INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

# Applied Data Science

## Capstone Project – Healthcare

# Cardiovascular diseases (CVDs)

**Capstone Project** Submitted by:

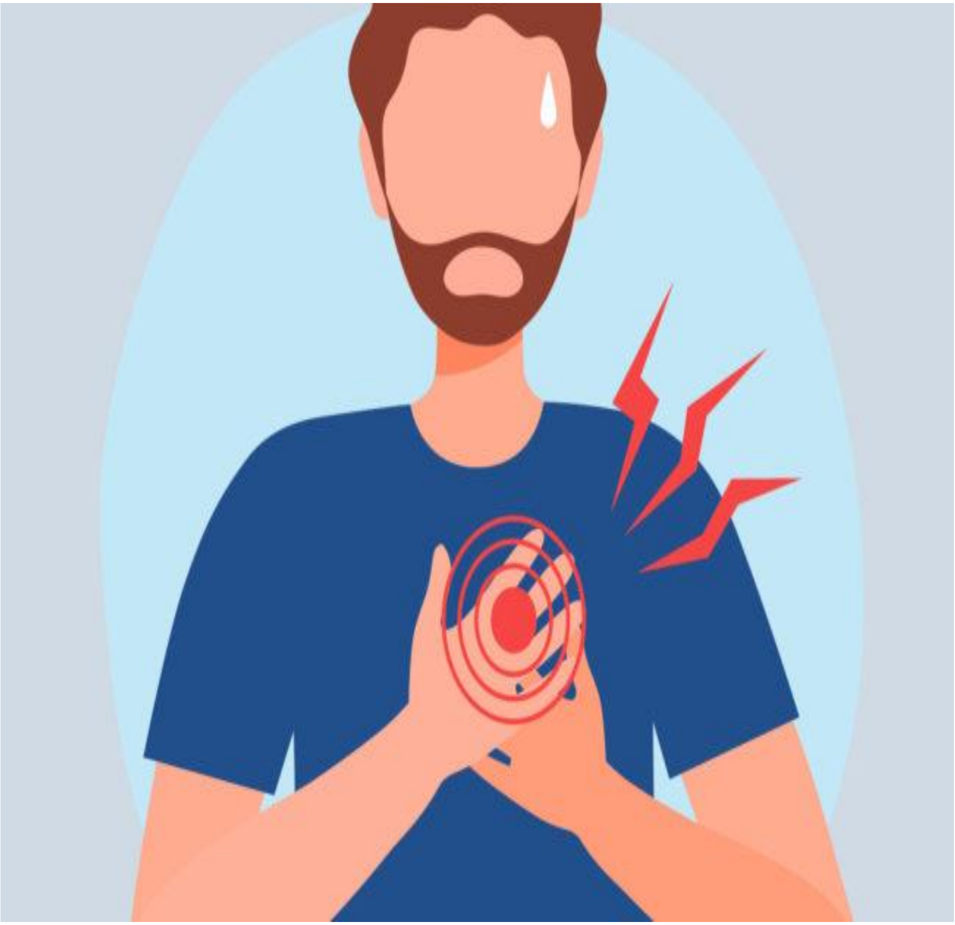
**PRAKASH S**

**Batch Code-IITPKD ADS Async Jun 2022**

INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

# Overview

- Cardiovascular diseases (CVDs) are the leading cause of death globally.
- An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.
- Over three quarters of CVD deaths take place in low- and middle-income countries.
- Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.
- Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.
- It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

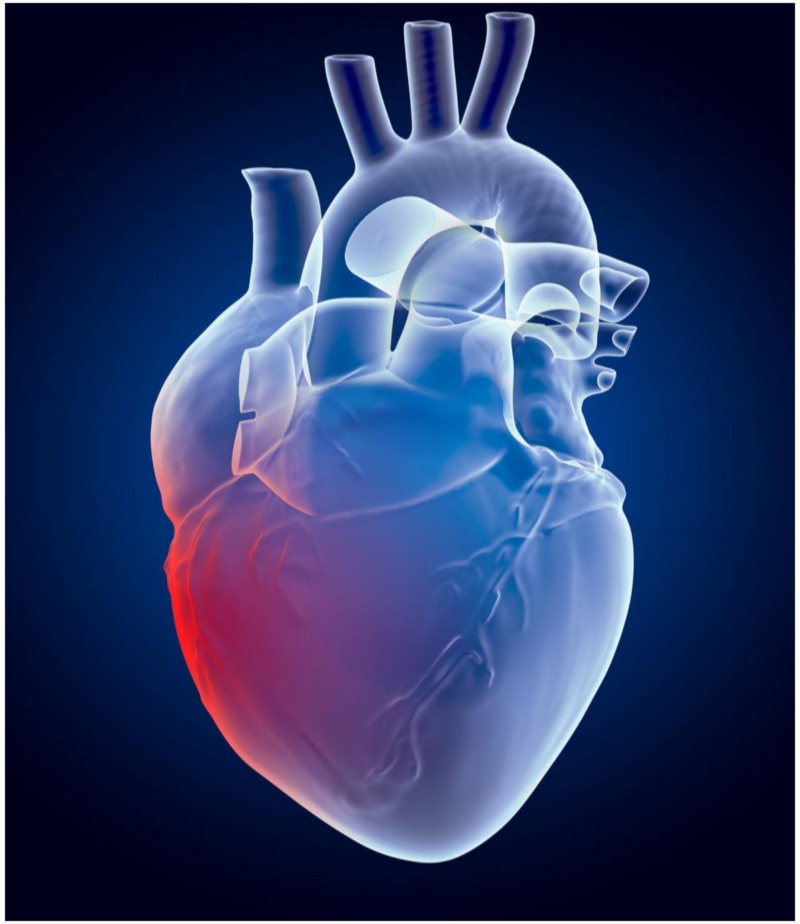


# Problem Statement

- Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

- Create a model for predicting mortality caused by Heart Failure.



## Project Scope

- Data Extraction and Data preparation for Analysis.
- Perform Exploratory Data Analysis to find patterns or trends on Patients data.
- Create a model for predicting mortality caused by Heart Failure.

## Cardiovascular diseases (CVDs)

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risks factors” can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications.

# Dataset

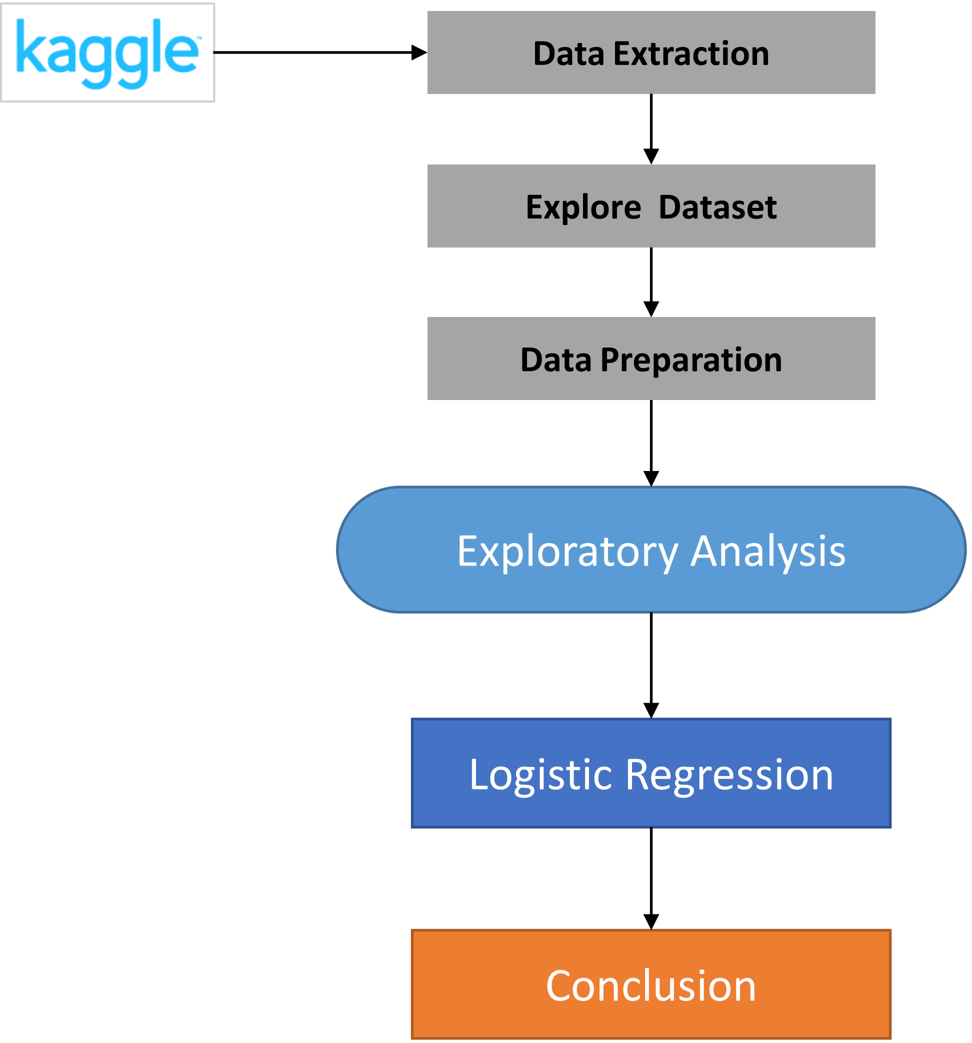
**Data Source** - kaggle

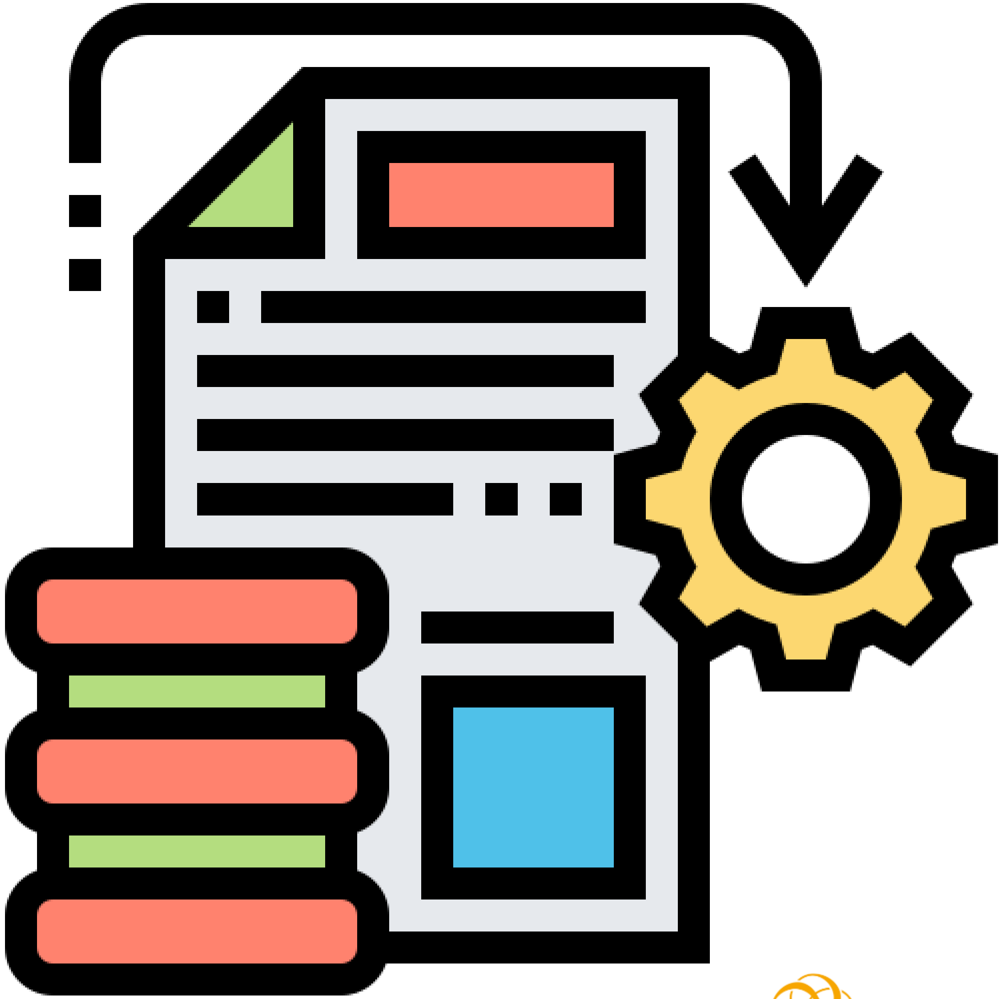
Data file containing 13 columns and 299 rows.

**Features** :

- Age : Age of the patient(Years)
- Anaemia : Decrease of red blood cells or hemoglobin (Boolean)
- Creatinine phosphokinase (CPK) : Level of the CPK enzyme in the blood (mcg/L)
- Diabetes : If the patient has diabetes (Boolean)
- Ejection fraction : Percentage of blood leaving the heart at each contraction (Percentage)
- High blood pressure : if the patient has hypertension (Boolean)
- Platelets : Platelets in the blood (Kiloplatelets/ml)
- Serum creatinine :Level of serum creatinine in the blood (mg/dL)
- Serum sodium - Level of serum sodium in the blood (mEg/L)
- Sex : Woman or Man (Binary)
- Smoking : If the person is smoking or not (Boolean)
- Time : Follow-Up period(Days)
- Death event : If the patient deceased during the follow-up period (Boolean)

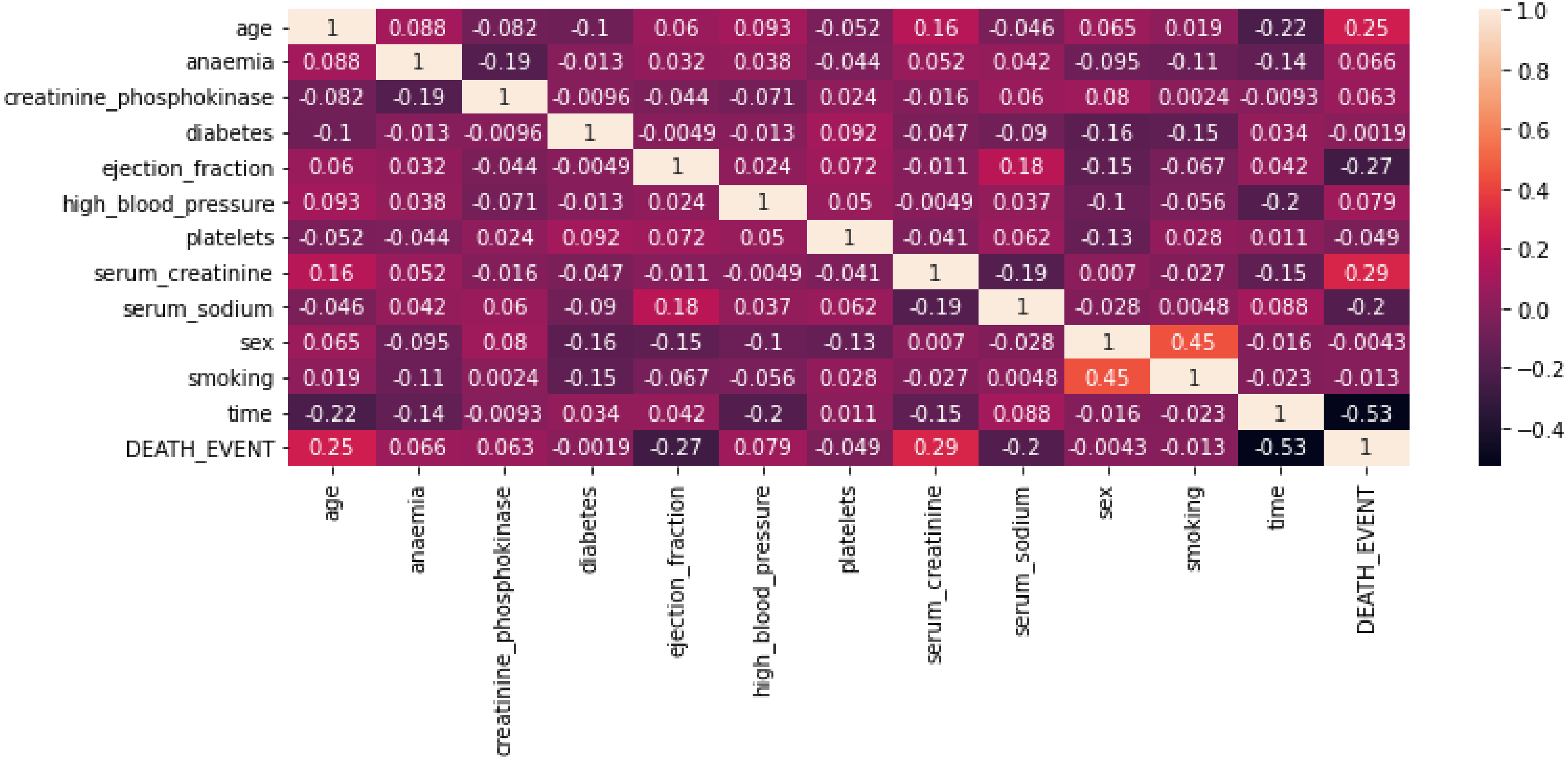
# Process Flow





# Analysis

❑ **HeatMap Visual of the data**

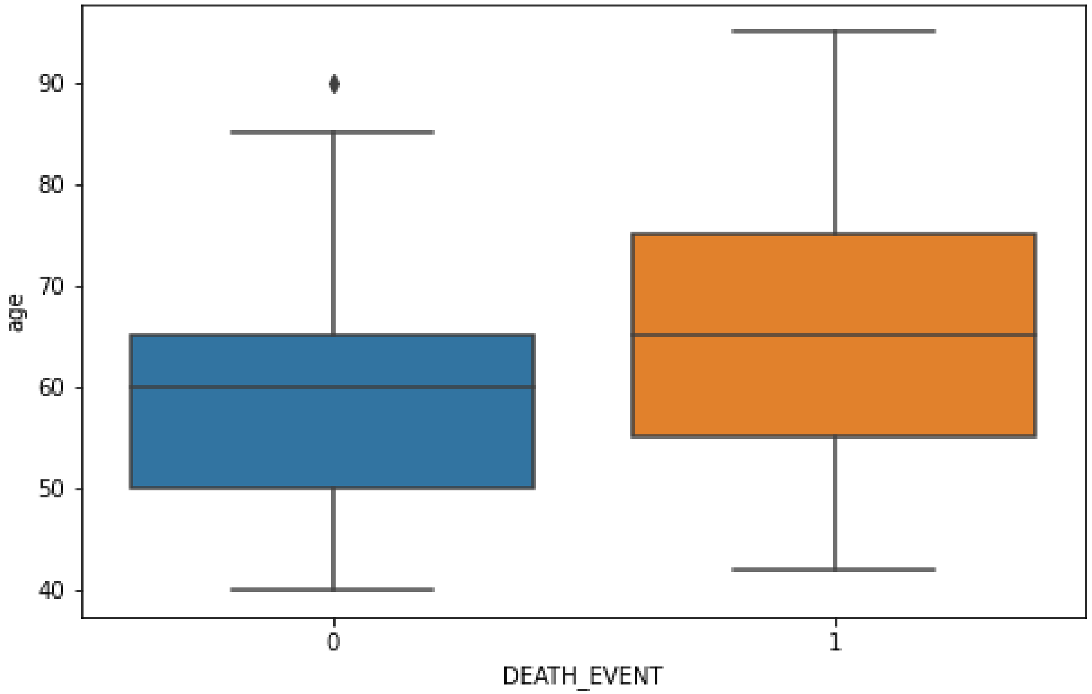


| | age | anaemia | creatinine_phosphokinase | diabetes | ejection_fraction | high_blood_pressure | platelets | serum_creatinine | serum_sodium | sex | smoking | time | DEATH_EVENT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | 0.088 | -0.082 | -0.1 | 0.06 | 0.093 | -0.052 | 0.16 | -0.046 | 0.065 | 0.019 | -0.22 | 0.25 |
| anaemia | 0.088 | 1 | -0.19 | -0.013 | 0.032 | 0.038 | -0.044 | 0.052 | 0.042 | -0.095 | -0.11 | -0.14 | 0.066 |
| creatinine_phosphokinase | -0.082 | -0.19 | 1 | -0.0096 | -0.044 | -0.071 | 0.024 | -0.016 | 0.06 | 0.08 | 0.0024 | -0.0093 | 0.063 |
| diabetes | -0.1 | -0.013 | -0.0096 | 1 | -0.0049 | -0.013 | 0.092 | -0.047 | -0.09 | -0.16 | -0.15 | 0.034 | -0.0019 |
| ejection_fraction | 0.06 | 0.032 | -0.044 | -0.0049 | 1 | 0.024 | 0.072 | -0.011 | 0.18 | -0.15 | -0.067 | 0.042 | -0.27 |
| high_blood_pressure | 0.093 | 0.038 | -0.071 | -0.013 | 0.024 | 1 | 0.05 | -0.0049 | 0.037 | -0.1 | -0.056 | -0.2 | 0.079 |
| platelets | -0.052 | -0.044 | 0.024 | 0.092 | 0.072 | 0.05 | 1 | -0.041 | 0.062 | -0.13 | 0.028 | 0.011 | -0.049 |
| serum_creatinine | 0.16 | 0.052 | -0.016 | -0.047 | -0.011 | -0.0049 | -0.041 | 1 | -0.19 | 0.007 | -0.027 | -0.15 | 0.29 |
| serum_sodium | -0.046 | 0.042 | 0.06 | -0.09 | 0.18 | 0.037 | 0.062 | -0.19 | 1 | -0.028 | 0.0048 | 0.088 | -0.2 |
| sex | 0.065 | -0.095 | 0.08 | -0.16 | -0.15 | -0.1 | -0.13 | 0.007 | -0.028 | 1 | 0.45 | -0.016 | -0.0043 |
| smoking | 0.019 | -0.11 | 0.0024 | -0.15 | -0.067 | -0.056 | 0.028 | -0.027 | 0.0048 | 0.45 | 1 | -0.023 | -0.013 |
| time | -0.22 | -0.14 | -0.0093 | 0.034 | 0.042 | -0.2 | 0.011 | -0.15 | 0.088 | -0.016 | -0.023 | 1 | -0.53 |
| DEATH_EVENT | 0.25 | 0.066 | 0.063 | -0.0019 | -0.27 | 0.079 | -0.049 | 0.29 | -0.2 | -0.0043 | -0.013 | -0.53 | 1 |

# Correlated Features

- According to the heatmap visualisation, four features are correlated to DEATH_EVENT. Those are Age, Ejection Fraction, Time(Follow-Up period), Serum Creatinine.

## 1. Death Event vs Age

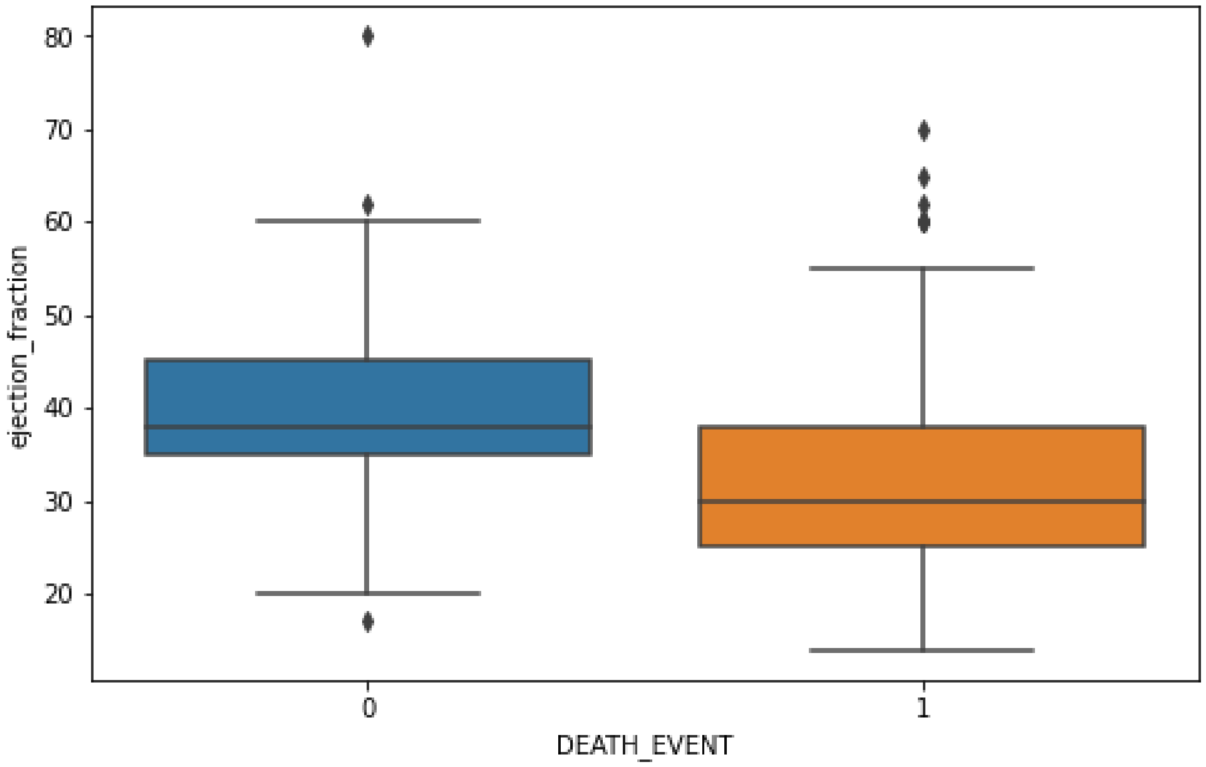


**Observation :**

- Elder peoples are more affected from heart disease into death.

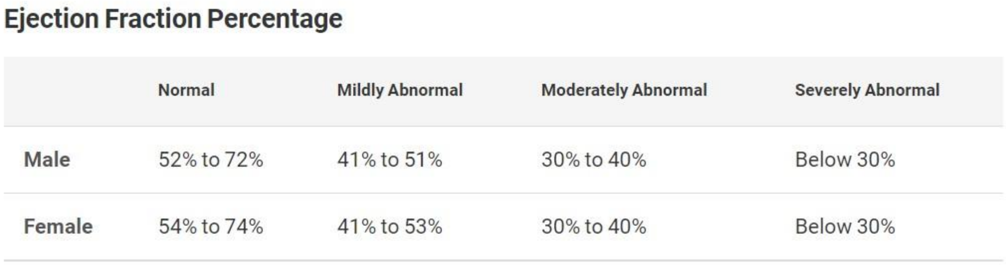
# Correlated Features

## 2. Death Event vs Ejection Fraction



**Ejection fraction :**

- Ejection fraction refers to how well your heart pumps blood. It is the amount of blood pumped out of your heart’s lower chambers (ventricles) each time it contracts.



**Ejection Fraction Percentage**

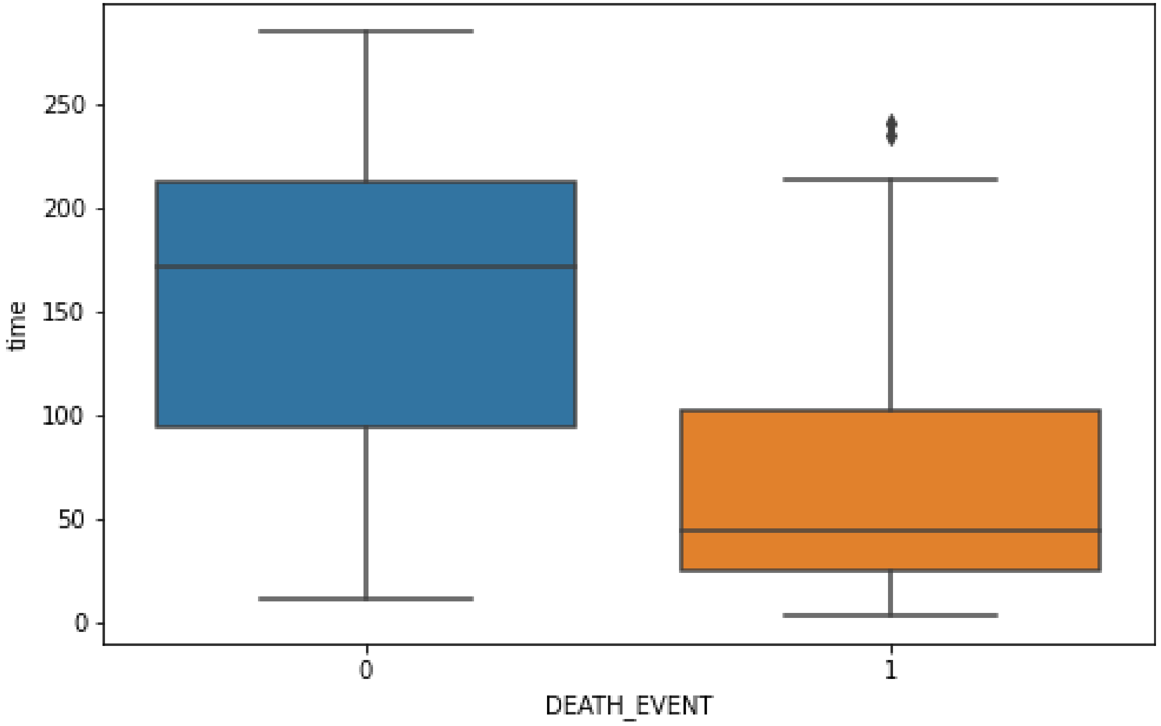
| | Normal | Mildly Abnormal | Moderately Abnormal | Severely Abnormal |
|---|---|---|---|---|
| **Male** | 52% to 72% | 41% to 51% | 30% to 40% | Below 30% |
| **Female** | 54% to 74% | 41% to 53% | 30% to 40% | Below 30% |

**Observation :**

- The person who is having low ejection fraction cause death.

# Correlated Features

### 3. Death Event vs Time(Follow-up Period)

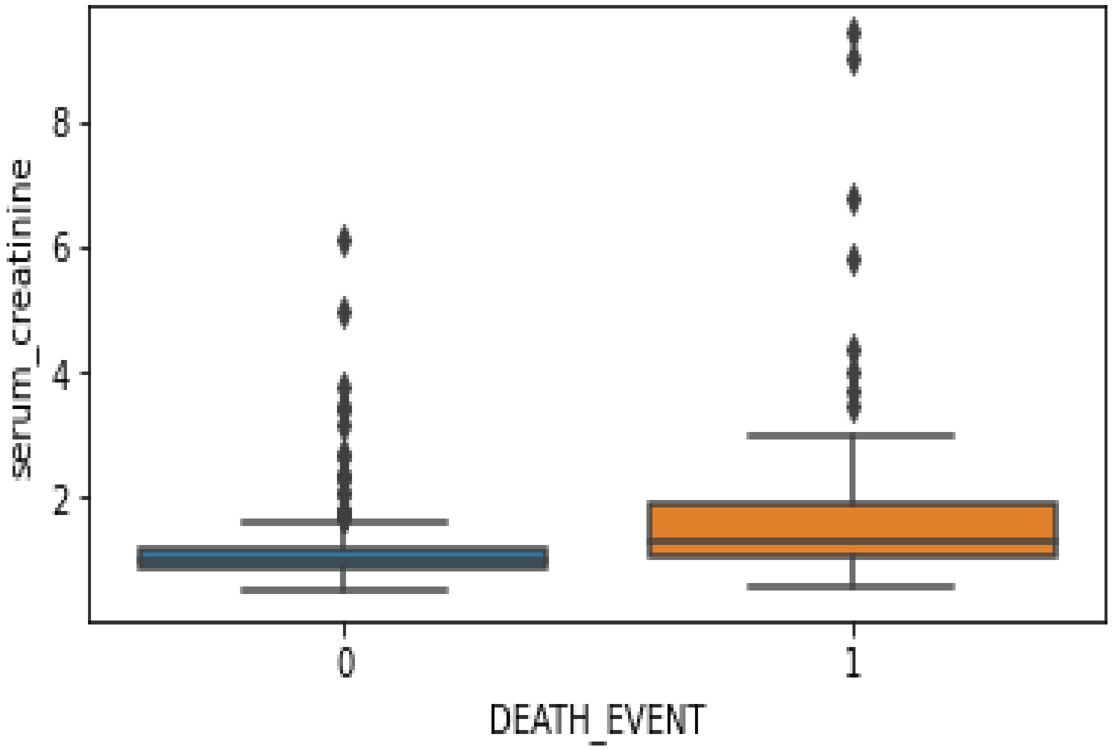


**Observation :**

- The person who done regular check-up in hospital lives more.

# Correlated Features

## 4. Death Event vs Serum Creatinine



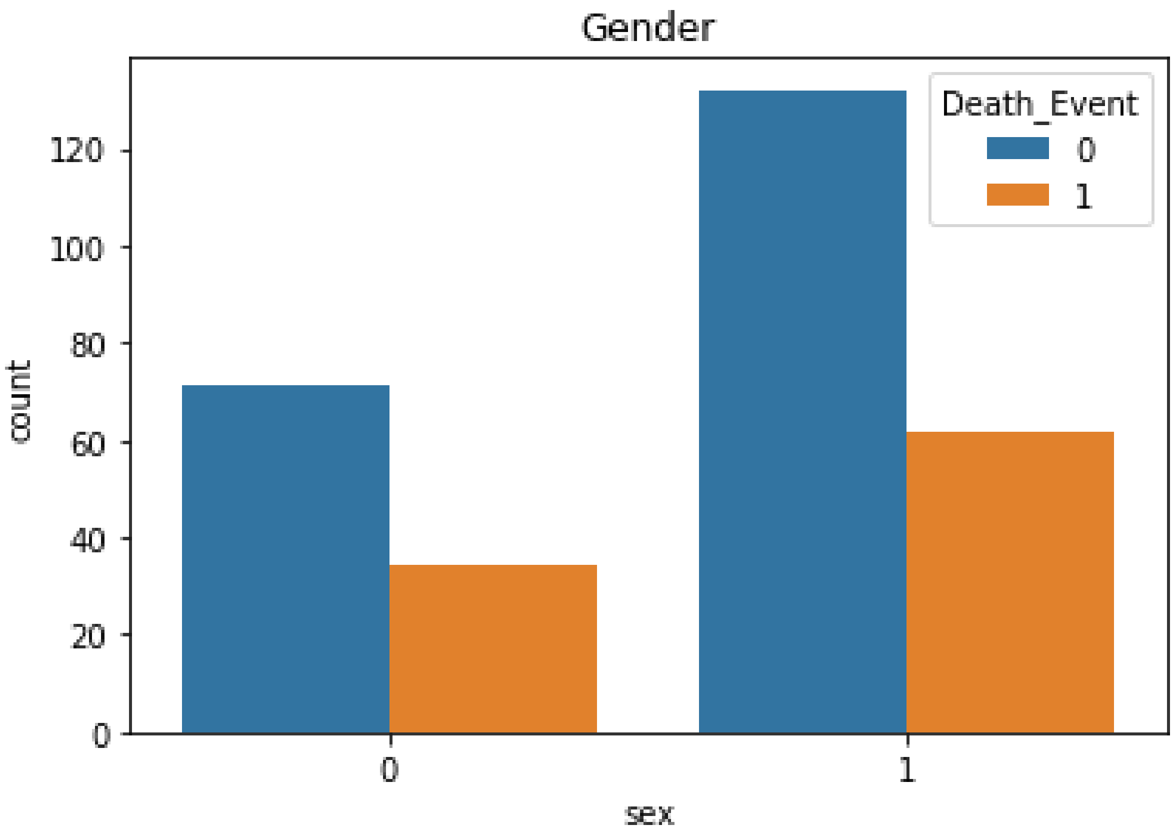
### Serum creatinine

- Serum creatinine is reported as **milligrams of creatinine to a deciliter of blood (mg/dL) or micromoles of creatinine to a liter of blood (micromoles/L)**. The typical range for serum creatinine is: For adult men, 0.74 to 1.35 mg/dL (65.4 to 119.3 micromoles/L).

**Observation :**

- The person who is having high serum creatinine level cause death compared to low serum creatine level.

# Analysis on Gender



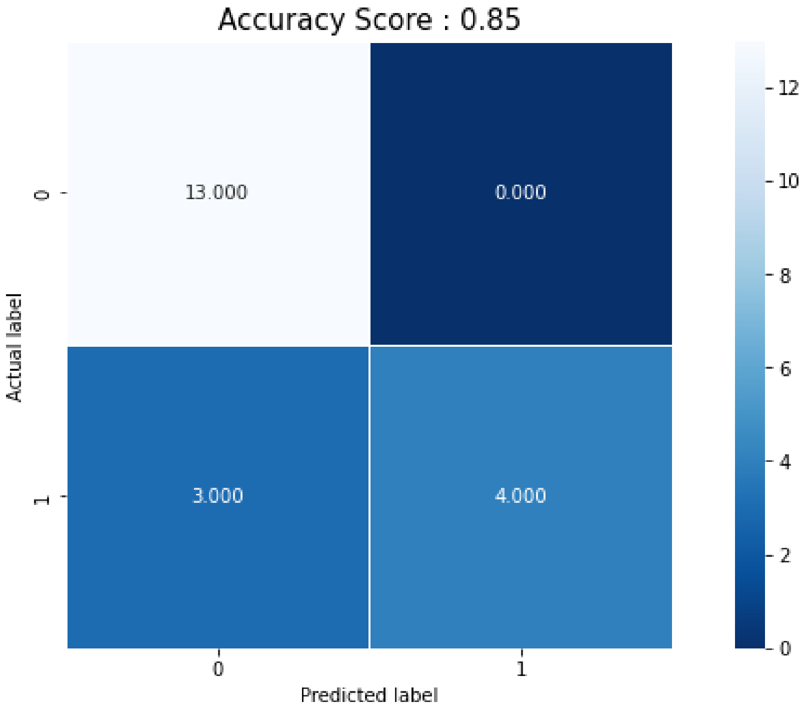
**Observation :**

- Death events are more happened for male compared to female. And male is more affected by heart disease.
- Around 90 people were died because of heart disease.

# ML Model 1

**Logistic Regression**

- Machine learning model was created by Logistic Regression initially. And it has 85% accuracy comparatively this accuracy is low.
- The features in this data is comparatively low. The dataset contains 13 columns only.
- Here we used 12 features as parameters excluded Death_event.

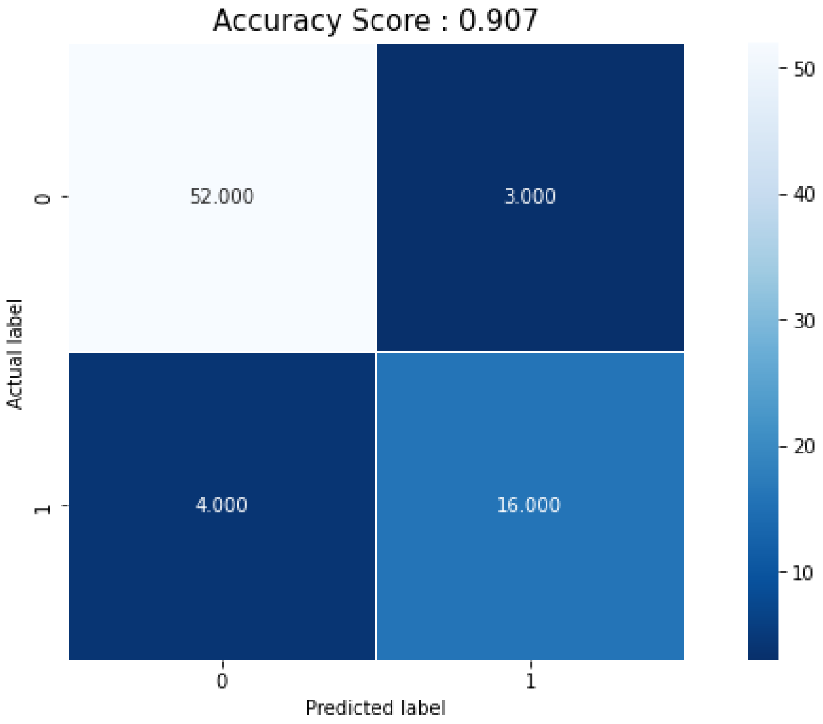


INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

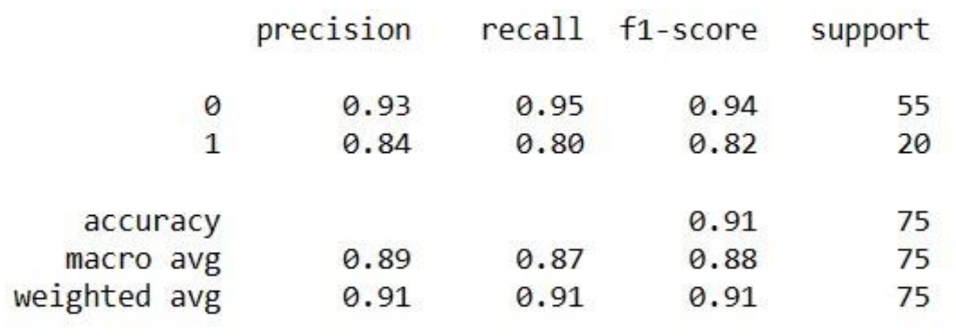
# ML Model 2

## Random Forest Classification

- Random forest classification is used here to get more accuracy to predict the mortality caused by heart disease. The accuracy is 90% in this model.
- The features in this data is comparatively low. The dataset contains 13 columns only.
- Here we used 12 features as parameters excluded Death_event.



# Classification Report



```
              precision    recall  f1-score   support

           0       0.93      0.95      0.94        55
           1       0.84      0.80      0.82        20

    accuracy                           0.91        75
   macro avg       0.89      0.87      0.88        75
weighted avg       0.91      0.91      0.91        75
```

- The precision is the ratio P =tp/(tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier to not label a sample as positive if it is negative.
- The recall is the ration R = tp/(tp + fn) where tp is the number of true positive and fn is the false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
- The F1 score can be interpreted as a weighted harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. F1-score = 2*(R * P)R+P

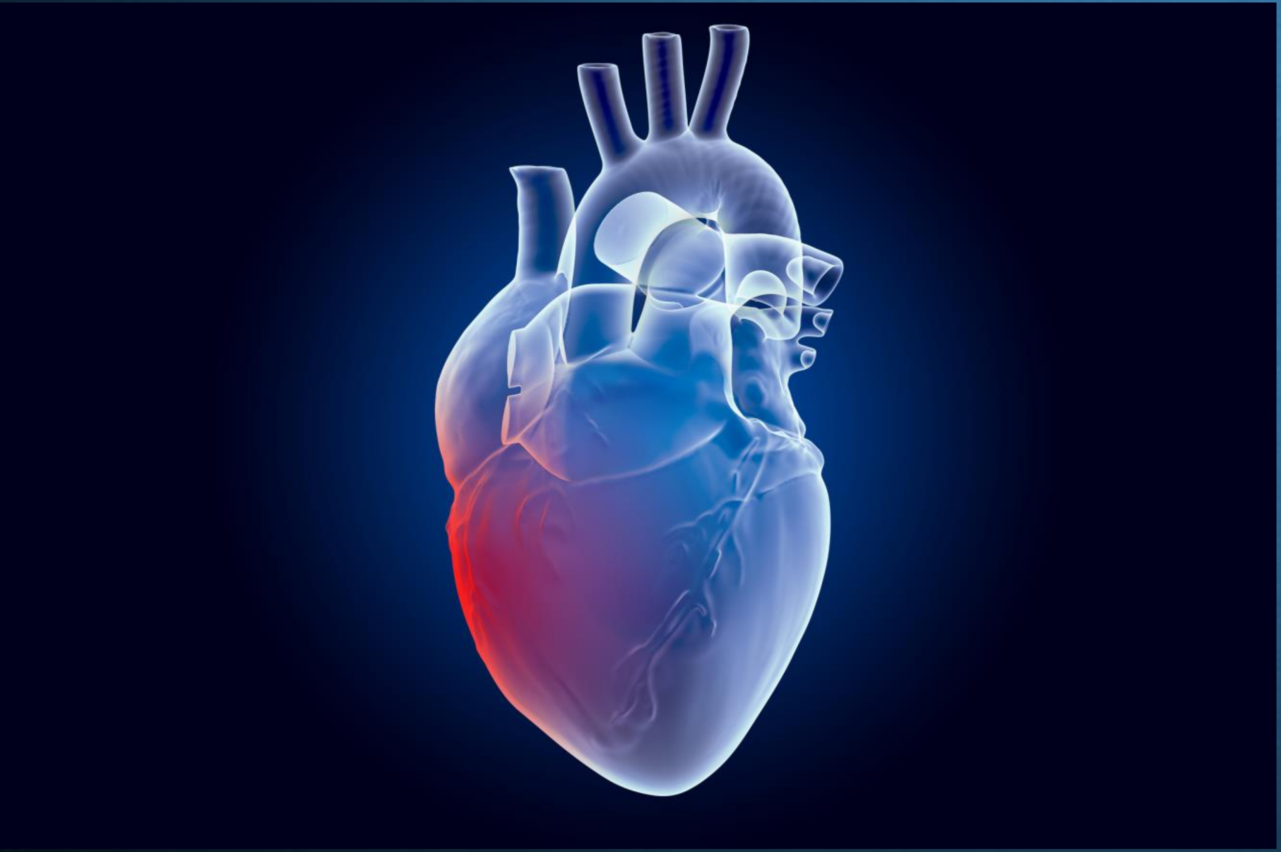
# Recommendation

- Maintain the **Serum Creatinine** level. For adult men, 0.74 to 1.35 mg/dL (65.4 to 119.3 micromoles/L) For adult women, 0.59 to 1.04 mg/dL (52.2 to 91.9 micromoles/L).

- Maintain the **Ejection Fraction** level. For male 52% - 72% and for women 54%-74%. Low ejection fraction is one of the things tends to death.

- Elder peoples should do check-up their health regularly.

- Diabetes and high blood pressure will also cause heart diseases. Maintain the diabetics and high blood pressure.



# THANK YOU..

**Capstone Project** Submitted by:
**PRAKASH S**
**Batch Code-IITPKD ADS Async Jun 2022**

INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD