Cardiovascular System Disease

FDAE Team 7
Lee Choonggi
Lee Pei Xin
Muhammad Rais Bin Aminuddin

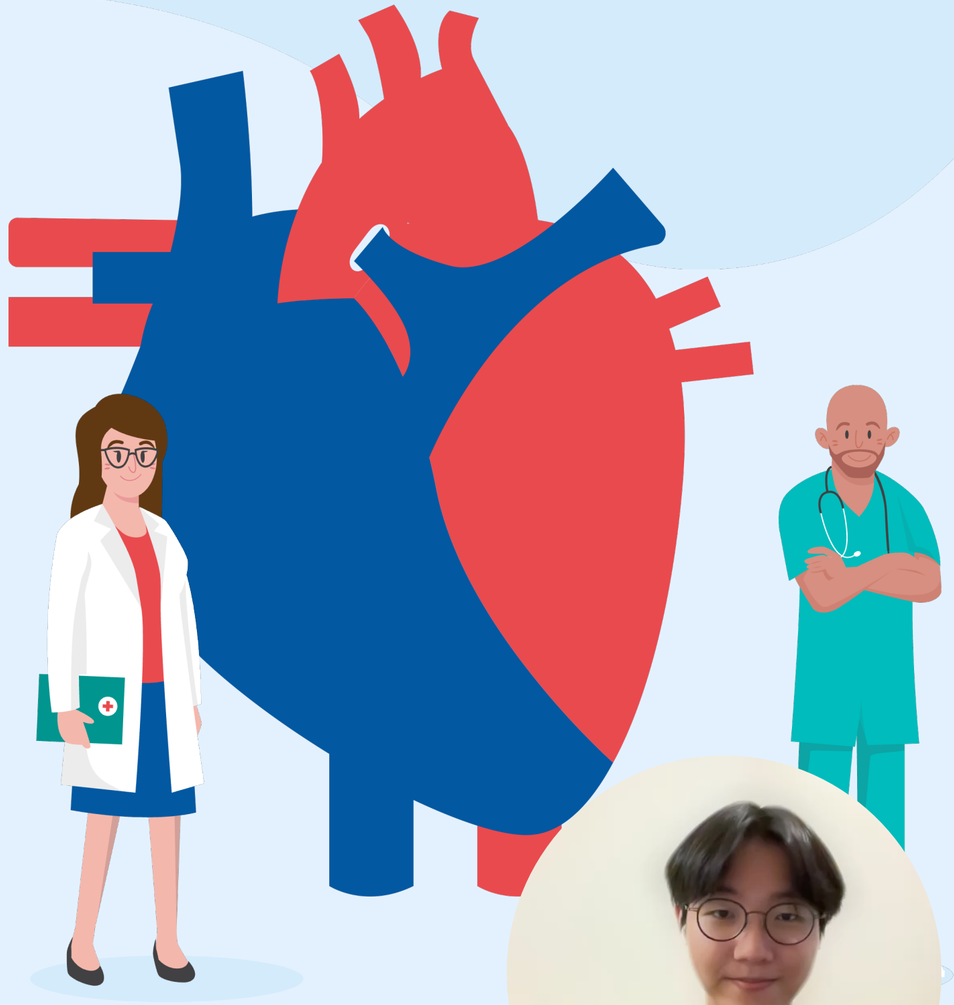


Table of contents

01 Practical motivation & Problem formulation

02 Data preparation
- includes data cleaning
- categorising data

03 Exploratory Data Analysis
- Analytical Visualisation
- Statistical Description

04 Machine Learning
- Logistic Regression
- Random Forest
- SVM

05 Takeaways

06 Conclusions

Sample collection & Practical Motivation

30% of population

Suffer from heart disease in 2022

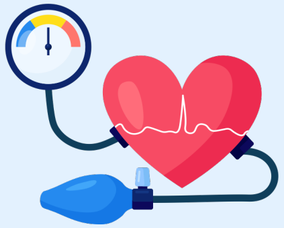
What if we could make a change?

What if we could develop tools that help doctors identify those at risk early on?

Problem Formulation

Problem statement

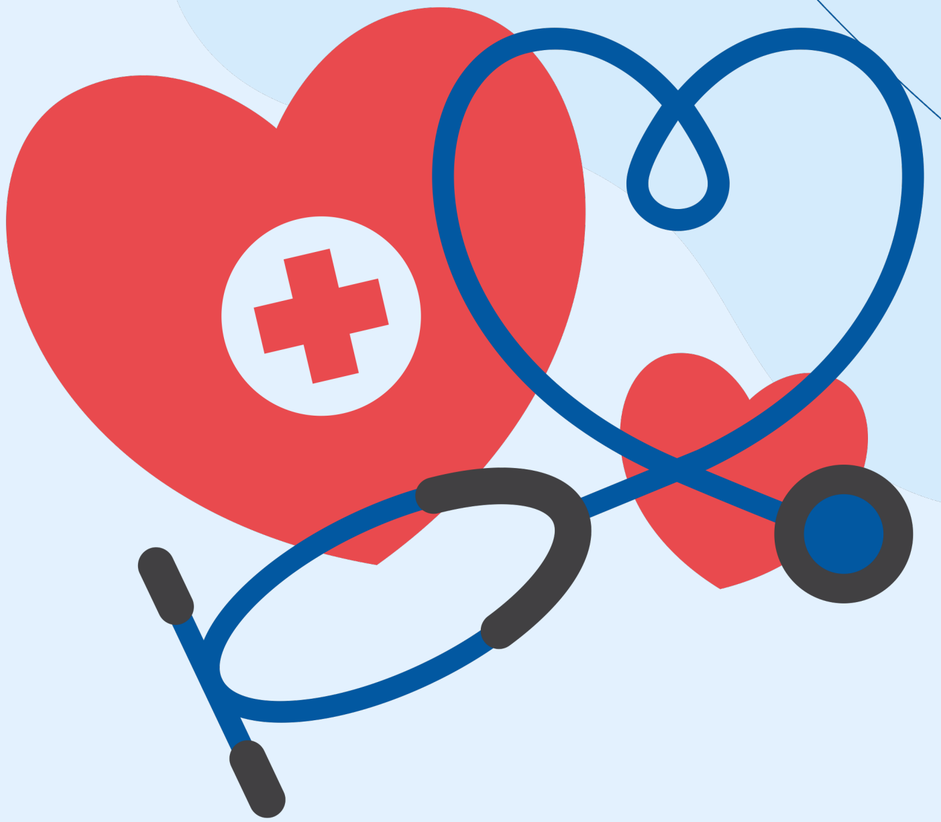
To **develop a predictive model** based on **physical characteristics** like age, blood pressure, cholesterol levels, etc, that can accurately identify the risk of heart disease for an individual.



Save life and improve quality of lives

01

About our dataset

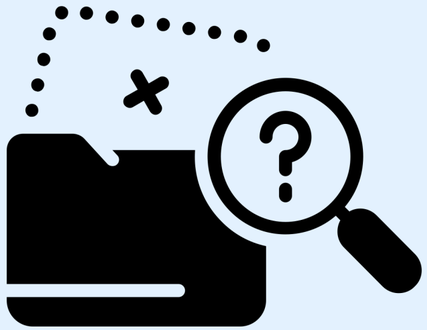


Source: Kaggle

Title: Heart

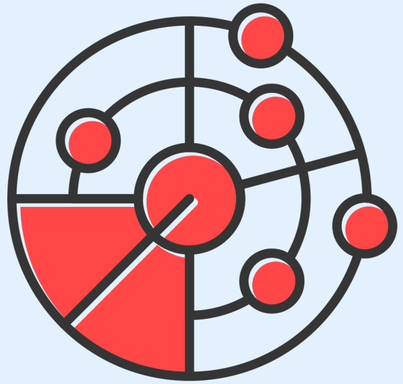
Author: desalegngeb

Data cleaning



check missing data (NaN)

drop out NaN and values out of range



Check for whether all values fell under the correct range.

- e.g.: age not exceeding 100

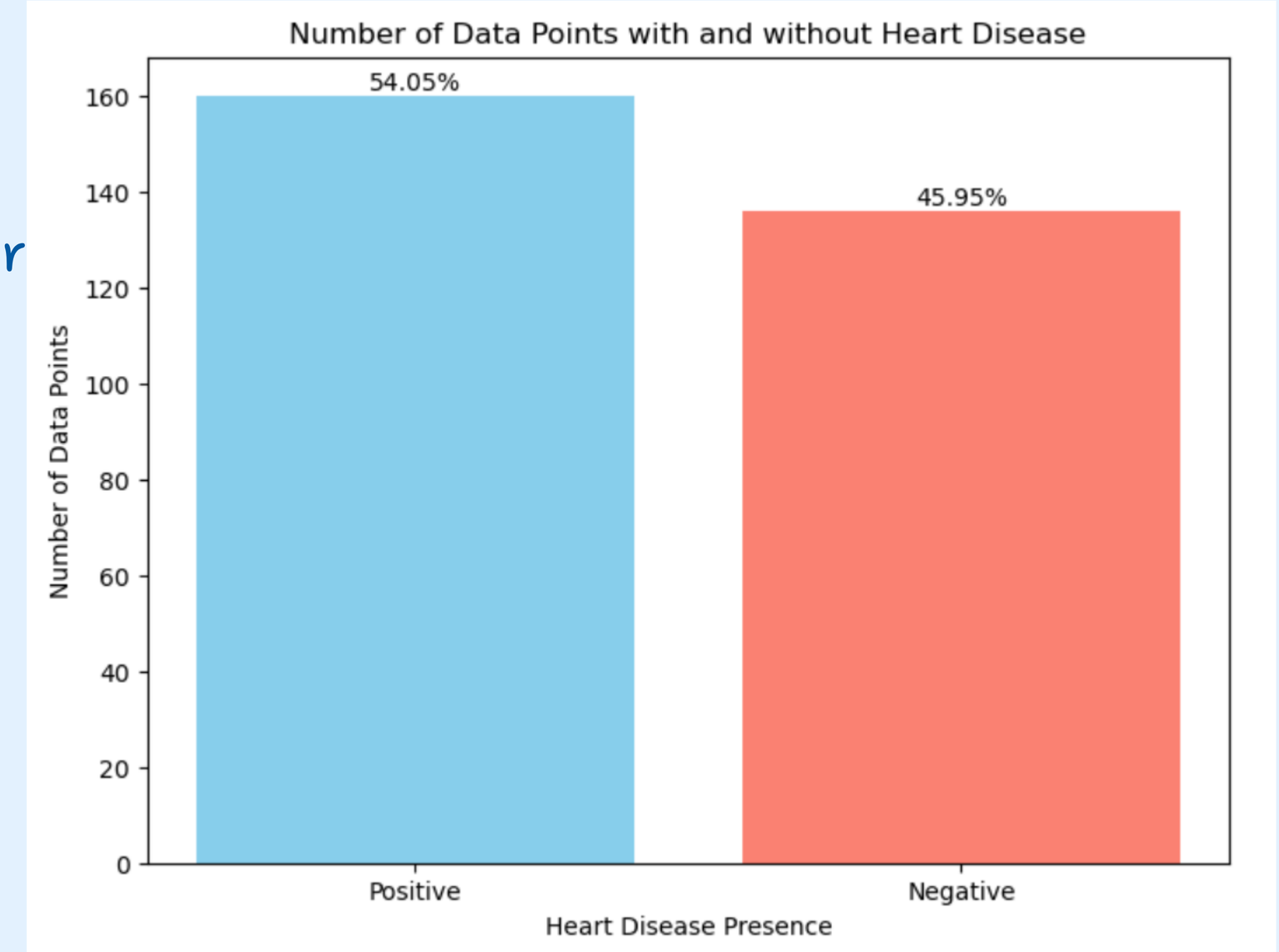
Original #datapoints =303
After dropping, #datapoints = 296

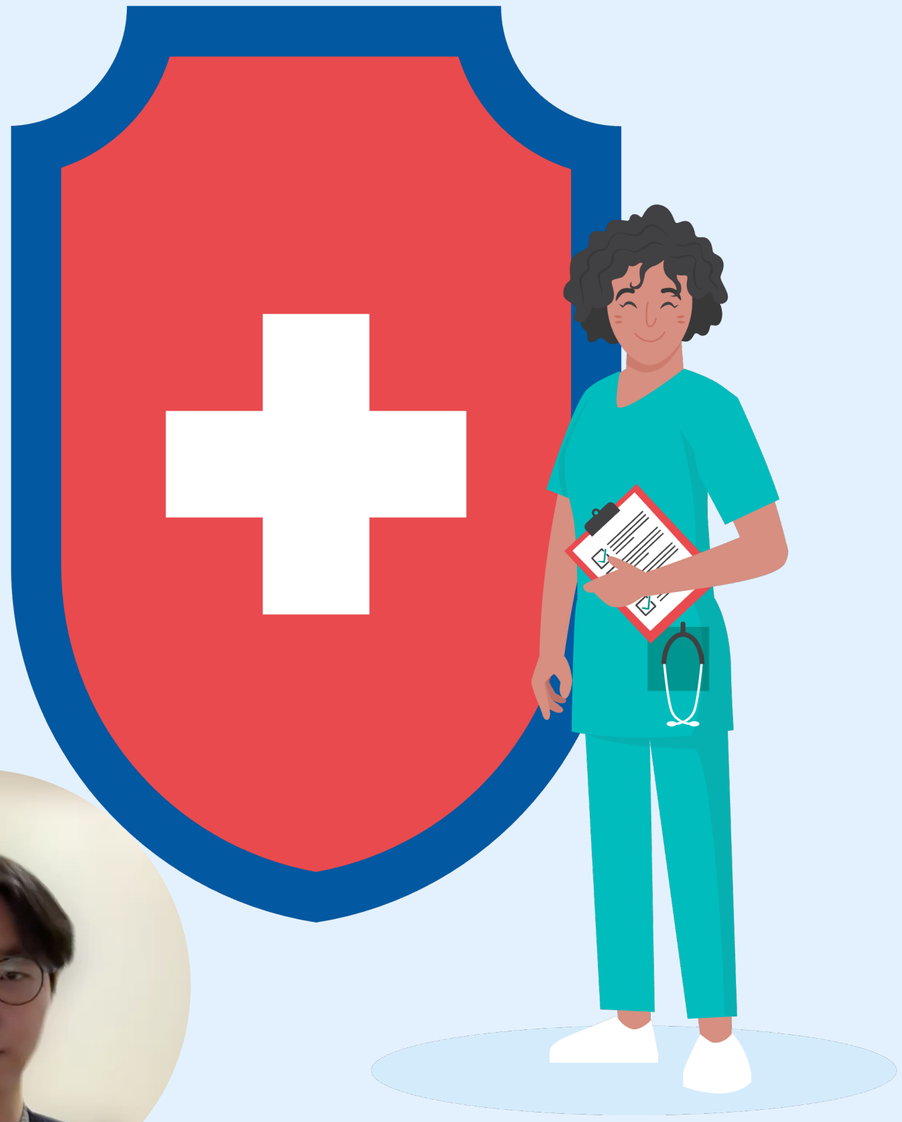
Data cleaning

Explore Data Balancing

A rather balanced dataset for our prediction factors.

54% vs 46%





02

Exploratory Data Analysis

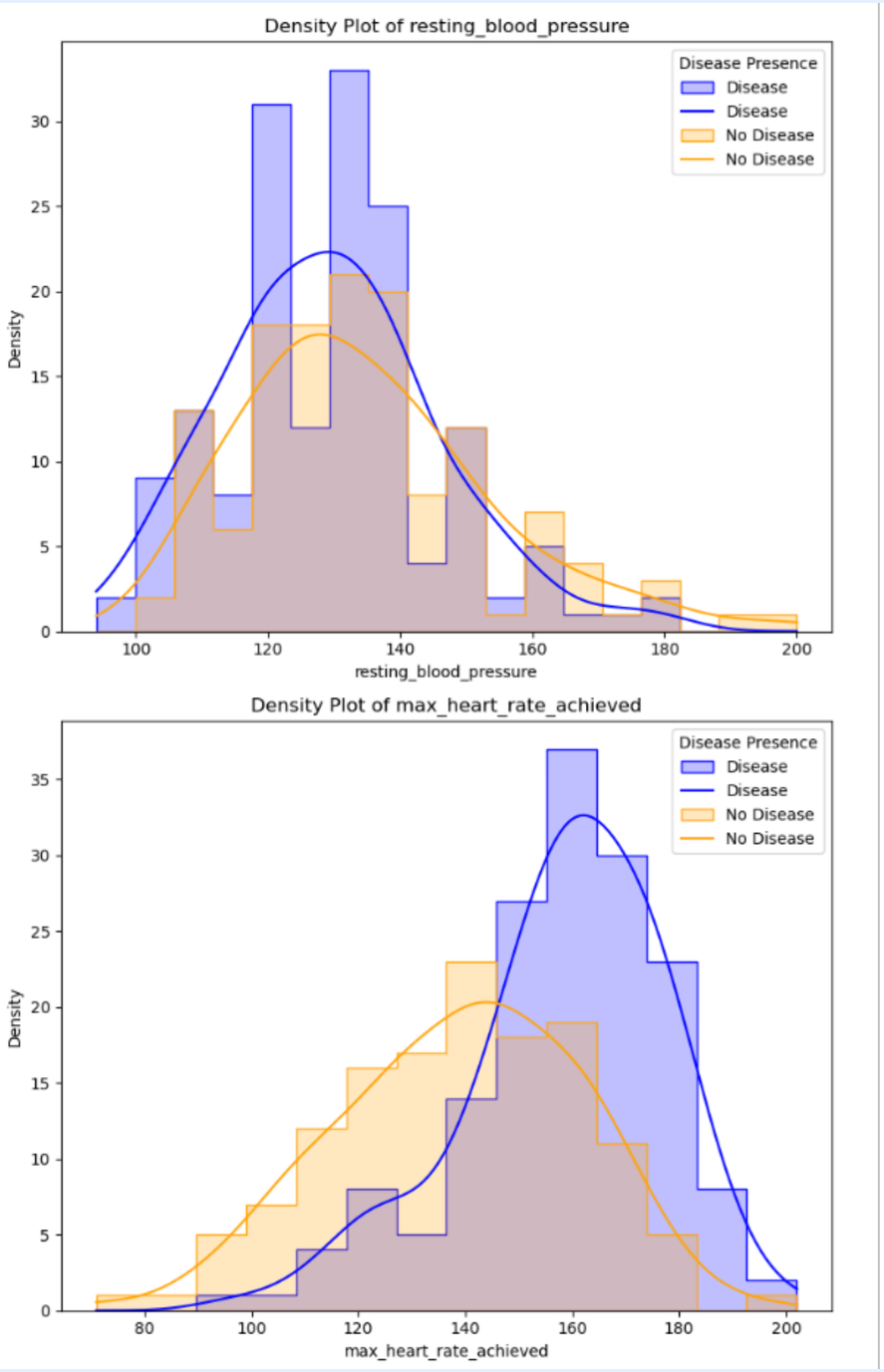
Numeric Variables

Histogram & KDE Plot

1. a smooth, continuous representation of data distributions.

2. clearer insights into the underlying data patterns.

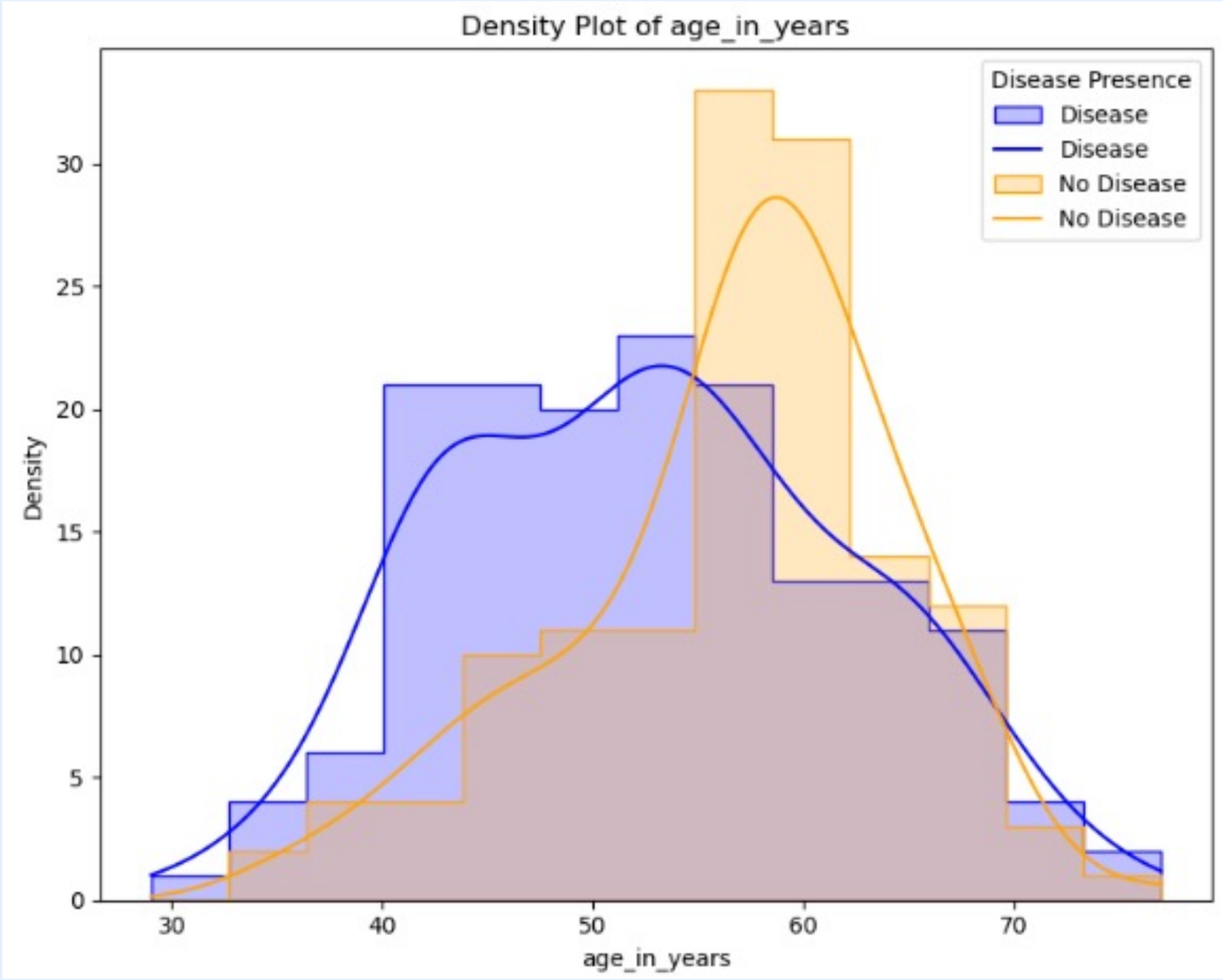
Example of our plots



Interesting Insights!!!

Dual peaks

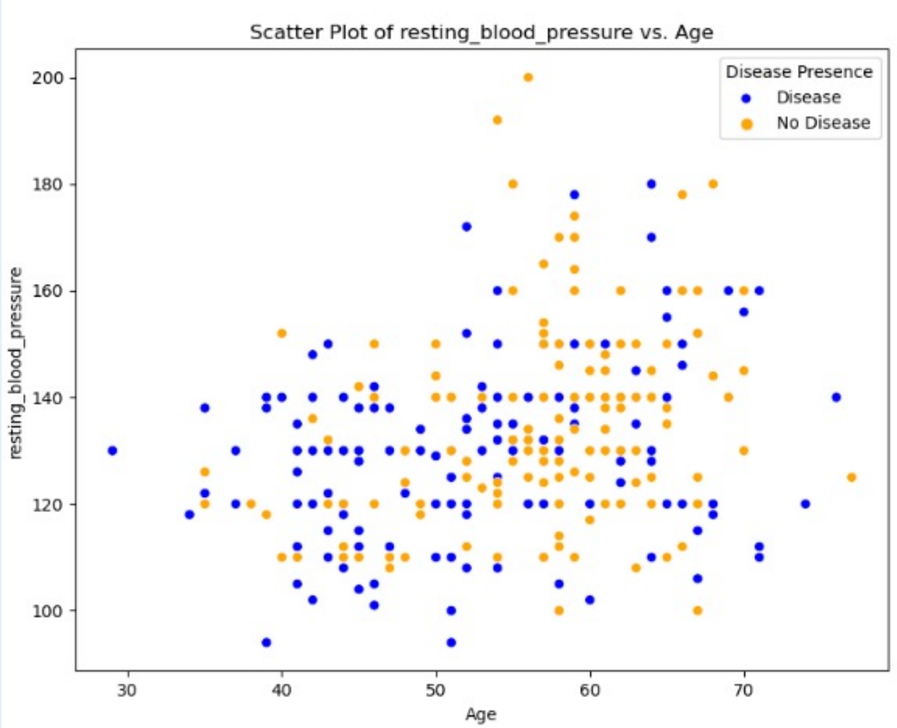
Younger individuals also have a high risk of getting heart disease.

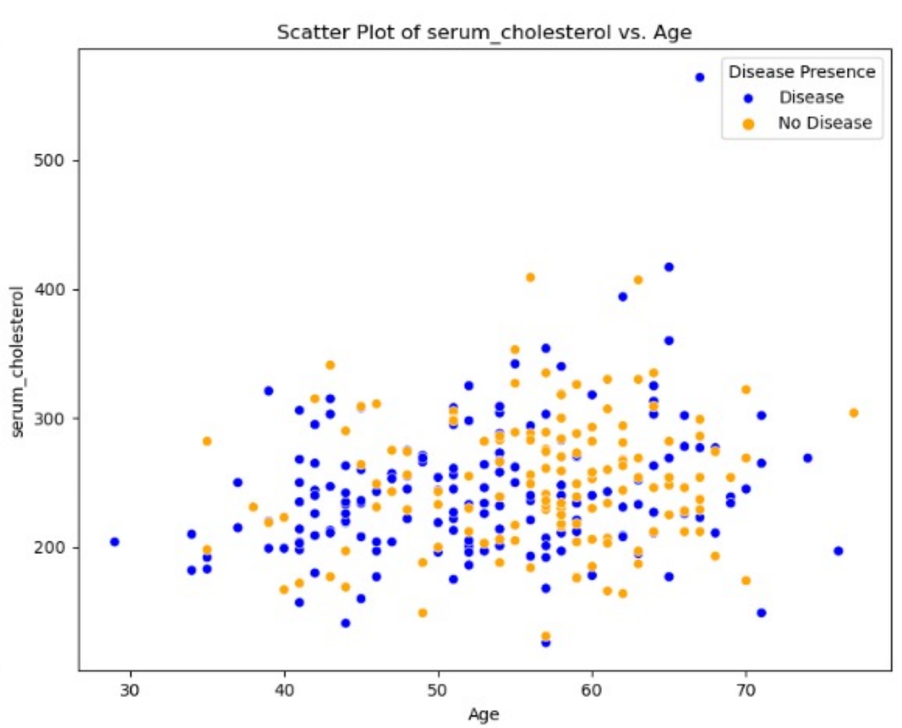


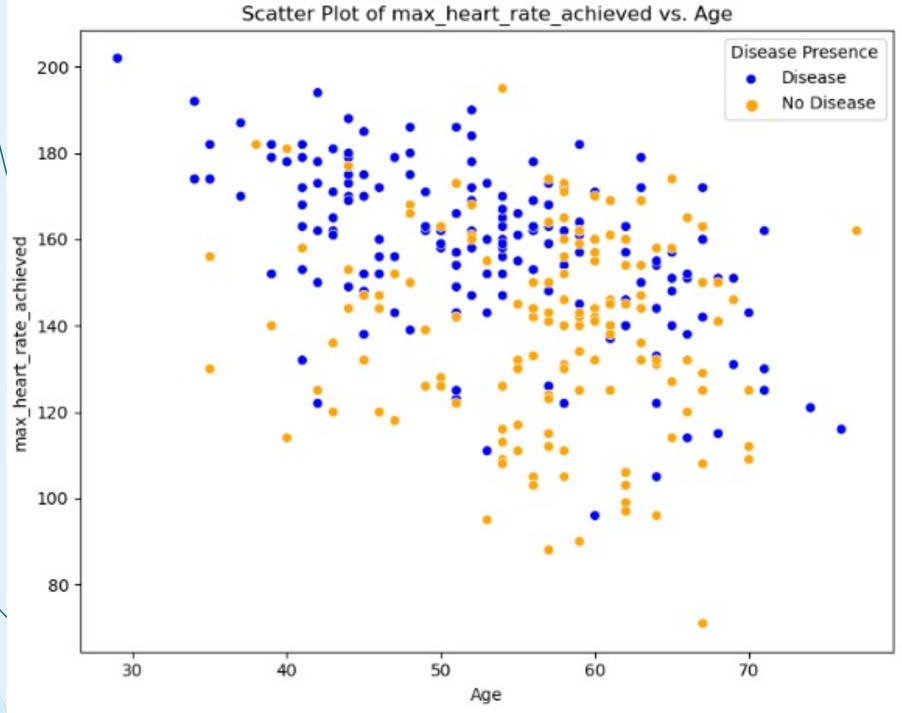
Scatterplot

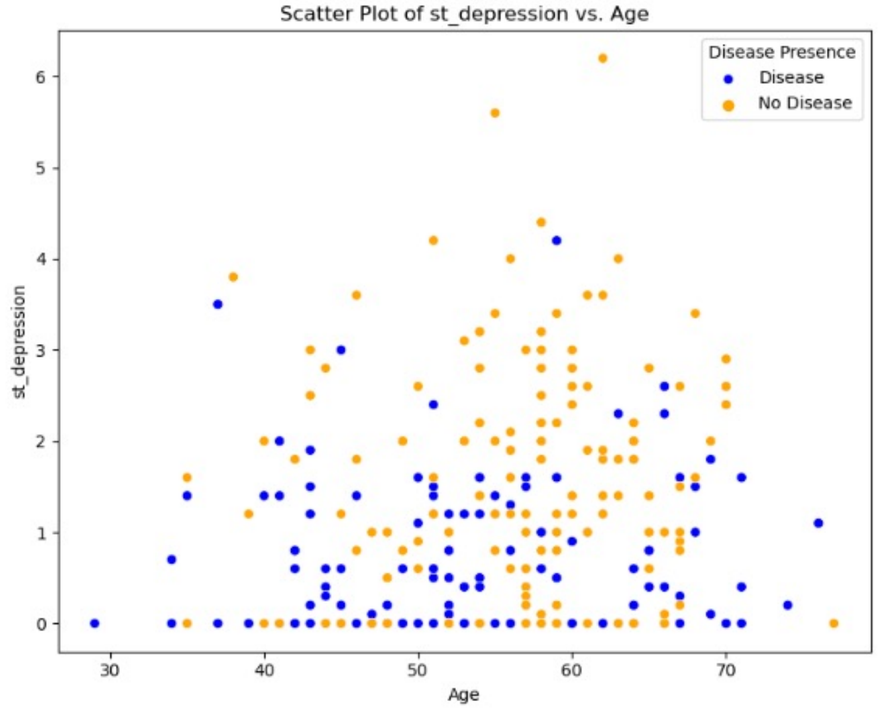
Numeric Factors vs Age

Tried to Observe Trends





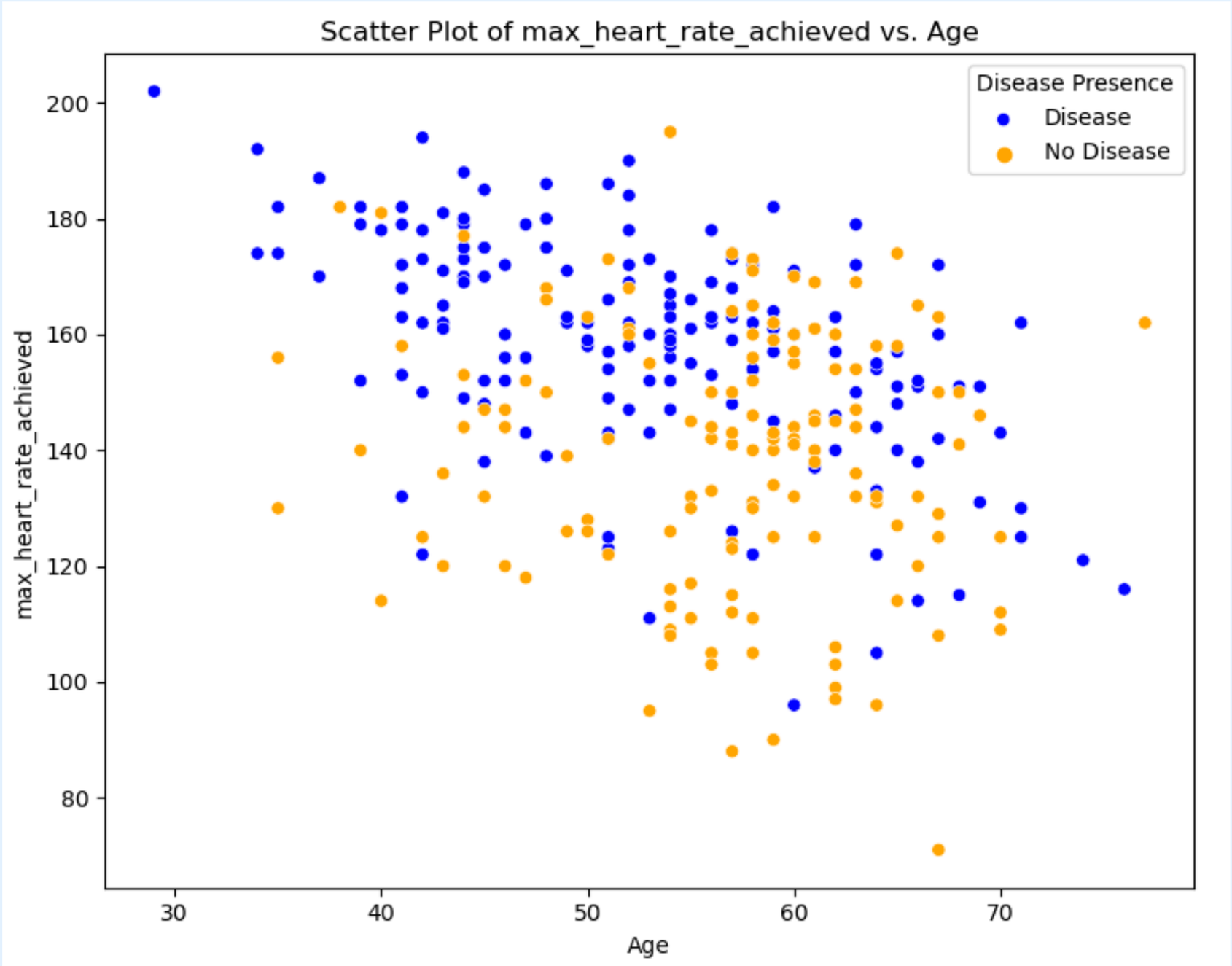




Scatterplot

Explanation of early peak

Younger individuals with elevated maximum heart rate are more prone to heart conditions



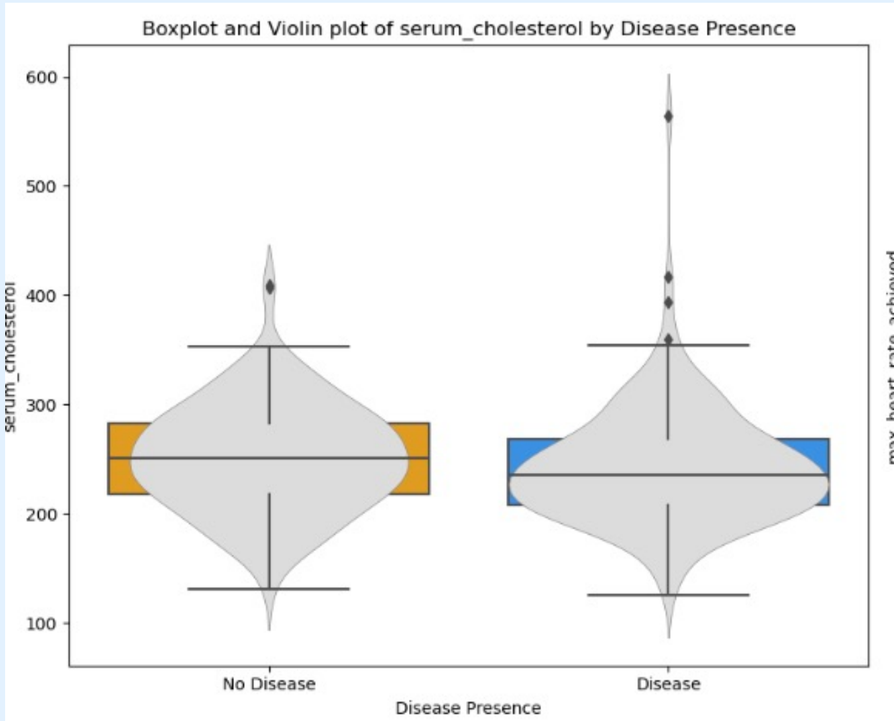
Numeric Variables

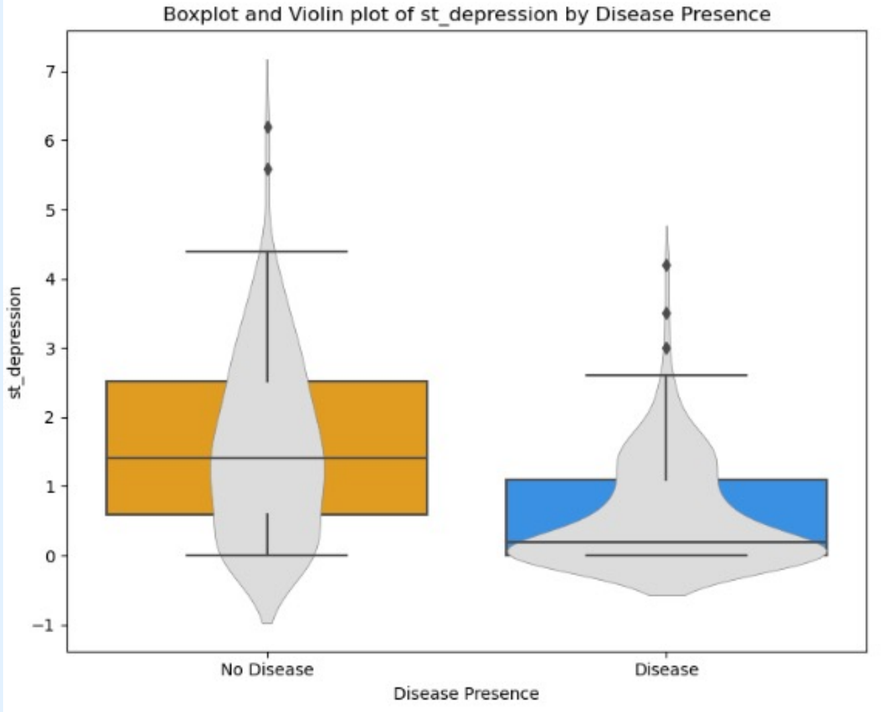
Box and Violin plot

1. display the central tendency, spread, and outliers

2. a clearer visualization of data distribution characteristic.

Example of our plots





Some Considerations

Outliers

We decided not to remove outliers based on the following reasons:

Clinical Relevance

- rare medical conditions
- could lead to exclusion of **critical case**

Diagnostic Accuracy

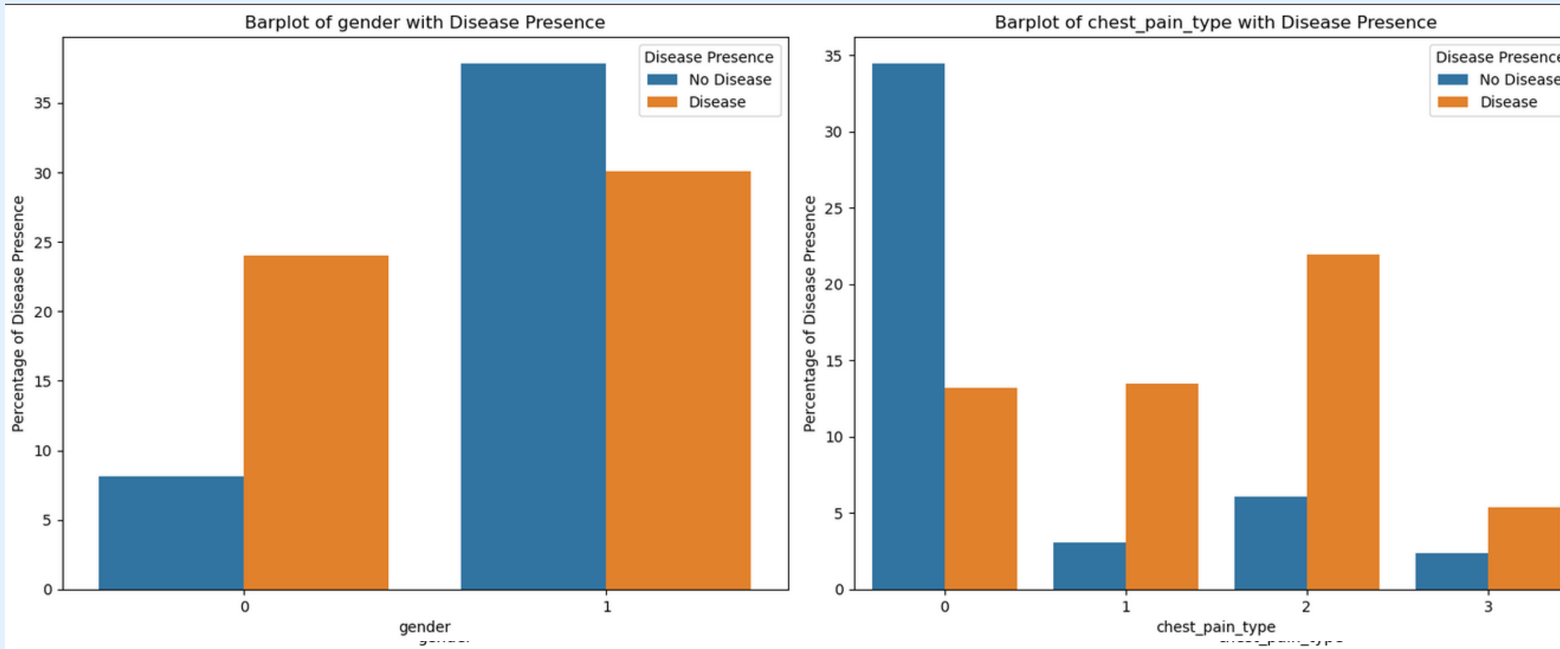
- could still be valid indicators of health status
- could **distort accuracy**
- **misclassification of patients**

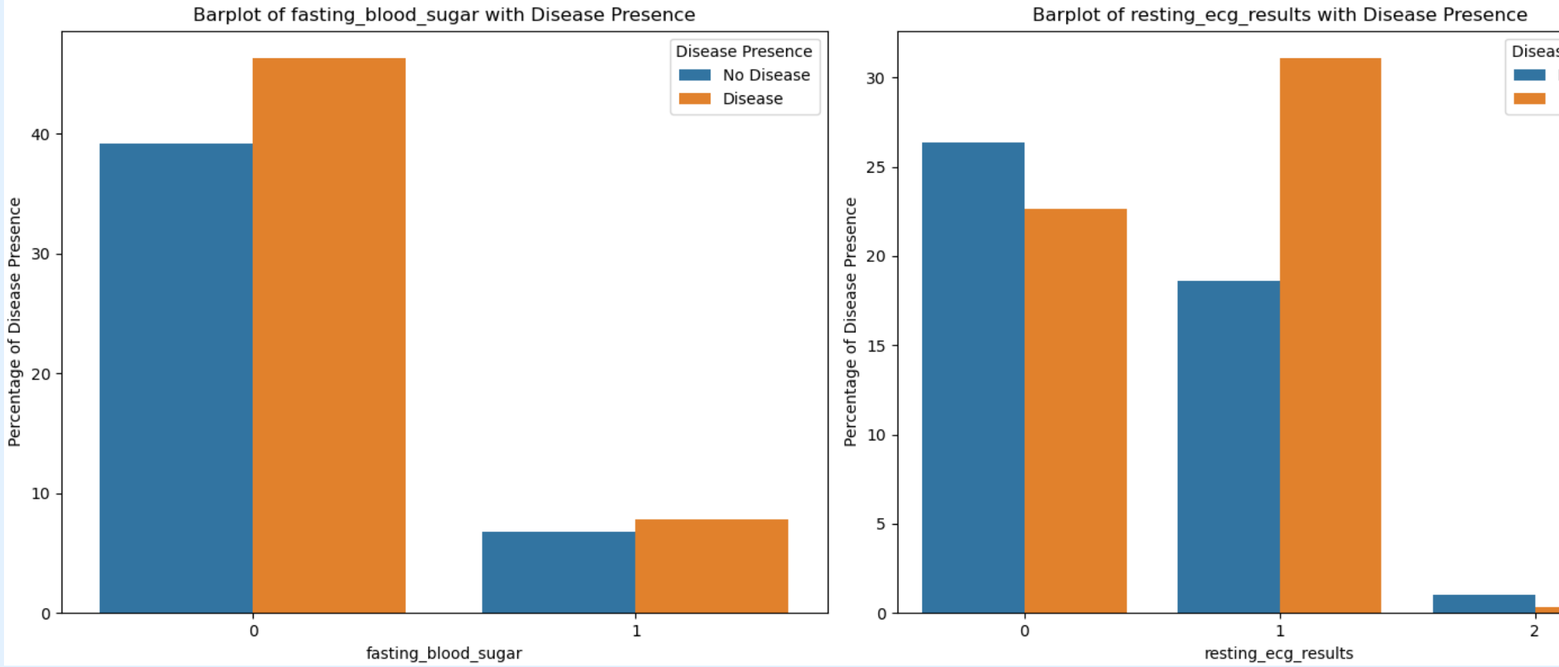
Predictive Modelling

- could influence the performance of predictive models
- **improves the model's ability to discriminate between high-risk and low-risk individuals**

Categorical Data

Barplots elegantly compares disease presence and absence, offering a clear visual distinction.

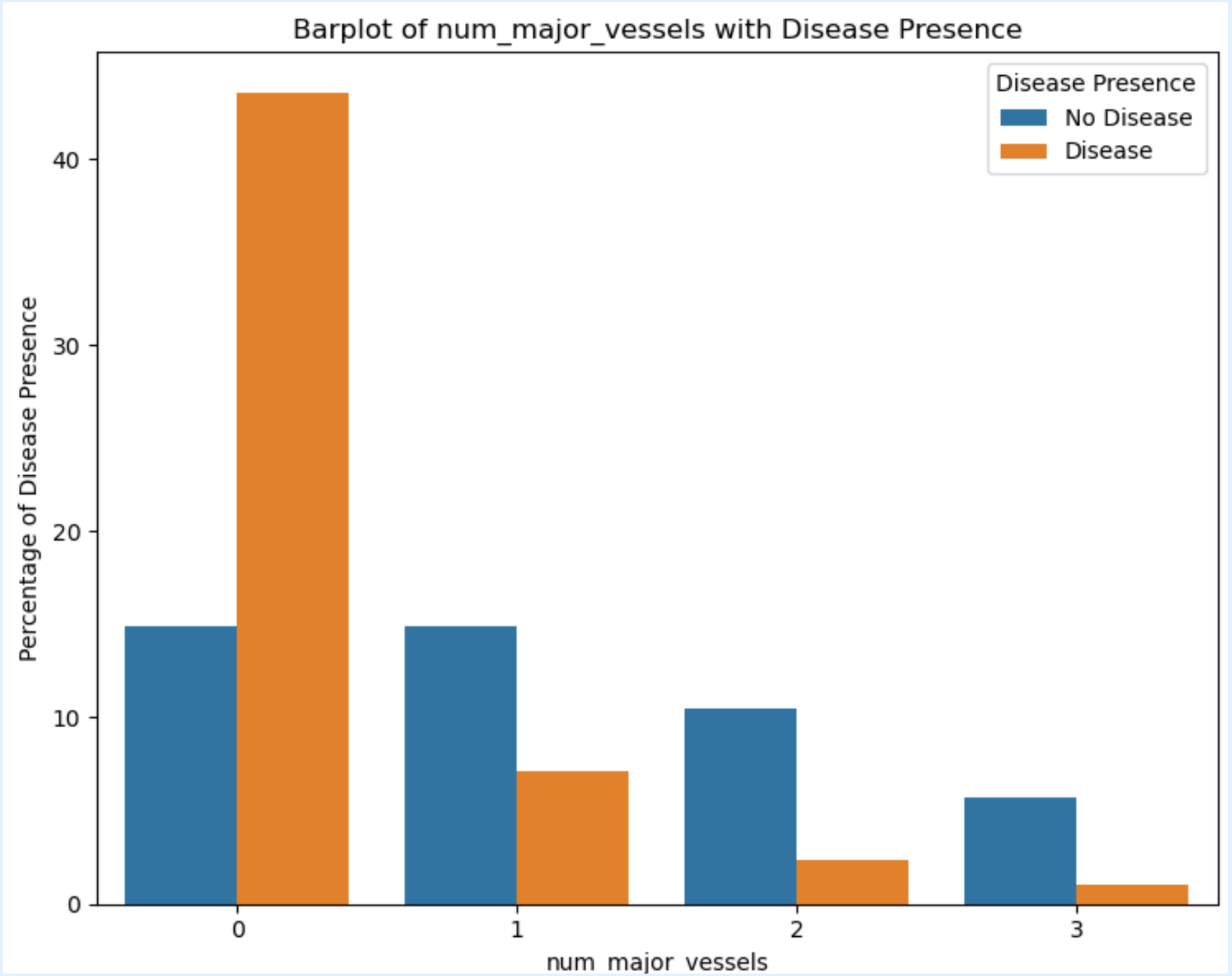




Categorical Data

0 major vessels

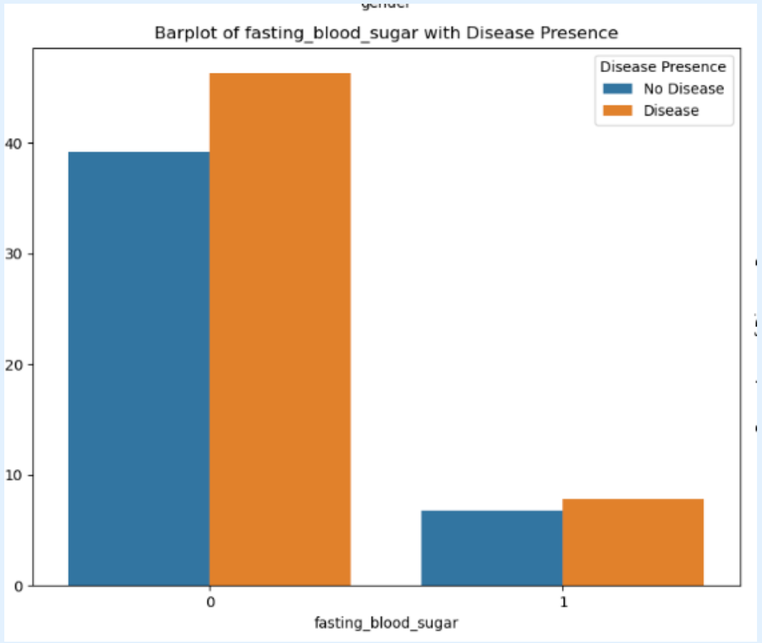
we may notice substantial differences in the counts of major vessels across different colored fluoroscopy images

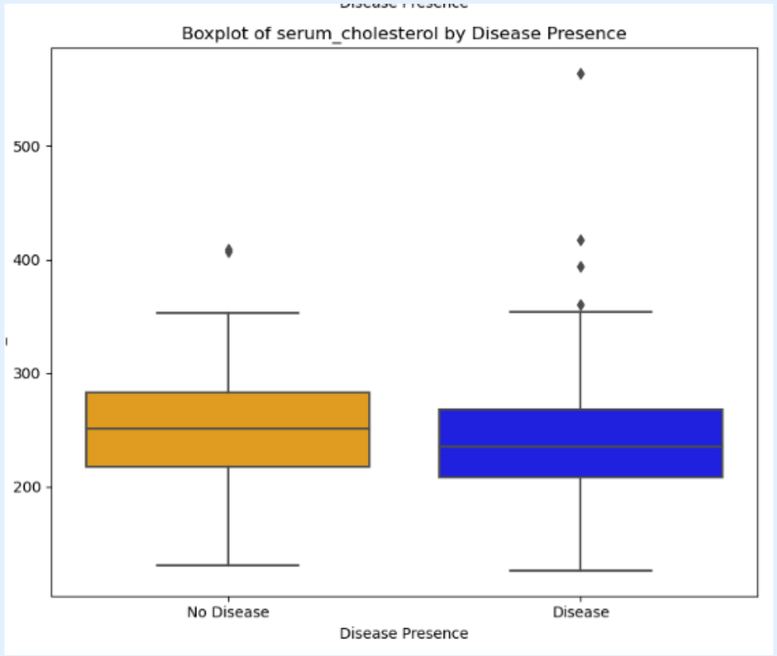


Utilising tools to drop the data

T-test

compare **two groups** in order to determine if they are statistically different from each other.

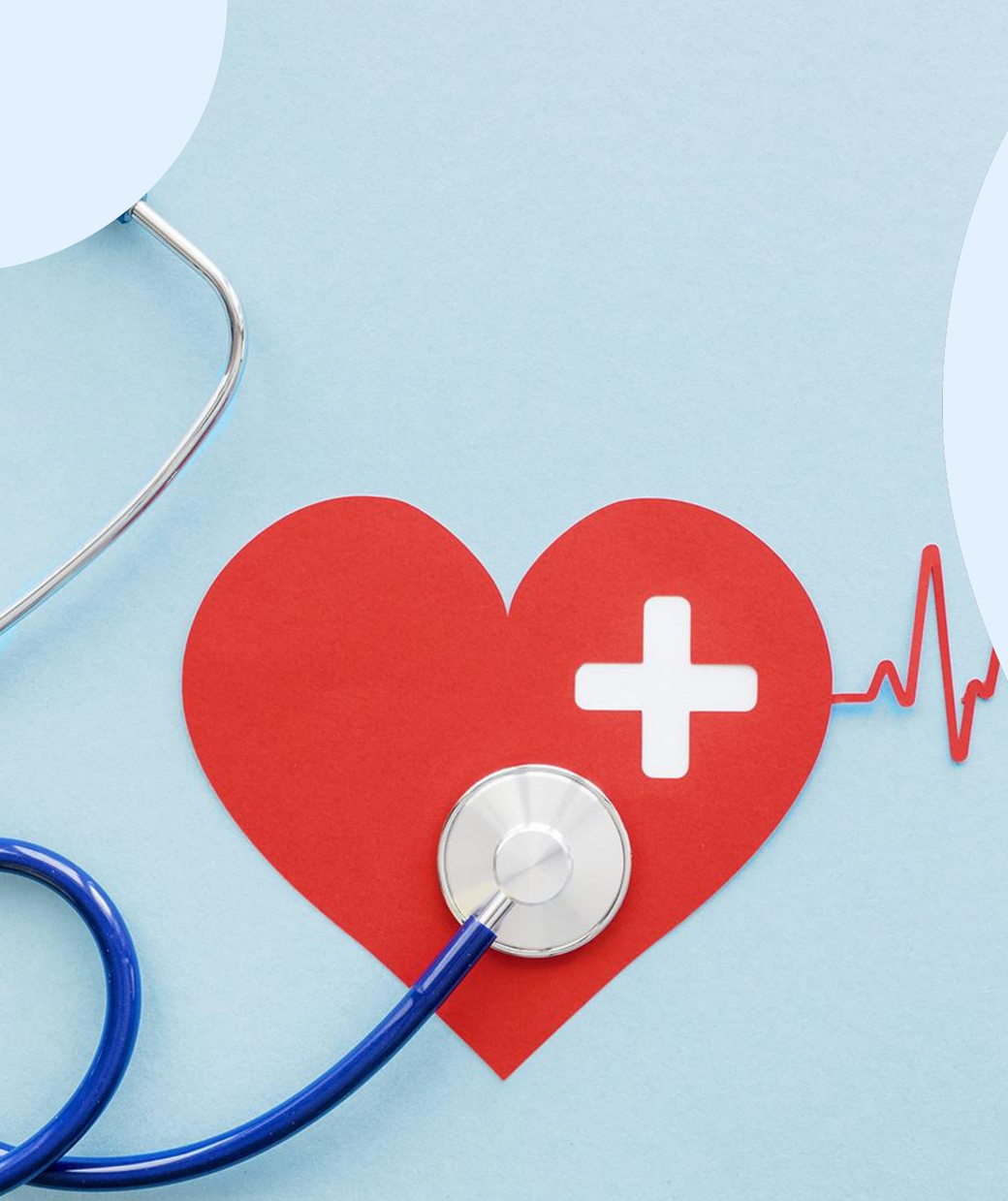




Variables with p-values>0.05
Dropping 'serum_cholestoral' and 'fasting_blood_sugar']

EDA

- important features like dual peaks
- identified critical features for predictive modelling
- filtering out less informative ones

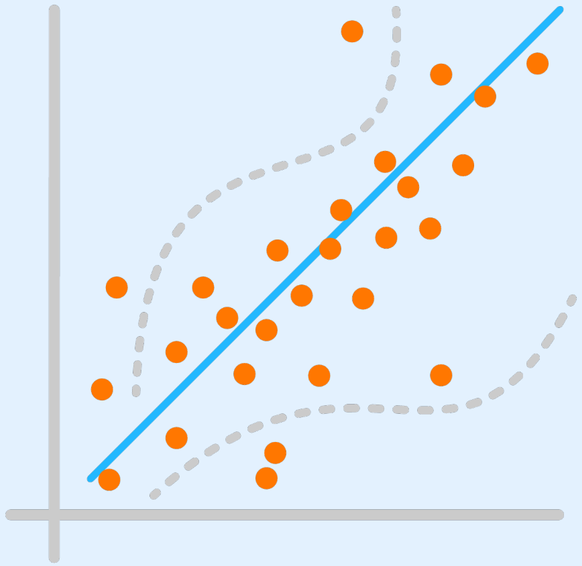


03

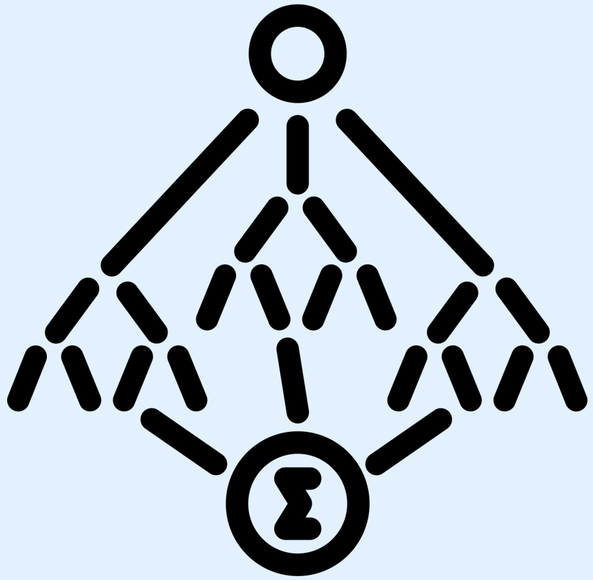
Machine Learning

- ROC
- Logistic Regression
- Random forest
- Support Vector Machine(SVM)

Tools that we are using



Logistic Regression



Random Forest

Support Vector Machine (SVM)

Logistic Regression

the goal is to predict the **probability** of a particular binary outcome.

It predicts probability that a node belongs to that binary class.

Equation for logistic Regression

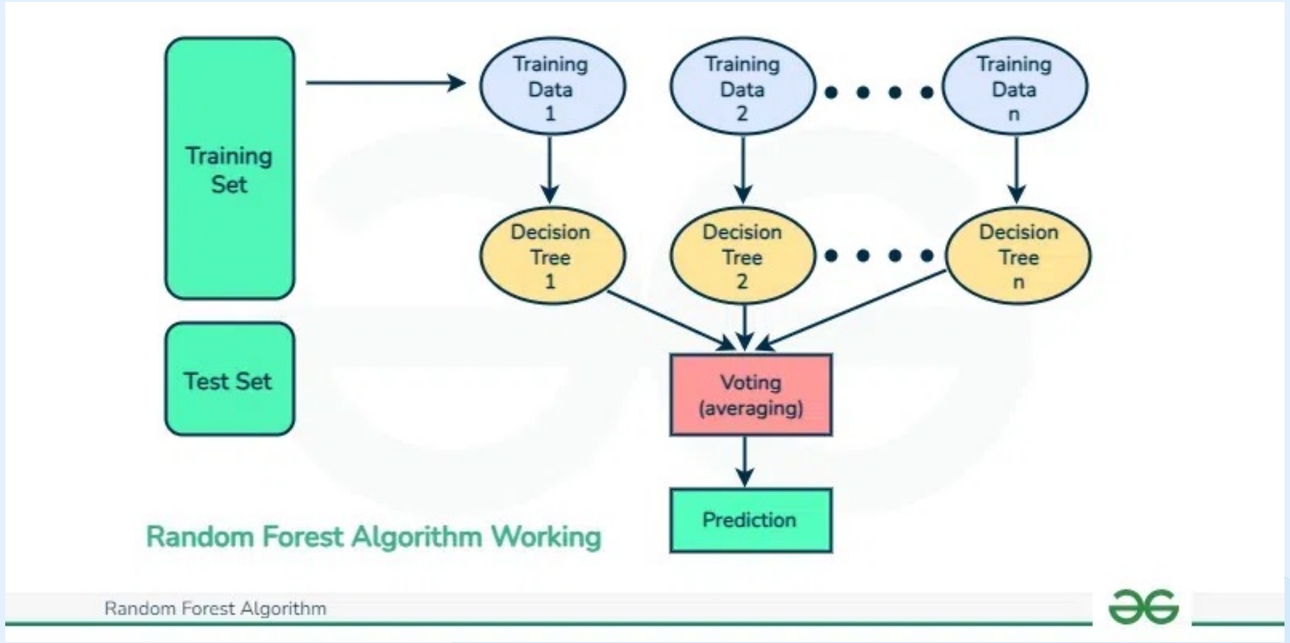
$$f(x) = \frac{1}{(1 + e^{-x})}$$

Random Forest

Multiple decision trees using random subsets of the data and features.

Merges trees later on via voting.

Tree for logistic Regression



Hyperparameter Tuning

searching over specified parameter distributions using random combinations, with cross-validation for evaluation.

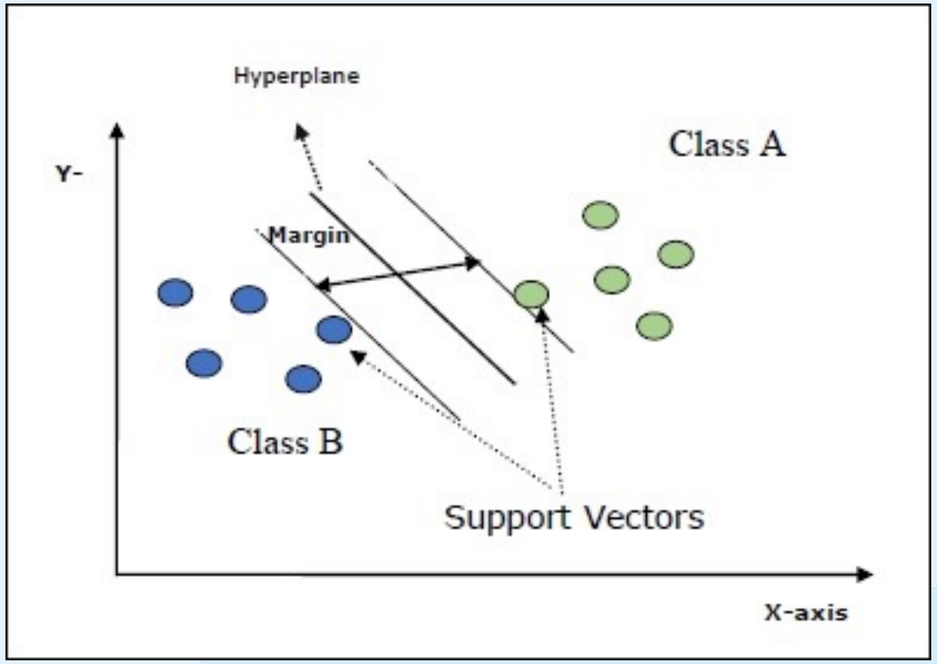
```
Best Parameters: {'min_samples_split': 5, 'min_samples_leaf': 6, 'max_depth': 6}
```

Support Vector Machine

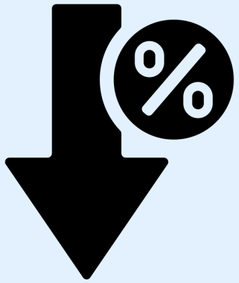
SVM transforms data into a higher-dimensional space to find a separator between different categories.

It works by maximizing the margin between the classes.

Diagram for SVM

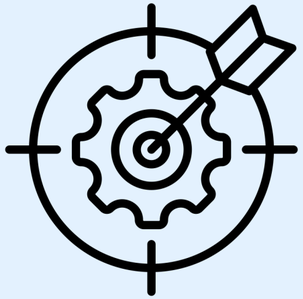


Model Comparing



Low False Negative Rate

Do not wish to has disease as absence of disease.



High Accuracy

Lower our chances of misdiagnosis. We want to be right most of the time.

Feature Importance Spread

All factors were found to be significant, so should not neglect any.

Random forest

Low False Negative Rate & High Accuracy

Random Forest

Offers the best balance of:

1. High accuracy
2. Low False Positive Rate

1. Logistic Regression

Test Accuracy: 0.8333333333333334

Test False Negative Rate (FNR): 0.06451612903225806

2. Random Forest

Test Accuracy: 0.85

Test False Negative Rate (FNR): 0.06451612903225806

3. Support vector machine SVM

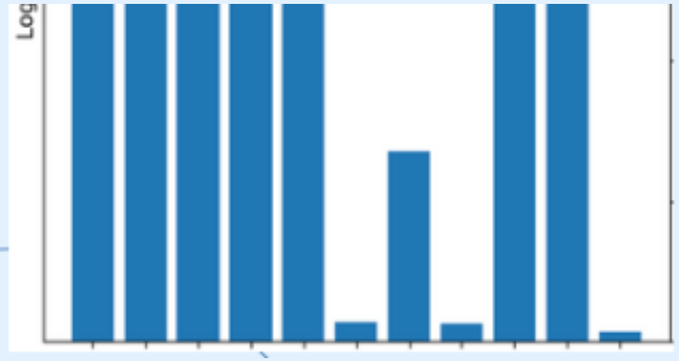
Test Accuracy: 0.8688524590163934

Test False Negative Rate (FNR): 0.13157894736842105

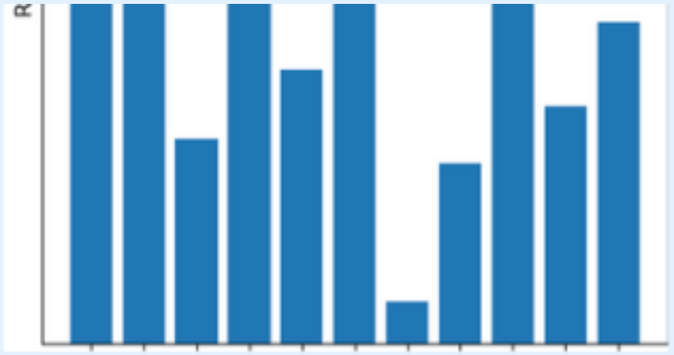
Even Spread on Feature Importance

All Factors are statistically significant, so shouldnt be completely irrelevant.

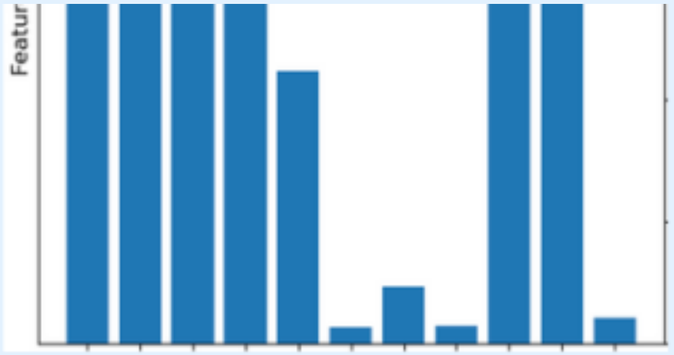
only Random Forest Captures this idea.



Logistic Regression



Random Forest



SVM

Random Forest in Action

Putting our model into practice

```
new_patient1_data = pd.DataFrame({
    'age_in_years': [41],
    'gender': [0],
    'chest_pain_type': [3],
    'resting_blood_pressure': [145],
    'resting_ecg_results': [1],
    'max_heart_rate_achieved': [150],
    'exercise_induced_angina': [0],
    'st_depression': [1.4],
    'peak_exercise_st_slope': [2],
    'num_major_vessels': [0],
    'thalassemia': [2]
})

predict_if_got_disease(new_data)
```

```
Prediction: Disease is present
Confidence Level: 85.62%
```

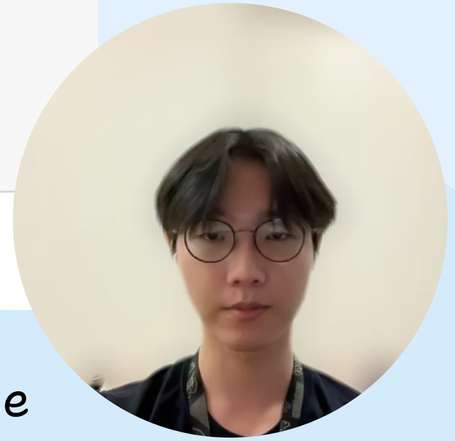
Patient 1 has disease

```
new_patient2_data = pd.DataFrame({
    'age_in_years': [50],
    'gender': [1],
    'chest_pain_type': [0],
    'resting_blood_pressure': [0],
    'resting_ecg_results': [1],
    'max_heart_rate_achieved': [154],
    'exercise_induced_angina': [2],
    'st_depression': [2],
    'peak_exercise_st_slope': [0],
    'num_major_vessels': [0],
    'thalassemia': [0]
})

predict_if_got_disease(new_data)
```

```
Prediction: Disease is not present
Confidence Level: 99.85%
```

Patient 2 does not have disease



Reflections

What we Learnt from this Project

Hard Skills

1. Statistical Analysis
2. Classification models
3. Hyperparameter tuning

Soft Skills

1. Communicating ideas
2. Delegating work
3. Open-mindedness

References:

1. https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/notebook%20
2. https:/www.myheart.org.sg/health/heart-disease-statistics/#:~:text=In%20Singapore%2C%2023%20people%20die,to%20heart%20diseases%20or%20stroke.%20
3. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/%20
4. https://www.ibm.com/topics/random-forest%20
5. https://www.investopedia.com/terms/t/t-test.asp%20
6. https://www.investopedia.com/terms/a/anova.asp%20https://www.ibm.com/spss-modeler/saas?topic=models-how-svm-works

Thank you !