RESEARCH ARTICLE OPEN ACCESS

Performance Analysis of Heart Disease Diagnosis Using Machine Learning Algorithm

G. Jaya Sree [1], Ch. Durga Prasanna [2], B. Yogya Rachana [3], B. Anusha [4], N Md Jubair Basha [5]

[1], [3], [3], [4], [5] Dept. of CSE, Kallam Haranahareddy Institute of Technology, Andhra Pradesh - India

ABSTRACT

Many people across the globe have died from heart disease in the past decade. About every 60 seconds in India, one person dies of heart disease. For a significant reduction in fatalities due to heart disease, a reliable and quick method of detecting this illness is needed.This paper analyses the detection of heart disease using machine learning algorithms and python programming. Over the post decades, heart disease is common and dangerous disease caused by fat containment. This disease occurs due to over pressure in the human body. Using different types of parameters in the dataset we can predict the cardiac-disease. In this paper it is observed a dataset consists of 12 parameters and 70000 individual data values to analyse the performance of patients. The main objective of this paper is to get a better accuracy to detect the heart-disease using algorithms in which the target output counts that a person having heart disease or not.

I. INTRODUCTION

The basic concept is to use machine learning methods to aid in the detection of cardiac disease. The human heart is the human body's most important component. To put it simply, it controls the movement of blood through our body. Any abnormality to the heart may result in other sections of the body's organs or systems showing signs of discomfort. The study classifies any kind of disruption to the regular operation of the heart as a Heart illness. Heart disease is one of the leading causes of mortality in the modern world. Unhealthy lifestyle, including smoking, drinking, and an excessive consumption of fat, may lead to heart disease[2]. Heart disease claims the lives of approximately 10 million people each year across the globe, according to the World Health Organization. Early diagnosis and a healthy lifestyle are the only methods to avoid the kinds of heart problems that can be prevented[6].There is one major obstacle today in the healthcare system: providing high-quality services while accurately diagnosing the illness[1]. Although cardiac problems are now considered the primary cause of mortality in the world, they may also be treated and controlled. It is all about the right timing of discovery of the illness.In order to prevent terrible outcomes, the planned study is designed to identify certain cardiac conditions at an early stage. Medical professionals have generated enormous records of medical data, which you may use to study and extract important information from. Data mining methods are used to locate important and previously hidden information within huge quantities of data. For the most part, the medical database is made up of discrete data. So, because of this, decisions that use discrete data will be difficult. Machine learning (ML), a subset of data mining, can effectively handle large size datasets with well-formatted data. Machine learning is utilised in diagnostics, detection, and prediction of different illnesses in the medical sector. The primary objective is to offer a diagnostic tool for physicians to locate early-stage cardiac disease[5]. In turn, this will assist in treating patients while avoiding harsh repercussions. While ML can find the discrete patterns that are buried in the data, it also analyses the data to find these patterns. ML methods assist in the prediction and early detection of cardiac disease[9].Naive Bayes, ANN, KNN,Decision Tree, Logistic Regression, and Support Vector Machines are all used to predict heart disease, especially in the early stages[3].

II. RELATED WORKS

According to a literature review, a vast array of diseases linked to the heart may be referred to as heart disease. Abnormal disorders that directly affect the heart and all its components are medical problems of this kind. Heart disease is a significant health issue in the current environment. The goal of this research is to examine the many Machine Learning (ML) methods developed in the last few years and assess their potential for heart disease prediction. Data Mining has been used before in the construction of these works. One of these study papers suggests that Shadab et al, Carlos et al, and the like utilised just one method of data mining for heart disease detection, while, in the other paper, at least two data mining methods are used. BoualiH,AkaichiJ[10] various methods have been suggested to use machine learning algorithms to predict cardiac disease. Studies examine a variety of different ML algorithms, all of which are appropriate for classifying heart disease. Researchers conducted an investigation to analyse Decision Tree, KNN, and K-Means algorithms and studied how accurate each of them are at classifying data[8].This study shows that decision tree accuracy was best when it was inferred that future optimization of various methods and parameter settings might make it more efficient.[7]

T. Nagamani, et al.[2] have developed a method that integrates data mining with Map Reduce. This method produced a higher accuracy than the traditional fuzzy artificial neural network in identifying 45 out of 45 cases of testing. Using dynamic schema and linear scaling, the algorithm's accuracy was enhanced. ML model Fahd Saleh Alotaibi has developed has been intended to test the merits of five alternative algorithms[3]. Higher accuracy was obtained with the usage of the Rapid Miner tool, which led to a better outcome when compared to Matlab and Weka. A comparison was made of the accuracy of the following machine learning classifiers: Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM. The decision tree algorithm performed best.

Anjan Nikhil Repaka[4], hereafter abbreviated as ANR, developed a system in which the following methods are used: NBI (Naive Bayesian Inferences) for classifying datasets, and AES (Advanced Encryption Standard) for encrypting sensitive data transfers.

III. METHODOLOGY AND ANALYSIS

A. Data collection

Overall process of predicting heart disease carries following procedure:

We have collected data from dataset provider –Kaggle.com. the dataset which is published by Svetlana Ulianova as in the title of Cardiovascular Disease dataset, 2019.The dataset collected consists of 70,000 records of patients data carries 11 features and

Dataset is the information or a tool essential to do any kind of research or a project

B. Data Preprocessing

Segregation of target data and feature data as training and test data.

Scaling the values in the data to be values between 0 and 1 in which and scale all the values before training the Machine Learning models.

C. Applying Algorithms

Comparing 6-machine learning algorithms such as SVM, Decision tree,ANN , Naïve bayes,Logistic regression and K-nearest neighbor to get the better accuracy to which highest parameter may cause disease.

For each algorithm, there is a pseudo code helpful to develop any kind of programming language. In python, there is a simple way to establish any kind of algorithm in which simple and short code easier to predict accuracy.

ii.

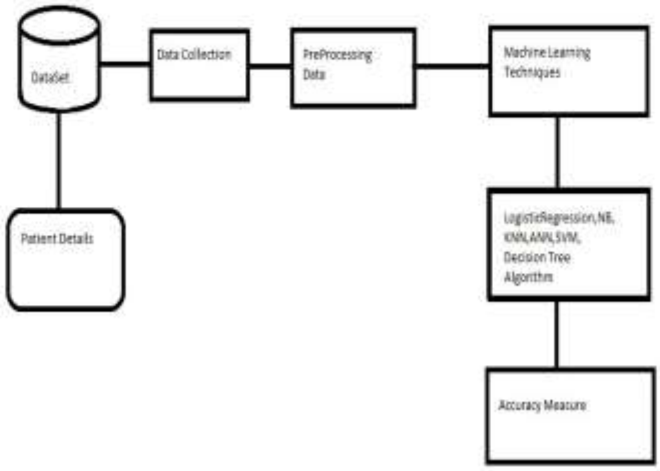

Fig. 1A model for Heart Disease Diagnosis

IV. IMPLEMENTATION

The algorithms used in this paper is highly helpful to predict the accurate result to detect heart disease in which factors that cause a disease can be detected. The following algorithms have built in this paper.

A. Decision Tree:

Decision Tree is used mostly for addressing Classification issues, although it may also be utilised for Regression analyses. The tree-structured classifier divides the dataset into internal nodes, with feature nodes as the divisions, branch nodes as the decision rules, and leaf nodes as the conclusions. The two nodes that may be found in a Decision tree are the Decision Node and the Leaf Node. Choice nodes are used to make any decision, and their outputs have no further branches. Leaf nodes, on the other hand, are only an output of decisions, and do not branch further. Features of the dataset are used while making choices or carrying out a test. An alternative graphical method to obtaining the potential answers to a problem or choice given the circumstances is called a scatterplot. Decision trees are often known as "tree-like structures" since they start with the root node, then branch out, resulting in a tree-like structure. Classification and Regression Tree (CART) is the method we utilise to construct a tree. A decision tree is an algorithm that has two stages: first, it asks a question, and second, depending on the response (Yes/No), it continues to divide the tree into subtrees. The image to the right demonstrates the basic structural outline of a decision tree:

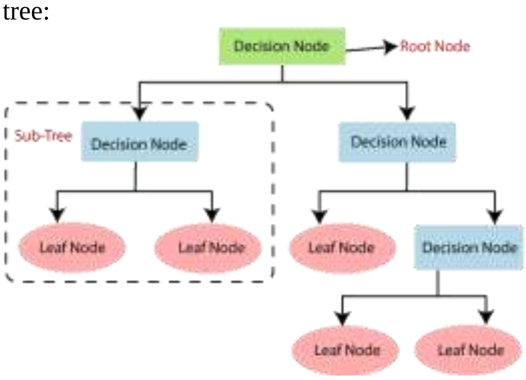

Fig.2 Decision Tree Process

B. Support Vector Machines (SVM):

This classification algorithm is known as the Support Vector Machine (SVM). SVM is a supervised classification technique that relies on vectors for supervised training. This is useful for numeric prediction, as well as categorization. The state-of-the-art SVM application has found its way into a variety of industries (i.e., object recognition, speaker identification, and hand-written digit recognition). Despite its excellent accuracy, SVM training is rather slow. This is because it doesn't overfit like many other classifiers. This classifier separates the training data into hyperplanes since the data in its original dimension cannot be easily separated. In order to separate the data more effectively, SVM utilises non-linear mapping, thereby creating hyperplanes. The finest separating hyperplane is then searched for after that. The specifics of this method may be found in the literature, books on data mining, and in a variety of other publications.

C. K-Nearest Neighbour:

Lazy learners, in contrast to eager classifiers like Rule-Based, Decision Tree, and SVM, follow a pattern of requiring a large amount of training data before beginning the learning process. In this scenario, the KNN is an ineffective learner since it is waiting for the test set to become available instead of acting on the training set. Generalization isn't built; only training tuples or instances are stored. KNN calculates distance between a given sample and training samples by measuring how similar the training samples are to the sample. In this way, the KNN approach is built on analogies. For a given unknown test tuple, KNN identifies the k tuples that are closest to the test tuple, and these tuples are the K-nearest neighbours. Euclidean distance is a good way to quantify the similarities between two datasets.

To introduce tuples, we may say "Let two tuples x1 = (x11, x12, …, x1n) and x2 = (x21, x22, …, x2n)."

To calculate the Euclidean distance (dist) between x1 and x2, equation (1) is used. (1 + 2) (1-2) Two plus two equals four (1)

Experimentally, we set k equal to one and increase by one until we find the lowest error rate.

D. Logistic Regression:

Logistic regression is a common supervised learning method, which falls within the subcategory of machine learning called supervised learning. Prediction of the categorical dependent variable using a specific collection of independent factors is a task performed by the model. Categorical dependent variables may be predicted using logistic regression. To arrive to a categorical or discrete value, the result must be categorical or discrete. It may be 0 or 1, or true or false, or many other things. However, it provides probabilities instead of precise numbers like 0 and 1.

Logistic Regression is just slightly different from the Linear Regression. They are often used in different ways. Logistic regression is utilised for addressing classification issues, while the application of logistic regression to solve regression problems is known as linear regression.

The "S" shaped logistic function may predict two maximum values instead of fitting a regression line (0 or 1).

The logistic curve predicts what percentage of the population a certain percentage of the population falls into, and is therefore used in applications such as determining whether or not a certain cell or population is malignant, a mouse is fat, etc.

Logistic regression is a notable machine learning technique because it can utilise continuous and discrete datasets to produce probabilities and identify new information.

InLogistic Regression, various kinds of data may be utilised to categorise the observations, and determining the best variables is quite simple.

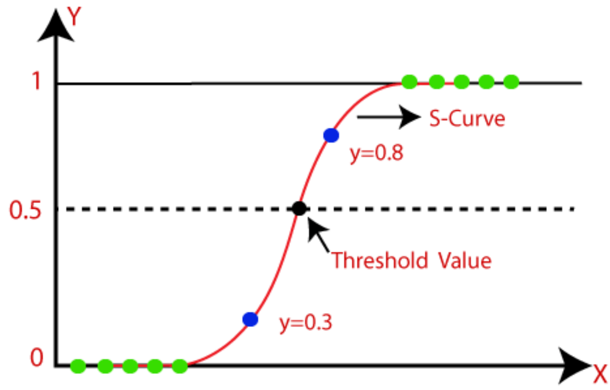

Fig. 3 Logistic Regression Graph

E. ANN:

The fundamental and advanced principles of ANNs are presented in the tutorial. For novices as well as professionals, our Artificial Neural Network lesson is created. Artificial neural networks are an artificial intelligence (AI) sub-field that take their form after the human brain. Artificial neural networks are computer networks based on the structure of the human brain that is constructed using biological neural networks. Artificial neural networks are similar to human brains, in that they contain neurons coupled to each other at different levels. This is called a node. This article is an introduction to artificial neural networks, covering every facet that is relevant to this kind of network. Learn how neural networks, Adaptive resonance theory, Kohonen self-organizing map, and building blocks, among other techniques, are used in this guide.

F. Naïve Bayes Algorithm:

Naïve Bayes is a supervised learning method, which is based on Bayes' theorem and is often used for classification tasks. It is most often employed in text classification applications when the training dataset has a large number of features. Simple and effective classifier, Naive Bayes is the kind of model which helps create rapid machine learning models capable of making quick predictions. This classifier operates on the basis of probabilities. Naïve Bayes Algorithm is often used for such things as spam filtering, sentiment analysis, and article classification. Naïve Bayes gets its name from its naïveté. Naïve Bayes is a two-word name, whose first letter is the beginning of each of the two terms. Naive: Naïve refers to assumptions about characteristics that are unrelated to other features. Considerations such as colour, form, and flavour will lead to the conclusion that spherical, red, and sweet fruit are all correctly classified as apples. In other words, each characteristic is able to recognise it as an apple without needing to rely on any other features. The term "Bayes" derives from the Bayes' Theorem.

Theorem of Bayes: Prior knowledge is sometimes referred to as "prior probability," and Bayes' theorem, also known as Bayes' Rule or Bayes' Law, is used to calculate the likelihood of a hypothesis after the data has been analysed. The conditional probability decides the outcome. Bayes' theorem is described as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

V. DISCUSSION OF RESULTS

A. Decision Tree:

```
In [93]: from sklearn.tree import DecisionTreeClassifier

In [94]: tree=DecisionTreeClassifier()

In [95]: tree.fit(x_train,y_train)

Out[95]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                       splitter='best')
```

B. Support Vector Machine(SVM):

```
In [78]: from sklearn.naive_bayes import GaussianNB

In [79]: nb=GaussianNB()

In [80]: nb.fit(x_train,y_train)

Out[80]: GaussianNB(priors=None, var_smoothing=1e-09)
```

C. K-Nearest Neighbor:

```
In [32]: from sklearn.neighbors import KNeighborsClassifier

In [33]: knn=KNeighborsClassifier(n_neighbors=5)

In [34]: knn.fit(x_train,y_train)

Out[34]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')
```

D. Logistic Regression:

```
In [12]: from sklearn.linear_model import LogisticRegression

In [14]: lr=LogisticRegression()

In [15]: lr.fit(x_train,y_train)

Out[15]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='warn',
                   n_jobs=None, penalty='l2', random_state=None, solver='warn',
                   tol=0.0001, verbose=0, warm_start=False)
```

E. ANN:

```
In [47]: from sklearn.neural_network import MLPClassifier

In [48]: mn=MLPClassifier()

In [49]: mn.fit(x_train,y_train)

Out[49]: MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
              beta_2=0.999, early_stopping=False, epsilon=1e-08,
              hidden_layer_sizes=(100,), learning_rate='constant',
              learning_rate_init=0.001, max_iter=200, momentum=0.9,
              n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
              random_state=None, shuffle=True, solver='adam', tol=0.0001,
              validation_fraction=0.1, verbose=False, warm_start=False)
```

F. Naive Bayes Algorithm:

```
In [78]: from sklearn.naive_bayes import GaussianNB

In [79]: nb=GaussianNB()

In [80]: nb.fit(x_train,y_train)

Out[80]: GaussianNB(priors=None, var_smoothing=1e-09)
```

In Fig. 4: Discussion of Results illustrates various Machine Learning Models on the training set and see which yields the highest accuracy and comparision of accuracy of Decision tree, Support Vector Machine, K-Nearest neighbor, Logistic Regression, Artificial Neural Network.

G. Representation of Accuracy Levels:

	Algorithm	Accuracy	Sensitivity	Specificity	MCC
0	Logistic Regression	0.819672	0.666667	0.967742	0.667530
1	KNN	0.639344	0.533333	0.741935	0.281703
2	ANN	0.786885	0.600000	0.967742	0.612902
3	SVM	0.557377	0.333333	0.774194	0.119897
4	Naive Bayes	0.803279	0.700000	0.903226	0.617326
5	Decision Tree	0.770492	0.766667	0.774194	0.540860

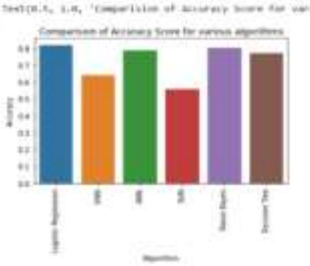

Fig. 4Comparision of Accuracy Levels for Various Alogorithms

VI. CONCLUSION

This paper presents an heart disease spirals out of control when exacerbated. Heart problems are difficult to diagnose and lead to thousands of deaths each year. If you disregard the early warning signs of heart disease, you will be putting your health at risk. Nowadays, an inactive lifestyle and high stress levels are making things worse. Early detection minimises the chance of the illness spiralling out of control. To be safe, it is always recommended to exercise on a regular basis and kick bad habits as soon as possible. The ultimate goal is to determine whether data mining methods may be helpful in accurately predicting cardiac disease. Our aim is to be more accurate and efficient with fewer characteristics and tests. These characteristics are the only important aspects of the research that I have considered. I used four different types of machine learning to conduct my research: K-nearest neighbour, Naive Bayes, decision tree, and random forest. Before being utilised in the model, the data were pre-processed. Naïve Bayes, K-nearest neighbour, and randomforest all show promising results for this modelling method. After developing four methods, I discovered the accuracy of K-nearest neighbours (k = 7) to be the greatest. This study may be improved by using more data mining methods such as time series, clustering, and association rules, as well as support vector machine, genetic algorithm, and neural network. Although this research used a limited number of variables, more sophisticated and multi-variate models would be needed in order to achieve a better level of accuracy for early heart disease prediction.

REFERENCES

[1] AvinashGolande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa PrincyR,J.Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn SystAppl.2017;9:1–16.

[7] PahwaK, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p.500–504.

[8] Pouriyeh S, Vahid S, SanninoG, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p.204–207.

[9] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p.1880–84.

[10] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications.