Prediction of Heart Diseases In Comparison With Different Machine Learning Algorithms

K.Jayakiran, M.Pranavi, M.Novika, V.Tejaswini, N.Rajesh

Abstract: ***Machine Learning is one of the field of computer technological know-how has the potential to expertise to "examine" and additionally get admission to the statistics. Predication of heart disease based on gadget studying set of rules for the actual time data is constantly an enthusiastic case. The coronary heart illnesses prediction utility is an give up person support and on line session task The task can be performed through the usage of device different algorithms to locate the accuracy of the dataset which incorporate the attributes like BP, sugar ,heartbeat and many other. The dataset may be carried out in R language to discover the accuracy based at the algorithm. Here, In this dataset we have were given 14 attributes with three hundred instances has been used as the primary dataset for the training and testing out of the developed gadget. There are distinct classifiers, specifically Decision Tree (DT), Naive Bayes (NB), Random Forest(RF), K –Nearest Neighbour (KNN) , and Neural Network(NN) been used to evaluate the accuracy of the algorithm.***

KeyWords : Decision Tree, Naive Bayes, Neural Network.

I. INTRODUCTION

The purpose of Machine Learning is to apprehend the shape of the records and match that information into models that may be understood and utilized by the humans. Machine gaining knowledge of is the sub department of Artificial intelligence and the improvement idea of gadget mastering is deep learning. Machine learning can be capable of predict the future based at the past or historic facts. 1.1 Terminology :

1.Characteristics: features of information
2.Feature extraction: extracting the features from the given facts
3.Feature vector: combing all the functions
4.Sample: a records to be processed

1.2 Types of Machine Learning Algorithms:

1.Supervised Learning :
Algorithms: Linear regression, Decision tree, Neural community.
2.Unsupervised Learning:
Algorithms: Association Rules, K-method clustering
3.Reinforcement:
Algorithms: Utility learning, Q learning.
4.Semi Supervised Learning.
Algorithms: Graph based techniques, self education.

Revised Manuscript Received on April 06, 2019.
K.Jayakiran, computer science engineering, KLEF, Vaddeswaram,India.
M.Pranavi, computer science engineering, KLEF, Vaddeswaram,India.
M.Novika, computer science engineering, KLEF ,Vaddeswaram, India.
V.Tejaswini,computer science engineering, KLEF ,Vaddeswaram, India.
N.Rajesh, Associate Professor, computer science engineering, KLEF.

1.3 Designing a getting to know system :

1.Choosing the education experience.
2.Choosing the target feature.
3.Choosing the illustration for the target function.
4.Choosing the feature approximation algorithm.
5.The very last design.

1.4 Perspectives in ML :

1.Constant studying:
A device mastering set of rules ought to learn constantly until maximum accuracy is reached
2.Feature learning:
Extract the capabilities even for disordered records sets
3.Fast processing:
A system getting to know algorithm ought to perform at better levels
4.Large seek space:
Choosing the quality direction from the massive facts set

II. LITERATURE SURVEY

The paper published in the year 2017 by the authors Himanshu Sharma, Department of Computer Engineering and Applications, National Institute of Technical Teachers' Training and Research, and M A Rizvi, Department of Computer Engineering and Applications, National Institute of Technical Teachers' Training and Research.
In this, the authors did a survey about the machine learning algorithms how they will help in prediction of heart disease. They used Decision tree, Support Vector Machine, K Nearest Neighbors and Deep learning techniques. They had concluded that Deep learning might be the best for heart disease prediction as its accuracy is high.[1]
The paper published in the year 2016 by the authors Sonam Nikhar, M.Tech Student, Department of CSE, Ramdeobaba College of Engineering and Management, Nagpur, India and A.M. Karandikar Assistant Professor, Department of CSE, Ramdeobaba College of Engineering and Management, Nagpur, India.
In this, the authors had taken machine learning algorithms Naive Bayes and Decision tree classifier were they had explained the two algorithms clearly in a detailed way.
From the experiment results they had concluded that using Naïve Bayes and Decision tree classifier with better information and better calculations gives better results in the diagnosis of heart disease and better accuracy as compared to other classifiers. They had even concluded that the Decision tree classifier's accuracy is better than that of Naïve Bayes [2]

The paper published in the year 2016 by author Ashok Kumar Dwivedi
In this, the author had done research on the performance of different machine learning techniques. The techniques are Artificial Neural Network, Support Vector Machine, Naive Bayes classifier, Logistic regression classifier, k-nearest neighbor, Classification trees, Classification performance measurement and F1 measure. The author concluded that the exertion can be extended for prediction of heart disease by gathering the measurable from various clinics that can provide the more exhaustive model and the recitation can be used for operational heart disease predictions [3]

III. PROPOSED SYSTEM

In this paper the basic idea is about to predict the heart disease by comparing different algorithms in machine learning to get the good accuracy among all the algorithms. The algorithms represented in the paper are:
1. Logistic Regression
2. Decision Tree
3. K-Means Clustering
4. Naive Bayes
5. Neural Networks

IV. METHODOLOGY

We will describe algorithms that we have used to analyse the heart disease by using algorithms. Firstly we take the dataset which consists of 14 attributes and 303 instances. Secondly, we used the attributes to perform the algorithms for prediction. We used the algorithms like

1. Logistic Regression
2. Decision Tree
3. K-Means Clustering
4. Naive Bayes
5. Neural Networks

Here First we described about the concept and later we performed the algorithms with the help of attributes and displayed the results as the final.

4.1 Dataset

Name	Type	Description
Age	Continuous	In Year
Sex	Discrete	0:Female 1:Male
CP	Discrete	1:Typical angina ; 2:Atypical angina; 3:Non-anginal pain 4:Asymptom
Trestbps	Continuous	In mm Hg
Chol	Continuous	In mg/dl
Fbs	Discrete	Fasting blood sugar 120 mg/dl 1:True; 0:False
Exang Continuous Maximum heart rate achieved	Discrete	1:Yes 0:No
Thalach	Continuous	Maximum heart rate reached
Old peak ST	Continuous	Relative to rest
Slope	Discrete	1:Up slope; 2:Flat; 3:Down slope
Ca	Continuous	Ranged between 0 to 3
Thal	Discrete	3:Normal; 6:Fixed defect; 7:Reversible defect
Class	Discrete	0:No Presence ; 1:Least likely to have disease ; 2,3:Middle stages; 4:Most likely to have disease

4.2 Logistic Regression

It is a regression version in which the structured variable is express. Output takes only values zero and one. In case of coronary heart ailment prediction it is beneficial only to decide whether someone get heart disease or not.

Disadvantage:

In case of coronary heart disease prediction we need to expect five lessons which is not feasible using logistic regression**.**

4.3 Decision Tree

Decision tree is used to resolve the complex troubles. It resembles tree-like shape. The essential additives of choice tree are selection node, nodes, and root. Algorithms particularly used by selection tree are ID3, CART, CY3, C5.0 and J48.These algorithms used by selection tree are used to analysis the dataset. Decision tree includes particularly three nodes, they may be:

CHANCE NODES
DECISION NODES
END NODES

and the representation of those nodes vary from every other together with risk nodes are depicted as circles, selection nodes are depicted as squares and this nodes will provide the decision, cease node shows the very last output of the of choice tree**.**

Entropy:

Among all the algorithms inside the decision tree ID3 makes use of entropy to calculate the homogeneity of the given samples. The entropy is said to be homogeneous if and most effective the entropy is zero and if the samples are partitioned similarly then the entropy is one.

$$E(S) = \sum_{i=1}^{c} -pi \log_2 pi$$

Entropy for 2 attributes:

$E(T,X)=\sum_c P(c)E(c)$

Information Gain:

The very first step to calculate the records gain is to calculate the entropy of the target. Mainly the statistics advantage is primarily based at the decrement in entropy after the dataset is break up. From all this, we came to recognize that the decision node of the given dataset.

Gain (T, X) =Entropy (T)-Entropy (T, X)

Decision Tree Representation:

- Root node : Represents whole pattern / populace
- Splitting : Dividing a node into 2 (or) greater sub nodes.
- Decision node : When sub node splits into further sub nodes.
- Leaf / Terminal node : Which can't be split.
- Pruning : When we put off sub node of a choice node.
- Branch : Part of the decision tree

4.4 K-Means Clustering

It is for partitioning in which each cluster center is represented by the imply cost of items in cluster
Input: Data set and quantity of clusters
Output: A set of k clusters

Working:

Randomly selects the items as preliminary clusters for the ultimate items, an

object is assigned to cluster to which it's far maximum comparable based totally on equidistance among cluster item and cluster facilities

okay-approach set of rules improves by computing new imply values or centroids all the objects are again reassigned the use of the updated means

Disadvantage:

If the quantity of clusters are much less then the accuracy can be much less

Answer:

Choose the cluster variety based totally on possibility distribution

4.5 Naive Bayes

In machine learning, Naive Bayes is a classification technique is based on Bayes theorem and not only simple classifier but it is also a simple probabilistic classifier with strong and independent assumptions. It has various applications like spam filtering, personal email sorting and language detection etc. It has an ability to solve the real world complex problems. The most important aspect of naive Bayes is even the attributes are independent still we can predict the classification and it is also used for large data for many dimensions. The memory and CPU consumption are relatively more, therefore, we use the naive Bayes technique when there is a limited resource.

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability and the probabilities of observing various data.

P(h/D)=(P(D/h)*P(h)

• A concept learning algorithm considers a finite hypothesis space H defined over an instance space X.

• The task is to learn the target concept

C:X--->[0,1]

• The learner gets a set of training examples.

• Brute force bayes concept learning algorithm finds the maximum a posterior hypothesis.

Bayes learning features :

- Each observed training is an example
- Prior knowledge can be combined with observed data to determine final probability
- Bayesian can accommodate hypothesis
- New instances can be classified by combining the predictions of multiple hypothesis

Bayes learning is based on two kinds of probability:

1. Prior probability: it is the probability that is calculated from prior information about that event p(A)
2. Conditional probability: It expresses the probability that event A will occur when we know that event B is already occurred p(A/B)

Naïve Bayes classifier:

Applied to learn tasks and each instance x is described by the conjunction of attribute values and target function f(x) of any value in some finite set V. It is based on the simplifying the assumption that all the attribute values are conditionally independent given the target value.

4.6 Neural Network

•These are inspired via the organic gaining knowledge of device consisting of human brains.

•Human brain has numerous talents together with processing statistics can take immediately decisions below complex surroundings and situations.

•This is viable since big networks of parallel and disbursed computational elements prompted neurons.

•Human brain is expected to comprise massive interconnection of neurons.

•These parallely inter connected neurons enhance brain and its processing capability.

•Artificial neural networks evolve with the idea to simulate the human brains.

• Here we've got 3 layers:

1. Input layer : The enter layer receives the enter from external source.
2. Hidden layer : These hidden layers computes a unmarried actual valued output primarily based on weighted aggregate.
3. Output layer : These units determine the class of the characteristic.

There are two styles of Neural network

1. FNN : Feed Forward Neural Network. Information drift from left to proper.
2. FBN : Feed Backward Neural Network. Information float from right to left.

V. EXPERIMENTAL RESULTS

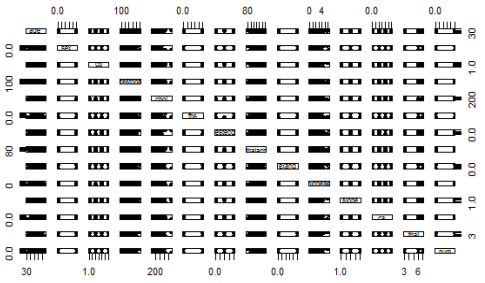

Figure 1:Logistic Regression

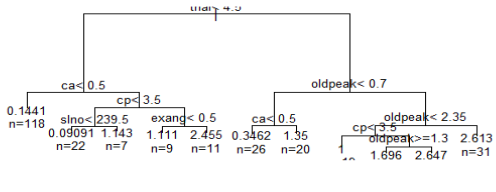

Figure 2:Decision Tree based on heart disease

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21
 33  31  42   9  46  29  45  19  34  25  18  15  25  29   3   9  18   9  19  29  39
 22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42
 38  15  34  42  44  19   7   9  43  35  14  19  18  29   9  43  42  12  25  45  29
 43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63
 35   3  19  34  30  43  38  33  44  18   8  46   6  42  18  36  15  39  22   7  42
 64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
 38  18  42  15  36  26   9   4  14  12  50   8   4  42  10  46  43  39  10   8  26
 85  86  87  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
 29   8   8  10  10  15  45  50  44  10   6  42  45  29   8   8  46  10  38  34  29
107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127
  6  50  12  18  22  25  46  22  11  18   3  44  14  34  14  45  39   2  33  21  32
128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
```

Figure 3.1:K-Means Clustering

```
within cluster sum of squares by cluster:
 [1] 23.056247 12.247856 38.860658 39.049608 20.447494 17.637884 12.037588 31.131041
 [9] 27.047341 57.003014 24.694632 46.292510 35.077630 19.193128 20.125284 17.021402
[17] 40.879534 37.516090 29.425911 31.426820 27.743169 21.138709 26.555383  9.249512
[25] 28.722295 15.384470 22.063506 16.553979 38.230296 30.537148 18.614245 14.221519
[33] 51.237533 40.172821  8.144577 35.626740 28.253464 37.684199 25.974621 18.968080
[41] 27.508707 28.165141 21.327562 63.411446 35.895523 27.756223 10.008449 18.134424
[49]  9.034326 24.344776
 (between_SS / total_SS =  69.4 %)
```

Figure 3.2:K-Means Clustering

```
  oldpeak
Y         [,1]      [,2]
  0 0.4640351 0.6230282
  1 0.9578947 1.0584618
  2 1.6565217 0.9930788
  3 2.1821429 1.5939826
  4 1.9333333 1.3190906

  slope
Y         [,1]      [,2]
  0 1.394737 0.5740467
  1 1.605263 0.5945461
  2 1.913043 0.4170288
  3 2.035714 0.5762036
  4 2.000000 0.5000000

  ca
Y         [,1]      [,2]
  0 0.3153153 0.7004152
  1 0.7297297 0.9021163
  2 1.1739130 0.9840627
  3 1.3928571 1.0659472
  4 1.4444444 1.3333333
```

Figure 4:Naive Bayes

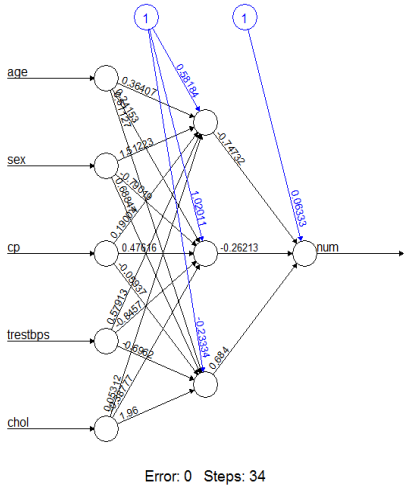

Figure 5.1:Neural Network with five parameters

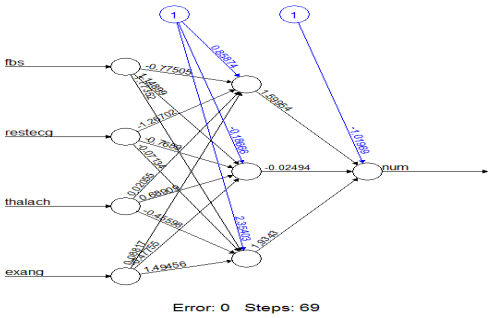

Figure 5.2:Neural Network with four parameters

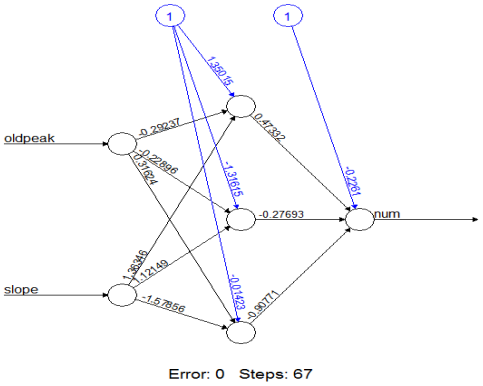

Figure 5.3: Neural Network with two parameters

VI. CONCLUSION

In the above data approximately prediction of coronary heart disease we have taken a dataset with 14 attributes and 303 rows. With the statistics we completed unique strategies like selection tree ,Naive Bayes , Neural network and one of a kind plots like field plot, bar plot and so on.

By acting the techniques we came to recognise the accuracy of given data in Decision tree, Naive Bayes, Neural Network we were given accuracies like

Methods	Accuracy
Decision tree	88.09
Naive Bayes	86.76
Neural Network	89.93

Here we finished some statistical checks to find out viable output predictions with the assist of some in build dataset and we as compared with the strategies with the help of the facts set. While we examine the strategies different techniques may be better or poor . In the future we will carry out the prediction for the one of a kind troubles like prediction of sicknesses in leafs, and sensible device may be evolved that could lead to choice of right remedy strategies for a patient who is laid low with analysis and many others. Data mining is very good assist in figuring out the line of treatment to be followed through extracting knowledge from such appropriate databases.

REFERENCES

1. Himanshu Sharma; M A Rizvi," Prediction of Heart Disease using Machine Learning Algorithms: A Survey" Volume: 5 Issue: 8
2. Sonam Nikhar1; A.M. Karandikar2," Prediction of Heart Disease Using Machine Learning Algorithms", Vol-2, Issue-6, June- 2016, ISSN : 2454-1311
3. Ashok Kumar Dwivedi1," Performance evaluation of different machine learning techniques for prediction of heart disease" 6 September 2016
4. Chaitrali S. Dangare, Sulabha S. Apte" Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques "Volume 47– No.10, June 2012
5. Saima Safdar1 ; Saad Zafar1 ; Nadeem Zafar2 ; Naurin Farooq Khan1," Machine learning based decision support systems (DSS) for heart disease diagnosis: a review", DOI 10.1007/s10462-017-9552-8, 2017.

AUTHORS PROFILE:

K JayaKiran, is a student at the department of Computer Science and Engineering at K L Educational Foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing his research work in knowledge engineering.

M Pranavi, is a student at the department of Computer Science and Engineering at K L Educational Foundation, Deemed to be University, Vaddeswaram ,Andhra Pradesh. She is doing her research work in knowledge engineering.

M Novika, is a student at the department of Computer Science and Engineering at K L Educational Foundation, Deemed to be University, Vaddeswaram ,Andhra Pradesh. She is doing her research work in knowledge engineering.

V Tejaswini is a student at the department of Computer Science and Engineering at K L Educational Foundation, Deemed to be University, Vaddeswaram ,Andhra Pradesh. She is doing her research work in knowledge engineering.

N.Rajesh is working as Assistant professor in the department of computer science and Engineering at K L Educational Foundation,Deemed to be University,Vaddeswaram,Andhra Pradesh.