Coronary Heart Disease Prediction

By: Nataliya Volkova and Thanojkumar Guntupalli

1. Introduction

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 34 seconds in the United States from cardiovascular disease. About 697,000 people in the United States died from heart disease in 2020 - that's 1 in every 5 deaths.

The "heart disease" term covers several types of heart-related conditions, including arrhythmias, congenital heart defects, disease of the heart muscle, coronary heart disease, and others. Coronary artery disease (CAD) is the most common type of heart disease in the United States. It is also called coronary heart disease or ischemic heart disease. In the United States, coronary heart disease is the most common heart disease, killing 382,820 people in 2020. About 20.1 million adults aged twenty and older have coronary heart disease (about 7.2% of the population). In 2020, about 2 in 10 deaths from coronary heart disease happened in adults less than 65 years old.

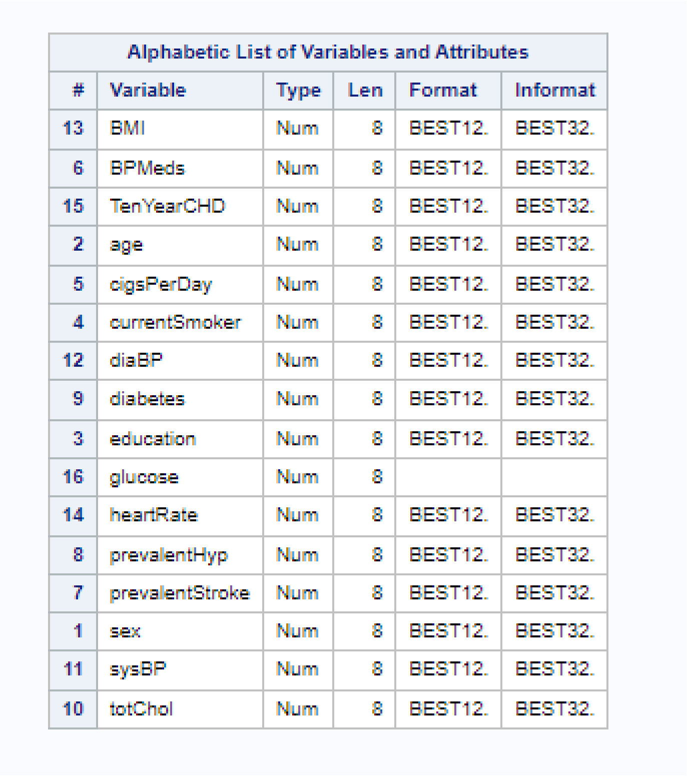
Coronary heart disease is the main cause of heart attacks all over the world. Every year, about 805,000 people in the United States have a heart attack, including 605,000 that are a first heart attack and 200,000 happen to people who have already had a heart attack. About 1 in 5 heart attacks are silent - the damage is done, but the person is not aware of it. In addition, heart disease cost the United States about $229 billion from 2017 to 2018. This includes the cost of health care services, medicines, and lost productivity due to death. That is why early prevention and prediction of heart disease play a crucial role today. Traditionally, to predict heart disease, healthcare providers rely on such risk factors as blood pressure, cholesterol levels, diabetes, smoking status, age, and family history.

In this particular research, our goal was to predict coronary heart disease (CHD) by building a model and to find significant predictors that cause CHD.

2. Dataset and Methods

2.1. Dataset

The dataset was downloaded from the Kaggle website. It consists of 4238 observations and 15 attributes: sex, age, education, number of cigarettes per day, BP meds (whether a patient is on blood pressure medication), prevalent stroke, prevalent hypertension, diabetes, total cholesterol level, systolic blood pressure, diastolic blood pressure, BMI, heart rate, glucose level, and ten years risk of CHD (coronary heart disease).



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
13	BMI	Num	8	BEST12.	BEST32.
6	BPMeds	Num	8	BEST12.	BEST32.
15	TenYearCHD	Num	8	BEST12.	BEST32.
2	age	Num	8	BEST12.	BEST32.
5	cigsPerDay	Num	8	BEST12.	BEST32.
4	currentSmoker	Num	8	BEST12.	BEST32.
12	diaBP	Num	8	BEST12.	BEST32.
9	diabetes	Num	8	BEST12.	BEST32.
3	education	Num	8	BEST12.	BEST32.
16	glucose	Num	8		
14	heartRate	Num	8	BEST12.	BEST32.
8	prevalentHyp	Num	8	BEST12.	BEST32.
7	prevalentStroke	Num	8	BEST12.	BEST32.
1	sex	Num	8	BEST12.	BEST32.
11	sysBP	Num	8	BEST12.	BEST32.
10	totChol	Num	8	BEST12.	BEST32.

2.2. Methods

After importing our dataset into SAS, we cleaned it and formatted the data. In order to better understand the data, we performed an exploratory data analysis where we explored the relationship between variables through building different types of plots, performing t-test and Wilcoxon test, looking at the distribution of the data.

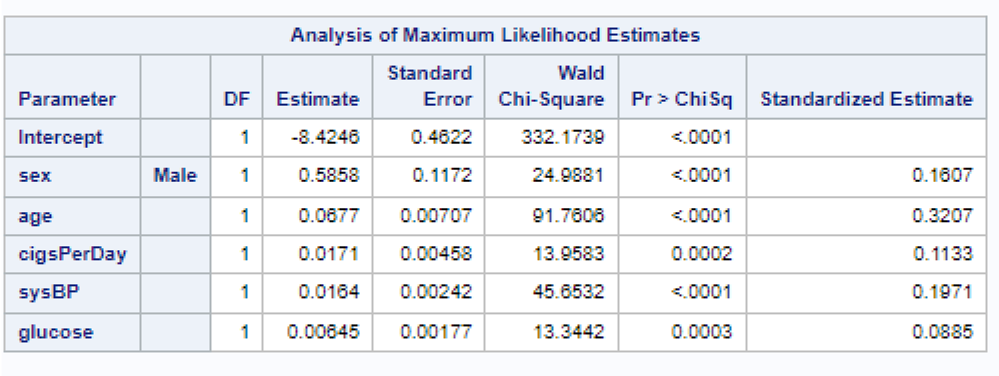
Since our dependent variable (ten-year coronary heart disease) is binary, we have decided to build a logistic regression model. We split the dataset into two sets: 80% training set and 20% validation set. We used a stepwise selection method to choose statistically significant predictors. From the model output, we found that significant predictors are age, systolic blood pressure, sex, glucose, and number of cigarettes per day. The model was built on the training set, and the validation set was used to check the model's accuracy. To further assess the model, we also looked at the regression output, including testing the global null hypothesis, analysis of maximum likelihood estimates, risk limits analysis, Hosmer-Lemeshow Fit Test, classification table, and ROC curve.

We were looking for other methods which would allow us to get better results, and we decided to build a classification model. We assessed the model based on the classification subtree, confusion matrix, and ROC curve.

3. Results

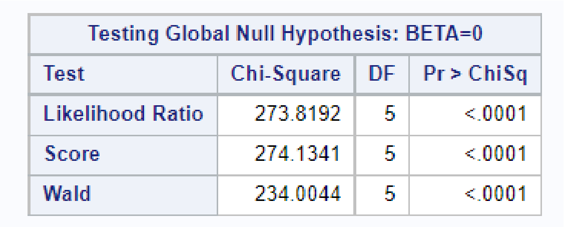
After using a stepwise selection, the formula of our model built on the 80% training set would be as follows:

$$\log\left(\frac{Pchd}{1-Pchd}\right) = -8.4246 + 0.5858\text{male} + 0.0677\text{age} + 0.0171\text{cigsPerDay} + 0.0164\text{sysBP} + 0.00645\text{glucose}$$



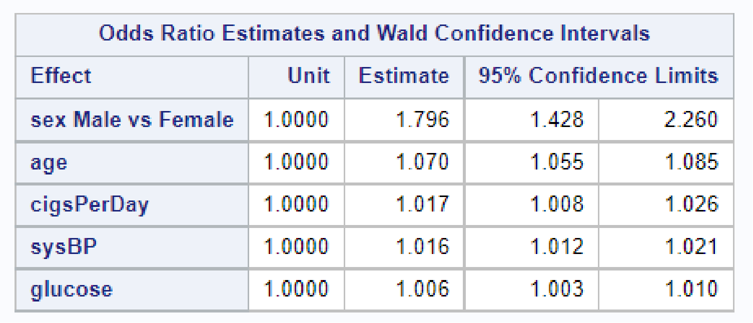
Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept		1	-8.4246	0.4622	332.1739	<.0001	
sex	Male	1	0.5858	0.1172	24.9881	<.0001	0.1607
age		1	0.0677	0.00707	91.7606	<.0001	0.3207
cigsPerDay		1	0.0171	0.00458	13.9583	0.0002	0.1133
sysBP		1	0.0164	0.00242	45.6532	<.0001	0.1971
glucose		1	0.00645	0.00177	13.3442	0.0003	0.0885

To better understand the model fit, we need to interpret the output. Testing Global Null Hypothesis table would show if the model is predictive or not. (H0: all the partial slopes are 0, Ha: at least one of the partial slopes is not 0). Based on the output, we reject the null hypothesis and say that we do not have enough evidence to suggest that all the partial slopes are 0.



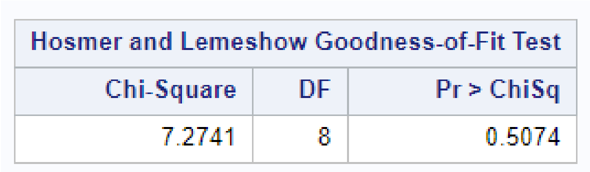
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	273.8192	5	<.0001
Score	274.1341	5	<.0001
Wald	234.0044	5	<.0001

Odds Ratio Estimate and Wald Confidence Intervals table would show the 95% confidence interval for the predictors; when the interval does not include 1, we understand that every increase in the predictor is significant. So, an increase in all our predictors is significant. Also, the Estimate column shows the odds of having coronary heart disease when a predictor increased by 1 unit.



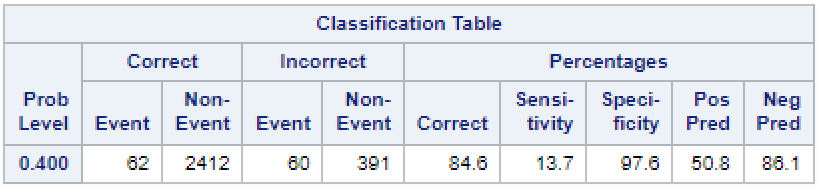
Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
sex Male vs Female	1.0000	1.796	1.428	2.260
age	1.0000	1.070	1.055	1.085
cigsPerDay	1.0000	1.017	1.008	1.026
sysBP	1.0000	1.016	1.012	1.021
glucose	1.0000	1.006	1.003	1.010

To check the goodness of fit, Hosmer-Lemershow goodness-of-fit test was used. (H0: the data fits the expected values, Ha: the data does not fit the expected values). Based on the output, we fail to reject the null hypothesis and say that we do not have enough evidence to suggest the data does not fit the expected values.



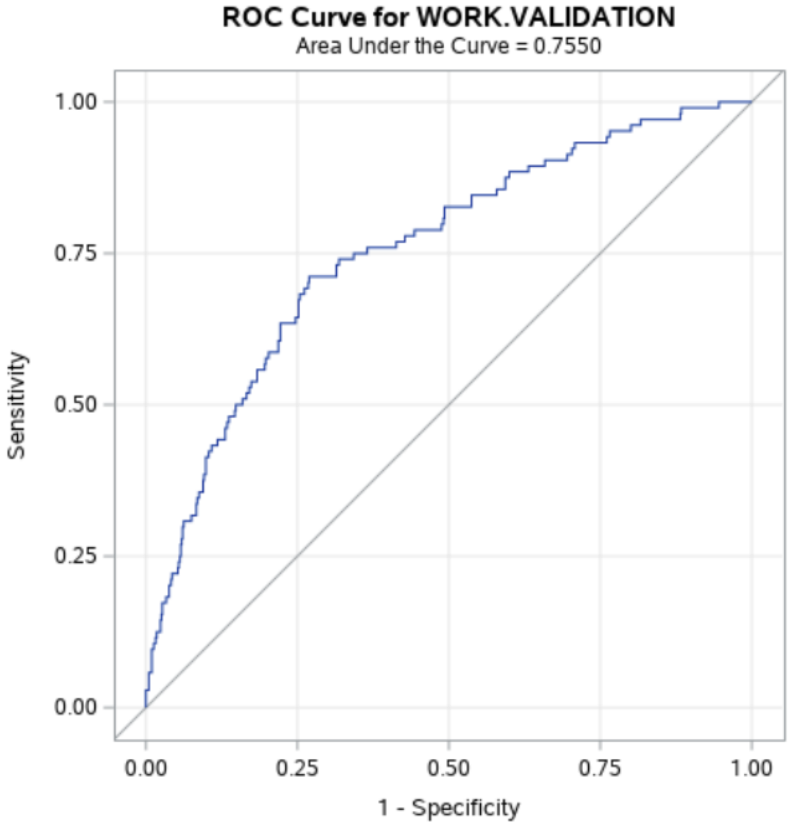
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.2741	8	0.5074

From the classification table, we found out that sensitivity is 13.7%, specificity is 97.6%, positive predictive value is 50.8%, and negative predictive value is 86.1%

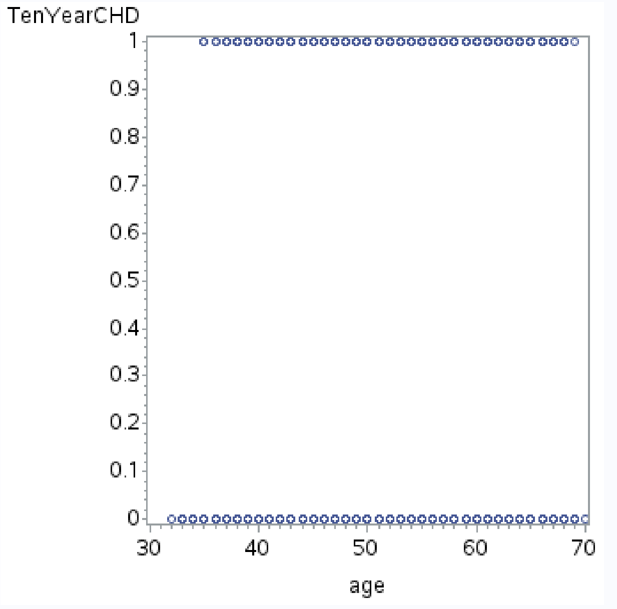


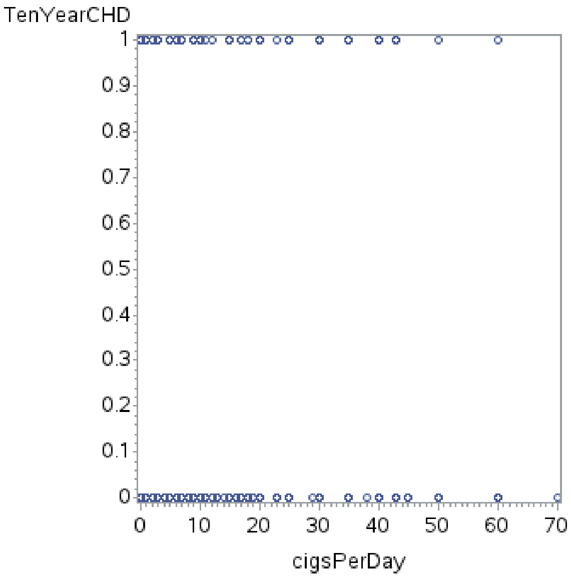
Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.400	62	2412	60	391	84.6	13.7	97.6	50.8	86.1

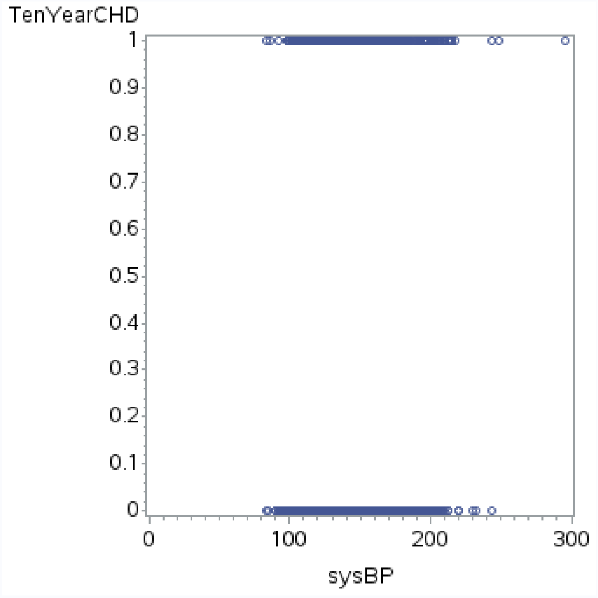
ROC curve is a graphical presentation of the true positive rate (sensitivity) against the false positive rate (1-specificity). The area under the ROC curve (AUC) shows the overall quality of the model. So, AUC for the model based on the validation set is 0.7550, which shows a fair fit.

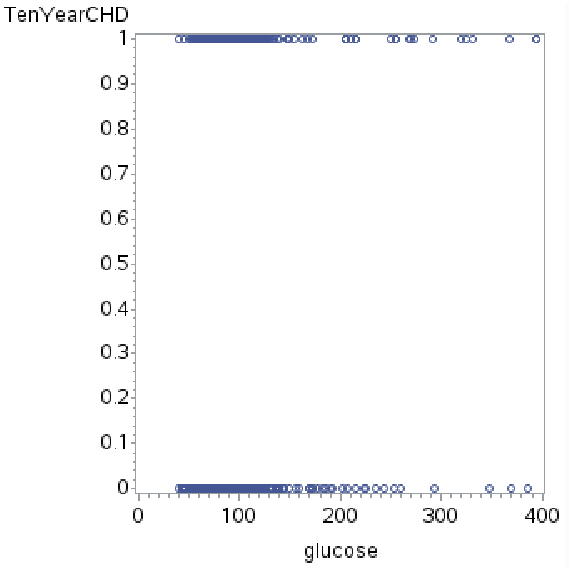


We plotted the response variable vs all significant predictors. From the below scatter plots, we can see that there is no issue of complete separation of the data in our model.





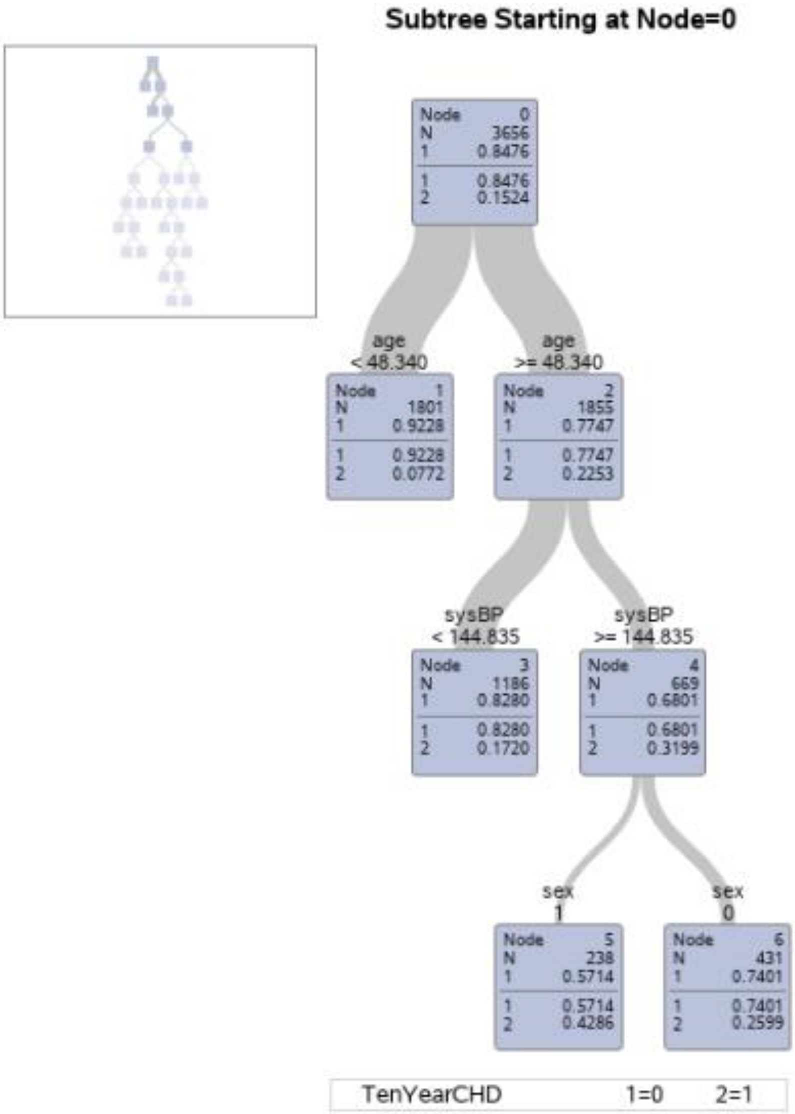




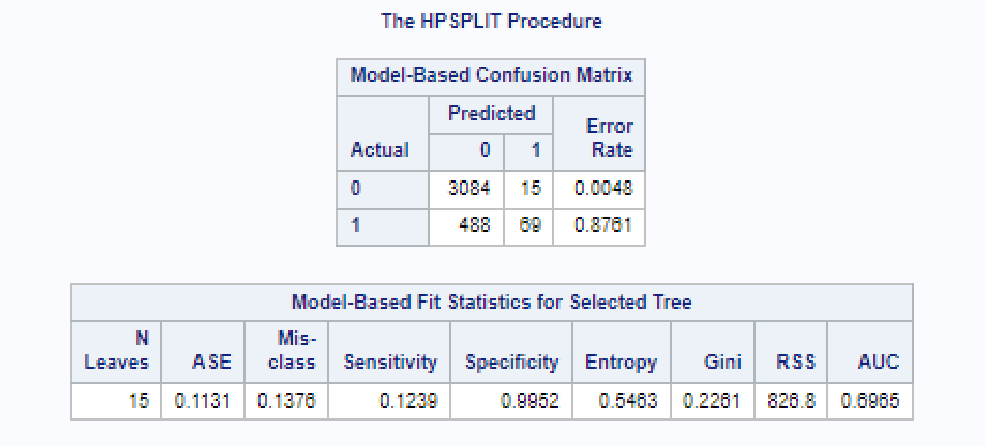
In an attempt to improve the accuracy, we decided to build a classification model. We interpreted the model based on the classification subtree, confusion matrix, and ROC curve.

The subtree below shows that the data is divided first by age (the tree uses 1 for not having CHD, and 2 for having CHD). People older than 48.34 years have more chances of having CHD. For example, the probability of having CHD for people younger than 48.34 is only 0.0772, and not having is 0.9228. For comparison, the probability of having CHD for patients older than 48.34

is increasing and equals to 0.2253, not having is 0.7747. The next division is made by systolic blood pressure, where we see that having a pressure more or equal to 144.835 increases the chances of CHD. The last division is made by sex, where men have more chances of getting CHD in comparison with women.



From the ROC curve, we get the AUC value of 0.70, which is less compared to the logistic model. The specificity (0.9952) is slightly higher than the logistic model, but the sensitivity (0.1239) of the classification model is very small.



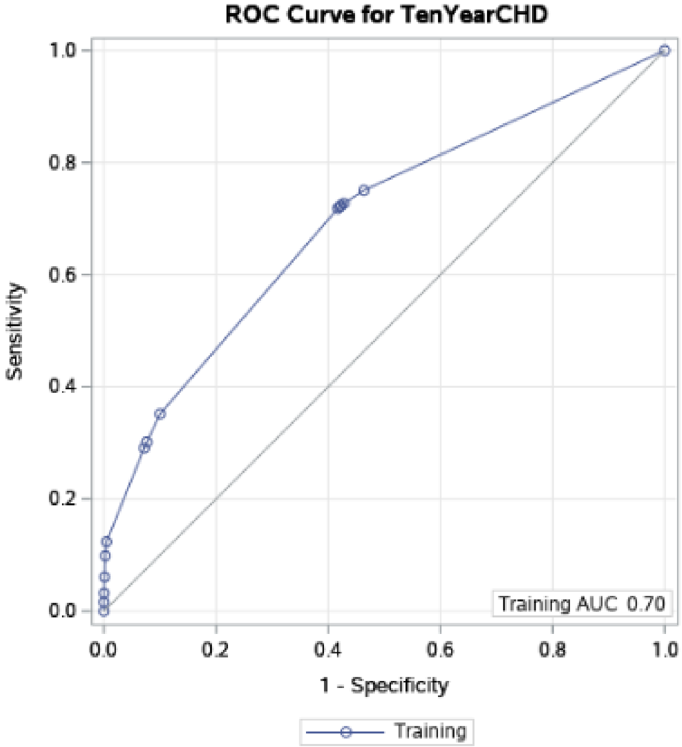
The HPSPLIT Procedure

Model-Based Confusion Matrix

	Predicted		
Actual	0	1	Error Rate
0	3084	15	0.0048
1	488	69	0.8761

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
15	0.1131	0.1376	0.1239	0.9952	0.5463	0.2261	826.8	0.6965



4. Discussion and conclusion

It is difficult to overestimate the importance of finding methods that will let healthcare providers find supplementary tools to predict heart disease. Ultimately, it will not only allow people to live longer and healthier lives but also decrease the budget expenditure for all the expenses associated with medical costs, loss of productivity, and other related expenses.

In this project, we have built two models: a logistic regression model and a classification model. After comparing them, we found out that the logistic model has better prediction based on AUC value. Also, we found that significant predictors for predicting CHD in the logistic regression model were age, systolic blood pressure, sex, glucose, and number of cigarettes per day.

To improve the model, more data might be needed. For future research and improvements of our model, we can use a cross-validation technique to boost the accuracy. Also, some other models, like Random Forest, can be implemented and compared in order to find the best model.

References

1. Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2020. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2022. Accessed February 21, 2022. https://wonder.cdc.gov/mcd-icd10.html
2. Heart Disease Facts. Centers for disease control and prevention. https://www.cdc.gov/heartdisease/facts.htm
3. Tsao CW, Aday AW, Almarzooq ZI, Beaton AZ, Bittencourt MS, Boehme AK, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation.* 2022 https://www.ahajournals.org/doi/10.1161/CIR.0000000000001052
4. Coronary Artery Disease (CAD). Centers for disease control and prevention. https://www.cdc.gov/heartdisease/coronary_ad.htm