, preventing features with larger ranges from dominating the learning process.

Model Selection and Evaluation:
Now our data is ready for model implementation. We used three models logistic regression as our base model, followed by decision tree and xg boost. Three models were used for the binary classification problem: Logistic Regression, Decision Tree, and XGBoost. In the context of medical data and the importance of correctly diagnosing cardiovascular disease, performance metrics like F1 score are considered. The balance between precision and recall is crucial in medical applications. High precision ensures confidence in positive predictions, reducing false alarms and unnecessary stress for patients. On the other hand, high recall ensures that the model is sensitive enough to detect all actual positive cases, reducing the risk of missing critical diagnoses. Striking a balance between precision and recall, as captured by the F1 score, is essential to develop a reliable and accurate predictive model.

ML Models Used:
Our baseline logistic model had an F1 score of 66% on both the training and test data, showing that it's not overfitting. However, for a medical problem like predicting the risk of heart disease, this score is considered low. We need a higher F1 score to ensure more accurate predictions. Then we used the decision tree model. Though the model was giving back a high F1 SCORE, it was overfitting even after hyperparameter tuning as well. At last, we chose cross-validate hyperparameter-tuned XGBoost classification model as the final model, giving an F1 score of 85%, showing an Improvement of 29.797% against tuned logistic Regression.

Feature Importance:
Lastly, we checked for the most important features for model explainability and found 'bpmeds,' 'age,' and 'cig per day' as the most important features, which play the most crucial role in predicting the 10-year risk of coronary heart disease.

Model Saving:
At last, we saved this model as a pickle file for model deployment.