5-2020

Life and death: Quantifying the risk of heart disease with machine learning

Jack Scott Glienke
University of Northern Iowa

LIFE AND DEATH:

QUANTIFYING THE RISK OF HEART DISEASE WITH MACHINE LEARNING

A Thesis Submitted

in Partial Fulfillment

of the Requirements for the Designation

University Honors

Jack Scott Glienke

University of Northern Iowa

May 2020

This Study by: Jack Glienke

Entitled: Life and Death: Quantifying the Risk of Heart Disease with Machine Learning

has been approved as meeting the thesis or project requirement for the Designation University Honors

_____	_____
Date	Marius Somodi, Honors Thesis Advisor, Department of Mathematics
_____	_____
Date	Dr. Jessica Moon, Director, University Honors Program

Abstract

Coronary heart disease has long been a key area of focus in the discussion of public health. As such, numerous studies have been conducted throughout history with the sole intention of identifying risk factors leading to the onset of cardiovascular conditions. A plethora of statistical procedures can be used to identify an individual's risk of developing heart disease, yet regression models tend to be the default tool used by researchers. Using the data obtained from the most influential cardiovascular study to date, the Framingham Heart Study, this analysis uses machine learning techniques to generate and test the predictive power of four different classification methods: logistic regression models, decision trees, random forests, and support vector machines. The findings of this study indicate that logistic regression is the most accurate classification technique; it correctly predicts whether an individual will develop coronary heart disease more than 84% of the time.

1. Introduction

In this day and age, technological advancements are continually reshaping the manner in which society approaches everyday life. From where people work to the tools they have at their disposal, the constant influx of new technology is altering entire industries around the globe at unprecedented rates. The medical field is not immune to this concept. As a result, society's understanding of various health-related issues is constantly being improved. If one takes this view and applies it specifically to the study of coronary heart disease, these advancements are quite apparent. Although new technology and various other studies have since expanded upon its initial research, the Framingham Heart Study remains one of the most influential cardiovascular studies ever conducted. This longitudinal study laid the foundation for the current understanding of heart disease, and its impacts are widespread. With that being said, this renowned research still has areas which can be further developed.

One such area that is largely under-examined is the quantification of an individual's risk of developing heart disease through the use of machine learning techniques. While it is true that the Framingham Heart Study has developed an online model (called the Framingham Risk Calculator) to help quantify this prospect, the researchers elected to use Cox regression to achieve their goal. As such, there is still a considerable opportunity for various machine learning techniques to further examine these data. With the assistance of four classification procedures (logistic regression, decision trees, random forest analysis, and support vector machines), this study analyzes the effectiveness of machine learning techniques in predicting the onset of coronary heart disease. It is hypothesized that logistic regression will outperform the other models constructed throughout this study. Because this method requires the most rigorous prep-

work and offers more than a dozen viable models to choose from, this hypothesis assumes that it will yield the most compelling results.

Upon completion of the aforementioned analyses, this study confirmed the initial belief: logistic regression techniques yielded the most accurate classification results, although it only slightly outperformed the other methods. More specifically, this model accurately classified data 84.86% of the time. At the same time, the machine learning technique with the least predictive power was random forest classification. This method wielded an accuracy rate of 83.53%. This paper will provide a background on the factors influencing the development of heart disease, discuss the history and contributions of the Framingham Heart Study, detail the models generated through throughout this project, and interpret the results of this statistical analysis.

2. History of Framingham Heart Study

While the medical community now possesses a wealth of information surrounding the development and prevention of coronary heart disease, this was not the case a mere seventy years ago. In the 1940s in the United States, cardiovascular disease was the leading cause of mortality; it accounted for half the deaths nationwide (Mahmood et al.). Despite this daunting statistic, a surprisingly small amount of action was taken to resolve this issue. Both the lack of knowledge and technology limited corrective action during this time. As such, many Americans continued to see early death from heart disease as unavoidable (Mahmood et al.). Unfortunately, this belief wasn't challenged until President Franklin Roosevelt died from cardiovascular complications in 1945. Beginning in 1948, the American government took matters into their own hands. On June 16th, the "National Heart Act" established the National Heart Institute in an attempt to reduce the impact of cardiovascular disease that was felt throughout the country (Mahmood et al.). The

passing of this law would forever change the medical understanding of America's leading cause of death.

Immediately following the National Heart Act's inception, researchers began to organize the Framingham Heart Study. In the early stages of this process, analysts disagreed whether the study should be observational or preventative in nature, but the former methods were eventually selected (Mahmood et al.). As such, the Framingham Heart Study officially originated in 1948 with the purpose of identifying the impact and developmental risk factors associated with heart disease in the American population (Randall et al.). The original participants of this study, later deemed the "Original Cohort," consisted of 5,209 men and women between the ages of thirty and sixty-two ("About the Framingham Heart Study").

After determining the scope and the participants involved in the study, strict data collection methods were put into place. According to Connie Tsao and Ramachandran Vasan, each participant in the Framingham Heart Study was required to have a physical examination every two years. During these visits, individuals underwent in-depth cardiovascular exams and discussed all medical and family history with a doctor (Tsao and Vasan). Furthermore, researchers were continually in contact with all study participants through various questionnaires and phone calls. As a result of these diligent efforts, crucial information was unveiled. More specifically, researchers utilized the data obtained from the Original Cohort to identify numerous behavioral risk factors such as smoking and dietary habits (Randall et al.). Furthermore, these analyses provided evidence against the widely accepted belief that systolic blood pressure played little role in the development of coronary heart disease (Mahmood et al.).

While the aforementioned findings were groundbreaking at the time of their discovery, researchers continued to refine their approach to this important topic. As a result, the

Framingham Heart Study began to take a new form. Beginning in the early 1970s, researchers switched their focus from detection to prevention. Using the numerous heart disease risk factors that were identified earlier in the study, researchers developed "risk scores" with the help of cumbersome multivariable risk functions and multiple cross-classification (Mahmood et al.). These risk scores were the first attempt to identify specific individuals which may be at risk of coronary heart disease in the future; these values were based upon large tables that compared various combinations of heart disease risk factors. Although these methods were rather impractical when analyzing more than a few risk factors at once, they laid foundation for what would later become known as "Framingham Risk Scores" (Mahmood et al.). Although this concept underwent numerous revisions, it eventually found its most popular form in 1998. In this version, researchers established simplistic tables which allowed doctors to look-up values for various risk factors, thus easily determining whether an individual was at risk for heart disease (Mahmood et al.).

Using the above method as an important basis, researchers eventually developed the Framingham Risk Calculator ("Cardiovascular Disease (10-Year Risk)"). This model is currently located both online as well as in numerous different phone applications. Using Cox regression techniques, researchers developed a model which would return the probability of contracting coronary heart disease when several common health factors – such as age and blood pressure – were treated as inputs (D'Agostino et al.). Thus, this model allows for a convenient way to quantify an individual's risk of developing heart disease, and it is discussed in greater detail later in this paper.

At roughly the same time that the Framingham Heart Study began to examine preventative measures for coronary heart disease, it expanded its research within another realm

as well. In 1971, the study grew in size to incorporate the descendants of the Original Cohort ("About the Framingham Heart Study"). This expansion led to the development of the Offspring Cohort, which consisted of 2,656 additional participants (Tsao and Vasan). In addition to continuing the previous research, this also allowed the analysts to study heredity as a component of cardiovascular disease. Since this addition in 1971, various other cohorts have been added to the study in an attempt to increase diversity across the volunteers (Tsao and Vasan). These new groups are called the Omni Cohorts, and the most recent addition was in 2003 (called the Second Generation Omni Cohort).

3. Heart Disease and Mortality

Although the Framingham Heart Study is one of the premier longitudinal heart studies to date, it would have been impossible for a single analysis to fully research every aspect of heart health. As such, specific areas of focus were selected, and the remaining topics were left to be more adequately fulfilled by various other studies. Two topics that the Framingham Heart Study researched as secondary subjects are the effects of gender and ethnicity in the development of cardiovascular conditions. The current research concerning these two topics are detailed in the following paragraphs.

When the initial studies concerning coronary heart disease first began, very little was known about this widespread and deadly condition. While this general lack of knowledge has since been eradicated, new questions have emerged. Of the continued research being conducted in the field of cardiovascular health, a prominent area of focus is the difference in coronary heart disease across gender. Although the overall findings from different research studies examining this relationship differ slightly, one theme remains constant: coronary heart disease is more prevalent in men (Claassen et al.; Pilote et al.). According to the study conducted by Claassen,

Sybrandy, Appelman, and Asselbergs, the prevalence of coronary heart disease for men residing in the United States was 37.4%; the corresponding rate for US women was 35%. The mortality rates from coronary heart disease differ based upon gender as well. The same study concluded that men and women with coronary heart disease faced respective death rates of 48.2% and 51.8% (Claassen et al.). Furthermore, the number of deaths associated with cardiovascular disease have slowly been decreasing in men over the last thirty years, but women have not yet reaped the same benefit (Pilote et al.). One potential reason for these significant cross-gender differences may be the presence of distinct risk factors. Claassen and colleagues found that hypertension and diabetes were more common in women who were diagnosed with cardiovascular disease. Conversely, smoking was more prevalent in men (Claassen et al.).

Going beyond the effects of gender, researchers have also investigated the role ethnicity and socioeconomic status play in determining the risk of heart disease. Although various studies have sought to unveil this relationship, conflicting results have at times arisen. While this remains the case, several studies have presented compelling arguments which indicate a link between the aforementioned variables. One of these studies was conducted by Jess Kraus, Nemat Borhani, and Charles Franti. Interestingly enough, this analysis used data which was partially obtained from the Framingham Heart Study. Moreover, the study considered data from five common ethnic groups: (1) white, (2) black, (3) Asian, (4) Spanish American, and (5) American Indian. As the following figure represents, the researchers concluded that the risk of a coronary heart disease event is inversely related to socioeconomic status (Kraus et al.).

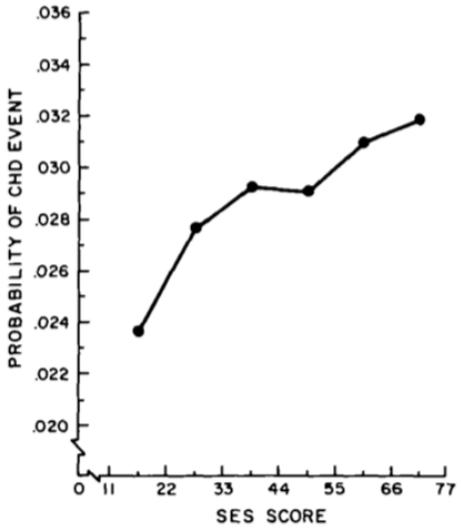

Figure 1. Heart disease and socioeconomic status from: Kraus, Jess F., et al. "Socioeconomic Status, Ethnicity, and Risk of Coronary Heart Disease." *American Journal of Epidemiology,* vol. 111, no. 4, 1980, p. 411.

Using factors based on education levels and occupation, the researchers in the above analysis determined socioeconomic scores ranging from 11-77 for each participant; high scores corresponded to low socioeconomic status. As one can see, the lower the socioeconomic class, the greater the probability of experiencing a coronary heart disease event. This relationship persisted across four of the five ethnic groups considered in this study (Kraus et al.). More specifically, the researchers determined that only the black ethnic group did not have an inversely proportionate relationship between coronary heart disease and socioeconomic status (Kraus et al.).

4. Other Contributions

Since the work presented in this thesis is most comparable to the regression model available on the Framingham Heart Study webpage, it is important to fully understand this technique. Prior to the year 2008, the models that were commonly used to determine an individual's risk of developing cardiovascular disease were developed through multiple cross-classification techniques (Mahmood et al.). Because these tables were increasingly difficult to

use when considering numerous risk factors, there existed an opportunity for a better model to capitalize on this shortcoming. As such, D'Agostino and colleagues developed several different models that were not subject to this restraint in 2008 (D'Agostino et al.).

Using nearly 8,500 observations obtained from the Framingham Heart Study, Cox regression techniques were used to develop two different risk algorithms (D'Agostino et al.). The first of these models treated many of the commonly cited risk factors as potential predictors of cardiovascular disease; as such, cholesterol levels, blood pressure, and various other medical-based values were needed as inputs for the model (D'Agostino et al.). Because of this, the researchers elected to investigate the construction of a precise prediction model that only used "non-laboratory-based predictors" (D'Agostino et al.). In other words, the second model that resulted from this study only included variables which could easily be measured without the assistance of a physician. Some of these variables include the participant's age, gender, smoking habits, body mass index, and diabetic diagnosis.

Both of the aforementioned prediction models are now available on the Framingham Heart Study website. Although the two models have several predictors in common, slight differences remain ("Cardiovascular Disease (10-Year Risk)").

5. Description of Data

Because this study could profoundly impact the manner in which medical professionals identify individuals at risk of coronary heart disease, it was paramount that the data used for this analysis came from a reliable source. As such, this study utilized data from the most influential and respected cardiovascular study throughout history: the Framingham Heart Study. When the data used in this study was first obtained, it consisted of 4,240 observations across sixteen variables. It is key to note, however, that several observations within this dataset contained

missing values because of collection constraints within the original study. Since the machine learning functions within RStudio are programmed to handle missing values differently, it was important to eliminate this source of variability from the present study. As such, any observation that contained a missing value was removed from the dataset and not included in this analysis. The result was a final dataset consisting of 3,658 observations across sixteen variables.

In each of the classification models that this study generated, "TenYearCHD" was treated as the dependent variable. This variable was binary in nature; a value of "1" implied that the individual developed coronary heart disease within the ten-year period following the collection of this data, and a value of "0" implied a healthy diagnosis. Furthermore, the remaining fifteen variables in this dataset were treated as predictors in each of the classification models. Before these variables are explained in detail, however, one should understand two concepts. First and foremost, each of the variables included in this analysis are commonly cited as risk factors for coronary heart disease. As such, each one would likely hold some degree of predictive capabilities when considered alone. However, in the presence of several other predisposing characteristics, some variables will no longer be statistically significant in the classification process. Secondly, the predictor variables within this study are a combination of demographic, behavioral, and medical risk factors. Furthermore, some of the values were self-reported by each participant, while others were recorded by a doctor. The limitations associated with self-reported values should be considered when examining the predictive power of the resulting classification models.

Taking the above considerations into account, the remaining variables included in this dataset can now be examined. Each model generated throughout this study utilized the following fifteen variables as predictors: (1) "Male," (2) "Age," (3) "Education," (4) "CurrentSmoker,"

(5) “CigsPerDay,” (6) “BPMeds,” (7) “PrevalentStroke,” (8) “PrevalentHyp,” (9) “Diabetes,” (10) “TotChol,” (11) “SysBP,” (12) “DiaBP,” (13) “BMI,” (14) “HeartRate,” and (15) “Glucose.” The first three variables listed above are demographic in nature, and they are rather straightforward. The “Male” variable indicates the participant’s sex, with a value of “1” implying that the individual is a male. Next, the “Age” variable is a truncated version of each participant’s age at the time of their exam. The final demographic variable is “Education.” This variable can take on the integers one through four, where each value corresponds to a different amount of education that the individual completed. The meaning of these values are as follows: (“1”) some high school, (“2”) high school diploma or GED, (“3”) some college, and (“4”) college degree.

In addition to the demographic variables detailed above, the dataset used throughout this study also included two behavioral variables: “CurrentSmoker” and “CigsPerDay.” Both of these values were self-reported by the participant. The “CurrentSmoker” variable is a binary value which identifies whether the individual smokes cigarettes; a value of “1” implies that the participant smokes on a daily basis. The “CigsPerDay” variable quantifies the previous variable by providing the average number of cigarettes a participant smokes each day. It is key to note that the “CigsPerDay” variable must be non-zero if the participant identifies as a smoker.

Lastly, this dataset considered ten medical variables. Four of these (“BPMeds,” “PrevalentStroke,” “PrevalentHyp,” and “Diabetes”) are binary; a value of “1” implies the presence of the corresponding variable, and a value of “0” implies its absence. The six remaining variables are continuous, and each of these were recorded by a medical professional. For instance, “TotChol,” “BMI,” “HeartRate,” and “Glucose” were measures of the participant’s cholesterol levels, body mass index, resting heart rate in beats per minute, and glucose levels respectively. While each of these values are rather intuitive, two variables in particular likely

need further explanation: “SysBP” and “DiaBP.” The variable “SysBP” represents an individual’s systolic blood pressure, and “DiaBP” represents their diastolic blood pressure. While these values are closely related, there is one key difference between the two. Systolic blood pressure measures the amount of force exerted on a participant’s artery walls when their heart contracts; diastolic blood pressure is the amount of force exerted when the heart is resting (Sheps).

6. Methods

This study relied upon four different machine learning techniques to analyze the data detailed in the prior section. More specifically, logistic regression models, decision trees, random forests, and support vector machines were used to classify individual observations to one of two possible groups: (1) low risk of developing coronary heart disease, or (2) high risk of developing coronary heart disease. The following paragraphs detail the construction of each of the aforementioned models.

6.1 Logistic Regression

This study began with the analysis of several logistic regression models. With that being said, it is key to note that numerous preliminary steps needed to be taken prior to constructing this classification model. More specifically, the first step in this process was to separate the dataset into two distinct groups: (1) a training subset, and (2) a testing subset. To satisfy the previous statement, a random sample consisting of 70% of the dataset was assigned to the training set; the remaining 30% was then assigned to the testing set. These subsets were then used throughout this study to construct the various classification models and compare their predictive power.

With the data now sufficiently partitioned, the logistic regression model could be constructed. To begin this process, the best subset of predictors needed to be selected from the entire dataset. To do this, a preliminary logistic regression model was constructed using all of the data and "TenYearCHD" as the response variable; the fifteen remaining variables were treated as predictors. From here, RStudio's "step function" was utilized to create a reduced model which consisted of only statistically significant predictors. Beginning with the original model detailed above, one independent variable was systematically dropped at each iteration to reduce the overall Akaike information criterion (AIC) of the model. Because AIC scores are used to determine the relative quality of a statistical model, the group of variables which minimizes this value is the best subset of predictors that can be used in the logistic regression model.

The results of the previous process indicated that the following eight variables should be used as predictors: (1) "Male," (2) "Age," (3) "CigsPerDay," (4) "SysBP," (5) "Glucose," (6) "PrevalentStroke," (7) "PrevalentHyp," and (8) "TotChol." Since this process was performed on the entirety of the dataset, the final step in this process was to fit the prescribed model to the training data. In doing this, three of the predictors that were identified using backwards stepwise selection were no longer statistically significant. As such, "PrevalentStroke," "PrevalentHyp," and "TotChol" were dropped from the model. The final logistic regression model consisted of only five predictors: (1) "Male," (2) "Age," (3) "CigsPerDay," (4) "SysBP," and (5) "Glucose." The resulting coefficients are displayed in the table below.

Intercept	**Male**	**Age**	**CigsPerDay**	**SysBP**	**Glucose**
-8.962230	0.567732	0.067032	0.017191	0.018933	0.008323

Table 1. Final logistic regression model

6.2 Decision Tree

With the logistic regression model completed, this study then turned its attention to the construction of classical decision trees. Because the dataset was partitioned into training and testing subsets in the previous portions of the analysis, there was no need to repeat this step. In constructing the decision tree on the training subset, "TenYearCHD" was once again treated as the response variable; the remaining fifteen variables were used as predictors. This information was then used to generate a decision tree with the assistance of various RStudio functions. While the decision tree that resulted from this process was not overwhelmingly large, it was imperative to prune this model to avoid overfitting the data. To do this, the Cp values of the original model were analyzed; the smallest tree falling within one standard error of the smallest xerror is a good candidate to be the optimal decision tree. After this value was identified as 0.011494, the original decision tree was pruned according to this Cp value. The optimal decision tree based upon the 70% training set is shown below.

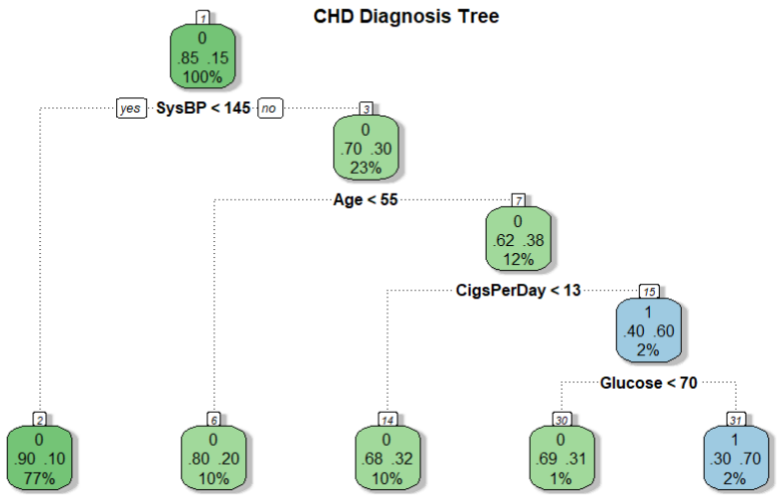

Figure 2. Optimal classical decision tree

As can be seen in the above graphic, the optimal decision tree used four variables to classify data: (1) "SysBP," (2) "Age," (3) "CigsPerDay," and (4) "Glucose." The values depicted

within each node provide important information regarding each classification. For instance, the uppermost number at each node represents the model's prediction. A value of "0" implies a healthy diagnosis/low risk of coronary heart disease, and a value of "1" represents a high risk of developing the condition. Additionally, the two values shown below this indicate the proportion of datapoints at that node that correspond to a "0" or "1" prediction. Lastly, the percent listed at the bottom represents the percent of all the data used in that node.

6.3 Random Forest

The third portion of this analysis involved generating random forests. While the underlying concepts for this type of model are quite similar to that of decision trees, the dataset needed to be slightly altered before generating this model. More specifically, the response variable ("TenYearCHD") needed to be transformed to characters. Additionally, the "mtry" value within this function was set to four. This instructs RStudio to consider four variables at the nodes of each decision tree that the model generates. Although there is not consensus regarding this concept, it is generally viewed as good practice to set the "mtry" value to the integer nearest the square root of the number of predictors used in the model. After accounting for the previously mentioned changes, a random forest model consisting of 500 decision trees was created for the training subset.

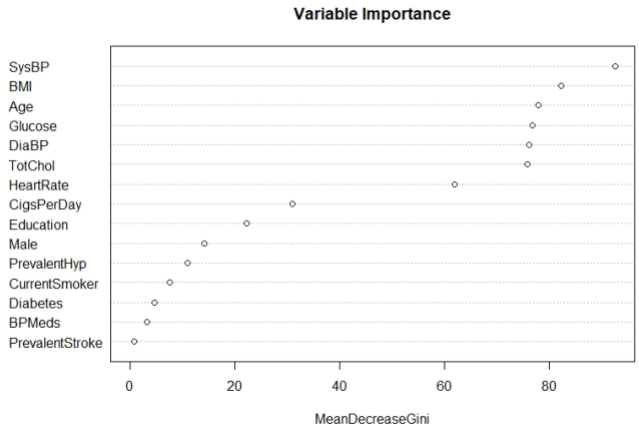

Figure 3. Variable importance in random forest model

As the above image indicates, not all fifteen of the variables included in the model were critically important. In other words, “SysBP,” “BMI,” “Age,” “Glucose,” “DiaBP,” “TotChol,” and “HeartRate” seemed to have the most influence on this model’s classification of data. This is something that should be considered when comparing the accuracy rates of each model later in this study.

6.4 Support Vector Machine

After successfully constructing the above random forest model, the next step in this study was to create and tune a support vector machine. Treating “TenYearCHD” as the response variable once again, RStudio constructed a hyperplane that separated the training data into the most homogeneous groups possible. Once this model had been created, it was tuned to achieve its maximum potential. To do this, the “gamma” and “cost” parameters were allowed to fluctuate. RStudio then tried various combinations of these parameters and returned the most accurate model. This process yielded a gamma value of 0.001 and a cost of 100.

After generating the previous classification models using a 70% training set, this study sought to obtain even more accurate results. To do this, all of the prior calculations were repeated using a 65% training subset. Because the dataset used in this study had a substantial number of observations, a larger testing set allowed each model the opportunity to classify more data without leaving the training set with too few observations to be accurate. To avoid redundancy, the calculations involving the 65% training set were not included in this paper. Instead, the models resulting from these procedures are found in Appendix A.

7. Results

The final stage of this study was to compare and contrast the predictive power of the models generated throughout the previous sections of this paper. In order to do this, each model

was used to classify the data stored within the testing subsets. Then, the actual results were compared to the predicted results using a confusion matrix. The model which yielded the highest accuracy rating was then deemed the best classification model for the given data. The accuracy rates for each of the models constructed using the 70% training data are discussed first.

Before the logistic regression model could be used to classify the data stored in the testing subset, the model was first used to obtain the predicted probability that each observation would develop coronary heart disease. From here, these predictions were divided into two groups: (1) Probability > 50%, and (2) Probability < 50%. Any observation belonging to the first group was classified as high-risk for developing coronary heart disease; the remaining observations corresponded to a healthy classification. The 50% cutoff value used in this calculation was a very natural choice since it indicated that an individual would "more than likely" or "less than likely" develop a heart condition. From here, the resulting model classifications were compared to the actual data stored in the testing subset using a confusion matrix shown below. A value of "1" indicates the presence of CHD. Of the 1,098 observations that the logistic regression model classified, 924 were correct. This corresponded to an 84.15% accuracy rate (see figure 4).

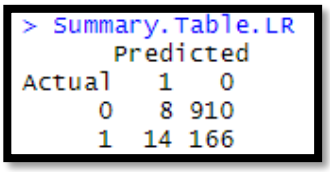

Figure 4. Confusion matrix for logistic regression model

The next machine learning technique analyzed in this study was the classical decision tree. Unlike the previous example, no additional prep-work was needed before this model could be used to classify data residing in the testing subset. With that being said, one must recall that this method yielded two different decision trees which could be used to classify data: (1) the

preliminary tree, and (2) the pruned tree. Typically speaking, pruned decision trees are more accurate; however, this is not always the case. As such, a confusion matrix comparing the actual and predicted values was generated for both the preliminary and pruned decision trees; the tables are shown below. A value of "1" implies the onset of cardiovascular disease. Interestingly, these calculations indicated that both decision trees accurately assigned data to the correct class 918 times out of 1,098 attempts. This corresponds to a respectable 83.61% accuracy rate. Since the pruned decision tree is more simplistic and equally accurate, we prefer it to the preliminary decision tree (see figures below).

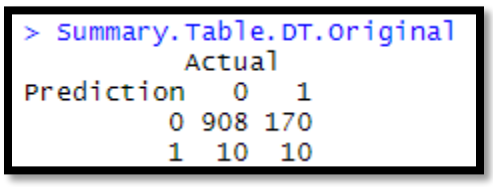

Figure 5. Confusion matrix for original tree

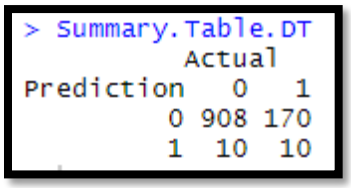

Figure 6. Confusion matrix for pruned tree

After completing the above calculations, the process was repeated using the random forest classification model. While it was assumed that this model would wield a more impressive accuracy rating than the previously discussed decision tree, this was not the case. In fact, this model performed approximately 0.1% worse. As the following confusion matrix exemplifies, this method correctly classified only 917 observations in 1,098 attempts. This translated into the least precise model generated throughout this study with an 83.52% accuracy rating. Once again, a value of "0" implied a healthy diagnosis, whereas a value of "1" implied the onset of cardiovascular disease.

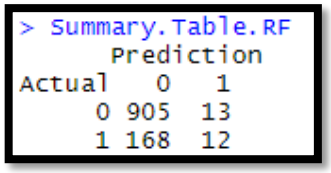

Figure 7. Confusion matrix for random forest model

The final 70% training set model needing assessed was the support vector machine. This study analyzed the accuracy rates of both the initial and tuned models to see whether accuracy improved. Furthermore, both the preliminary and tuned models were each assessed using a linear kernel and a radial kernel to classify the data. Since four distinct tables are necessary to depict each model's accuracy rate, only the confusion matrices corresponding to the final linear (left) and radial (right) models are shown below. With that being said, it is important to understand one concept before evaluating these support vector machines. When a support vector machine is created using a radial kernel, two parameters control how the model is fit to the data. As such, tuning this machine will determine the optimal combination between these two parameters which leads to the best data classification. When a linear kernel is used to develop an SVM model, however, only one parameter is used. With this knowledge in mind, interpretation of the resulting models is much easier. After being tuned, the accuracy rate of the radial model increased by roughly 0.31%. Interestingly, the final versions of these models boasted identical accuracy rates: 83.61%. Although these rates were identical, the manner in which the individual data was classified remained distinctly different (see figures below).

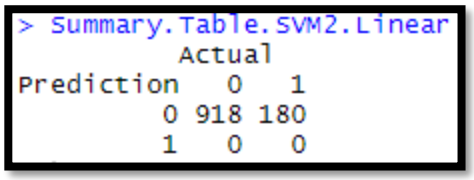

Figure 8. Confusion matrix for linear SVM

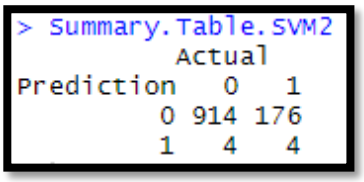

Figure 9. Confusion matrix for radial SVM

Since the calculations regarding accuracy rates for the 65% training set models are similar to those described above, this paper does not discuss this process in detail. Instead, the accuracy rates for each model can be found in the image below. Additionally, the confusion matrix for each model can be located in Appendix A under the corresponding heading.

Accuracy Rates

	Logistic Regression	Decision Tree	Random Forest	Tuned SVM
70% Training Set	84.15%	83.61%	83.52%	83.61%
65% Training Set	84.86%	84.23%	83.53%	84.23%

Error Rates

	Logistic Regression	Decision Tree	Random Forest	Tuned SVM
70% Training Set	15.85%	16.39%	16.48%	16.39%
65% Training Set	15.14%	15.77%	16.47%	15.77%

Figure 10. Accuracy and error rates for all models constructed in this study

8. Predictions

For the final portion of this study, each of the previously generated models were used to predict the probability of an individual developing cardiovascular disease. In order to do this, three fictitious patients were considered. After analyzing the original dataset, a range of typical values for each predictor was identified. Then, observations were assigned to the hypothetical patients so that one individual had low values for each risk factor, the next had typical values, and the last patient had elevated values. The models developed throughout this study were then used to make various predictions on this data. In order to avoid redundancy, this analysis will only discuss the predictions made using the models which were fit to the 65% training data (since these were the most accurate models identified above).

When viewing the predictions for each of these models simultaneously, one concept is very noticeable: the logistic regression model consistently predicted lower probabilities of developing cardiovascular disease. After this, however, the relationship between the remaining models is less clear. The identification of this link may indicate why the logistic regression model had the highest classification accuracy rate. The predicted probabilities for each of the three hypothetical patients can conveniently be compared and contrasted using the following table.

	Logistic Regression	Decision Tree	Random Forest	SVM
Low Values	2.7%	10.8%	9.2%	10.7%
Typical Values	20.5%	10.8%	14.8%	14.4%
Elevated Values	70.4%	92.3%	63.0%	46.5%

Table 2. Each model's predicted probability of developing coronary heart disease

9. Conclusions

Although the development of coronary heart disease is very complex in nature, this study showed that it is possible to generate models which accurately identify high-risk individuals. While the machine learning techniques that were assessed throughout this study had very similar accuracy rates, the logistic regression models were deemed the most precise, regardless of the training subset size. More specifically, the 70% logistic regression model accurately classified data 84.15% of the time; the 65% logistic regression model had an improved accuracy rate of 84.86%. These insights have the potential to drastically change the medical field. With access to the aforementioned models, early detection of coronary heart disease becomes possible. As a result, physicians around the world would be better equipped to combat the ongoing heart disease problem.

As the previous statements allude to, these findings support the study's initial hypothesis: logistic regression models will provide the most accurate quantification method. With that being said, it was intriguing to see that fitting each model on a smaller training set led to increased classification power. One possible explanation for this deals with the sampling technique. Because this study only analyzed the results of models based upon two different training subsets, it is entirely possible that the 35% testing set consisted of more ideal observations to classify. In

other words, the random chance associated with sampling could have slightly influenced these results so that the 65% training models were slightly more accurate.

Overall, each of the models generated throughout this study exceeded initial expectations. With that being said, there is always room for improvement. One area in particular which could prove highly beneficial is the collection of more data. Although 3,658 observations may seem substantial, more data is always useful when dealing with a complex response variable. Furthermore, it is advisable to reconduct this research using more variables. While each of the predictors used in this dataset are common risk factors which can allude to the development of coronary heart disease, this was far from an exhaustive list. As such, it is possible that the models within this study could see significant improvements with this addition.

Appendix A – 65% Training Set Models:

1. Logistic Regression Model

1.1 Model Coefficients

Intercept	Male	Age	CigsPerDay	SysBP	Glucose
-8.708082	0.531153	0.067377	0.019520	0.017988	0.007305

1.2 Confusion Matrix

```
> Summary.Table.LR2
      Predicted
Actual    1    0
     0    8 1071
     1   16  186
```

2. Classical Decision Tree

2.1 CP Pruning Values

```
> printcp(Framingham.Tree2) #use the CP of smallest tree within one xstd of least xerror

Classification tree:
rpart(formula = TenYearCHD ~ ., data = Framingham[Training.Data.DT2,
    ], method = "class")

Variables actually used in tree construction:
[1] Age        BMI        CigsPerDay Glucose    SysBP

Root node error: 355/2377 = 0.14935

n= 2377

        CP nsplit rel error xerror     xstd
1 0.010329      0   1.00000 1.0000 0.048951
2 0.010000      5   0.93521 1.0338 0.049624
```

2.2 Confusion Matrices (Preliminary Tree: left, Pruned Tree: right)

```
> Summary.Table.DT2.Original
          Actual
Prediction    0    1
         0 1050  188
         1   29   14
```

```
> Summary.Table.DT2
          Actual
Prediction    0    1
         0 1079  202
         1    0    0
```

3. Random Forest

3.1 Variable Importance Values

```
> importance(Framingham.RandomForest2, type=2)
                 MeanDecreaseGini
Male                     13.409742
Age                      73.910525
Education                20.618303
CurrentSmoker             7.025016
CigsPerDay               30.440015
BPMeds                    2.904890
PrevalentStroke           1.071584
PrevalentHyp             10.383178
Diabetes                  4.029806
TotChol                  72.010173
SysBP                    85.244356
DiaBP                    70.549742
BMI                      78.323543
HeartRate                57.086570
Glucose                  71.705883
```

3.2 Confusion Matrix

```
> Summary.Table.RF2
      Prediction
Actual    0    1
     0 1059   20
     1  191   11
```

4. Support Vector Machine

4.1 Radial Model Before Tuning

```
> summary(SVM.Fit2)

Call:
svm(formula = TenYearCHD.SVM2 ~ ., data = Framingham.SVM2[Training.Data.SVM2, ], kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  988

 ( 633 355 )


Number of Classes:  2

Levels:
 0 1
```

4.2 Radial Confusion Matrix Before Tuning

```
> Summary.Table.SVM3
          Actual
Prediction    0    1
         0 1072  200
         1    7    2
```

4.3 Radial Model After Tuning

```
> summary(SVM.Best2)

Call:
best.svm(x = TenYearCHD.SVM2 ~ ., data = Framingham.SVM2[Training.Data.SVM2, ], gamma = 10^(-5:1),
    cost = 10^(-2:2), kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  0.01

Number of Support Vectors:  710

 ( 355 355 )


Number of Classes:  2

Levels:
 0 1
```

4.4 Radial Confusion Matrix After Tuning

```
> Summary.Table.SVM4
          Actual
Prediction    0    1
         0 1079  202
         1    0    0
```

4.5 Linear Model

```
> summary(SVM.Fit2.Linear)

Call:
svm(formula = TenYearCHD.SVM2 ~ ., data = Framingham.SVM2[Training.Data.SVM2, ], kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  906

 ( 551 355 )


Number of Classes:  2

Levels:
 0 1
```

4.6 Linear Confusion Matrix

```
> Summary.Table.SVM3.Linear
          Actual
Prediction    0    1
         0 1079  202
         1    0    0
```

Appendix B – 70% Training Set Models:

1. Logistic Regression Model

1.1 Model Coefficients

Intercept	Male	Age	CigsPerDay	SysBP	Glucose
-8.962230	0.567732	0.067032	0.017191	0.018933	0.008323

1.2 Confusion Matrix

```
> Summary.Table.LR
      Predicted
Actual   1   0
     0   8 910
     1  14 166
```

2. Classical Decision Tree

2.1 CP Pruning Values

```
> printcp(Framingham.Tree) #use the CP of smallest tree within one xstd of least xerror

Classification tree:
rpart(formula = TenYearCHD ~ ., data = Framingham[Training.Data.DT,
    ], method = "class")

Variables actually used in tree construction:
[1] Age        CigsPerDay Glucose    SysBP

Root node error: 377/2560 = 0.14727

n= 2560

        CP nsplit rel error xerror     xstd
1 0.011494      0    1.0000 1.0000 0.047559
2 0.010000      4    0.9496 1.0557 0.048630
```

2.2 Confusion Matrices (Preliminary Tree: left, Pruned Tree: right)

```
> Summary.Table.DT.Original
          Actual
Prediction   0   1
         0 908 170
         1  10  10
```

```
> Summary.Table.DT
          Actual
Prediction   0   1
         0 908 170
         1  10  10
```

3. Random Forest

3.1 Variable Importance Values

```
> importance(Framingham.RandomForest, type=2)
                 MeanDecreaseGini
Male                   14.1968529
Age                    77.9491151
Education              22.1302139
CurrentSmoker           7.5307989
CigsPerDay             30.9197180
BPMeds                  3.1206762
PrevalentStroke         0.8227094
PrevalentHyp           10.8941591
Diabetes                4.6569196
TotChol                75.8425533
SysBP                  92.5906278
DiaBP                  76.1293684
BMI                    82.3038732
HeartRate              61.8655147
Glucose                76.8044469
```

3.2 Confusion Matrix

```
> Summary.Table.RF
      Prediction
Actual   0   1
     0 905  13
     1 168  12
```

4. Support Vector Machine

4.1 Radial Model Before Tuning

```
> summary(SVM.Fit)

Call:
svm(formula = TenYearCHD.SVM ~ ., data = Framingham.SVM[Training.Data.SVM, ], kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1053

 ( 676 377 )


Number of Classes:  2

Levels:
 0 1
```

4.2 Radial Confusion Matrix Before Tuning

```
> Summary.Table.SVM1
          Actual
Prediction   0   1
         0 913 178
         1   5   2
```

4.3 Radial Model After Tuning

```
> summary(SVM.Best)

Call:
best.svm(x = TenYearCHD.SVM ~ ., data = Framingham.SVM[Training.Data.SVM, ], gamma = 10^(-5:1), cost = 10^(-2:2),
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  874

 ( 497 377 )


Number of Classes:  2

Levels:
 0 1
```

4.4 Radial Confusion Matrix After Tuning

```
> Summary.Table.SVM2
          Actual
Prediction   0   1
         0 914 176
         1   4   4
```

4.5 Linear Model

```
> summary(SVM.Fit.Linear)

Call:
svm(formula = TenYearCHD.SVM ~ ., data = Framingham.SVM[Training.Data.SVM, ], kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  974

 ( 597 377 )


Number of Classes:  2

Levels:
 0 1
```

4.6 Linear Confusion Matrix

```
> Summary.Table.SVM2.Linear
          Actual
Prediction   0   1
         0 918 180
         1   0   0
```

Works Cited

"About the Framingham Heart Study." *Framingham Heart Study,* Boston University & the National Heart, Lung, & Blood Institute, https://www.framinghamheartstudy.org/fhs-about/, Accessed 20 February 2020.

"Cardiovascular Disease (10-Year Risk)." *Framingham Heart Study,* Boston University & the National Heart, Lung, & Blood Institute, https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/, Accessed 20 February 2020.

Claassen, Mette, et al. "Gender Gap in Acute Coronary Heart Disease: Myth or Reality?" *World Journal of Cardiology* vol. 4, no. 2, 2012, pp. 36-47.

D'Agostino, Ralph, et al. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study." *Circulation,* vol. 117, no. 6, 2008, pp. 743-753.

Kraus, Jess F., et al. "Socioeconomic Status, Ethnicity, and Risk of Coronary Heart Disease." *American Journal of Epidemiology,* vol. 111, no. 4, 1980, pp. 407-414.

Mahmood, Syed, et al. "The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective." *Lancet,* vol. 383, no. 9921, 2014, pp. 999-1008.

Pilote, et al. "A Comprehensive View of Sex-Specific Issues Related to Cardiovascular Disease." *CMAJ: Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne*, vol. 176, no. 6, 2007, pp. S1–44.

Randall, Otelio S., et al. *Library of Health and Living: Encyclopedia of the Heart and Heart Disease*. 2nd ed., Facts On File, 2011, pp. 172-173.

Sheps, Sheldon G. "Pulse Pressure: An Indicator of Heart Health?" *Mayo Clinic,* Mayo Foundation for Medical Education and Research, https://www.mayoclinic.org/diseases-

conditions/high-blood-pressure/expert-answers/pulse-pressure/faq-20058189, Accessed 25 February 2020.

Tsao, Connie W., and Ramachandran S. Vasan. “Cohort Profile: The Framingham Heart Study (FHS): Overview of Milestones in Cardiovascular Epidemiology.” *International Journal of Epidemiology*, vol. 44, no. 6, 2015, pp. 1800-1813.