Advanced Analytics and Machine Learning Assignment 2

An investigation into Coronary Heart Disease using advanced machine learning techniques

Terry McElroy

40126429

May 2020

Abstract

The Framingham Heart Study was established in 1948 to conduct research into cardiovascular diseases. This seminal work allowed a deeper understanding of the causes for the number one reason for premature death. This analysis uses data from a later cohort study by the same institution. The aim of this analysis is to predict who will have Coronary Heart Disease and what the key factors are in a patient's diagnosis. This was done using many different machine learning techniques. From the analysis carried out we have found that the research that was found from the initial study in 1948 were correct with their causal links between medical health and diagnosis. The prediction into which patients had CHD was insufficient, this is down to the data's target variable being ambiguous and possibly a result from another analysis of patients that will go on to develop heart disease. More research into patients that are diagnosed would improve the predictive power from the techniques used.

Word Count: 4180

Table of Contents

1. Introduction

1.1. Research Problem

According to the UN's WHO, Cardiovascular diseases are the number on cause of death worldwide (WHO, 2017). This group of disease is a general term for conditions affecting the heart and blood vessels around the body (NHS, 2018). The most common of this disease group is Coronary Heart Disease (CHD), with this status it also makes it the largest killer worldwide (British Heart Foundation, n.d.). However, CHD is preventable through lifestyle changes. This research is based on using a data driven approach to establish what factors are responsible and predict based on the data which patients are diagnosed with CHD.

1.1.1. Origins of Heart Disease Research

The premature death of President Roosevelt in 1945 was the catalyst that started research into heart disease. During this era heart diseases, were responsible for one in every two deaths (Mahmood et al, 2014). It is for these reasons that led to the inception of the Framingham Heart Study (FHS) that examined a cohort's health for a ten year period and was responsible for determining what was causing and what could be done to prevent cardiovascular disease (Mahmood et al., 2014; Hajar, 2017).

The initial experiment determined that the major risk factors for heart disease were obesity, high blood pressure, high cholesterol, smoking, diabetes, and being inactive (Hajar, 2017). The studies demography was of white middle class people; however, the results have been found to hold water for other ethnic groups too.

Knowing the causal factors for heart disease can allow patients to take remedial actions to lessen the effects and live with the heart disease (Cohn, 2003). Furthermore, taking these measures also limits the chance of patients without any of these symptoms developing in the future. This knowledge could then limit the amount of deaths that are experienced and remove heart disease as the deadliest disease that modern society faces.

1.2. Research Data

The data that will be used for this research is from the Framingham Heart Study (FHS). This study is a long-term study to identify characteristics in the participants to establish a trend between lifestyle and medical history that correspond to a CHD diagnosis (FHS, n.d.). The data that is being used is accessible from Kaggle (Ajmera, 2017). Table 1 below shows the data variables that are being used in the analysis along with the data type and a brief description.

Variable	Type	Description
Male	Binary	1 = True (i.e. Male) 0 = False (i.e. Female)
Age	Numeric	Age of the patient
Education	Category	What level of education they have (1- High School Drop-out; 2- High School Grad; 3- Vocational College Graduate; 4- University Graduate
Current Smoker	Binary	1 = Patients smokes; 0= patient **does not currently** smoke
Cigs per Day	Numeric	The average number of cigarettes the patient smokes every day
BPMeds	Binary	Blood Pressure Medication; 1 = yes; 0= No
Prevalent Stroke	Binary	1= yes, they have strokes; 0= No they do not have strokes
Prevalent Hypertension	Binary	1= yes, they experience hypertension; 0= No they do not experience hypertension
Diabetes	Binary	1 = Yes, they have diabetes; 0= No they do not have diabetes
Total Cholesterol	Numeric	The total amount of both good and bad cholesterol measured in mg/dL (milligrams per decilitre)
sysBP	Numeric	Systolic Blood Pressure; pressure blood exerts on arteries when injected by the heart measured in mmHg (millimetre of Mercury
diaBP	Numeric	Diastolic Blood Pressure; pressure blood exerts arteries between heart beats measure in mmHg
BMI	Numeric	Body Mass Index
Heart Rate	Numeric	Number of heart beats per minute
Glucose	Numeric	Concentration of glucose in the blood by ratio to volume measured in mg/dL
Ten-year CHD	Binary Target	Does the person develop Coronary Heart Disease within 10 years? 1= yes 0= No

Table 1. Data dictionary for the Framingham heart Study data (Ajmera, 2017).

2. Methodology

The aim of this analysis is to be able to predict the patients that have CHD, and to establish which of the medical data is key to establishing patients with the disease. There are many analytical methods that can be used to determine both queries. To begin with, predicting the patients that have this aliment will be carried out by two methods; Logistic Regression and Linear Discriminant Analysis (LDA), both methods are well suited in classification problems. To establish the causal links, bootstrap aggregation will be used.

2.1. Logistic Regression

Logistic regression is a branch off from linear regression. This method uses the linear regression equation but applies it in a logarithmic form (Equation 1). Since the target of this method is to be placed into a category there is not a definite answer that it belongs to one category over another, however, logistic regression uses probability to determine the likelihood that it will belong to a certain category (Figure 1, James et al., 2013).

$$P(y) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}$$

Equation 1. The logistic regression equation

Figure 1 shows an example on defaulting on loan payments. The x axis shows the probability of default and the y axis show the predicting variables (in this case it is just account balance). If the probability is 0.5 or greater then this shows the maximum likelihood and the person will default on there loan. However, if lower than 0.5 then the maximum likelihood states that they will not default on there loan. This cut-off probability can be altered to make the model more accurate.

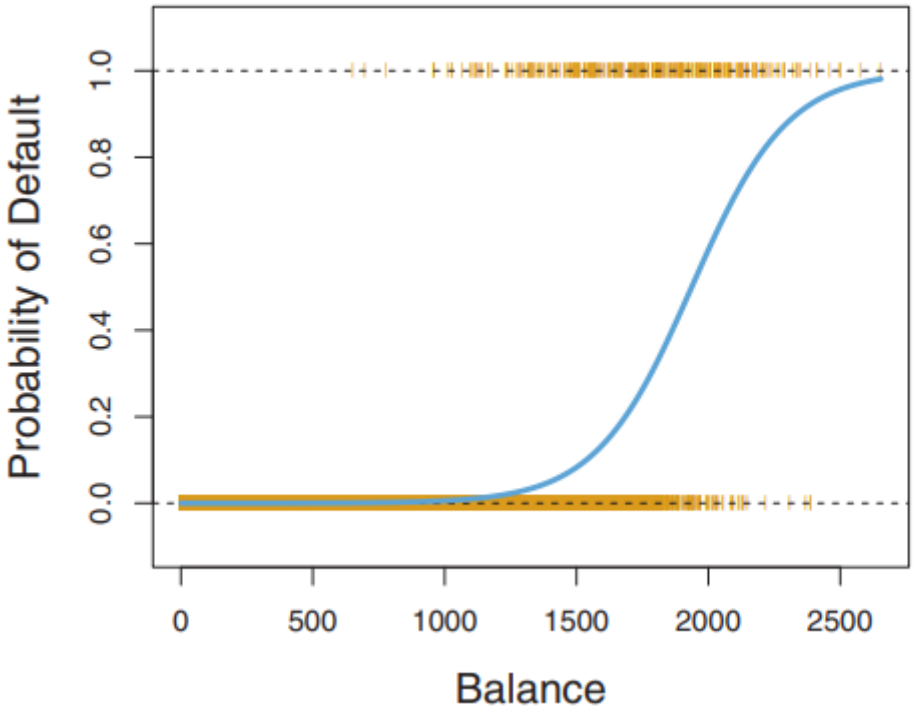

Figure 1.This shows an example of logistic regression and how it determines the category that it belongs to based on the probability. Figure replicated for James et al. (2013, Pg. 131)

2.2. Linear Discriminant Analysis (LDA)

This method is another way to determine which class a row of variables will likely belong too. It does by discriminating the data to those that belong to a certain category (Equation 2). The category is determined by the target variable. For this method to work then the distribution of the predictors (the x terms in equation 1 and 2) must/can assumed to be normal. This then allows the use of Bayes Theorem to work out the probability that it belongs to a certain category (James et al, 2013). This is a very similar approach to Logistic regression.

$$D = v_1 x_1 + v_2 x_2 + \cdots + v_n x_n + a$$

Equation 2. The discriminant function that LDA uses

This will affect the data by dividing the entries into different camps based on the predictor values (Figure 2). This works by reducing the mean for each category whilst maximising the distance between the means between each category. This will also result in the dimensionality of the problem to be reduced and will group values for predictors to be enclosed within a certain group (James et al., 2013; Mahapatra, 2018).

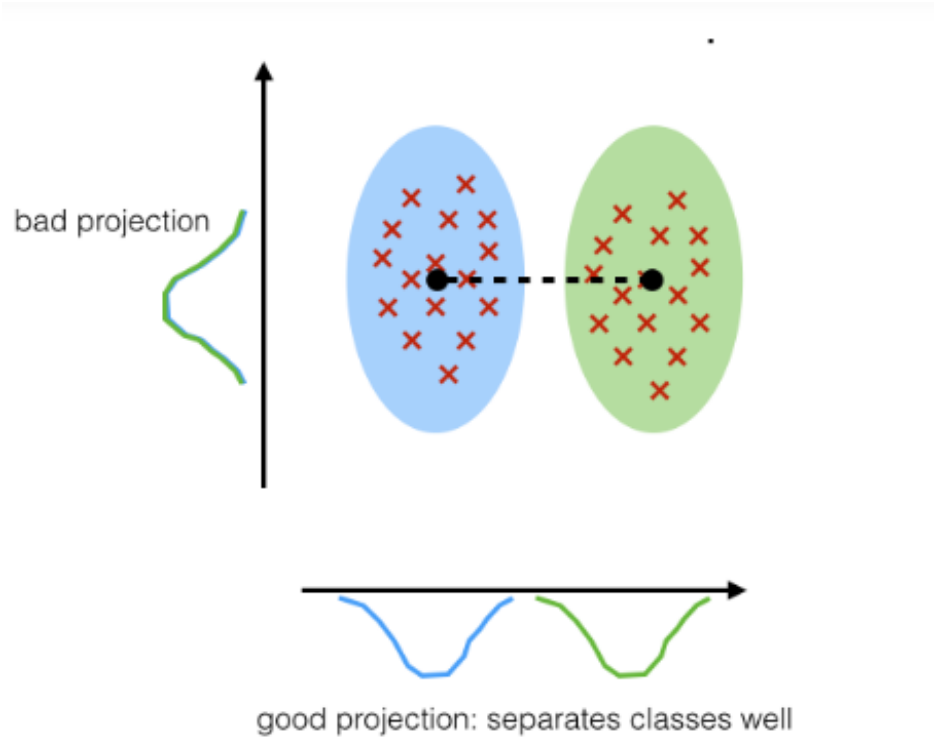

Figure 2. A graphical representation of LDA. Figure replicated from Mahapatra (2018).

2.3. K-Fold Cross Validation

This is a resampling method that is used to establish a testing of the predictive models. It does this by dividing the data into many, k, subgroups known as folds (Figure 3). This then allows the predictive model to build the data based on k-1 folds and test the accuracy with the remaining fold. This process is similar to the validation set approach. However, this method repeats the building and testing of the data k times and gets an average accuracy for the process. By doing so it allows each fold to be the test data at some point. By employing this technique, it will reduce the bias in testing the model that would be experienced through a simple validation set approach (Hastie et al., 2009; James et al., 2013). Common k-values for this type of cross validation are 5 or 10.

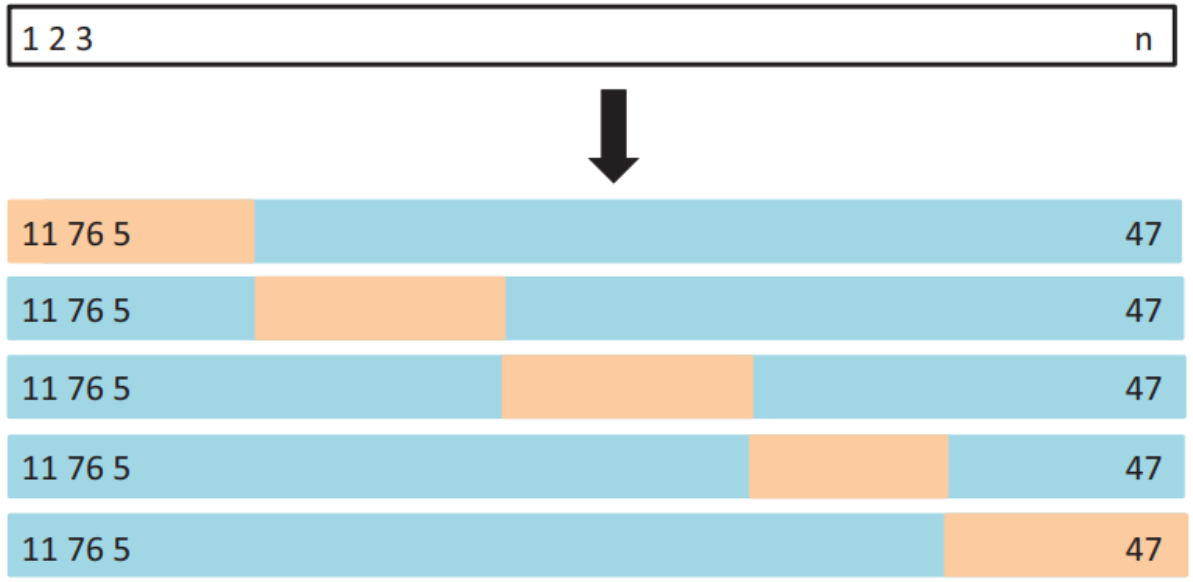

Figure 3. This is a representation of how the testing group changes throughout a cross validation approach. Figure replicated from James et al. (2013, Pg. 181).

2.4. Bootstrap Aggregation (Bagging)

This is an ensemble method, uses tree models. Bagging is a method that is used to reduce the variance of tree models. In R this method can be used to create a model that shows the most influential variables in the prediction for a classification problem. Bagging essentially divides the data into several smaller data frames that it will build and test a model on each frame. This process will repeat itself until all data frames have been tested, similar to cross validation. The trees that have been built are not pruned so have a low bias but large variance, however averaging out over hundreds of trees enable the variance to be reduced and gives a model with both low variance and bias. This is something that is wanted from a model for predictions. By creating several hundred trees it reduces the interpretability of the model, however, it is possible to establish the influential factors in the prediction. This gives us the ability to establish which factors are most important in the prediction. (James et al., 2013; Brownlee 2019)

2.5. K-Means Clustering

This is an unsupervised learning technique. Clustering will take the data and segment observations into a set number of clusters. For K-means the number of groups must be stated. Once the number of clusters is known the algorithm will establish a centre point for each cluster. After this the algorithm will use an iterative process to segment data to their closest cluster centre. Preforming this type of analysis allows insight into how the data can be grouped (Figure 4).

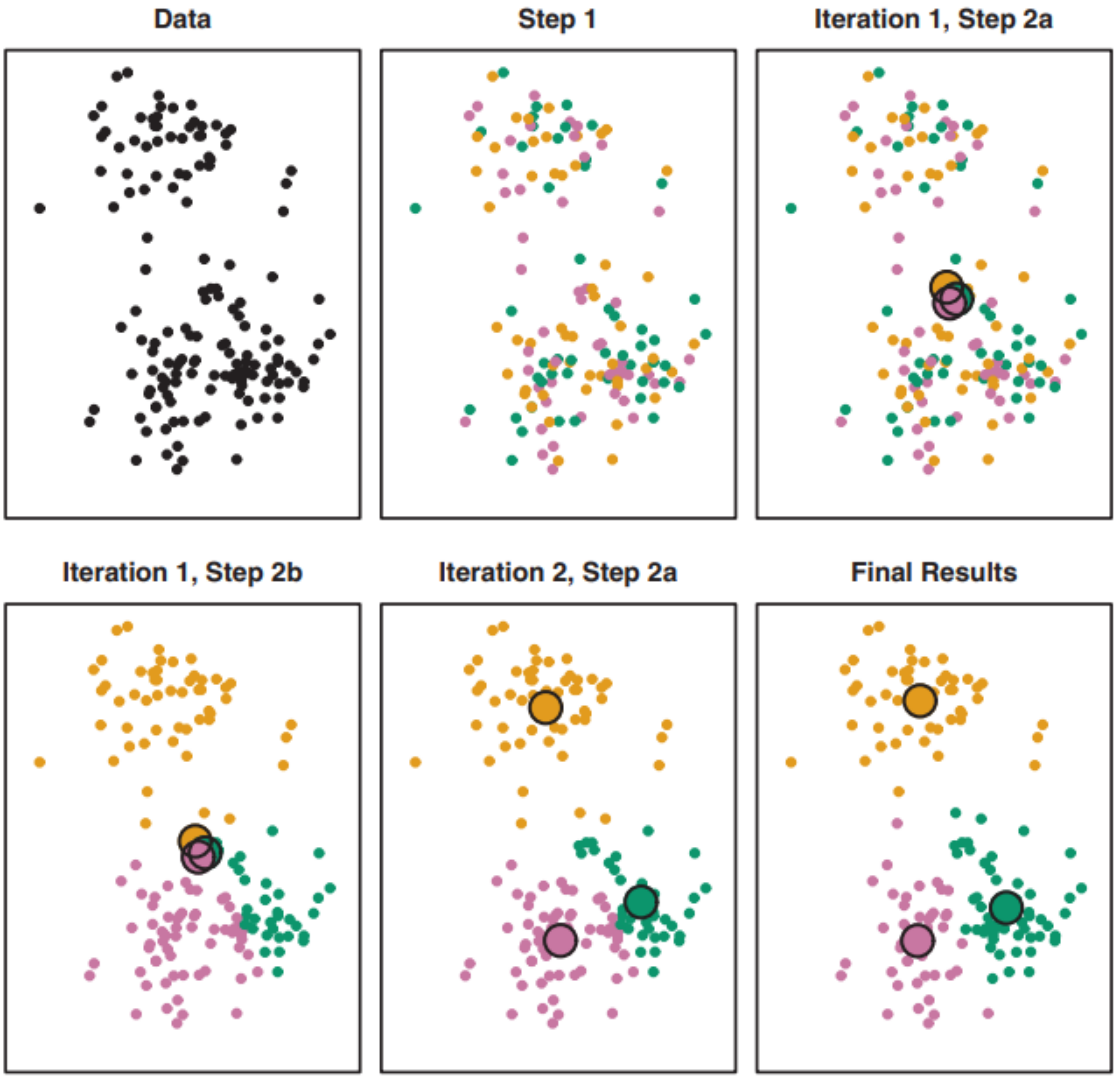

Figure 4. A visual representation of how the clustering algorithm works. Figure replicated from James et al. (2013, Pg.389)

3. Data Quality

Before the analysis can be carried out a check on the data quality must happen. This step ensures that the data that is being used is not going to have any outliers that influence the model towards them instead of the majority of the data.

3.1 Data Characteristics

As mentioned in 1.2, the data that is being used is from the Framingham Heart Study; the data is made up of 16 variables (see table 1) for 4238 patients. The patient data is based on personal and medical data that relates to the patient. The objective is to use the personal and medical data to predict whether they are a sufferer of Coronary Heart Disease. Table 2 below shows the summary statistics for the variables.

Variable	Summary Statistics				Missing Values
Male	**1**		**0**		**0**
	1819		2419		
Age	**Minimum**	**Mean**	**Maximum**		**0**
	32	49.58	70		
Education	**1**	**2**	**3**	**4**	**105**
	1720	1253	687	473	
Current Smoker	**1**		**0**		**0**
	2094		2144		
Cigs per Day	**Minimum**	**Mean**	**Maximum**		**29**
	0	9	70		
BPMeds	**1**		**0**		**53**
	124		4061		
Prevalent Stroke	**1**		**0**		**0**
	25		4213		
Prevalent Hypertension	**1**		**0**		**0**
	1316		2922		
Diabetes	**1**		**0**		**0**
	109		4129		
	Minimum	**Mean**	**Maximum**		**50**

Total Cholesterol	107	236.7	6960	
sysBP	**Minimum**	**Mean**	**Maximum**	**0**
	83.5	132.4	295	
diaBP	**Minimum**	**Mean**	**Maximum**	**0**
	48	82.89	142.5	
BMI	**Minimum**	**Mean**	**Maximum**	**19**
	15.54	25.8	56.8	
Heart Rate	**Minimum**	**Mean**	**Maximum**	**1**
	44	75.88	143	
Glucose	**Minimum**	**Mean**	**Maximum**	**388**
	40	81.97	394	
Ten-year CHD	**1**		**0**	**0**
	644		3594	

Table 2. Table of summary Statistics for the Framingham Heart Study data

From the table it is not overly obvious which variables have outliers that could potentially skew the model to these points. To correct for this, a graphical approach can be used. It is important to note that some of the modelling techniques do not allow missing data, this will result in any row of data with missing data to be omitted from the building and testing of the model. In the case of omitting all rows with missing data there is now 3656 patients being used to develop certain models.

3.2. Identifying outliers

To efficiently identify outliers the variables can be plotted in a histogram, the default plotting that R uses will automatically develop the axis'. If the x axis for a variable continues for long after the obvious bars, we can conclude that there is an outlier in this variable. Furthermore, algorithm 1 and 2 can be used to remove any outliers. Algorithm 1 is a clamp transformation that establishes an accepted minimum and maximum value that is +/- 3 standard deviations. If the data is outside these boundaries, then it is overwritten to become the accepted minimum or maximum value. This overwriting will only effect 0.27% of the data so it will not reduce the predictability of the models that are built from it. Algorithm 2 is then able to compare the variable when the outliers are included and not, this will then allow a judgement to be made to determine if the outliers should be altered or not. The results of the process are shown in Figure 5.

```
clamped <- function(x){

  x <- as.numeric(x)
  clamp(x, lower = (mean(x, na.rm = T) - 3*sd(x, na.rm = T)),
        upper = (mean(x, na.rm = T) + 3*sd(x, na.rm = T)))

}
```

Algorithm 1. This is a clamp transformation function that limits the upper and lower bounds

```
compare <- function(x1,x2 = NULL, title = NULL){
  p1 <- hist(as.numeric(x1))
  p2 <- hist(as.numeric(x2))
  plot(p1, col = "orange", main = paste("Comparison of ",
                                        title), xlab = title)
  plot(p2, col = "green", add = T)
}
```

Algorithm 2. Graphically comparing the variables with and without outliers

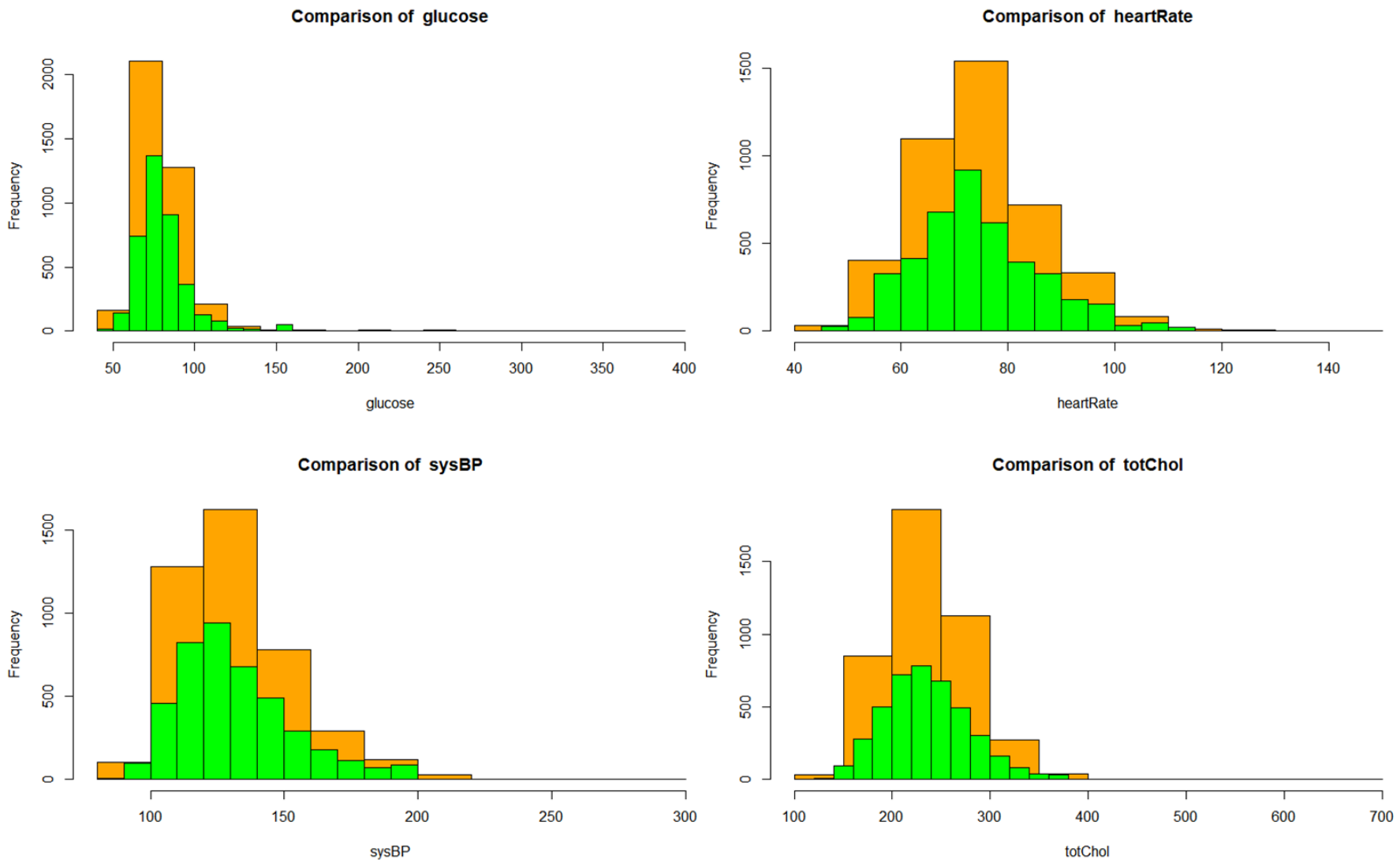

Figure 5. This shows the graphical comparison for the variables whose outliers were changed

4. Results

With outliers taken care of it is now possible to perform the analysis to answer the research problem; Can we use the data to predict patients with heart disease and what factors are most influential on the target of do they develop heart disease? To begin the analysis the prediction models will first be carried out. After a reasonable model has been developed an analysis on influential factors can be compiled, after this we should be able to conclude that we are able to predict persons with heart disease and be able to give remedial advice on how to prevent the development of coronary heart disease.

4.1. Logistic Regression

The research question regarding whether a patient develops CHD is a two-response class problem, either the patient has CHD, or they do not, there is no in-between. This then allows logistic regression to be used to build a model to predict this outcome. However, before a comprehensive logistic regression model can be built it is advisable to find the p-values to determine whether an individual variable has a relationship with the target. The p-value for the variables must be less than 0.05 to make the conclusion that there is a relationship between the variables. If the variable has a relationship, then the alternative hypothesis will be accepted and if there is no mathematically relationship then the null hypothesis is accepted (no mathematical relationship). Table 3 shows the result of this test along with the testing method.

Variable	Method	P-Value	Accepted Hypothesis
Male	Chi-squared	5.00E-04	Alternative Hypothesis
Age	T-test	1.04E-48	Alternative Hypothesis
Education	Chi-squared	5.00E-04	Alternative Hypothesis
Current Smoker	T-test	2.06E-01	Null Hypothesis
Cigs per Day	T-test	5.03E-04	Alternative Hypothesis
BPMeds	Chi-squared	5.00E-04	Alternative Hypothesis

Prevalent Stroke	Chi-squared	5.00E-04	Alternative Hypothesis
Prevalent Hypertension	Chi-squared	5.00E-04	Alternative Hypothesis
Diabetes	Chi-squared	5.00E-04	Alternative Hypothesis
Total Cholesterol	T-test	6.05E-07	Alternative Hypothesis
sysBP	T-test	4.92E-32	Alternative Hypothesis
diaBP	T-test	6.42E-16	Alternative Hypothesis
BMI	T-test	8.45E-06	Alternative Hypothesis
Heart Rate	T-test	1.29E-01	Null Hypothesis
Glucose	T-test	1.05E-06	Alternative Hypothesis

Table 3. This shows the P-Values for each variable with respect to the ten-year CHD variable. T-tests are used for the continuous data whereas the chi-squared methods are for the categorical data

These results were then verified by doing an individual logistic regression between Ten-year CHD and each variable. Now that we have the p-values we can conclude that heart rate and current smoker do not have a link to CHD, and it would not make sense to include them in the model. The first model includes all variables except the two mentioned. This results in a model that returns a very rudimental model where most of the variables are not being used in the prediction, this is shown by their p-values in figure 6. To correct for this stepwise selection can be used to determine the variables that work best together to model the outcome. By using this method, the variables that are included in the model now are; Male, Age, Cigs Per Day, sysBP, Diabetes, Prevalent Stroke, and BPMeds. With this new and improved formula all the variables are responsible for the building of the model. Using the predict function the accuracy of the model is 85.16% whenever the probability cut-off is at the optimal level of 0.51 (figure 7).

```
Call:
glm(formula = All.Form, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4111  -0.5940  -0.4257  -0.2844   2.8698

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.430537   0.679069 -12.415  < 2e-16 ***
male1             0.533287   0.108947   4.895 9.83e-07 ***
age               0.062627   0.006722   9.317  < 2e-16 ***
education2       -0.184584   0.123026  -1.500 0.133520
education3       -0.179094   0.149217  -1.200 0.230054
education4       -0.061864   0.164461  -0.376 0.706798
cigsPerDay        0.019823   0.004192   4.729 2.26e-06 ***
BPMeds1           0.198477   0.231908   0.856 0.392084
prevalentStroke1  0.695271   0.490559   1.417 0.156394
prevalentHyp1     0.212586   0.139929   1.519 0.128701
diabetes1         0.344012   0.281350   1.223 0.221435
totChol           0.001991   0.001168   1.704 0.088306 .
sysBP             0.015470   0.004003   3.865 0.000111 ***
diaBP            -0.002974   0.006352  -0.468 0.639573
BMI               0.003553   0.012660   0.281 0.778959
glucose           0.007010   0.003298   2.126 0.033525 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. This is the console output for the first logistic regression model.

```
          target
pred.fit    0    1
       0 3567  604
       1   27   40
[1] 0.851109
[1] 1.421189
```

Figure 7. The console output for the logistic regression model with the improved formula. This shows the confusion matrix of predicted to actual and shows the accuracy and the MSE

However, the purpose of the model development is to make predictions on fresh data that is unseen. To do this a validation test approach was used to partition the data into testing and training data sets at a ratio of 75/25 respectively. Whenever this is done a prediction can be made that properly tests the model. Using the same formula as before we get a testing accuracy of 83.68% whenever the probability cut-off is optimised at 0.42 (figure 8). Whilst this is not as accurate as before it is making predictions on previously unseen data, so it is quite useful. To better the prediction again k-fold cross-validation can be used. For this analysis K=10 was used for the fold number. This resulted in the accuracy of this model being 84.98% and with a kappa value greater than zero this shows that it is better than random chance.

```
          target
pred.fit   0   1
       0 864 166
       1   7  23
[1] 0.8367925
[1] 1.463208
```

Figure 8. The console output for the logistic regression with a validation set approach.

4.2. Linear Discriminant Analysis

The interest of the research question is to make the most accurate model possible. To so LDA is also applied to the data to see if there will be a more accurate prediction, thus resulting in an ability to diagnose patients earlier. With LDA there is no need to find p-values when trying to build a model, so it is fine to include all variables in the model. However, for LDA to work there cannot be any missing values. This results in the data only having 3656 patients in the training data. Building the model with these conditions gives an accuracy of 85.28% whenever the probability cut-off is optimised at 0.38 (Figure 9)

```
          Target
predProb    0    1
       0 3092  531
       1    7   26
[1] 0.8528446
[1] 0.1471554
```

Figure 9. The console output for the first LDA model, accuracy of 85.28% with MSE of 0.1471.

Again, the model needs to be able to predict based on unseen data. To do so simple a validation test will be used to produce the most accurate formula before using cross validation. Whenever the formula is kept the same however the training and testing data are used it yields an accuracy of 82.99% with the optimal cut-off probability at 0.56 (Figure 10).

```
          Target
predProb   0   1
       0 740 144
       1  12  21
[1] 0.82988
[1] 0.17012
```

Figure 10. The console output for LDA using a validation test for all variables, accuracy of 82.99% with MSE of 0.1701.

Changing the formula that is inputted into the model can improve the accuracy. Using the formula, the most accurate logistic regression. By doing so and using the validation test approach we get an accuracy on predictions of 83.09% whenever the optimal probability cut off is 0.51 (Figure 11). This model is more accurate than the previous model and has a lower MSE too, this then allows the K-fold cross validation to be carried out on this model. When this is preformed the accuracy of the 88.46%. This is the most accurate model that has been achieved for this dataset.

```
          Target
predProb   0   1
       0 748 151
       1   4  14
[1] 0.8309706
[1] 0.1690294
```

Figure 11. The console output for LDA with changed formula and a validation test approach, accuracy of 83.09% with MSE of 0.169.

4.3. Determining the Causal Factors for CHD

The next part of the research question is; what factors are most influential in determining a diagnosis of CHD. As described in 2.4 bagging is the development of a random forest to develop a model for prediction. However, in this analysis we are not worried about the predictive ability, we want the most influential variables and since hundred of trees are produced in this model it will allow them to be easily assessed. Using the omitted data and all the variables allows the figure 12 to be developed. With this we can also query to see the importance for each class (Table 4). From this we can conclude that the variables that are required are Blood pressure, age, gender, cholesterol and BMI, to make a prediction. This would confirm what has been found in at the start of the Framingham Heart Study (Hajar, 2017). To make a prediction on those that are diagnosed with CHD we need Age, Gender, glucose level, Systolic blood pressure, and Prevalent Stroke.

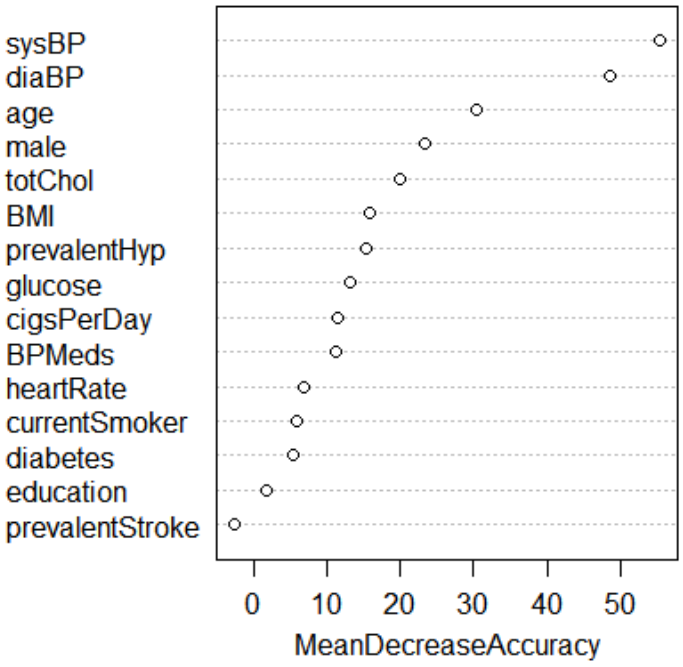

Figure 12. This is the bagging graph and shows how removing each variable will affect the overall accuracy.

Variable	0	1	Mean Decreased Accuracy	Mean Decrease Gini
Male	19.91121	11.10154	23.38705	16.77702
Age	16.15909	38.06771	30.4031	117.6559
Education	2.813384	-2.0726	1.688471	33.93391
Current Smoker	5.369255	-0.42774	5.759774	5.488631
Cigs per Day	14.56458	-4.80041	11.50795	47.37361
BPMeds	11.15531	1.226838	11.24732	4.382028
Prevalent Stroke	-2.95164	0.148164	-2.65549	2.497469
Prevalent Hypertension	15.70315	-5.58254	15.27073	7.029353
Diabetes	5.521826	-0.36806	5.378409	2.293376
Total Cholesterol	24.05151	-5.13219	19.92475	123.4634
sysBP	52.64092	1.556657	55.46612	135.6719
diaBP	51.23534	-18.591	48.66016	107.1871
BMI	20.35877	-7.70855	15.8236	132.7848
Heart Rate	8.427519	-1.37016	6.946951	90.29853
Glucose	9.225303	11.03234	13.16102	116.8193

Table 4.Table displaying the importance for each variable under different circumstances

4.4. Clustering

The clustering the patients is to create subcategories of typical CHD patients. To begin the number of clusters, must be decided on. The best way to do this is plot a graph showing the what each clusters dissimilarity will be if that is the chosen K, the lower the dissimilarity the more alike they are (Figure 13). Knowing that diagnosed patients are optimally cluster into seven groups, it can give an insight into the outcomes of the disease. A further analysis with data that includes the outcome could allow an insight into how bad the disease will affect a person's life.

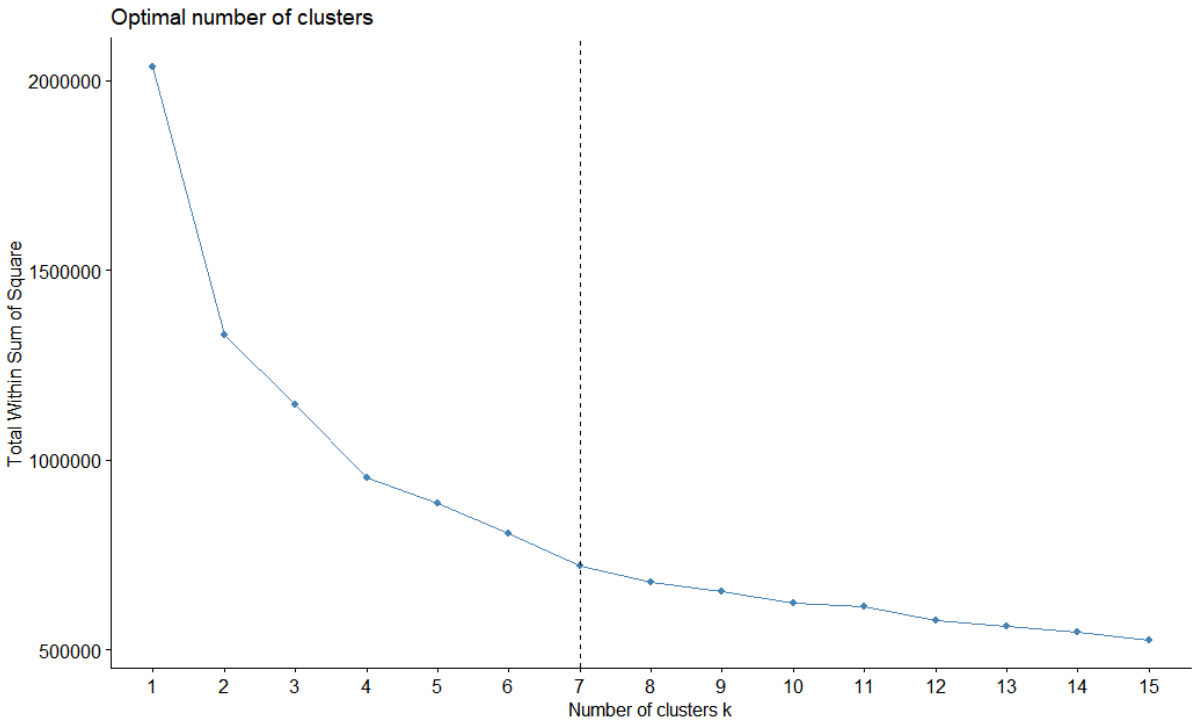

Figure 13. This is a graph showing the total within sum of squares (a measure of dissimilarity) and the different number of clusters there can be. K = 7 is the best trade off because afterwards the curve flattens out so it will not change the variance much.

5. Limitations

With the analysis there have been a few shortcomings. Firstly, the analysis for predicting whether a patient has CHD or not. The data that we have has roughly a split of 85% without to 15% with, the accuracies that has been achieved through most of the analysis is approximately 85% accurate. With further examination of the confusion matrices (Figure 7) we can see that the majority of the confirmed cases are inaccurate for predicting patient that have CHD. This could be a limit from the data that has been selected. The target variable from Ajmera (2017) is Ten-year CHD, this data may have already been used to predict whether they will develop CHD within 10 years and therefore is an ambiguous target to predict. To correct for this an analysis with data that has patients that are a definitely diagnosed with CHD could be carried out to improve predictive power. It should be noted that the analysis of the key factors will not be affected.

Furthermore, the data that has been used is very medical focused. There are studies that show that a person's lifestyle also are key contributing factors to being diagnosed with CHD. An examination of these factors in a future analysis would make the interpretation of the models a lot easier to understand. This would also allow a more approachable clustering method because people would understand where they would stand with regards to chance of developing CHD and their own lifestyle.

6. Conclusion

The research problem was to predict patients that have CHD and determine what the key factors are for people suffering from CHD. From the analysis we have demonstrated that it is possible to highlight the key aspects of those that are suffering from CHD. However, when it comes to predicting whether the person has CHD a different dataset may be advisable due to the ambiguity of the target variable. The analysis that was preformed to predict CHD demonstrates how different machine learning techniques can be applied to predicting categorical outcomes, however because of the imbalance in the classes a more rigorous analysis would need to be carried out to determine whether they will develop CHD. To further this study into CHD, it would be interesting to see if lifestyle and demographics are important in the prediction for CHD.

7. References

Ajmera, A. (2017) 'Framingham Heart Study dataset'. Available at: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset (Accessed: 2nd May 2020)

British Heart Foundation (n.d.) *Facts and Figures,* Available at: *https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures* (Accessed: 2nd May 2020).

Brownlee, J. (2019) *Bagging and Random Forest Ensemble Algorithms for Machine Learning,* Available at: *https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/* (Accessed: 10th May 2020).

Cohn, J.N., Hoke, L., Whitwam, W., Sommers, P.A., Taylor, A.L., Duprez, D., Roessler, R. and Florea, N., 2003. Screening for early detection of cardiovascular disease in asymptomatic individuals. *American heart journal*, *146*(4), pp.679-685.

FHS (n.d.) *About the Framingham Heart Study,* Available at: *https://framinghamheartstudy.org/fhs-about/* (Accessed: 2nd May 2020).

Hajar, R., 2017. Risk factors for coronary artery disease: historical perspectives. *Heart views: the official journal of the Gulf Heart Association*, *18*(3), p.109.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd edn., New York: Springer.

James, G., Witten, D., Hastie,T. and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R, New York: Springer.

Mahapatra, S. (2018) *(Linear Discriminant Analysis) using Python,* Available at: *https://medium.com/journey-2-artificial-intelligence/lda-linear-discriminant-analysis-using-python-2155cf5b6398* (Accessed: 10th May 2020).

Mahmood, S.S., Levy, D., Vasan, R.S. and Wang, T.J., 2014. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, *383*(9921), pp.999-1008.

NHS (2018) *Cardiovascular Disease,* Available at: *https://www.nhs.uk/conditions/cardiovascular-disease/* (Accessed: 2nd May 2020).

WHO (2017) *Cardiovascular diseases (CVDs),* Available at: *https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)* (Accessed: 2nd May 2020).