BT2101 Project Report

Project title: Heart Disease Classification

Team Members:

Hu Xinbei (A0187644R), Liao Meng (A0164769L),

Shi Wen (A0188604X), Tan Tian Le (A0188014H), Xu Pengtai (A0187551X)

1.Introduction

1.1 Motivation

Heart disease is one common ailment among the population and it has led to many deaths and complications worldwide. There are concerns that the rate of heart disease increases with age and other health measures, but how accurate are they in predicting the prevalence of heart diseases? In this report, we will investigate and predict the prevalence of heart diseases using the risk factors identified by many healthcare practitioners today.

1.2 Objectives

Through this, we hope to gain better insights into the risk factors of heart diseases, so that healthcare professionals will be better prepared to handle cases of patients at risk. In particular, we hope to achieve these objectives:

- Understand how each risk factor play their role and significance in determining the prevalence of heart diseases.
- Select the best set of predictors to classify patients with heart diseases.
- Identify the best prediction method and refine it to best classify patients with heart diseases.

Achieving these objectives is paramount to the healthcare industry as we increase patients' satisfaction and save lives. This will not only help people take extra precautionary steps to improve their health status before it grows to become a heart disease, but also drive down the healthcare costs for these people with early precautionary measures that can help to prevent a major accident.

1.3 Hypothesis

Based on our understanding of heart disease, which is highly influenced by conventional beliefs, we hypothesise that age, sex and resting blood pressure are the most important predictors for heart diseases. This is the key thing that we will seek to know in the early stages, before we continue to predict the prevalence of heart diseases using various methods with the significant predictors selected.

2. Data Description

2.1 Attributes Explanation

As shown in Table 2.1.1 below, the dataset has 14 attributes, and the first 13 are used as independent variables to predict the 14th variable "target", which indicates presence/absence of heart disease.

Column	Attribute	Variable Type	Categorisation
1	age	Continuous	29 – 77
2	sex	Categorical	1 = male; 0 = female
3	cp (chest pain type)	Categorical	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
4	trestbps (resting blood pressure)	Continuous	94 – 200
5	chol (serum cholesterol)	Continuous	126 – 564
6	fbs (fasting blood sugar) > 120 mg/dl	Categorical	1 = true; 0 = false
7	restecg (resting ECG results)	Categorical	0: normal 1: having ST-T wave abnormality 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach (maximum heart rate achieved)	Continuous	71 – 202
9	exang (exercise-induced angina)	Categorical	1 = yes; 0 = no
10	oldpeak (ST depression induced by exercise relative to rest)	Continuous	0 – 6.2
11	slope (slope of the peak exercise ST segment)	Continuous	0 – 3
12	ca (number of major vessels)	Continuous	0-4
13	thal (thalassemia)	Categorical	3 = normal 6 = fixed defect 7 = reversible defect
14	target	Categorical	0: Absence of Heart Disease 1: Presence of Heart Disease

Table 2.1.1 Attributes of the Data

2.2 Descriptive Analytics for Attributes

It is key to ensure that the correlations in every pair of continuous variables are not too significant because one of the key assumptions needed for predictions using our selected models is that all features are independent of each other. From the correlation plot (Figure 2.2.1 below), it is observed that the correlation between continuous variables slope and oldpeak is considered high with an approximate value of around 0.6 according to the colour intensity scale. Therefore, it may be necessary for us to only include one of the two in future analysis.

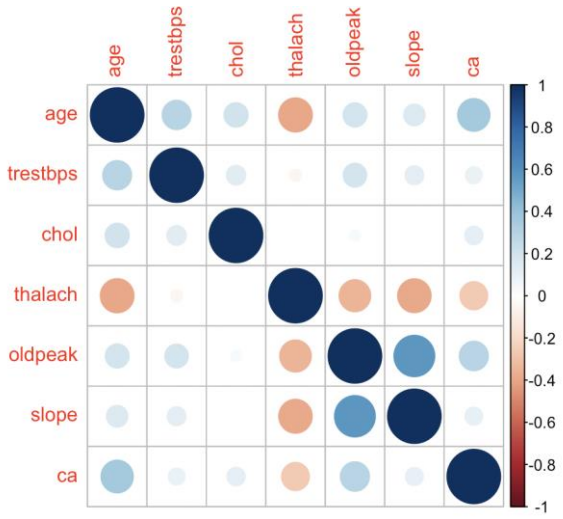

Figure 2.2.1 Correlation Plot

Figure 2.2.2 Bar Plot for 'Target'

The bar plot for the binary dependent attribute "target" (Figure 2.2.2 above) shows that the dataset is quite balanced. Hence, problems of imbalanced dataset should not be a concern. Previously, we have hypothesised that the 3 variables (age, sex and resting blood pressure) might be key factors affecting the outcome variable. We can now have a rough understanding on the correctness of our hypothesis by referring to the bar chart and histograms plotted.

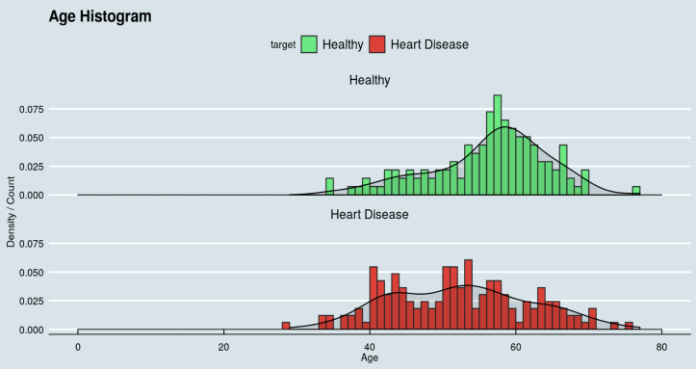

Figure 2.2.3 Histogram on Age

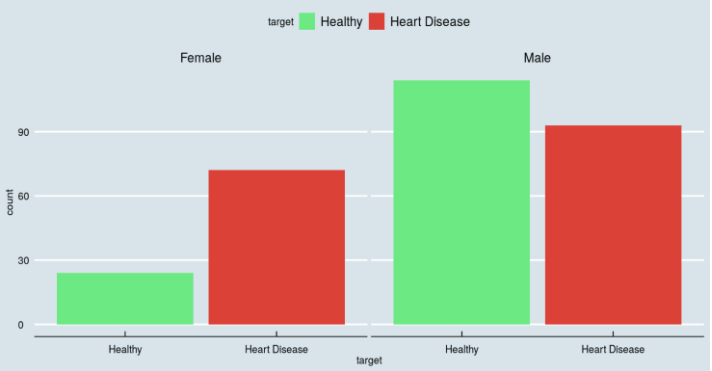

Figure 2.2.4 Bar Chart for Sex

Looking at the histogram on age (Figure 2.2.3 above), for the 2 outcomes of dependent variable, the distributions of age are rather similar. Thus, age may not play a key role in affecting target.

The bar chart plotted for sex is shown in Figure 2.2.4 above. From the bar chart, more females tend to have heart disease while more males tend not to have heart disease. Thus, sex may be statistically significant in this case.

For the resting blood pressure, the 2 histograms plotted for 2 outcome groups are very similarly distributed as suggested by Figure 2.2.5 below. Thus, it may be wrong that resting blood pressure is an important factor affecting the dependent variable, target.

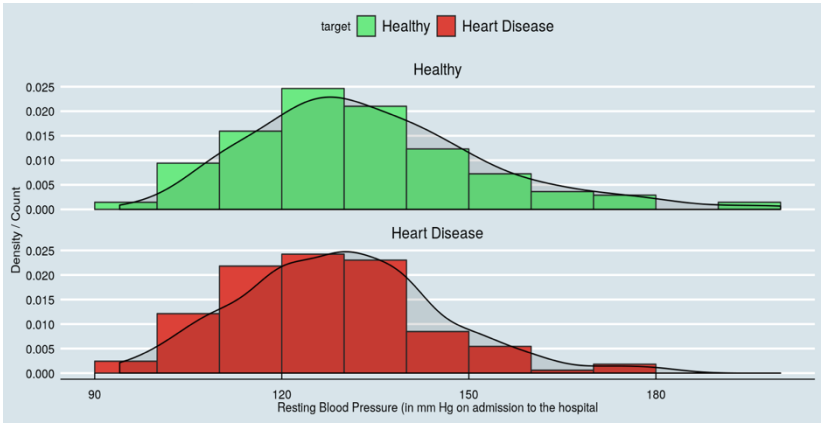

Figure 2.2.5 Histogram for Resting Blood Pressure

3. Methods and Results

3.1 Feature Selection

We used logistic regression to do feature selection. Variables which are statistically significant are kept in the model while others are dropped. The result summary of logistic regression is shown below.

Variables	Coefficient (Std error)
Intercept	1.10 (3.36)
Age	-0.00057 (0.024)
Male (binary)	-1.51 (0.52) **
Chest Pain Type 1 (binary)	0.98 (0.56) *
Chest Pain Type 2 (binary)	1.95 (0.48) ***
Chest Pain Type 3 (binary)	2.02 (0.65) **
Resting Blood Pressure	-0.017 (0.011)
Fasting Blood Sugar > 120 mg/dl (binary)	0.18 (0.57)
Resting Electrocardiographic Results 1 (binary)	0.57 (0.37)
Resting Electrocardiographic Results 2 (binary)	-0.28 (2.27)
Maximum Heart Rate Achieved	0.017 (0.011)
Exercise Induced Angina	-0.76 (0.43) *
ST Depression Induced by Exercise Relative to Rest	-0.49 (0.23) **
Slope of the Peak Exercise ST Segment 1 (binary)	-0.72 (0.86)
Slope of the Peak Exercise ST Segment 2 (binary)	0.20 (0.94)
Number of Major Vessels (0-3) Colored by Flourosopy	-0.83 (0.20) ***
Normal Thalassemia (binary)	1.81 (2.38)
Fixed Defect Thalassemia (binary)	1.85 (2.29)

Significance indicator: *p <0.1, **p<0.05, ***p<0.001

Table 3.1.1 Result Summary for Logistic Regression

Based on the result, five variables -- “Male”, “Chest Pain Type”, “Exercise Induced Angina”, “ST Depression Induced by Exercise Related to Rest” and “Number of Major Vessels (0-3) Coloured by Fluoroscopy” -- are statistically significant. The threshold for statistical significance used here is 0.1 instead of conventional 0.05. This is because we are worried of the presence of omitted variable bias, which is not accounted for by the machine learning algorithms being used later for prediction. For the three variables that we have in our hypothesis, only sex is significant. Both age and resting blood pressure are not.

To better understand the data, we also applied K-means cluster analysis (K is set to 2 in this case as there are two outcome groups). Figure 3.1.2 below is the result we got before feature selection. We see that among the subjects, there does not seem to be two distinct groups. However, after selecting the significant variables, two distinct groups start to appear as shown in Figure 3.1.3 below, which confirmed that correct variables have been chosen.

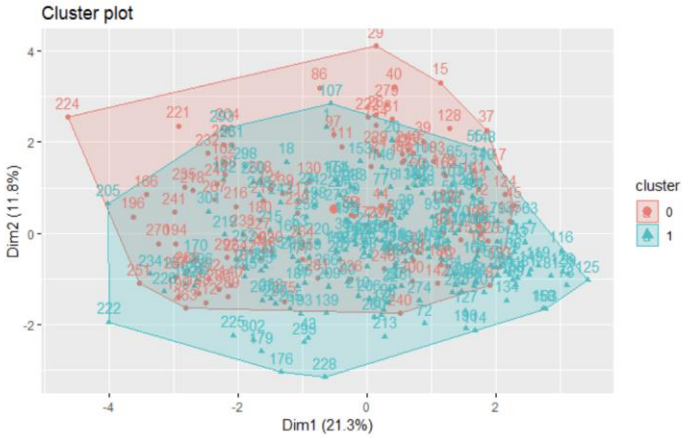

Figure 3.1.2 2-Means Cluster Plot before Feature Selection

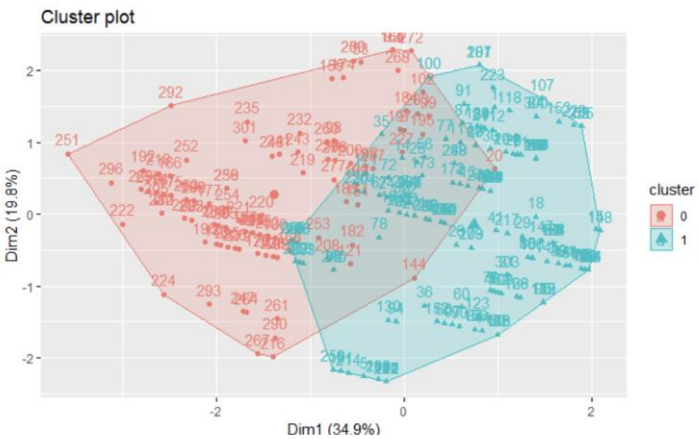

Figure 3.1.3 2-Means Cluster Plot after Feature Selection

A major strength of econometric models is that we are able to understand the importance of each factor through their individual coefficients. We have plotted the coefficient for each predictor as shown in Figure 3.1.4 below. Take sex as an example. With the coefficient of male being -1.4117, we can conclude that males are 68% less likely to get heart disease than females, which is actually quite surprising, as we expected males to be more prone to heart diseases.

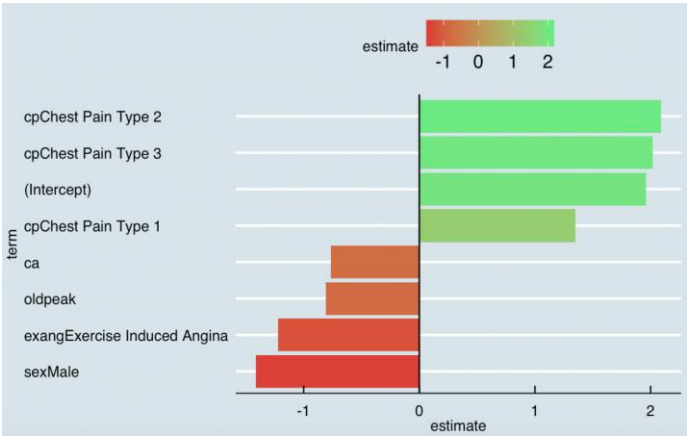

Figure 3.1.4 Coefficients of Variables

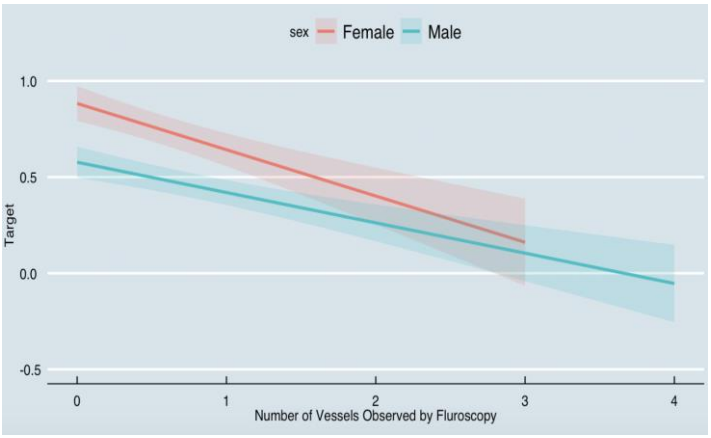

Figure 3.1.5 Change in Probability of Heart Disease

To understand the coefficient for continuous variables, we managed to come up with smoothed lines to represent the probability. Figure 3.1.5 above is an example of the plots. From the plot, we see that as number of vessels observed by fluoroscopy increases, probability of heart disease decreases. A limitation of the smoothed line is that the probability gets below zero after a certain threshold, which is clearly not correct. However, we do still think that the line is valuable in suggesting an overall trend.

Having selected the significant variables, we will then apply a few machine learning algorithms which we have selected based on the characteristic of the data. We employed 70/30 data split (use 70% of the dataset to build models and use the remaining 30% for performance testing) to obtain train data and test data. During the model building, repeated cross validation was used to assess and improve the model.

3.2 Support Vector Machines (SVM)

One of the methods we used to do the classification is Support Vector Machine (SVM). A major advantage of the SVM method is that it is a non-linear learning algorithm. This means that it can handle non-linear relationship between the independent and dependent variables (through the trick of Kernel function). As a result, both linear and non-linear models are possible for SVM. We have built models for both and compared the results in the following section. Both the linear and non-linear models are further tuned with grid search.

3.2.1 Standardisation

As we can see from the descriptive analysis earlier, attributes have different ranges. This means that we need to standardise the data for each attribute to ensure that our model is not skewed. To achieve this,

we used the preProcess() method in the train() function of caret's package by specifying the parameters "centre" and "scale", which basically normalises our attributes' data.

3.2.2 Result Comparison for Various SVM Models

We used the receiver operating characteristic (commonly known as ROC) values to compare each SVM model. ROC value refers to the area under the graph of a plot based on the sensitivity and specificity of the predicted classification. A summary of the results is shown in Table 3.2.2.1 below.

Models\Metrics	**ROC**
Linear	86.80%
Linear with Grid Search	87.23%
Radial Kernel	87.22%
Radial Kernel with Grid Search	**87.39%**

Table 3.2.2.1 Summary of SVM

Confusion matrix for SVM prediction result		Actual	
		Healthy	Have heart diseases
Predicted	Healthy	33	5
	Have heart diseases	7	46

Table 3.2.2.2 Confusion Matrix of SVM Radial with Grid Search

Radial kernel with grid search is the model that has scored the highest in ROC. As such, the model is used to represent SVM to compete with other models. Confusion matrix obtained by the model is shown in Table 3.2.2.2 above.

3.3 Random Forest

In addition to SVM, we used random forest, which is an ensemble learning method of decision trees. A major strength of the random forest model is that it does not require many assumptions. Random forest is not probabilistic, hence it does not require the common probabilistic assumptions such as independence. The only key assumption is that samples are representative, which is a common assumption for many models. The confusion matrix obtained by the model is shown in Table 3.3.1 below.

Confusion matrix for Random Forest prediction result		Actual	
		Healthy	Have heart diseases
Predicted	Healthy	33	5
	Have heart diseases	7	46

Table 3.3.1 Confusion Matrix for Random Forest

3.4 Naive Bayes Classifier

The third model we used is the Naive Bayes classifier. A key assumption of the method is that all features are independent of each other. Upon closer inspection of the dataset, we realised that there are both categorical and continuous variables in the dataset. Thus, we need to make another assumption that $P(Xi \mid Y=yk)$, the conditional probability of P(X) for continuous variables, given that the final outcome Y occurs, follows a normal (Gaussian) distribution. The confusion matrix obtained by Naive Bayes classifier is shown in Table 3.4.1 below.

Confusion matrix for Naive Bayes prediction result		Actual	
		Healthy	Have heart diseases
predicted	Healthy	39	10
	Have heart diseases	5	34

Table 3.4.1 Confusion Matrix for Naïve Bayes Classifier

3.5 Comparing the performance of models

The performance metrics of the three models we have chosen, based on their respective confusion matrix, are calculated and shown in Table 3.5.1 below. We concluded that SVM and Random Forest appear to be the more accurate models for this data set, since they have a much higher sensitivity than Naive Bayes. We are mainly comparing sensitivity (i.e. true positive rate) because for this data set false negative (predict the person to be healthy, when actually the person has heart disease) is more serious and should be minimised.

Models \ Metrics	Accuracy	Sensitivity	Specificity	Precision
Naive Bayes	82.95%	77.27%	**88.64%**	**87.18%**
SVM Radial with Grid Search	**86.81%**	**90.20%**	82.50%	86.79%
Random Forest	**86.81%**	**90.20%**	82.50%	86.79%

Table 3.5.1 performance metrics for each model

3.6 Ensemble model

Having trained three models that obtain considerably good prediction results, we want to further explore the possibility to improve the accuracy of our prediction. Hence, we attempted to build an ensemble model. Among the three main model ensembling techniques: bagging to decrease the model's variance;

Boosting to decreasing the model's bias, and stacking to increasing the predictive force of the classifier, we chose the stacking methods as we main focus on improving the predictive force.

We developed our ensemble model using the "caretEnsemble" package. First, we developed the layer of base models (SVM Radial, Naive Bayes and Random Forest) using the "caretList" method. We included ten-fold cross validation in the process as specified in the "trControl".

Then we experimented with the final output model. The final model was developed using the "caretStack" method. By comparing the result of putting the generalized linear model and the random forest model as the final output model, we realised that the random forest model performed better as shown by the higher ROC, sensitivity and specificity (Table 3.6.1). Hence, our final ensemble model used random forest as the output model and the aforementioned three models as the base model.

	ROC	Sensitivity	Specificity
GLM	0.8777474	0.7493391	0.8480618
Random forest	0.9002122	0.7810784	0.8555652

Table 3.6.1 Performance of Final Output Model

However surprisingly, the ensemble model performs worse than the individual models as shown in Table 3.6.2 below. The model's sensitivity is relatively low at 78.11% and sensitivity is of paramount importance in this study. When applying the models to the same set of testing data, ensemble model gives the lowest AUC of 0.8481. Hence this again indicates that the ensemble model is not the best binary classification model for our study.

	SVM Radial	Naive Bayes	Random Forest	Ensemble model
Sensitivity	90.20%	78.38%	90.20%	78.11%
AUC	0.8889463	0.8765496	0.9039256	0.8481405

Table 3.6.2 Performance of the Different Models

We proceeded to analyse the reason behind its poor performance. The reason lies behind the mechanism of how stacked ensemble model works. Often times the stacked model will outperform each of the individual models due its smoothing nature and ability to highlight each base model where it performs best and discredit each base model where it performs poorly. For this reason, stacking is most effective when the base models are significantly different. It gives the best output when each of the base models capture a different aspect of the data. However, as we can see from the correlation matrix (Table 3.6.3 below), the base models are highly correlated with correlation of more than 0.9. Hence, in this case the ensemble model tends to discredit each base model and gives a poor performance.

Correlation	SVM Radial	Random Forest	Naive Bayes
SVM Radial	1.0000000	0.9263731	0.9164968
Random Forest	0.9263731	1.0000000	0.9255797
Naive Bayes	0.9164968	0.9255797	1.0000000

Table 3.6.3 Correlation Matrix of the Base Models

4. Limitations

Possible limitations of our models are as follows:

- Our dataset is a relatively small dataset (around 300 entries). This limits the accuracy of our model as the model could only be trained and tested against a small dataset. In addition, the dataset is sampled from a limited number of medical institutions, hence might not accurately reflect the general conditions of people with/without heart disease. This means that the external validity of our models is not guaranteed, i.e. we cannot ensure that our models will yield high accuracy when we use them to predict data obtained outside those medical institutions.
- There might still be omitted variable bias after we adjusted the significance level. As we only included five variables, the chance of having omitted variables is high, and since initially there are only 13 features available for selection, it is possible that some variables other than these 13 features could influence the risk of getting heart disease.
- There might be some problem with model training mechanism. In order to have reproducible results, it is necessary to set the seed number for splitting the datasets. However, different seed numbers will give different coefficient and prediction accuracy. As we can only stick with one seed number, the features selected and the models can hence be slightly different if we change the seed number.
- We only experimented with three models and there might be other machine learning models such as neural network that may have higher prediction accuracy.

5.Conclusion

In this project, we have explored different risk factors contributing to heart diseases using the Heart Disease Dataset obtained from UCI Machine Learning Repository. Our original hypothesis was that age, gender and resting heart pressure are significant contributors. However, our analysis showed that among the three--which are beliefs commonly held by most people--only gender is significant. There are also other significant factors such as chest pain type. By identifying the significant factors, it helps doctors and individuals to better interpret their health reports and predict the likelihood of future heart disease. This process is further aided with the classification model that we have come up with. By comparing various models, we have come to realise that the SVM (radial with grid search) and random forests models give the

best predictions. These models can them be utilised by practitioners to have to better predict heart diseases and take necessary precautionary actions in advance.

References

- Gandhi, R., & Gandhi, R. (2018, June 07). Support Vector Machine - Introduction to Machine Learning Algorithms. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
- K-means Cluster Analysis. (n.d.). Retrieved from https://uc-r.github.io/kmeans_clustering
- Caret Ensemble. Mayer, Z. (2016, January 31). Retrieved from https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html
- Patel, S., & Patel, S. (2017, May 18). Chapter 5: Random Forest Classifier. Retrieved from https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1