**Heart Failure Prediction Using Machine Learning Techniques with Comparative Study**

Submitted by

**V SURESH KUMAR**

Submitted To

**EICT IITR**

For the fulfillment of

**Applied Data Science Course**

**Index**

|  |  |  |
| --- | --- | --- |
| S.No | Content | Page No |
| 1 | Introduction | 3 |
| 2 | Problem Statement | 6-7 |
| 3 | Methods and Methodology | 8-19 |
| 4 | Resources | 20-21 |
| 5. | Software used | 22-23 |
| 7. | Individual Details | 24 |
| 8. | Milestones | 25 |

**Introduction**

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 worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

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.

According to the National Heart, Lung and Blood Institute:

Heart disease is a catch-all phrase for a variety of conditions that affect the heart’s structure and function. Coronary heart disease is a type of heart disease that develops when the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. **It is the leading cause of death in the United States**.

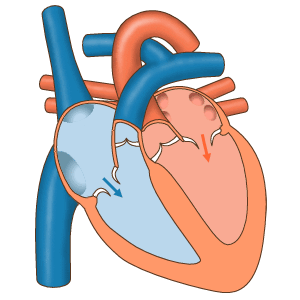
(Emphasis by me. Source: <https://www.nhlbi.nih.gov/health-topics/espanol/enfermedad-coronaria>)

Also, according to the World Health Organization, cardiovascular diseases are the **leading cause of death globally** (source: [https://www.who.int/news-room/fact- sheets/detail/cardiovascular-diseases-(cvds](https://www.who.int/news-room/fact-%20sheets/detail/cardiovascular-diseases-(cvds))).

## Blood and heart

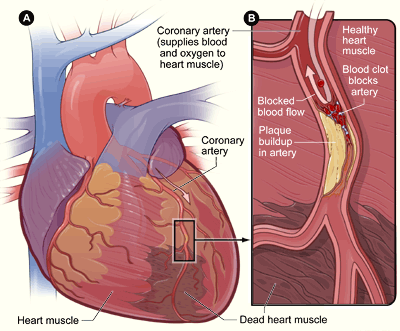
Blood is very important to ensure the proper functioning of the body. Its functions cover the transport of oxygen and nutrients to the cells of the body as well as the removal of the cellular waste products.

Blood is transported to the rest of the body because it is pumped by the heart. This organ receives oxygen-poor blood and sends it to the lungs to oxygenate it. And sends the oxygen-rich blood that comes from the lugns to the rest of the body.



## Coronary heart disease

The heart also needs oxygen and nutrients to function properly, these come through arteries known as coronary arteries. When we talk about a coronary disease, we often mean a difficulty of the blood flow in these arteries due to the accumulation of substances on their walls.



By NIH: National Heart, Lung and Blood Institute - <http://www.nhlbi.nih.gov/health/health-topics/topics/heartattack/Public>Domain, <https://commons.wikimedia.org/w/index.php?curid=25287085>. Death of heart cells due to an ischemia in the coronary arteries.

In the worst case, the impact of leaving the cells of the heart without nutrients and oxygen is a heart attack, in other words, the death of part of the heart cells. This, in turn, would have an impact on the rest of the body because the pumping of the heart would be affected.

**Key facts**

1. Cardiovascular diseases (CVDs) are the leading cause of death globally.
2. 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.
3. Over three quarters of CVD deaths take place in low- and middle-income countries.
4. Out of the 17 million premature deaths (under the age of 70) due to no communicable diseases in 2019, 38% were caused by CVDs.
5. 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.
6. It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin.

**What are cardiovascular diseases?**

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. They include:

1. coronary heart disease – a disease of the blood vessels supplying the heart muscle;
2. cerebrovascular disease – a disease of the blood vessels supplying the brain;
3. peripheral arterial disease – a disease of blood vessels supplying the arms and legs;
4. rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria;
5. congenital heart disease – birth defects that affect the normal development and functioning of the heart caused by malformations of the heart structure from birth; and
6. deep vein thrombosis and pulmonary embolism – blood clots in the leg veins, which can dislodge and move to the heart and lungs.

Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can be caused by bleeding from a blood vessel in the brain or from blood clots.

**Problem statement**

* In the given study, we have a binary classification problem.
* We will make a prection on the target variable **HeartDisease**
* Lastly we will build a variety of Classification models and compare the models giving the best prediction on Heart Disease.

Target Variable:

Target variable, in the machine learning context, is the variable that is or should be the output. For example it could be binary 0 or 1 if you are classifying or it could be a continuous variable if you are doing a regression. In statistics you also refer to it as the response variable.

In our study our target variable is **Heart Disease** in the contex of determining whether anybody is likely to get hearth disease based on the input parameters like gender, age and various test results or not.

|  |  |
| --- | --- |
| Age | Age of the patient  Age > 39 and Age < 96 |
| Anaemia | Anaemia was found in the patient or not  If Anaemia = 1 otherwise =0 |
| Creatinine\_phosphokinase | Creatine phosphokinase (a.k.a., **creatine kinase**, CPK, or CK) is an enzyme (a protein that helps to elicit chemical changes in your body) found in your heart, brain, and skeletal muscles.  Range from 23 to 7861 |
| Diabetes | Diabetes was found in the patient or not  If Diabetes =1 otherwise =0 |
| Ejection\_fraction | A normal LVEF reading for adults over 20 years of age is **53 to 73 percent**. An LVEF of below 53 percent for women and 52 percent for men is considered low. An RVEF of less than 45 percent is considered a potential indicator of heart issues  Range from 14 to 80 |
| High\_blood\_pressure | High blood Pressure was found in the patient or not  If High blood Pressure =1 otherwise =0 |
| Platelets | A normal platelet count ranges from 150,000 to 450,000 platelets per microliter of blood  Range from 25100.to 850000. |
| Serum\_creatinine | The amount of creatinine in your blood should be **relatively stable**. An increased level of creatinine may be a sign of poor kidney function |
| Sex | Male =1 , Female =0 |
| Smoking | Smoking =1 , Non Smoking =0 |
| Time | Range from 4 to 285 mins |
| DEATH\_EVENT | If occurred =1 otherwise =0 |

* Complete analysis of Heart Disease UCI dataset.
* To predict whether a person has a heart disease or not based on the various biological and physical parameters.

**Related Study**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Reference** | **Problem** | **Solution methodology** |
| 1 | Berry, Colin, David R. Murdoch, and John JV McMurray. "Economics of chronic heart failure." *European journal of heart failure* 3.3 (2001): 283-291 | Health-care expenditure on CHF in developed countries represents 1-2% of the total health care  budget. | To Reduce the CHF expenditure of developed countries |
| 2 | Mosterd, Arend, and Arno W. Hoes. "Clinical epidemiology of heart failure." *heart* 93.9 (2007): 1137-1146. | HF is a staggering clinical and public health problem. The  study of the epidemiology of HF demonstrated that although  HF is associated with significant mortality, morbidity, and  healthcare expenditures, particularly among those aged ≥65  years, this burden is not related to an increase in the incidence  of the disease. | Using ML Model to predict accuracy rate to solve this Heart Failure |
| 3 | Kannel, William B., and Albert J. Belanger. "Epidemiology of heart failure." *American heart journal* 121.3 (1991): 951-957 | Clinical Epidemiology of Heart Failure | Using ML Model to predict accuracy rate to solve this Heart Failure |

**Method And Methodology**

**Working flow**



EDA:

In this Exploratory Data Analysis (EDA) and a variety of Model Classifications including Logistic Regression (LR), Support Vector Machine (SVM), AdaBoosting (AB), GradientBoosting (GB), K-Nearest Neighbors (KNN), Random Forest (RF), Desicion Tree (DT), XGBoost (XGB), this study will examine the dataset named as "Heart Failure Prediction" under the 'heart\_failure\_clinical\_records' "csv" file at Kaggle.

This study, in general, will cover what any beginner in Machine Learning can do as much as possible for a better understanding with the given dataset not only by examining its various aspects but also visualising it. Later S/he will be familiar with eight (8) Classification Algorithms in Machine Learning.

Downloading Required Libraries :

* numpy
* pandas
* matplotlib
* seaborn
* plotly
* scipy
* pyforest
* sklearn
* cufflinks
* plotly

**Important steps taken before model creation**

* Defining User Defined Functions
* General visualization of data
* Examination of Numerical features
* Examination of skewness
* Dummy variable operation
* Train/Test Split
* Feature scaling

Feature scaling (Normalization) is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

For machine learning, in general, it is necessary to normalize features so that no features are arbitrarily large (centering) and all features are on the same scale (scaling).

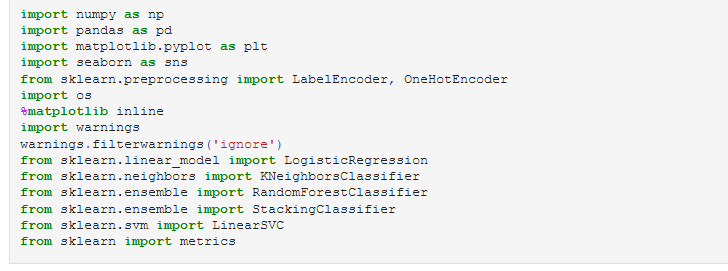
In general, algorithms that exploit distances or similarities (e.g. in the form of scalar product) between data samples, such as K-NN and SVM, are sensitive to feature transformations. So it is generally useful, when you are solving a system of equations, least squares, etc, where we have serious issues due to rounding errors.

However, Graphical-model based classifiers, such as Fisher LDA or Naive Bayes, as well as Decision trees and Tree-based ensemble methods (RF, XGB) are invariant to feature scaling, but still, it might be a good idea to rescale/standardize the data.

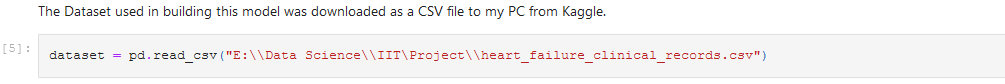
**Implementation Steps**

* Implementation of logistic regression
* Implementation of KNN
* Analysis of scaling factors of logistic regression and knn
* Checking accuracy scores
* Modeling logistic regression with best parameters using GridSearchCV
* ROC and AUC Curve
* Determination of the Optimal Threshold
* Modeling K-Nearest Neighbour with default parameters
* Cross- Validating KNN
* Elbow method for choosing reasonable K values
* GridsearchCV for choosing reasonable K values
* ROC and AUC curve
* Comparative study of Logistic Regression and KNN

Step 1: Import Libraries

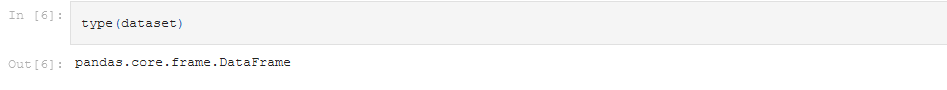


Step 2: Import the Dataset



Step 3: Data Cleaning and EDA

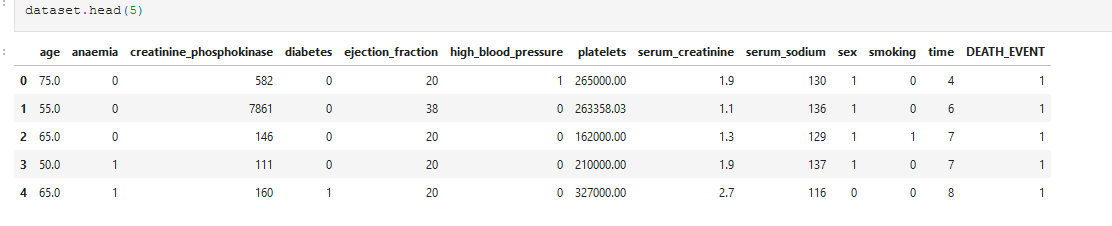
* type(dataset)



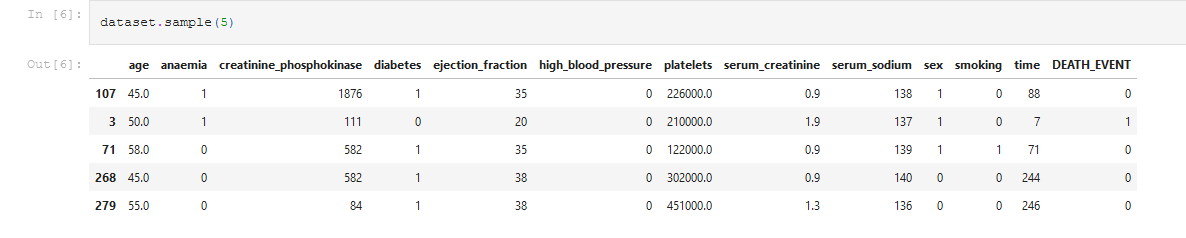
* dataset.shape



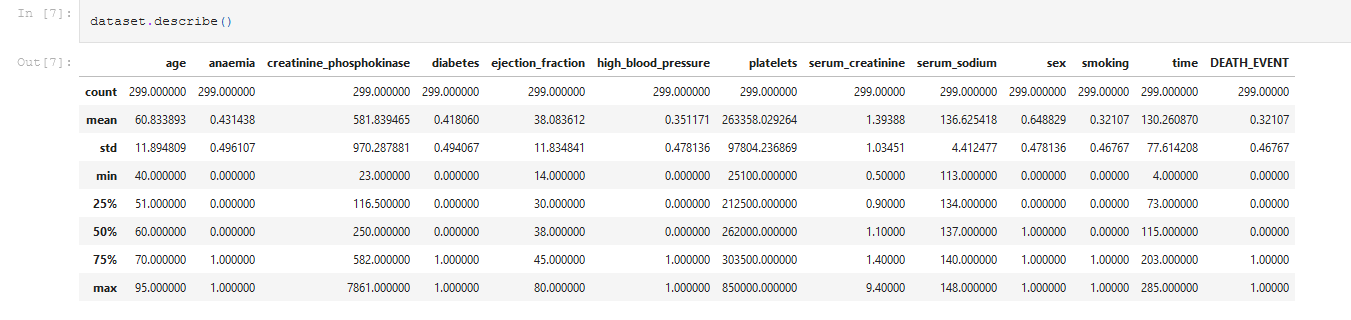
* dataset.head(5)



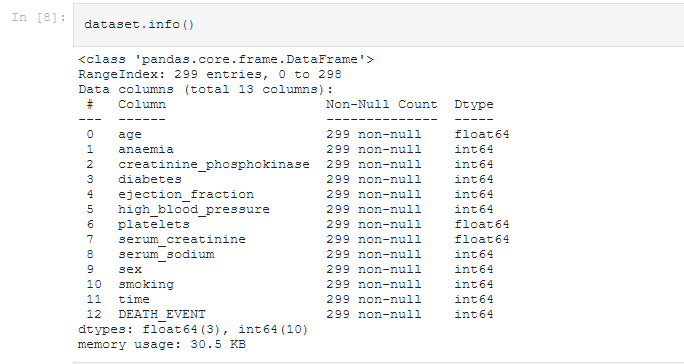
* dataset.sample(5)



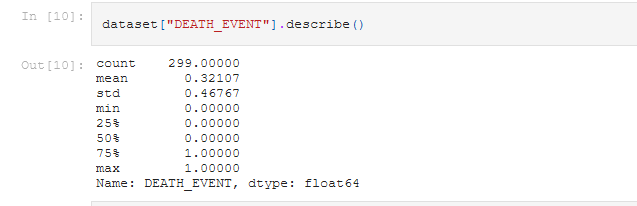
* dataset.describe()



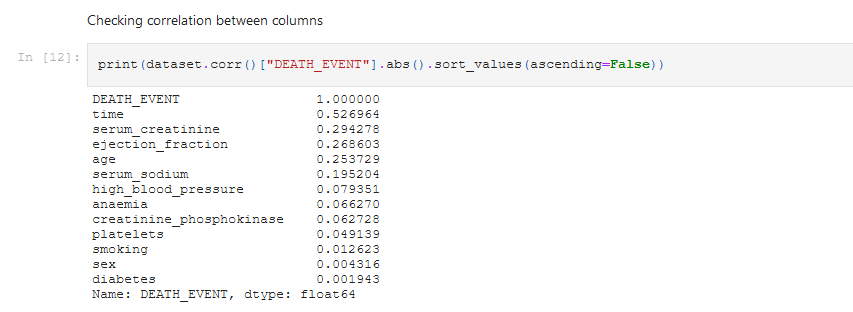
* dataset.info()



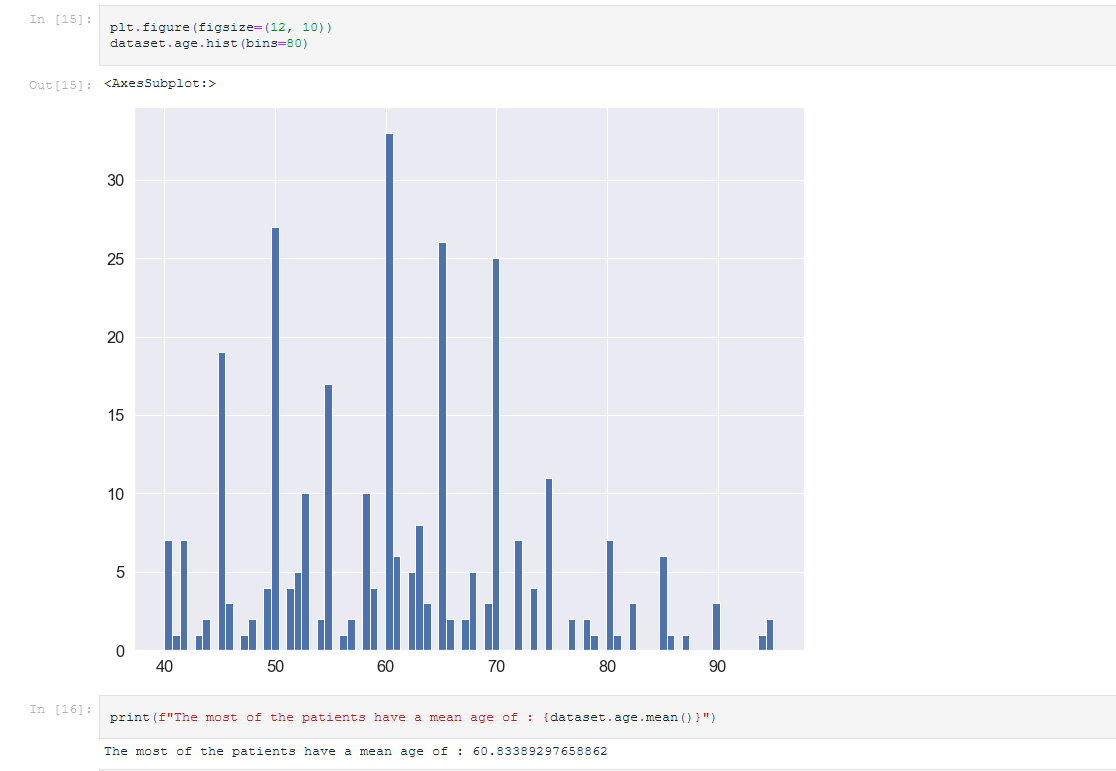
* Analysing the 'target' variable

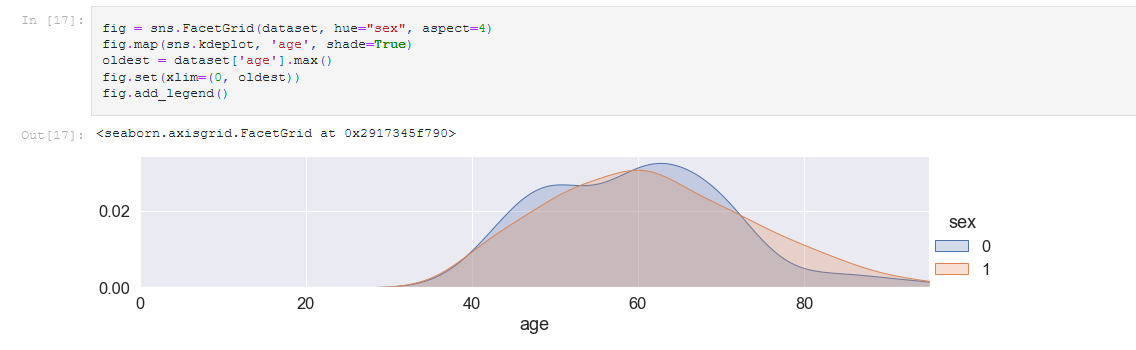


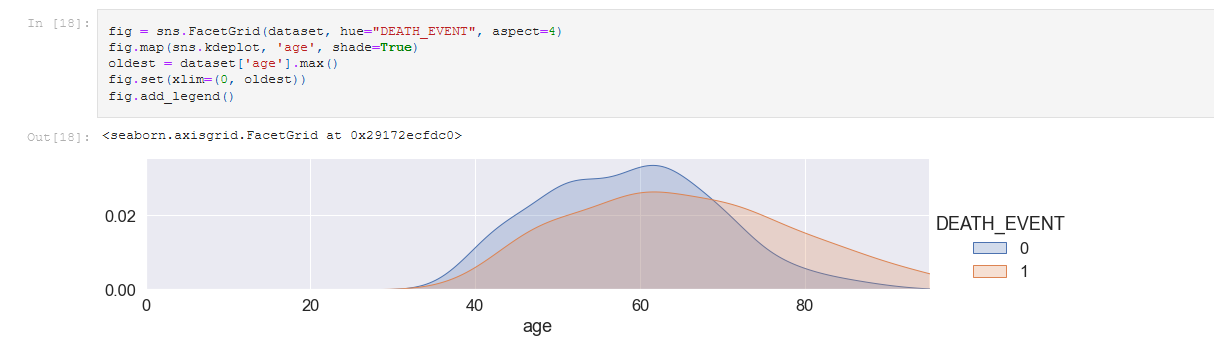
* Checking correlation between columns

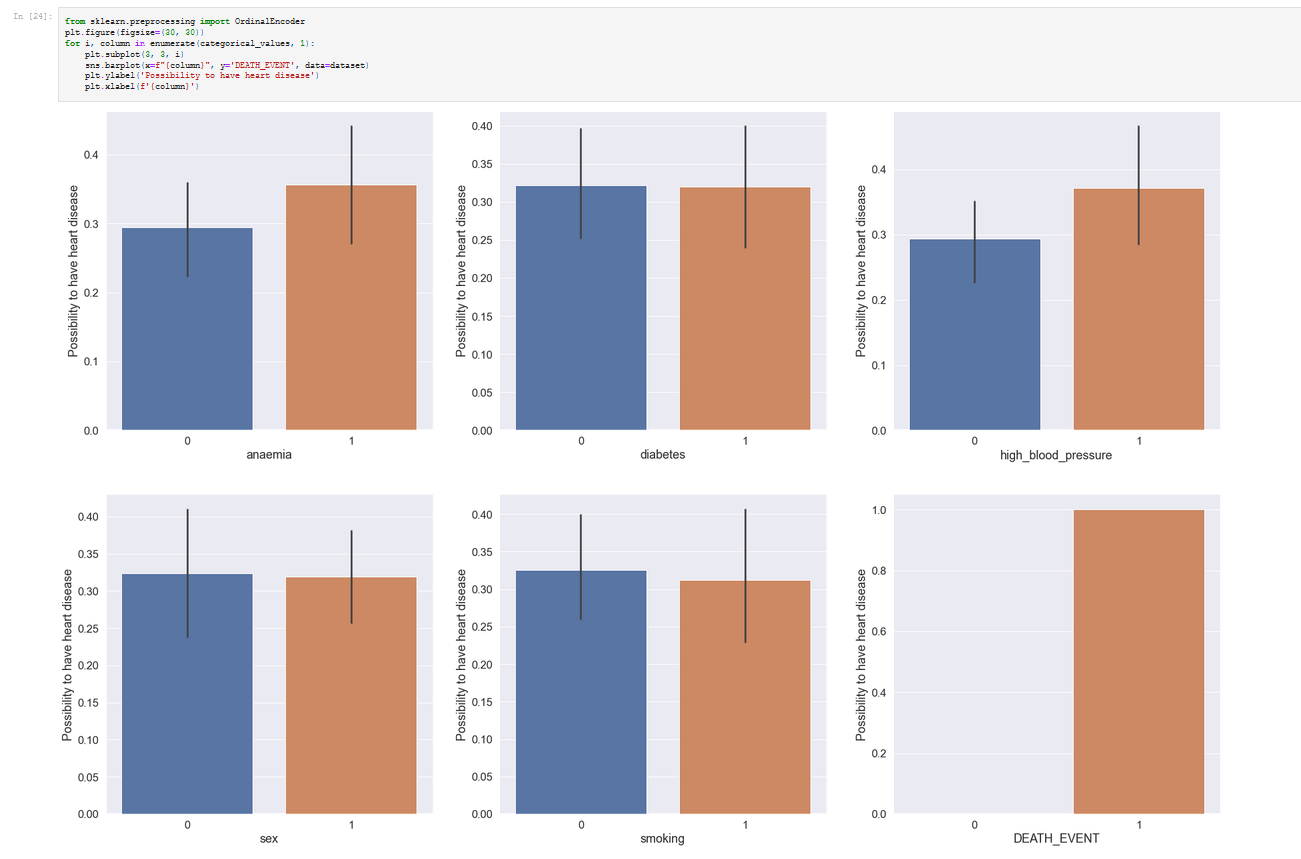


* Data Visualization

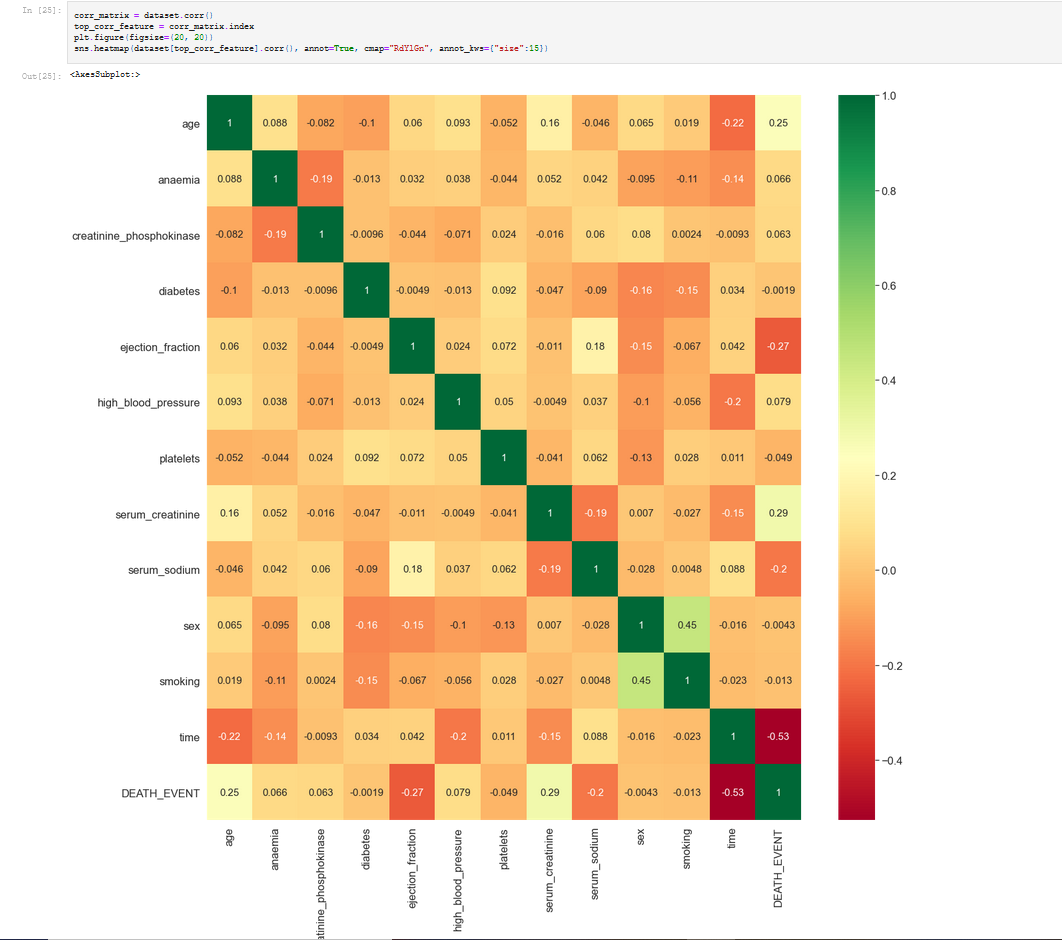




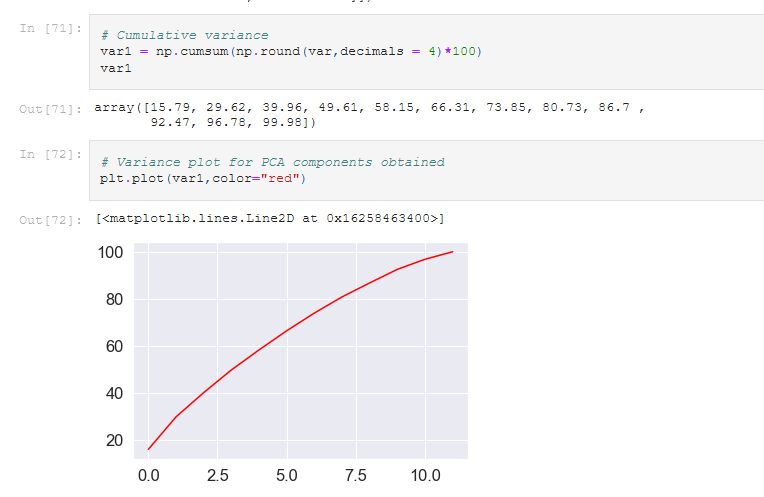




* Correlation Matrix



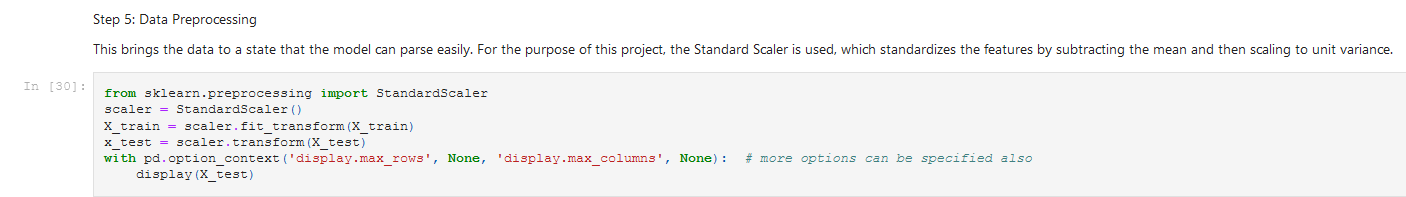
* Data Processing Using PCA

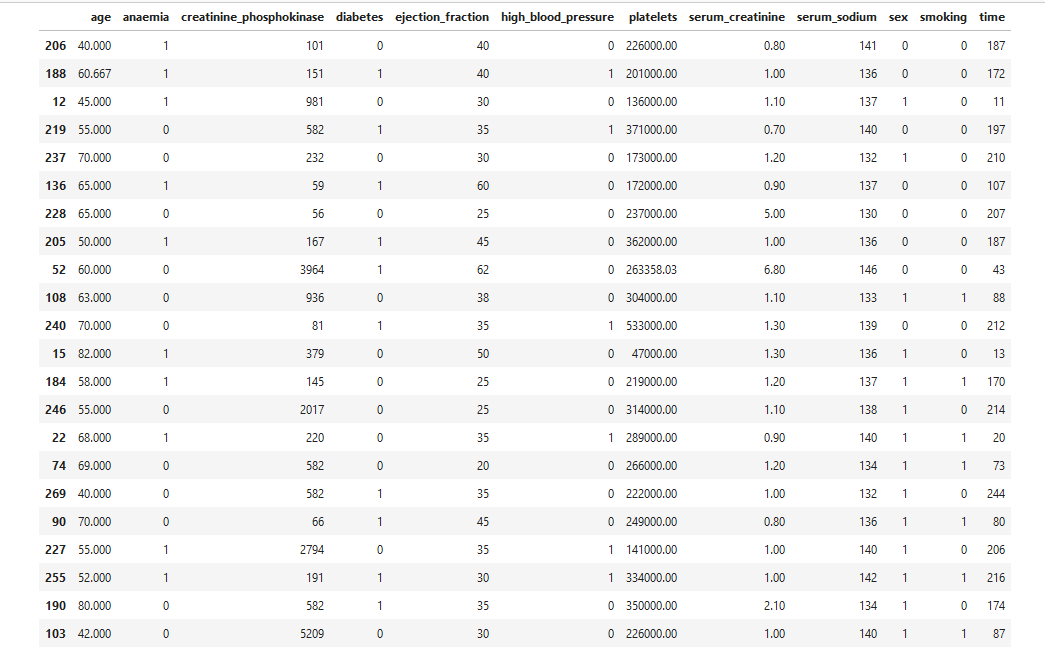


Step 4: Splitting the Train and Test Data



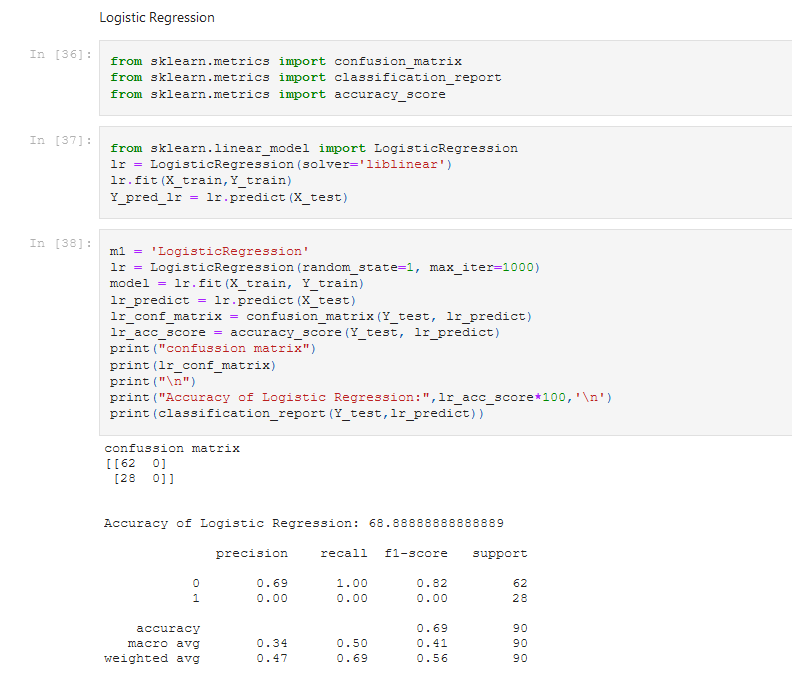
Step 5: Data Preprocessing



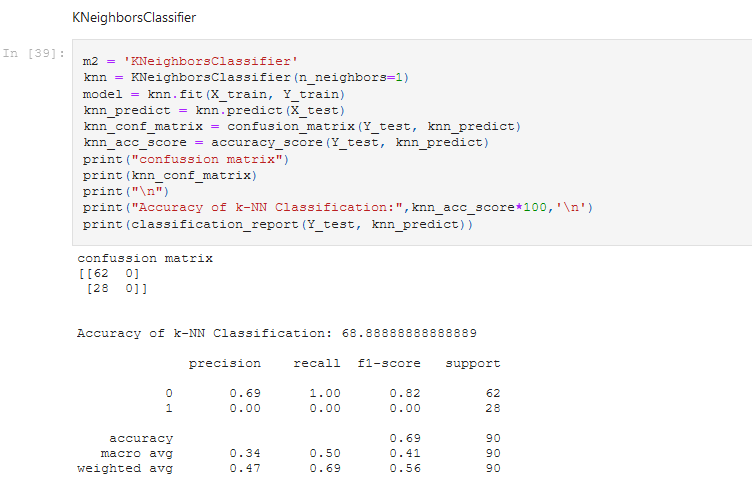


Step 6: Model Selection

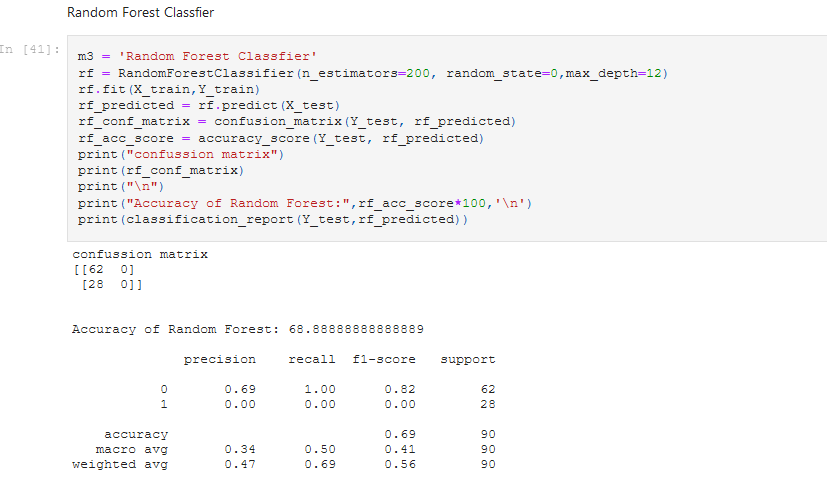
* Logistic Regression



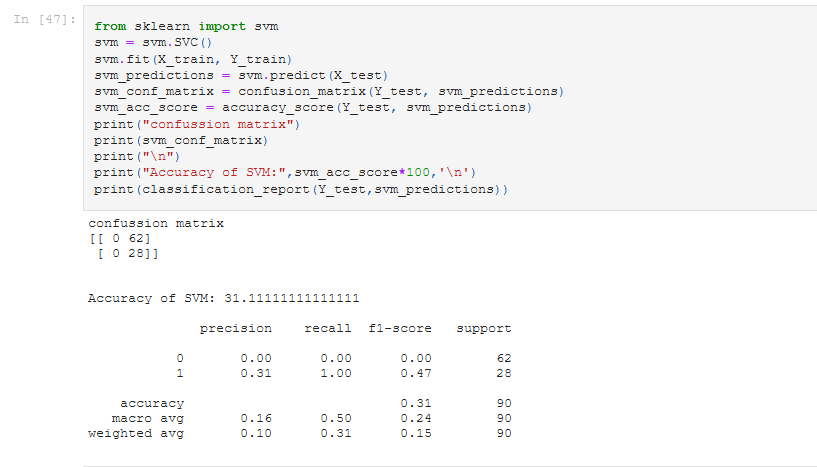
* KNeighborsClassifier



* Random Forest Classfier

****

* SVM



Output Final :



1. **Resources**

## Collecting Data :

The data used for training and testing is the [Heart Disease UCI](https://www.kaggle.com/ronitf/heart-disease-uci) downloaded from Kaggle.  
This database contains 76 attributes, but all published experiments refer to using a subset of 12 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient.



Data Set Dimensions

|  |  |  |
| --- | --- | --- |
| S.No | Attribute | Value |
| 1 | Data Set | Heart Failure Clinical Records |
| 2 | Repository | Heart Disease UCI |
| 3. | No. of Features | 13 |
| 4. | Shape of data set | 299 rows, 13 Columns |
| 5. | No. of Elements | 3887 |
| 6 | Dimensions | 2 |

Statistical Summary of age of a person and Death\_Events

|  |  |  |  |
| --- | --- | --- | --- |
| SNo. | Attribute | Age | Death\_Event |
| 1 | Count | 299 | 95 |
| 2 | Mean | 60.83 | 0 |
| 3 | Std | 11.89 | 0 |
| 4. | Min | 40 | 0 |
| 5 | 25% | 51 | 0 |
| 6 | 50% | 60 | 0 |
| 7 | 75% | 70 | 1 |
| 8 | Max | 95 | 1 |

1. **Software Used**

1.Python 3.8.8

2. Jupyter Notebook

3Import the following Libraries

* importnumpyasnp
* importpandasaspd
* importmatplotlib.pyplotasplt
* importseabornassns
* fromsklearn.preprocessingimportLabelEncoder,OneHotEncoder
* importos
* %matplotlib inline
* importwarnings
* warnings.filterwarnings('ignore')
* fromsklearn.linear\_modelimportLogisticRegression
* fromsklearn.neighborsimportKNeighborsClassifier
* fromsklearn.ensembleimportRandomForestClassifier
* fromsklearn.svmimportLinearSVC
* fromsklearnimportmetrics
* fromsklearn.utilsimportshuffle
* fromsklearn.model\_selectionimporttrain\_test\_split
* fromsklearn.preprocessingimportStandardScaler
* fromsklearn.metricsimportconfusion\_matrix
* fromsklearn.metricsimportclassification\_report
* fromsklearn.metricsimportaccuracy\_score
* fromsklearn.decompositionimportPCA

**Conclusion**

* In this study respectively,
* We have tried to a predict classification problem in Heart Disease Dataset by a variety of models to classifiyHeart Disease predictions in the contex of determining whether anybody is likely to get hearth disease based on the input parameters like gender, age and various test results or not.
* We have made the detailed exploratory analysis (EDA).
* There have been NO missing values in the Dataset.
* We have decided which metrics will be used.
* We have analyzed both target and features in detail.
* We have transformed categorical variables into dummies so we can use them in the models.
* We have handled with skewness problem for make them closer to normal distribution; however, having examined the results, it's clear to assume that handling with skewness could NOT make any contribution to our models when comparing the results obtained by Logistic Classifier without using Power Transform. Therefore, in this study we have continue not handling with skewness assuming that it's useless for the results.
* We have cross-checked the models obtained from train sets by applying cross validation for each model performance.
* We have examined the feature importance of some models.
* Lastly we have examined the results of all models visually with respect to select the best one for the problem in hand.
* Any contribution will be appreciated.
* By the way, if you enjoy reading this analysis, you can show it by supporting

**Individual Details**

**Name: V SURESH KUMAR**

**E-mail Id: sureshtrichy3@gmail.com**

**Phone Number: 94436-59512**

1. **Milestones**

|  |  |
| --- | --- |
| **Milestones** |  |
| Define a problem | Heart Failure Prediction |
| Understanding the business problem | To use the data set to predict the heart failure |
| Get the Data | We got the data from kaggle  heart\_failure\_clinical\_records.csv |
| Explore and pre-process data | Using Visualization toExplore and pre-process data |
| Choosing the python platform | Python 3.8.8 |
| Create Features | In the pre-processing PCA use to determine the Score rates.  Using the Logistic,KNN, Random Forest and SVM Model to predicted the accuracy rates |
| EDA | Using info () to determine the non-null values are appearing in the data set. |
| Create Model | Logistic Regression  k-NN Classification  Random Forest  SVM |
| Model Evaluation | Logistic Regression Accuracy:68.89 %  k-NN Classification Accuracy :68.89 %  Random Forest Accuracy :68.89 %  SVM Accuracy :31.11% |
| Report Writing | Logistic Regression Accuracy : 68.89 |
| Project submission | Heart failure accuracy rate is : 68.89% base on the data set. |

**References**

**[1]** Kannel, William B., and Albert J. Belanger. "Epidemiology of heart failure." *American heart journal* 121.3 (1991): 951-957.

[2] Mosterd, Arend, and Arno W. Hoes. "Clinical epidemiology of heart failure." *heart* 93.9 (2007): 1137-1146.

[3] Berry, Colin, David R. Murdoch, and John JV McMurray. "Economics of chronic heart failure." *European journal of heart failure* 3.3 (2001): 283-291.