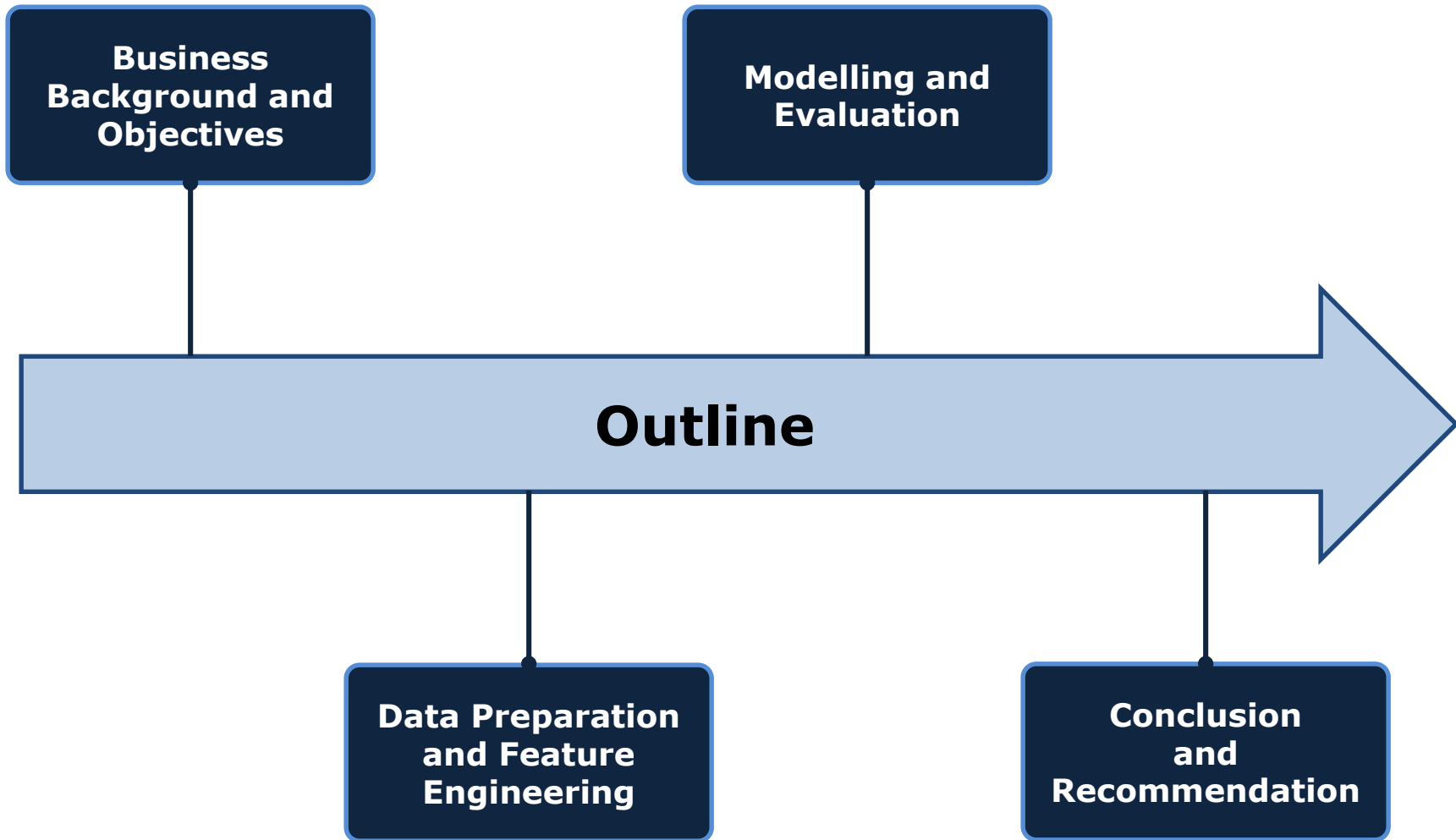


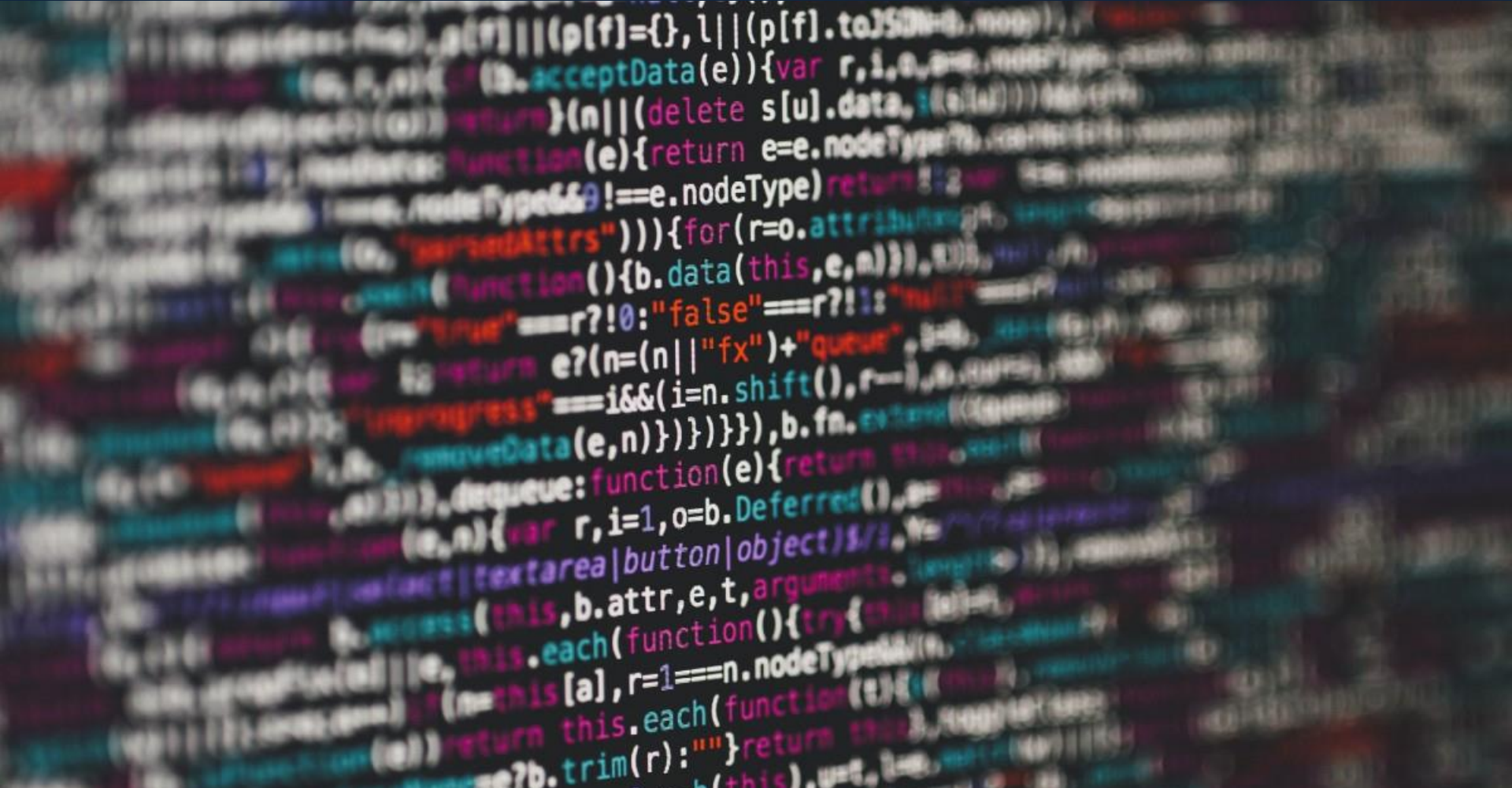


Heart Failure Analysis and Prediction with Logistic Regression

By: Aldimeola Alfarisy



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.

Objectives

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



Image source:

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

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**

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
```

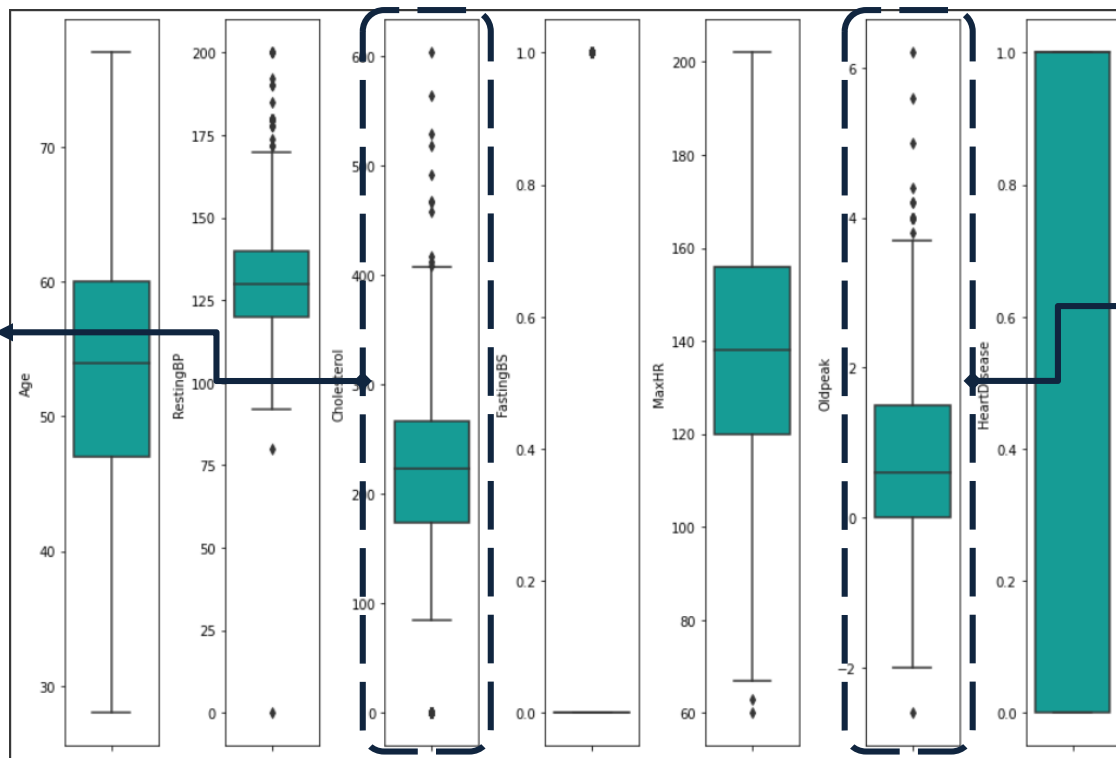
Result

- No Missing Value
- No Duplicated Value

Data Preparation and Feature Engineering

Boxplot for Numerical Data

Outliers
Cholesterol
19,93%
Trim data



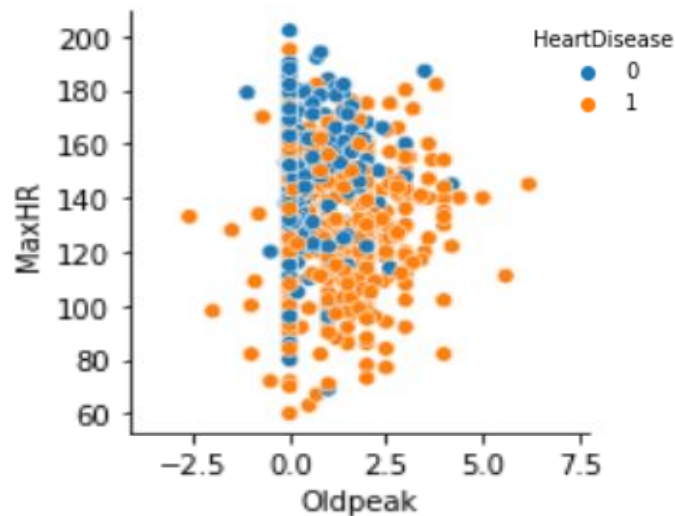
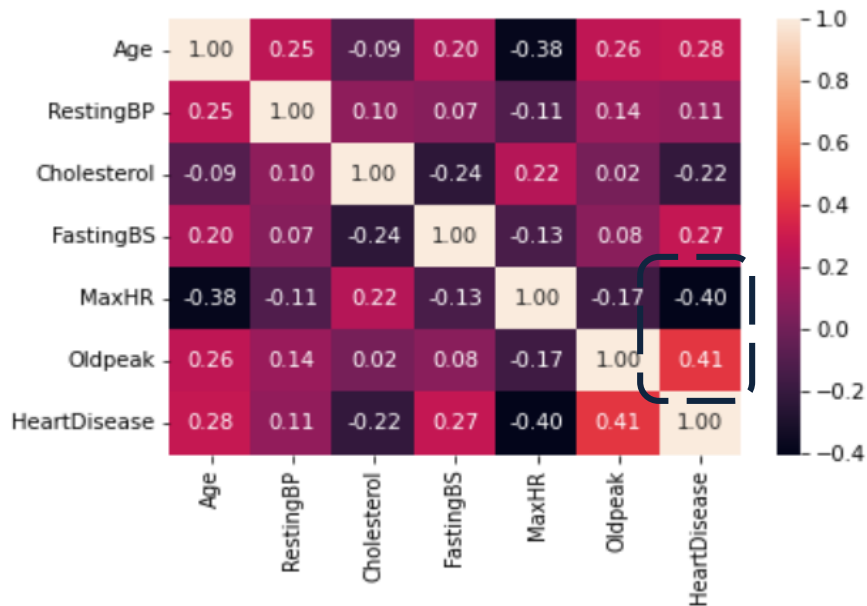
Outliers
Oldpeak
1,74%
Drop data

After filtering row becomes 902 (before 918)

Data Preparation and Feature Engineering

Exploratory Data Analysis (EDA) Insights

Numerical Data

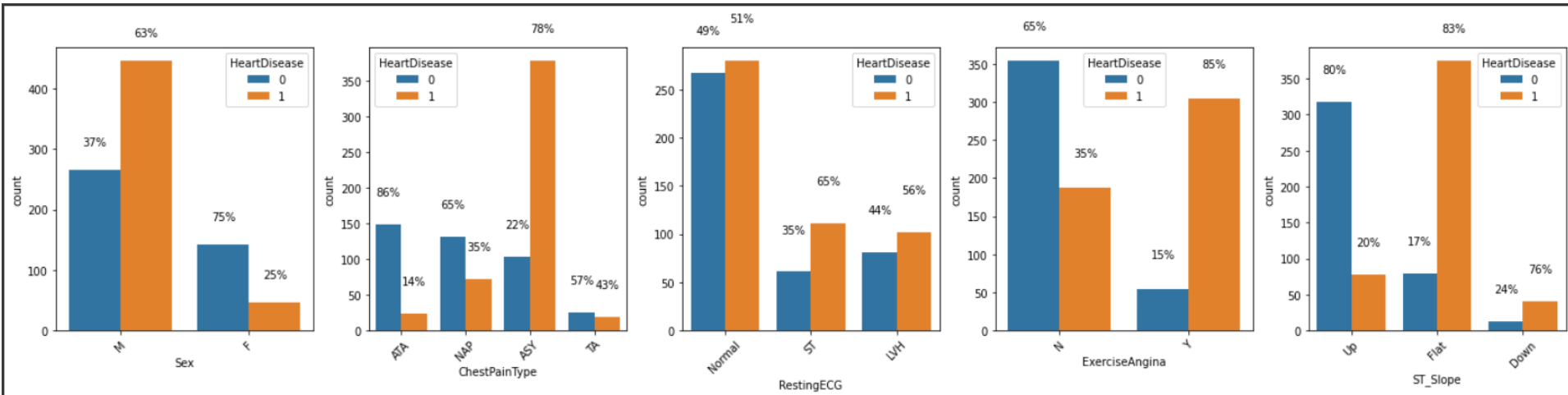


- There is no redundant feature(s)
- Oldpeak has highest correlation to heart disease (41%)
- Meanwhile MaxHR has high correlation to heart disease, but in negative way (-40%)
- The higher Oldpeak and the lower MaxHR tend have heart disease and heart failure

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)

Sex	M	F
ExerciseAngina	N	Y



Sex	0	1
ExerciseAngina	0	1

Feature Engineering (One Hot Encoding)

Column name	Values
ChestPainType	ASY, ATA, NAP, TA
RestingECG	LVH, Normal, ST
ST_Slope	Down, Flat, Up

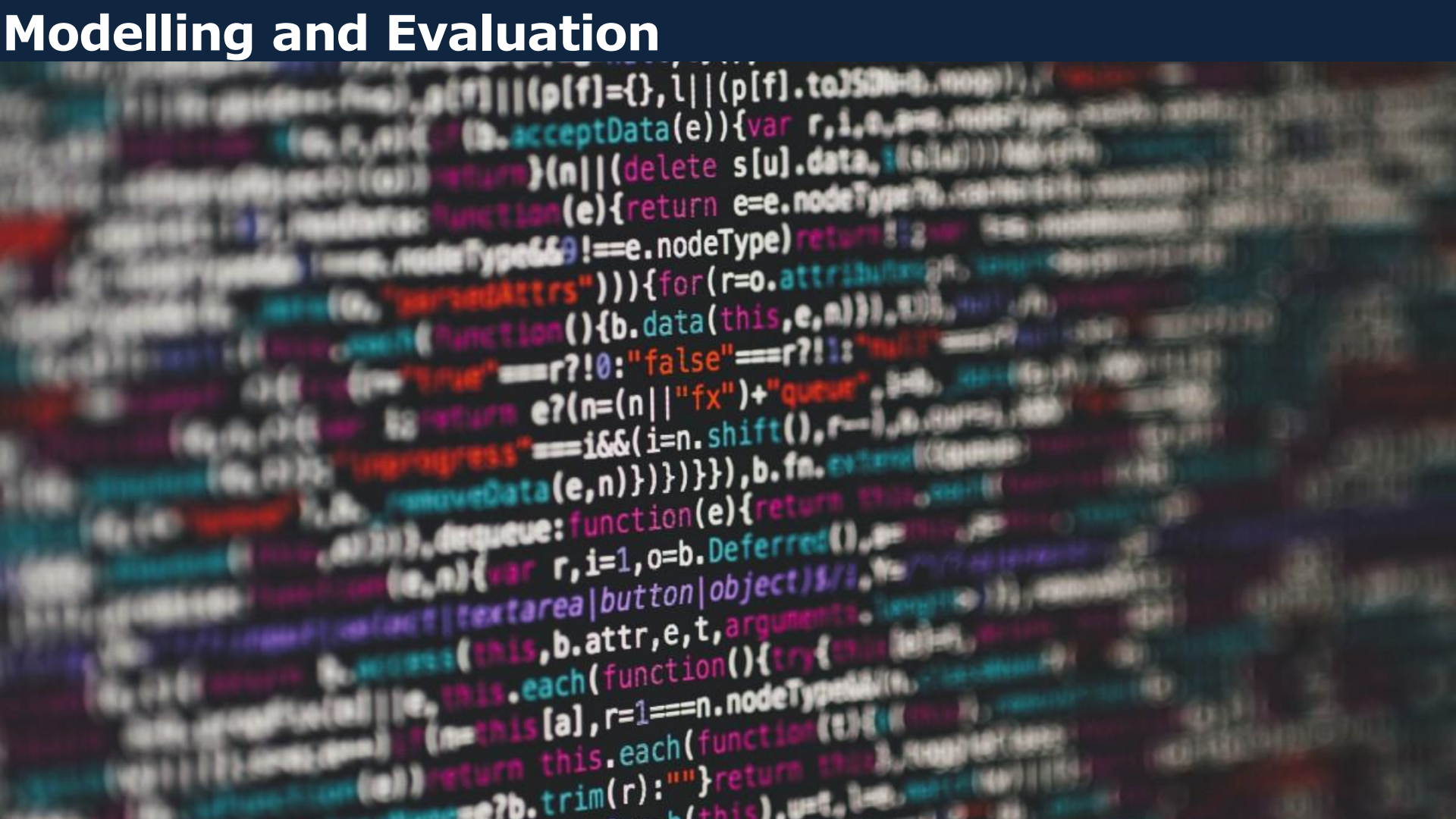


	ChestPainType_ASY	ChestPainType_ATA	ChestPainType_NAP	ChestPainType_TA
0	0	1	0	0
1	0	0	1	0
2	0	1	0	0
3	1	0	0	0
4	0	0	1	0

	RestingECG_LVH	RestingECG_Normal	RestingECG_ST
0	0	1	0
1	0	1	0
2	0	0	1
3	0	1	0
4	0	1	0

	ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up
0	0	0	1
1	0	1	0
2	0	0	1
3	0	1	0
4	0	0	1

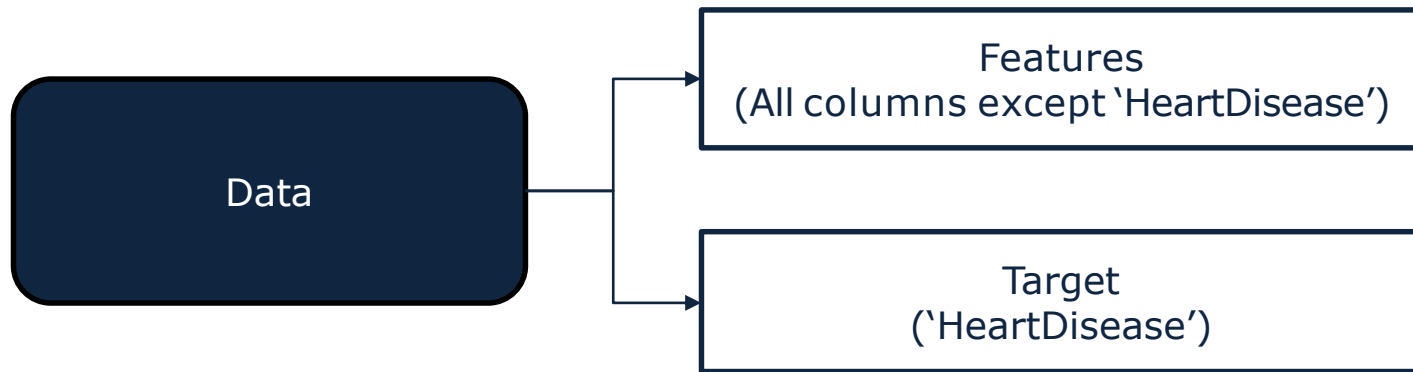
Modelling and Evaluation

The background of the slide is a blurred image of computer code, likely JavaScript or a similar language, with various syntax-highlighted elements in colors like blue, green, and red. The code is out of focus, serving as a decorative backdrop for the title.

```
... (p[f]={}), i||(p[f].toJSON=...), ...  
... (e, p, n) { if (b.acceptData(e)) { var r, i, o, ...  
... (n) { delete s[u].data, ...  
... function(e) { return e=e.nodeType ...  
... !=e.nodeType) return ...  
... ("parsedAttrs")) { for (r=0, ...  
... function() { b.data(this, e, n) ...  
... "true" == r ? !0 : "false" == r ? !1 : "null" == ...  
... return e ? (n = (n || "fx") + "queue", ...  
... "ingress" == i && (i = n.shift(), r--), ...  
... enqueueData(e, n) } } } } }, b.fn.extend({ ...  
... dequeue: function(e) { return ...  
... (e, n) { var r, i=1, o=b.Deferred(), ...  
... |button|object) % 1, ...  
... success: (this, b.attr, e, t, arguments, ...  
... this.each(function() { try { ...  
... (n=this[a], r=1==n.nodeType ...  
... return this.each(function(t) { ...  
... return this.trim(r): "" } return this ...  
... (this). ...
```

Modelling and Evaluation

Split Data



```
feature_data_fix_train, feature_data_fix_test, target_data_fix_train, target_data_fix_test
```



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

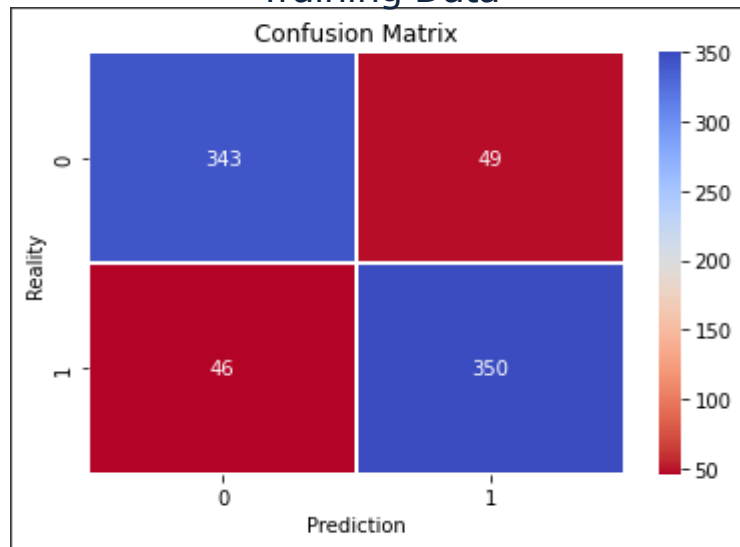


80% Train Data
20% Test Data

Modelling and Evaluation

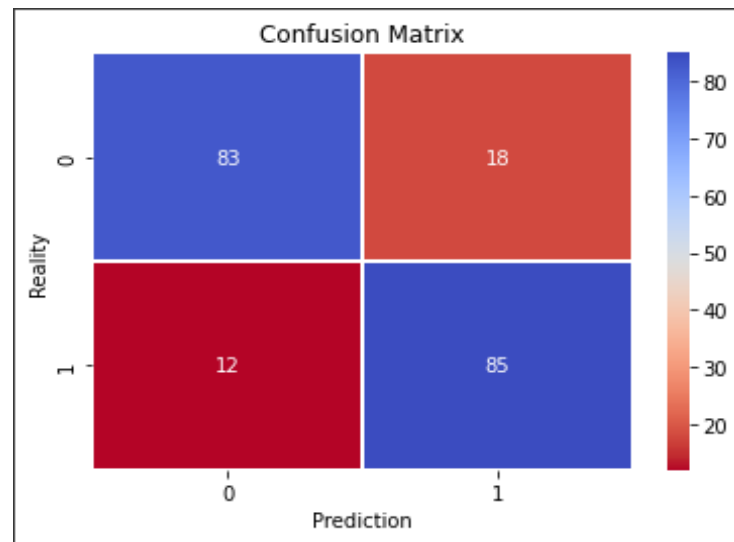
Logistic Regression

Training Data



```
Accuracy Training data : 0.8794416243654822
Recall Training data : 0.8838383838383839
Precision Training data : 0.8771929824561403
F-1 Training data : 0.8805031446540881
```

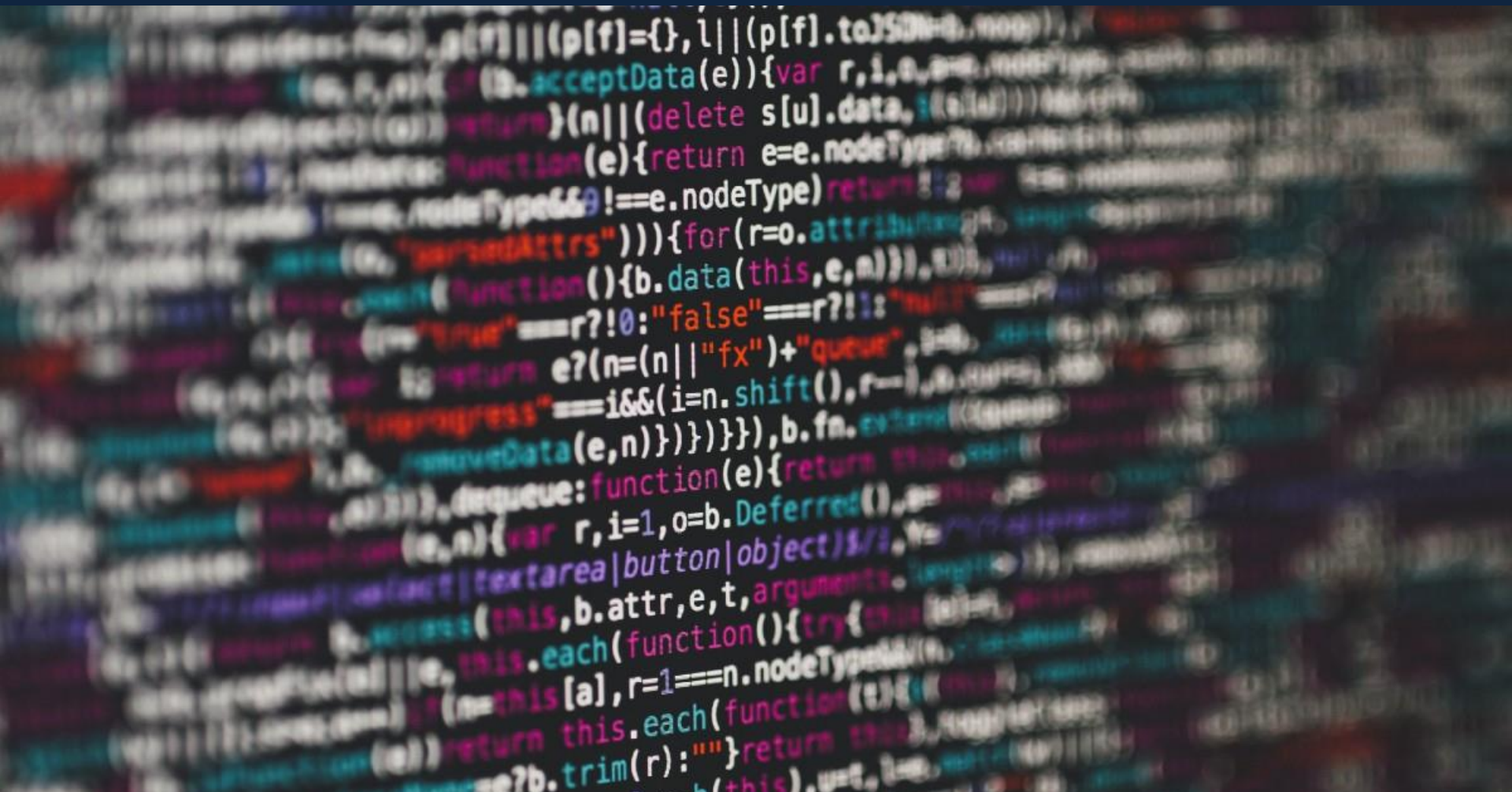
Test Data



```
Accuracy Test data : 0.8484848484848485
Recall Test data : 0.8762886597938144
Precision Test data : 0.8252427184466019
F-1 Test data : 0.85
```

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](#)