e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

COMPARITIVE ANALYSIS FOR VARIOUS CLASSIFICATION MACHINE LEARNING ALGORITHMS FOR DETECTING HEART DISEASES

Akshat Chaturvedi*1, Shaik Haseeb Ur Rahman*2

*1,2Student, Department Of Computer Science And Engineering, Vellore Institute Of Technology, Vellore, Tamil Nadu, India-632014

ABSTRACT

These days heart diseases are of major concern in people. In this covid pandemic period, the people with heart diseases are more prone to death then others. Also, in this period when the medical infrastructure is collapsing it has become too difficult to get in touch with hospitals or some medical persons. It would be very helpful if we can find out whether a person is suffering from heart disease or not at home. There are various Machine Learning algorithms that we can use to predict the possibility of heart disease in a person. This research provides a comparative analysis of various Machine Learning algorithms used for predicting heart disease in a person. To draw comparison, we have used the elementary classification algorithms like Decision Trees, Support Vector Machine, KNN Classification & Logistic Regression and along with that Ensemble Learning algorithms like Random Forest & XGBoost Classifier. The result shows that Logistic Regression Model gives us the best accuracy for the dataset.

Keywords: Heart Disease, Machine Learning, Ensemble Learning, Support Vector Machine, Xgboost.

I. INTRODUCTION

The rise in digital technologies and internet world has increased the amount of data everywhere. Whether it is an IT Industry or Medical Industry or Agricultural Industry, now we have access to data more than we ever had. So, by making good use of this data we can develop new & better applications that has a positive effect in the world. A vast amount of data like patient's medical records are stored in data warehouses and medical databases. In this project we aim to use the patient's data to predict whether he is a victim of heart disease or not. The information in the data has the details of patient such as age, sex, cholesterol level, resting ecg, etc. The disease prediction is a very sensitive task because in the worst-case scenario it can also cost someone their life. So, along with the predictions we'll compare the accuracies of the algorithms and then give the best algorithm for the prediction. These prediction APIs can be stored in a person's personal desktops or the hospital reception computers, so that they don't have to consult a doctor to check whether they are prone to heart disease or not.

II. METHODOLOGY

For the implementation of this project work, we have used Google Colaboratory. We'll start by enabling GPU in the notebook. GPU gives us better performance so that we get less waiting time while the code is executing. In this research we worked with the Heart Diseases Dataset which we have taken from UCI Machine Learning Repository. The dataset has 13 Independent Features and 1 Dependent Feature. The dataset has 303 entries corresponding to each feature. The reason we chose this dataset because it is perfect for analyzing models based on their Binary Classification ability. https://archive.ics.uci.edu/ml/datasets/heart+disease

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

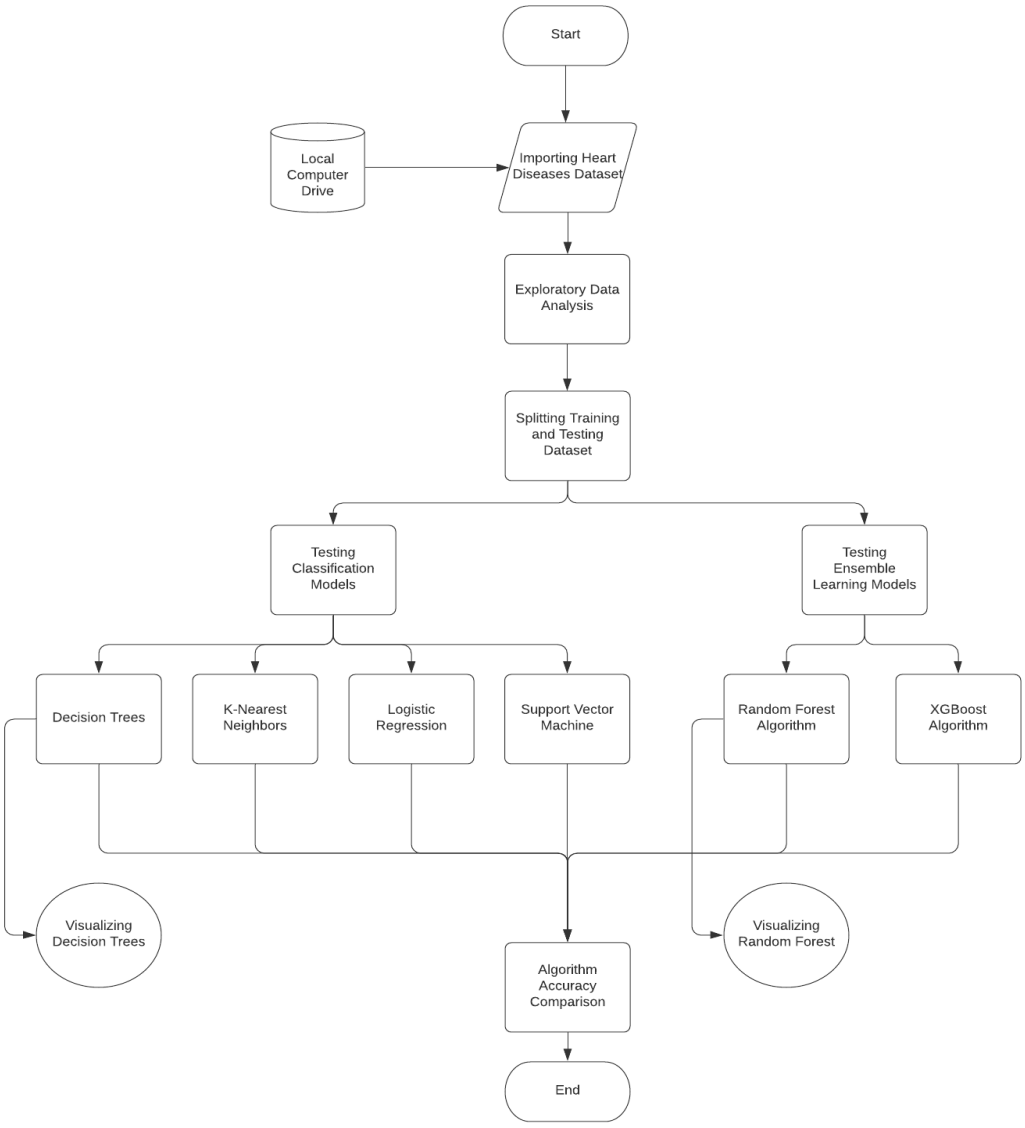

Figure 2.1 Workflow of the proposed methodology

The first and foremost thing is to import the dataset (csv file) into our notebook.

EXPLORATORY DATA ANALYSIS

The first task in EDA is to check for Missing or Nan values in the dataset. We can find the number of Nan values corresponding to each feature easily by basic python coding. It is always best to visualize the null values with graphical visualization for that we've used Heat Map to plot null values & also missing number matrix plot.

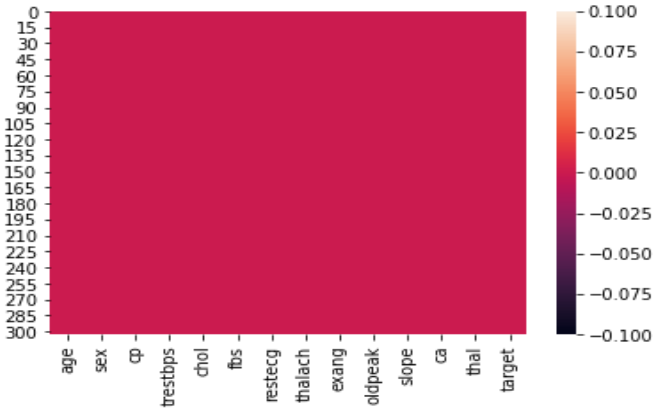

Figure 2.2 Heatmap visualization of Null Values

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

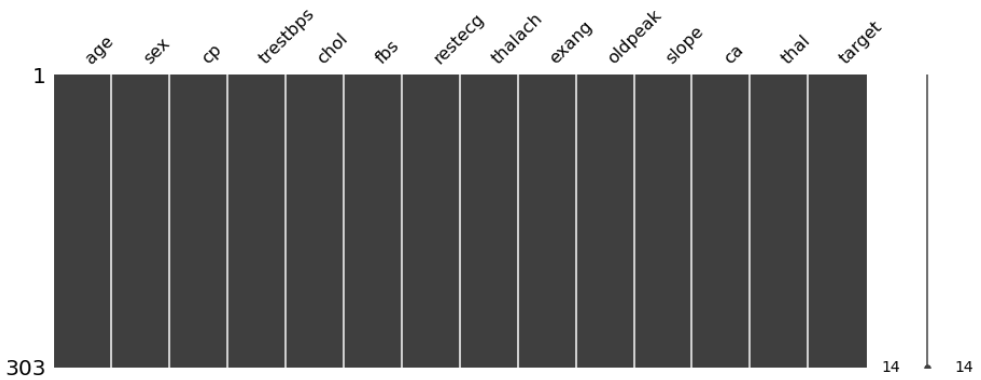

Figure 2.3 Matrix Plot for null counts

Inference: As there are no Null values found in the dataset we can process further. Now we'll use the data.describe() method to find the Statistical Measures corresponding to each column. The Mean, Standard Deviation, Median & Quartiles deviations gives us a lot of details about the distribution of the dataset. The dataset has 165 (around 54%) 'True' values and 138 (around 46%) 'False', so we can confirm that the dataset is balanced.

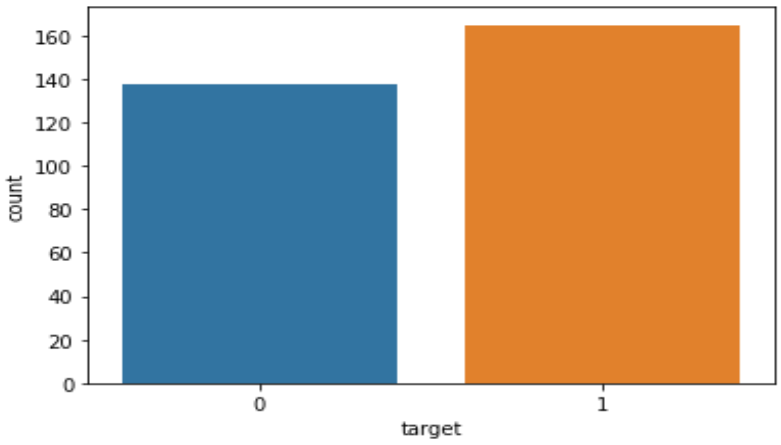

Figure 2.4 Count Plot of Target Feature

Analyzing the dataset based on demographics

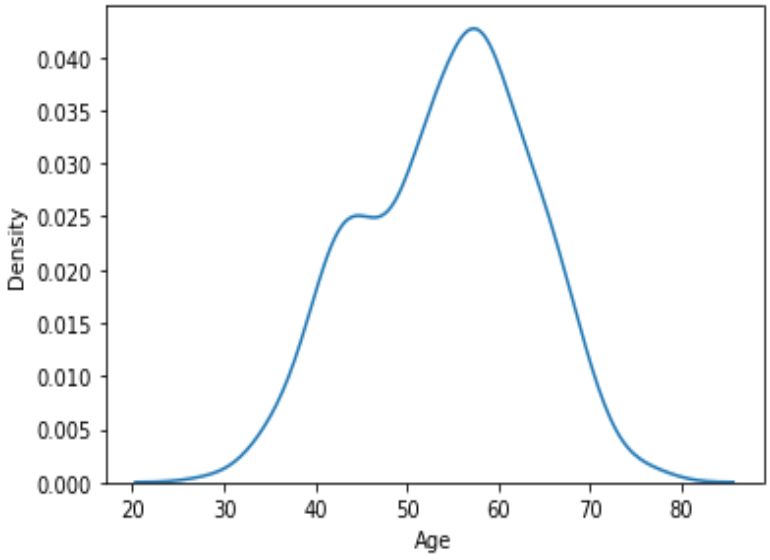

Figure 2.5 Age – Kernel Density Plot

Inference: The maximum number of people suffering from Heart Disease are in the age group (55-60)

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

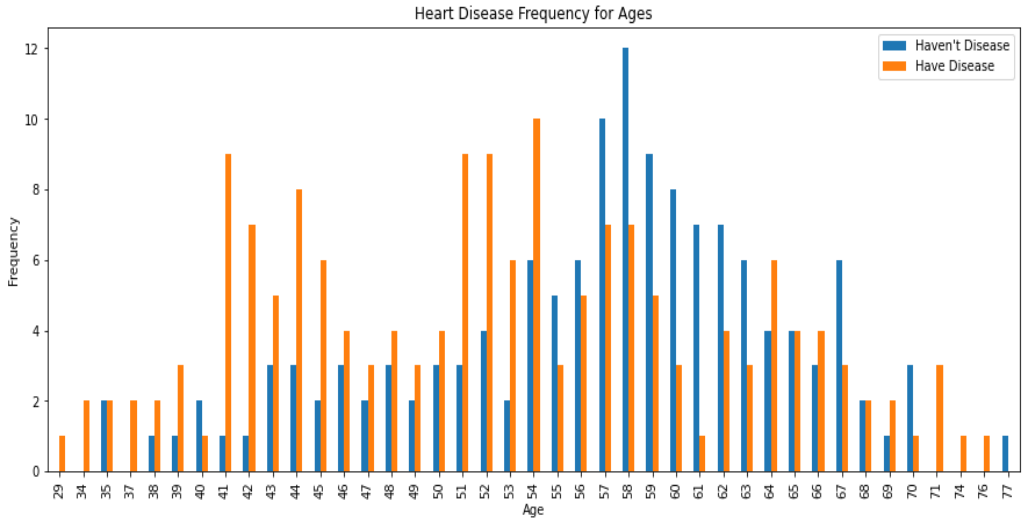

Figure 2.6 Age-Frequency Bar Plot

We can see from the Graph that in our data the out of all the people from a specified age how many are affected by diseases and how many aren't affected.

PARTIONING THE DATASET

- First, we've removed the Target feature and store it as our label.
- Then we've used train_test_split method available in sklearn.model_selection library to split our dataset into training and testing.
- We've kept the length of our test dataset as 0.25 (0.25*303 = 76 - rows in our case) because it works best in most of the scenarios.

RUNNING CLASSIFICATION ALGORITHMS

For comparing classification accuracies, for each algorithm we've first imported the classifier of that algorithm from sklearn library and then fitted it onto our training dataset. We then used our testing dataset to predict the labels, after prediction of labels we've compared it with original labels of testing dataset to get accuracy score. We generated classification report for every algorithm to get the values of respective Precision, Recall & F1-scores. Since, in this paper our task is to classify whether a person has Heart Diseases or not, it is very important for us to judge the values of Precision and Recall along with accuracy. Because a patient with Heart Disease getting classified as Healthy will be a major drawback.

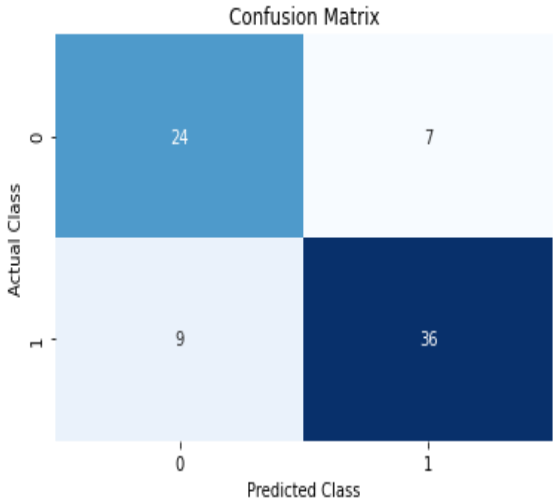

Figure 2.7 Decision Trees Confusion Matrix

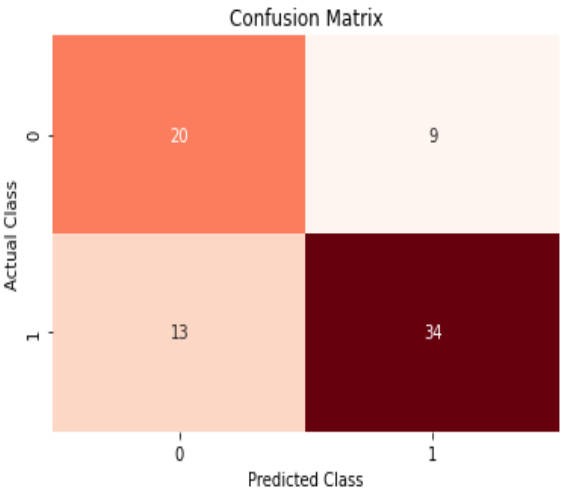

Figure 2.8 KNN Classifier Confusion Matrix

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

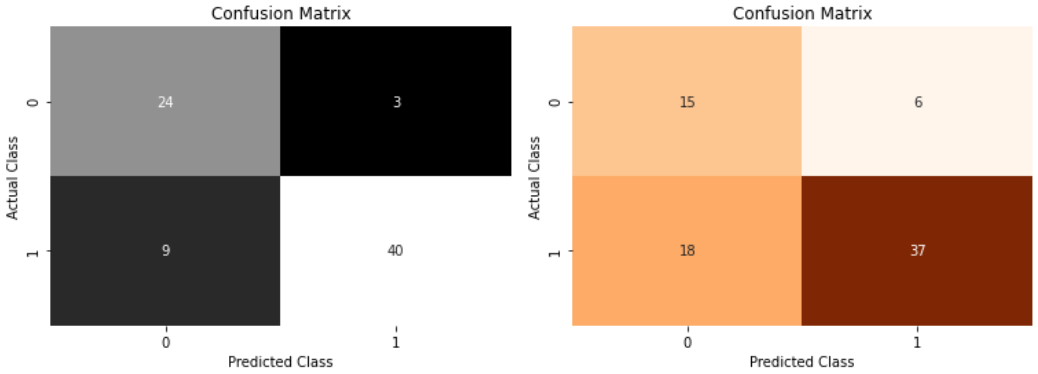

Figure 2.9 Logistic Regression Classifier Confusion Matrix **Figure 2.10** SVM Classifier Confusion Matrix

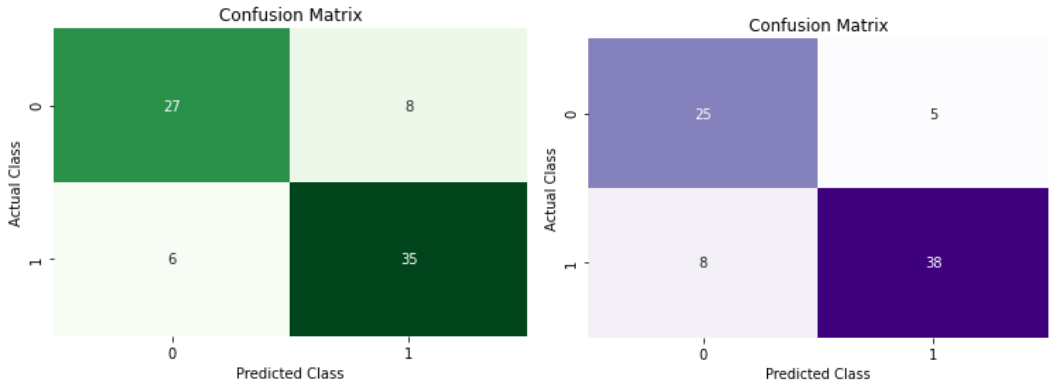

Figure 2.11 Random Forest Confusion Matrix **Figure 2.12** XGBoost Classifier Confusion Matrix

III. MODELING AND ANALYSIS

Visualizing different steps in Decision Tree

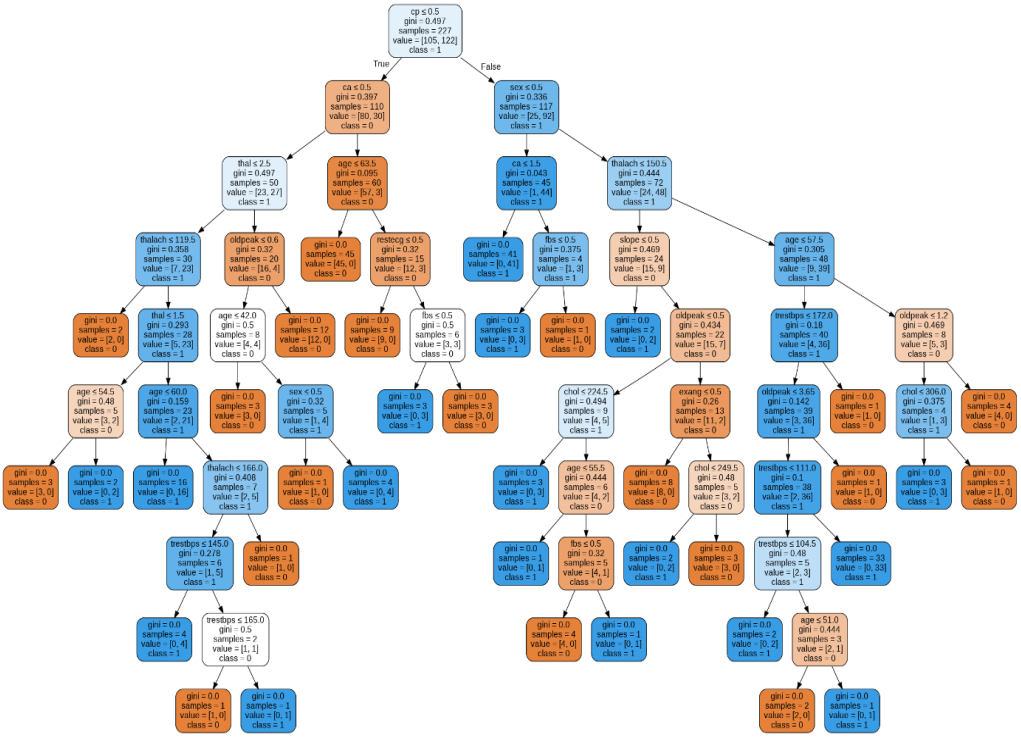

Figure 3.1 Decision Tree for Heart Disease Dataset

K Nearest Neighbors Analysis

We ran KNN in a with different values of K to get best accuracy score. We ran the loop from 1 to 25, in each iteration we provided the value of n_neighbors as i and then trained it with our training data & got the accuracy.

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

```
Accuracy of our KNN model for K value 1 is:  0.5263157894736842
Accuracy of our KNN model for K value 2 is:  0.618421052631579
Accuracy of our KNN model for K value 3 is:  0.6578947368421053
Accuracy of our KNN model for K value 4 is:  0.6842105263157895
Accuracy of our KNN model for K value 5 is:  0.6447368421052632
Accuracy of our KNN model for K value 6 is:  0.6447368421052632
Accuracy of our KNN model for K value 7 is:  0.6710526315789473
Accuracy of our KNN model for K value 8 is:  0.6578947368421053
Accuracy of our KNN model for K value 9 is:  0.6710526315789473
Accuracy of our KNN model for K value 10 is:  0.6842105263157895
Accuracy of our KNN model for K value 11 is:  0.7105263157894737
Accuracy of our KNN model for K value 12 is:  0.6842105263157895
Accuracy of our KNN model for K value 13 is:  0.6973684210526315
Accuracy of our KNN model for K value 14 is:  0.6973684210526315
Accuracy of our KNN model for K value 15 is:  0.6973684210526315
Accuracy of our KNN model for K value 16 is:  0.6973684210526315
Accuracy of our KNN model for K value 17 is:  0.7105263157894737
Accuracy of our KNN model for K value 18 is:  0.6710526315789473
Accuracy of our KNN model for K value 19 is:  0.6710526315789473
Accuracy of our KNN model for K value 20 is:  0.6710526315789473
Accuracy of our KNN model for K value 21 is:  0.6710526315789473
Accuracy of our KNN model for K value 22 is:  0.6842105263157895
Accuracy of our KNN model for K value 23 is:  0.6710526315789473
Accuracy of our KNN model for K value 24 is:  0.6842105263157895
```

Figure 3.2 Different accuracies with different K values

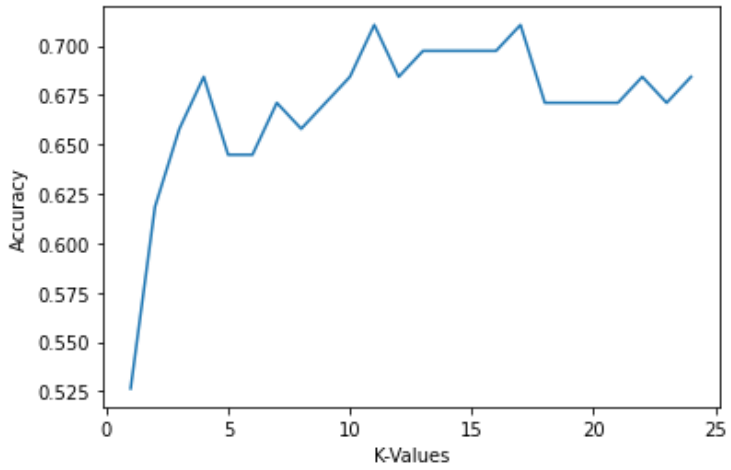

Figure 3.3 Accuracy vs K-value Graph

As we can see from the graph, we got best accuracy for K=11 which is around 0.7105.

IV. RESULTS AND DISCUSSION

ELEMENTARY CLASSIFICATION MODELS

Model	Accuracy	Execution Time (ms)	F1-Score
Decision Tree	0.7894	4.3952	0.82
K-Nearest Neighbors	0.7105	4.5602	0.76
Logistic Regression	0.8421	48.4745	0.87
Support Vector Machine	0.6842	10.7893	0.76

ENSEMBLE LEARNING MODELS

Model	Accuracy	Execution Time (ms)	F1-Score
Random Forest	0.81	23.1913	0.83
XGBoost	0.82	43.2531	0.85

e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

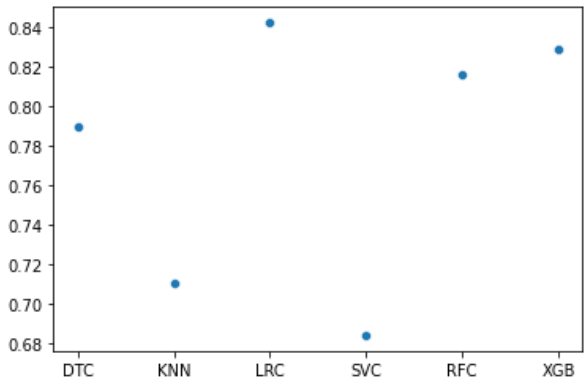

Figure 4.1 Scatter plot of the accuracies

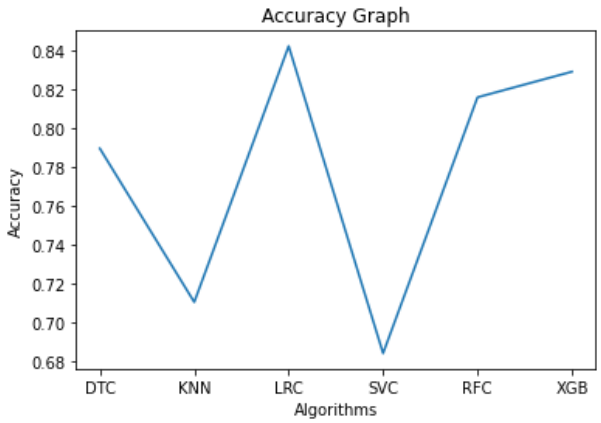

Figure 4.2 Algorithms vs Accuracy Line Graph

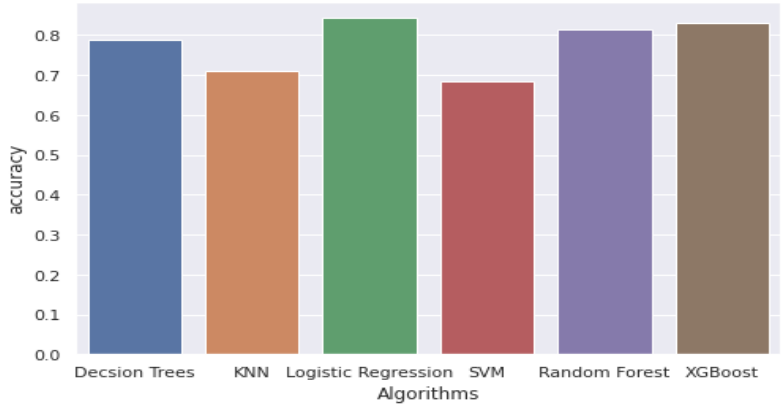

Figure 4.3 Bar Plot of Accuracy vs Algorithms

V. CONCLUSION

Heart Diseases is a very hazardous disease nowadays because not only a large number of people are suffering from it but also because it invites many other deadliest diseases like COVID. We know that the number of people getting sick are increasing at an alarming rate & we can't always rely on our Medical Infrastructure. But the good thing is that the data is also increasing, so if we can make use of that data, it will be very beneficial for us. In our study we have used Heart Disease Dataset and made predictions on the data by utilizing 6 different types of Machine Learning algorithms. The outcome of our comparative analysis shows that Logistic Regression Classifier gives us the best accuracy & F1-score. A future scope for this project will be- to make available an API with front-end website which can be used by people at their homes or hospital receptions to predict whether a person is having heart disease or not.

VI. REFERENCES

[1] Zunaidi, W.H.A.W., Saedudin, R.R., Shah, Z.A., Kasim, S., Seah, C.S. and Abdurohman, M., 2018. Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(6), pp.2677-2682.

[2] Shaikh, A., Mahoto, N., Khuhawar, F. and Memon, M., 2015. Performance evaluation of classification methods for heart disease dataset. *Sindh University Research Journal-SURJ (Science Series)*, *47*(3).

[3] Ranganatha, S., Raj, H.P., Anusha, C. and Vinay, S.K., 2013. Medical data mining and analysis for heart disease dataset using classification techniques.

[4] Shouman, M., Turner, T. and Stocker, R., 2012. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, *2*(3), pp.220-223.

[5] Deepika Bansal, Rita Chhikara, Kavita Khanna, Poonam Gupta, Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia, Procedia Computer Science, Volume 132, 2018, Pages 1497-1502, ISSN 1877-0509.