**Data visualization and Heart Failure prediction using machine learning algorithms and python**

**EED363 – APPLIED MACHINE LEARNING**

**(Spring 2021)**

*Project report submitted in partial fulfillment of the requirement for the degree of*

Bachelor of Technology

In

Electronics and Communication Engineering/Electrical and Electronics Engineering

**Submitted by**

Datla Gautam Varma (171011099)

Under supervision of

Prof Madan Gopal

Electrical Engineering

**Department of Electrical Engineering**

**School of Engineering**

SNU

**Abstract:**

Among various life-threatening diseases, heart disease has garnered a great deal of attention in medical research. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise. Now the question we tend to face is if there is a method using which we can predict such a life-threatening disease based on a few existing symptoms and signs. The answer is yes! Data mining and machine learning have made it possible. Machine learning and artificial intelligence are playing a pivotal role in the present-day scenario. It has attained such importance mainly because of its wide range of applications and its incredible ability to adapt and provide solutions to complex problems efficiently, effectively and quickly.

So, in this project first we try to explore the dataset and then using the existing machine learning algorithms such as K-NN, Naïve bayes and decision tree algorithms on the data set we try to figure out which algorithm would turn out to be the best fit for the model giving the least possible errors while predicting. The coding for this project has been done using python3 in PyCharm IDE.

**1) Introduction**

A heart attack is a medical emergency that usually occurs when a blood clot blocks blood flow to the heart. During a heart attack, the blood supply that normally nourishes the heart with oxygen is cut off and the heart muscle begins to die. This blockage is most often a build-up of fat, cholesterol and other substances, which form a plaque in the arteries that feed the heart (coronary arteries). Now for such a lethal disease, Patients' signs, body characteristics, and clinical laboratory test values can all be quantified in electronic medical records, which can then be used to conduct biostatistics analysis to uncover trends and associations that would otherwise go undetected by physicians. Machine learning, in particular, can use data to predict patients' survival and identify the most important features among them.

**Problem statement**

Despite the fact that we are familiar with a large number of algorithms, no consensus has emerged to direct the selection of new algorithms for clinical use in the field of cardiovascular medicine. Although it is possible to select optimal algorithms for research questions and to replicate algorithms in various clinical datasets, clinical understanding and judgement for algorithm implementation are extremely difficult. Now, this is what motivated me to take up this project. In this project I try to build a prediction model that would suit the best by testing out various classification algorithms on data set and choosing the best fit that would give us minimal error.

**2) Literature Review:**

**A) Heart Failure prediction and visualization data set:**

The data set I have used for this project was taken from Kaggle. This particular data set has 300 samples with 12 features for each sample. These 12 features for each sample would basically be the risk factors for a heart failure, using which we will be building our model. The features that are considered in this data set are:

**i) Age:**

Age is an important factor which can be used to predict if a person is susceptible to heart stroke or not. Adults whose age is 65 and above are more likely than younger people to suffer from cardiovascular disease, which is problems with the heart, blood vessels, or both. Aging can cause changes in the heart and blood vessels that may increase a person's risk of developing cardiovascular disease.

**ii) Anemia:**

Anemia is found in about one-third of all cases of congestive heart failure (CHF). Anemia in heart failure is considered to develop due to a complex interaction of iron deficiency, kidney disease, and cytokine production, although micronutrient insufficiency and blood loss may contribute. Thus, it could prove to be another important factor which we are to consider while predicting heart failure.

**iii) Creatine phosphokinase:**

Creatine kinase or creatine phosphokinase is an enzyme chiefly found in the brain, skeletal muscles, and heart. An elevated level of creatine kinase is seen in heart attacks, when the heart muscle is damaged, or in conditions that produce damage to the skeletal muscles or brain.

**iv) Diabetes:**

Diabetes and heart failure are closely related: patients with diabetes have an increased risk of developing heart failure and those with heart failure are at higher risk of developing diabetes. Furthermore, antidiabetic medications increase the risk of mortality and hospitalization for heart failure in patients with and without pre-existing heart failure.

**v) Ejection Fraction:**

This refers to how well your left ventricle (or right ventricle) pumps blood with each heartbeat. Most times, EF refers to the amount of blood being pumped out of the left ventricle each time it contracts. An EF that is below normal can be a sign of heart failure. If you have heart failure and a lower-than-normal (reduced) EF (HF-rEF), your EF helps your doctor know how severe your condition is.

**vi) High Blood Pressure:**

The extra strain and damage caused by high blood pressure (HBP or hypertension) causes the coronary arteries that serve the heart to narrow over time due to an accumulation of fat, cholesterol, and other substances known as plaque. The flow of blood through the heart muscle is disrupted when an artery becomes blocked due to plaque buildup or a blood clot, thus causing a heart attack.

**vii) Serum Creatinine:**

A significant subset of patients with heart failure (HF) experience small to moderate rise in serum creatinine (RSC) in the setting of otherwise beneficial therapies such as aggressive diuresis or renin-angiotensin-aldosterone system (RAAS) inhibition.

**viii) Serum Sodium:**

Hyponatremia or low serum sodium level is typically defined as a serum sodium concentration of <135 mEq/L and is one of the most common biochemical disorders featured in heart failure patients, with a prevalence close to 25% [2–4]. HF affects cardiac output by either decreasing heart rate or reducing the stroke volume.

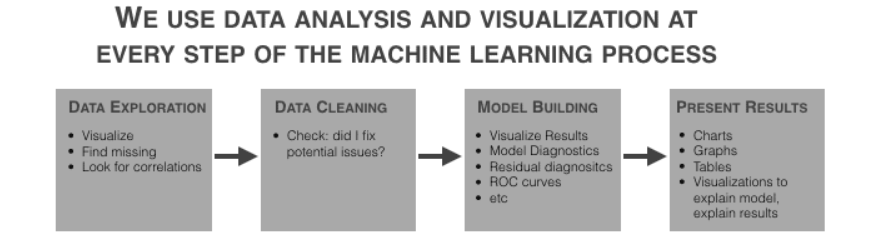
**ix) Smoking:**

Smoking causes the blood vessels in your body, including those in your heart, to constrict (narrow). This will not only make the symptoms of your heart failure worse but also may cause a heart failure.

**X) Platelets:**

An excessive increase in the platelet count can lead to certain conditions, including stroke, heart attack or a clot in the blood vessels thus can be an indicative factor of heart failure.

**B) EDA (Exploratory data analysis):**



It's a method of visualizing, summarizing, and analyzing data that's hidden behind rows and columns. EDA is a critical step in data science because it helps us to gain specific information and statistical measurements that are critical for business continuity, stockholders, and data scientists. It performs to define and refine our important features variable selection, that will be used in our model.

In EDA for our dataset, we will first identify various parameters for all the features such as their probability distribution, mean, variance and their correlation with the output. Also, we will be identifying and removing the outliners in each feature using boxplot and also identify and remove null values from the data set if any.

C) **Extra Tree Classifier for Feature Selection:**

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is given a random sample of k features from the feature collection, from which it must choose the best feature to divide the data based on certain mathematical parameters (typically the Gini Index).

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature.

To perform feature selection, each feature is sorted in descending order by Gini Importance, and the user chooses the top k features.

**D) Python Libraries:**

For visualization and EDA, I have used various python libraries such as:

**i) Pandas:** Powerful data structure for data analysis, time series, and statistics.

**ii) NumPy:** NumPy is the fundamental package for array computation with python.

**iii) Matplotlib:** Python plotting package

**iv) Seaborn:** Statistical data visualization

**v) Plotly:** An open-source, interactive data visualization library for Python.

**vi) Sklearn: Python** modules for machine learning and data mining.

**E) Testing and Training Data:**

In statistics and machine learning we usually split our data into two subsets: training data and testing data (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. When we do that, one of two things might happen: we overfit our model or we underfit our model. We don’t want any of these things to happen, because they affect the predictability of our model.

In this project, From **Sklearn, sub-library mode selection,** I have used the test\_train\_split to split the dataset into testing and training data.

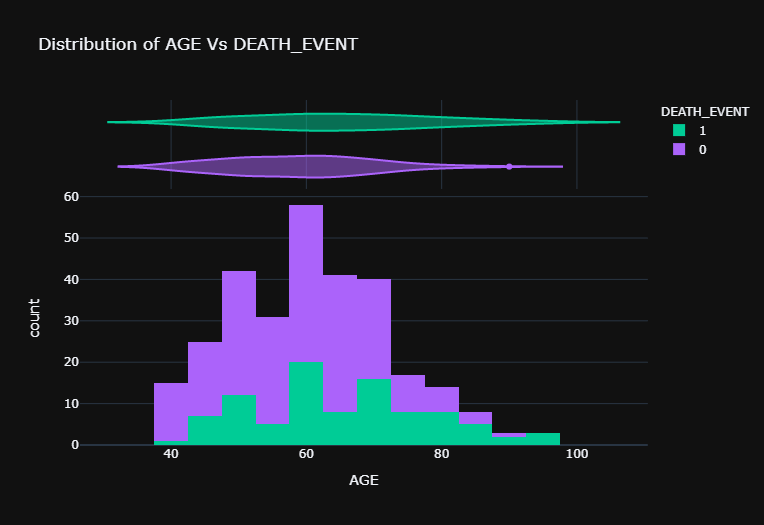
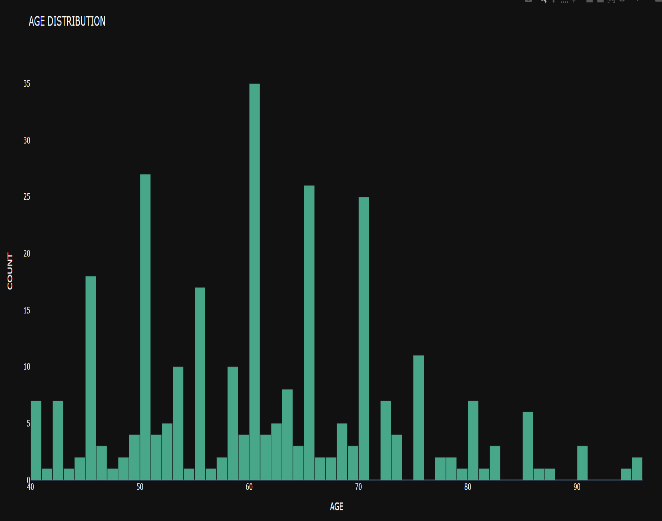
**F) Logistic regression:**

Logistic regression is basically a supervised classification algorithm. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

**3) Work Done:**

To begin with I first start data analysis of the data set. As mentioned earlier, the data set has 12 features with which are to deal with. EDA of the feature set is as follows,

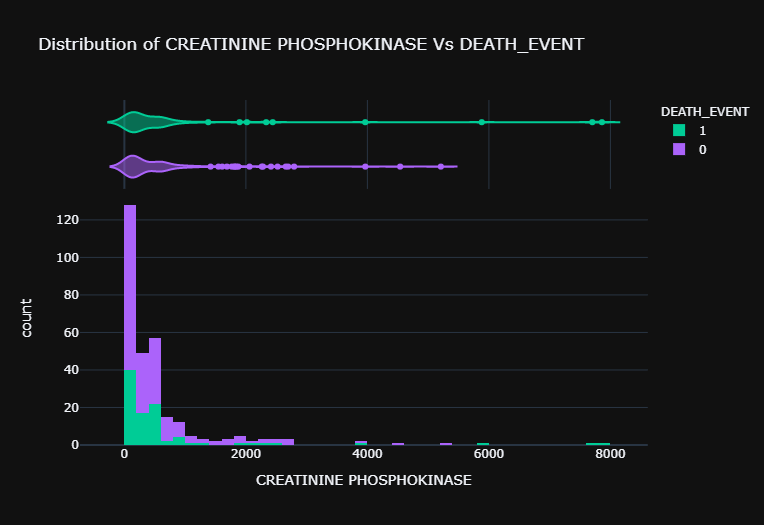
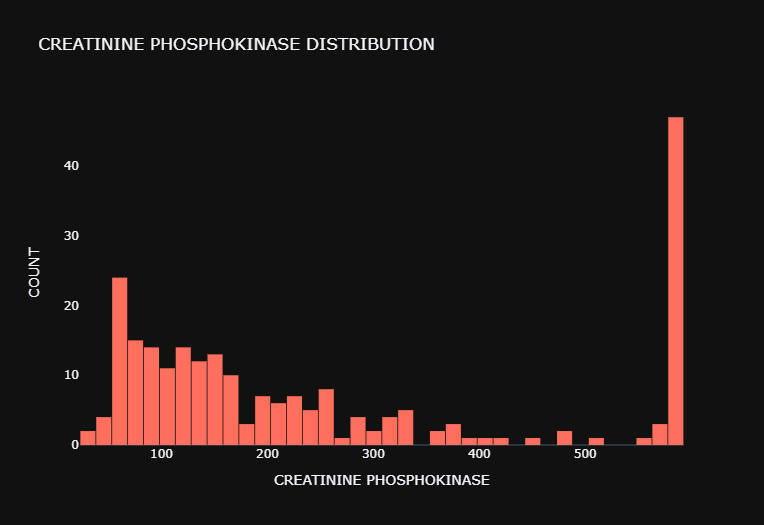
**1) AGE**



**Fig1. Age distribution**

The Average was 60.8399 with the maximum being 95 and minimum being 45. Standard deviation is 11.894. In the second graph we can see the relationship between the death event and age, here Wider sections of the violin plot represent a higher probability of observations taking a given value, the thinner sections correspond to a lower probability and the value of probability is given by pde value for given x.

**2) CREATININE PHOSPHOKINASE**

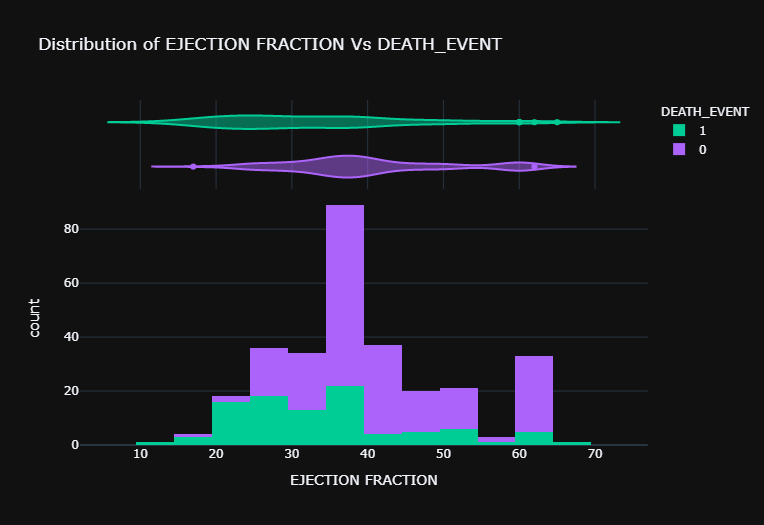
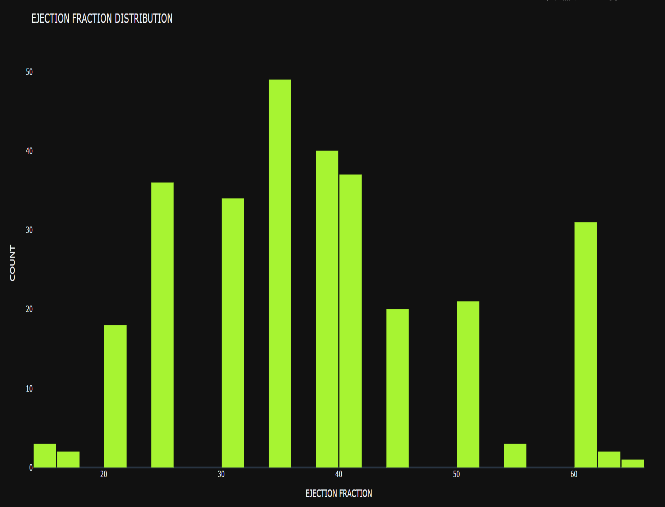


**Fig2. Creatinine Phosphokinase distribution**

The average and standard deviation of creatinine phosphokinase are 581.8394 and 0.4963 respectively.

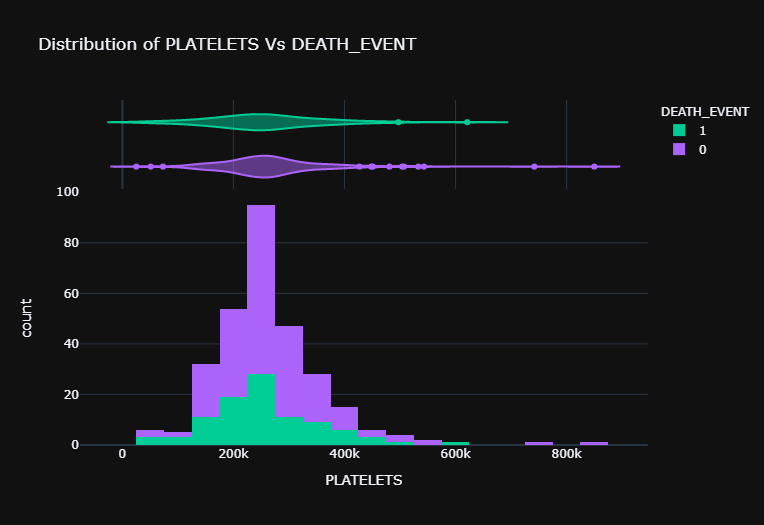
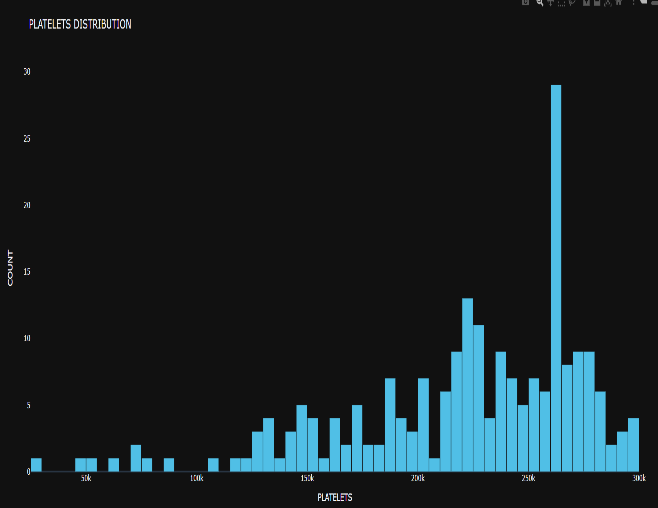
And also, we can notice that the highest probability of the death event at the lower values of creatinine phosphokinase.

**3) EJECTION FRACTION:**

**Fig3. Ejection fraction distribution**

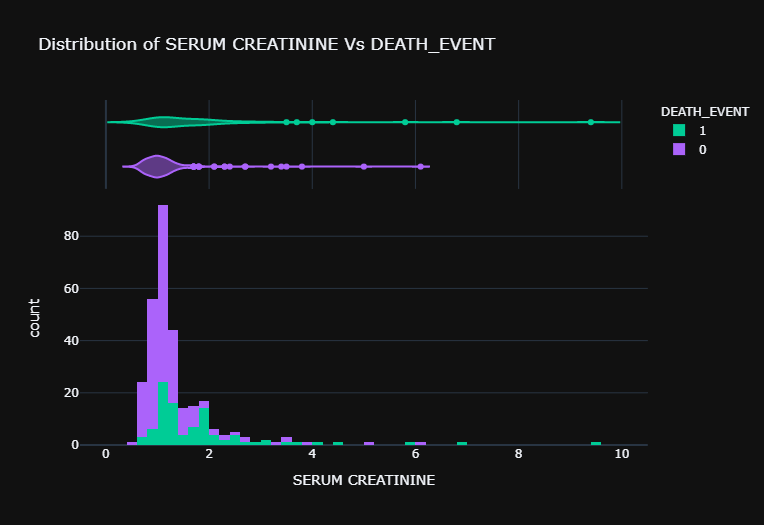
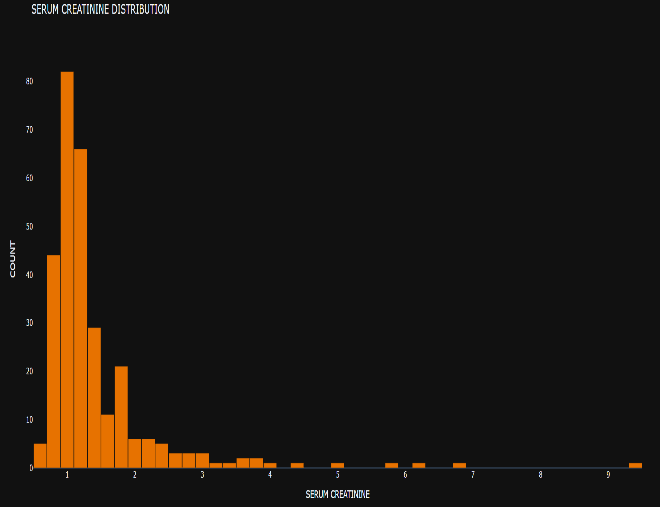
The mean, median, mode of this feature is 75.9133, 38 and 35 respectively. While the standard deviation of this feature is 11.8348

**4) PLATELET DISTRIBUTION:**

**Fig4. Platelet distribution**

The mean median mode and standard deviation of platelet distribution are 263358.092, 262000, 263358.03 and 97804.24 respectively.

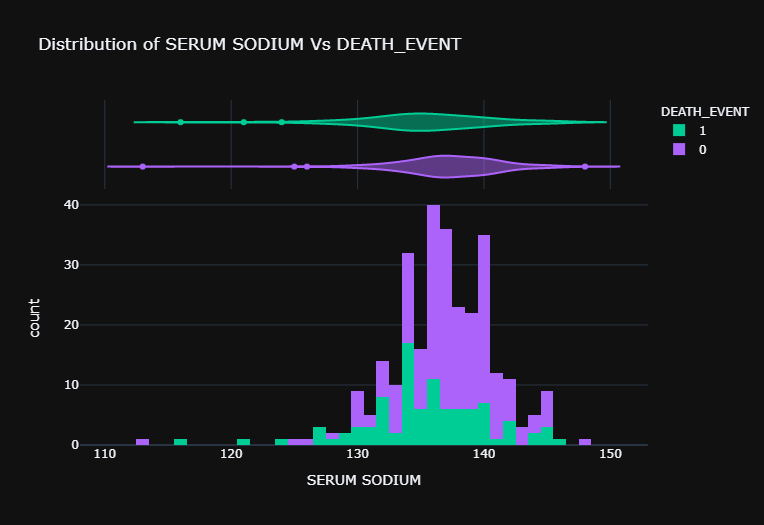
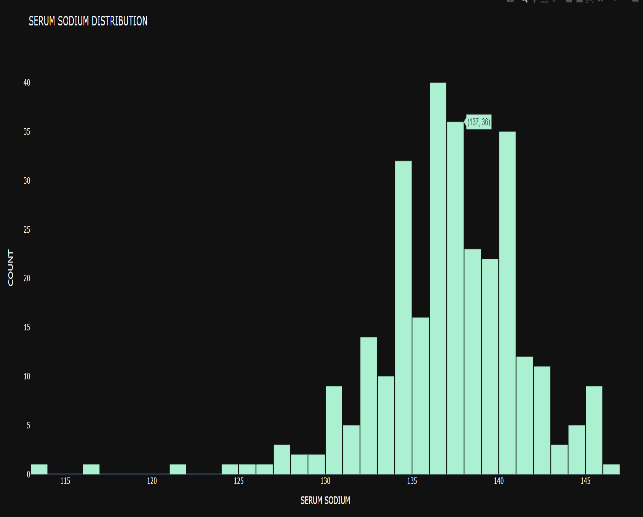
**5) SERUM CREATININE:**



**Fig5.Serum Creatinine distribution**

For serum creatinine the mean median mode and standard deviation are 1.393, 1.1, 1 and 1.0345 respectively.

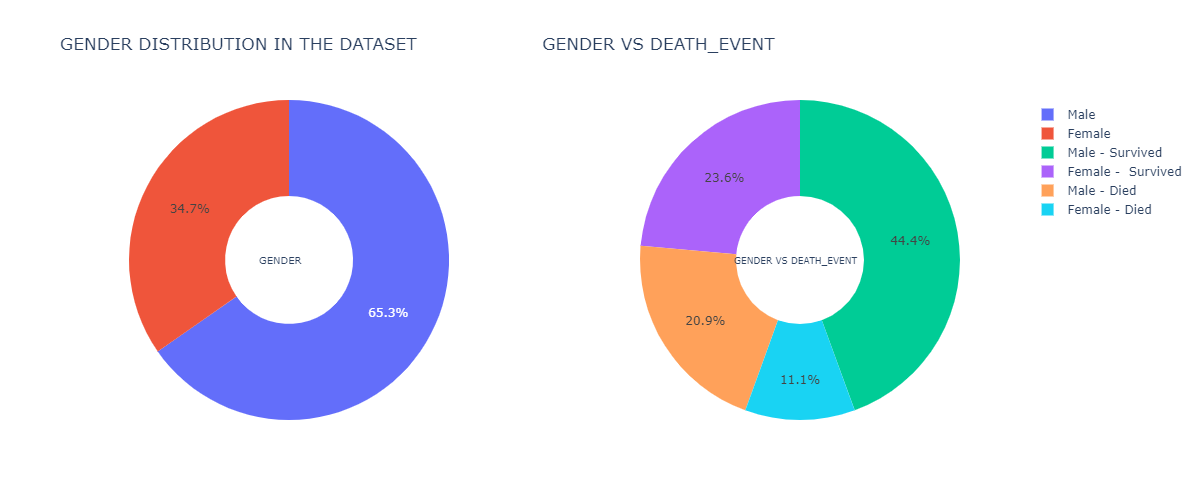
**6) SERUM SODIUM**



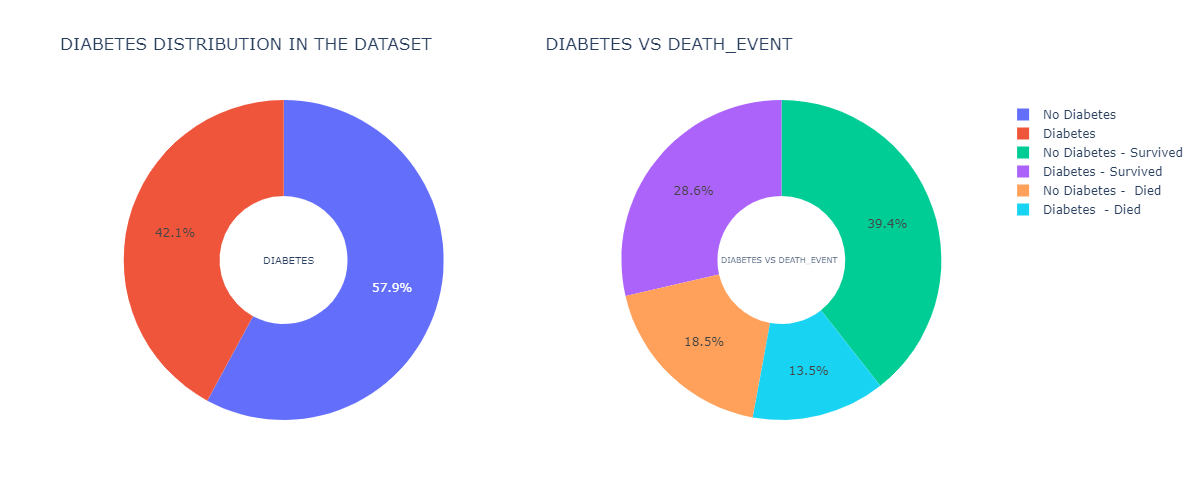
**Fig6. Serum sodium distribution**

For serum sodium the mean median mode and standard deviation are 136.625, 137.1, 136 and 4.4124 respectively.

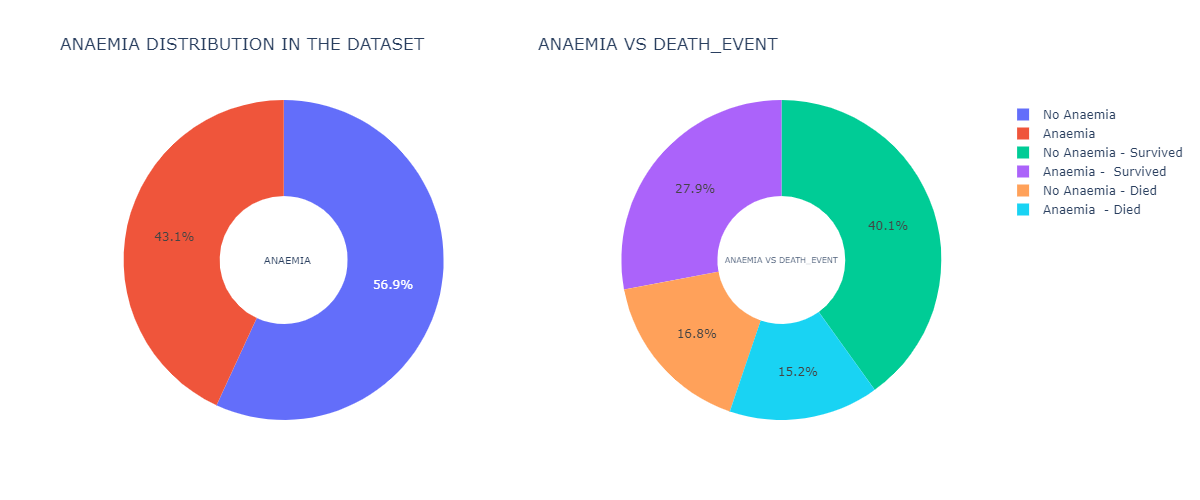
**7) GENDER, DIABETES and ANEMIA**



**Fig8. Gender distribution**

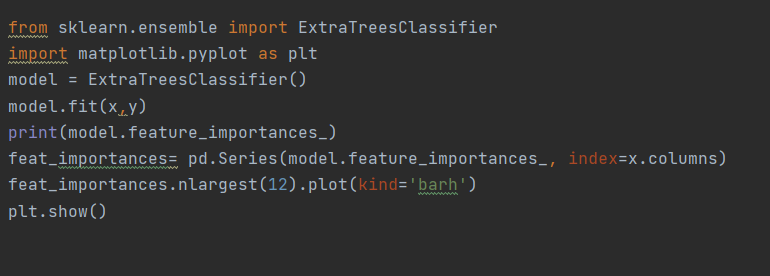


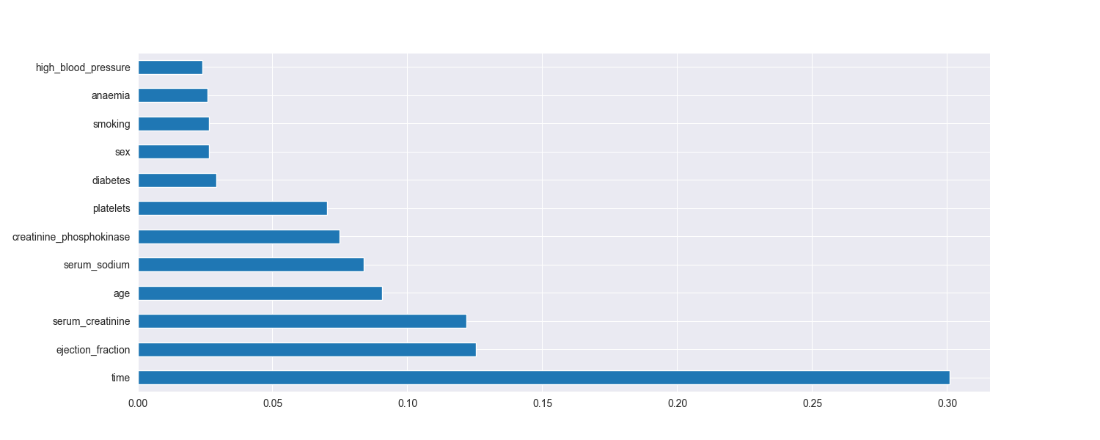
**Fig9. Diabetes distribution**



**Fig10. Anemia Distribution**

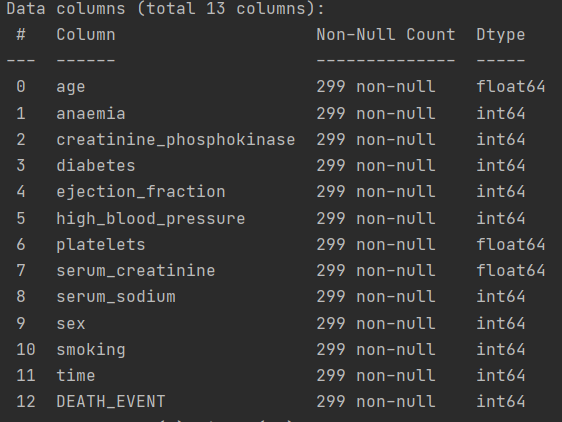
After having an in-depth analysis of all the features and their distributions, I then selected the most important features using extra tree classifier.



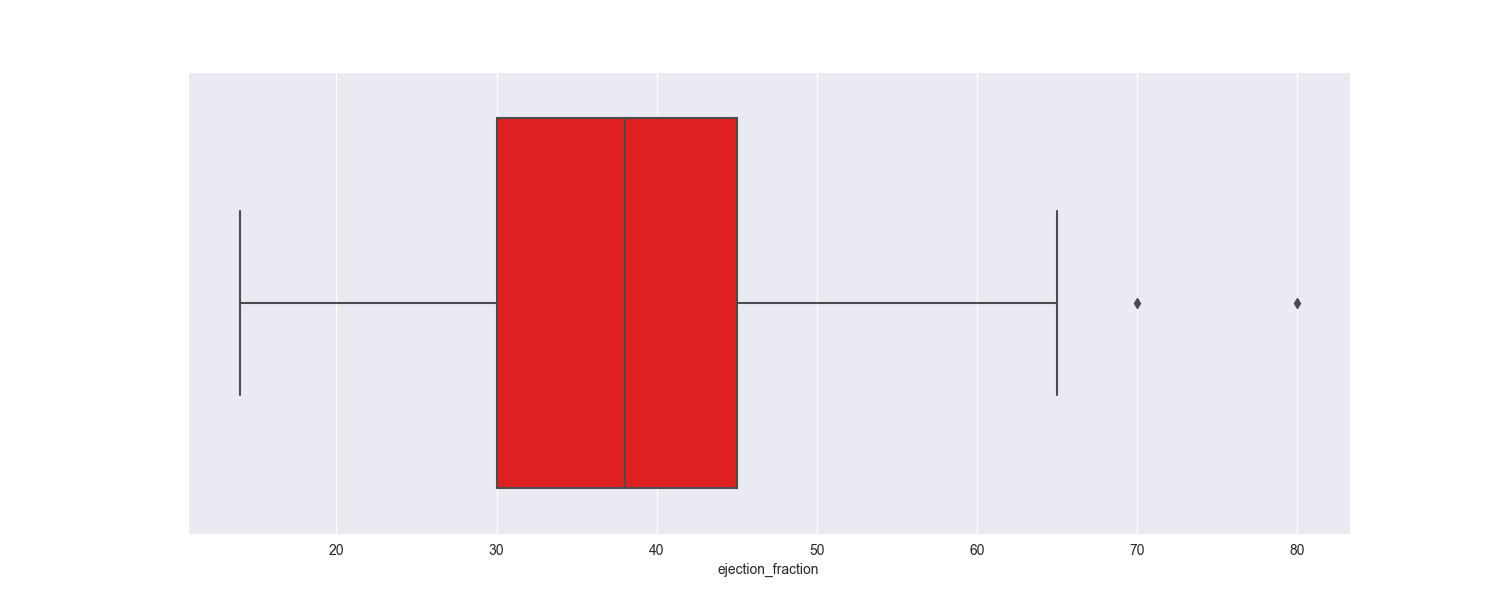
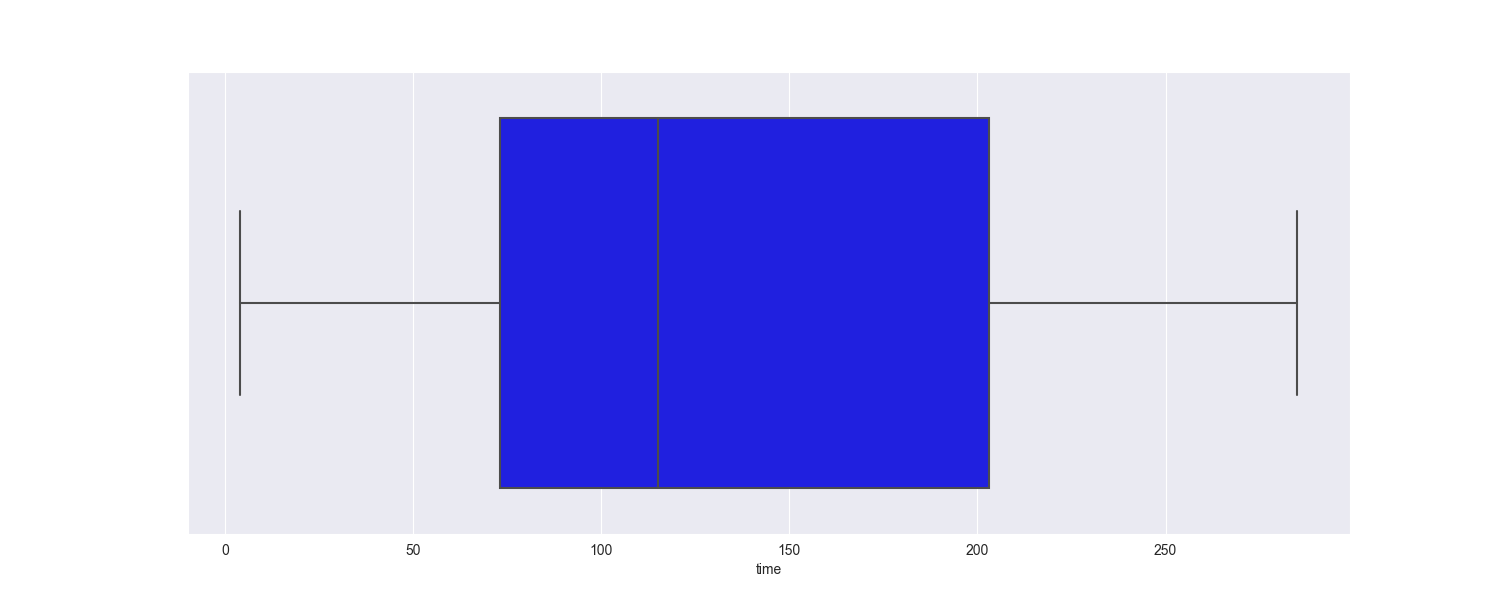


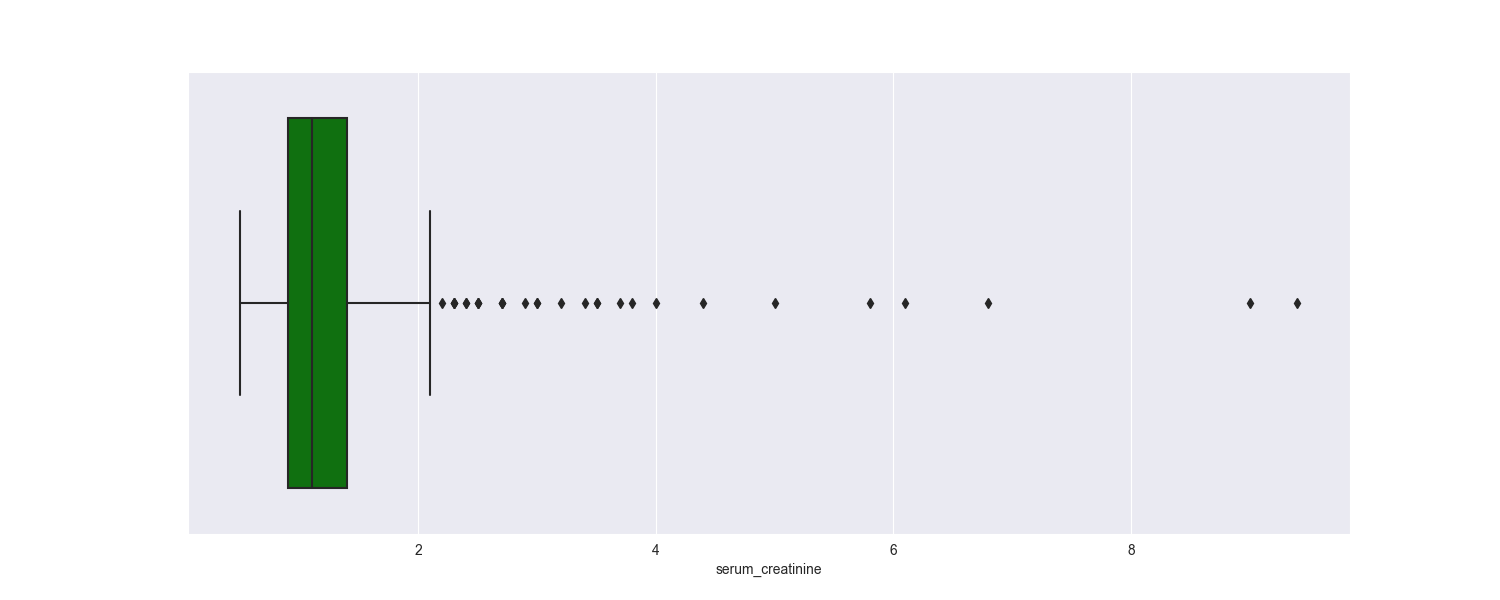
**Fig11. Code snippet and figure of the output for the feature selection using extra tree classifier**

Now, once the most important contributing features were identified we now need to check for any null values in the data set and remove outliners if any using the box plot distribution.

**Fig12. Checking null count using dataset.info () in python**

Since there are no null points in the data set, now we can proceed to see the box plot of the three most important features namely, time ejection fraction and Serum creatinine and the identify and remove outliners if any.

**Fig 13. Ejection fraction boxplot****Fig14. Time box plot**

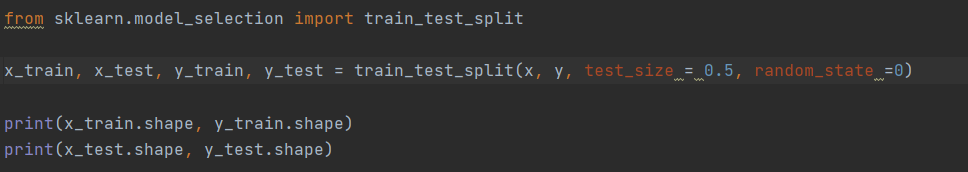


**Fig 15. Serum creatinine box plot**

Now from the time box plot we can see that there are no outliners for ejection fraction we have two outliners and for serum creatinine we have many. Before dealing with outliers, we require knowledge about the outlier, the dataset and possibly some domain knowledge. Removing outliers without a good reason will not always increase accuracy. Without a deep understanding of what are the possible ranges that# exist within each feature, removing outliers becomes tricky.After a little research I found out that all the values in serum creatinine fall into possible range of values. So, they are not outliers but rather are actual data points that helps in predicting the death event. So, we only remove the two outliners of the ejection fraction from the data set.

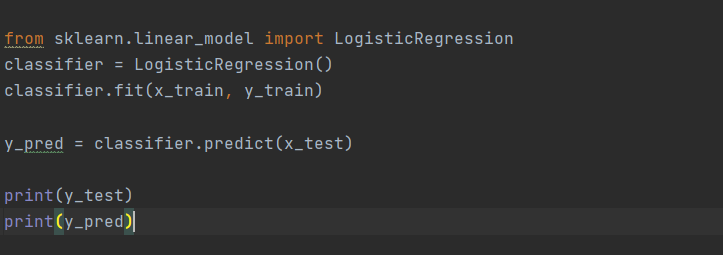
**Fig16. Code snippet for removing outliers.**

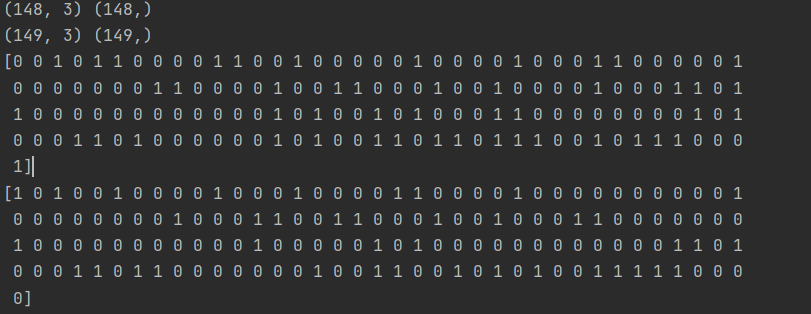
Since we have now dealt with all the outliers and null points in the data set and also have an exact idea of the statistics and distribution of all the features, we may now divide the data into test and train data.From Sklearn, sub-library model, I have imported the train\_test \_split so the I can divide the data set into testing and training data.



**Fig17. Code snippet for splitting data**

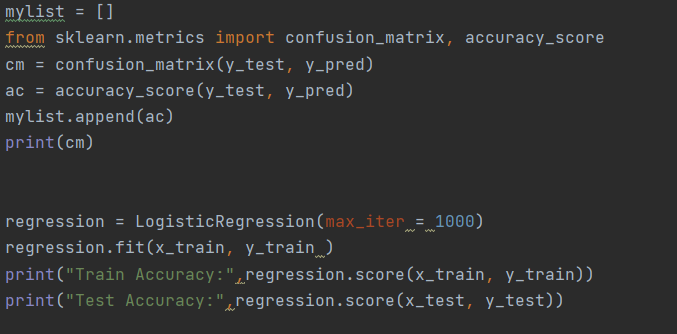
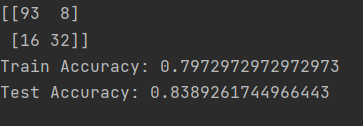
Once the data has been divided into testing and training, we apply the logistic regression on training data then predict the test set.



 **Fig18. Code snippet for logistic regression**

**Fig19.Test and train data sizes, Test outputs and train outputs for test size = 0.5**

Once the logistic regression successfully implemented, next I tried to calculate a few important parameters to check how good the algorithm is working on data set. These parameters include the confusion. matrix, tprate, fprate , testing and training accuracy



**Fig20. Code snippet for calculating confusion matrix, test accuracy and train accuracy**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Size** | **Test acc** | **Train Acc** | **TP** | **FP** | **FN** | **TN** | **tprate** | **fprate** |
| 0.1 | 90% | 82.77% | 19 | 2 | 1 | 8 | 0.95 | 0.2 |
| 0.2 | 88.3% | 80.1% | 40 | 3 | 4 | 13 | 0.909 | 0.187 |
| 0.3 | 84.44% | 80.19% | 60 | 5 | 9 | 16 | 0.869 | 0.238 |
| 0.4 | 84.48% | 79.77% | 78 | 7 | 11 | 23 | 0.876 | 0.233 |
| 0.5 | 83.89% | 79.72% | 93 | 8 | 16 | 32 | 0.853 | 0.2 |
| 0.6 | 78.81% | 83.7% | 110 | 11 | 18 | 40 | 0.859 | 0.215 |
| 0.7 | 81.7% | 76.4% | 127 | 14 | 24 | 43 | 0.841 | 0.245 |
| 0.8 | 85% | 72.88% | 151 | 13 | 21 | 53 | 0.872 | 0.197 |

**Table1. Parameters obtained from simulation for different test data set sizes using logistic regression**

From the above results we can clearly say that test size of 0.1 and 0l2 are giving us the best results.

**4) PLAN FOR NEXT HALF SEMESTER:**

In the first half I have performed EDA (Exploratory Data analysis) on the chosen data set, then divided the data set into test and train data and finally used logistic regression and calculated the accuracy. In the next half semester, I will implement K-NN, SVC and Decision tree classifier and finally find out which machine learning algorithm would be the best fit for the given data set.

**5) REFERENCES:**

1) Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,**16 (2020). <https://doi.org/10.1186/s12911-020-1023-5>

2) Larxel. (2020, June 20). Heart failure prediction. Retrieved March 10, 2021, from <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>.

3) L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, China, 2018, pp. 157-160, doi: 10.1109/ICRIS.2018.00049.

4) Brownlee, J. (2020, August 17). How to develop an extra trees ensemble with python. Retrieved March 10, 2021, from <https://machinelearningmastery.com/extra-trees-ensemble-with-python/>

5) ML: Extra Tree classifier for feature selection. (2020, July 01). Retrieved March 10, 2021, from <https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/>