1. Problem:

The reason we chose this dataset is that we are interested in finding the pattern of cardiovascular patients. We previously learned from the Center for Disease and Control that complication of cardiovascular disease, heart attack, has been a leading cause of death globally. And many patients did not realize that they had the disease. The main reason for unawareness of disease was due to the false negative diagnostics made by their medical doctors. We are motivated in creating this machine learning model so that people can self-identify the presence of developing cardiovascular disease (CVD). Also, doctors can use the methods we created as an assurance to check for potential mis-diagnostics which possibly reduce the fatal rate. In addition, identification of individuals can lead to earlier intervention and better management of risk factors, ultimately reducing the burden of CVD. The specific binary outcome is to determine whether a person has cardiovascular disease based on the biomarkers(predictors) given. The background for the dataset will be further learned in the original articles and researches related to cardiovascular disease indicators have been discussed in many articles, which can be accessed by the website PubMed.

2.Data:

Dataset: https://www.kaggle.com/datasets/yassinehamdaoui1/cardiovascular-disease

Data was collected in Western Cape in South Africa. The raw dataset is preprocessed by the data collector before we have access. We found no missing values. Variable for identity is removed since it is not related to the outcome. We previously didn't remove identity since it would be removed automatically in the variable selection stage. We are now removing it in the beginning because we found that variable to be meaningless in our project. One variable, famhist, was changed from "Present" to "YES", and "Absent" to "NO", which was not a necessary change but to match with our response that would be discussed later. After reviewing the data, there are 10 variables and 462 observations in our dataset. They are systolic blood pressure (sbp), cumulative tobacco, low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease, a factor with levels "Absent" and "Present" (famhist), type-A behavior, obesity, current alcohol consumption, age, and response, coronary heart disease (Chd); famhist is a categorical variable; and others are numeric variables. Chd is our binary response. Chd also had a change that was not necessary. Chd was changed either from "1" to "YES", and "0" to "NO" or kept the original form depending on the coding setting. Some models were visually better with changed form and some did not matter.

3. Approach:

To obtain the best fitting results, 7 machine learning models for classification are selected to compare their results, which are Logistic Regression, Linear Discriminant Analysis(LDA), Quadratic Discriminant Analysis(QDA), K-Nearest Neighbor (KNN), Tree, Decision Tree, Random Forest and Boosting model. Bagging is another common method listed in the proposal but we chose not to use it since our variables are far less than 4000. And the evaluation metric is set as sensitivity. Specificity is also computed as some models showed identical sensitivity. Specificity is only used to evaluate the models with the same sensitivity, and for reference purposes as specificity needs to be reasonable to make sure sensitivity results are not due to overfit. We planned to use accuracy as our metric in the proposal. But accuracy consists of more than what we are looking for, so we changed to sensitivity. We believe that high sensitivity can diagnose people who have disease as positive and more accurate. The diagnosis of cardiovascular disease is important for the true patients since they will be facing fatal risks such as sudden heart attack, and we want them to be aware of that rather than not knowing.

The potential useful predictor will be retracted based on 5 selection criteria, subset of lowest BIC(Formula 1)forward stepwise selection with highest R-square value(Formula 2), backward

stepwise selection with lowest mallow's cp value(Formula 3), Lasso regression(Formula 4) and Ridge regression(Formula 5). 5 formulas for selected predictors are shown below.

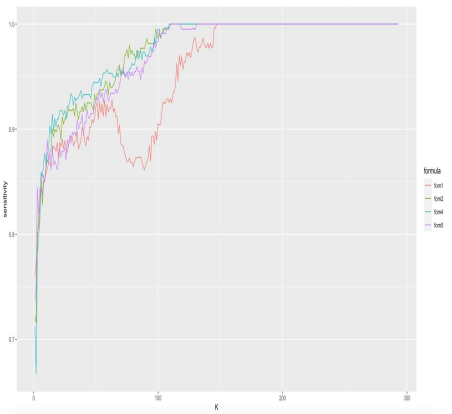
	Tabacco	Age	Obseity	Type-a	Sbp	Ldl	Alcohol	Adposity	Family history
Formula 1	✓	✓							✓
Formula 2	✓	✓	✓	✓	✓	✓			✓
Formula 3	✓	✓	✓	✓	✓	✓			✓
Formula 4	✓	✓	✓	✓	✓	✓	✓	✓	✓
Formula 5	✓	✓			✓	✓			✓

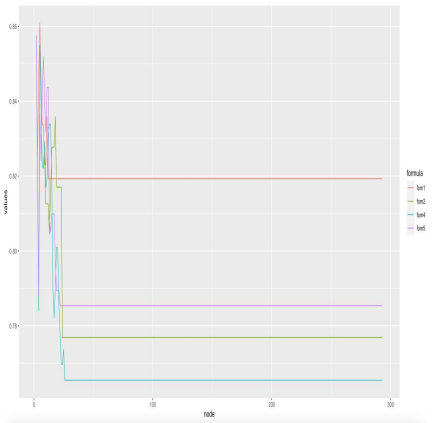
Since there are complete duplicate predictors obtained between forward selection and backward selection, the result of backward selection was removed from the total sets, resulting in a total of 4 formulas.

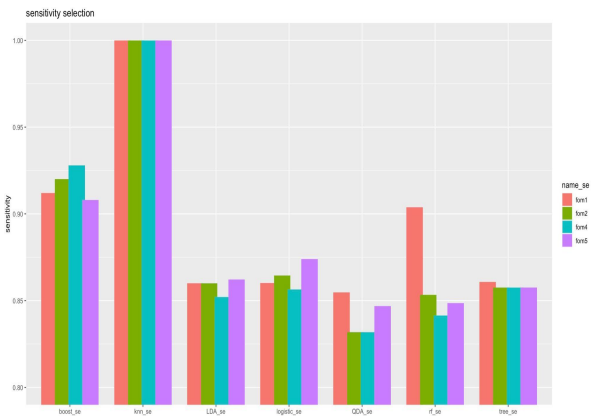
The best formula for each method will be selected by their evaluation index, sensitivity, on the training dataset(80%), meaning there will be 28 total combinations of formulas and methods. If the sensitivity is the same, the second index, specificity, will be tested.

Because of the small sample size (n = 462), application of simple splitting of training and testing dataset can cause high variance in the evaluation metric. Changing from the proposal, we only conducted Cross-validation which can provide a more reliable estimate of the model performance by using all available data. As a result, cross validation with 5-folds will be conducted within 80% of the entire dataset, meaning that 64% of the entire set will be used for training, and 16% of the entire set would be validation dataset for cross validation purpose only. 20% of the entire set would be left and not used until step 4.

When setting up the cross-validation, the hyperparameters are taken into consideration. K in KNN model and number of tree nodes in the decision tree are tested for the best value. And the ranges for K and nodes are the same as 1 to 293, where 293 is the size of 4 folds(16% for each fold) in our training data of cross validation(80%). Two hyper parameters selection graphs are following.







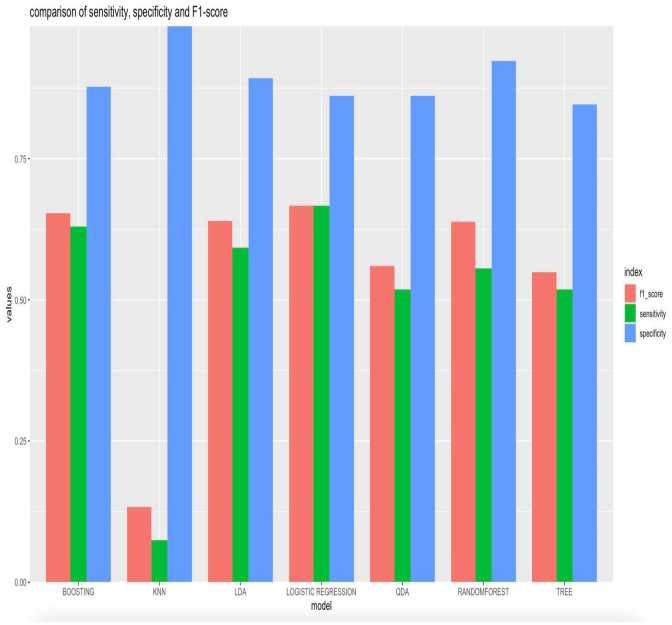
The training results are shown left.

The sensitivity within each method using different formulas is recorded. However, for KNN, the highest sensitivity of 4 formulas is the

same as 1 in the cross-validation dataset. This may result from the overfitting problem. For further selection, the second evaluation index specificity is tested on the cross-validation dataset. Out of 4 formulas, Formula5 with K=110 has the highest specificity.
In boosting, the number of best trees is also tested by cross validation. We found that formula 4 with 347 trees is the best, with a shrinkage of 0.01.

In summary for the model selection, The best combinations are logistic regression with formula5, LDA with formula5, QDA with formula1, decision tree with formula1 nodes = 5, random forest with formula1, boosting with formula4 best trees = 347 and Knn with formula 5 and K=110. They will be recorded and used for evaluation in step 4.

4. Evaluation and Results:



We will now use the 20% of the entire set(92 observations) to calculate the sensitivity of each model with a matched formula that allows us to determine the performance for each model.

The logistic regression model achieved a sensitivity(green) of 67%, specificity of 86% and F1-score 0.67, which has the highest sensitivity among all models. This means that the model correctly diagnosed 67% of individuals with heart disease, while correctly identifying 86% of those without heart disease. These statistics indicate that our diagnostic methods are reliable and have the potential to aid in early detection and treatment of heart disease. Overall, the chosen predictors and model selection process have allowed researchers to develop an accurate and effective diagnostic tool. Although we are not interested in f1 score and specificity, they are also computed because they need to be in a reasonable number.

Limitations and Messages for Future Researchers

One of the limitations of our study is the sample size. Although we took measures to ensure the quality of our data, the limited number of observations could potentially affect the accuracy of our results. Another potential weakness is the reliance on self-reported data, which could introduce bias and errors in our analysis. Additionally, our study only focused on a specific population, which may not be representative of other groups or regions.

To address the limitations of our study, we suggest future research could expand the sample size to improve the statistical power of the analysis. In addition, future studies could utilize more objective measures to reduce the reliance on self-reported data. This could include collecting biological samples or using wearable devices to monitor health outcomes. Another area for future research could be to examine the impact of different predictors on the accuracy of the diagnostic tool. This could involve exploring the influence of environmental factors, lifestyle choices, and genetic predispositions. Finally, it will be interesting to investigate the generalizability of our findings to other populations and regions to better understand the potential impact of our diagnostic tool on a larger scale. Also, using our models as a basic, further development can be done for implementing severity tests for patients.