AUT

**Student ID Number:** 21151144

# Assignment 1

# Data Exploration and Classification

**Semester 1, 2025**

**Student Name:** Mirasha Fernando

**Student ID:** 21151144

**PAPER NAME:** Foundations of Data Science

**PAPER CODE:** COMP615

**Due Date:** Sunday 16 April 2025 (before midnight)

**TOTAL MARKS:** 100

**INSTRUCTIONS:**

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Using any other unfair means.
2. **Please email DCT.EXAM@AUT.AC.NZ and Teaching Team COMP615 <teachingteam.comp615@aut.ac.nz> if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately. Attach your assignment files as evidence of timely submission.**
3. **Attach all your code for all the datasets in the Appendix section**.

# Contents



## List of figures

# Task 1: Introduction

Heart diseases remain as one of the leading causes of death globally, with approximately eighteen million deaths each year according to the World Health Organization. Early diagnosis and timely treatment of heart diseases are particularly important for improving patient outcomes and reducing healthcare costs. Throughout this assignment, I aim to analyse the UCI Heart Disease dataset to explore patterns and build a classification model that predicts the presence of heart disease in patients based on various clinical and demographic attributes.

The core problem this dataset addresses is the early identification of individuals at risk of heart disease using historical data collected from patients which includes age, sex, resting blood pressure, cholesterol levels, chest pain type, and many more. This analysis has a global impact as the findings and models developed can be adapted to support early screening efforts in healthcare systems around the world.

The main goals of this project are,

- ✓ To conduct an exploratory data analysis (EDA) to understand the key attributes and their relationships with heart disease.
- ✓ To build and evaluate classification models that can predict heart diseases accurately.
- ✓ To identify which variables strongly contribute to the prediction, thereby supporting more informed medical decision-making.

Research questions that I am attempting to answer includes,

- ✓ Which patient attributes are most associated with heart disease?
- ✓ Can a classification model accurately predict whether a patient has heart disease?
- ✓ Which model suits the best for the prediction task?

Assumptions that are met,

- ✓ Missing values are handled.
- ✓ The dataset represents a general patient population.
- ✓ Independent variables

# Task 2: Data Exploration

This dataset consists of 14 attributes (13 features + 1 target) and 303 instances. Some features are categorical, and others are integers. Below gives a detailed description of the variables.

| Attributes | Description | Data type |
|---|---|---|
| **1. Age** | Age of the patient in years | Continuous |
| **2. sex** | Patient's sex<br>(0 = female, 1 = male) | Categorical |
| **3. cp** | Chest pain type<br>0 = typical angina<br>1 = atypical angina<br>2 = non-anginal pain<br>3 = asymptomatic | Categorical |
| **4. trestbps** | Resting blood pressure(mm Hg) | Continuous |
| **5. chol** | Serum cholesterol in mg/dl | Continuous |
| **6. fbs** | Fasting blood sugar > 120 mg/dl<br>(0 = false, 1 = true) | Categorical |
| **7. restecg** | Resting electrocardiographic results<br>0 = normal<br>1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)<br>2 = showing probable or definite left ventricular hypertrophy by Estes' criteria | Categorical |
| **8. thalach** | Maximum heart rate achieved during stress test | Continuous |
| **9. exang** | Exercise-induced angina (chest pain brought on by exercise)<br>(0 = no, 1 = yes) | Categorical |
| **10. oldpeak** | ST depression induced by exercise relative to rest (how much the ST segment is depressed during exercise) | Continuous |
| **11. slope** | Slope of the peak exercise ST segment<br>0 = upsloping<br>1 = flat<br>2 = down sloping | Categorical |
| **12. ca** | Number of major vessels (0-3) coloured by fluoroscopy | Categorical |
| **13. thal** | Thalassemia status<br>1 = normal<br>2 = fixed defect<br>3 = reversible defect | Categorical |
| **14. num** | Diagnosis of heart disease<br>0 = no heart disease<br>1–4 = presence of heart disease (increasing severity) | Categorical |

*Table 1 - Attributes*

```
Number of missing values:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          4
thal        2
num         0
dtype: int64
```

*Figure 1 - missing values.*

Figure 1 shows that columns ca and thal has missing values. In this case it is better to remove the rows with missing values. The dataset is now down to 297 rows.

Below figure gives the summary statistics of the continuous numerical features.

| | age | trestbps | chol | thalach | oldpeak |
|---|---|---|---|---|---|
| count | 297.000000 | 297.000000 | 297.000000 | 297.000000 | 297.000000 |
| mean | 54.542088 | 131.693603 | 247.350168 | 149.599327 | 1.055556 |
| std | 9.049736 | 17.762806 | 51.997583 | 22.941562 | 1.166123 |
| min | 29.000000 | 94.000000 | 126.000000 | 71.000000 | 0.000000 |
| 25% | 48.000000 | 120.000000 | 211.000000 | 133.000000 | 0.000000 |
| 50% | 56.000000 | 130.000000 | 243.000000 | 153.000000 | 0.800000 |
| 75% | 61.000000 | 140.000000 | 276.000000 | 166.000000 | 1.600000 |
| max | 77.000000 | 200.000000 | 564.000000 | 202.000000 | 6.200000 |

*Figure 2 - summary statistics of continuous variables.*

If we look at the mean and standard deviation of age, we can see that majority of patients are middle-aged to older adults. Resting blood pressure (trestbps) seems to be elevated on average as the mean is above 120 mm Hg and the max is 200 mm Hg which indicates some extreme hypertensive cases. Mean cholesterol is above desirable levels (<200 mg/dl), with a wider spread (standard deviation = 51.9976). The maximum is 564 which is a significant outlier, suggesting some patients have extremely high risk. Average maximum heart rate (thalach) under stress tests is within range and is wide, indicating variability. The median is 0.8, but the distribution is right skewed in ST depression (oldpeak).

Cholesterol and oldpeak both show positive skew, driven by high maxima (564 and 6.2 respectively). These outliers might influence models, therefore its best if we treat them. Trestbps also has some outliers (200 mm Hg).

We will now use some visualizations to better understand the variables, their spread, outliers, and skewness.

Below gives the boxplots of the continuous variables.

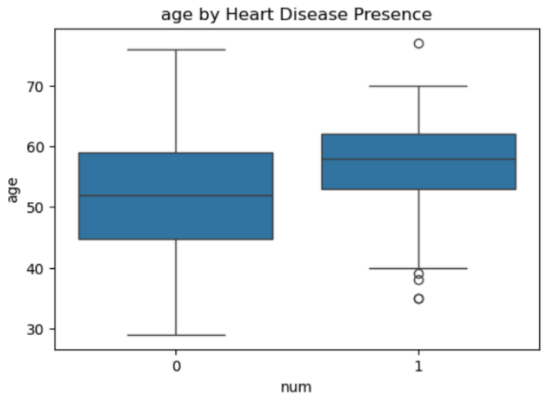


*Figure 3 - boxplot of age vs heart disease presence presence.*

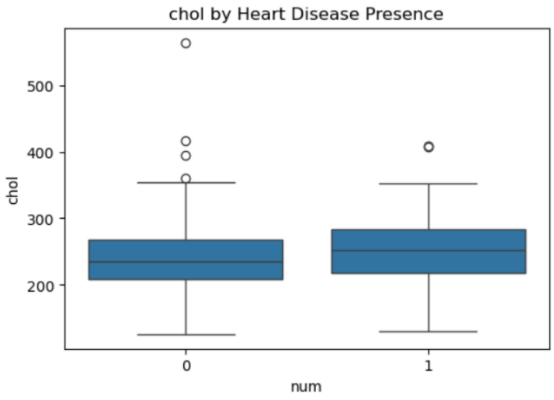


*Figure 4 - boxplot of cholesterol vs heart disease*

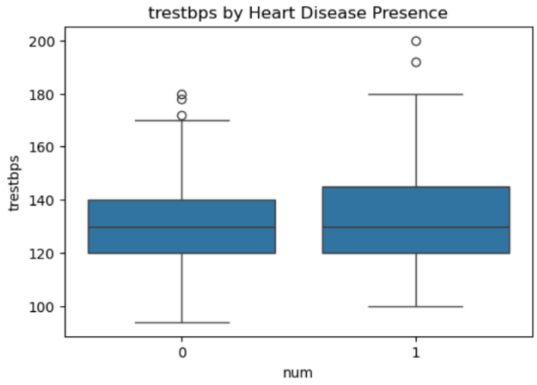


*Figure 5 - boxplot of trestbps vs heart disease presence*

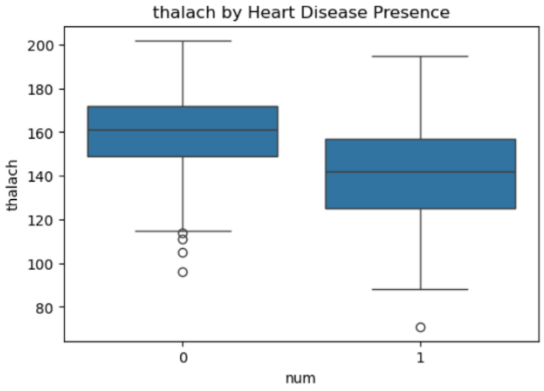


*Figure 6 - boxplot of max heart rate vs heart disease presence*

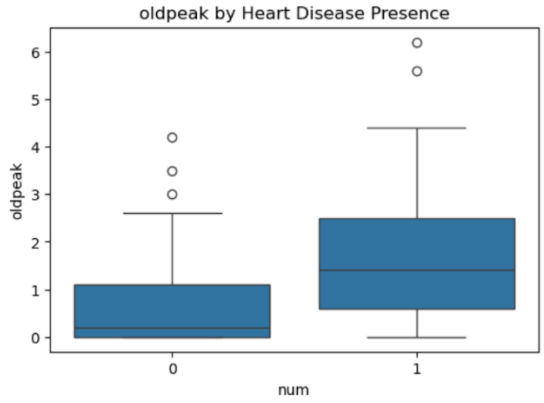


*Figure 7 - boxplot of ST depression vs heart disease presence.*

Figure 3 shows us the boxplots of age; heart disease tends to be more common in older patients as it is right skewed with a higher median and more older individuals. In the chol boxplot, we can see several outliers but “no disease” has more extreme high outliers which suggests that cholesterol levels alone may not clearly distinguish between disease and no disease. Still, individuals without disease can also have very high levels of cholesterol (564), suggesting other factors are likely interacting.

Resting blood pressure does not appear to vary much between the two groups. The interquartile ranges overlap, and the median is also similar. If we look at the maximum heart rate boxplot, we can see that “with disease” group has a lower median which explains that individuals with heart disease tend to achieve lower maximum heart rates while healthy individuals depict higher heart rates. ST depression seems to have a strong

association with the presence of heart disease. “No disease” is concentrated near zero with a tight spread.
Now let us look at the histograms of the same variables to examine their spread and skewness.

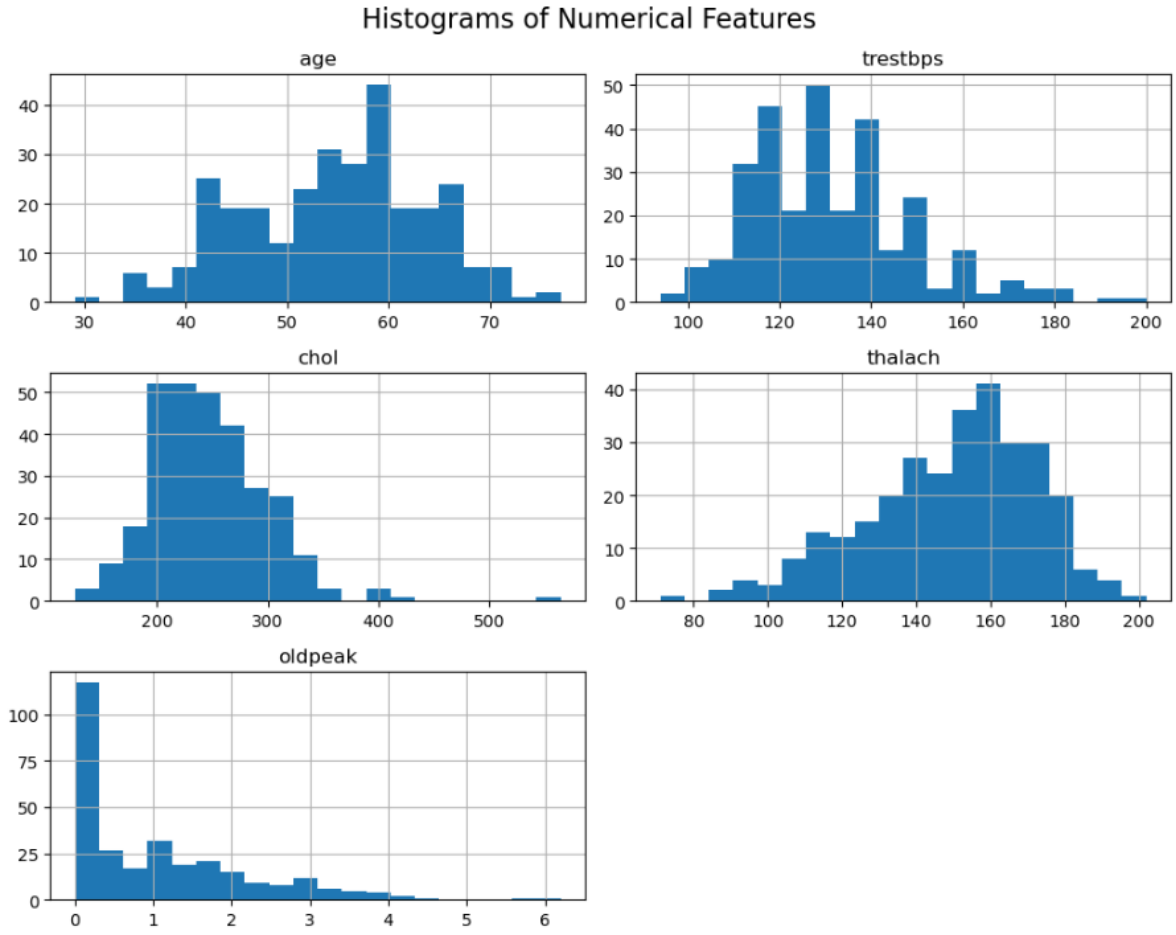


*Figure 8 - histograms of the continuous variables.*

The age histogram has a slight bell-shaped distribution with a majority of patients ranging from 50 – 60 years. Age is a meaningful feature. Older individuals have a higher risk of heart disease. Resting blood pressure ranges from 120-140 mm Hg. We can note the maximum value of two hundred (outlier) which makes the histogram skewed. Cholesterol (chol) histogram has a clear bell-shaped curve with several outliers. Mostly the cholesterol levels range from 200-300 mg/dl. The maximum heart rate histogram has a left-skewed distribution with a peak around 140-160 bpm. ST depression is strongly right skewed. This can be considered as a key variable in this analysis.

We can further confirm if these variables have outliers by using their z-scores. Below gives a list of data where the absolute z-score is greater than the default threshold (3).

```
Outliers in 'trestbps':
     trestbps
126       200
188       192

Outliers in 'chol':
     chol
48    417
121   407
152   564
181   409

Outliers in 'thalach':
     thalach
245       71

Outliers in 'oldpeak':
     oldpeak
91       6.2
123      5.6
```

*Figure 9 - outlier detection using z-score.*

We can see that trestbps has two outliers which explains the skewness depicted in the histograms. Chol has four outliers with an extreme value of 564, this should be treated as it can has a major effect when it comes to building a model. Thalach only has one outlier and by considering the histogram, if we can replace this outlier with the median, it will give a more normal distribution to thalach. Oldpeak has two outliers with an extreme value of 6.2. the distribution is strongly skewed therefore a transformation might reduce the skewness.

After considering the above visualizations and stats, we can conclude that it is better to transform oldpeak as it is highly skewed. We will replace all other outliers in chol, thalach and trestbps with the median and check whether the distribution is normal. Below figures show the distributions after outliers are treated.

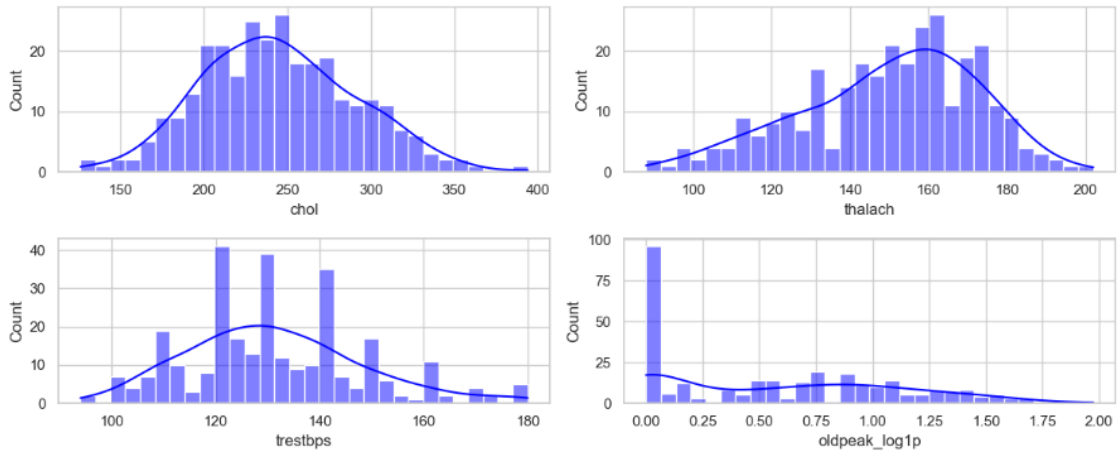


*Figure 10 - distribution of chol, thalach, trestbps, and oldpeak after treating for outliers.*

As you can see after the transformations, the chol distribution appears to be less skewed, closer to normal, and approximately symmetric. The thalach distribution also looks more stabilized compared to the original distribution. The trestbps distribution without the outlier appears to be equally spread and approximately symmetric. The log transformation on oldpeak successfully reduced the skewness and made the variable more suitable for modelling. Although it still appears to be right skewed, the long tail is compressed and the spike at 0 is less dominant. Now all our variables are cleaned and ready for exploration.

We can use correlation matrix to understand the relationship between the variables.

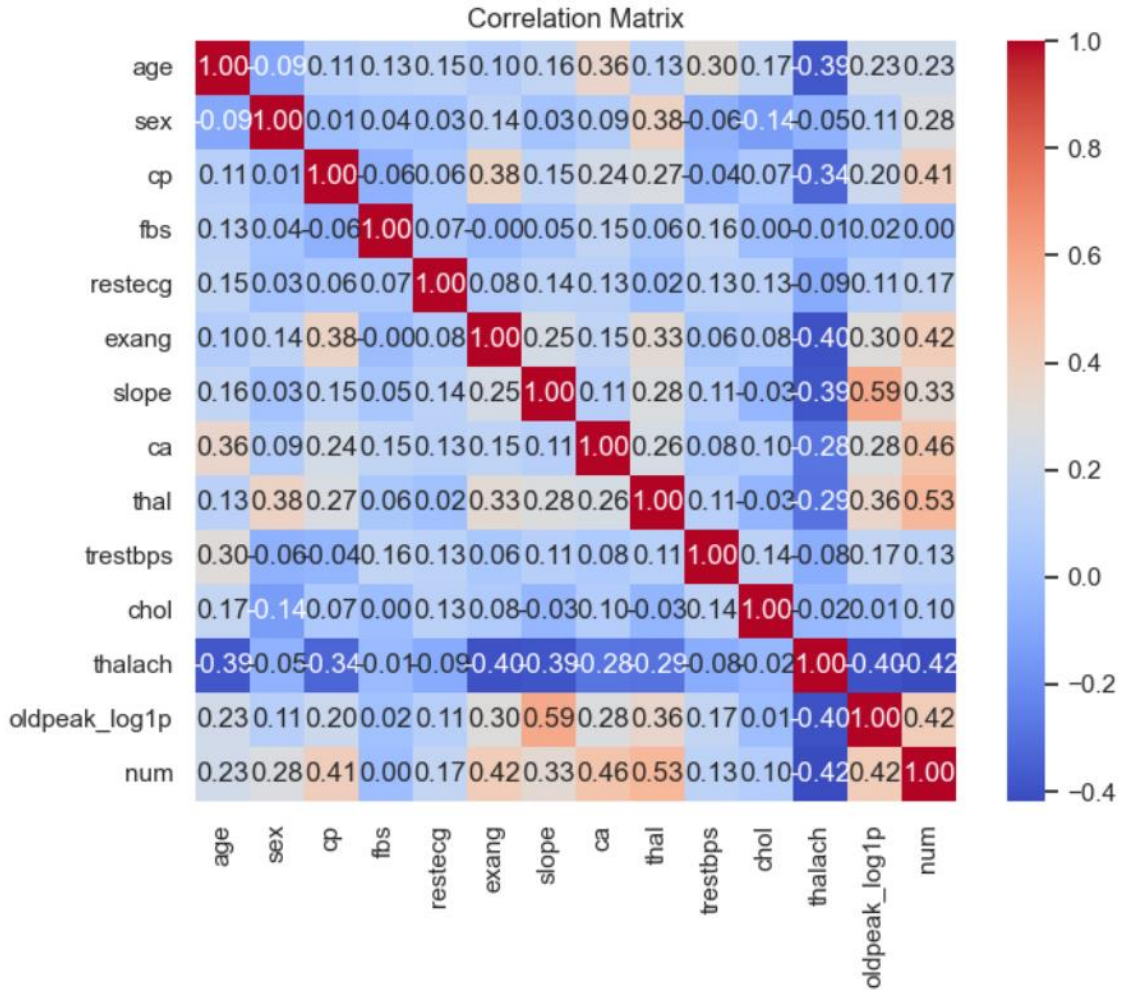


*Figure 11 - correlation matrix of all the variables.*

First if we look at the correlation between the target variable and the rest, we can select below variables as the predictors with highest correlation with the target variable.

**Age, sex, cp, exang, slope, ca, thal, slope, thalach, and oldpeak**

Compared to other variables these have high correlation therefore it is important to consider them when modelling. But if we actively look, we can see that slope and oldpeak has a high correlation between each other. Therefore, it is better to select only on of them for modelling as the multicollinearity between them might affect the accuracy of the model. On this occasion we will go forward with oldpeak instead of slope.

The following features were selected based on statistical insights from the dataset.

- **Age** – Risk of heart diseases generally increase with age. It is a known factor in heart disease.
- **Sex** – Biological sex influences heart disease prevalence and symptoms.
- **Chest pain type (cp)** – The type of chest pain is a key clinical indicator of potential heart problems.
- **Maximum heart rate achieved (thalach)** – Heart's response to stress. Strong negative correlation with the target. Lower max heart rate has higher chance of disease.
- **ST depression (oldpeak)** – High correlation with target.
- **Exercise-induced angina (exang)** – Presence strongly associated with disease.
- **Number of major vessels (ca)** – Strong positive correlation. Likely very predictive.
- **Thalassemia status (thal)** – Strongest predictor therefore very important.
- **Target (num)** – This is the outcome variable.

The above variables were chosen because they are not only clinically significant, but they also showed meaningful variation and patterns during the exploratory data analysis.

By conducting visualisations and computing summary statistics, we were able to identify patterns, outliers, and feature behaviours relevant to predicting the presence of heart disease.

Histograms and descriptive statistics depicted that features like cholesterol (chol) and ST depression (oldpeak) were right skewed, indicating the presence of outliers and non-normality. Maximum heart rate (thalach) appeared more normal distributed although it showed few extreme values. Applying log transformations to the variables that showed skewness and outliers, made them more stabilized and normal.

Using the correlation matrix, we were able to examine the relationships between each predictor and the target variable. It was notable that thal (thalassemia), ca (number of major vessels), cp (chest pain type), exang (exercise-induced angina), oldpeak (ST depression), and thalach (maximum heart rate) showed relatively strong correlations with the target variable. These variables are likely to be important predictors in a classification model.

Finally, we were able to clean the dataset with no missing values and treat variables with unusual patterns by transforming them. We were able to improve the distribution for model training and select key features that are important for modelling.

# Task 3: Classification Models

a) You are required to report your preprocessing steps. The steps should include identifying any missing/duplicate data or outliers. Provide explanations of how you dealt with them.

As we checked during the data exploration, there were no duplicates, and we removed the rows with missing values. Below given is proof that there are no duplicates and missing values.

```
Duplicates: 0
Missing values:
 age              0
sex               0
cp                0
trestbps          0
chol              0
fbs               0
restecg           0
thalach           0
exang             0
oldpeak           0
slope             0
ca                0
thal              0
num               0
chol_log          0
oldpeak_log1p     0
oldpeak_boxcox    0
dtype: int64
```

*Figure 12 - duplicates and missing values.*

We also found that there were several outliers and skewed variables. Therefore, we decided to do a log transformation on oldpeak, and it made the distribution closer to normal and the spike at zero became less dominant. For chol, thalach, and trestbps we decided to replace the outliers with the median and it made the distributions stabilized. The extreme values were replaced by median and made the distribution more normal reducing the skewness.

b) Create a model using the Decision Tree algorithm. Adjust two suitable parameters (one at a time) to reduce the tree's size and improve your model's accuracy. Report the accuracy score for each parameter using the plots. Provide the final optimised classification tree and describe its structure.

To create the most perfect model, it is appropriate to use the features we selected from our data exploration.

To create the model and improve its performance and size, two parameters were adjusted one at a time. Below figure gives us two plots depicting the accuracy at each test value.

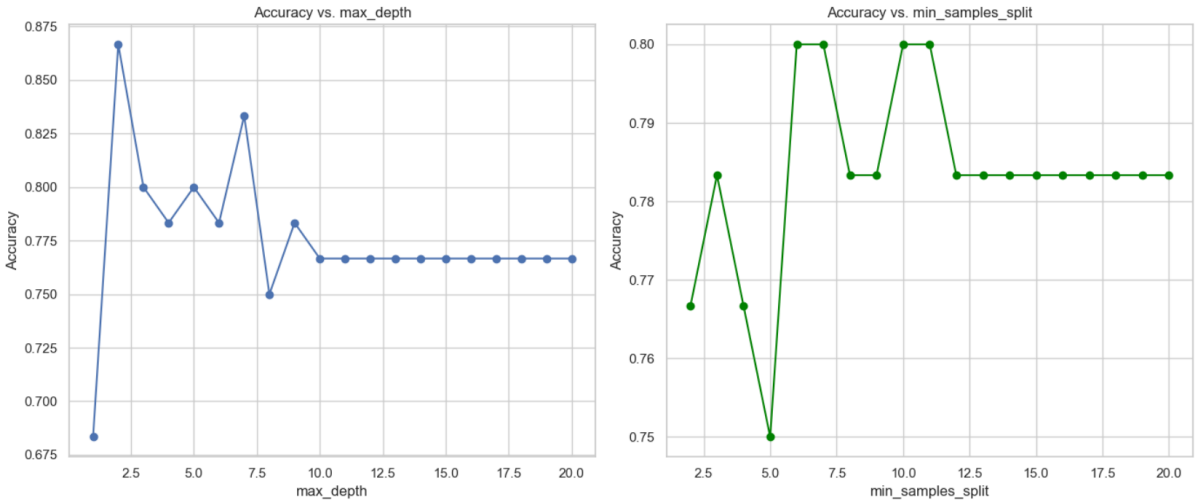


*Figure 13 - accuracy of each parameter.*

- Parameter one

  max_depth – controls the maximum depth of the tree. Tested values from 1 to 20. Maximum depth of 2.0 has the highest accuracy of approximately 0.865 and will help create an accurate model.
- Parameter two

  Min_samples_split – gives the minimum number of samples that is required to split a node. Tested values from 2 to 20. A split value around 6.0 with the highest accuracy of 0.80 improves generalization.

Now we can create the final optimised classification tree using the above values.

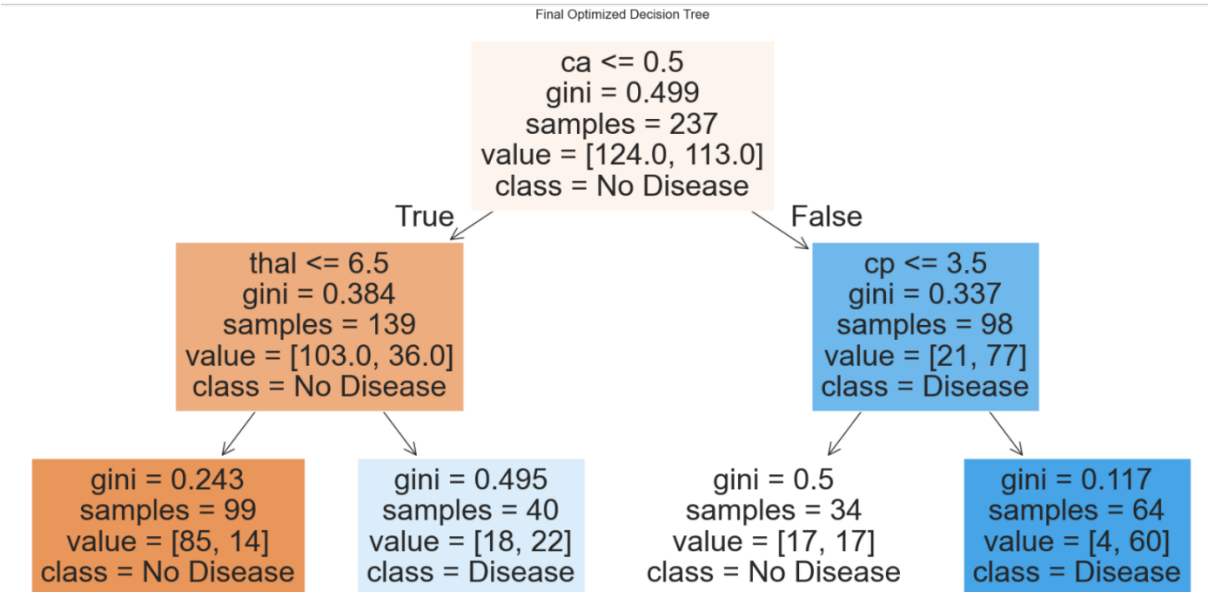


*Figure 14 - final optimised classification tree.*

As we can see the final decision tree optimized its splits using the most predictive features: ca, thal, and cp.

The tree starts with the root node with the most influential feature: ca. patients with <= 0.5 major vessels were predicted as not having heart disease. If we move onto the left subtree (low ca) which is split by thal <= 6.5 means patients with low thalassemia values

(thal <= 6.5) were mostly classified as not having heart disease. The right subtree (high ca), split by cp <= 3.5 means patients with high chest pain scores i.e. cp >= 3.5, were strongly predicted to have heart disease.

Each node in the tree has four types of information.

1. Gini index: lower values indicate better classification.
2. Samples: number of observations at the node.
3. Value: counts of each class [No disease, Disease].
4. Predicted class: the majority class at that node.

This model achieved a balance between simplicity and predictive power. It only uses three key features which makes it easy to interpret. It produces relatively pure splits in most terminal nodes.

c) Describe the role of the two parameters in the model building you used in part b) above. Do you expect that using the same values obtained for this dataset will improve the accuracy of other datasets? Justify your answer.

**Parameter one: max_depth**

This parameter controls the maximum depth of the tree which tells us how many levels it can grow. This limits how complex the tree can become. A larger max depth allows the tree to capture more patterns but risks overfitting while a smaller max depth reduces overfitting. The max depth used in this model is 2.0 which balances the complexity with generalization.

**Parameter two: min_samples_split**

This parameter gives us the minimum number of samples required to split an internal node. It prevents the tree from splitting nodes that do not have enough data to justify a meaningful split. It helps avoid overfitting to small subgroups. Tuning this helped improve robustness in our model.

**Parameter values from this dataset will not improve accuracy on other datasets.** The values of max_depth and min_samples_split depend on the structure and complexity of the dataset. A dataset with more features or more complexity might require different values. Other datasets might contain distinct levels of class, relationships, or patterns. A tree like this might underfit more complex data elsewhere.

Therefore, using the same parameter values without tuning on a new dataset may lead to overfitting or underfitting. So, it is always best to re-tune parameters using cross validation when applying the model to a different dataset. This ensures that the parameters best suit the specific characteristics of the new dataset.

d) Find the feature importance based on the final classification model and explain your findings.

Figure 15 & 16 gives the feature importances of the features used in the model and the plot.

```
ca                0.406790
cp                0.196577
oldpeak_log1p     0.134318
thal              0.120636
thalach           0.063117
age               0.049068
sex               0.018447
exang             0.011046
dtype: float64
```

*Figure 15 - feature importances.*

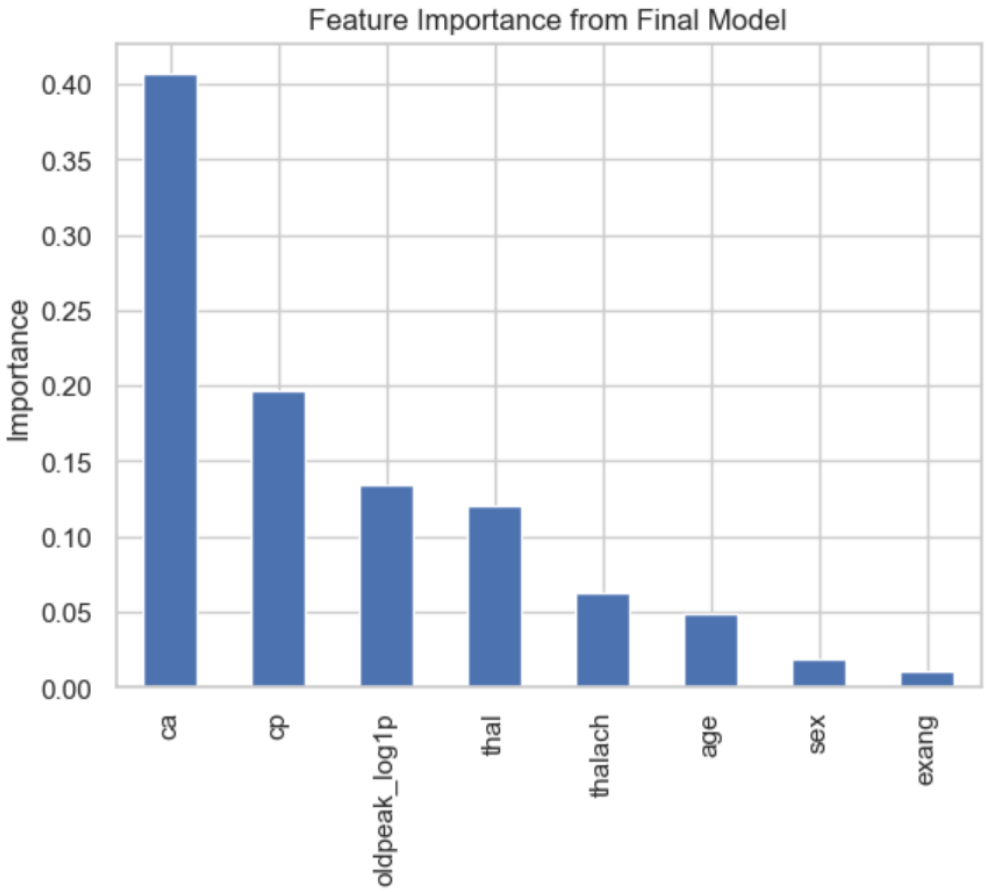


*Figure 16 - feature importances in a plot.*

The most important feature contributing over 40% of the model's decision power is ca (number of major vessels coloured by fluoroscopy). blocked vessels are a strong indicator of heart disease risks. The second most important variable which is cp has approximately 20% importance. This makes sense as types of chest pain are clinically relevant symptoms of heart issues. Oldpeak (ST depression induced by exercise) has an importance of about 13% and thal (thalassemia) has an importance of about 12%. Thalassemia is a blood disorder that influences heart health. Thalach, age, sex, and exang has low importances. These features still add value but are less impactful compared to the top four.

The model relies mostly on clinical test results (ca, cp, thal, oldpeak) rather than basic demographics (age, sex) which is an encouraging sign of model validity.

e) Generate and carefully examine the Confusion Matrix and explain your findings. Provide the model summary report and discuss the metrics (accuracy, precision, recall, and F1- score).

Below gives the confusion matrix of the model we created.

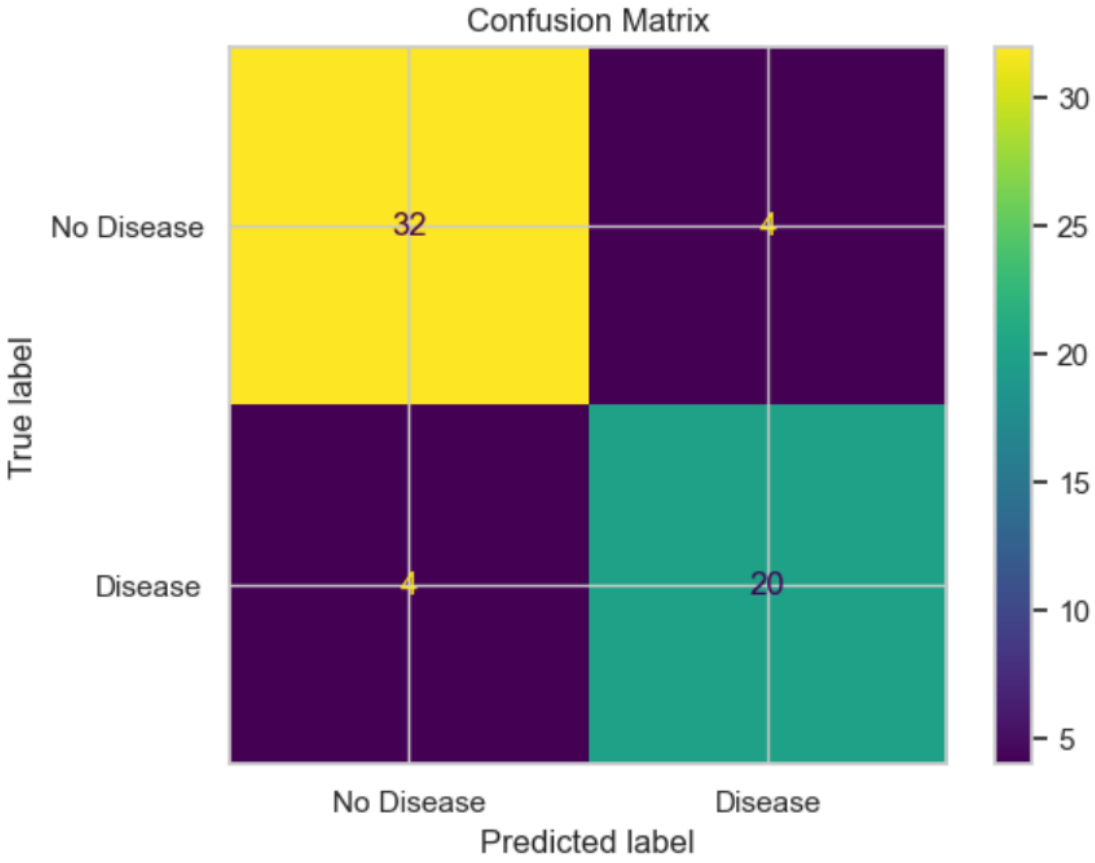


*Figure 17 - confusion matrix.*

A confusion matrix helps to access the model's performance beyond just accuracy. Given below is a breakdown of the confusion matrix.

- True negative: the model predicted No Disease and actual was No Disease – 32.
- True positive: the model predicted Disease and actual was Disease – 20.
- False positive: the model predicted Disease but actual was No Disease – 4.
- False negative: the model predicted No Disease but actual was Disease – 4.

Figure 18 gives the model summary report with the metrics: accuracy, precision, recall, and F1- score.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.89      0.89        36
           1       0.83      0.83      0.83        24

    accuracy                           0.87        60
   macro avg       0.86      0.86      0.86        60
weighted avg       0.87      0.87      0.87        60
```

*Figure 18 - model summary report.*

Here is a breakdown of the classification report.

- Precision: 0.83 for class 1, which means when the model says this person has heart disease its correct 83% of the time. When the model says this person does not have heart disease, it is correct 89% of the time.
- Recall: 0.83 for class 1, which means it detects 83% of all the people who actually have the disease. The models detect 89% of all the people who does not have the disease.
- F1-Score: this is 0.83 which is good as the model has a balanced ability to both catch and correctly predict heart disease.
- Weighted average: class zero has more instances (36 out of 24) therefore it is given more weight.
- Accuracy: overall accuracy of the model is 87% which is good.

The model performs very well as he overall accuracy is 87% and handles both classes reasonably. Its slightly better at predicting people who do not have heart diseases (class 0 – No Disease). The recall for class one is not perfect as four out of twenty-four people with heart disease were missed (false negatives), which might be an issue when it comes to medical settings.

# Task 4: Results and Discussions

In this analysis we built a decision tree classifier to predict the presence of heart disease using selected features from the UCI Heart Disease dataset. The features used – ca, cp, oldpeak_log1p, thal, thalach, age, sex, and exang – were chosen based on their strong correlation with the target (num) variable. The goal was to optimize model performance by tuning parameters and evaluating performance through the confusion matrix and the various metrices in the summary report (accuracy, precision, recall and F1-score).

After cleaning the dataset by handling missing values, duplicates and encoding categorical variables, we were able to apply a decision tree classifier. We used two parameters max_depth and min_samples_split and tuned them one at a time accordingly. These parameters directly affect the size and complexity of the model. Max depth limits how deep the tree can grow. It helps prevent overfitting by simplifying the model. On the other hand, min_samples_split sets the minimum number of samples required to split an internal node. This controls how finely the data is partitioned at each node.

We used 10-fold-cross-validation as the testing/training validation model. This method splits the data into ten equal parts, trains the model on nine parts, and tests it on the remaining one. This process is rotated ten times. A reliable estimate of the model effectiveness on unseen data is given by the average performance across all folds.

After parameter tuning the final optimized decision tree model achieved an overall accuracy of 87% suggesting that the model performs reliably. These metrics indicate that this model is well-balanced. It is notable that the precision and recall scores are relatively high. By limiting max depth and adjusting min_samples_split, I was able to reduce overfitting and improve interpretability while maintaining accuracy.

The confusion matrix also supports this as eight samples were misclassified out of sixty test samples which is a low number but compared to medical context it can be risky as it is risky to miss someone who needs care. The feature importance analysis showed that ca (number of major vessels coloured by fluoroscopy) and cp (chest pain type) were the most influential featuring in predicting heart disease. This aligns with medical understanding and confirms that model relies more on clinical insights.

A potential future research point would be assessing the model on heart disease data from different hospitals or demographic groups to evaluate its generalizability. Also looking for new knowledge such as new clinical metrics could enhance the model performance further.

Finally, the optimized decision tree model showed impressive performance with balanced metrics making it a reliable tool for preliminary heart disease screening. However, it is recommended to prioritize even high recall even at the cost of slightly lower precision and undergo extensive validation across diverse populations.