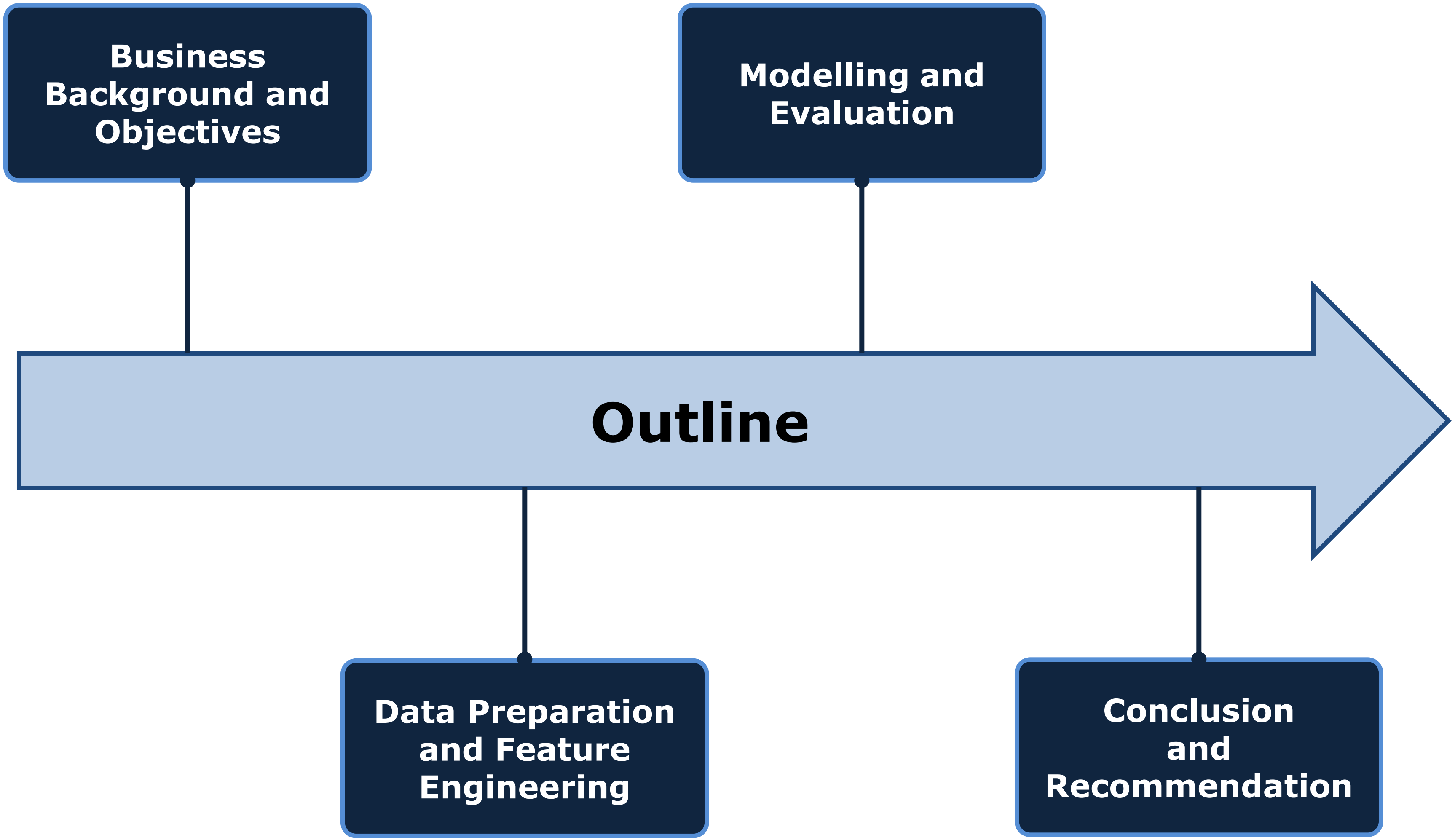
Heart Failure Analysis and Prediction with Logistic Regression

By: Aldimeola Alfarisy



Outline

Business Background and Objectives

Data Preparation and Feature Engineering

Modelling and Evaluation

Conclusion and Recommendation

Business Background and Objectives

Business Background and Objectives

Background

Cardiovascular diseases (CVDs) are the leading cause of death globally with taking an estimated 17.9 million lives each year. More than four out of five CVD deaths are due to heart attacks and one third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs. early detection and management wherein a machine learning model can be of great help.

Image source: https://news.harvard.edu/gazette/story/2022/04/infertility-history-linked-with-increased-risk-of-heart-failure/

Objectives

- What factors affect heart failure?
- How accurate a machine learning model can predict to heart disease early detection?

Data Preparation and Feature Engineering

Data Preparation and Feature Engineering

Dataset Information

918 rows

11 feature

Numerical Feature

- Age
- RestingBP
- Cholesterol
- FastingBS
- MaxHR
- Oldpeak

Categorical Feature

- Sex
- ChestPainType
- RestingECG
- ExerciseAngina
- ST_Slope

1 target

Heart Disease

Source dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

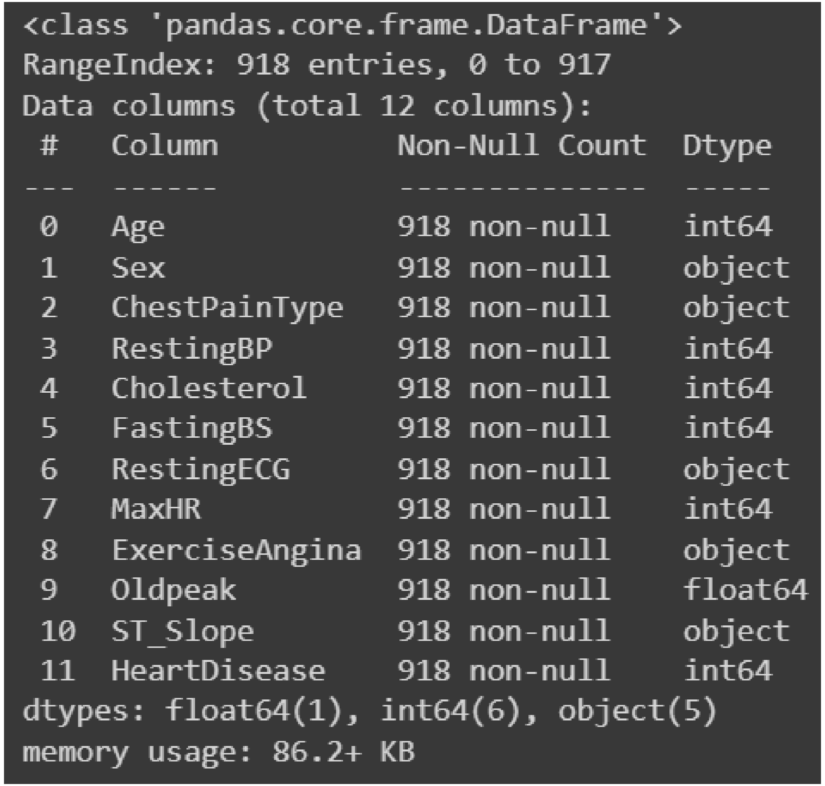
Data Preparation and Feature Engineering

Dataset Attribute Information

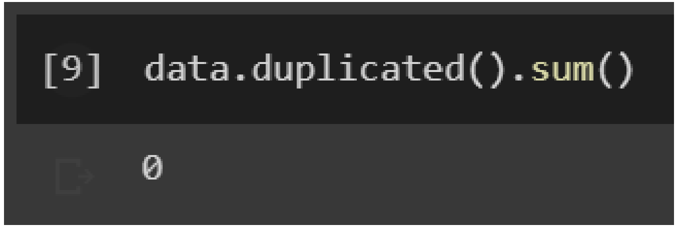
Column name	Description
Age	Age of the patient [years]
Sex	Sex of the patient [M: Male, F: Female]
ChestPainType	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	Resting blood pressure [mm Hg]
Cholesterol	Serum cholesterol [mm/dl]
FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina [Y: Yes, N: No]
Oldpeak	ST [Numeric value measured in depression]
ST_Slope	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	Output class [1: heart disease, 0: Normal]

Data Preparation and Feature Engineering

General Info



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             918 non-null    int64  
 1   Sex             918 non-null    object 
 2   ChestPainType   918 non-null    object 
 3   RestingBP       918 non-null    int64  
 4   Cholesterol     918 non-null    int64  
 5   FastingBS       918 non-null    int64  
 6   RestingECG      918 non-null    object 
 7   MaxHR           918 non-null    int64  
 8   ExerciseAngina  918 non-null    object 
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object 
 11  HeartDisease    918 non-null    int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```



```
[9]  data.duplicated().sum()

     0
```

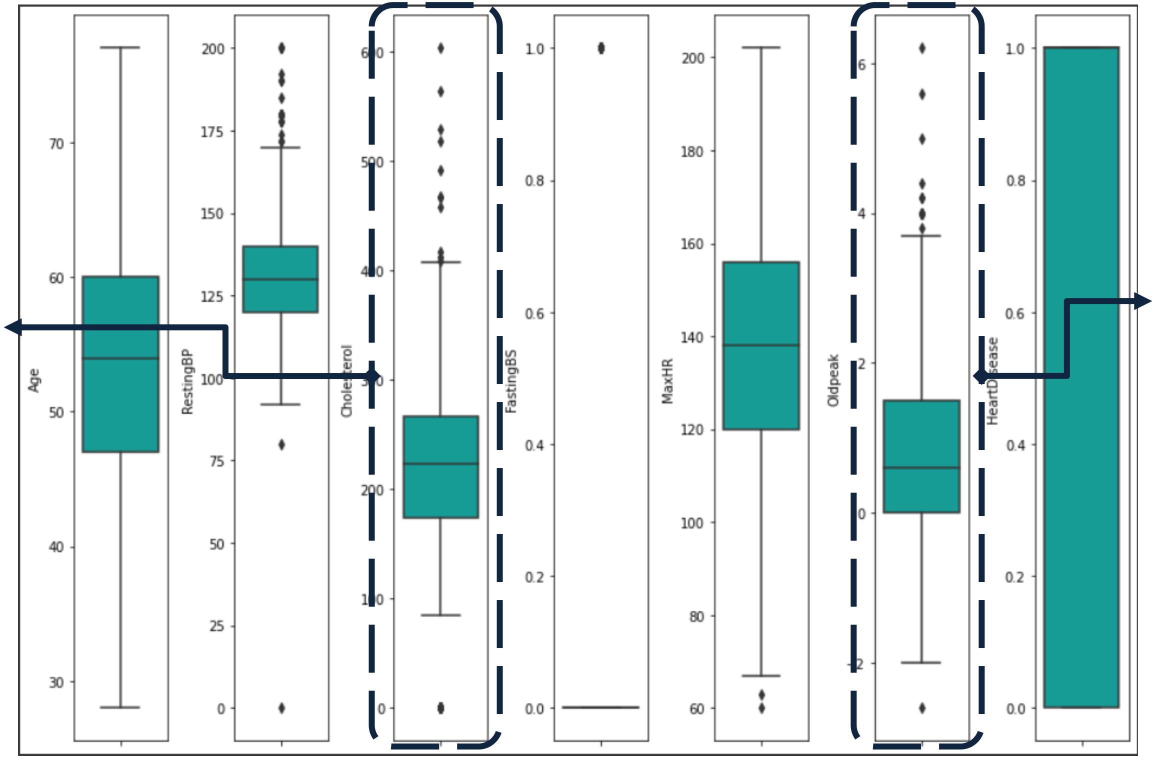
Result

- **No Missing Value**
- **No Duplicated Value**

Data Preparation and Feature Engineering

Boxplot for Numerical Data

Outliers
Cholesterol
19,93%
Trim data



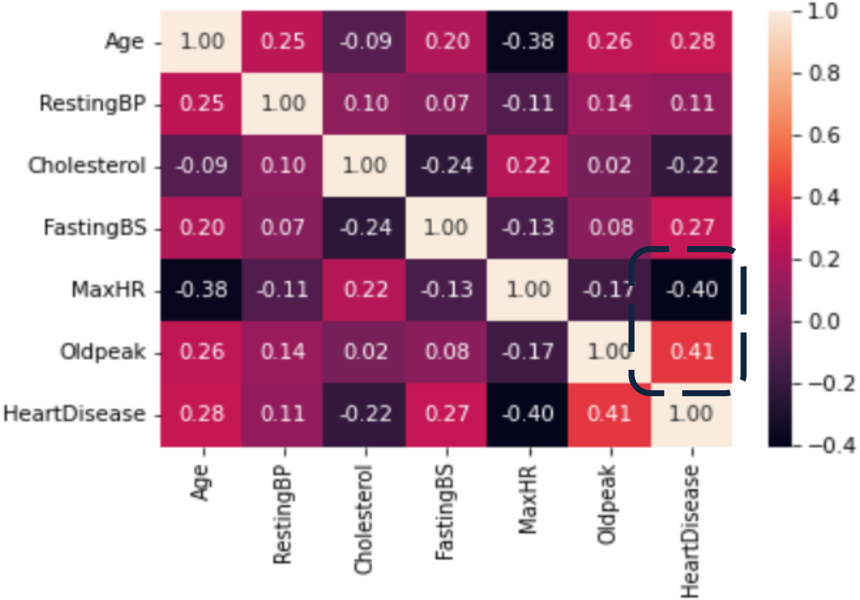
Outliers
Oldpeak
1,74%
Drop data

After filtering row becomes 902 (before 918)

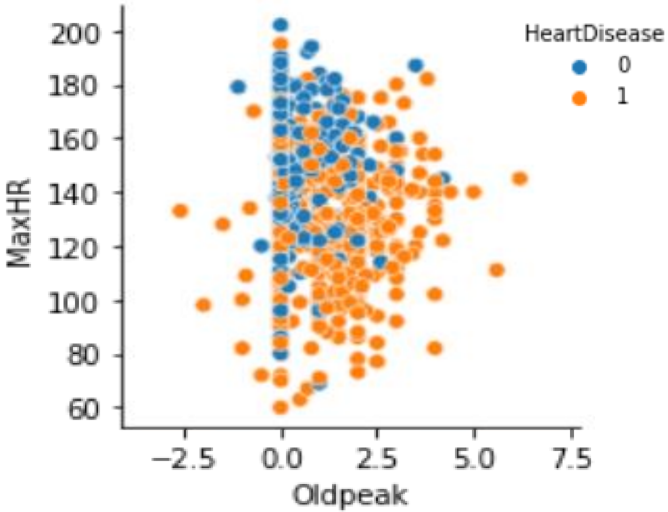
Data Preparation and Feature Engineering

Exploratory Data Analysis (EDA) Insights

Numerical Data



	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.00	0.25	-0.09	0.20	-0.38	0.26	0.28
RestingBP	0.25	1.00	0.10	0.07	-0.11	0.14	0.11
Cholesterol	-0.09	0.10	1.00	-0.24	0.22	0.02	-0.22
FastingBS	0.20	0.07	-0.24	1.00	-0.13	0.08	0.27
MaxHR	-0.38	-0.11	0.22	-0.13	1.00	-0.17	-0.40
Oldpeak	0.26	0.14	0.02	0.08	-0.17	1.00	0.41
HeartDisease	0.28	0.11	-0.22	0.27	-0.40	0.41	1.00

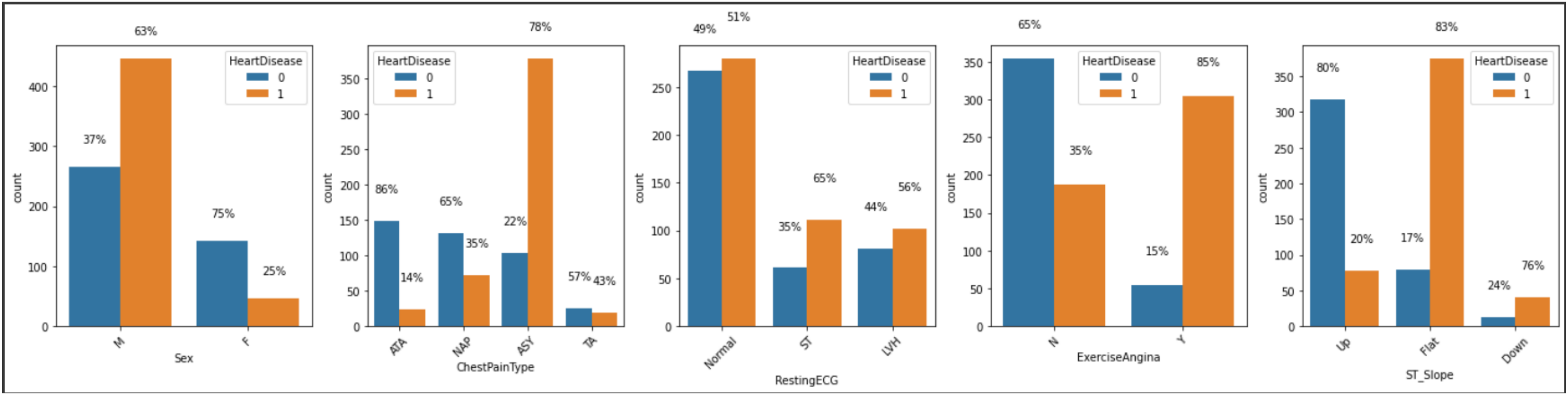


- There is no redundant feature(s)
- Oldpeak has highest correlation to heart disease (41%)
- Meanwhile MaxHR has high correlation to heart disease, but in negative way (-40%)
- The higher Oldpeak and the lower MaxHR tend have heart disease and heart failure

Data Preparation and Feature Engineering

Exploratory Data Analysis (EDA) Insights

Categorical Data



- The highest category that risky have heart disease:

Sex	: Male
ChestPainType	: ASY (Asymptomatic)
RestingECG	: Normal
ExcerciseAngina	: Yes
ST_Slope	: Flat

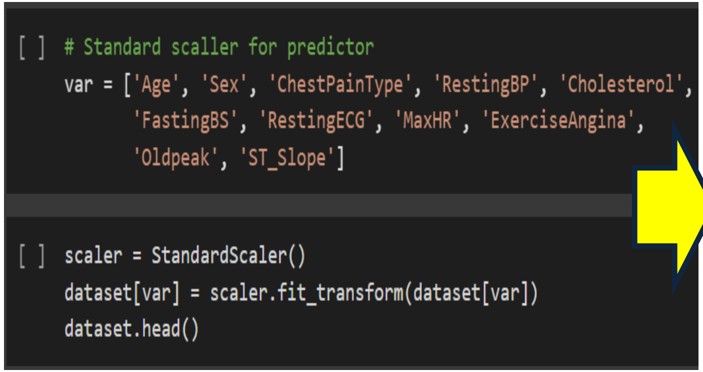
- Those condition needs to be proven by further check other aspects (ex. Lifestyle)

Data Preparation and Feature Engineering

Feature Engineering (Label Encoding)

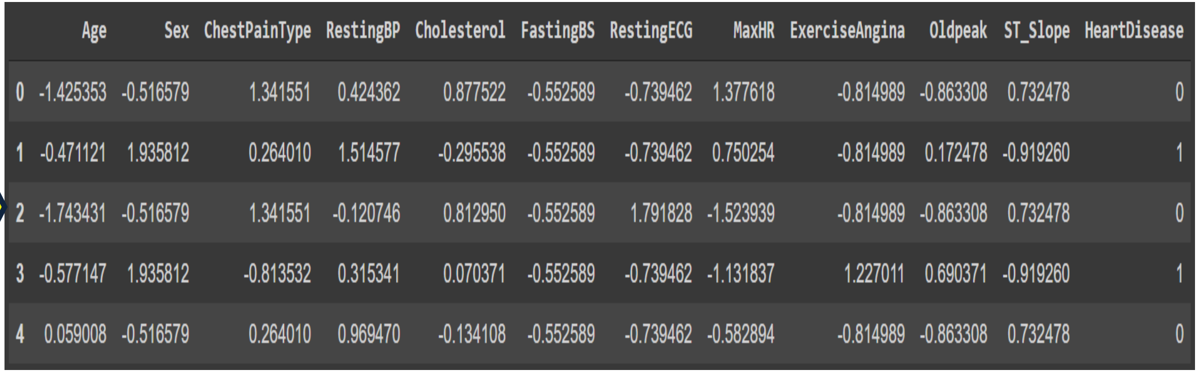
Features	0	1	2	3
Sex	M	F		
ChestPainType	ASY	NAP	ATA	TA
RestingECG	Normal	ST	LVH	
ExerciseAngina	N	Y		
ST_Slope	Up	Flat	Down	

Feature Engineering (Standardization)



```
[ ]  # Standard scaller for predictor
     var = ['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol',
            'FastingBS', 'RestingECG', 'MaxHR', 'ExerciseAngina',
            'Oldpeak', 'ST_Slope']
```

```
[ ]  scaler = StandardScaler()
     dataset[var] = scaler.fit_transform(dataset[var])
     dataset.head()
```

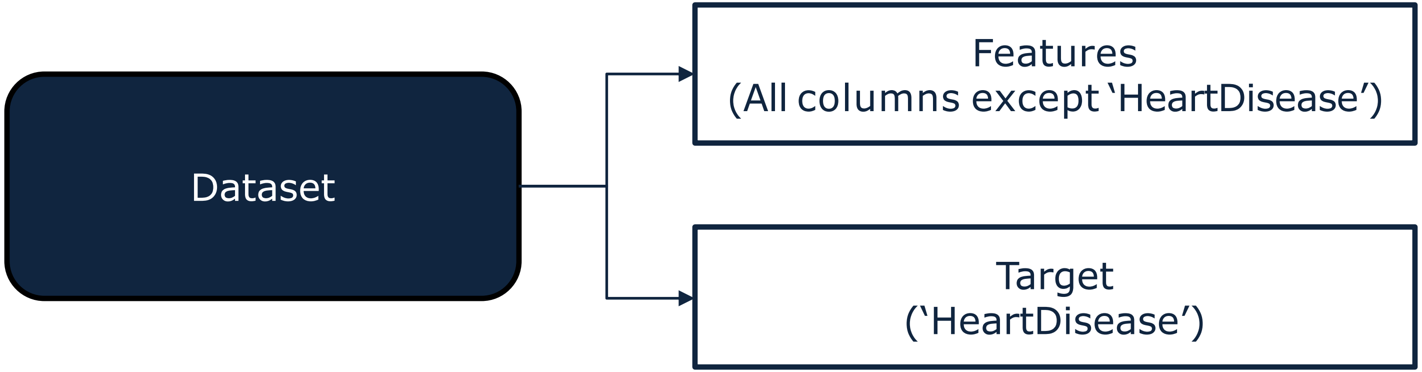


	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	-1.425353	-0.516579	1.341551	0.424362	0.877522	-0.552589	-0.739462	1.377618	-0.814989	-0.863308	0.732478	0
1	-0.471121	1.935812	0.264010	1.514577	-0.295538	-0.552589	-0.739462	0.750254	-0.814989	0.172478	-0.919260	1
2	-1.743431	-0.516579	1.341551	-0.120746	0.812950	-0.552589	1.791828	-1.523939	-0.814989	-0.863308	0.732478	0
3	-0.577147	1.935812	-0.813532	0.315341	0.070371	-0.552589	-0.739462	-1.131837	1.227011	0.690371	-0.919260	1
4	0.059008	-0.516579	0.264010	0.969470	-0.134108	-0.552589	-0.739462	-0.582894	-0.814989	-0.863308	0.732478	0

Modelling and Evaluation

Modelling and Evaluation

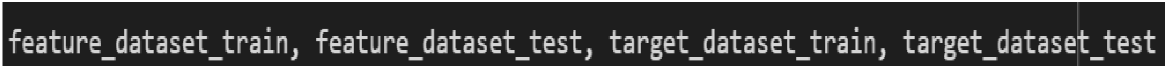
Split Data



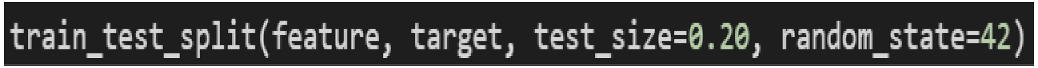
Dataset

Features
(All columns except 'HeartDisease')

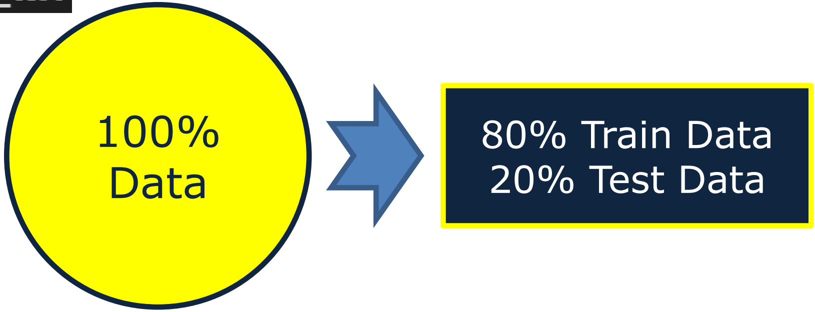
Target
('HeartDisease')



```
feature_dataset_train, feature_dataset_test, target_dataset_train, target_dataset_test
```



```
train_test_split(feature, target, test_size=0.20, random_state=42)
```



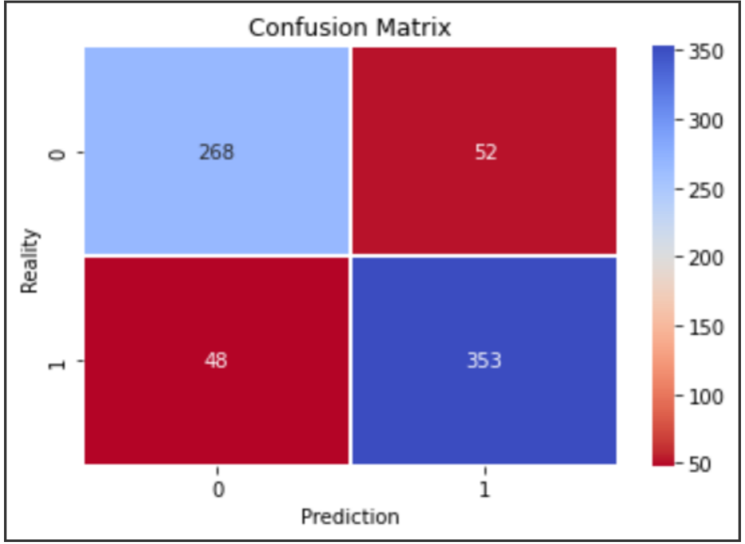
100%
Data

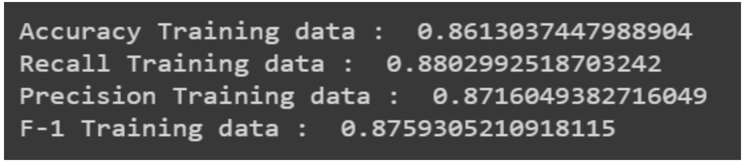
80% Train Data
20% Test Data

Modelling and Evaluation

Logistic Regression

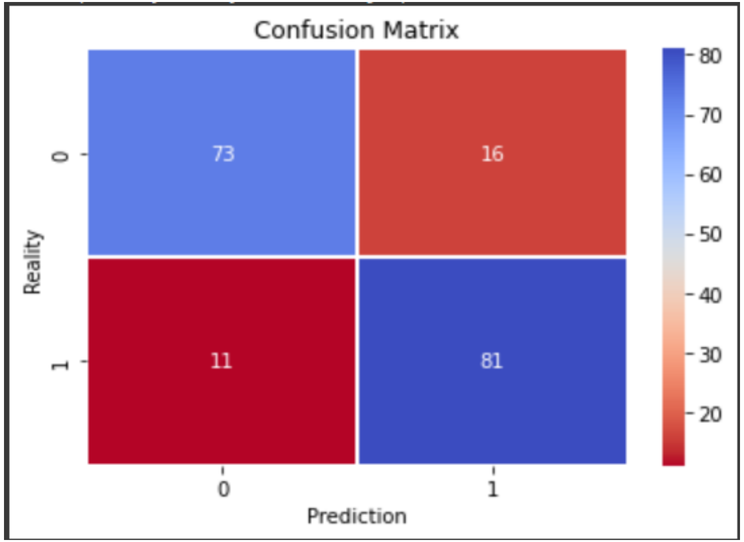
Training Data

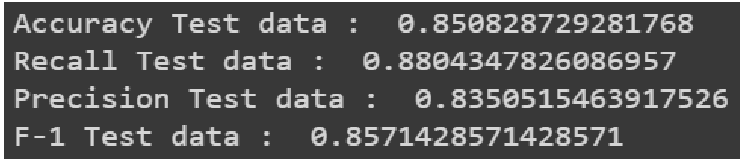




```
Accuracy Training data :  0.8613037447988904
Recall Training data :  0.8802992518703242
Precision Training data :  0.8716049382716049
F-1 Training data :  0.8759305210918115
```

Test Data





```
Accuracy Test data :  0.850828729281768
Recall Test data :  0.8804347826086957
Precision Test data :  0.8350515463917526
F-1 Test data :  0.8571428571428571
```

Model from test data not far from the model training data and has high accuracy and precision

Conclusion and Recommendation

Conclusion and Recommendation

Conclusions

- All features in the dataset are used to analysing (no redundant features)
- Individual's Old peak is the highest factor that cause heart disease and affects heart failure
- Logistic regression model is capable to used due to high accuracy and precision

Recommendations

Adding more features about lifestyle. Example:

- Smoker status
- Daily food

THANK YOU

Email : aldimeolaalfarisy@yahoo.com
LinkedIn : aldimeolaalfarisy
Github : aldimeolaalfarisy