Machine Learning

(ST3189)

University of London

Name: Malcolm Teo

UOL Student Number: 200615961

Table of Contents

Introduction

With the advent of more sophisticated technology, there has been a general consensus that a variety of heart disease such as coronary heart disease (Jones & Greene, 2013) and rheumatic heart disease (Liang, Yu, Lu, Zheng, & Yang, 2023) has been on the decline over the years. Coupled with the fact of advanced modern medicine, greater understanding of prevention measures, and greater standard of living across societies, it would come as no surprise that we would observe a fall in heart diseases. However, even with those mitigating factors, we still face conflicting reports of a rise in those same heart diseases whereby (M Ahern, et al., 2011) pushes the point that there is an increase in cardiovascular related problems from high-income countries, underscoring a positive correlation in standard of living and cardiovascular disease. There has also been a rise in obesity (Sarma, Sockalingam, & Dash, 2021) in locations with higher incomes, thereby leading to long-term health complications such as coronary heart disease (Agha, 22 June 2017). To address the conflicting points laid out by the different studies, we will be conducting our own test through the use of various machine learning techniques such as classification, regression, and unsupervised learning.

Research Question

To address the claims made by the different studies, we have identified four main aims:

I. **Research Question 1**: What are some early detectors of heart disease that can be detected?
II. **Research Question 2**: How confident can we predict heart disease in a person who has not encountered any symptoms?
III. **Research Question 3**: Does the age of the person play a role in the risk of the onset of heart disease?
IV. **Research Question 4**: Are we able to speed up the detection process using imaging technology?

Research Objective

We propose the following research objective to address the respective research questions:

I. **Research Objective 1**: To identify the statistical correlations of heart diseases categorical and continuous variables
II. **Research Objective 2**: To evaluate the effectiveness of predictive models in diagnosing patients who experienced asymptomatic chest pain.

III. **Research Objective 3**: To determine how strong a role "age" plays as a determinant of heart diseases.

IV. **Research Objective 4**: To consider the possibility of machine vision or machine detection in the identification of heart diseases.

Strategies

We will be using various machine learning techniques such as classification, regression, and unsupervised learning to address the different research objectives. We will be using the following strategies for the different research objectives:

Research Objective	Classification	Regression	Unsupervised Learning
1	• Decision Tree	• Decision Tree • Regression Line • Mean Absolute Error (MAE) / Root Mean Squared Error (RMSE)	
2	• Linear Discriminant Analysis (LDA) • Random Forest • Confusion Matrix	• Random Forest • Regression Line	• Random Forest • K-Means
3	• Gradient Boosting Machine (GBM)	• MAE / RMSE / Mean Squared Error (MSE) • GBM	• GBM
4	• Machine Vision		• Machine Vision

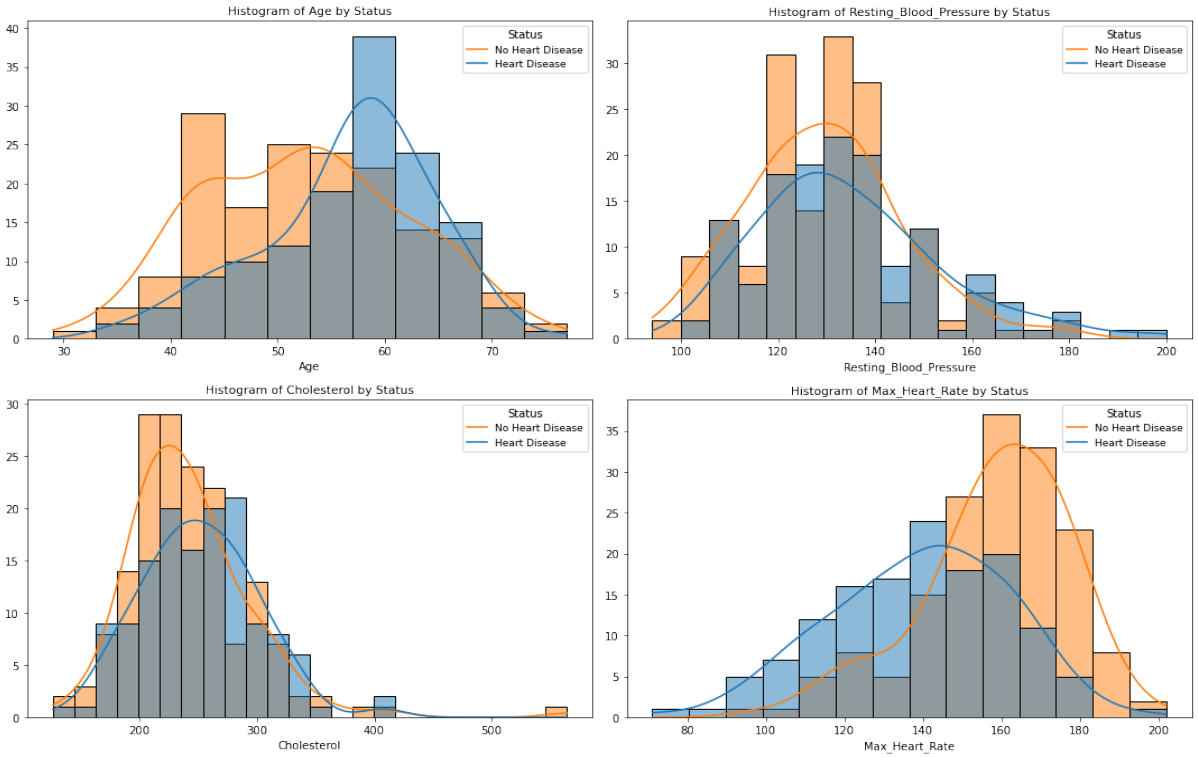
Data Analysis

Our research involves the use of two different datasets to fulfil our objective, The first dataset relates to heart disease where we will use statistical techniques to address research questions 1 – 3. For the final research question, we will use a digit recogniser as a basis for our machine vision recognition. This aims to investigate the potential of employing imaging techniques such as 3D or magnetic resonance imaging (MRI) to pinpoint heart diseases.

Data Collection

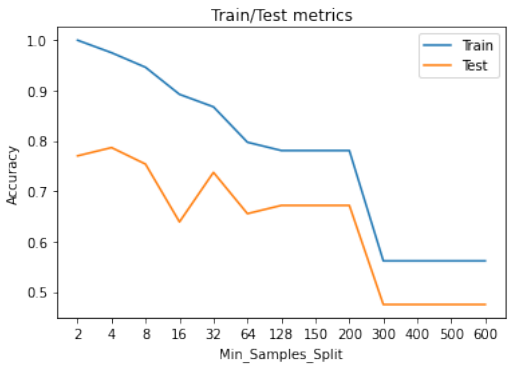
Before commencing on the research objectives, we divide our data based on the heart disease status, then visualising it using histograms and line charts across the various categories. Our key observations include:

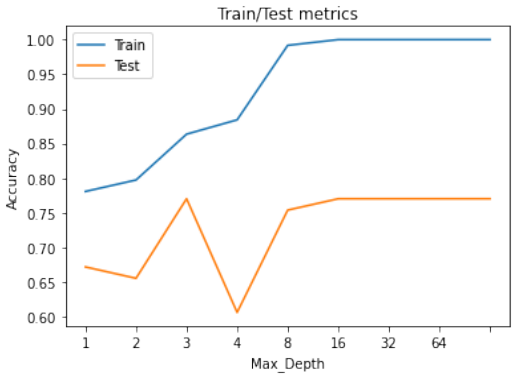
- Participants with heart disease have an average age of around 60 years old, while those without heart disease averages at 50 years old
- Resting blood pressure has no discernible difference between the two groups
- Cholesterol numbers are almost identical in their respective status
- Maximum heart rate among participants suffering from heart disease is lower, at approximately 150, compared to 170 for those without heart disease.



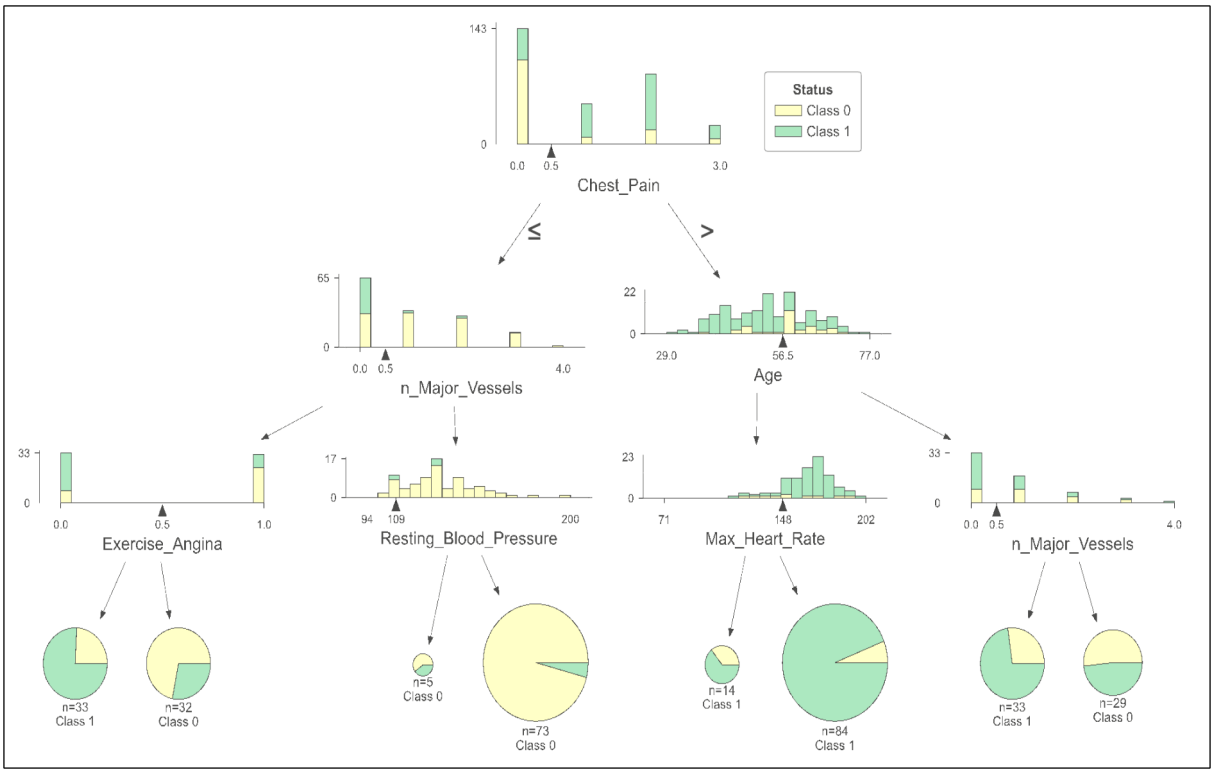
Research Question 1

From the histogram, we used a decision tree model to identify significant variables that influences the participant's likelihood of having heart disease. The dataset was divided into training and testing subsets (80% and 20% respectively). Through our testing, we determined that a minimum split of approximately 220 and a tree depth of 3 yields optimal results for the decision tree. The dependent variables, which included continuous and ordinal variables such as exercise angina and chest pain, were set against the target variable "status". This resulted in the following observations:





- Participants without heart disease exhibited higher resting blood pressure surpassing 109
- Participants devoid of heart disease had higher maximum heart rate

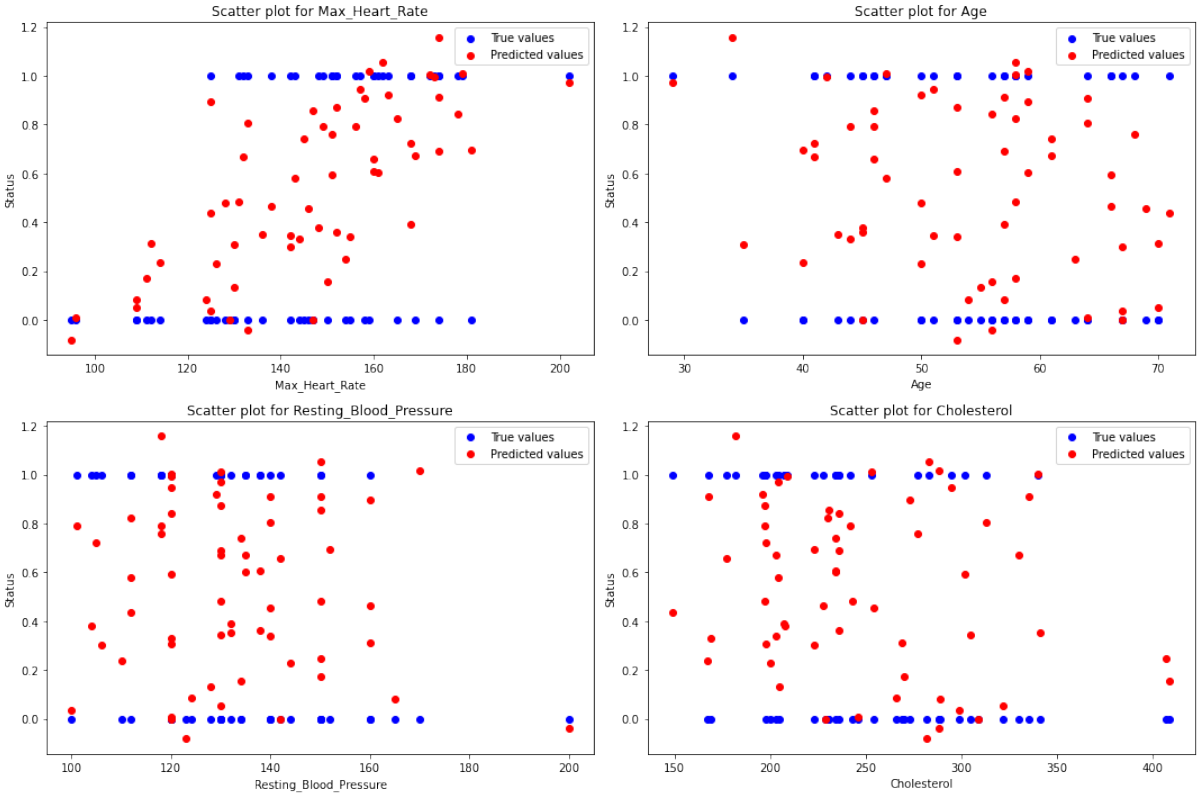


After generating the decision tree, we proceeded to determine the regression line and create scatterplots for various variables relative to the status variable. Reviewing our regression line shows an intercept value close to 0.2046, indicating an absence of heart disease. However, it is worth noting that exercise angina has the strongest negative impact at -0.2358, while chest pain had the highest positive impact at 0.149.

```
Intercept: [0.2046]
Coefficients:
Max_Heart_Rate : 0.0038
Exercise_Angina : -0.2358
Age : -0.0022
Resting_Blood_Pressure : -0.0013
Cholesterol : 0.0004
n_Major_Vessels : -0.1222
Chest_Pain : 0.149
```

From the scatterplot, we note the following observations:

- Max heart rate has a positive correlation with heart disease
- There is no discernible relationship between age, resting blood pressure, and cholesterol levels in the presence of heart disease

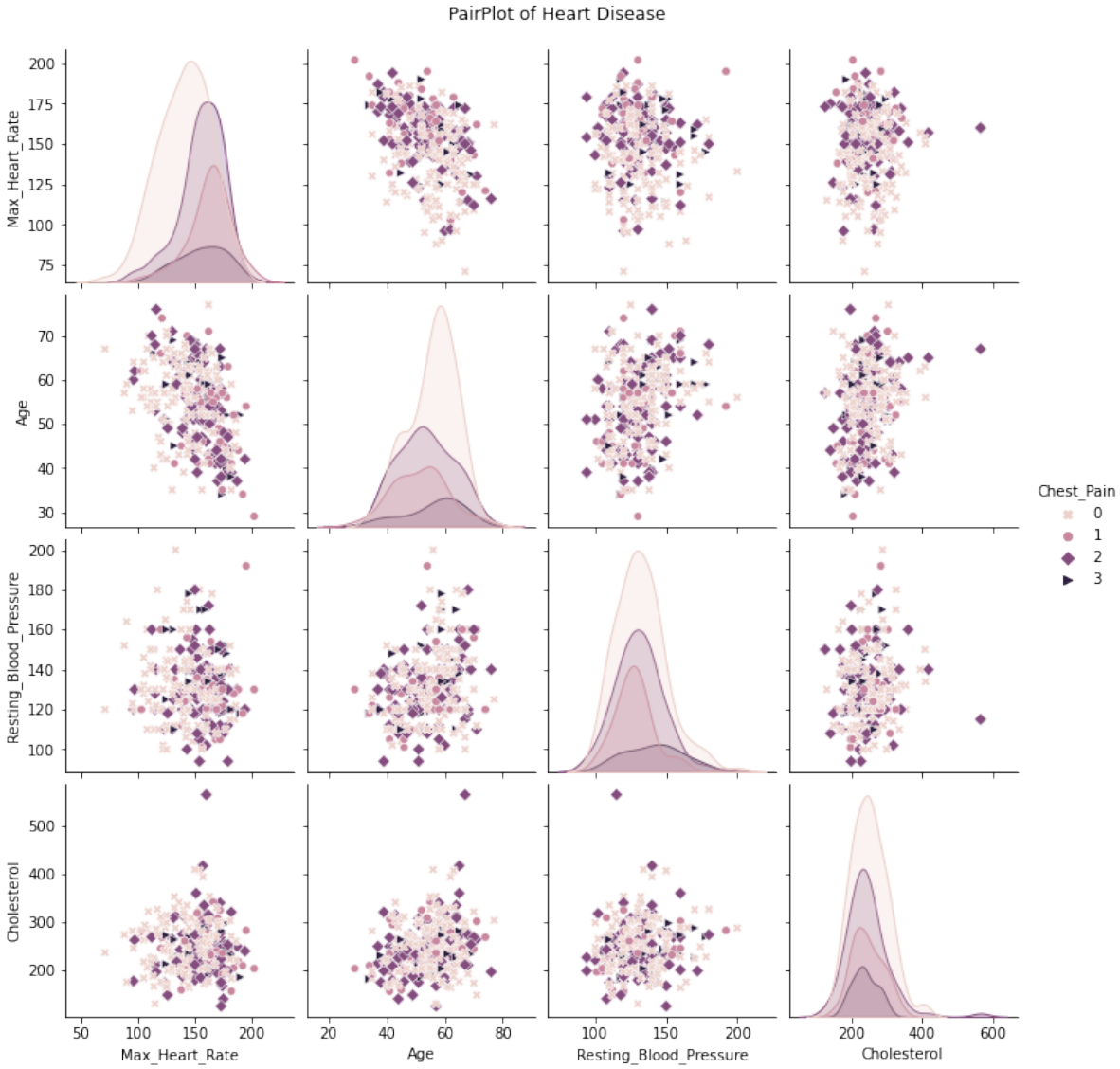


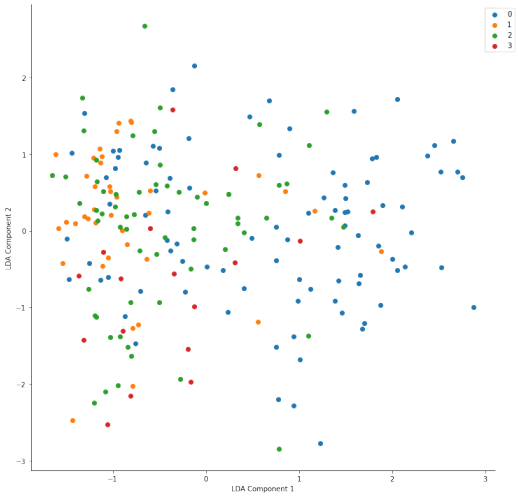
Concluding on our first research question, we evaluated the performance of MAE and RMSE resulting in values of 0.3305 and 0.4276 respectively. Additionally, our logistic regression achieved an accuracy of 73.77%. The values suggest that our model performs reasonably well with minor errors. Therefore, from our analysis and observation, it is evident that maximum heart rate serves as an indicator for persons with heart disease.

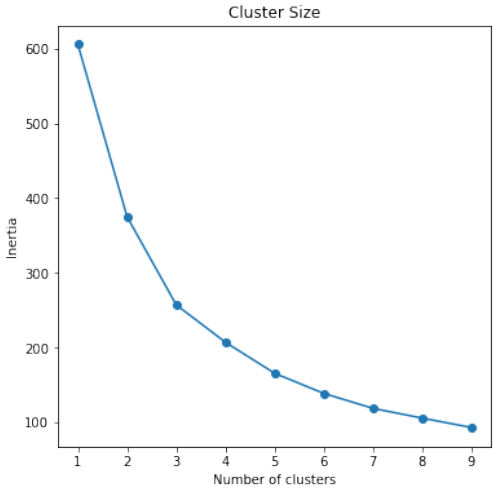
Research Question 2

To access our confidence in predicting individuals who has yet to exhibit symptoms of heart disease, we start off by comparing pair plots of various variables. We observe the following:

- Age and resting blood pressure exhibit a positive correlation
- Age and cholesterol levels have a positive correlation
- Max heart rate displays a left-skewed distribution and negatively correlation with the age, indicating that the older the individual, the more likely they are to experience non-asymptomatic chest pain.

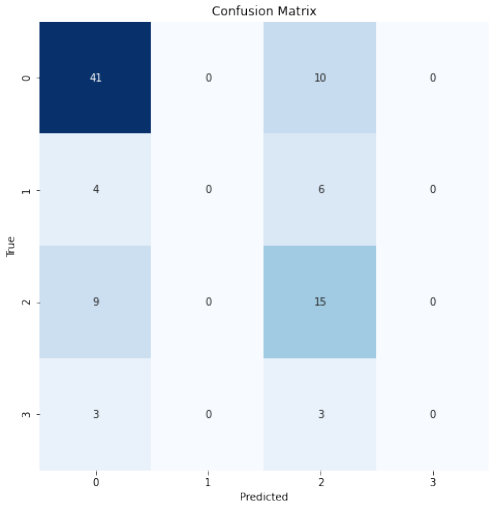


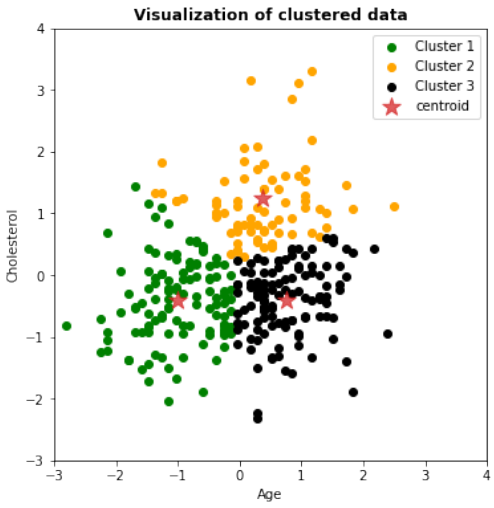


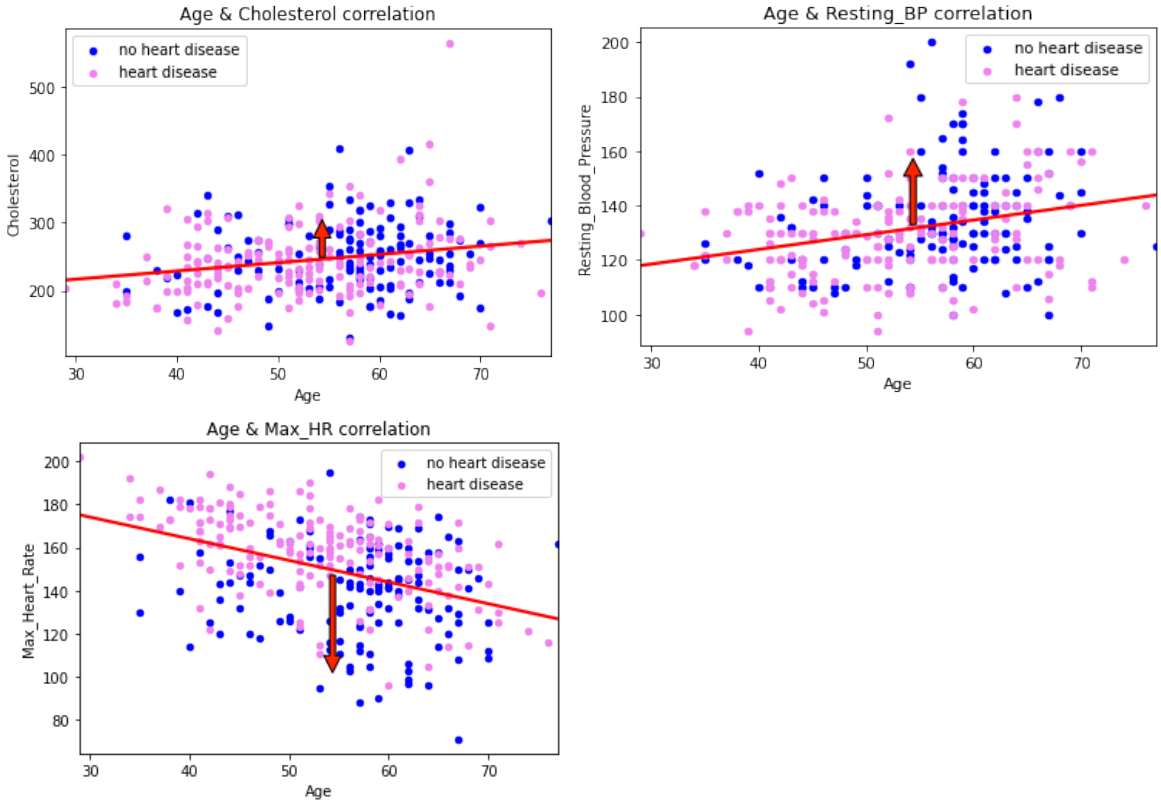


The analysis of LDA indicates that there is no clear distinction within the classes of "Chest_Pain", suggesting a potential lack of data for effective discrimination. Subsequent random forest analysis and confusion matrix both yields a prediction accuracy of 62%, which reveals the low confidence in identifying atypical angina and asymptomatic chest pain. Transitioning to K-means clustering, we establish three clusters based off the diagram. Observation from the K-means cluster leads to the following conclusions:

- Age and cholesterol levels exhibit positive but weak relationship, with a higher rate of heart disease among younger participants.
- Age and resting blood pressure exhibits a positive and strong relationship, with no discernible correlation between individuals with heart disease
- Age and maximum heart rate exhibit a negative and strong relationship, with a clear distinction of individuals without heart disease exhibiting lower maximum heart rates.







Research Question 3

To determine the importance of age as a variable, we compare two models using Bayesian information criteria (BIC), with age being our differentiating factor. Our hypothesis is:

- H0: The full model and reduced model fit the data equally well
- H1: The full model fits better than the reduced model

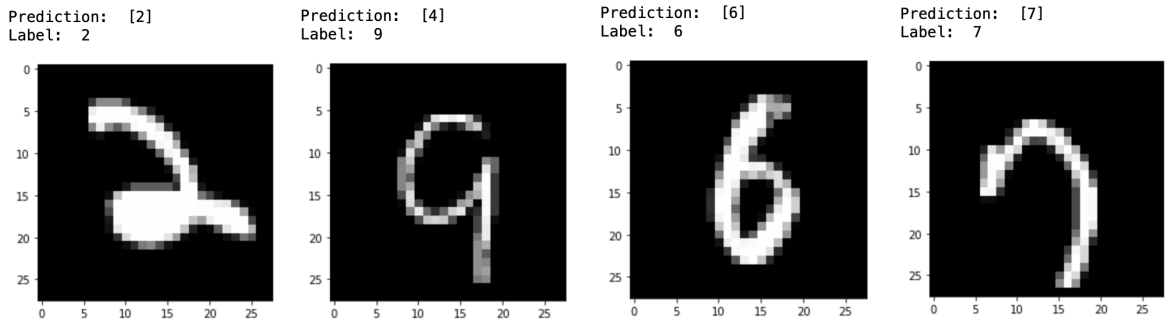
The results shows that the model including age has a slightly lower BIC value at 869 compared to 874 without age. We further access the RMSE and perform a likelihood ratio test using the chi-squared test. Both the training and testing models shows similar RMSE values at 0.5183 and 0.5432 respectively. The chi-squared value of 0.307, which is greater than 0.05, leading us to not reject the null hypothesis. Thereby, suggesting that age has minimal impact on heart disease presence.

Employing the use of support vector machine (SVM) for enhanced predictability, we first determine the optimal kernel among linear, sigmoid, radial basis function (RBF), and polynomial. The 3-fold RBF kernel yielded the lowest mean squared error (MSE), MAE, and RMSE at 6521.31, 639.34, and 807.55 respectively, with the highest r-squared value at 0.3210.

Further utilisation of gradient boosting machine (GBM) resulted in improved performance, evidenced by lower MSE, MAE, and RMSE at 5210.60, 603.36, and 721.84 respectively, while maintaining the r-squared value. This indicates that the GBM model is better able to capture relationships between the features and target variable, age. Therefore, the model performs better with GBM and reduced the error metrics with higher predictive accuracy.

Research Question 4

As society progresses with better standard of living and healthcare, there is a greater risk of negligence for one's health. Coupled with an ever-increasing aging population as seen in records numbers across the world (Tahir, 2024), there is an urgent need to quickly identify heart disease. Implementing machine vision allows for quicker identification of heart diseases. As a proof of concept, we utilised a digit recogniser dataset, with the use of various algorithms to obtain the required parameters for machine vision. The training and testing datasets were iterated 500 times, with the aim for practicality and viability. It was able to achieve an accuracy score of 84.49% thereby validating the approach which was used. Checks on the data does show occasional errors however, the results do reaffirm the method that was used.



Considerations

Our approach has several strengths, including the utilisation of a diverse range of techniques for trend discovery and analysis and the use of valid foundational data. However, several considerations should be taken into account. Firstly, our heart disease study relies on an online data source rather than a first-hand data, thereby introducing potential errors. Secondly, we lack insight as to the lifestyle choices and ethnicity of the participants, thereby limiting the generalisation of our results to only specific demographic groups. Thirdly, the absence of access to the original participants raises the possibility of bias being at play. Lastly, in attempting to validate the notions of heart disease, bias will inherently arise from the selection of reports that contradicts our hypothesis, potentially skewing our conclusions.

Additionally, with any machine learning tool, a bias-variance trade-off will always be present. The selection of complex data will lead to overfitting and the selection of simplistic data will lead to underfitting. Hence, striking a balance between the bias and variance in our models is crucial to generalise the unseen data. Therefore, the findings in which we derive upon may not be representative of all ethnicity groups, and biases will be present from our conclusions.

Conclusion

Our research endeavours to have a comprehensive understanding of heart disease by introducing a variety of machine learning techniques. Through the decision tree and regression analysis, we are able to identify crucial indicators such as the a high maximum heart rate indicating heart disease and a high resting blood pressure not indicating heart disease. Data from the pair plot also suggests correlation between age, cholesterol levels, and maximum heart rate, in relation to asymptomatic individuals. However, data from SVM suggests that age does not play a significant factor into one's likelihood of developing heart related diseases.

Therefore, our findings suggest that a rise in standard of living over the past decades would have been a contributory factor in the increase of heart related disease due to changes in dietary habits, particularly evident in developed countries. This underscores the importance of regular exercise in reducing one's likelihood of developing heart disease, evident by the substantial negative impact of exercise angina observed in the regression line.

References

1. David S. Jones, Jeremy A. Greene, “The Decline and Rise of Coronary Heart Disease: Understanding Public Health Catastrophism”, American Journal of Public Health 103, no. 7 (July 1, 2013): pp. 1207-1218. DOI: 10.2105/AJPH.2013.301226

2. Liang, Y., Yu, D., Lu, Q., Zheng, Y., & Yang, Y., “The rise and fall of acute rheumatic fever and rheumatic heart disease: a mini review”, Frontiers in Cardiovascular Medicine, no 10 (Sec. Pediatric Cardiology), 2023. DOI: 10.3389/fcvm.2023.1183606

3. Ahern, R.M., Lozano, R., Naghavi, M. et al. “Improving the public health utility of global cardiovascular mortality data: the rise of ischemic heart disease”, Population Health Metrics 9, 8 (2011). DOI: 10.1186/1478-7954-9-8

4. Sarma, S., Sockalingam, S., & Dash, S., “Obesity as a multisystem disease: Trends in obesity rates and obesity-related complications”, (23 Feb 2021) DOI: 10.1111/dom.14290

5. Agha, Riaz, “The rising prevalence of obesity: part A: impact on public health”, International Journal of Surgery Oncology 2(7) e17, (Aug 2017). DOI: 10.1097/IJ9.0000000000000017

6. Tahir, F. (27 February 2024) 15 countries with declining birth rates in 2024 https://finance.yahoo.com/news/15-countries-declining-birth-rates-171707716.html