C. can be changed based on how strict

C. Models We have started with running a logistic regression model that had an accuracy linear regression on the data. In order to have a baseline that we can compare to, we assumed a baseline that predicted all cases as true negative and no true positive. A Classification and Regression Tree (CART) model, as well as a Random Forest were also created.

We examined the predictive accuracy of each method using the framingham dataset, where the dataset was divided into two subsets with a split ratio of 65%, a training set comprised of 2385 observations and a test set of 1273 observation. Using each prediction method, a model was developed for predicting the probability of 10-Year CHD risk label from the observations in the training sample. We then applied the developed model to each subject in the test set to estimate each patient’s predicted probability of having 10-year CHD risk. The tree-based methods considered all 16 variables, described in the dataset section, as candidate predictor variables. Two separate logistic regression models were fit to predict the probability of 10-Year CHD risk. First, we fit a logistic regression model that contained all 16 variables as main effects. No feature reduction was performed.

Second, we fit a logistic regression model that contained 5 predictor variables that previously had been identified as important predictors of 10-year CHD using data from the Framingham Heart Study (sex, age, currentSmoker, cigsPerDay, prevalentHyp, sysBP, BMI, prevalentStroke, BPMeds). However, the rest of the models were fit using the entire variable set. We describe here the used models: Logistic RegressionLogistic regression is a specific type of Generalized Linear Models (GLM) used to

predict probabilities for binary classification, in this study, " 10-year CHD" risk label that is originally coded as 0 and 1.

Formally, the algorithm is defined as following: Where h(x(i))is the sigmoid logistic function. We predict one class if the probability is above a certain threshold and the other class otherwise. The threshold can be changed based on how strict we want to be on the data, in our case threshold= 0. 5. Classification and regression trees (CART) Classification and regression trees use binary recursive partitioning methods to partition the sample into distinct subsets.

At the first step, all possible splits of all continuous variables (above vs. below a given threshold) and of all categorical variables are considered. Using each possible split, all the possible ways of partitioning the sample into two distinct subsets are considered. The binary partition that results in the greatest reduction in impurity is selected. This process is then repeated iteratively until a predefined stopping rule is satisfied 8.

CART enables better interpretability of decision rules. Advantage of such tree-based methods is that it doesn't assume linearity or parametric form of the relationship with outcome variable and predictors 7. Classification and regression trees were built using the the " rpart" package of the R statistical programming environment. In our study, the default criteria in the rpart package was used in the rpart package, the complexity parameter cp= 0. 01, and a 10-fold cross validation was used.

Random Forest Random forest is a large collection of uncorrelated trees that are built and then averaged, it can be used for both regression and

classification. It is based on a group of n trees grown randomly, which output either a class for classification problems or a continuous variable for regression problems. In our experiment we have set the tree number to n= 200, both n= 100 and n= 300 yielded a declining performance.

The idea is to reduce variance as the bias of averaged bagged trees is the same as that of an individual tree 7. They’re similar to CART in that they capture high-order interactions between variables and handle mixed predictors. Random forests are better than CART in that they’re less prone to overfitting and because of using out- of-bag samples, cross validation is already built-in. Finally, boosting fits additional predictors to residuals from initial predictions (ICME ML Workshop, 2015):