**Curneu Med Tech Assignment :**

**Name :** Sahana S R

**Reg. No. :**1832043

**Dept:** MSc. Data Science 3rd year

**Heart Failure Prediction**

**Aim :**

Predict Heart Failure, create a model to assess the likelihood of a death by heart failure event from the given data

**Given description about the dataset:**

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.

**Exploratory data analysis:**

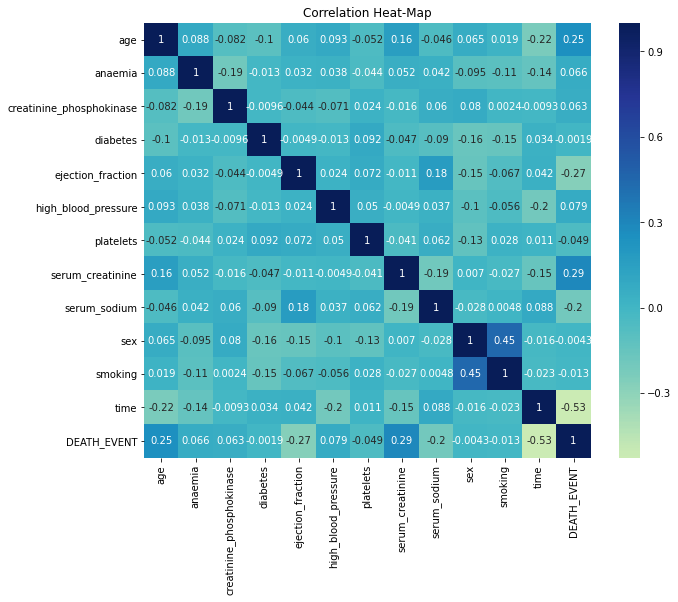
**Checking Null values:**

There are no null values in the data

**Graphs:**



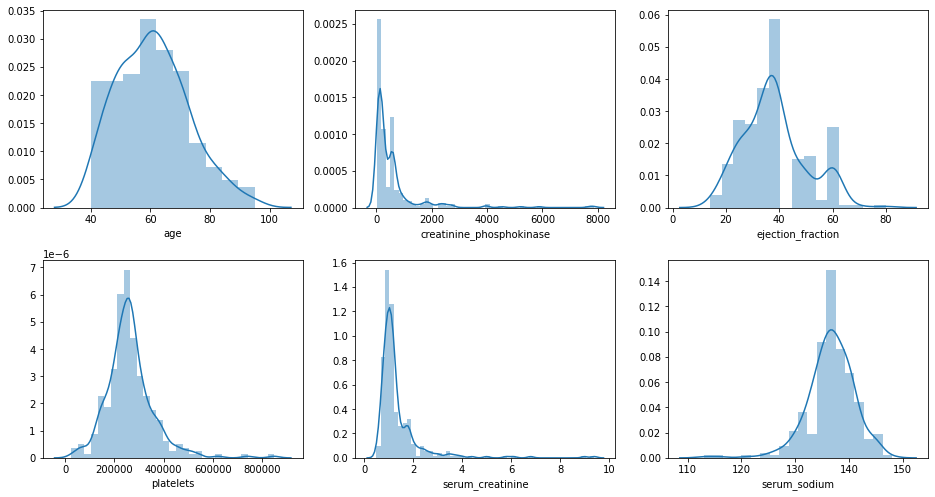
It can be seen from the above graph that the dataset is imbalanced



The above correlation plot does not show any major amount of inter-collinearity in the data



The above plot shows the distribution of people under different catagories

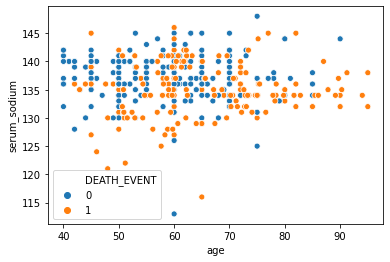
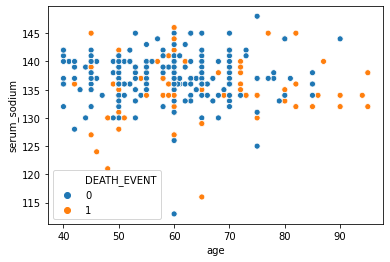


The above plot shows the histogram for each numeric category to understand their distribution and scale better

**Balancing the data:**

We use the SMOTE algorithm to balance the data. Smote stands for Synthetic Minority Oversampling Technique. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

The dataset before and after smote is implemented is shown below



From the graph below we can see that the data is now balanced



**Train Test Split**

Of the given data 80% of the data goes to training and 20% to test set

**Scaling:**

Standard scaling technique is used to scale the data

**Model fitting:**

We fit the following models to find the best one out of it by performing hyper parameter tuning on each model. All the parameters of each model are passed into the GridSearch algorithm of python inorder to find the best parameters that produce the outcome

* Random Forest Classifier
* Logistic regression classifier
* Decision tree classifier
* KNN classifier
* Stochastic Gradient Descent Classifier

After fitting all the models we got the following output and hence random forest is a best possible fit for our dataset

