Heart disease Prediction Using Machine Learning Algorithms

Kavitha Bai A S[1], Rachitha M V[2], Noor Basha[3]

[1,2,3] *Assistant Professor in CSE, Vemana Institute of Technology, Bengaluru, Karnataka, India.*

E-mail: kavithabai.pawar@gmail.com, rachithamv@gmail.com, md.noor202@gmail.com

Abstract: According to the Centers for Disease Control and Prevention (CDC), heart disease is one of the leading causes of death for people of most races in the world. About half of all world population (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator includes diabetic status, obesity, not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition. Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. This presents various attributes related to heart disease, and the model on basis of logistic regression, K-neighbor Classifier, Decision tree model.

Keywords: **Heart disease, Machine learning, Prediction, logistic regression.**

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

The classification is common in every study to predict that the patient's common pattern in which detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Researchers found that, throughout life, men were about twice as likely as women to have a heart attack. That higher risk persisted even after they accounted for traditional risk factors of heart disease, including high cholesterol, high blood pressure, diabetes, body mass index, and physical activity. It is considered as one of the benchmark datasets when someone is working on heart disease prediction. Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports. Many studies have been performed and various machine learning models are used for doing the classification and prediction for the diagnosis of heart disease. An automatic classifier for detecting congestive heart failure shows the patients at high risk and the patients at low risk. Then for improving the performance electrocardiogram (ECG) approach is suggested in which deep neural networks are used for choosing the best features and then using them.

Terabytes of data are produced and stored day-to-day life because of fast growth in Information Technology. The data which is collected is converted into knowledge by data analysis by using various combinations of algorithms. For Example: the huge amount of the data regarding the patients is generated by the hospital such as X-ray results, lung results, heart paining results, chest pain results, personal health records (PHRs), etc. There is no effective use ofthe data which is generated from the hospitals.

DOI: 10.55041/IJSREM14688

II. LITERATURE REVIEW

Noor Basha et al [1] made use of many classification algorithms to predict the severe heart syndromes based on risk rate, where the author specifically used machine learning approach. The approach can be used for various datasets to predict the conditional requirements. Support Vector machine approach to predict the syndrome.

Noor Basha et al [2] a novel distance-based clustering algorithm called the entropic distance based K-means clustering algorithm (EDBK) is proposed to remove the outliers in effective way. The entropic distance between attributes of data points and some basic mathematical statistics operations.

III. Proposed System

3.1 MACHINE LEARNING:

Machine Learning is a system that can learn from example through self- improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights.Machine learning is closely related to data mining and Bayesian predictive modelling. The machine receives data as input, use an algorithm to formulate answers. A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

3.2 Pandas

Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data_science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like Active State’s Active Python.

3.3 Numpy

NumPy, which stands for Numerical_Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays_can be performed. NumPy is a Python package. It stands for ‘Numerical Python’. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

3.4 Sklearn

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

3.5 Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to thedata structures from pandas. Seaborn aims to make visualization the central part of exploring and understandingdata. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

3.6 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots likeline, bar, scatter, histogram etc.

3.7 k-nearest neighbors

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

3.7 Decision Tree Classifier

Decision tree classifiers provide a readable classification model that is potentially accurate in many different application contexts, including energy-based applications. The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. Each leaf represents class labels associated with the instance. Instances in the training set are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Starting from the root node of the tree, each node splits the instance space into two or more sub-spaces according to an attribute test condition. Then moving down the tree branch corresponding to the value of the attribute, a new node is created. This process is then repeated for the sub tree rooted at the new node, until all records in the training set have been classified. The decision tree construction process usually works in a top-down manner, by choosing an attribute test condition at each step that best splits the records.

IV. EXPERIMENT RESULTS AND DISCUSSION

Data mining technology afford an effective approach to latest and indefinite patterns in the data. These hidden patterns can be used for health diagnosis in medicinal data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most important reason of victims in the countries like India, UnitedStates.

To initiate with the work, we can use different types of techniques and algorithms. In thisprocess, machine learning techniques are used to increase the accuracy rate. In machine learning technique we can use the following algorithm:

1. Logistic Regression
2. Comparing and Confusion Matrix

Sklearn Logistic Regression

The term regression can be defined as the measuring and analyzing the relation between one or more independent variable and dependent variable. Regression can be defined by two categories; they are linear regression and logistic regression. Logistic regression is a generalized by linear regression. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression i.e., discrete variable changed into continuous value.

Logistic regression basically is used to classify the low dimensional data having nonlinear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main motto of Logistic regression is to determine the result of each variable correctly Logistic regression is also known as logistic model/ logit model that provide categorical variable for target variable with two categories such as light or dark, slim/ healthy.

The logistic regression is also known as sigmoid function which helps in the easyrepresentation in graphs. It also provides high accuracy. In this algorithm first the data should be imported and then trained. By using equation, the logistic regression algorithm is represented in the graphs showing the difference between the attributes. From the training data we have to estimate the best and approximate coefficient and represent it.

Comparing and Confusion Matrix

The comparison of confusing matrices, is the summary for the prediction of the result which we classified. Based on the classification of attributes the correct and incorrect predictions are marked with count values. A confusion matrix, which is represented in table format will explains about the performance of characterization model on the trained dataset. The most of the performance measures are calculated using this confusionmatrix.

The most important organ of the human body is heart. The function of the heart is to pump the blood and circulates entire body. It is protected by rib cage and it is surroundedby two layered tissue membrane called Pericardium. It is a four chambered organ which separates oxygenated and deoxygenated blood. Heart is having the five types of blood vessels, arteries, veins, capillaries, arterioles and venules. The size of the human heart is about the size of the fist and weight approximately 300grams, the weight in femalesbeing about 25% lesser than males. Arteries and veins are present in heart which helps to collect the blood from all the parts and purifies it and the circulates to all the body parts. The nutrients and oxygen present in the body parts are provided by the blood and alsowill helps in the removal of metabolic wastes. Now a day the life span of human beingis reduced due to the heart diseases. The factors which may lead to heart disease are obesity, high cholesterol, smoking, increase in blood pressure, diabetes and other factors.

As per the WHO (World Health Organization) records, each and every year millions of people died with different types of cardio attacks such as heart stroke, chest pain, etc. We here proposed the collection of relevant data from the hospitals where the data is so enormous. Now we have to separate the data regarding the patients related to heart diseases. We train the data as per proposed algorithm of machine learning using logistic regression. For the purpose of detecting the heart disease, we can enter patient medical details into the trained data. Training data we have to estimate the best and approximate coefficient and represent it.

Comparing and Confusion Matrix

The comparison of confusing matrices, is the summary for the prediction of the result which we classified. Based on the classification of attributes the correct and incorrect predictions are marked with count values. A confusion matrix, which is represented in table format will explains about the performance of characterization model on the trained dataset. The most of the performance measures are calculated using this confusion matrix.

The most important organ of the human body is heart. The function of the heart is to pump the blood and circulates entire body. It is protected by rib cage and it is surroundedby two layered tissue membrane called Pericardium. It is a four chambered organ which separates oxygenated and deoxygenated blood. Heart is having the five types of blood vessels, arteries, veins, capillaries, arterioles, venules. The size of the human heart is about the size of the fist and weight approximately 300grams, the weight in femalesbeing about 25% lesser than males. Arteries and veins are present in heart which helps to collect the blood from all the parts and purifies it and the circulates to all the body parts. The nutrients and oxygen present in the body parts are provided by the blood and alsowill helps in the removal of metabolic wastes. Now a day the life span of human beingis reduced due to the heart diseases. The factors which may lead to heart disease are obesity, high cholesterol, smoking, increase in blood pressure, diabetes and other factors.

As per the WHO (World Health Organization) records, each and every year millions of people died with different types of cardio attacks such as heart stroke, chest pain, etc. We here proposed the collection of relevant data from the hospitals where the data is so enormous. Now we have to separate the data regarding the patients related to heart diseases. We train the data as per proposed algorithm of machine learning using logistic regression. For the purpose of detecting the heart disease, we can enter patient medical details into the trained data.

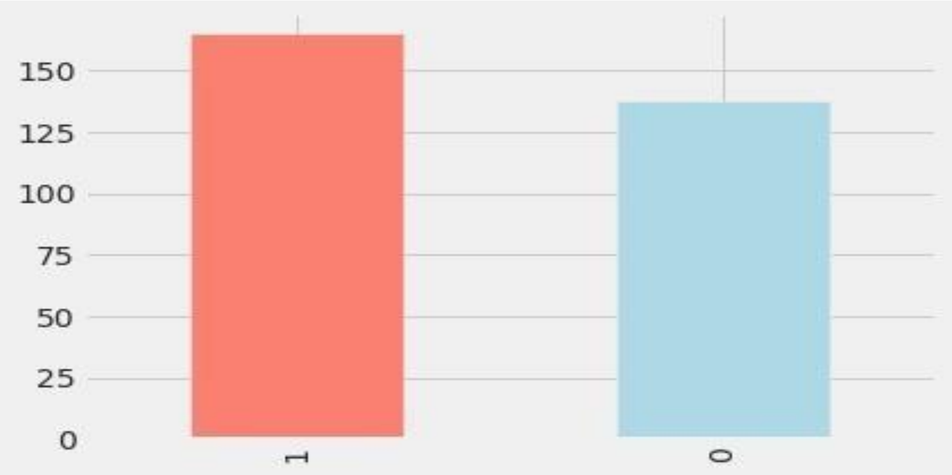
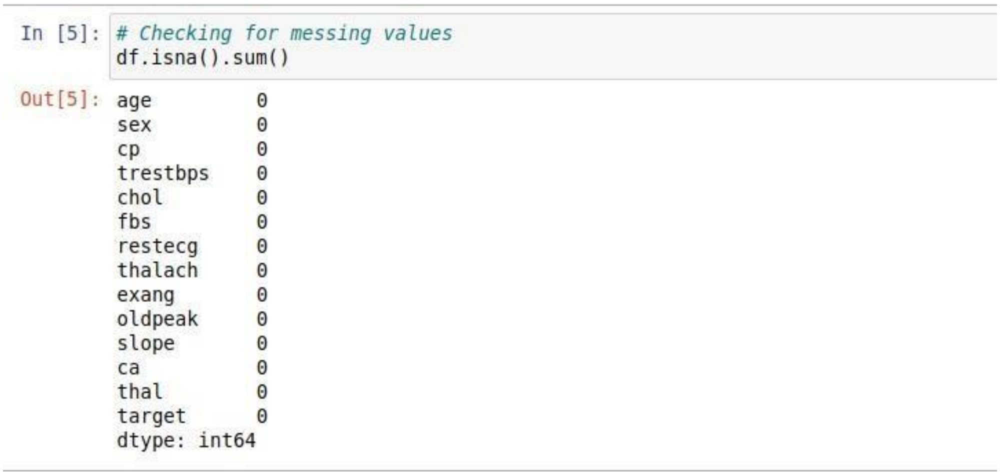


Fig 4.1: Target counts

Target count shows target counts of the disease, we have 165 people with heart disease and 138people without heart disease.



```
In [5]: # Checking for messing values
        df.isna().sum()

Out[5]: age         0
        sex         0
        cp          0
        trestbps    0
        chol        0
        fbs         0
        restecg     0
        thalach     0
        exang       0
        oldpeak     0
        slope       0
        ca          0
        thal        0
        target      0
        dtype: int64
```

Fig 4.2: Missing Value Analysis

We have to remove null values from the dataset to make dataset much accurate.

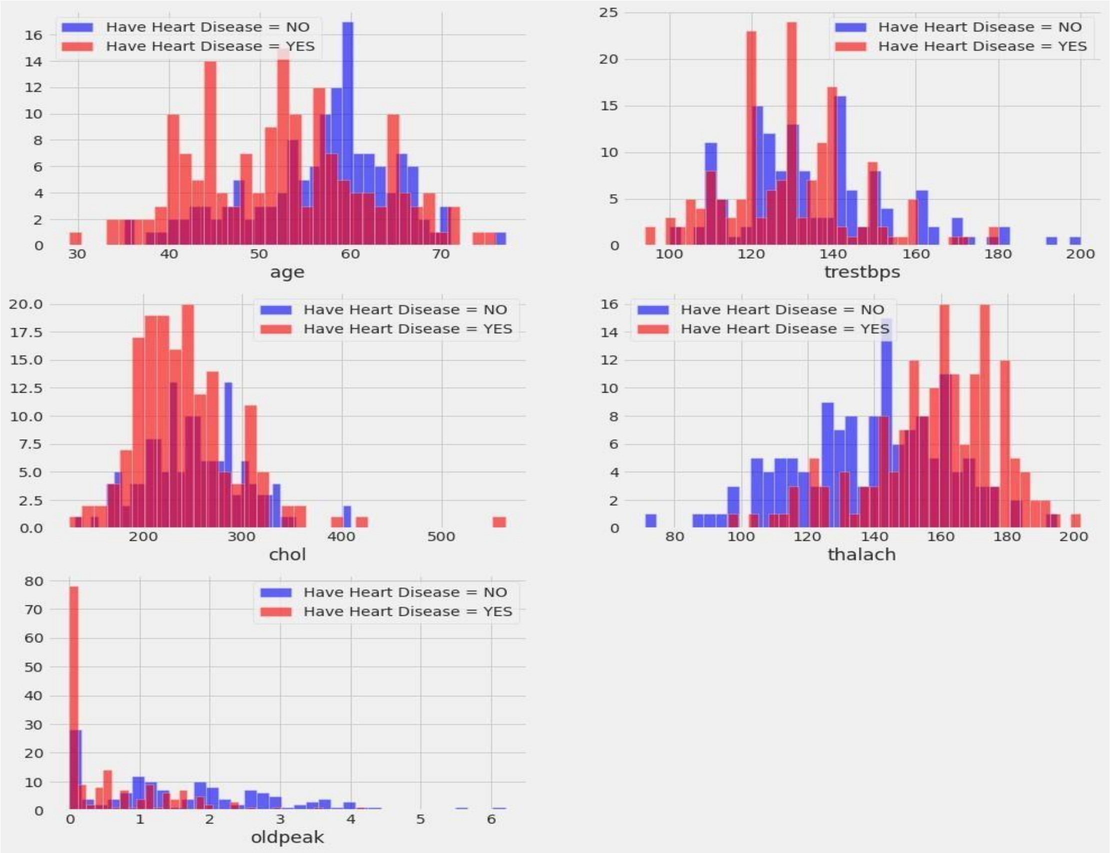


Fig 4.3: Visualization of the Dataset

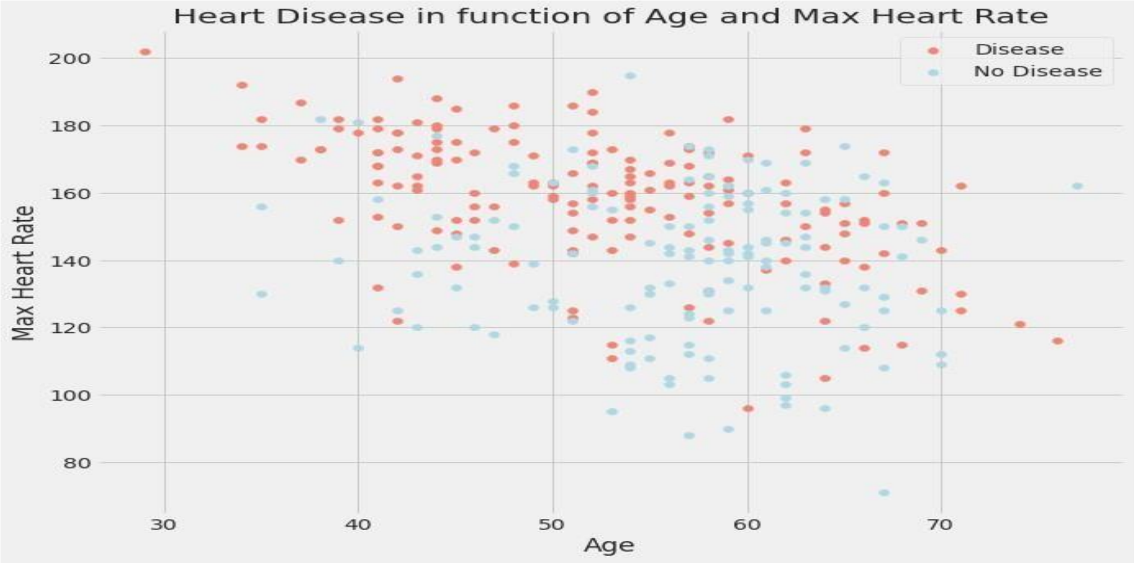


Fig 4.4: Relation Between Heart Rate and Diseases Observation:

As we can see the There is High chance of having heart disease if person is detected with high heart rate.

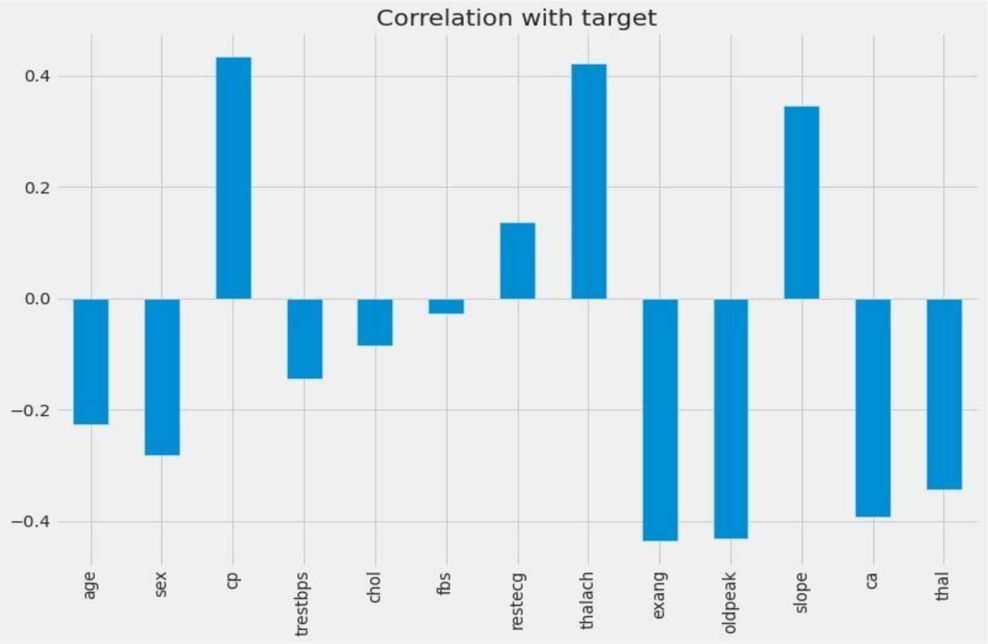
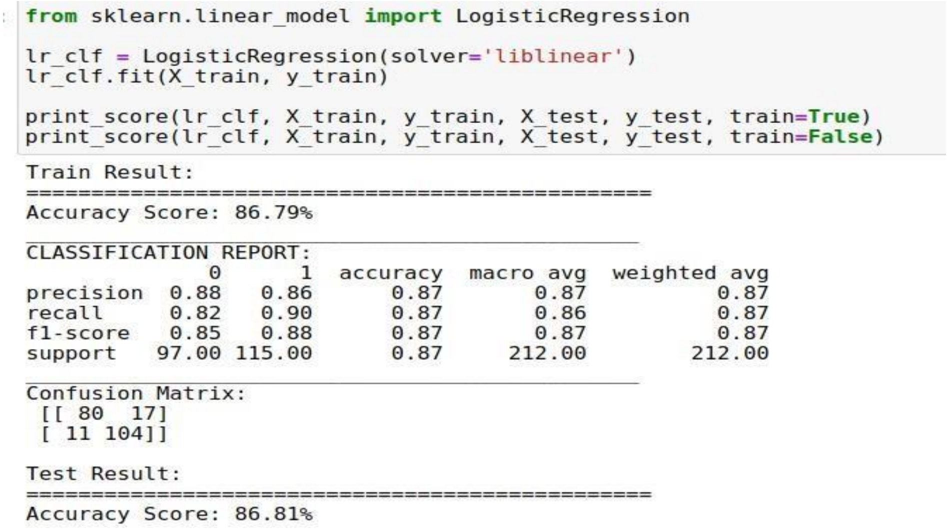


Fig 4.5: Correlation Analysis

- fbs and chol are the least correlated with the target variable.
- All other variables have a significant correlation with the target variable.



```
from sklearn.linear_model import LogisticRegression

lr_clf = LogisticRegression(solver='liblinear')
lr_clf.fit(X_train, y_train)

print_score(lr_clf, X_train, y_train, X_test, y_test, train=True)
print_score(lr_clf, X_train, y_train, X_test, y_test, train=False)
```

```
Train Result:
================================================
Accuracy Score: 86.79%
_______________________________________________
CLASSIFICATION REPORT:
              0       1  accuracy  macro avg  weighted avg
precision  0.88    0.86      0.87       0.87          0.87
recall     0.82    0.90      0.87       0.86          0.87
f1-score   0.85    0.88      0.87       0.87          0.87
support   97.00  115.00      0.87     212.00        212.00
_______________________________________________
Confusion Matrix:
 [[ 80  17]
 [ 11 104]]

Test Result:
================================================
Accuracy Score: 86.81%
```

Fig 4.6: Results and Accuracy

Results and accuracy, this model achieved 86.81% accuracy.

V. CONCLUSION

As identified through the logistic regression, it is a more efficient than the data mining techniques as it is combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases. The amount of heart diseases can exceed the control line and reach to maximum point. Heart disease are complicated and each and every year lots of people are dying with this disease by using this all systems one of the major drawbacks of these works is mainly focus only to the application of classify techniques and algorithms for heart disease prediction, by all these studying various data cleaning and mining techniques that prepare and build a dataset appropriate for data mining. So, by using Machine Learning in the logistic regression algorithms to predict if patient has heart disease or not is effective. Any non-medical employee can use this software and predict the heart disease and reduce the time complexity of the doctors. The framework can also be extended for use on other models such as neural networks, ensemble algorithms, etc.

REFERENCES

[1] Noor Basha, Ashok Kumar P.S, P Venkatesh, “Early Detection of Heart Syndrome Using Machine Learning Technique,” 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT) Pages 387-391.

[2] Noor Basha, Ashok Kumar PS “Distance-based K-Means Clustering Algorithm for Anomaly Detection in Categorical Datasets” International Journal of Computer Applications, volume 183, issue11, page 9-14.

[3] Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01

[4]]Prerana T H M, Shivaprakash N C et al “Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PACAlgorithm, Comparison of Algorithms and HDPS”,Vol 3, PP: 90-99 ©IJSE,

[5] Noor Basha, K Manjunath, Mohan Kumar Naik, PS Ashok “Analysis and Forecast of Heart Syndrome by Intelligent Retrieval Approach” Intelligent Computing and Innovation on Data Science, Springer, Singapore, pages 507-515.

[6] F. Zabihi, and Babak Nasiri. "A Novel History-driven Artificial Bee Colony Algorithm for Data Clustering." Applied Soft Computingvol. 71, pp. 226-241, 2018.

[7] Purushottam, Kanak Saxena, Richa Sharma Efficient heart disease prediction system using Decision tree (2015)

[8] Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)

[9] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis andPrediction Using Machine Learning and Data Mining Techniques: AReview (2017)

[10] Noor Basha, PS Ashokkumar, P Venkatesh ”Reduction of Dimensionality in Structured Data Sets on Clustering Efficiency in Data Mining “ IEEE International Conference on Computational Intelligence and Computing Research (ICCICI), pages 1-4.