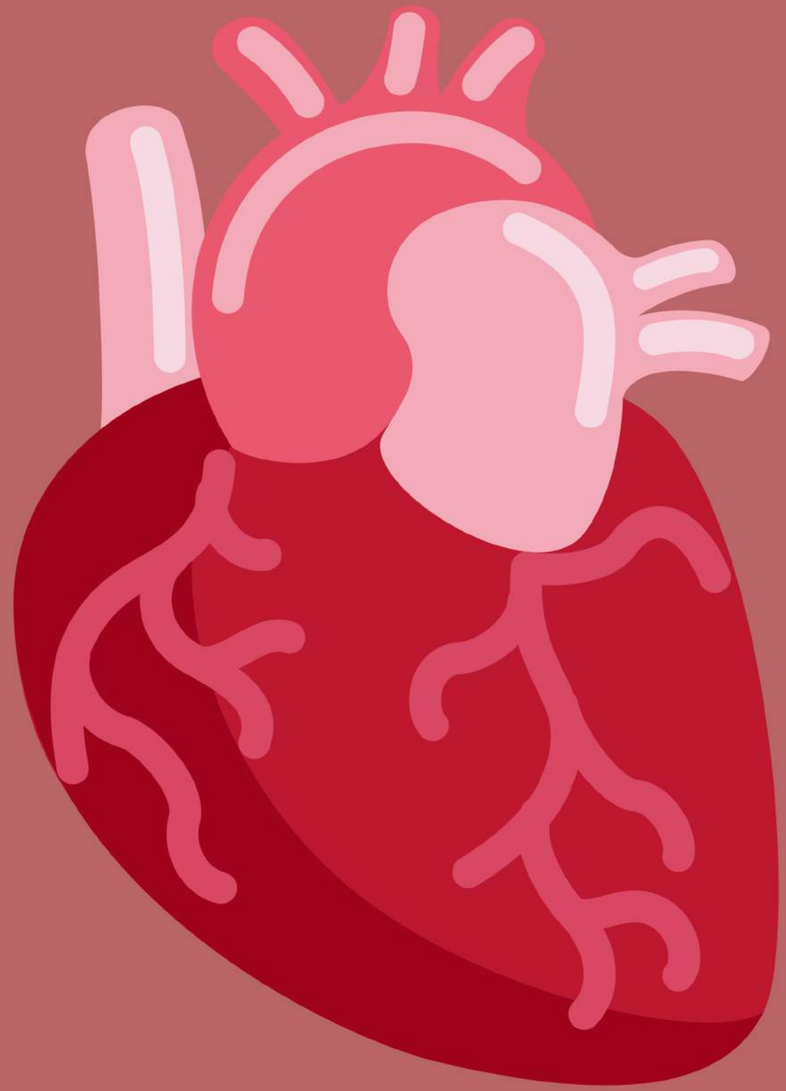
Heart Disease Classification



29 March 2023 | Machine Learning - DF

Group 1:

Matthew
21/478240/PA/20736

Rabbani Nur Kumoro
21/472599/PA/20310

William Hilmy Susatyo
21/472585/PA/20380

Introduction

- The heart is a vital organ for every human body (Dangare et al., 2012).

- Heart diseases are the leading cause of death globally, with 17.9 million deaths occurring each year, one-third of which occur below the age of 70 (World Health Organization, n.d.).

- Carmat, a French company, recently received approval in December 2020 to sell its artificial heart in Europe, offering hope for preventing, helping, and potentially curing heart diseases (Bloomberg, 2020).

1
Dataset

Data Acquisition

- The dataset used in diagnosing heart disease is the Heart Disease Dataset, which combines four different datasets, but sometimes only the UCI Cleveland dataset is used.

- The UCI Cleveland dataset is publicly available and contains 297 instances with 14 features/attributes.

- The dataset has one dependent variable, "Diagnosis," and the remaining 13 attributes are independent variables.

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

Figure 1. First 5 Rows of the Dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

Figure 2. Dataset Description

Columns - Categorical

No	Features	Type Value	Description	Value
1	sex	Discrete	Male or female representative	0 = female, 1 = male
2	cp	Discrete	Chest pain type	0 = typical angina 1 = atypical angina 2 = non-anginal pain 3 = asymptomatic
3	fbs	Discrete	The patient's fasting blood sugar	0 = false (< 120 mg/dl) 1 = true (> 120 mg/dl)
4	restecg	Discrete	ECG's outcome, where each integer represents the level of pain.	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
5	exang	Discrete	Exercise induced angina	1 = yes; 0 = no
6	slope	Discrete	The patient's state at the height of exercise	0 = downsloping 1 = flat 2 = upsloping
7	ca	Discrete	The quantity of main vessels that fluoroscopy can colour	0 - 3
8	thal	Discrete	A blood disorder	0: normal 1: fixed defect (no blood flow in some parts of the heart) 2: reversible defect (a blood flow is observed but it is not normal)

Table 1. Categorical Columns Description

Columns - Numeric

No	Feature	Type Value	Description	Value
1	age	Continuous	Patient range shown by age	Age 28 - 77
2	trestbps	Continuous	Represent the resting heart rate (in mm Hg on admission to the hospital)	Multiple continue value in mmHg
3	chol	Continuous	The patient's cholesterol measurement in mg/dl	Multiple continue value in mm/dl
4	thalach	Continuous	The patient's maximum heart rate	Low: < 50 beats/min Normal: 51 - 119 beats/min High: 120 - 180 beats/min
5	oldpeak	Continuous	ST depression induced by exercise relative to rest	Multiple decimal number values between 0 and 6.2

Table 2. Numerical Columns Description

2
Pre-Processing

Data Types and Null Values

```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
```

```
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
target      0
dtype: int64
```

Figure 3. Dataset Info

Skewness

```
age  =  -0.2024633654856539
sex  =  -0.791335191480832
cp  =  0.48473236883889675
trestbps  =  0.7137684379181465
chol  =  1.1434008206693387
fbs  =  1.986651930914452
restecg  =  0.16252224492761935
thalach  =  -0.5374096526832253
exang  =  0.7425315444212832
oldpeak  =  1.269719930601997
slope  =  -0.5083156098165442
ca  =  1.3104221354767875
thal  =  -0.47672219490975737
target  =  -0.17982105403495655
```

Figure 4. Skew Data

3
Exploratory Data Analysis (EDA)

Univariate Analysis

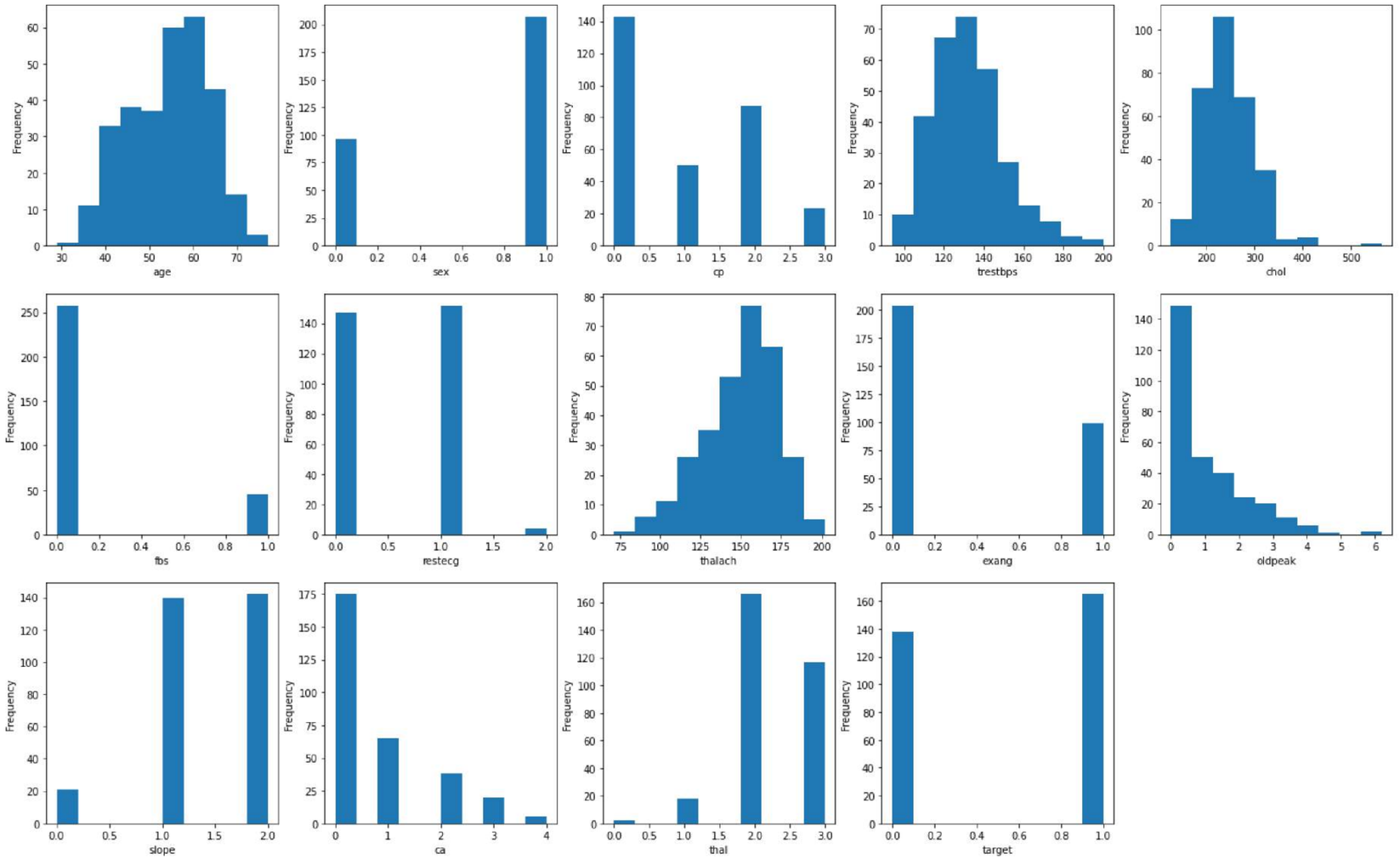


Figure 5. Histograms

Categorical Features vs Target

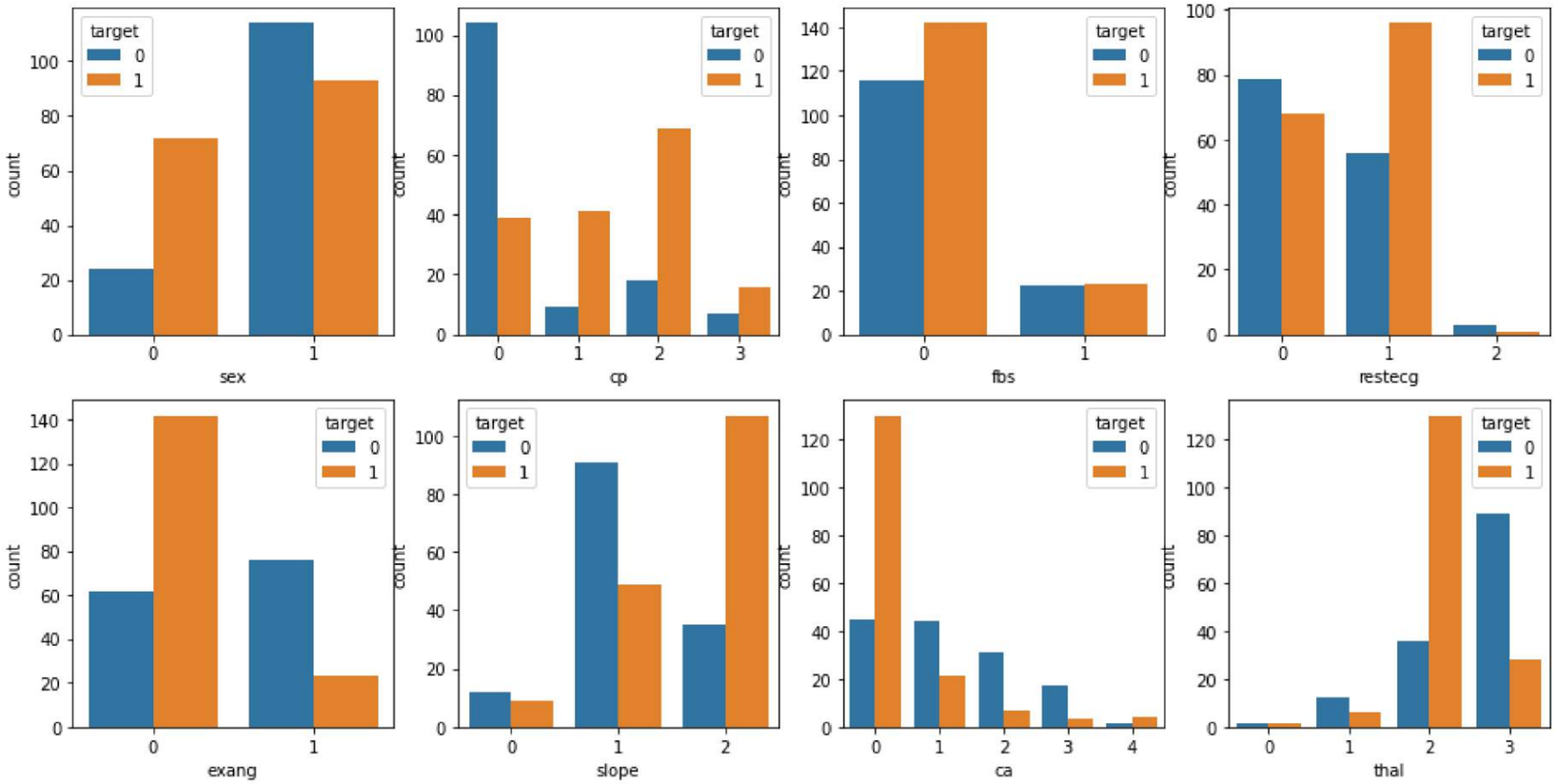


Figure 6. Grouped Bar Chart

Numerical Features vs Target

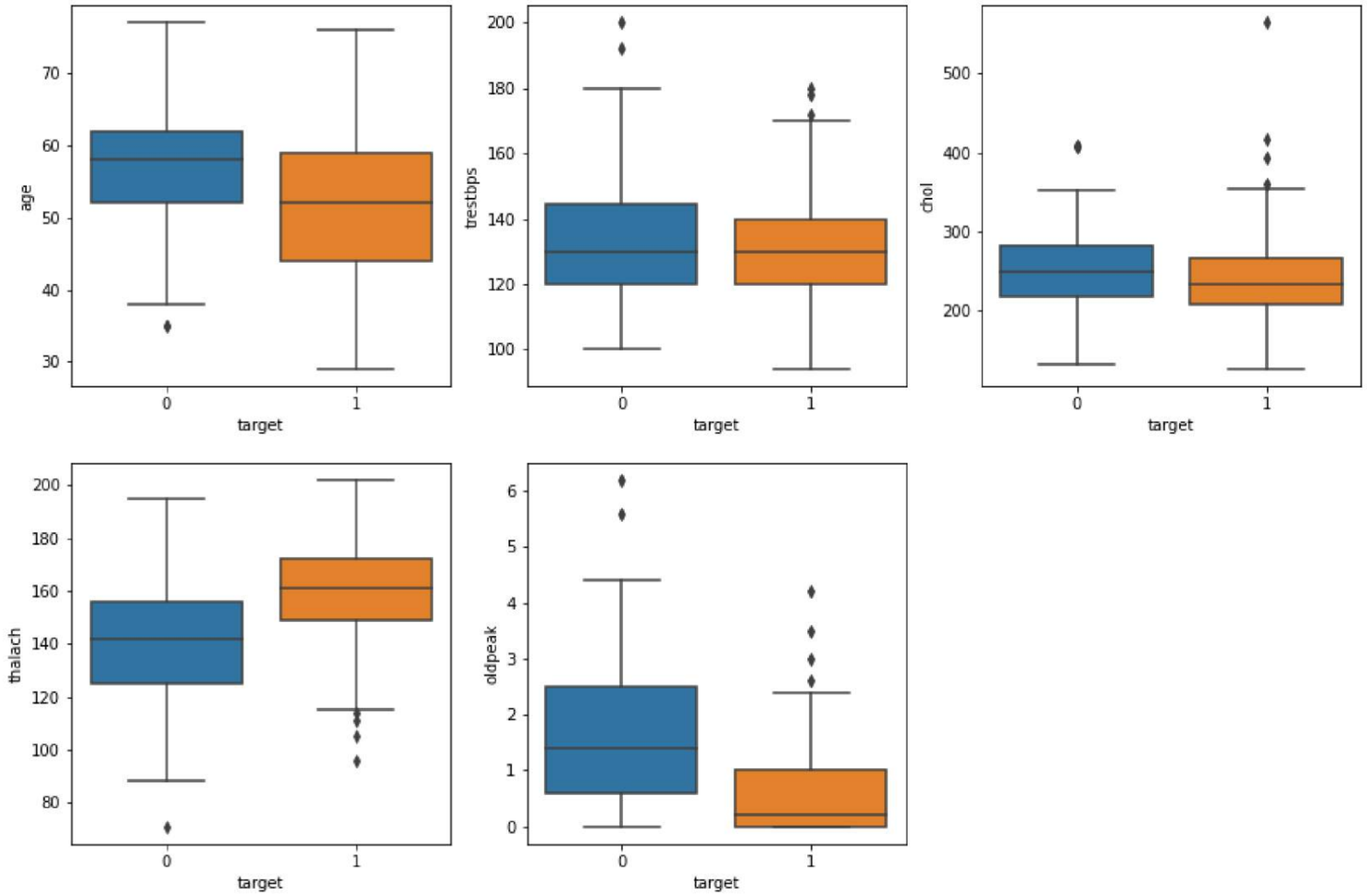


Figure 7. Box Plot

4
Feature Engineering

Feature Importances

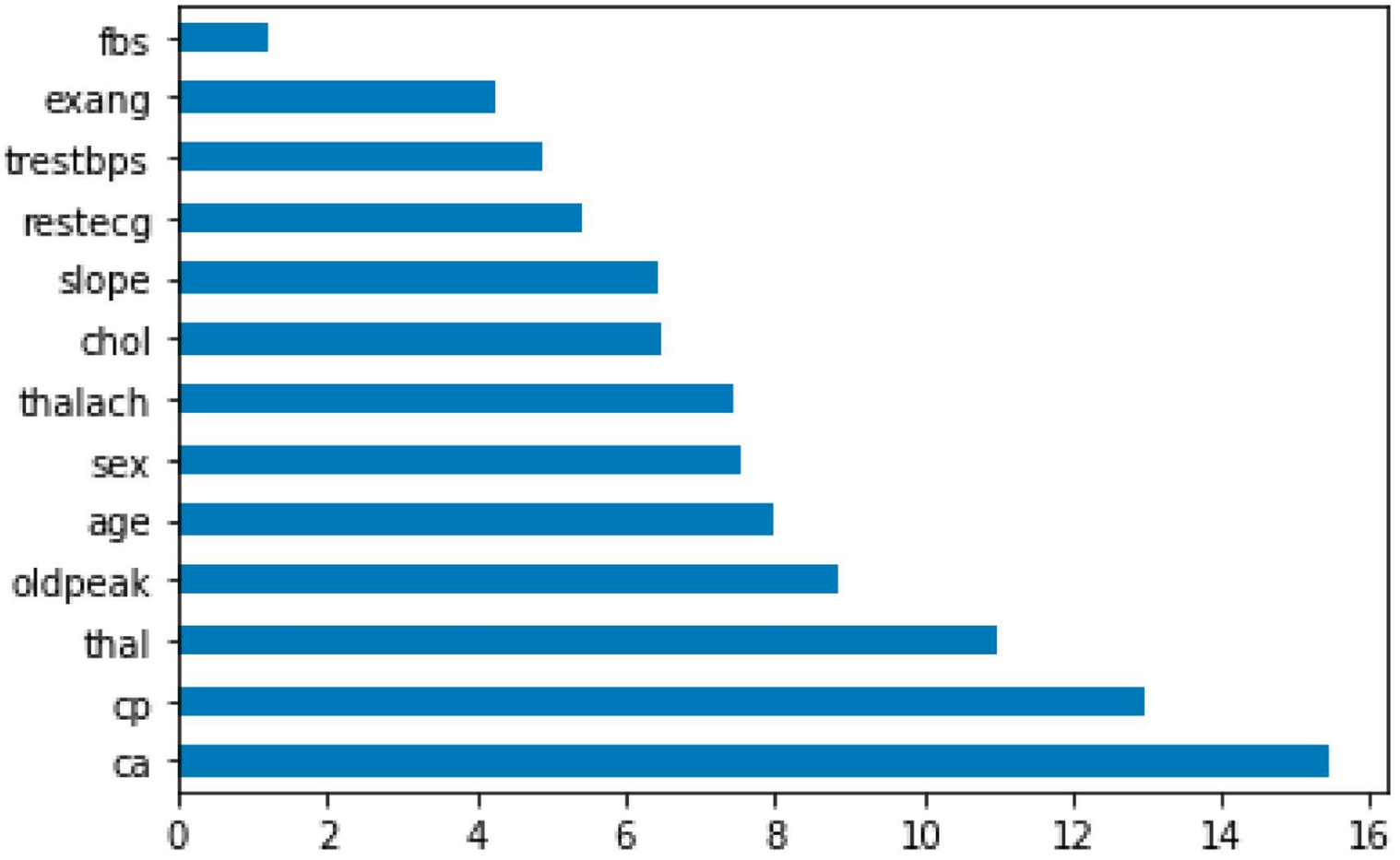


Figure 8. Graph Plot

Correlation Matrix

Figure 9. Heatmap

5
Modelling

K-Nearest Neighbors

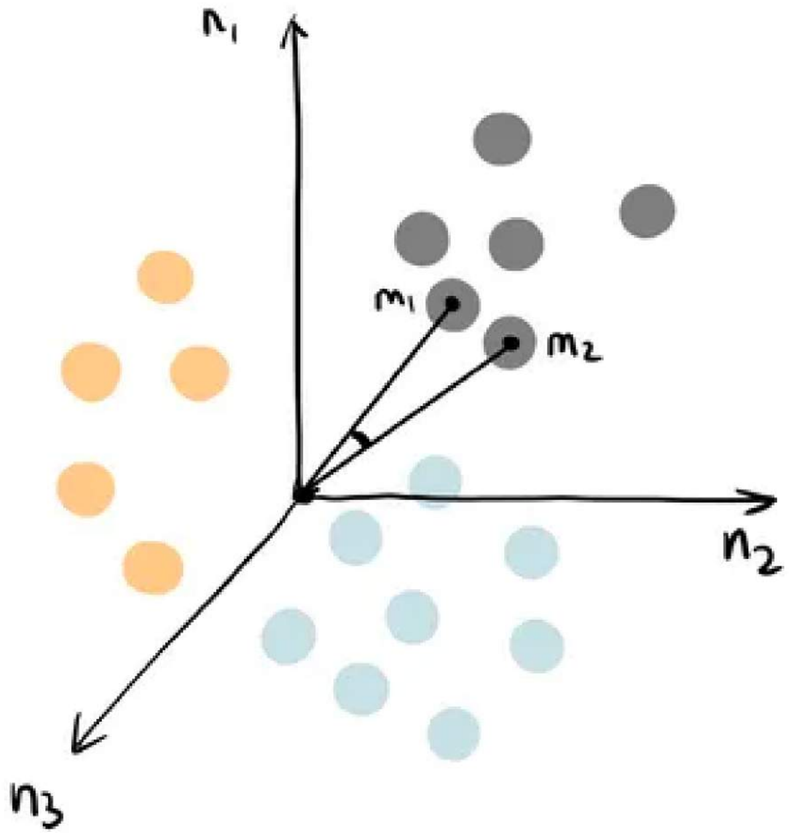


Figure 10. K-Nearest Neighbors Theorem

K-Nearest Neighbors

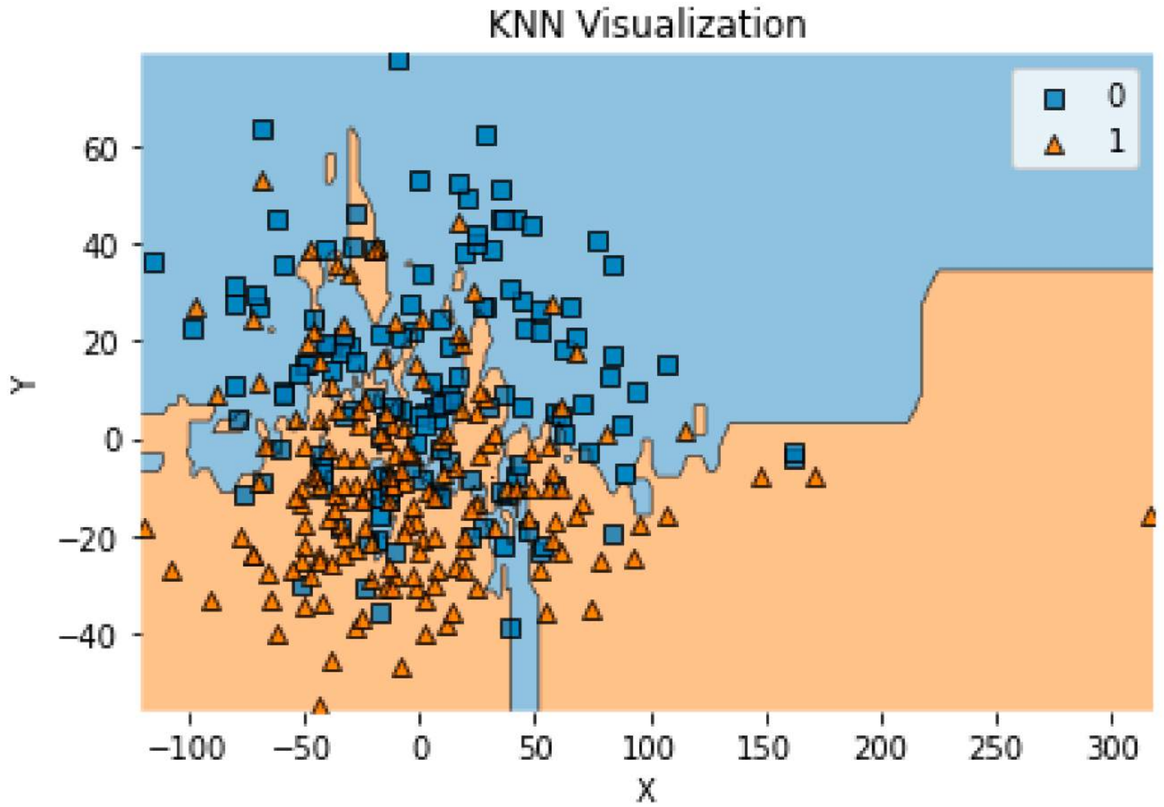


Figure 11. Explainability of K-Nearest Neighbors Model

Naïve Bayes

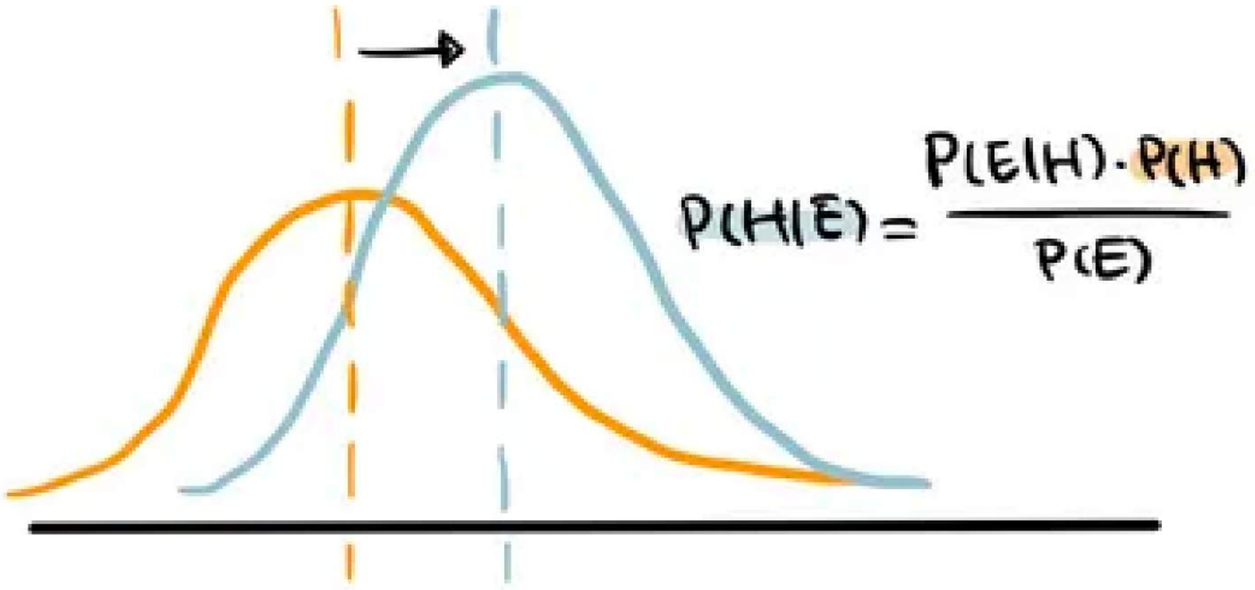


Figure 12. Naïve Bayes Theorem

Naïve Bayes

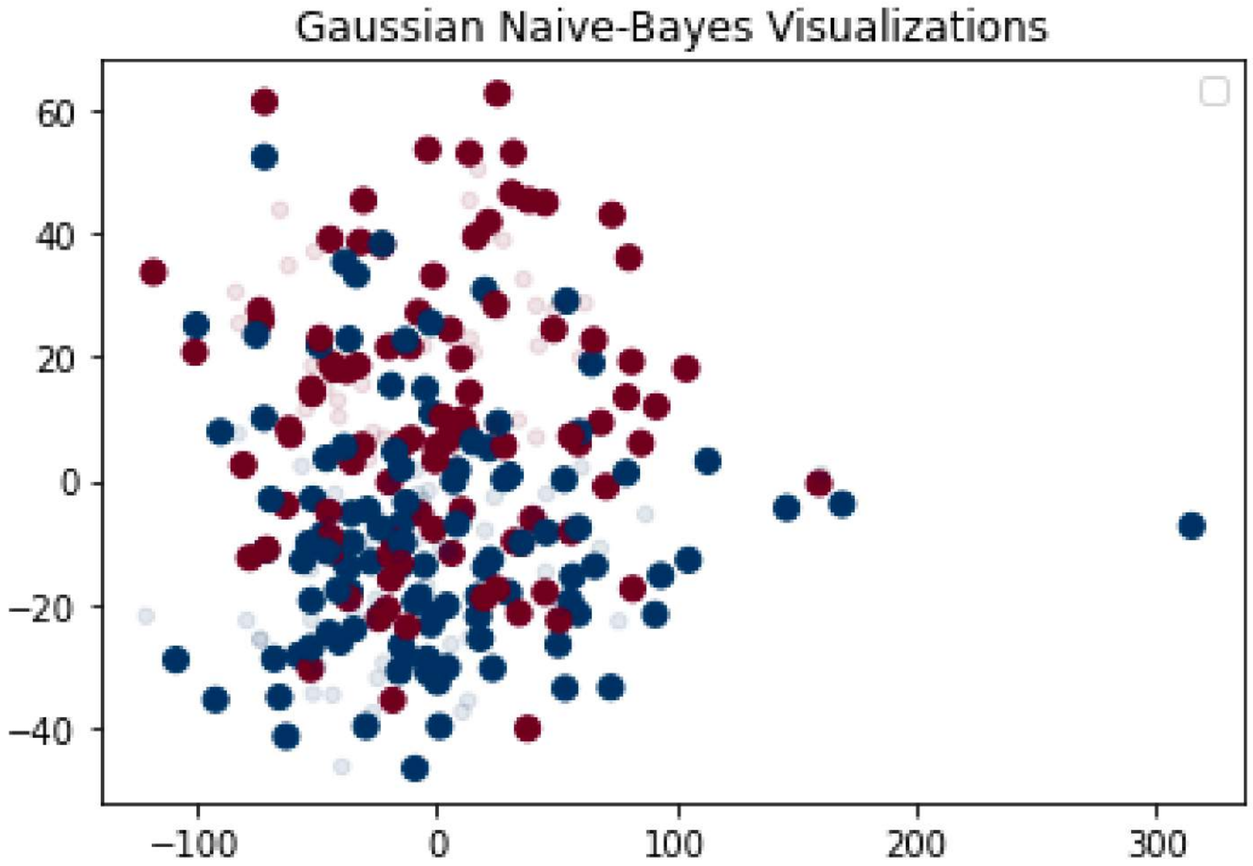


Figure 13. Visualization of Gaussian Naïve Bayes Model

6
Performance Measure

Train Test Split

- For our model, the data is split into two subsets: **a training set that consists of 67% of the data, and a testing set that consists of 33% of the data.**

- The goal of splitting a dataset is to obtain an unbiased estimate of the model's performance on new, unseen data.

- This is achieved by training the model on the training set and evaluating its performance on the testing set, which contains data that the model has not seen before.

Hyperparameter Tuning
Grid Search

K-Nearest Neighbors

Parameters	Value
leaf_size	1
p	1
n_neigbors	7

Table 3. Best Parameters for K-Nearest Neighbors

Naïve Bayes

Parameter	Value
var_smoothing	1e-10

Table 4. Best Parameters for Naïve Bayes

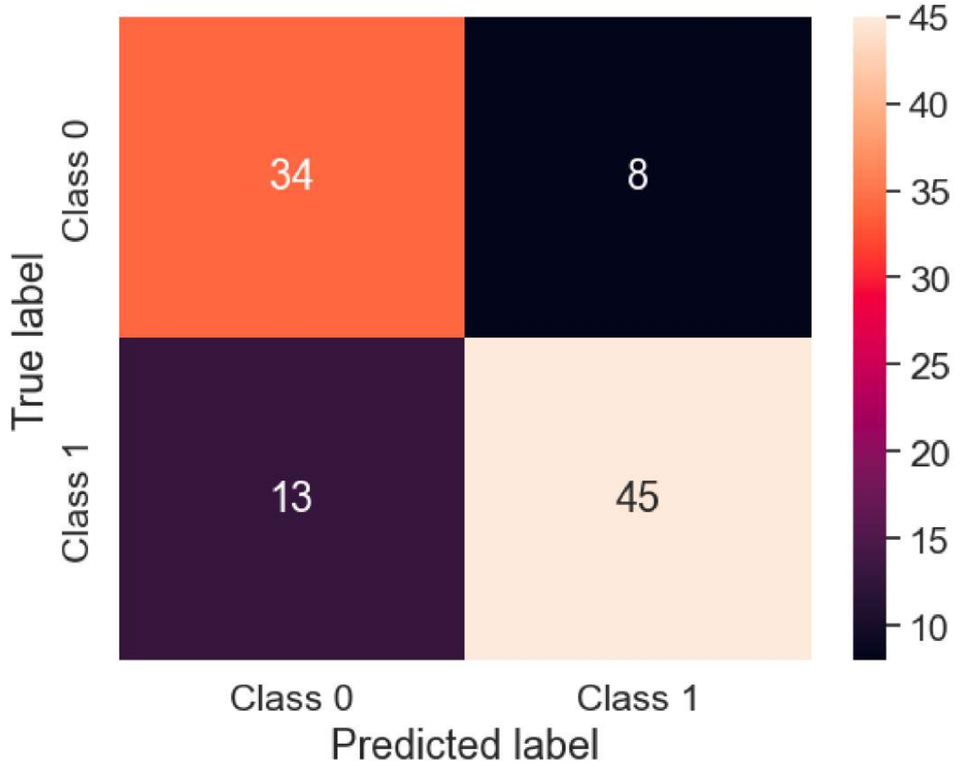
Performance Comparison

Models	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	80%	0.8	0.79	0.79
Naïve Bayes	88%	0.88	0.88	0.88

Table 5. Performance Comparison between the Machine Learning Models

Confusion Matrix

K-Nearest Neighbors



Naïve Bayes

Figure 14. Confusion Matrix Comparison of each Algorithms

7
Conclusion

Key Takeaways

- Naïve Bayes performed the best out of the two methods, while K-Nearest Neighbors is not far behind.
- Each algorithm has its preferences and requires different data processing and feature engineering techniques.
- Understanding the characteristics of each algorithm allows us to balance the trade-off and select the appropriate model according to the dataset.

References

"Cardiovascular Diseases", World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases.

Bloomberg. (2020, December 23). CARMAT Receives the CE Marking for Its Total Artificial Heart.

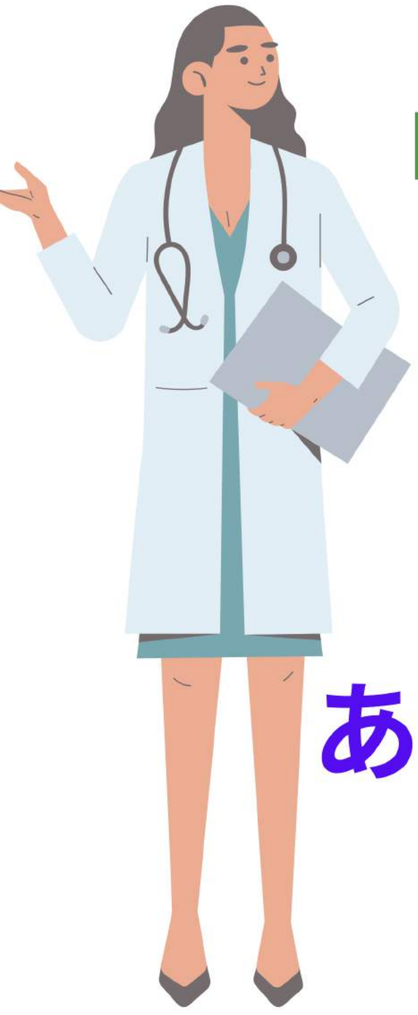
Dangare, C. S., Apte, S. S., & Student, M. E. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 975-888.

Terima Kasih

감사합니다

Hartstikke Bedankt!

شُكْرًا



Merci

Thank You

Gracias

Matur Nuwun

Благодарствую

ありがとう

धन्यवाद

谢谢