IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋇ Impact Factor 7.39 ⋇ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11696

Heart Disease Prediction Using Machine Learning Algorithms

Devi Kannan[1], Akashdeep Boxi[2]

[1]Assistant Professor, Dept. of Computer Science and Engineering, Bangalore

[2]Student, Dept. of Computer Science and Engineering, Bangalore, Karnataka, India

Abstract: The heart is an important organ in living things. Heart-related disease diagnosis and prognosis calls for greater research precision, excellence, and correctness because even the smallest error can cause fatigue issues or individual death, there are many heart-related deaths are becoming more common, and their number is growing day, exponentially. A illness awareness prediction system is absolutely necessary to address the issue.Machine learning, a subset of artificial intelligence (AI), offers excellent assistance in making predictions about any form of event using data from real-world occurrences. In this study, utilising the UCI repository dataset for training and testing, we measure the accuracy of machine learning methods for predicting cardiac disease. These algorithms include k-nearest neighbour, decision tree, linear regression, and support vector machine (SVM). The greatest tool for implementing Python programming is the Anaconda (Jupytor) notebook, which has a variety of header files and libraries that improve the accuracy and precision of the task.

Keywords:supervised; unsupervised; reinforced; linear regression; decision tree; python programming; jupytor Notebook; confusion matrix;

I. INTROUDUCTION

The maintenance of the heart is crucial because it is one of the largest and most important organs in the human body. Since the majority of diseases are heart-related, it is important to predict heart diseases, and for this reason comparative studies in the field are required. Since most patients die today because their diseases are discovered too late due to instrument inaccuracy, it is important to learn about more effective disease prediction algorithms.

One effective testing tool is machine learning, which is based on training and testing. Machine learning is a particular subset of Artificial Intelligence (AI), a large field of learning in which machines imitate human abilities. On the other hand, machine learning systems are taught how to process and use data; as a result, the fusion of the two fields of technology is also known as machine intelligence.

In line with the definition of machine learning, which states that it learns from natural
phenomena and things, this project uses biological parameters as testing data, such as cholesterol, blood pressure, sex, age, etc., and on the basis of these, comparison is done in terms of the accuracy of algorithms. For example, in this project, we used four algorithms: decision tree, linear regression, k-neighbor, and SVM.

The accuracy of four different machine learning algorithms is calculated in this study, and the result is used to determine which strategy is the most accurate.This paper's introduction to machine learning and cardiac disorders is included in Section 1. The classification using machine learning was described in Section II. Section III provided examples of the researchers' relevant work. The mechanism for this prediction system is covered in Section IV. The algorithms utilised in this research are discussed in Section V. Section VI provides a quick overview of the dataset and its analysis in relation to the project's findings. The summary of this work is concluded with Section VII, which also offers a brief prediction of its potential future reach.

II. MACHINE LEARNING

One effective technique is machine learning, which relies on two concepts: testing and training. The system learns from data and experience directly, and based on this training, tests should be applied to various needs in accordance with the necessary algorithms.
Three categories of machine learning algorithms exist:

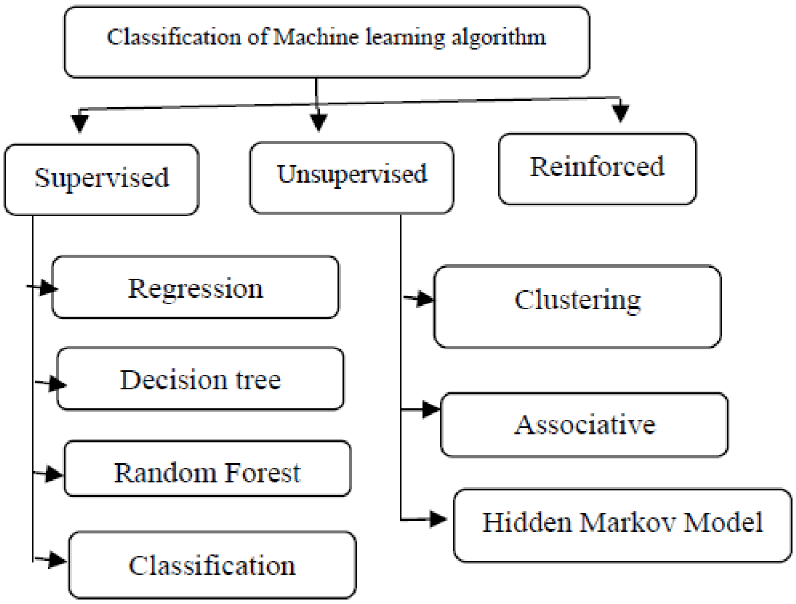


Fig.1 Classification of machine learning

A. Supervising

Supervised learning is defined as learning under proper supervision or as learning while a teacher is present. We have a training dataset that serves as the teacher for making predictions on the given dataset, therefore there is always a training dataset when testing a dataset. The "train me" principle underpins supervised learning. The processes involved in supervised learning include:

- Classification
- Random Forest
- □Decision tree
- □Regression

Regression is a phenomenon that helps people identify patterns and calculates the likelihood of unforeseen outcomes.The system is able to recognise numbers, their values, and the grouping of numbers to represent width, height, etc. The supervised machine learning algorithms are as follows:

- Linear Regression
- □Logistical Regression
- □Support Vector Machines (SVM)
- □Neural Networks
- □Random Forest
- □Gradient Boosted Trees
- □Decision Trees
- □Naive Bayes

B. Unsupervised Learning

Unsupervised learning is defined as learning without instructor supervision, where no teacher is providing direction. When a dataset is given, unsupervised learning automatically analyses it to uncover patterns and relationships between them. Then, in accordance with these relationships, it classifies fresh data and stores it in one of those associations. Unsupervised learning is predicated on the idea of being "self sufficient."

Assume, for instance, that there are combinations of the fruits mango, banana, and apple. Using unsupervised learning, it will classify the fruits into three distinct clusters based on their relationships with one another, and it will automatically transmit fresh data to one of the clusters.

Mango, banana, and apple are mentioned in supervisor learning, however unsupervised learning mentioned three distinct clusters. The procedure of unsupervised algorithms is as follows:

- Dimensionality
- Clustering

There are following unsupervised machine learning algorithms:

- t-SNE
- k-means clustering
- PCA

C. Reinforcement

The ability of an agent to interact with the environment and ascertain the result is known as reinforced learning. It is built on the "hit and trial" theory. Each agent receives positive and negative points in reinforced learning, and on the basis of the positive points, reinforced learning produces the dataset output that was trained on the basis of the positive awards, and on the basis of this training, performs the dataset testing.

III. RELATED WORK

One of the main reasons researchers are working on this is because the heart is a basic organ of the human body. It plays a crucial role in the blood pumping process, which is as necessary to the body as oxygen is. Consequently, many researchers are working on it. Analysis of heart-related issues is always necessary, whether for diagnosis, prognosis, or prevention of heart disease. This work was influenced by a number of domains, including artificial intelligence, machine learning, and data mining.

Any algorithm's performance is dependent on the dataset's bias and variance[4]. According to research by Himanshu et al.[4] on machine learning for heart disease prediction, naïve bayes perform better with low variance and high biasness than high variance and low biasness, which is knn.

Low biasness and high variance cause KNN to suffer from the issue of over fitting, which is why KNN performance degrades. There are several benefits to utilising low variance and high biasness because it takes less time to train and test algorithms due to the small size of the dataset, but there are also some drawbacks. As the dataset size grows, asymptotic errors appear. In these situations, low biasness, low variance based algorithms perform effectively. One nonparametric machine learning approach is the decision tree, however as we all know, overfitting is a problem that can be resolved using other overfitting removal methods. Support vector machines, which have an algebraic and statics foundation, build linear separable n-dimensional hyperplanes to classify datasets.

Because the nature of the heart is complex, it must be handled cautiously in order to prevent mortality. Heart disease severity is categorised using a variety of techniques, including knn, decision trees, generic algorithms, and naive bayes [3]. According to Mohan et al.[3], you may combine two separate ways to create a single strategy called a hybrid approach, which has the highest accuracy of all other approaches at 88.4%.

Data mining has been used by some researchers to make predictions about cardiac disorders. In their research, Kaur et al.[6] explain how the intriguing pattern and knowledge are gleaned from the sizable dataset. They compare the accuracy of different machine learning and data mining methods to determine which is the best one, and the results are in favour of svm.

SVM was shown to be the best among the machine learning and data mining algorithms developed by Kumar et al. [5]; other algorithms included naivy bayes, knn, and decision tree. These algorithms were trained using the UCI machine learning dataset, which contains 303 samples and 14 input features.

Using CAD technology, Gavhane et al.[1] have developed a multi-layer perceptron model for the prediction of human cardiac illnesses and the accuracy of the method. If more people use prediction systems to diagnose their illnesses, then more people will be aware of the ailments, which will lower the death rate for cardiac patients.

One or two illness prediction algorithms are the work of some researchers. In their project, Krishnan et al.[2] shown that decision trees are more accurate than the naive bayes classification algorithm.

Many academics, including Kohali et al., are working on machine learning algorithms for disease prediction of all kinds. [7] research on predicting heart disease using logistic regression, diabetes using support vector machines, and breast cancer using Adaboot classifier came to the conclusion that the accuracy of the logistic regression was 87.1 percent, the accuracy of the support vector machines was 85.71 percent, and the accuracy of the Adaboot classifier was up to 98.57 percent, which is good from a prediction standpoint.

DOI: 10.17148/IJARCCE.2022.11696

A survey report on the prediction of cardiac illnesses has demonstrated that hybridization performs well and provides better prediction accuracy than the older machine learning algorithms[8].

IV. METHODOLOGY OF SYSTEM

The first step in system processing is data gathering, for which we use the dataset from the UCI repository that has been thoroughly confirmed by numerous researchers and UCI authorities [15].

A. Data Collection

Data collection and selection of the training and testing datasets are the first steps in the prediction system process. In this project, we used 37% of the dataset for system testing and 73% of the dataset for system training.

B. Attribute Selection

Dataset attributes are characteristics of the dataset that are utilised for systems, and for the heart many attributes are like heart bit rate of the individual, gender, age, and many more displayed in TABLE.1 for prediction system.

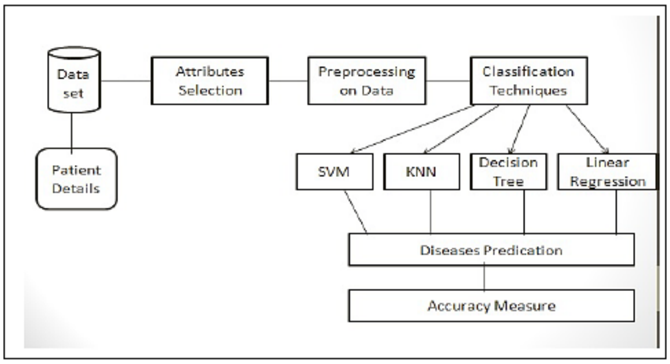
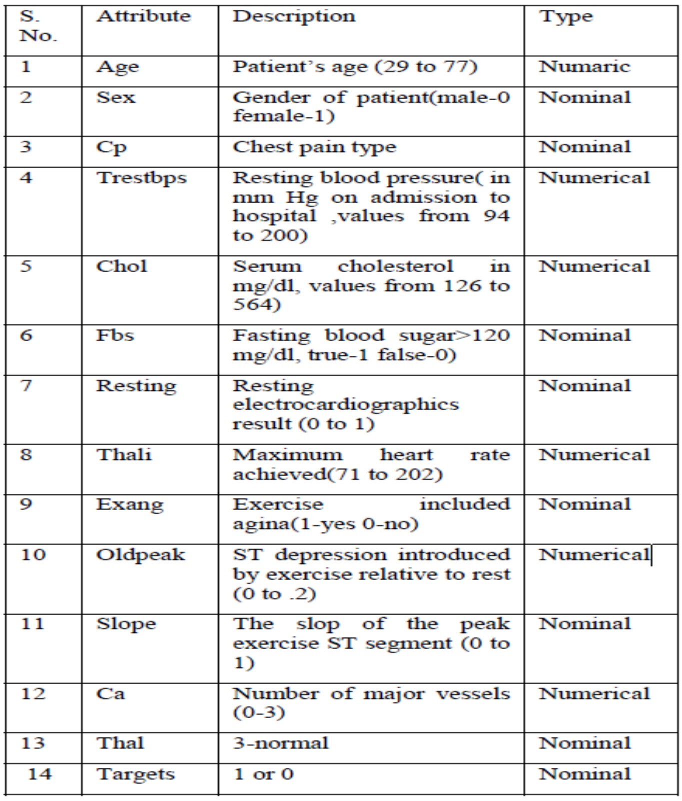


Fig.2 Architecture of Prediction System

TABLE.1 Attributes of the Dataset



S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numaric
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Resting	Resting electrocardiographics result (0 to 1)	Nominal
8	Thali	Maximum heart rate achieved(71 to 202)	Numerical
9	Exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	Thal	3-normal	Nominal
14	Targets	1 or 0	Nominal

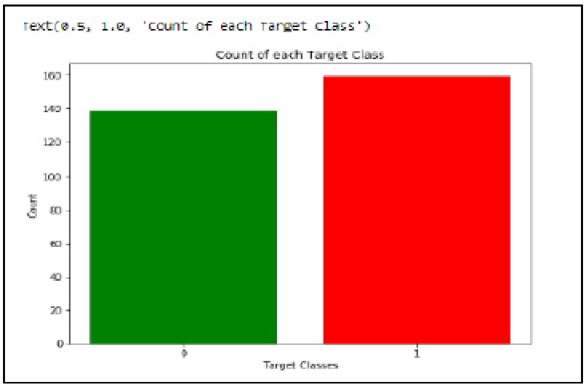
DOI: 10.17148/IJARCCE.2022.11696

C. Preprocessing of data

Preprocessing is required for the machine learning algorithms to produce prestigious results. For instance, the Random Forest technique does not allow datasets with null values, so we must manage null values in the original raw data.For our project, we must use the following code to transform some categorised values into dummy values in the form of "0" and "1":

D. Data Balancing

Since the data balancing graph shows that both the target classes are equal, data balancing is crucial for accurate results. The target classes are shown in Fig. 3 with "0" denoting patients with heart disease and "1" denoting patients without heart disease.



```
Text(0.5, 1.0, 'Count of each Target Class')
```

Fig.3 Target class view

E. Histogram of attributes

Histogram of attributes shows the range of dataset attributes and code which is used to create it.dataset.hist()

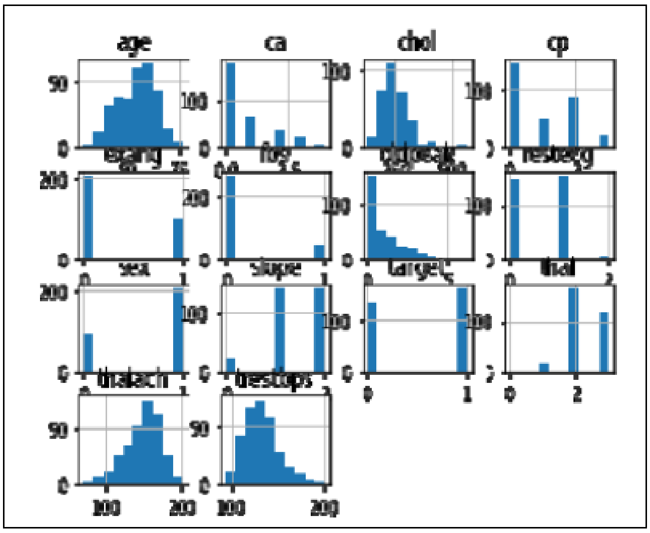


Fig.4 Histogram of attributes

V. MACHINE LEARNING ALGORITHMS

A. Linear regression

The supervised learning method is what it is. It is based on the relationship between independent and dependent variables, as seen in Fig. 5. The independent and dependent variables "x" and "y" are demonstrated to be related by a line equation, which is linear in nature, which is why this method is known as linear regression.

DOI: 10.17148/IJARCCE.2022.11696

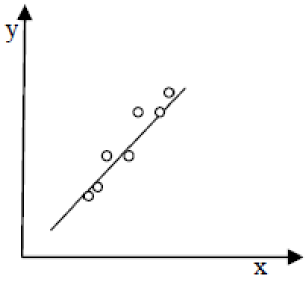


Fig.5 relation between x and y

As shown in Fig. 5, it provides a relation equation to forecast the value of a dependent variable, "y," based on the value of an independent variable, "x." As a result, it is determined that the linear regression approach provides a linear relationship between x(input) and y. (output).

B. Decision tree

On the other hand decision tree is the graphical representation of the data and it is also the kind of supervised machine learning algorithms.

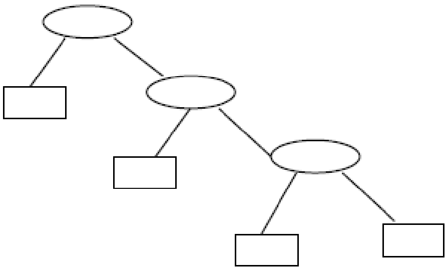


Fig.6 Decision tree

For the tree construction we use entropy of the data attributes and on the basis of attribute root and other nodes are drawn.

$$\text{Entropy} = -\Sigma\ P_{ij} \log P_{ij} \qquad (1)$$

The probability of the node, Pij, is used in the entropy (1) equation above to determine the entropy of each node. The root node of the tree is chosen as the node with the highest computed entropy, and this procedure is repeated until all of the tree's nodes have been calculated or the tree has been built.

One of the reasons why decision trees are less accurate than linear regression is because they overfit when there is an imbalance in the number of nodes, which is bad for calculations.

C. Support Vector Machine

It is a type of machine learning technique that relies on the idea of a hyperplan, which classifies the data by building a hyperplan between it.

(Yi, Xi) is the training sample dataset where i=1, 2, 3,..., n and Xi is the ith vector whereas Yi is the target vector The quantity of hyperplanes determines the kind of support vector; for instance, if a line is employed as a hyperplane, the technique is known as a linear support vector.

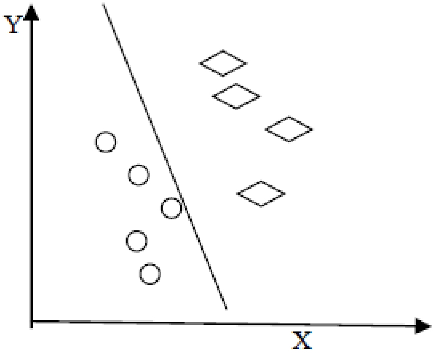


Fig.7 Linear Regression

DOI: 10.17148/IJARCCE.2022.11696

D. K-nearest Neighbour

It classifies different types of data with each other based on the distance between the locations where the data are located. The user determines the number of neighbours for each other's data sets, which is a very important factor in the analysis of the dataset.

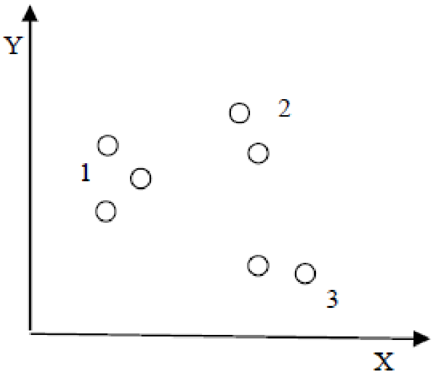


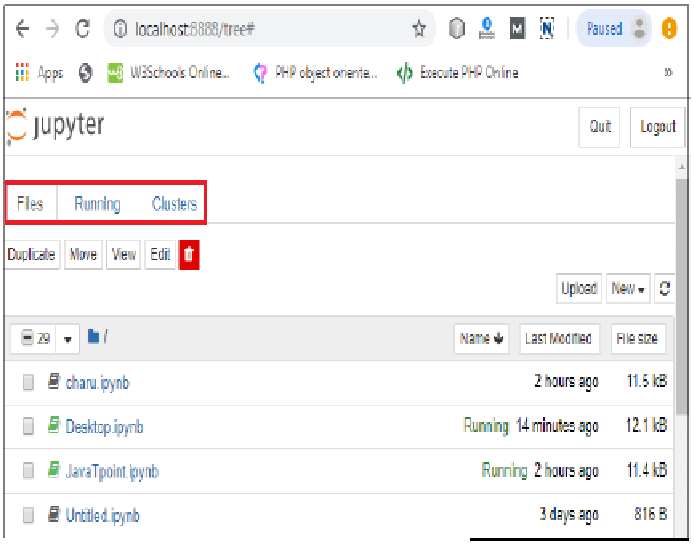
Fig.8 KNN where k=3

There are three neighbours in the aforementioned Fig., which indicates that there are three separate types of data present. Each cluster is represented in two dimensions by coordinates (Xi,Yi), where Xi is the x-axis, Y is the y-axis, and I is 1, 2, 3,..., n.

VI. RESULT ANALYSIS

A. About Jupytor Notebook

Jupiter notebook is a convenient tool for python programming projects and is used as a simulation tool. Jupytor notebook includes code as well as rich text features including equations, links, and many other types of data. These documents are the ideal place to combine an analysis description and its results since they include rich text elements with code, and they also enable real-time data analysis.A web-based interactive tool for creating images, maps, charts, visualisations, and narrative prose is called Jupyter Notebook.



Name		Last Modified	File size
charu.ipynb		2 hours ago	11.5 kB
Desktop.ipynb	Running	14 minutes ago	12.1 kB
JavaTpoint.ipynb	Running	2 hours ago	11.4 kB
Untitled.ipynb		3 days ago	816 B

Fig.9 Jupyter Notebook

B. Accuracy calculation

Accuracy of the algorithms are depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).

Accuracy= (FN+TP) / (TP+FP+TN+FN) (2)

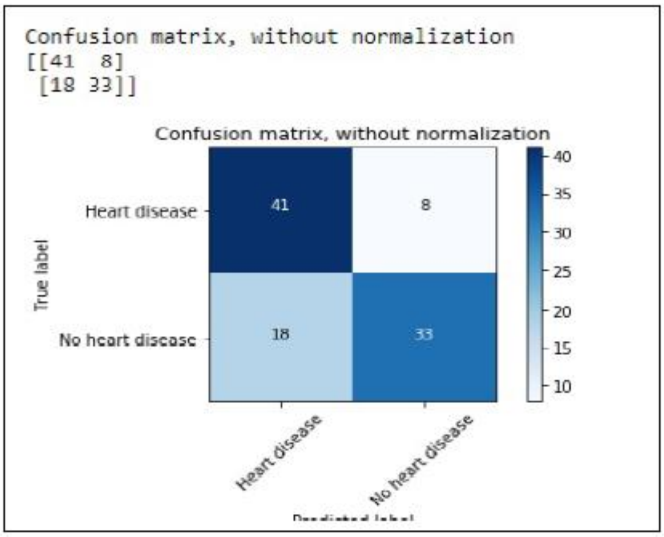
The numerical value of TP, FP, TN, FN defines as:

TP= Number of person with heart diseases

TN= Number of person with heart diseases and no heart Diseases

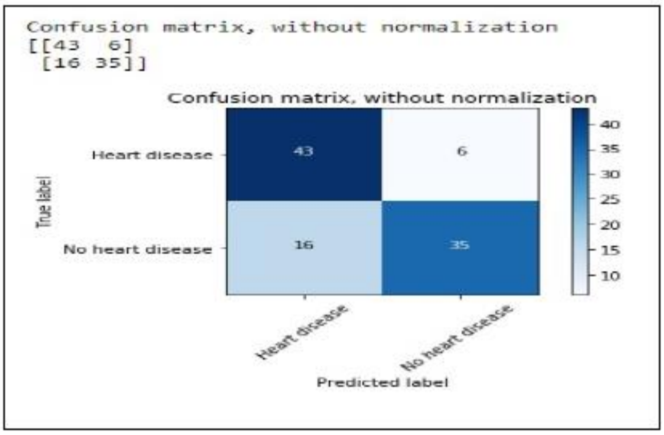
FP= Number of person with no heart diseases

FN= Number of person with no heart diseases and with heart Diseases



```
Confusion matrix, without normalization
[[41  8]
 [18 33]]
```

Fig.10 Confusion matrix for Decision tree



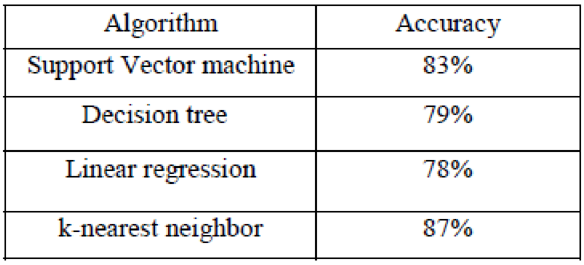
```
Confusion matrix, without normalization
[[43  6]
 [16 35]]
```

Fig.11 Confusion Matrix for linear regression

C. Result

Following the testing and training phases of the machine learning approach, we discover that the accuracy of the knn is significantly more effective than that of other algorithms. As shown in Figs. 6 and 7, where the number of counts for TP, TN, FP, and FN are given, accuracy should be calculated with the help of the confusion matrix of each algorithm. Using equation (2) of accuracy, value has been calculated, and it is concluded that knn is the best among them with 87 percent accuracy. The comparison is shown in TABLE 2.

TABLE.2 Accuracy comparison



Algorithm	Accuracy
Support Vector machine	83%
Decision tree	79%
Linear regression	78%
k-nearest neighbor	87%

VII. CONCLUSION AND FUTURE SCOPE

Since the human heart is one of the body's most significant organs and heart disease prediction is a major human concern, algorithm accuracy is one of the factors considered when evaluating an algorithm's performance.The dataset utilised for both training and testing purposes affects how accurate machine learning algorithms are. KNN is the best method, according to our examination of the algorithms using the dataset whose attributes are presented in TABLE.1 and the confusion matrix.

We find that the accuracy of the knn is much higher than that of other algorithms after the testing and training phases of the machine learning approach.

REFERENCES

1. Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine Learning Algorithms” ICIICT, 2019.

2. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease using Machine Learning”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

3. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.

4. Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

5. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

6. Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

7. Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

8. M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.

9. S. Kumra, R. Saxena, and S. Mehta, “An Extensive Review on Swarm Robotics,” pp. 140–145, 2009.

10. Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “ A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” Advances in Computational Sciences and Technology , 2017.

11. Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique” Journals of Computer Science & Electronics , 2016.

DOI: 10.17148/IJARCCE.2022.11696

12. Chavan Patil, A.B. and Sonawane, P.“To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients” International Journal on Emerging Trends in Technology, 2017.

13. V. Kirubha and S. M. Priya, “Survey on Data Mining Algorithms in Disease Prediction,” vol. 38, no. 3, pp. 124–128, 2016.

14. M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, “Prediction of risk score for heart disease using associative classification and hybrid feature subset selection,” Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.

15. https://archive.ics.uci.edu/ml/datasets/Heart+Disease