WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Prediction by Machine Learning

Harshita Bishnoi[1] , Saranya Raj[2]

1"Computer Science & Engineering" "KCC Institute of Technology & Management" Greater Noida, UP,201306
harshitabishnoi0209@gmail.com
2"Department of Computer Science & Engineering" "KCC Institute of Technology & Management" Greater Noida, UP,201306
saranyaraj777@gmail.com

ABSTRACT –

These days life continues so fast that people forget how health is important. Due to the fast pace, health is neglected. Because of this busy lifestyle, people are unable to focus on their health. Because of which an increase in diseases is being seen. As per the WHO data, cardiovascular diseases are responsible for about 31% of deaths occurring globally. Due to this busy lifestyle and unawareness regarding health, people are falling sick more often. But it becomes difficult to do analysis manually as health care organizations generate huge amounts of data. In the medical field, machine learning can be useful in making disease predictions and processing data. In this paper we will study heart disease, the risk factors for heart disease and techniques of machine learning. By means of these methods, we predicted the occurrence of heart disease and did a comparison of the machine learning techniques used for the experiment. These research paper aims to enhance the heart disease prediction using machine learning and investigate the efficiencies of different algorithms used.

Keywords*:* supervise learning, unsupervise learning, reinforce method; regression; python programming, a decision trees, cardiovascular diseases (CVD).

INTRODUCTION

In our body, the heart is one of most important part complicated organs in human body, look after your heart as it works to keep the body healthy. A significant number of diseases are associated with the heart, making early predictions of heart conditions crucial. Comparative study in this field is necessary, as many patients lose their lives because their illnesses are diagnosed at an advanced stage due to the limited accuracy of medical instruments. Therefore, improving disease prediction through more efficient algorithms is a vital area of research.
Machine learning is an advanced technology that plays a crucial role in medical diagnostics by means of training and testing procedures. It refers to computer programming that makes the computer do its task without human assistance and broadly focuses on enabling machines to replicate human cognitive functions. The creation of the systems for data processing is what the term basically means. That is why the combination with AI is often called machine.
A machine algorithm are designed to learn from real-world patterns, this study utilizes biological parameters such as cholesterol levels, blood pressure, age, and gender as testing data. Based on these parameters, a comparative analysis is performed to evaluate the accuracy of different algorithms. In this area of study, four machines are learning algorithms.
The study involved four machine learning algorithm which is Tree, Regression, K-Nearest Neighbors and Support Vector Machine methods are used to identify heart diseases problems.

The four machine learning techniques outlined in this paper are used effectively. They evaluate different machine learning techniques. Section I introduces heart disease problems and machine learning. Section II discusses the classification of the machine algorithm. Section III reviews previous research in the field. Section IV outlines the method used for disease prediction. Section V explains this algorithm implemented in the study. Section VI Looks at data and the results. Lastly, Section VII presents the closing statement along with possible future studies. HealthCare remains a fundamental concern for humanity. According to the WHO, good health is a right that everyone has, and every person should have access to essential healthcare services. For regular health monitoring, Heart-related diseases are seen in nearly 31% of the world wide deaths. Early detection and treatment of cardiovascular. It remains difficult to fight illness, especially in underdeveloped countries. Lack of diagnostic facilities, medical professionals and other resources and other important things., which influence heart accuracy to diagnosis the diseases.
To address this problem, advances in computer technology and machine learning are being applied to the development of medical support systems for early heart disease detection. Identifying cardiovascular conditions in the early stages can significantly reduce mortality rates. Machine learning techniques help analyze extensive healthcare datasets, recognize patterns, and make predictions. Given that medical data is often vast and structurally complex, machine learning algorithms effectively process and extract valuable insights from such data. These algorithms are learned from historical data and make real- time predictions. Implementing this algorithm which is similar heart disease prediction system to help cardiologists in decision making, more precise diagnoses, allowing prompt treatment and possibly saving numerous lives.

MACHINE LEARNING EXPLANATION

Machine Learning is a highly advanced technology that functions on training and testing are important concepts. The systems learned by extracting patterns from data and previous experiences, enabling it to apply this acquired knowledge to testing processes based on the requirements of different algorithms.
Machine learning algorithms are categorized into three types:

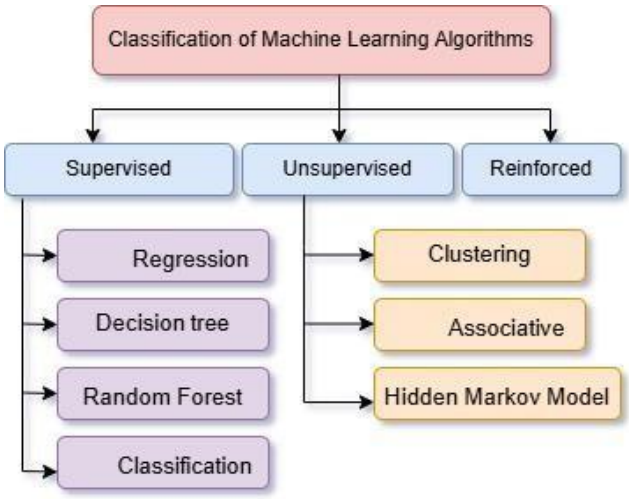


Fig.1 Classification of machine learning

Supervise Learning

Modeling would help to predict accurate choices for an event of input data and the corresponding outcome. The data is divided into training and test sets. These models are trained using labeled datasets, The training set is used to develop the type, while the accuracy is tested with new data. The dataset. It has both the model and its output. The model of this approach involves classified and regression method.

Unsupervise Learning

The training data are not labeled or classed in the dataset.
Find the hidden pattern in the data. These model learned to create a pattern. It is clearly foreseeable hidden patterns for any new input data set, but after data exploration, it makes inferences to characterize hidden patterns. With this method, the dataset shows no answers. One example of an unsupervised learning strategy is the clustering approach.

Reinforce Learning

The model learns from the experience because it doesn't use tagged databases or link outcomes to data. By evaluating and testing many possibilities, the model in this method enhances this presentation is like on its association to the environment and determines whether to explain its flaws and obtain the best results. Typical supervised learning methods for determining the likelihood of cardiovascular illness occurrence are classification algorithms.

RELATED WORK

The heart is like one of the body's primary organ and vital to the pumping of blood, which is just as vital to the body as oxygen, it is always necessary to preserve it. This is one of the reasons that researchers are doing work on this topic. Therefore, several scientists are focusing on it. Analysis of heart-related issues is constantly required, whether it helps diagnosis, prognosis as you can say, heart disease prevention.
This work involves contributions from a few disciplines, including data analysis, machine learning, and machine intelligence.

The characteristics of the data can influence how an algorithm performs and investigated in a study the machine learning for the prediction of cardiac disease. So unlike k-Nearest Neighbors (KNN), which has the high variance and low bias. And the nave baye algorithm performs effectively with the low variance and the high biased compared to KNN leading to decline in performance.
Utilizing algorithms with low variance and high bias has certain advantages, such as reduced training and testing time, particularly when working with small datasets. However, using a small dataset also has drawbacks. As dataset size increases, asymptotic errors may arise, and algorithms with both low bias and low variance tend to perform better in such cases.

One non-parametric machine learning technique is the decision- tree algorithm, which is known to be prone to overfitting. However, this issue can be addressed using specific techniques designed to mitigate overfitting. The Support Vector Machine algorithm, however, which is based on algebraic and statistical principles, constructs an n-dimensional hyperplane to classify datasets in a linearly separable manner.

The human heart is a complex organ that requires careful monitoring, as any mishandling can lead to fatal consequences. The severalty of heart disease can be classified using various techniques as the hybrid approach which is a technique that combines two distinct approaches into a single framework and becomes more efficient than other ways. The hybrid approach has accuracy of 88.4% and other ways have accuracy of 81%. This hybrid and more improvement in not only clustering but also classification. It includes Decision Trees, Genetic Algorithms, K-Nearest Neighbors (KNN), etc. Many researchers have explored data mining techniques for the prediction of heart disease.

The studies conducted in this field demonstrate how meaningful patterns and insights can be extracted from large datasets. They also performed an accurate comparison of different machine learning and data mining algorithms for analysis and research of performance of the various machine learning and data collecting techniques on the UCI Machine Learning dataset, a total of 304 samples with 15 input features were used. The findings showed that the Support Vector Machine (SVM) performed better than Naive Baye, K-Nearest Neighbors (KNN), Decision Tree, and others. Using a multi-layer perceptron model to predict heart disease.

Evaluating its accuracy with the help of CAD technology. The use of such predictive systems can increase awareness about heart disease, ultimately contributing to a reduction in mortality rates associated with heart-related conditions.

Some researchers have analyzed a limited number of algorithms for disease prediction. Krishnan et al. [2] demonstrated in their study that the Decision Tree algorithm achieved higher accuracy compared to the Naïve Bayes classification method. Machine learning techniques have been applied to predict various diseases. I studied the forecasting of breast cancer, method of detection of diabetes using Support Vector Machine (SVM) and heart disease forecast using logistic regression, use the classifier AdaBoost.

As per their findings, logistical regression has an accuracy of 87.1% Support vector machine which has also consider and it
reached 85.71%, and the AdaBoost classifier demonstrated a high accuracy of 98.57%, making it a strong candidate for predictive applications.

A survey on heart disease prediction revealed that traditional machine learning algorithms struggles to achieve high accuracy. However, hybrid approaches have shown improved predictive performance, making them a more reliable option for disease detection [8].

METHODOLOGY OF SYSTEM

The following steps outline the methodology used to develop a heart disease prediction model. This process involves data collection, preprocessing, feature selection, model training, and evaluation to accurately predict heart disease risk based on various medical attributes.

Collection of Data

The Data collected and the separated for the training and testing datasets constitute the initial stage of a prediction system. 73% of data of these projects is used for training purposes, with the remaining 37% going toward system testing. A well-balanced dataset helps improve the model's accuracy and reliability. The training data allows the model to learn patterns, while the testing data evaluates its performance on unseen cases. Proper data distribution ensures that the system can generalize well to new predictions.

Attribute Selection

The attributes of a dataset represent its key properties, which are utilized by the system. As indicated in TABLE 1 for the prediction system, a person's age, gender, heart rate, and other characteristics are all taken into consideration when predicting heart disease. These qualities are essential for spotting possible hazards and assessing trends. The quality and applicability of the chosen attributes have a significant impact on the prediction system's accuracy. Proper feature selection helps improve the model's efficiency and reduces computational complexity.

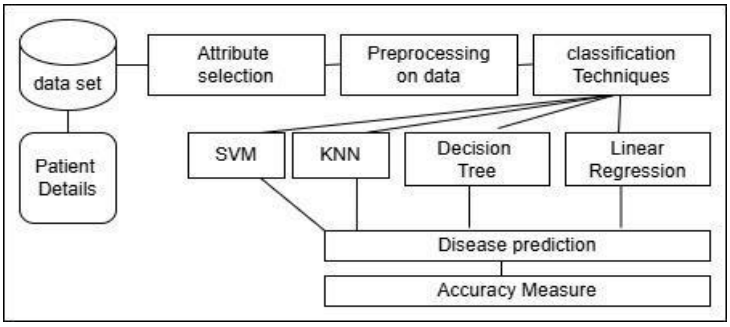
Data Processing

Preprocessing is a crucial step in ensuring accurate and reliable results from machine learning algorithms. It involves handling missing values, transforming data, and standardizing features to enhance model performance. For instance, algorithms like Random Forest do not support datasets with null values, making it necessary to manage and clean missing data. Additionally, for this project, categorical variables are converted into numerical representations, such as "0" and "1," using dummy encoding to make the data suitable for machine learning models.

Data Balancing

Data balancing is a critical process in achieving accurate and unbiased results in machine learning models. It ensures that both target classes are equally represented, preventing model bias toward the majority class. The target classes are balanced, as seen in Fig. As shown in Fig. 3, 0 is used for people with heart disease and 1 for others.

Fig.2 Architecture of Prediction System



S N	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numerical
2	Sex	Gender of patient(male-0 female-1)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbp	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

TABLE.1 Attributes of the Dataset

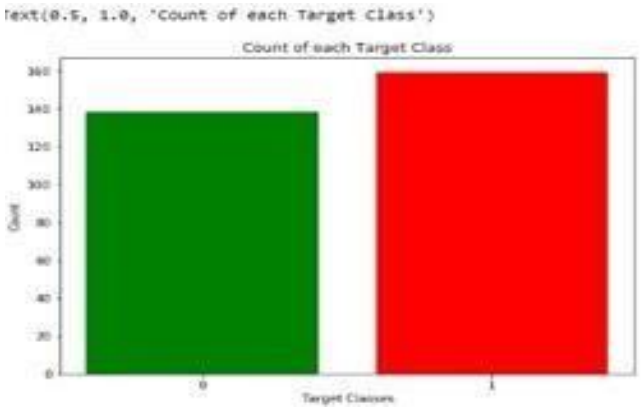


Fig.3 Target class view

ALGORITHMS USED IN ML

Linear Regression Algorithm

A Linear regression is a supervised learning technique for finding a relationship between an independent and dependent variable. Fig. In 5 "x" means the independent variable, while "y" is the dependent variable, with their relationship expressed through a linear equation. This approach is termed linear regression because it models a straight-line relationship between input (x) and output (y). By using this equation, the model can predict the value of "y" based on a given "x," making it a fundamental technique for analyzing and forecasting trends in data.

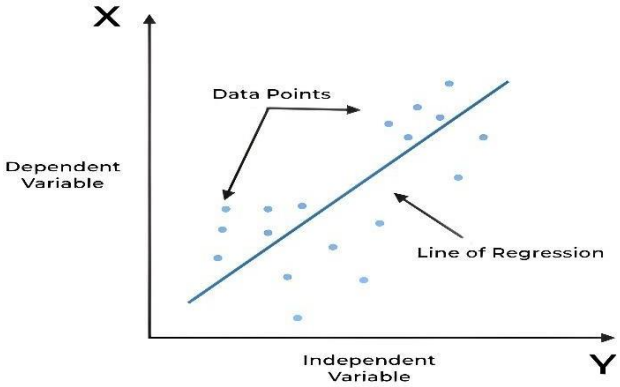


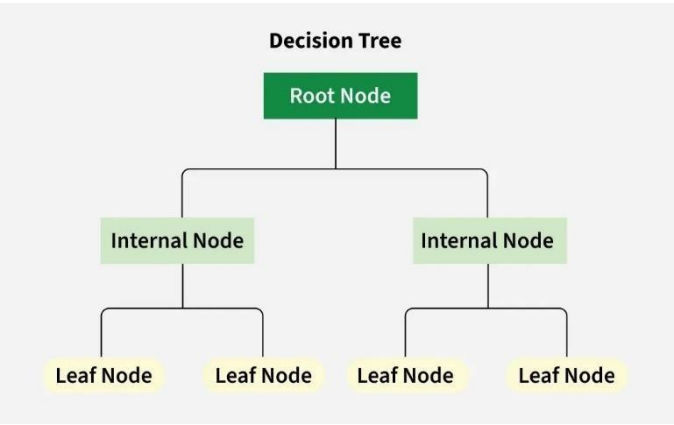
Fig.5 Linear Regression

According to Fig. The linear regression approach gives a formula an equation or relation to forecast the value of "y" a dependent variable established on the value of "x" an independent variables and then the straight relationship between input (x) and output (y).

Decision Tree Method

A Decision Tree is a supervise method of machine learning to displays data in a tree framework, which is hierarchical.it segregates data according to the feature values in a branch-like structure for classification and regression The construction of the tree is guided by entropy and information gain, determining the root node and subsequent decision nodes. This method helps in making clear, interpretable decisions by breaking down complex data into simpler parts. Additionally, decision trees handle both numerical and categorical data effectively, making them a widely used technique in predictive modeling.

Fig.6 Decision tree



Entropy(s)= -∑ Pij log Pij (1)

In the given entropy equation (1), PijP_{ij}Pij represents the probability of a node, It is employed to determine each node's entropy. The node having maximum entropy is selected as the root node in this algorithm. The calculations are repeated until all nodes are processed, or the tree is fully constructed.

Support Vector Machine Algorithm

The Support Vector Machine method is a type of a machine learning method which classifies the data by a hyperplane. It works by determining the best border to maximize the distance between the data points for different categories.. A training sample dataset is represented as (Yi,Xi)(Y_i, X_i)(Yi,Xi), where i=1,2,3,…,ni = 1,2,3, \dots, ni=1,2,3,…,n, with XiX_iXi as the input feature vector and YiY_iYi as the target label. The number and type of hyperplanes determine the classification approach; for instance The technique is known as Linear SVM if a straight line is utilized as the hyperplane.

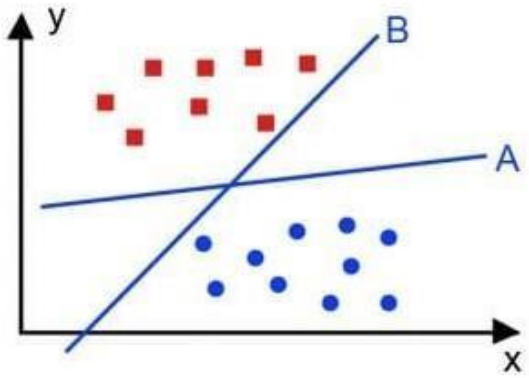


Fig.7 Support Vector machine

K-nearest Neighbor Algorithm

In the K-Nearest Neighbor is a supervised method of machine learning algorithm which can usually applied to the classification or regression problem in KNN, if target value of the data point is not present, it finds the nearest data points in the training set and gives them an average value of the recognized data points. Regression assesses the meaning of K labels, whereas Classification predicts or returns the most common value of K labels.

This method is employed for classification when prior knowledge of data is not available. To determine the nearest data or distance to a point, one can use distance metrics like Manhattan or Euclidean distances. If you have a very large data set and high-noise data, it can still perform better and yield superior predictions.

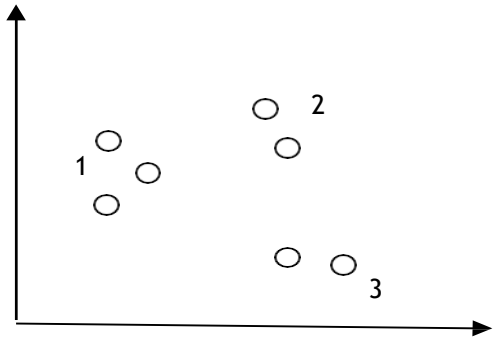


Fig 8 KNN

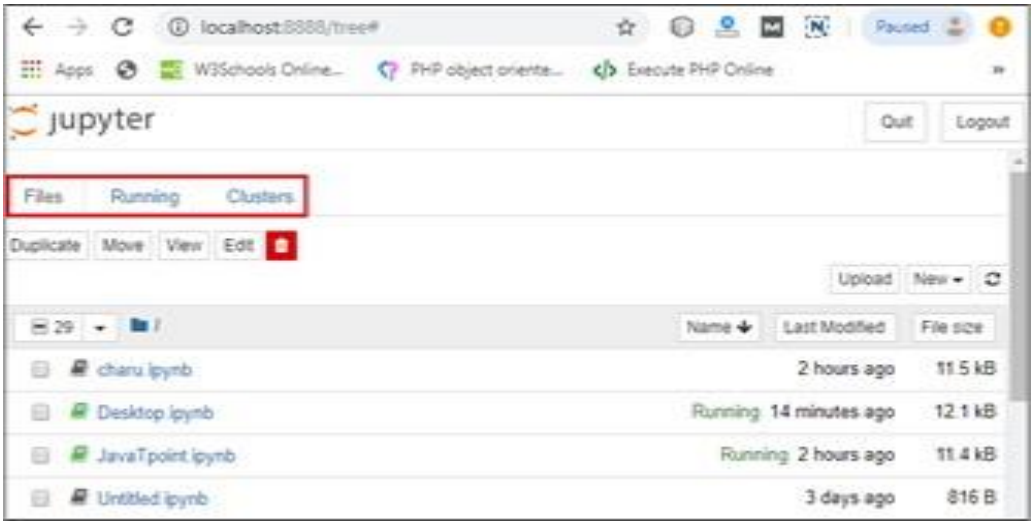
The above figure, k=3k=3k=3 indicates that three neighbors are considered, meaning the data is classified into three different groups. Each cluster is represented in a two- dimensional space, where the coordinates are denoted as (Xi,Yi)(X_i, Y_i)(Xi,Yi). Here, XiX_iXi corresponds to the x- axis, YiY_iYi to the y-axis, and i=1,2, 3,,ni = 1,2,3, \dots, ni=1,2,3,…,n represents the data points within the clusters.

RESULT AND SUMMARY

Jupiter Notebook

A common simulation tool that works good with Python programming assignment is Jupiter Notebook. Many text elements are supported, as well as Figures, equations, links, and other executable code. The combined elements make Jupiter Notebook an ideal platform for integrating analysis

descriptions, executing real-time data analysis, and presenting results in an interactive format. As an open-source, web-based tool, it enables users to



create visualizations, plots, maps, and narrative text, making it a powerful resource for data science and machine learning applications.

Fig 9 Jupiter Notebook

Precision Calculation

Factual Description of TP, FP, TN, FN Measurement Metrics is very important. As per the name, True Positive is expected to be the right one. These values play a crucial role in evaluating the performance of a model by determining how well it correctly classifies data and minimizes errors.

(FN+TP) / (TP+FP+TN+FN) is the accuracy.
TP, FP, TN and FN are the numeric values that have the following meanings
TP= Total number of heart patients.
TN refers to the total number of non-cases.
FP = Total number of heart disease-free individuals.
The FN refers to the number of people with and without heart disease.

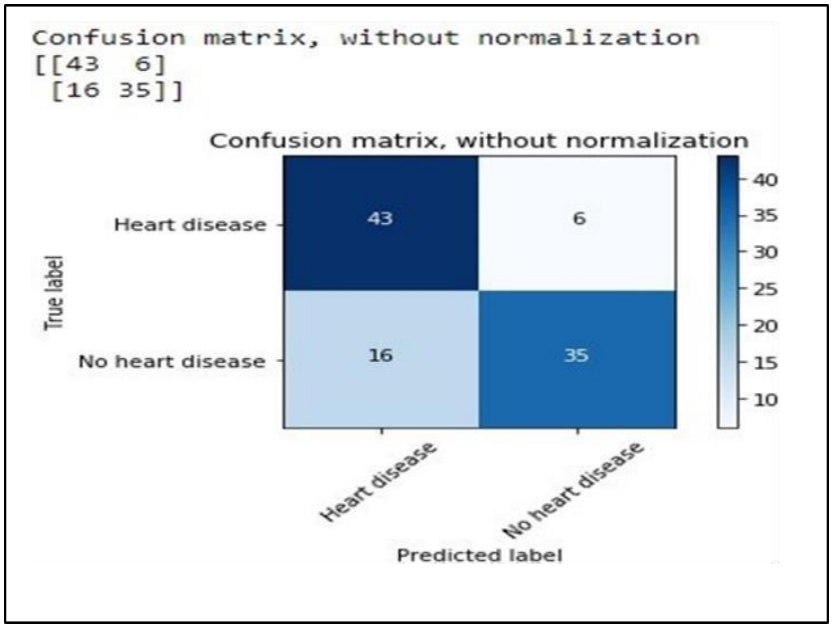


Fig.10 Confusions Matrix for linear regression

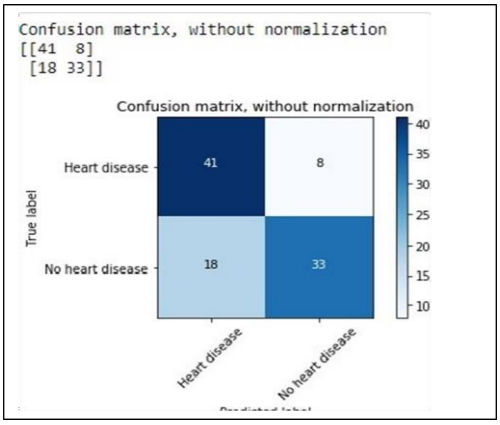


Fig.11 Confusions matrix for Decision tree

Accuracy Comparison

The assessment of several algorithms according to how well they classify data. Key metrics including the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are useful to measure it. by evaluating the accuracy of several different methods.

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

Table 2 Accuracy Comparison

RESULT & OUTPUT

After training and testing the machine learning technique, the accuracy of K-Nearest Neighbors (KNN) algorithm is most effective than other algorithms. In the above Figs. In 6 and 7, to calculate accuracy, of the True Positive, True Negative , False Positive, and False Negative counts are calculated through confusion matrix of each algorithm. Equation is used to calculate accuracy. The results confirm that KNN outperforms other models, achieving an accuracy of 87%. The comparative analysis is presented accordingly.

FUTURE SCOPE AND CONCLUSION

Predication of heart illness is an important part of medical care. The accuracy of machine learning algorithms plays a key role in assessing their performance, as it directly impacts the reliability of disease prediction. The accuracy of these algorithms depends on the dataset used for training and testing. After comparing several algorithms and using the dataset parameters shown in table and assessed using a confusion matrix, KNN was found to be the best and successful model.

There could be new advances in the future which make use of more machine learning techniques to better forecast heart disease. Early detection methods and improved algorithms can help minimize death rates by increasing awareness and enabling timely medical intervention. Additionally, integrating deep learning models and hybrid approaches may further refine the predictive capabilities of these systems. The use of real-time patient data and advanced medical imaging techniques could provide more precise and dynamic results. Furthermore, collaboration between healthcare professionals and data scientists can lead to the development of more robust and efficient diagnostic tools.Wearing high-tech equipment that keeps a steady watch on heart health can also avoid and diagnose the disease early.

REFERENCES

[1] Technique" Journals of Computer Science & Electronics , 2016.

[1] 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.

[2] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.

[3] 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.

[4] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

[5] 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.

[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

[12] Wolgast G, Ehrenborg C, Israelsson A, Helander J,

Johansson E & Manefjord H(2016). Wireless

[13] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device - kinect. International Journal of Scientific and Research Publications, 4(1), 1-4

[14] Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office

[15] M.Satish, D Sridhar, "Prediction of Heart Disease in Data Mining Technique", International Journal of Computer Trends & Technology (IJCTT), 2015.

[16] Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.