Diagnosing Cardiovascular Disease Based on Random Forests and Logistic Regression ECS784P

Rashid Al-Mahmud (140707874), Muhammad Najeeb Ul Haq (180746620), Nasra Samadi (150335032)

Abstract—This paper contains findings of predictions made to diagnose heart disease. Random Forests as well as Logistic Regression are utilized as machine learning methodologies. This dataset consists of 14 dimensions and has features which are correlated with the risk of heart disease. We assess the models, explain why they were specifically used and presents the final results. In this report data management and data cleansing is also discussed as well as exploratory data analysis of the dataset. We will also show the process of reaching conclusions based on inference and analysis. A literature review of related works is also presented in this report to expand the scope of discussion for the reader. An evaluation of both Random Forests and Logistic Regression is also conducted to identify pros and cons of both models and more specifically the completion time of each models. Traditionally diagnosing heart disease was based on mostly qualitative study of patients with doctors however the recent advancements in Machine Learning and Data Science techniques can now give a quantitative value to the risk or likelihood of someone being diagnosed with heart disease. Random forests is an example of supervised learning algorithm and it can be used for both classification and regression problems, Logistic Regression is a classification technique in supervised learning. The findings are presented in tabular form as well as visuals to aid the reader.

Keywords—Random forests, Logistic regression, Heart disease, ROC

I. INTRODUCTION

This project will investigate the prediction of heart disease using Random Forests and Logistic Regression models which will both attempt to output prediction of whether a person has heart disease. Random forest will create decision trees based on the data to output a classification of whether a person will or will not have the disease. This is a real-world problem and the report findings could be used in the real-life applications, given, there needs to be industrial level research and optimization of algorithms, but we attempt to try and create a foot stone for this type of data and issue. The definition of what characterizes heart disease and background into the issue is described below. The high-risk factors that we found out online after research for a heart disease are: high blood pressure, smoking, high cholesterol, weight, family history and diabetes [], this was from one source, from another source we found out major risk factors that can't be changed includes: male gender (they have a higher risk of a heart attack than females.), heredity and increasing age. As we mentioned heredity as one of the major risk factors that can't be changed so from our attributes that we have in our dataset thalassemia is one.

Heart disease

Heart disease is a condition that affects the functionality of the heart, this state describes a multitude disorders that influence the heart. Some of these affects can be fatal and currently this is a major cause of the mortality in the UK. (Ons.gov.uk, 2019) 'Ischemic heart diseases remained the second leading cause of death in 2017, accounting for 10.9% of deaths registered. Heart disease is used to generalize an array of different problems one could have with one's heart for example blood vessel disease, also coronary heart artery disease, arrhythmias (issues with the rhythm of the heart) and other defects.

The NHS spends around £6.8 billion on cardiovascular (heart) disease, this is a substantial portion of the NHS budget. It is hoped through providing the professional practitioners a quantitative value based on machine learning that we can help with the diagnosis of heart disease. It is hoped that this report will help with efficient and effective prevention of heart disease will be developed that is personal and tailored to each individual need.

OBJECTIVES

- Identify heavily weighted coefficients/features that contribute heart disease.
- Use a minimum of two classification/prediction models on heart disease to produce an outcome.
- Conduct Exploratory Data Analysis to find correlation between certain features in the dataset that have a higher relation to each other in determining heart disease or not using data analysis.
- Utilize ROC curve to visually demonstrate accuracy.
- Analyze strengths and weaknesses of models and evaluate both.
- Highlight possible improvements for future data analysis.

II. LITERATURE REVIEW

In this section of the report, related discussions and a literature review will be presented and dissected. The work discussed will be evaluated in relation to the research problem being investigated.

The first report to be discussed is titled 'Big Data Analytics in Heart Attack prediction' *by Cheryl Ann Alexander and Lidong Wang* [1]. The aim of the report was to discuss big data analytics in relation to predicting and preventing heart disease, the report also discusses privacy concerns for the patient and challenges as well as future trends in the health industry towards a more technology-based approach. The report discusses that the 5 V's of big data will continue to aid medical procedures and aid further in the prevention of

major diseases such as cardiovascular disease. The 5 V's of big data include; Velocity; data being generated at a high speed, Variety; data being structured, semi structured or unstructured, Value; referring to the value added, Variability; referring to the changes during processing and lifecycle and Veracity; regarding data consistency and data credibility. The report highlights this growing trend and suggests that this will inevitably help with management of chronic diseases and providing more frequent heart data for data analysis. This establishes a strong foundation for the purpose of our report and highlights that with enough knowledge of the features and variables related to heart disease, we can increase the practitioner's ability to make sensible and knowledgeable clinical decisions.

The second report that will be discussed as part of the literature review is titled 'Big Data Analytics in Healthcare: Promise and Potential' *by Wullianallur Raghupathi and Viju Raghupathi* [2]. This report discussed varies conceptual architectures for big data analytics in the healthcare industry. One applied conceptual architecture involved 4 stages, from Big Data Source to Big Data Transformation to Big Data Platforms and finally to Big Data Analytics Applications. Methods such as MapReduce for processing large datasets and a distribution of subtasks for healthcare features were suggested. The report notes that the data analytics will help reduce healthcare costs significantly in the future but there are still some challenges that remain such as privacy and established models for use. There is also an overwhelming amount of variety in the data within healthcare and this can be an obstacle when seeking relevant features; one method to overcome this can be feature extraction or feature selection with the help of Principle Component Analysis. This report provided further confidence in the basis for our report because it highlighted that cost reduction are likely, which is one of our projects aims.

III. DATA MANAGEMENT

Data Source and Description

This dataset was found from Kaggle and reference link has been provided [].
The dataset we can see we have 14 attributes along with the target attribute which tells us if this person/row have the heart disease or not. The shape of dataset is 303x14. The sex of person is 1 = male and 0 = female, the cp (chest pain) attribute has 4 values (1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic). The trestbps is a person resting blood pressure in mm Hg when he was admitted in hospital. chol is the person cholesterol measurement in mg/dl. fbs is the person fasting blood sugar whose values are (1= greater than 120 mg/dl and 0=false). The restecg is "resting electrocardiographic measurement" (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or "definite left ventricular hypertrophy by Estes' criteria"). Exang is the exercise induced angina (1=yes and 0=no). oldpeak is the ST depression induced by exercise relative to rest ("ST is the flat section between S and T wave. It represents the interval between ventricular depolarization and repolarization.). The slope column represents the slope of peak exercise ST segment, it has 3 values: 1=unsloping, 2=flat, 3=down-sloping. Attribute ca as mentioned above the number of major vessels (0-3). The second last attribute is thal which represents a blood disorder known as "thalassemia" and values it can have are mentioned above.

Table 1 Data Sample from original dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Dealing with Missing Data

Missing data values occurs when no data value is saved for a variable/feature in the dataset. Missing data can have a signficant effect on the conclusions that can be established from the dataset, hence missing data is a problem both in the training and testing stages. There are a number of ways to deal with missing data such as;

- Denial: Drop rows with missing values
- Average: Replace missing value with mode for discrete data and mean for continuous data
- KNN and impute: Find K nearest neighbours using available features
- Cluster and impute: Cluster the data using K-Means algorithm.
- Predictive model: setup a predictive model to predict value

In the dataset used for this report there were no missing values however there was 1 duplicate value, but we used the drop.duplicate() function to remove.

Feature Selection

This database contains 76 features, but the Cleveland data source concluded that with professional knowledge only a subset of 14 of them are relevant for machine learning due to the remaining features being redundant based on medical health.

EXTERNAL LIBARIES

NumPy: A library that includes support for large multidimensional arrays and matrices, extensively used as it provides access to very large library of high-level math functions for operations on these arrays and matrices.

Pandas: A popular data framework library in Python, one can manipulate and present data. A DataFrame is defined as a 2D data matrix that support many operations on it for convenience.

Scikit-learn: A machine learning library in Python that is extensively used within Data Science and Computer Science. It was used in this project for the implementation of the algorithms as it is easy and straightforward to implement.

Matplotlib: A library for visualisation of data, there are different types of graphs that can be developed using this library. The graph's axis and scales can be changed accordingly and some elements of customisation in colour are possible.

NORMALITY TESTING

The objective of normality testing is to identify whether a given sample of data is normally distributed, there are several statistical tests that attempt to do this. And they produce certain significance levels (also known as p value). Typical p values are 0.05 or 0.01, for example if we have a p value of 0.05 it means there is a 5% probability of a type 1 error.

IV. METHODOLOGY

The approach we take in diagnosis and prediction of heart disease is we use two models, namely; Random forests, this is a classification model with an ensembles approach. We also use the logistic regression model, a classification model using the data features to provide a forecast to whether a patient has or does not have heart disease.

There are many advantages and disadvantages to each model but based on the dataset, problem and approach the following models were chosen, explanation and reasoning behind these models are explained as follows:

Random Forests Algorithm

Before starting with a complicated definition, we would like to start with the simpler definition to allow the reader to have a simple interpretation of the method. This is a supervised algorithm used mainly for classification and also regression. A random forest is a classification technique in machine learning, this method will take in data and split data based on preset rules depending on criteria of the outcome, features are chosen at random to be split upon. This method will take multiple decision trees with high variance and low bias and combine them to an outcome of low variance and low bias improving accuracy of the result. Random forests are a form of classification, this means that this method will use the output of multiple trees to produce a mean output. This output should represent the best answer and is the most reliable, it will take the outcome of multiple decisions trees and produce the mean.

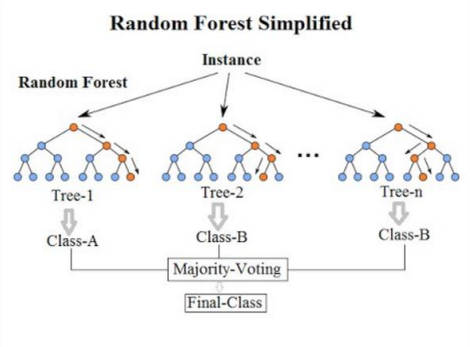


Figure 1 Illustration of the Random Forest classification

Logistic Regression Algorithm

Logistic regression is an optimization method that is based on boundaries in classifier models. A boundary is set usually between 0-1, when the weight and vectors are multiplied the outcome will correspond with the distance from the boundary. Whether the outcome is larger/smaller than the boundary will determine which section this particular instance of the predictor space is in, in our case our predictor space will be whether a patient HAS/HAS NOT heart disease. We use this model as a classifier and not as an optimizer. A P(X) value is given as a result of the logistic function determining the probability of an instance belonging to a section in the prediction space, 1-P(X) is the probability of that same instance belonging in the opposite categorical section.

$$f(t) = \frac{1}{1+e^{-t}}$$

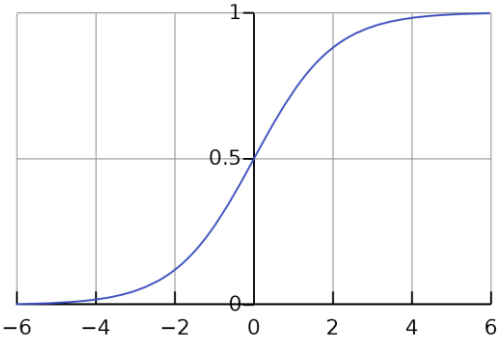


Figure 2 Sigmoid Function

Random Forests Justification

The accurate prediction of heart disease is one of aims of the report and random forests can provide high accuracy because ensemble methods can be created. In a singular fashion like a single decision tree the output is usually weak, meaning the result cannot hold much weight behind and it cannot in most cases be used a conclusion of a hypothesis. One-way of counteracting this is using a multitude of decision trees where the mean of each output will be used as the answer, this improves the accuracy and also stops the effect of overfitting. Overfitting is classified when the error

is higher in the test data compared to training data, this happens when the model memorizes the data.

Another reason for using random forest are that it does not require normalization, it does not suffer from having scalars that cannot be compared and also out range instances, they are split based on a criterion. And, trees used in the random forest model can be trained in parallel meaning at the same time which is good for computational speed, pruning could also be used to manage overfitting, which aids the model's overall performance.

Logisitc Regresion Justification

The model is based linearly so if the outcome is categorical or binary and linearly separable (the outcome is 0 or 1 to show if a patient has heart disease (1) or does not have heart disease (0), since we want our outcome to be a binary value, this model is useful and appropriate to use.

The logistic regression can also be regularized, this process includes a penalty addition to the algorithm which penalizes large coefficients.
During the process of fitting the data you need to perform a min-max normalization so it can fit the model, this will result in the removal of outliers which reduces the impact of anomalous data. It considers the multiple features and after normalization allows them to be able to enter the same plane as all values become between 0 and 1 which also allows better comparisons of features.

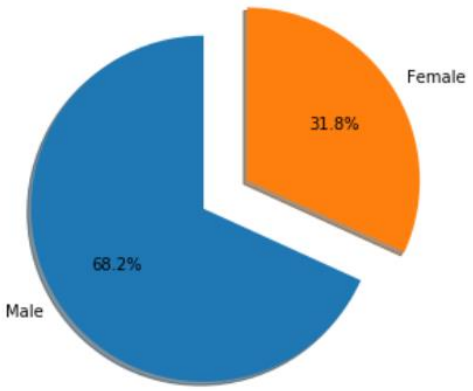
This model can be used simply and applied easily to complex datasets which is also a contributing factor to why we used it. A probabilistic return is favorable because we would like to quantify the accuracy of the diagnosis.

V. EXPLORATORY DATA ANALYSIS

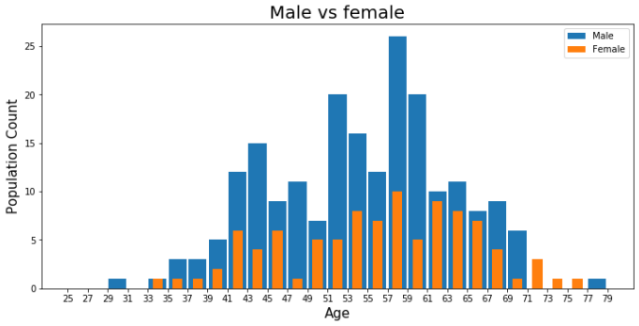
Now as we know a little about our dataset and what attributes we have and which ones are important, we will perform some experiments and conduct data analysis with our dataset so we can have a general idea of interrelationships

According to the dataset source and as mentioned in the introduction, being a male is a strong factor for heart disease. As a result, now we want to identify numbers of males and females within our dataset. We will use pandas value_counts() function on "sex" attribute, it returns a series containing counts of unique values in descending order in our case its only 1 and 0 (1=male, 0=female). We will verify this by another function as well called .count() and a condition where ['sex']==1. (See appendices for code.)

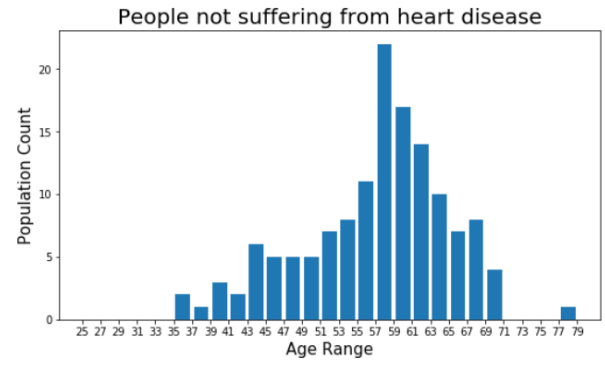
We will use pie-plot from matplotlib to visualise our findings and it will give us an idea how much percentage of population is male and female. As from the pie chart we found out that 68.2% of the population is male that is colour blue and 31.8% of the population is female.



To continue our analysis what we now want to do is look at the distribution of population with respect to age and after that we will look at the distribution with respect to gender as well. From the distribution where we have labelled female by orange and male by blue shows that the start from 29 to 31 we have no female in our dataset and after that the count of females is nearly half as that of males till 47-49(age group) where we have much less females as compared to males. We can see also that in certain number of bins/age-group we have nearly same number of males and females that is for 61-67. Another thing we noticed is that for age group 71-77 we only have females and there are no males in that age group but at the extreme 77-79 we have less number of males and no female in that range. From inspecting these distributions, we now have idea about our population, and we know much more in detail that most of the population is between 41-69 and in which age group we have no males and no females and where the count is less. It is evident that the dataset is slightly skewed in terms of age corresponding to gender.



Now we will investigate distributions of people that have a heart disease and people not suffering from a heart disease. For this we will use the age and target attributes, age because we want to look at the population's age relation to target attribute distribution. By looking at the plots we analysed that for age group 41-45, 51-55 and 57-59 the count of people suffering from heart disease is higher than the mean count, but there is no relation it's kind of random. For the second plot people not suffering from heart disease we can see that age group 57-59, the count is high. For this plot there is also no clear relation that we can talk about. We can say that probability of suffering from heart disease decreases after 60, because there is a decreasing trend.



In this stage of the Exploratory Analysis we find the correlation between all the attributes and the target doing this, we will have a rough idea about how a specific attribute is more related to the heart disease. For this we used pandas corr() function it will return us a dataframe where we can see the correlation of every attribute with other attributes, but for now we just want to see with the "target". We have sorted the output in descending order and colour labelled with intensities accordingly. It can be analysed that 'cp' chest pain and 'tahlach' maximum heart rate achieved are more correlated to target as compared to others. For most of the attributes there is negative correlation as well.

	target
target	1
cp	0.43208
thalach	0.419955
slope	0.34394
restecg	0.134874
fbs	-0.026826
chol	-0.0814372
trestbps	-0.146269
age	-0.221476
sex	-0.283609
thal	-0.343101
ca	-0.408992
oldpeak	-0.429146
exang	-0.435601

VI. TESTING AND RESULTS

This section will cover the test and result section of the report, it will explain the how the models were implemented, and the results obtained from the model implementation. Both models showed good levels of accuracy and have proved to be good tools in the diagnosis of the heart disease. The models could successfully return whether a person has heart disease or not with a binary output.

Random Forest Results

Implementation of the model was relatively simple and done with ease. For the case of the random forest model using the inbuilt function to fit the data set and run it was straight forward. We used the python library

'from sklearn.ensemble import RandomForestClassifier'

sklearn is a library that contains multiple predictive and classification models, this includes the random forest classifier. The model performed well and behaved expectedly, the function outputs a tree and classified whether group of people had heart disease, in the method it was explained how the random forest comes to these conclusions, the large scale version of tree is provided in the appendix and a result link is also provided below as well as a small scaled version.

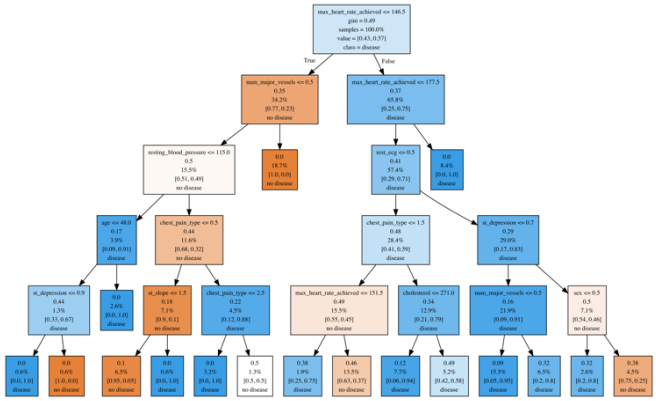


Figure 3 Random Forest train output

Result link

You can see that from the root node it specifies a starting feature to be max heart rate as the first rule to split the data, it would suggest that maximum heart rate has an impactful influence on the results of the heart disease. The trees first rule is to check if the patient's heart rate is less than or equal to 146.5 which allows it to make its first separation. Here we have a plot to show that people with a maximum heart rate have a higher risk of heart disease than other patients. Although this is not definitive to whether a patient has heart disease or not the graph shows that generally patients with a lower maximum heart rate have less heart chance of heart disease. The difference in the plot in figure 3 is marginal and doesn't provide a strong causation point but be a contribution to heart disease.

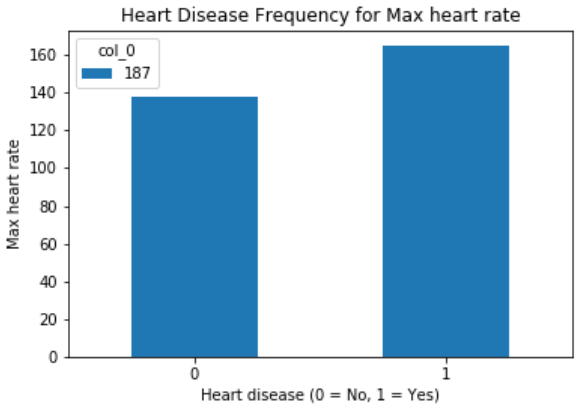


Figure 4 Heart Rate VS Target

Due to the sparser amount of data the splitting structure will not be identical and instead of the root node being the max heart rate it is a different feature; the features are also split at random which also means the tree structure will be different when you use the test data instead on training data.

The above procedure is done throughout all dimensions of the dataset and the bottom row shows the results, this is a culmination of factors done through the tree, the arrow pointing left = True and the arrow pointing right = false. For example if you look at the result on the bottom left tree you can see from the dimension of the age, the slope of the peak exercise ST segment, and a different partition of the age, if the criteria is met for the random forest you will find that the it predicts these group of patients do have heart disease. In the results of the decision tree you find that 2.8% of the heart disease patient data contain patients with the given criteria having heart disease. If you follow the tree through the branches and nodes you will find different criteria for the diagnosis of heart disease.

Below is the relation of the root node to heart disease and age, you can see that the relation is not explicit, but the visualization shows how much of an impact multiple dimensions have further down the tree for example in the layer two it splits the input major vessel and exercise features.

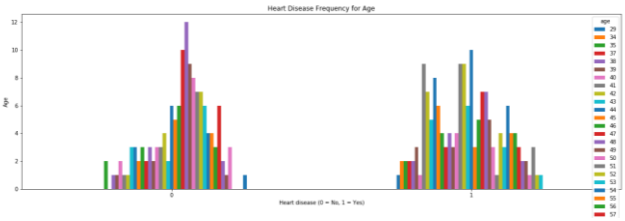


Figure 5 Age vs Heart Disease

The following tree is the result of the test data, it is smaller due to less data from the test data section the structure is different because the features are split randomly meaning feature order in the tree will not be the same as the training data output. The selection of the samples is random (reduces overfitting) so the structure of the partitions of the trees is different.

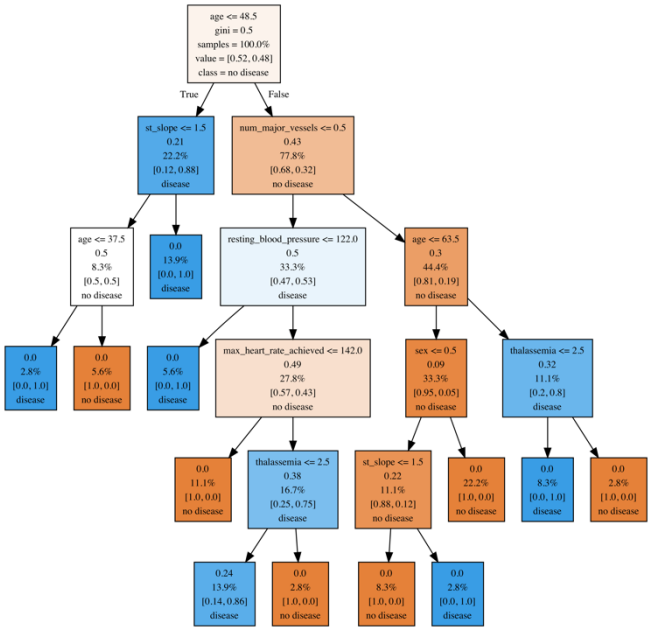


Figure 6 Test data Tree

Result Link

To test the accuracy of the decision tree we employed the confusion matrix, this is placing the correct and incorrect answers in the table. We are interested in two outputs of the confusion tree, the true positive rate (TPR) the amount of results predicted true and are true or sensitivity and true negative rate (TNR) which is the rate of correctly negative predicted results or specificity.

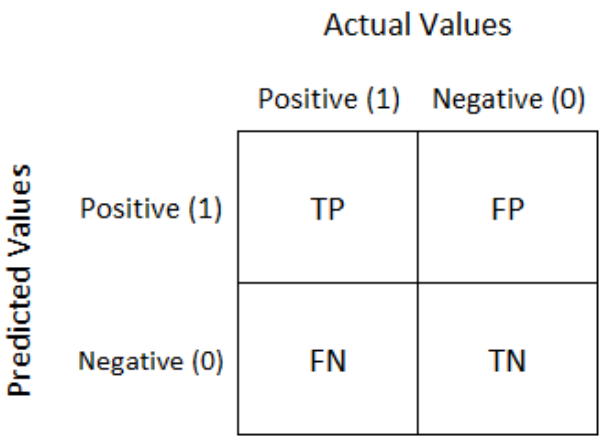


Figure 7 Confusion matrix

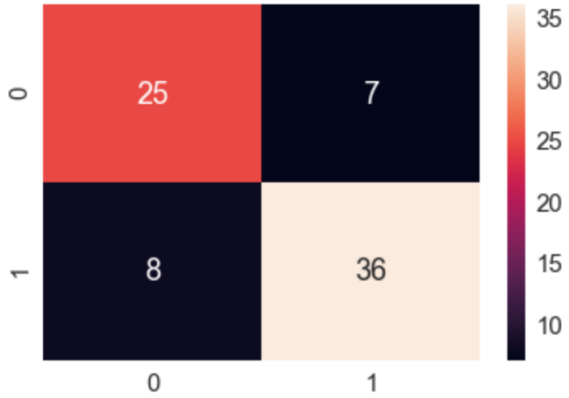
```
total=sum(sum(cm))

sensitivity = cm[0,0]/(cm[0,0]+cm[1,0])
print('Sensitivity : ', sensitivity )

specificity = cm[1,1]/(cm[1,1]+cm[0,1])
print('Specificity : ', specificity)
```

```
Sensitivity :  0.90625
Specificity :  0.7931034482758621
```

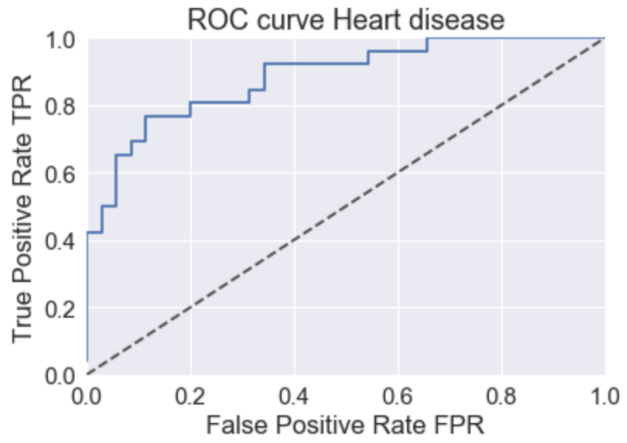
Figure 8 Sensitivity and Specificity

The sensitivity is 90%+ and the specificity is 79%+, this is a good result, but it could be improved exactly how it could be improved will be part of the future improvements.
Below is the plotted confusion matrix for the random forest decision tree. This shows overall accuracy of the random forest classifier.

```
Accuracy  80.26315789473685
```



We also used a receiver operator curve (ROC) which plots the TPR and false positive rate (FPR) against each other to give an indication to the how well the model performed. It is important to look for the area under the curve (AUC) value, this will highlight how much the of the data the model correctly classified, the plot should show a large area under the plotted to curve indicating a high TPR pointing towards a good model. Usually this curve will move around when you move the boundary in other classifiers but can be just as effective in decision trees as is still uses the values that are correctly predicted and incorrectly predicted. The higher the AUC the better the result.




```
auc(fpr, tpr)
```

```
0.8846153846153847
```

Figure 9 ROC plot

Logistic Regression Results

Now we can move onto our second classifier we used on the dataset Logistic Regression. This model is from same sklearn library and is practically applied in a similar fashion.

```
from sklearn. linear_model import LogisticRegression
```

The results will be more restricted to smaller output of either yes or no, in this case it will be either 0 or 1, this has been previously explained in the methodology section. The logistic regression method is simpler to apply, and the method used to classify the data is straight, multiplying the weight and predictors to plot an output between the given boundary as explained in more detail in previous sections. The process goes as: split the data between test and train in this case being 75% train and 25% test

```
# Here we simply assinging the data traing and test chunk to the x and y axis, we split the data 75% train and
# therefore 25% test size
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)
```

Figure 10 Splitting Data

Then we fit the data to the model with the given function

```
# Assignining the 'LogisticRegres
LogisticR = LogisticRegression()

# using the fit function to take
# function requires, we training
LogisticR.fit(X_train, y_train)
```

Figure 11 Fitting data

We use the fit function that will automatically apply any pre-processing required, this could include some regularization like the min-max so the dataset can be applied to the model. The min-max regularization is explained thoroughly in previous sections, when this is applied the model can then be executed.

After the model is executed, we use the confusion matrix to assess the model accuracy, you can see the following table and the rate at which the model performed. The confusion matrix can be applied again in the same way of the random forest example

```
#Plot the confusion matrix
sns.set(font_scale=1.5)
cm = confusion_matrix(y_pred, y_test)
sns.heatmap(cm, annot=True, fmt='g')
plt.show()
```

Figure 12 Confusion Matrix of Logistic regression model

This is the resulting confusion matrix and accuracy, the accuracy is higher, but this could be a result of the test data size being higher, in the random forest model we set the test data to 20%, more data could mean better accuracy in this instance not just better model performance.

```
Accuracy  82.89473684210526
```

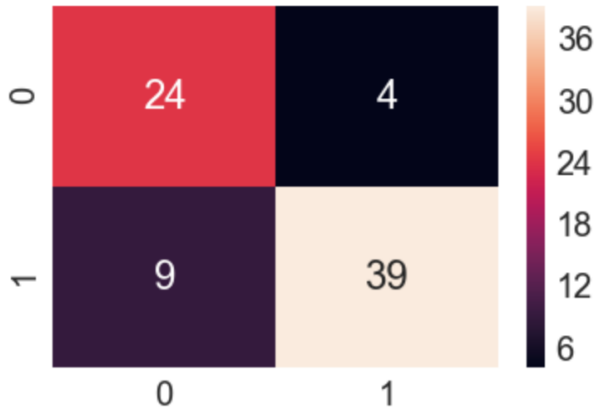
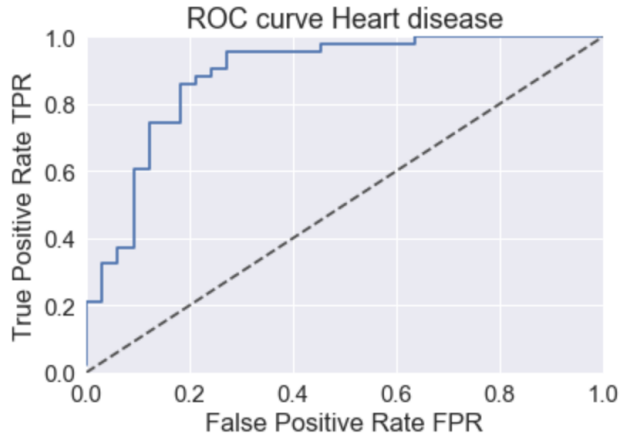


Figure 13 LR confusion matrix plot

We also have a ROC plot to show the rate at which the model performs when you move the boundary of the classifier. This produces an AUC score of 88%+ which is a reflection of a good model, considering the high dimensions of the dataset and the complexity of the problem.




```
auc(fpr, tpr)
```

```
0.8858350951374208
```

VII. CONCLUSIONS

We believe the that the aims of the project have been achieved, we set out with objective aims and they have been identified in the report along with evidence. One of the first aims was to the be able to diagnose and therefore predict cardiovascular disease given the patient data in the form of the heart.CSV file. We used two models (Random Forests and Logistic Regression) that output result with relatively high accuracy, and both performed well in the testing phase of the report showing high levels of accuracy and provided a strong AUC number in the ROC, so we believe this aim was achieved. Both models used in the report were evaluated and compared in terms of accuracy and theory. We use the two models so there could a certainty factor, the models both output > 70% accuracy and applying both indicate the given dimensions in the dataset are not redundant.

Another aim was to the identify key weigh/coefficients in the dataset that have a high impact on the outcome. This is essentially which weights provide the most performance increase in to the models. Through feature selection we found that that not one single feature had a determining influence on whether a person had heart disease or not, rather the features all have a contributing element to whether a patient is diagnosed or not with heart disease. Some of the features could more strongly correlated to heart disease for example chest pain type or cholesterol over for example sex, but if more research is done into each feature there could be a more leading singular factor in common with patients with heart disease. With the high level of accuracies achieved it is evident that if these models were deployed for GPs in the UK then we there would be significant reductions in costs and also time with patients for Doctors.

Some of the limitations of the project were the dataset could have been created by internal arrangements meaning we collected the dataset by ourselves which would have increased the validity of the dataset. Also, due to the nature of the problem, different datasets will have different features as there is a multitude of factors which lead to heart disease. Perhaps if the dataset were to change then the model performance could also have changed.

A simpler method such as Linear Regression could have been used to provide a baseline for the comparison of the more complex algorithms, this would show a more extensive evaluation. In contrast, using a more complex model such as Neural Networks and deploying them on people who have been to hospital due to heart problems could be more effective because Neural Networks have the advantage of learning the key features, however this could lead to overfitting.

Furthermore, consulting with professionals in the medical field and conducting greater primary and secondary research would have helped the scope of this report because there could be missing dimensions not taken in to account such as lifestyle choices.

VIII. REFERENCES

[1] ‘Big Data Analytics in Healthcare: Promise and Potential’ Wullianallur Raghupathi and Viju Raghupathi - 2014
URL:https://www.researchgate.net/publication/272830136_Big_data_analytics_in_healthcare_Promise_and_potential

[2] ‘Big Data Analytics in Heart Attack Prediction’ Cheryl Ann Alexander and Lidong Wan - 2017
URL:https://www.omicsonline.org/open-access/big-data-analytics-in-heart-attack-prediction-2167-1168-1000393.pdf

[3] ‘Health Deaths Registered’ Office for National Statistics 2017
URL:https://www.ons.gov.uk/peoplepopulationandcommuni

ty/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredinenglandandwalesseriesdr

[4] Python Data Analysis Library URL: http://pandas.pydata.org

[5] Scikit Learn: sklearn. ensemble.RandomForestClassifier Scikit-learn developer URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[6] Scikit Learn: sklearn. ensemble. LogisticRegressionClassifier Scikit-learn developer URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[7] Heart Disease Data Set https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[8] Diagnosis of heart disease via cnns. K Wang, Y Kong - Diagnosis of Heart Disease via CNNs, 2016 http://cs231n.stanford.edu/reports/2016/pdfs/331_Report.pdf

[9] A low-cost method for multiple disease prediction M Bayati, S Bhaskar, A Montanari - AMIA Annual Symposium 2015 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765607/

[10] Predictive data mining for medical diagnosis: An overview of heart disease prediction J Soni, U Ansari, D Sharma, 2011. https://pdfs.semanticscholar.org/fbd6/5a18f6653b56138cd5196d20e2f39de189e3.pdf

IX. APPENDICES

The notebook code for the implementation and analysis can be accessed with the following links. Python 3 was used with NumPy, Pandas, Scikit learn, Matplotlib and seaborn.

Notebook for Logistic Regression Implementation

https://drive.google.com/open?id=1OvrmA-Hlov3q931HMyQ0KHhFiJNqyzP9

Notebook for Random Forests Implementation

https://drive.google.com/open?id=1haAP0oPS2ennfIisTagQ8TqqCQ1MZBcU

Notebook for Exploratory Analysis

https://drive.google.com/open?id=13DN9DAC5UIvFDDw9H_N7CC8itv92WvS-