**Analysis Plan**

The 8 candidate predictors include age, BMI, systolic blood pressure (SBP), diastolic blood pressure (DBP), total cholesterol, behavioral pattern (A vs. B), smoking status (yes vs. no), and arcus senilis (present vs. absent). The demographic and clinical characteristics of the study population will be described using frequencies with percentages for categorical variables or median values with interquartile ranges (IQRs) for continuous variables. Univariate and multivariable logistic regression will be performed to evaluate the association between candidate predictors and CHD outcome. Additional bivariate analyses will also be explored to assess the association between predictors.

The data will first be split into 70% training and 30% testing, stratified by CHD to ensure balanced outcome values in both sets. A 5-fold cross-validation repeated 5 times will be performed on the train data for model tuning. The machine learning (ML) models used for predicting CHD include random forest (RF), extreme gradient boosting (XgBoost), and support vector machines (SVM) with three different types of kernel (linear, polynomial, and radial basis function).

The SVM is a prominent classification and regression technique. It works by finding the optimal decision boundary in a high-dimensional space to separate different classes of data points or predicted values. Three distinct kernel types within SVM tailored to specific characteristics of the data and the complexity of the problem. Firstly, linear SVM works well for datasets where underlying relationships are straightforward and linearly separable. Secondly, for datasets with more intricate decision boundaries, polynomial SVM can capture the non-linear patterns and complexities inherent in CHD data. Finally, another powerful kernel, the radial basis function, allows us to model and classify outcomes in scenarios where the underlying relationships are exceptionally intricate and convoluted.

The RF model is a machine learning model that uses decision trees to handle classification and regression problems. The RF collects assorted trees by using the bagging method (Bootstrap Aggregation) where each tree is built on each unique sample that was randomly drawn from the data with replacement. In addition to random sampling of the data, the RF also subsets the features for each split of the tree to further “de-correlate” the data to create a robust model. The RF will make predictions by taking the average of all the predictions from each tree. By collecting multiple uncorrelated trees, the RF keeps the bias low while lowering the variance. When constructing an RF model, many parameters can be considered to build the model, such as the max depth, number of trees, and max features.

The XgBoost model is another machine learning technique that utilizes boosting method to capture complex interactions between candidate predictors and with CHD. It works by iteractively correcting the error made in a previous tree to minimize bias but at the same time keep the bias low. XgBoost can provide insights into feature importance, handle missing data(even though we don’t have to deal with this issue much at our project), and handle non-linear relationships between predictors and the outcome.

The models will be tuned by optimizing the accuracy on the same cross-validation set. Lastly, we will select the best model based on model performance and fit it to the test set to evaluate out-of-sample predictions.