**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 the 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). 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. Three distinct kernel types within SVM tailor to specific characteristics of the data and 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. These model 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.