Heart Failure Prediction

Submited By : Salini U & Sreekutty

Data Overview

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide.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.Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.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.

Goals

1. Find out the count of Patients with Anemia and death Event as per Anaemia
2. Find out the count of Patients who Smoke and death Event as per Smoking
3. What is the percentage of people who died and are smokers
4. What is the percentage of people who died and are not smokers
5. Find out the count of Patients with high blood pressure and death Event as per Blood Pressure
6. Find out the count of Patients with diabetes and death Event as per diabetes
7. Find out the count of Patients with age after 60 and death Event for people with Age after 60
8. Find out the count of Patients according to sex and death Event as per sex
9. Create a model for predicting mortality caused by Heart Failure.

Specifications

1. Age - Age of the patient - Years - [40,…, 95]
2. Anaemia - Decrease of red blood cells or hemoglobin - Boolean - 0, 1
3. High blood pressure - If a patient has hypertension - Boolean - 0, 1
4. Creatinine phosphokinase(CPK) - Level of the CPK enzyme in the blood - mcg/L - [23,…, 7861]
5. Diabetes - If the patient has diabetes – Boolean - 0, 1
6. Ejection fraction - Percentage of blood leaving - Percentage - [14,…, 80]
7. Sex - Woman or man – Binary - 0, 1
8. Platelets - Platelets in the blood – kiloplatelets/mL - 25.01,…, 850.00]
9. Serum creatinine - Level of creatinine in the blood – mg/dL - [0.50,…, 9.40]
10. Serum sodium - Level of sodium in the blood - mEq/L - [114,…, 148]

11. Smoking - If the patient smokes – Boolean - 0, 1
12. Time - Follow-up period – Days - [4,…,285]
13. (target) death event -If the patient died during the follow-up period - Boolean- 0,1

Approach

- Pre processing
  - Attribute selection
  - Cleaning missing values
  - Training and Test data
  - Handling Outlier values

- Processing
  - Processing is applying different algorithms to the data to find the best results

Algorithms used

1. Logistic Regression
2. Random Forest Classifier
3. Support Vector Machine
4. XGBoost
5. Naive Bayes

Results

The data set used for is further splitted into two sets consisting of 80% data as training set and 20% data as testing set. Among the five algorithms applied Logistic Regression shown the best results. The efficiency of the five approaches is compared in terms of the accuracy.The accuracy of the prediction model/classifier is defined as the total number of correctly predicted/classified instances.Accuracy is given by using following formula:

Accuracy= (TP+TN/TP+FN+FP+TN) ∗100

where TP, TN, FN, FP represents the number of true positives, true negative, false negative and false positive cases. we can see the prediction and say that Logistic Regression model is better to perform than the remaining models. and the accuracy score is 80%.