Heart Failure Analysis and Prediction

By: Aldimeola Alfarisy

TABLE OF CONTENTS

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?
- What is the best machine learning algorithm model to predict heart failure?

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

Data Preparation and Feature Engineering

Dataset Information

918 rows

11 features

Numerical Feature	Categorical Feature
• Age	• Sex
• RestingBP	• ChestPainType
• Cholesterol	• RestingECG
• FastingBS	• ExerciseAngina
• MaxHR	• ST_Slope
• Oldpeak	

1 target

Heart Disease

Dataset source: kaagle.com

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
```

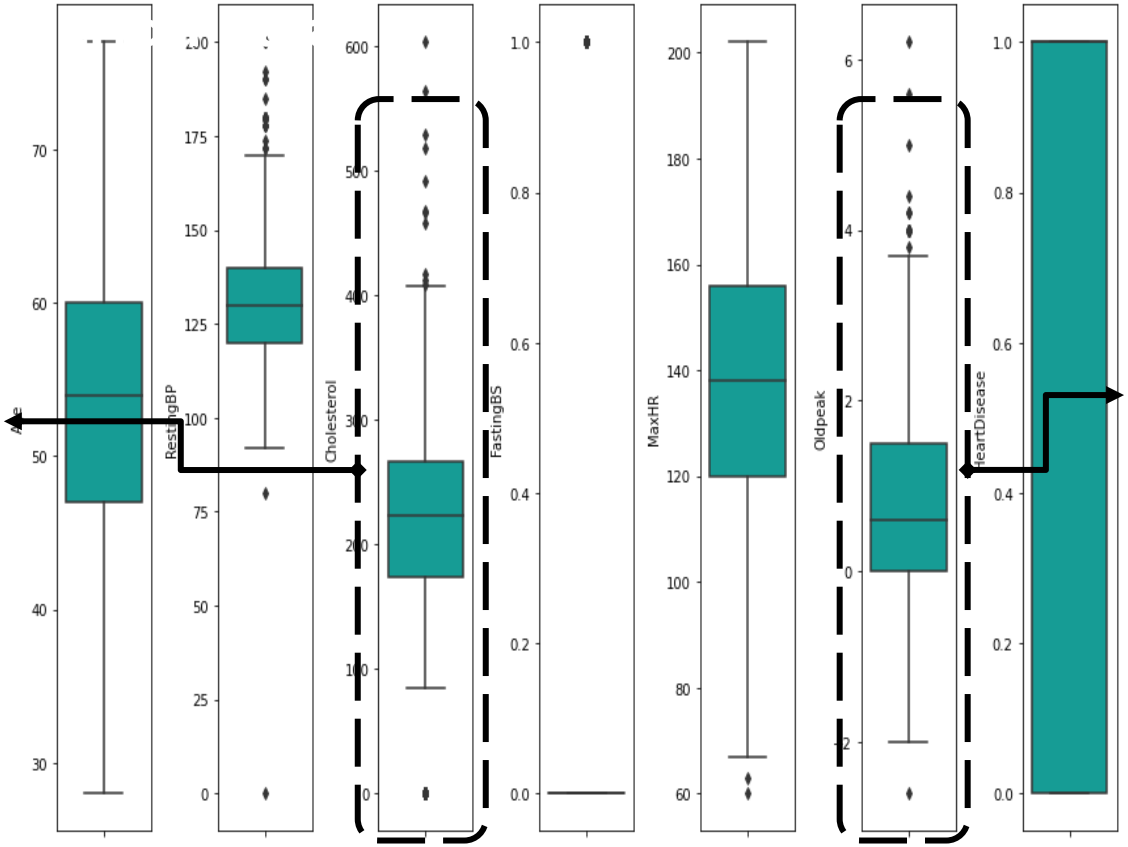
```
In [8]: data.duplicated().sum()

Out[8]: 0
```

Result

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

Data Preparation and Feature Engineering

Outliers
Cholesterol
19,93%
Trim data



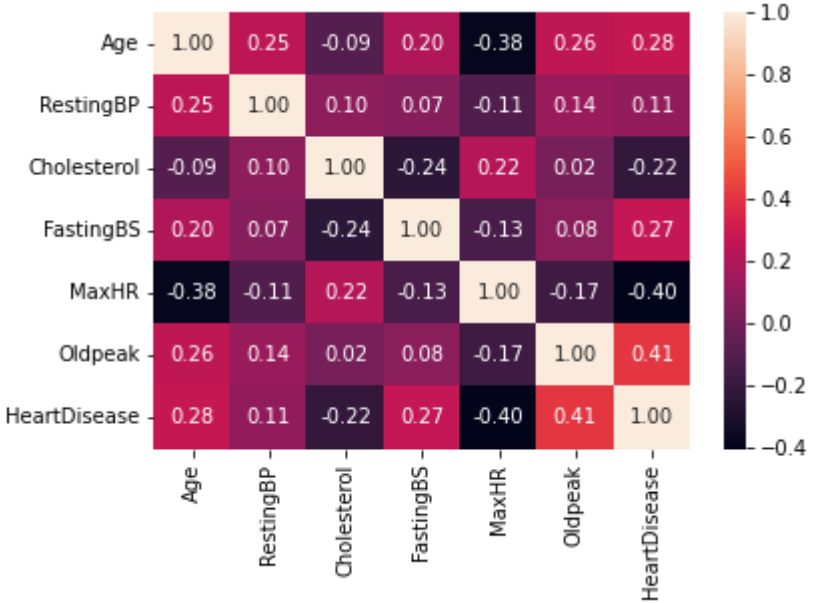
Outliers
Oldpeak
1,74%
Drop data

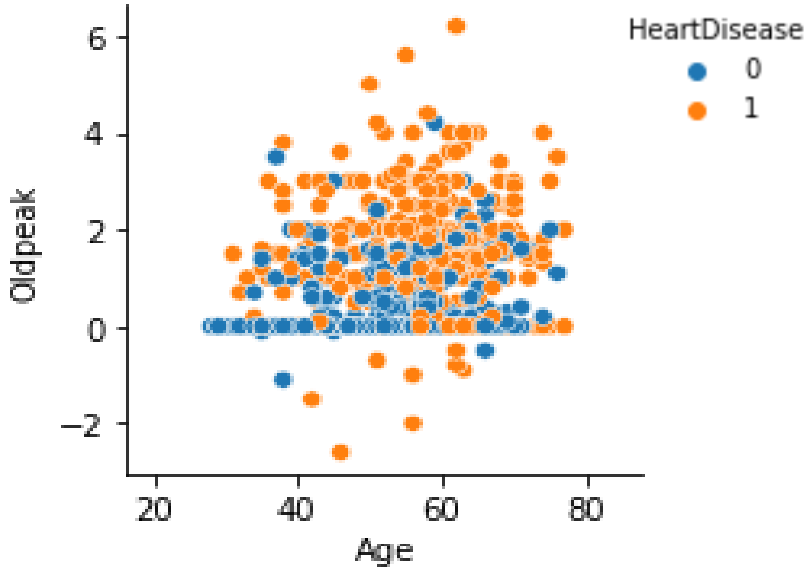
After filtering row becomes 902 before 918

Data Preparation and Feature Engineering

Exploratory Data Analysis (EDA) Insights

Numerical Data



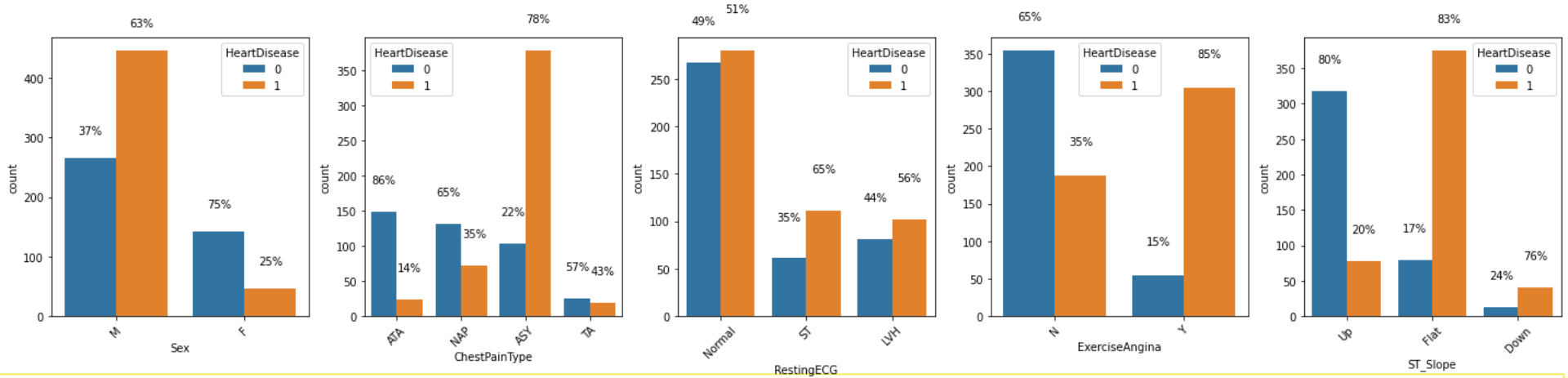


- There is no redundant feature(s)
- Oldpeak has highest correlation to heart disease (41%)
- Oldpeak has high correlation to age (26%), more mature people with high oldpeak, tend to have heart disease and heart failure

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

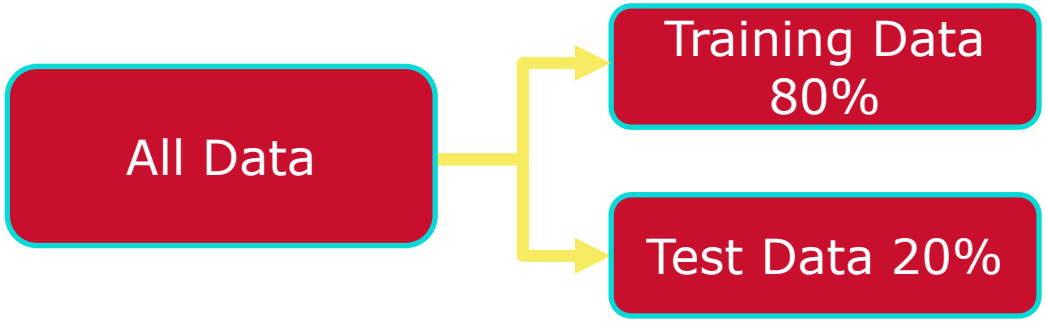
Sex	0	1
ExerciseAngina	0	1

One Hot Encoding

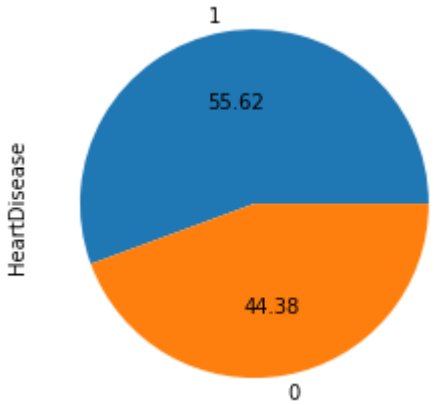
ChestPainType	ChestPainType_ASY	FChestPainType_ATA	ChestPainType_NAP	ChestPainType_TA
RestingECG	RestingECG_LVH	RestingECG_Normal	RestingECG_ST	
ST_Slope	ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up	

Modelling and Evaluation

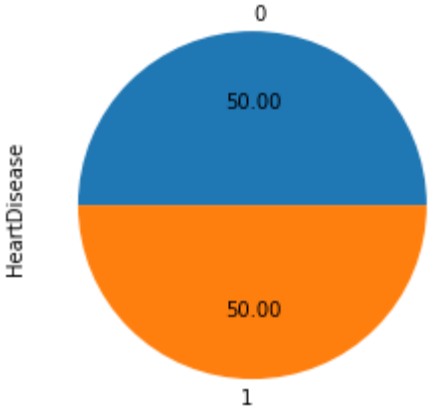
Data Split



Imbalance Data Training



Oversampling Use SMOTE



Modelling and Evaluation

Model Training with Hyper Parameter Tuning and Evaluation

Logistic Regression

```
logreg = LogisticRegression(random_state=42)
```

```
# Hyperparameter Tuning
parameters = {
    'penalty': ('l2','none'),
    'C': (1,2,5),
    'fit_intercept': [0],
    'solver': ('newton-cg', 'sag', 'lbfgs'),
    'max_iter': (2,9)
}

# note: I use recall
logreg_gridcv = GridSearchCV(logreg, parameters, cv=5, scoring='recall')
logreg_gridcv.fit(x_train, y_train)
```

Recall	91%
F-1 Score	85%
Precision	88%

Random Forest

```
RF = RandomForestClassifier(random_state=42)
```

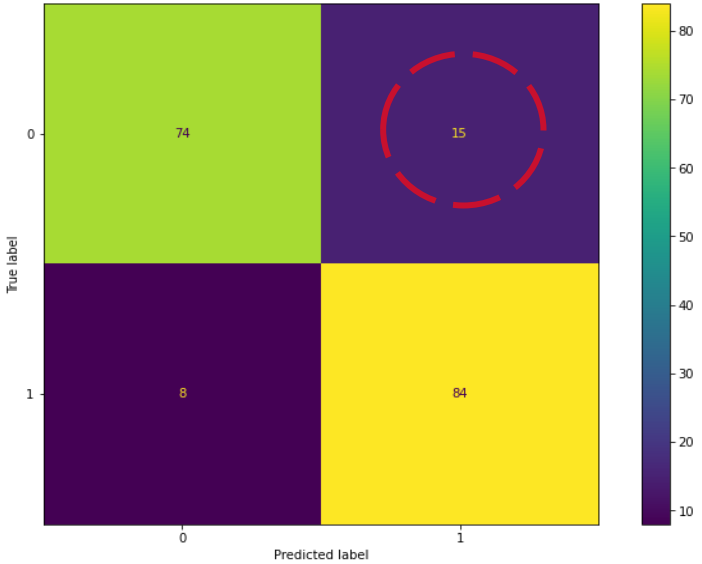
```
# Hyperparameter Tuning
parameters = {
    'max_depth': (1,3,5,7),
    'min_samples_split':(2, 5, 10),
    'min_samples_leaf': (1, 2, 4)
}

# note: I use recall
RF_gridcv = GridSearchCV(RF, parameters, cv=5, scoring='recall')
RF_gridcv.fit(x_train, y_train)
```

Recall	93%
F-1 Score	80%
Precision	86%

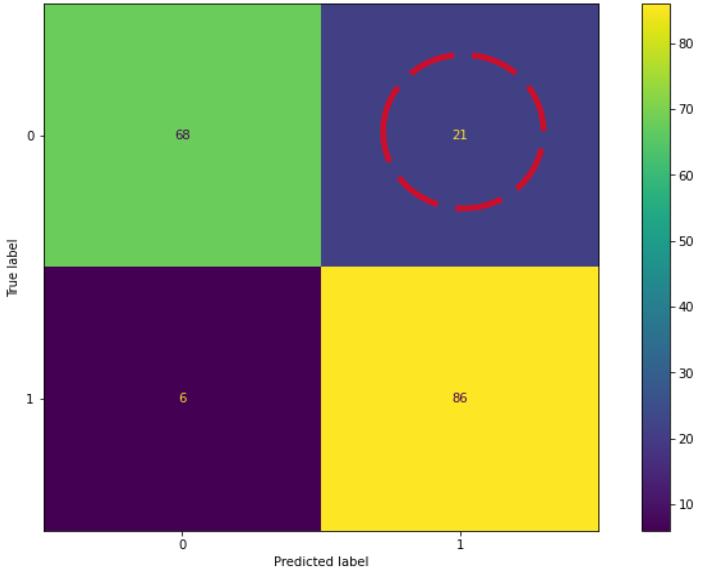
Modelling and Evaluation

The Best Algorithm

Logistic Regression



Random Forest

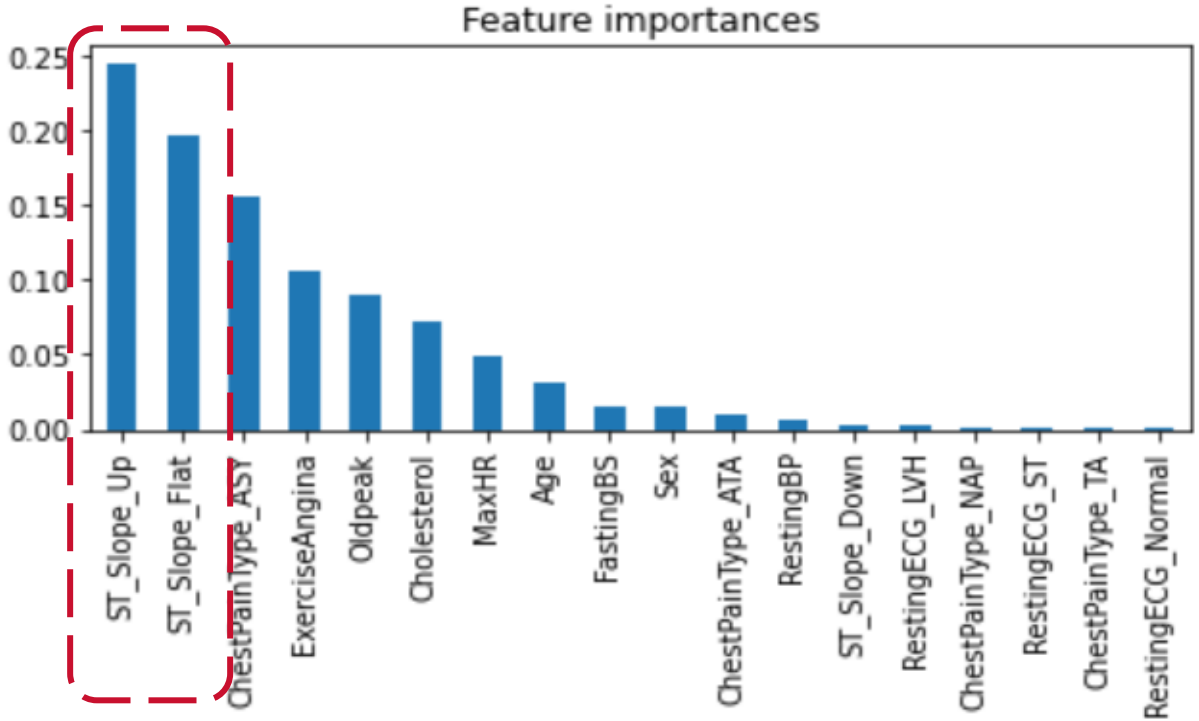


False positive in these confusion metric:
Model predicts people have heart disease, but they don't

Logistic Regression is suitable algorithm to predict the tendetion of heart failure

Modelling and Evaluation

Feature Importance



The slope of the peak exercise ST segment become the highest factor that cause heart failure

Conclusion and Recommendation

- All features in the dataset are used to analysing (no redundant features)
- Individual ST_Slope become the highest factor that can cause heart failure.
- Best algorithm to use is logistic regression

Adding more features about lifestyle. Example:

- Smoker status
- Daily food

THANK YOU

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