A COMPARISON OF MACHINE LEARNING TECHNIQUES IN PREDICTING 10-YEAR RISK OF CORONARY HEART DISEASE

A THESIS
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF SCIENCE

BY

ADEOLA M. OLANIYAN

WITH DR. REBECCA PIERCE AS ADVISOR

BALL STATE UNIVERSITY
MUNCIE, INDIANA
MAY 2021

Acknowledgment

I want to thank God Almighty for the opportunity to go through this graduate program and successfully writing this thesis.

I also want to thank my project supervisor, Dr. Rebecca Pierce for her unparalleled patience and understanding throughout the course of writing this thesis. She has been both a teacher, mother, and friend. I will never forget your sacrifices and generosity towards me. Your kindness is overwhelming.

To my thesis committee, Dr. Imon and Dr. Drew Lazar. Thank you for making me a better person.

To my husband, Patrick and daughter, Helena you both give me the joy that keeps me going. Thank you for your love, support, prayers and help all the way- from the beginning of my program.

To my course mates turned sisters and friends, thank you for your support, generosity, and help. May our bond of friendship never break.

I am most grateful to my family and loved ones who helped in one way or another during the entirety of my program. I owe you a debt of gratitude.

ABSTRACT

THESIS: A Comparison of Machine Learning Techniques in Predicting 10-year Risk of Coronary Heart Disease

STUDENT: Adeola M. Olaniyan

DEGREE: Master of Science

COLLEGE: Mathematics

DATE: MAY 2021

PAGES: 40

The COVID-19 pandemic health crisis has necessitated a re-evaluation of medical health conditions. Otherwise "silent" conditions have been thrust into more awareness leading to an increase in research to identify mitigating measures. Previous studies have been carried out to develop models in predicting 10-year risk of CHD in patients using the Framingham data set. The current study is a comparison of models developed for the Framingham data set using six machine learning techniques to predict the 10-year risk of CHD. The model with the lowest test error and the highest prediction accuracy result was selected as the preferred model.

The Framingham data set is obtained from an on-going longitudinal survey in Massachusetts. The supervised machine learning techniques utilized in this study include: multivariate logistic regression (MLR), linear discriminant analysis (LDA), classification tree, bagging, boosting and random forest algorithm. Both MLR and LDA are parametric models, while the other techniques are considered ensemble methods and non-parametric. The multivariate logistic regression model was selected as the preferred model due to its lowest test error of 0.149 and 85% prediction accuracy. The selected variables include: age, gender, systolic blood pressure,

blood pressure medication, body mass index (BMI) and glucose concentration level in the body. Age, BMI, and systolic blood pressure were identified as the three most significant and recurring features in all the machine learning technique models.

The analysis carried out does not reflect the age at which either a male or female patient's systolic reading can be interpreted to be in the high blood pressure range, leading to the risk of CHD (all other significant risk factors present). Rather, it identifies advancement in age as increasing the risk of CHD.

Contents

List of Figures

List of Tables

1. INTRODUCTION

The COVID-19 pandemic health crisis presently ravaging the entire world, leaving in its wake a rising mortality rate, has necessitated a re-evaluation of medical health conditions. Otherwise "silent" conditions have been thrust into more awareness leading to an increase in research into identifying innovative mitigating measures.

Heart diseases rank as the number one cause of death globally. According to the World Health Organization's (WHO) fact sheet, a higher number of people die yearly from cardiovascular diseases (CVDs) than from any other cause. The WHO records show close to 20 million people died from CVDs in 2016, representing approximately 31% of worldwide deaths. Of these deaths, close to 90% are due to heart attack and stroke (WHO Factsheet 2017). Coronary heart disease (CHD), as a type of CVD, is a disease that affects the supply of blood to the heart. Most people are unaware of the presence of this disease with only a small percentage experiencing noticeable symptoms. At the dawn of the 20th century, approximately 10% of deaths were attributed to CVDs and at the close of the century, the death rate caused by CVDs had jumped to 25%. If the trend continues as it is, in 5 years about 50% of deaths globally would be due to CVDs. If sufficient mitigating measures are not implemented before the close of year 2020, an estimated 25 million deaths will have been caused by CVDs (Ayatollahi et al., 2019).

Several factors can be attributed to an increased mortality risk. These include obesity, lack of adequate exercise, unhealthy lifestyle, and medications. Identification of risk factors contributing to the incidence of CVDs is one of the major highlights of achievements of the present age (Pencina et al., 2009). Over the last two decades, several methods have been proposed in studies to better promote cardiovascular (CV) health, especially in the elderly and those most susceptible. In a study to identify options for improving CV conditions, (Greenlund et al. 2012) proposed options such as early identification and treatment, integrated programs to effectively manage multiple conditions, clinical-community partnerships and policy as some of the options that can help to improve the health of young and older adults.

Prediction models are of the utmost advantage to healthcare professionals and patients who must make decisions about the use of certain modes and types of treatment, changes in lifestyle, or stopping treatments altogether (Shipe et al. 2019). While not a substitute for clinical knowledge, they can provide unbiased data about an individual's disease risk and susceptibility; what is more, prediction models prevent some common biases seen in clinical decision making. CVD prediction serves as one of the most effective CVD control “tools” in the world. Logistic regression analysis is often adopted for this situation because of the binary nature of the dependent variable being analyzed. Like most regression analyses, logistic regression is a type of predictive analysis.

Although academic research is expanding the use of predictive analysis to cover 30 years, this study will use logistic regression to predict the risk of CHD within 10 years. First, several studies will be reviewed to further highlight the overall impact of using predictive analysis to determine CHD. Then this study seeks to identify the role of other risk factors contributing to the increase in the risk of CHD. Data from the Framingham study, which first proposed the concept of risk factors, will be used for determining the risk of CHD within 10 years.

2. LITERATURE REVIEW

2.1 Risk factors associated with cardiovascular heart diseases

Risk factors are conditions that increase the risk of coronary heart disease (CHD). The risk of CHD increases in people with pre-hypertensive situations depending on the number of associated risk factors present. The more the presence of these factors the higher the risk of the disease. The understanding and identification of these risk factors are essential in the treatment of heart related morbidities. Hajar et al. (2017) traced the origins of risk factors to the Framingham study published in 1957. Their study highlighted the relationship between cigarette smoking, cholesterol level and blood pressure to the incidence of CHD. Their study identified these major factors, classifying them into modifiable and non-modifiable factors. Modifiable factors are controllable and generally include obesity, high cholesterol, smoking, high blood pressure, diabetes, physical inactivity, overweight and stress. On the other hand, non-modifiable factors are not controllable and they include factors such as age, family history and ethnic background.

Results from Ibekwe et al. (2015) showed varying prevalence of hypertension from modifiable risk factors, such as alcohol consumption, smoking and obesity accounting for 15.8%, 43.4% and 18.8% respectively. Hypertension is also considered a major risk factor as it accounted for the largest number of deaths in 2009 in the United States (Danaei et al. 2009). Diabetes mellitus (DM) is another major risk factor which can be controlled. Kannel et al. (1976) identified diabetes as a major risk factor after studying the Framingham data. The study concluded that people with diabetes were more likely to die of cardiovascular related diseases than those patients without diabetes. In addition, (Fox et al. 2007) in comparing DM with other risk factors associated with CHD, reported a significant increase in the presence of DM amongst people with cardiovascular disease. Specifically, compared with other risk factors such as high blood pressure, high cholesterol and so on, only diabetes reflected an increase over the period.

Akil et al. (2011) examined the relationship between obesity and cardiovascular diseases concluding that obesity served as a precursor to other risk factors such as increased (high) blood pressure. The authors discovered that CHD morbidity and mortality were evident in people with obese conditions, observing that a 10kg increase in body weight leads to a 12% increase in the likelihood of CHD. In addition, (Algoblan et al. 2014) examined the relationship between obesity and DM concluding that there is a strong relationship between obesity and diabetes. Pencina et al. (2019) in reviewing the importance of major risk factors for heart diseases, concluded that non-modifiable risk factors significantly increased the risk of heart disease, while control of modifiable risk factors led to a reduction in the risk of CHD. Furthermore, Brown, Gerhadt and Kwon (2020) in studying risk factors associated with cardiovascular diseases concluded that modifiable risk factors have a reduced but significant role.

Although a non-modifiable risk factor such as age is regarded as an independent risk factor, its association with CHD becomes prominent as age increases as seen in the study conducted by Rodgers et al. (2019). This study analyzed and concluded that age as a risk factor is compounded by other risk factors such as diabetes and frailty. However, Dhingra and Vasan (2012) contended that while age is considered a non-modifiable factor, the risk of CHD associated with increasing age can be reduced by adjusting other well-known modifiable factors.

The findings of Pencina et al. (2009) emphasized the significance of risk factors levels in early adulthood on the long-term dangers of CVD in addition to the considerable influence of CVD risk factors on all-cause mortality. Standard CVD risk factors (e.g., gender, age, antihypertensive treatment, total and HDL cholesterol levels, smoking, and diabetes) were the significant risk factors in both the 10-year model and 30-year model. Results in the 30-year model has estimates that are almost 10-times higher than the 10-year models. For instance, there is 1.4% risk for a 25-year-old- women with adverse lipid profile and hypertension in the 10-year model, but her risk is 12% in the 30-year model. However, the models were adjusted for the competing risk of non-cardiovascular death and not for cardiovascular risk alone.

Another major non-modifiable risk factor is family history. A patient's family history plays a major contributory role in increasing the risk of CHD. Bachmann et al. (2012) in a study to examine the relationship between family history and CHD concluded that a significant relationship existed between family history and sustained increase in CHD risks over a long period of time.

Several other factors are associated with the risk of CHD. The traditional CHD risk factors which account for the most prevalent cases are presently complemented by newly researched factors. For example, Haddad et al. (2017) also classified these risk factors into modifiable and non-modifiable factors in a study of CHD among patients with schizophrenia receiving antipsychotic medication. The study revealed that non-modifiable risk factors, including a longer duration of schizophrenia illness and a history of other medical illnesses, increases the 10-year risk of CHD. Furthermore, people with serious mental illness, such as schizophrenia, bipolar, or major depressive disorders, are at higher risks than the general population without these conditions (Haddad et al., 2017).

2.2 Risk prediction for cardiovascular heart disease

Risk predictions provide an opportunity to raise awareness concerning a particular disease. Several studies have reviewed the 10-year prediction using different models. For example, a study concluded that an accurate estimate of 10-year CHD risk can be obtained using traditional risk factors and coronary artery calcium. The risk score is then used to review individual cases and offer the most applicable treatment options (McClelland et al., 2015). A model by D'Agostino et al. (2008) examined the use of a sex-specific multivariable risk factor algorithm to assess general CVD risk and risk of individual CVD history.

The progression from a 5-year risk assessment to a 10-year risk prediction has not totally removed all limitations inherent in it. For instance, age is a strong predictor in 10-year risk models sampled from populations that cover the adult age bracket. Also, slight increases in the risk factors have little significant effect on 10-year risk. These anomalies are a function of the limits imposed on the risk estimates and the 10-year duration (Lloyd-Jones, 2010).

Publications from the researchers at the Chicago Heart Association Detection Project (Lloyd-Jones et al., 2007) in Industry also give useful information on the effect of risk factors on the long-term risk of cardiovascular disease. Lifetime risks for ages up to 85 compared to risks for ages between 40-59 weighted on a 0-5 elevated CVD risk factors are balanced out. In essence, risk factors for ages up to 85 are the standard risk of death for 40-59 years. They also assessed the effect on the 30-year risk of CVD using standard risk factors for coronary and all other causes of mortality in women aged 18–39. However, their model was not designed for individual-specific risk prediction in a clinical setting (Lloyd-Jones et al., 2007). Ultimately, improvements on these limitations will continue to be sought and developed either through increased lifetime risk estimation or improved methods of estimation.

3. METHODS

3.1 Description of the data

The Framingham Heart Study is considered the first in-depth study of heart diseases in a local population. The study was started in 1948 with 5,209 participants from the town of Framingham, Massachusetts. It is considered a longitudinal cohort study which is a type of epidemiological study that follows a group of individuals over time. The study researchers tailored their research to understanding heart disease rather than on prevention methods for heart disease. The Framingham study is now on its fourth generation of participants (Mahmood et al. 2014).

The basis for clinical risk scores was established from the discussion about the definition of the risk factor. A first attempt establishes the multivariable risk feature for coronary heart disease (CHD) in Framingham. It was based on a multivariate logistic model with the 7 risk factors of age, total cholesterol, weight, ECG abnormality, hemoglobin, cigarettes smoked, and systolic blood pressure. A simple way to identify individuals as low, moderate, or high risk for potential CHD was the 10-year risk projections included in the 1998 score (Wilson et al., 1998).

The study includes demographic risk factors, behavioral risk factors, medical history risk factors, and risk factors from the first physical examination of the patient. The analysis used 4,238 observations with 16 variables. Table 3.1 identifies each variable by risk factor category and name as well as providing a description and the type of data.

Table 3.1

Risk Factors for the Framingham Data Set

Risk Factor Category	Variable Name	Description	Data Type
Demographic	Gender	Categorical variable that indicates the patient's gender (Male='M', Female='F').	Nominal
	Age	Patient's age.	Continuous
	Level of Education	Categorical variable that indicates patient’s level of education, coded as: (1) for some high school, (2) for a high school diploma or GED, (3) for some college or vocational school, and (4) for a college degree.	Ordinal
Behavioral	Current Smoker	Categorial variable that indicates if patient is current smoker or not (1 = Yes, 0 = No).	Nominal
	Cigarette per Day	Continuous variable that indicates the average number of cigarettes a patient smokes per day.	Continuous
Medical History	Blood Pressure Medication	Categorial variable that indicates if a patient was on blood pressure medication or not (1 = Yes, 0 = No).	Nominal
	Prevalent Stroke	Categorial variable that indicates if a patient had previously had stroke or not (1 = Yes, 0 = No).	Nominal
	Prevalent Hypertensive	Categorial variable that indicates if a patient was hypertensive or not (1 = Yes, 0 = No).	Nominal
	Diabetes Status	Categorial variable that indicates if a patient had diabetes or not (1 = Yes, 0 = No).	Nominal
	Total cholesterol level	Variable that indicates the total cholesterol level of patient.	Continuous
	Systolic blood pressure	Variable that indicates the systolic blood pressure of patient.	Continuous
	Diastolic blood pressure	Variable that indicates the diastolic blood pressure of patient.	Continuous
First physical examination	Body Mass Index	Variable that indicates the body mass index of patient.	Continuous
	Heart rate	Variable that indicates the heart rate of patient.	Continuous
	Glucose level	Variable that indicates the glucose level of patient.	Continuous
Response Variable	10-year risk of coronary heart disease (CHD)	Categorical variable that indicates the 10-year risk of CHD of a patient.	Binary

3.2 Description of the data analysis methods

In analyzing the data, logistic regression and selected machine learning techniques are used to estimate a 10-year prediction. The selected machine learning techniques include linear discriminant analysis, tree-based methods, bagging, boosting, and random forest. The technique with the lowest prediction, or misclassification, error will be selected to make the 10-year prediction. Short descriptions of each technique as well as descriptions of some additional ideas used in the analysis are provided next.

3.2.1 Logistic regression

To illustrate the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval or ratio-level variables, logistic regression is often used. The model can also be used to classify data. Logistic regression models the likelihood that Y (the response variable) belongs to a specific category instead of explicitly modelling the Y response (James, 2013).

For data with a binary response variable, $Y,$ and one explanatory variable, $X,$ logistic regression models probability through the logistic function:

$$\mathrm{P(x)} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

The method of maximum likelihood is applied by taking the logarithm of both sides to obtain:

$$\mathrm{Log}\left(\frac{p(x)}{1-p(x)}\right) = \beta_{0+}\,\beta_1 \mathrm{x}$$

The log-odds or logit is on the left-hand side and on the right-hand side, it is expressed as a linear function of x. This equation represents the logistic regression model.

3.2.2 Estimating the regression coefficients

The unknown coefficients β_0 and β_1 can be estimated using available training data $(x_1, y_1), \dots, (x_n, y_n)$. This can be mathematically represented by the likelihood function equation:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'=0}} (1 - p(x_{i'}))$$

The values that maximize the likelihood function with respect to observations $(x_1, y_1), \dots, (x_n, y_n)$ are then the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of population parameter $\beta_0$and β_1.

3.2.3 Multiple logistic regression

For data with a binary response $Y,$ and one or more independent variables $X = (X_1, \dots, X_p),$ logistic regression models the probability as the logistic function:

$$\mathrm{P(X)} = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_p$ can be estimated using maximum likelihood to obtain $\hat{\beta}_0$, $\hat{\beta}_1, \ldots, \hat{\beta}_p$

3.2.4 Linear discriminant analysis

Linear discriminant analysis (LDA) is a method used to classify data and minimize dimensionality. LDA easily manages a situation of unequal in-class frequencies and its outcome on test data can be randomly generated. In any specific data collection, this approach maximizes the ratio of between-class variance to within-class variance, thus simultaneously ensuring optimal separation and small variability within classes (Balakrishnama et al., n.d.).

3.2.5 Confusion Matrix

A confusion matrix is a table used to show evaluation performance. It is a binary classifier for checking if a prediction of a response matches the actual value of the response. The 2x2 confusion matrix shows the potential types of expected values in one dimension while the other dimension indicates the same for the actual values (Lantz, 2013).

The accuracy of the prediction for a 2x2 matrix can be written as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}.$$

The error rate can be written as:

$$\text{Error rate} = \frac{FP+FN}{TP+TN+FP+FN} = 1 - \text{Accuracy}.$$

The terms TP (True Positive), TN (True negative), FP (False Positive), and FN (False Negative) refer to the number of times the predictions for the model fall into each of these groups. The table below illustrates the format of a 2x2 confusion matrix..

Predicted

		1	0
Actual	1	TP	FN
	0	FP	TN

3.2.6 Decision tree

For a classification tree, each observation is predicted to be part of the most common class of training observations belonging to its region. The classification error is the fraction of the training observations that do not belong to the most common class in that region. The classification error, E, is given by the equation below:

$$E = 1 -_{k}^{Max} (\hat{P}_{mk})$$

Here $\hat{P}_{mk}$ is the proportion of training observations from the kth class in the mth region. However, the classification error E as defined above is not adequate for tree-growing, the Gini index and cross-entropy are the preferred methods. The Gini index for node m is defined by:

$$\boldsymbol{G_m} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{C}} \hat{P}_{mk}(1 - \hat{P}_{mk})$$

In the case of a binary response, where p is the proportion of 1's in the node, Gini index is:

$$\boldsymbol{G_m} = 2p(1 - p)$$

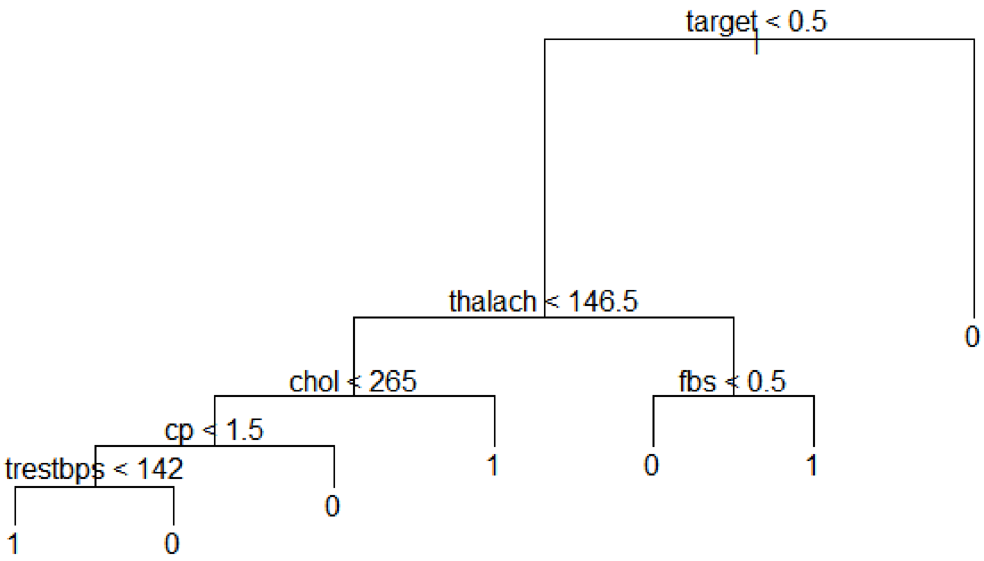
The Gini index is known as a measure of node purity — the smaller the value the greater the extent to which node m is primarily made up of a single-class of observations.

The cross-entropy is defined by:

$$\boldsymbol{D_m} = -\sum_{\boldsymbol{k=1}}^{\boldsymbol{C}} \hat{P}_k \log \hat{P}_k$$

From the formula, if $0 \leq \hat{P}_{mk} \leq 1$, it follows that $0 \leq -\hat{P}_{mk} \log \hat{P}_{mk}$. It can be illustrated that if the P_{mk}'s are all near zero or near one, the cross-entropy will take on a value near zero. Thus, like the Gini index, if node m is pure, cross-entropy will take on a small value.

When constructing a classification tree either the Gini index or the cross-entropy are usually used to determine the consistency of a particular split, as these two methods are more sensitive to node purity than the rate of classification error. Displayed below is an illustration of a classification decision tree, pruned to the best 5 features to avoid overfitting.



3.2.7 Bagging

Bagging is a statistical learning technique for reducing the model's variance. We could calculate $\hat{f}1$(x), $\hat{f}2$(x),… , $\hat{f}$B(x) using B bootstrap samples from the same training data set, then average them to obtain a single statistical learning model with low variance. It can be written as:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{\boldsymbol{k=1}}^{\boldsymbol{k}} \boldsymbol{\hat{f}^{*b}(x)}$$

Bagging is not prone to large variability, which is an issue with ordinary decision trees.

3.2.8 Boosting

The trees are sequentially cultivated with boosting, that is, each tree is cultivated using data from previously grown trees. Boosting does not require bootstrap sampling; instead, each tree matches the original data set with a changed version. The boosted formula is given as:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x)$$

A major feature of boosting is the ability to overfit B when it becomes too large. B is selected by cross validation. Even though this self-adjusting process occurs very slowly if it happens, it ensures B is selected. Also, boosting helps to decrease the model's bias. Depending on the nature of the problem, to achieve good efficiency, very small λ can require the use of a very large value of B. The complexity of the boosted ensemble is controlled by the number d of splits in each tree. An additive model is fit with each term as a single variable when d = 1, where d is the interaction depth that controls the boosted model interaction order.

Bagging and boosting are methods designed to improve the stability and the accuracy of machine learning algorithms. Combinations of multiple classifiers decrease variance, especially in the case of unstable classifiers, and may produce a more reliable classification than a single classifier.

3.2.9 Random forests

Random decision forest is a method which uses many uncorrelated trees working as a set to provide improvement to a model compare to a single tree. Here, the decision tree is built on bootstrapped training samples. A random sample of m predictors are selected as a split candidate from all the sets of p predictors in each decision tree. Each split used one of the m predictors, with a new sample of m predictors, that is, m $\approx \sqrt{p}$. The random forest method leads to a substantial reduction of variance by making each split using only a subset of the predictors.

3.2.10 Resampling methods

Resampling is a technique for estimating the test error by drawing a sample of observations from the training data then refitting the model with a randomly drawn sample of observations called the test set. Thus, the available training data is divided into test data to improve the accuracy and quantify the population parameter variability. The test validation set approach, which is a technique under the resampling method, was used. This method involves splitting the available data randomly into two components, training set and validation set. In this study, the data set of 3,656 observations (after the removal of the missing observations) were divided into fourths where ¾ (2,742 observations) is used as the training set and ¼ as the validation set (914 observations).

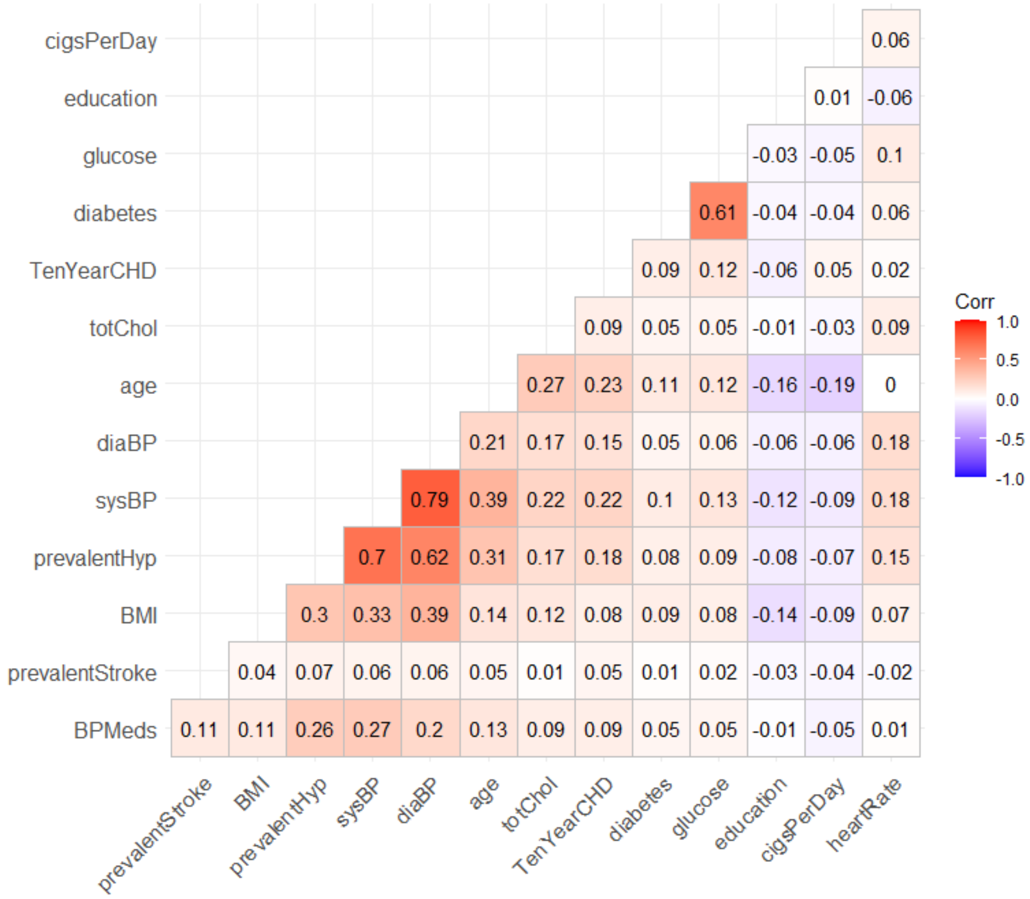
4. RESULTS

4.1 Data visualization

The data set included 1,622 men and 2,034 women with ages ranging 32 to 70 years. The average age is 50 years. The educational level of patients with the highest percentage in the study is high school diploma or general educational development (GED).

Table 4.1

Correlation Coefficient Estimates for Variables



	prevalentStroke	BMI	prevalentHyp	sysBP	diaBP	age	totChol	TenYearCHD	diabetes	glucose	education	cigsPerDay	heartRate
cigsPerDay													0.06
education												0.01	-0.06
glucose											-0.03	-0.05	0.1
diabetes										0.61	-0.04	-0.04	0.06
TenYearCHD									0.09	0.12	-0.06	0.05	0.02
totChol								0.09	0.05	0.05	-0.01	-0.03	0.09
age							0.27	0.23	0.11	0.12	-0.16	-0.19	0
diaBP						0.21	0.17	0.15	0.05	0.06	-0.06	-0.06	0.18
sysBP					0.79	0.39	0.22	0.22	0.1	0.13	-0.12	-0.09	0.18
prevalentHyp				0.7	0.62	0.31	0.17	0.18	0.08	0.09	-0.08	-0.07	0.15
BMI			0.3	0.33	0.39	0.14	0.12	0.08	0.09	0.08	-0.14	-0.09	0.07
prevalentStroke		0.04	0.07	0.06	0.06	0.05	0.01	0.05	0.01	0.02	-0.03	-0.04	-0.02
BPMeds	0.11	0.11	0.26	0.27	0.2	0.13	0.09	0.09	0.05	0.05	-0.01	-0.05	0.01

Table 4.1 shows the relationship between pairs of variables. Notable results include several strong positive correlations which are highlighted by a dark orange shade. This is followed by the light orange shade which indicates a weak positive relationship between two variables. Variables

with no relationship are identified with white shade. The light purple shade shows a weak negative relationship between two variables, while the dark purple shade indicates a strong negative relationship between two variables.

Figure 4.1.A

Systolic Blood Pressure and Age

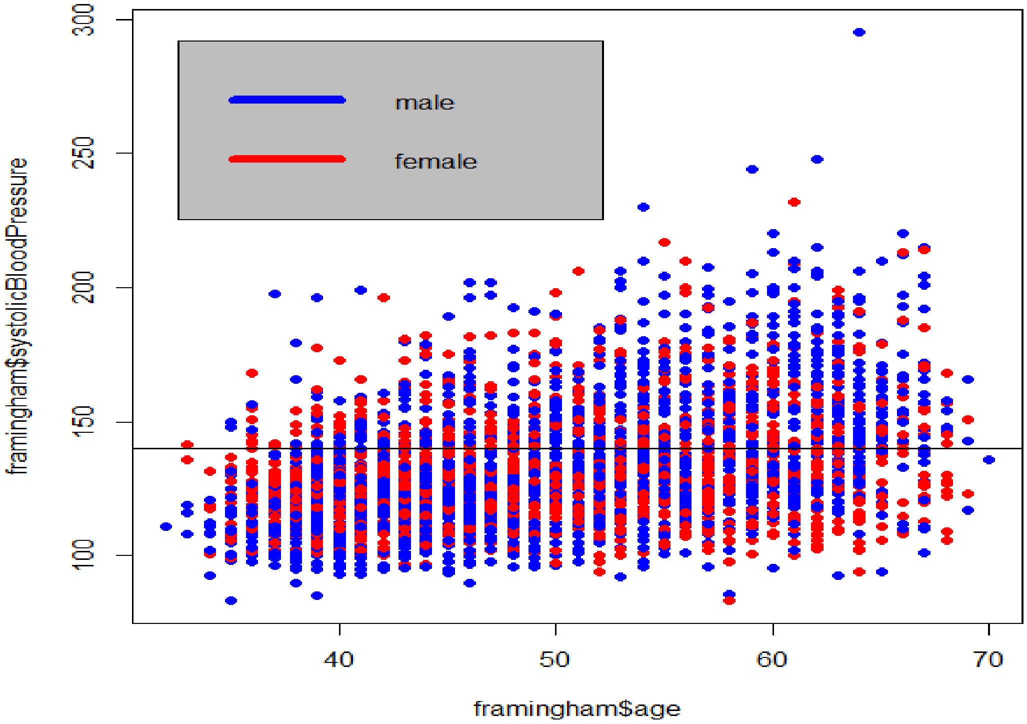


Figure 4.1.A shows a plot of systolic blood pressure versus age. The plot reveals that systolic blood pressure increases with age since the dots cover a wider range as age increases. Also, observe from the numerous blue dots above the line, the number of males with high systolic blood pressure (>140mmHg) was more than the females. Additionally, note a male patient with a systolic reading of about 300 was observed.

Figure 4.1.B

Diastolic Blood Pressure and Age

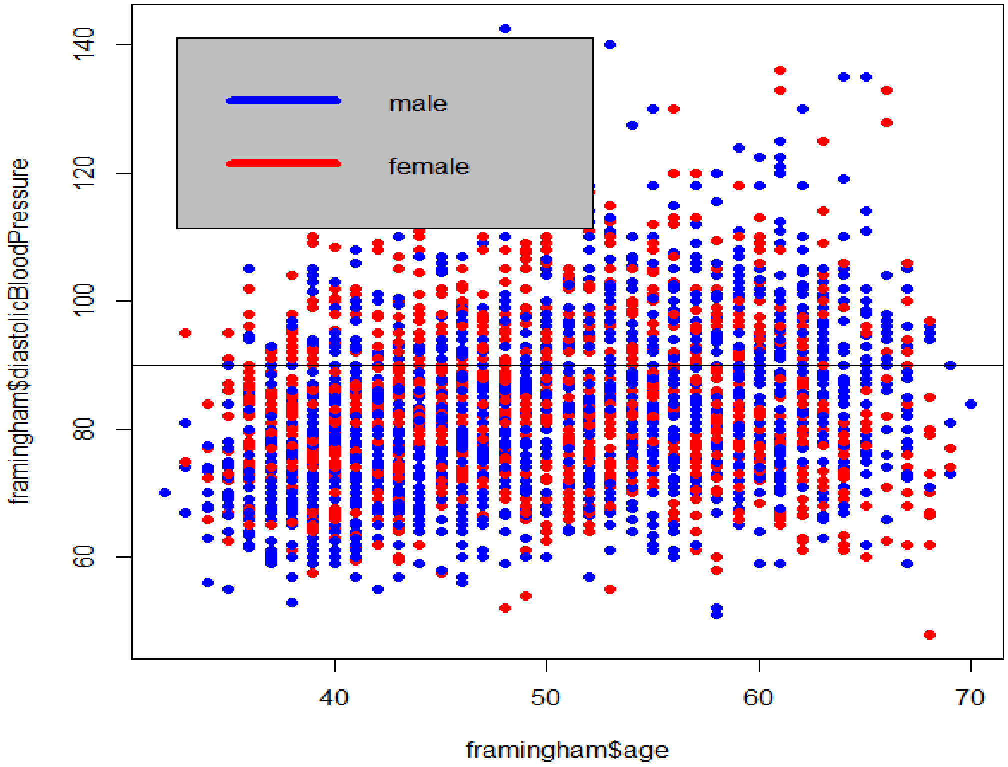


Figure 4.1.B shows a plot of diastolic blood pressure versus age. Unlike the systolic blood pressure reading, the plot shows about equal numbers for both male and female with greater than 90mmHg diastolic reading. There are two possible outliers, a male recording a little over 140mmHg and a female with a recorded reading at about 50mmHg.

4.2 Multivariate logistic regression

In predicting a 10-year risk of coronary heart disease (CHD), the multiple logistic regression model showed seven variables were significant at a p-value of 0.05 or lower. The variables include gender, age, number of cigarettes per day, BMI, systolic blood pressure and glucose level. Table 4.2A shows the specific p-value associated with each variable. The test error rate for the logistic regression is approximately 0.149.

Table 4.2A
Multivariate Logistic Regression Coefficient Estimates and P-values

Coefficients	**Estimate**	**Standard Error**	**P- value**
(Intercept)	-8.382795	0.488230	< 2e-16
Gender	0.391257	0.121698	0.001305
Age	0.072053	0.007423	< 2e-16
Cigarette per day	0.017459	0.004945	0.000415
BMI	0.028613	0.013736	0.037242
Systolic Blood Pressure	0.014432	0.002569	1.93e-08
Glucose	0.006807	0.001900	0.000339

Table 4.2A shows the coefficients that describes the risk given one of these variables. Each coefficient indicates the rate of increase in the risk of having CHD at every additional one unit of increase in variable. For example, every additional cigarette leads to an approximate increase in the odds of having the event by a factor of $e^{0.017459}$ ≈1.017 (that is, the odds go up by 1.7%). The coefficient for gender shows that males are more likely of having CHD than females by a factor of $e^{0.391257}$ ≈1.48. Also, for every 1 unit increase in BMI leads to the risk of having CHD by a factor of $e^{0.028613}$ ≈1.03. The estimated coefficients for systolic BP and glucose show minimal impact on the odds of having CHD.

4.2.1 Confusion matrices

For a threshold of 0.5, Table 4.2B shows the overall accuracy of the multivariate logistic model as approximately 85.12% ≈ [(768 + 10)/(768 + 129 + 7 + 10)]*100.

Table 4.2B

Confusion Matrix for Multivariate Logistic Model

Predicted

		1	0
Actual	1	768	129
	0	7	10

For a threshold of 0.1, Table 4.2C shows the overall accuracy of patients with 10-year risk of CHD as approximately 84.17% ≈[(117/(117+22)]*100.

Table 4.2C

Confusion Matrix for Patients with 10-year Risk of Coronary Heart Disease.

Predicted

		1	0
Actual	1	369	22
	0	406	117

4.3 Linear discriminant analysis (LDA)

While certain variables were not included in the logistic regression model due to non-significance, they were included in the LDA model. The additional variables included were diabetes, prevalent hypertension, prevalent stroke, and current smoker. For a threshold of 0.5, the overall prediction accuracy of the LDA is approximately 84.50 % ≈[(2266 + 51)/(2266+ 51 + 366 + 59)*100] which is slightly lower than the logistic regression. The test error for LDA is approximately 0.155.

4.4 The decision tree

The classification tree was grown on the training data using 500 trees with 8 variables at each split. To address overfitting, the cross-validation approach was used in cost complexity pruning the set to an optimal of 6. The overall prediction accuracy of the classification tree is 84.57%≈[(2312+7/ (2312+7+411+12)]*100. The test error rate is 0.154 which is almost the same as test error rate of LDA, but higher than the test error for the logistic regression. Figure 4.4A shows the value for the number of nodes needed for optimization to achieve the best tree pruning.

Figure 4.4A

Nodes Selection for Tree Pruning Optimization

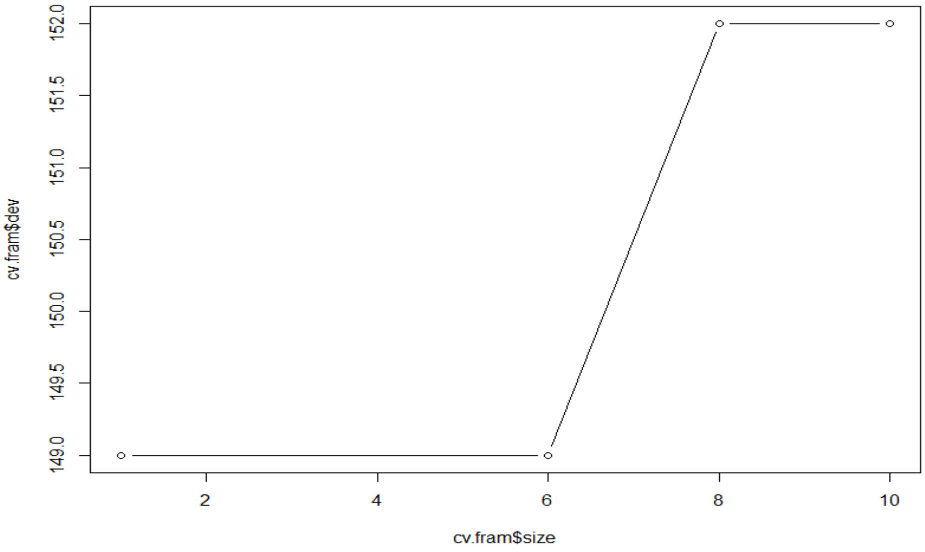
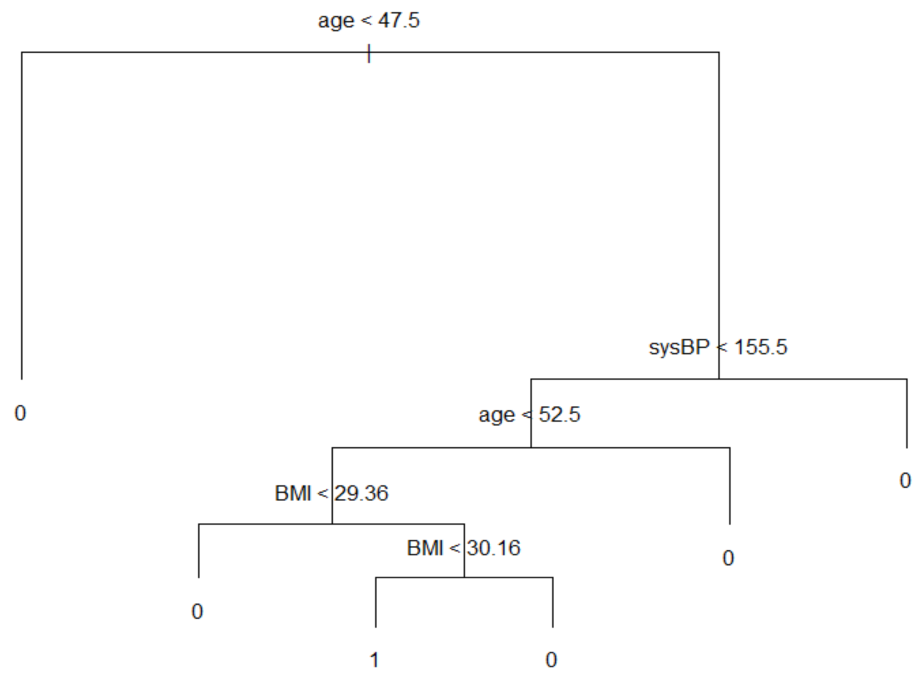


Figure 4.4B shows that for patients approximately 48 years and older whose systolic blood pressure is over 155.5 with a BMI greater than 30.16 are predicted to have a risk of 10-year CHD.

Figure 4.4B

The Classification Tree

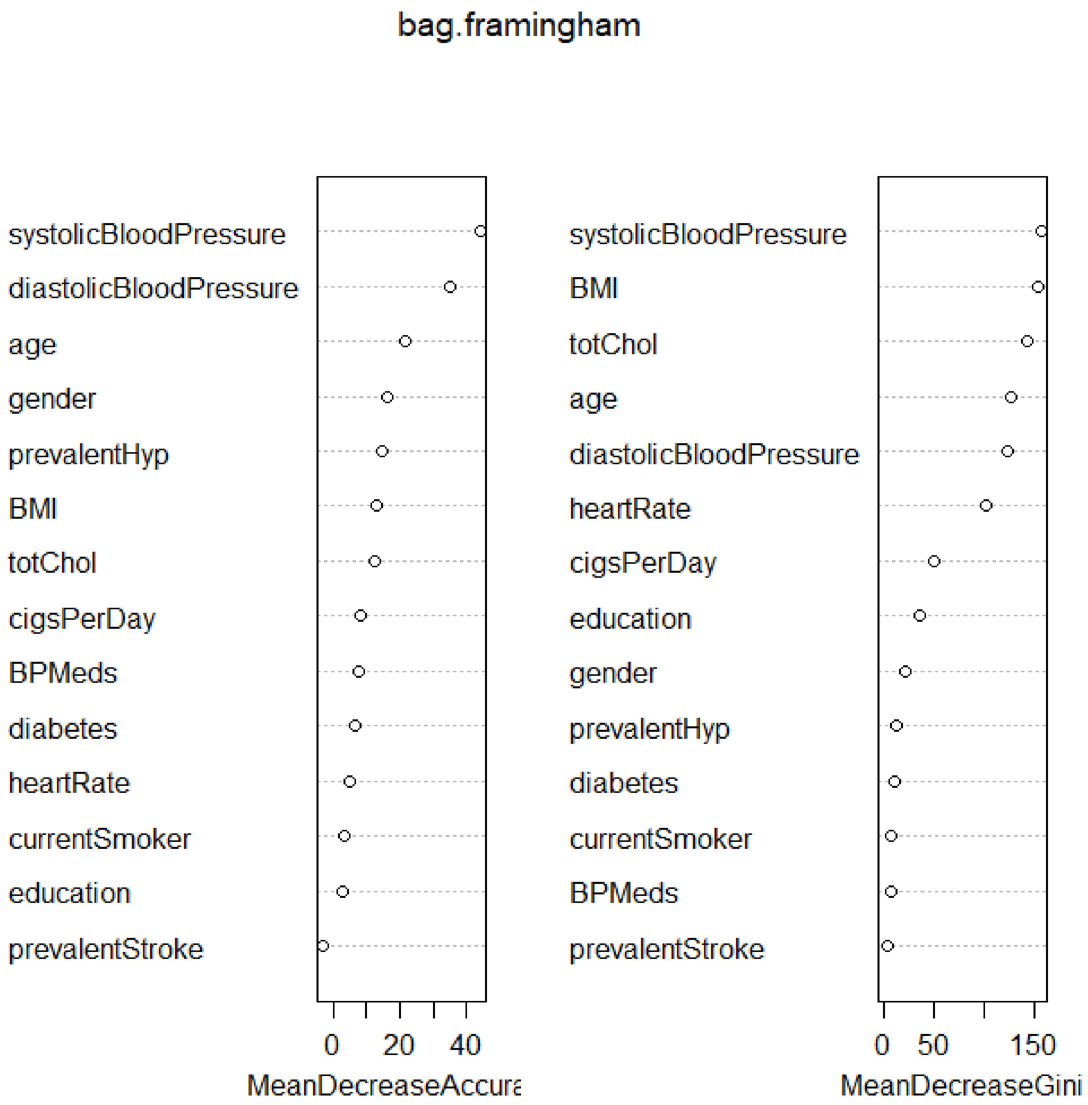


4.5 Bagging

The bagging algorithm uses all the features to find the best split in 500 trees. The bagging algorithm was built on a training data set using 500 trees with 14 variables at each split. The test error rate for the bagging algorithm was approximately 0.155, which is the same as LDA. Systolic BP, diastolic BP, age, and BMI were chosen as the important variables in predicting he response variable, 10-year risk of CHD. Figure 4.5A shows the order of the selection of these variables for the two important measures, decrease in misclassification error and decrease in node impurity as measured by the Gini Index.

Figure 4.5A

Mean Decrease Accuracy and Mean Decrease in Gini Index



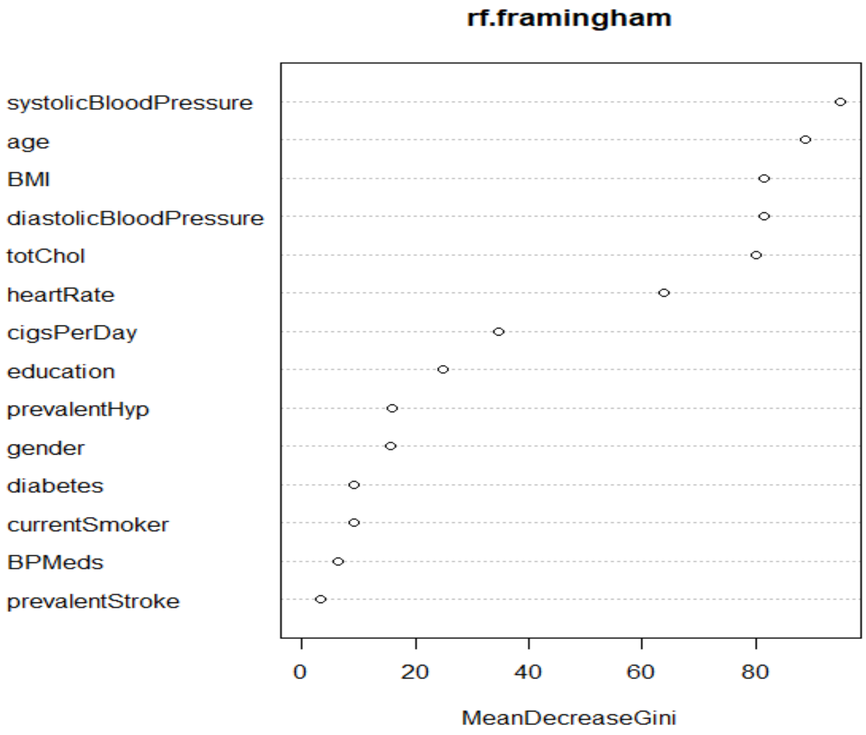
4.6 Random forest

Using the random forest method, several trials were considered to produce the iteration with the lowest misclassification error from 500 forest trees. Two iterations, otherwise called 2 weak learners, were considered to produce the misclassification error rate of 0.154. This test error rate shows that the random forest method produced a slight improvement to the classification tree than bagging.

Figure 4.6A shows that systolic BP, age and BMI were the first 3 important features for the response variable, 10-year risk of CHD. These variables were based on the indicator of decrease in the node impurity as measured by the Gini index.

Figure 4.6

The Random Forest Output



4.7 Boosting

The misclassification error rate of approximately 0.150 was estimated in the boosting algorithm. This test error rate shows that the boost algorithm is slightly better than both the classification tree and the random forest which each ad a test error test of 0.154.

Table 4.7

Descending Order of Variable Importance in Boosting Algorithm

Features	**Over-all**
Age	100.000
BMI	77.215
Systolic Blood Pressure	37.499
Diastolic Blood Pressure	37.494
Heart Rate	29.128
Total Cholesterol Level	22.826
Cigarette per day	20.877
Gender	6.382
Diabetes	6.041
Education	3.408
Blood Pressure Medication	2.176
Prevalent Hypertension	0.000
Current smoker	0.000
Prevalent stroke	0.000

Table 4.7 shows that age, BMI, systolic BP, and diastolic BP were the first 4 important features for the response variable, 10-year risk of CHD. The last three variables, which are prevalent hypertension, current smoker and prevalent stroke, show no significance in this model.

5. DISCUSSION

5.1 Relationship between identified variables

Table 4.1 shows the correlations between pairs of variables. Specifically, there is a positive high correlation coefficient of 0.79 between systolic blood pressure and diastolic blood pressure. Therefore, an increase in the systolic blood pressure may lead to an increase in the diastolic blood pressure of the patient. In addition, there is a significant relationship between diabetes and glucose with a high positive correlation coefficient of 0.61. This implies that an increase in glucose level in the body may lead to a high risk of diabetes. Furthermore, the results show a significant and positive relationship greater than 0.5 between prevalent hypertension and systolic blood pressure. Patients with high systolic blood pressure are at risk of hypertension. Likewise, patients with high diastolic blood pressure run the risk of hypertension. This is shown by the positive relationship of 0.62 between both variables. Also, from Table 4.1, there is a positive relationship of less than 0.5 between age and systolic blood pressure, BMI and systolic blood pressure, BMI and diastolic blood pressure. However, these relationships are weak.

Figure 4.1A shows the relationship between systolic blood pressure and age in both males and females, while Figure 4.1B shows the relationship between diastolic blood pressure and age of both males and females. Results from figure 4.1A indicate there were more male patients with higher systolic reading compared to females. While Figure 4.1B shows several males and females with high diastolic blood pressure as age increases.

5.2 Identification of significant variables

The various machine learning techniques identified different sets of variables as being significant for predicting the 10-year risk of CHD. The multivariate logistic regression model identified age, gender, systolic BP, BMI, glucose level, and cigarettes per day in predicting 10-

year risk of CHD. While the LDA model identified all the predictor variables to be significant in predicting the 10-year risk of CHD except diastolic BP, education, and heart rate.

Results from the other machine learning methods also varied. The classification tree shows that age, systolic BP and BMI are the most important variables in making the prediction of a 10-year risk of CHD. A BMI of 30.0 or higher is regarded as obesity (“Defining Adult Overweight and Obesity | Overweight & Obesity | CDC” 2020). The decision tree in Figure 4.4B shows that a patient with BMI greater than 30.16 is at risk of a 10-year CHD. The bagging algorithm selected age, gender, systolic BP, diastolic BP, BMI, and total cholesterol level as the important variables for predicting a 10-year CHD. Age, gender, and total cholesterol level among others were also discussed in the findings of Pencina et al. (2009) as the significant features in predicting both a 10-year risk and 30-year risk of CHD.

The random forest algorithm showed that systolic BP, age, BMI, diastolic BP, and total cholesterol level were the most important variables for predicting a 10-year risk of CHD using the important measure of the decrease in Gini Index. The boosting algorithm identifies age, BMI, systolic BP, diastolic BP, heart rate, total cholesterol, and number of cigarettes per days as significant predictors for a 10-year risk of CHD.

5.3 Selection of preferred model

Tables 5.3A and 5.3B show the test error rate and prediction accuracy, respectively, of the different machine learning algorithms. The test error rate is used to make both predictive assessment and model selection.

Table 5.3.A

Test Error Rate for the Machine Learning Technique

Machine Learning Algorithms	Test Error Rate
Multivariate Logistic Regression	0.149
Linear Discriminant Analysis	0.155
Classification Tree	0.154
Bagging Algorithm	0.155
Random forest Algorithm	0.153
Boosting Algorithm	0.150

Table 5.3.B

Prediction accuracy for the Machine Learning Techniques

Machine Learning Algorithms	Prediction Accuracy %
Multivariate Logistic Regression	85.12
Linear Discriminant Analysis	84.50
Classification Tree	84.57
Bagging Algorithm	84.49
Random forest Algorithm	84.62
Boosting Algorithm	85.01

The multivariate logistic regression model had the lowest test error rate of 0.149 and the highest prediction accuracy of 85.12%. Therefore, the multivariate logistic model was selected as the best model based on these identified conditions. However, the logistic regression is less robust than any of the ensemble methods. The analysis uses validation data for assessment from the same data set that is used to train the model. The parametric model might not do well on another data set with the same variables collected independently of this data. While you can interpret the parameter estimates in the logistic regression, the estimates might be quite different if the assumptions associated with the model are not met.

6. Conclusion

The aim of this study is to develop various statistical models using machine learning techniques or methods to predict 10-year risk of CHD. The various machine methods used included multivariate logistic regression, linear discriminant analysis, decision tree, bagging, boosting and random forest.

There were fifteen (15) predictor variables with a binary response variable. These variables were analyzed using each machine learning technique to identify the significant variables. Age, BMI and systolic BP were identified as significant and recurring variables in all the models for predicting 10-year risk of CHD. Calculation of both the test error rate and prediction accuracy were carried out to determine the most suitable model to be adopted. The multivariate logistic regression model was selected due to its lowest error rate and highest prediction accuracy results.

To summarize, accurate prediction of the risk of CHD is based on the patients age, BMI and systolic BP reading. While other variables can serve as underlying factors, these identified factors remain significant due to their high positive relationship as seen from the results of this study.

7. References

Akil, L., & Ahmad, H. A. (2011). Relationships between Obesity and Cardiovascular Diseases in Four Southern States and Colorado. *Journal of Health Care for the Poor and Underserved*, *22*(4A), 61–72. https://doi.org/10.1353/hpu.2011.0166

Algoblan, A., Alalfi, M., & Khan, M. (2014). Mechanism linking diabetes mellitus and obesity. *DMSO*, 587. 10.2147/dmso.s67400

Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health*, *19*(1). 10.1186/s12889-019-6721-5

Bachmann, J. M., Willis, B. L., Ayers, C. R., Khera, A., & Berry, J. D. (2012). Association Between Family History and Coronary Heart Disease Death Across Long-Term Follow-Up in Men. *Circulation*, *125*(25), 3092–3098. https://doi.org/10.1161/circulationaha.111.065490

Balakrishnama, S., Ganapathiraju, A., & Picone, J. (n.d.). Linear discriminant analysis for signal processing problems. https://doi.org/10.1109/secon.1999.766096

Brown, J. C., Gerhardt, T. E., & Kwon, E. (n.d.). Risk Factors For Coronary Artery Disease. Accessed December 2020

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, *117*(6), 743–753. https://doi.org/10.1161/circulationaha.107.699579

Danaei, G., Ding, E. L., Mozaffarian, D., Taylor, B., Rehm, J., Murray, C. J. L., & Ezzati, M. (2009). The Preventable Causes of Death in the United States: Comparative Risk Assessment of Dietary, Lifestyle, and Metabolic Risk Factors. *PLoS Med*, *6*(4), e1000058. https://doi.org/10.1371/journal.pmed.1000058

Defining Adult Overweight And Obesity | Overweight & Obesity | CDC. (2020, September 17). https://www.cdc.gov/obesity/adult/defining.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range.Accessed25February2021. Accessed 25 February 2021

Dhingra, R., & Vasan, R. S. (2012). Age As a Risk Factor. *Medical Clinics of North America*, *96*(1), 87–91. 10.1016/j.mcna.2011.11.003

Fox, C. S., Coady, S., Sorlie, P. D., D'Agostino, R. B., Pencina, M. J., Vasan, R. S., et al. (2007). Increasing Cardiovascular Disease Burden Due to Diabetes Mellitus. *Circulation*, *115*(12), 1544–1550. https://doi.org/10.1161/circulationaha.106.658948

Greenlund, K. J., Keenan, N. L., Clayton, P. F., Pandey, D. K., & Hong, Y. (2012). Public Health Options for Improving Cardiovascular Health Among Older Americans. *Am J Public Health*, *102*(8), 1498–1507. 10.2105/AJPH.2011.300570

Haddad, C., Hallit, S., Salameh, P., Bou-Assi, T., & Zoghbi, M. (2017). Coronary Heart Disease Risk in Patients with Schizophrenia: A Lebanese Cross-Sectional Study. *J Comorb*, *7*(1), 79–88. https://doi.org/10.15256/joc.2017.7.107

Hajar, R. (2017). Risk factors for coronary artery disease: Historical perspectives. *Heart Views*, *18*(3), 109. https://doi.org/10.4103/heartviews.heartviews_106_17

Ibekwe, R. (2015). Modifiable risk factors of hypertension and socio-demographic profile in Oghara, Delta State; prevalence and correlates. *Ann Med Health Sci Res*, *5*(1), 71. https://doi.org/10.4103/2141-9248.149793

James, G. (2013). *An Introduction to Statistical Learning (1st ed., pp. 137–140, 210–211, 251–253*). New York, NY: Springer Science & Business Media.

Kannel, W. B., McGee, D., & Gordon, T. (1976). A general cardiovascular risk profile: The Framingham study. *The American Journal of Cardiology*, *38*(1), 46–51. https://doi.org/10.1016/0002-9149(76)90061-8

Lantz, B. (2013). *Machine Learning with R: Expert techniques for predictive modeling* (pp. 541–548). Packt Publishing Ltd.

Lloyd-Jones, D. M. (2010). Cardiovascular Risk Prediction. *Circulation*, *121*(15), 1768–1777. 10.1161/circulationaha.109.849166

Lloyd-Jones, D. M., Dyer, A. R., Wang, R., Daviglus, M. L., & Greenland, P. (2007). Risk Factor Burden in Middle Age and Lifetime Risks for Cardiovascular and Non-Cardiovascular Death (Chicago Heart Association Detection Project in Industry). *The American Journal of Cardiology*, *99*(4), 535–540. https://doi.org/10.1016/j.amjcard.2006.09.099

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, *383*(9921), 999–1008. https://doi.org/10.1016/s0140-6736(13)61752-3

McClelland, R. L., Jorgensen, N. W., Budoff, M., Blaha, M. J., Post, W. S., Kronmal, R. A., et al. (2015). 10-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors. *Journal of the American College of Cardiology*, *66*(15), 1643–1653. 10.1016/j.jacc.2015.08.035

Pencina, M. J., D'Agostino, R. B., Larson, M. G., Massaro, J. M., & Vasan, R. S. (2009). Predicting the 30-Year Risk of Cardiovascular Disease. *Circulation*, *119*(24), 3078–3084. https://doi.org/10.1161/circulationaha.108.816694

Pencina, M. J., Navar, A. M., Wojdyla, D., Sanchez, R. J., Khan, I., Elassal, J., et al. (2019). Quantifying Importance of Major Risk Factors for Coronary Heart Disease. *Circulation*, *139*(13), 1603–1611. https://doi.org/10.1161/circulationaha.117.031855

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., et al. (2019). Cardiovascular Risks Associated with Gender and Aging. *JCDD*, *6*(2), 19. 10.3390/jcdd6020019

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. *J. Thorac. Dis*, *11*(S4), S574–S584. 10.21037/jtd.2019.01.25

Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, *97*(18), 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837