Prediction Of Heart Diesease Using Machine Learning Techniques

M L Varsha, M H P Kashyap, E Bodhith, M S R Prasad

Abstract: Now-a-days we are using modern techniques to diagnose the diseases at early stages to help the doctors to give the treatment precisely and to save the people. We are using machine learning and data mining techniques in various fields and now in medical field. Many techniques came into existence to diagnose heart diseases, The key challenge is to identify the early stages of heart disease and reliably so that we can reduce the death rate due to heart disease. Machine learning is a technology where machine can do on its own without any intervention of the human and it will learn from past experiments. We used Data mining techniques to label and extract the information from the data. We have uses Decision trees (DT), Random Forest, ANN, Naïve Bayes, SVM , to predict heart disease. The objective is to implement predictive analyzes using these data mining, machine learning algorithms, and to evaluate which mining algorithms are used in machine learning and to conclude that techniques are effective and efficient.

Key words: Data Mining, Decision Trees, Heart disease, Naïve Bayes, Support vector machine (SVM), Random forest

1. INTRODUCTION:

Machine learning is the Artificial Intelligence (AI) sub-domain. Machine Learning is the way to make computer intelligence. The methodologies we use to make computer learn are mostly from machine learning. Machine learning techniques are of four types:

- Supervised Learning
- Un supervised Learning
- Semi supervised learning
- Reinforcement Learning

Such training methods allow machines to understand past experiences and learn from them. In machine learning the most interesting point is machine learns from its past as human being. In prediction of heart diseases we use both supervised and unsupervised learning techniques to get better results and accurately. Since we have huge amounts of medical data sets, machine learning will help us identify patterns and valuable information from them. Although it has many possibilities, in the medical field, machine learning is mostly used to predict disease. Several experts have become involved in using machine learning to recognize diseases as it helps to reduce diagnostic time and improve accuracy and effectiveness. Identifying the illness is the most challenging task in the medical field. Machine learning techniques make it easy. It won't be wasted by this time and we can save a lot of lives. Here we took the heart disease dataset from the UCI server with 14 attributes to decide if the patient has the heart disease or not.. Most researchers used techniques for particle swarm optimization (PSO)[13] and optimization of ant colony (ACO) as well as supervised learning and unsupervised learning.

- *M L VARSHA1, M H P KASHYAP2, E BODHITH3, M S R PRASAD4*
- *K L Deemed To Be University, India, m.lakshmivarsha99@gmail.com*
- *K L Deemed To Be University, India, kashyapmathukumalli@gmail.com*
- *K L Deemed To Be University, India, bodhith9@gmail.com*
- *K L Deemed To Be University, India, msrprasad@kluniversity.in*

SUPERVISED LEARNING: The data set is trained against the algorithm in supervised learning and it checks the data given. For classification and regression, supervised learning methodology is used. Here, there is the output in the dataset. Examples of supervised learning techniques are the decision tree, Naïve Bayes, etc.

UNSUPERVISED LEARNING: The data in the dataset are grouped in unsupervised learning based on their similarities and predicts the out. It is based on clustering techniques like K-nearest neighbor , K-means, Hierarchical clustering etc.

SEMI SUPERVISED LEARNING: We use semi-supervised learning to predict the results.

REINFORCEMENT LEARNING [1]: In reinforcement learning, it learns from the environment and it learns from the mistakes and it gets the best output by going different possibilities.

2. LITERATURE SURVEY:

2.1 DECISION TREE:

Decision tree is one of the technique in supervised learning in which it looks like a tree. Here we construct decision tree by using entropy and information gain. By using this we select the root node. Classification technique is used by Decision tree.

$$\text{Entropy}(T) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

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

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

The tree of decisions is represented as the roots, the leaves. The efficiency and accuracy of the tree is calculated based upon the root node, the number of leaves and the tree length. If the root node changes the accuracy and efficiency of the tree changes. Decision tree is the one which represents the decision through tree like structure. It is capable of handling both numerical and categorical dat a. The complexity of the tree affects the efficiency of the tree. So we use pruning methods to get best accuracy and efficiency for the given data set.

2.2 NAÏVE BAYES[2]:

This classifier

is associate economical probabilistic illustration has attaine d substantial attention to its use for classification. This classifier learns the conditional chance of every Ai attribute given the label C from the training information. Classification is then performed by applying the Bayes principle to live the chance of C given the particular instances of A_1.....A_n associate then predicting the very best later chance category. The categorification objective is to properly predict the worth given a vector of predictors or attributes of a such as separate class variable. The Naive Bayes classifier particularly may be a theorem network wherever there are not any oldsters within the class and every attribute has the category as its solely parent. though the naive theorem (NB) algorithmic rule is basic, it's terribly powerful in several real-world datasets as a result of it will give higher applied math accuracy than well-known strategies like C4.5 and BP and is very effective in learning to mix classifier predictions in linear fashion mistreatment ensemble mechanisms like textile and boosting. it's straightforward and it needs less information to be trained and to be calculable and it got the most effective leads to most of the cases. Naive Bayes classifiers are especially climbable, requiring a variety of linear parameters within the wide range of variables (features / predictors) while obtaining downside information. Maximum likelihood coaching is often performed through mistreatment evaluating a closed-form expression that takes linear time as a substitute for high-priced repetitive approximation as used for several different classifier types. The technique for designing classifiers is straightforward: models assigning classification labels to confounding instances, delineated as vectors of characteristic values, are extracted from a finite set of class labels. There is not one algorithmic rule for the education of such classifiers, but a family of algorithms supported a common principle: all naive Bayes classifiers measure that the value of the selected characteristic is unbiased, provided the classification parameter, to the value of any completely different function. Of example, if it is red, round and around 10 cm in diameter, a fruit can be considered to be associated with an apple. A naive Bayes classifier sees each of these sides as making a contribution to the likelihood that this fruit may be correlated with apple without any potential associations between color, roundness, and diameter choices. Naive Bayes classifiers are often extremely effectively trained for a few varieties of chance models in a supervised learning environment. Parameter estimation for naive Bayes fashions uses the greatest possible strategy in several practical applications; in a few words, one works with the naive Bayes model without supporting theorem chance or using any Bayesian strategies

2.3 ARTIFICIAL NEURAL NETWORKS:

These are designed on the basis of human neural networks. It comprises of specially organized with computing elements that they can learn and acquire the knowledge from the dataset. An artificial neural network is associated with nodes. The behaviors of artificial neural networks largely mimic the activities of human brain. ANN does the computations, pattern recognising and so on. After building the artificial neural networks they are trained with the different datasets.

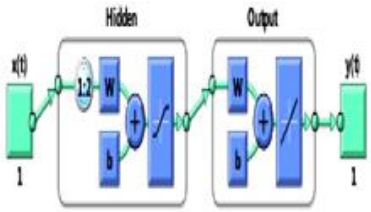

DIAGRAM-1[3]

We have two types of Artificial neural networks:

- Single layered neural networks.
- Multi layered neural networks.

Here we are using back propagation neural network. Firstly, we have to train the neural network by a dataset which is having both input and outputs. Firstly, the input is given to the neural network. The weights which are giving to neural networks are random numbers which are untrained. The value of weights which are giving to the neural network must have values in the middle of -1 to +1. Weight training is used to decrease the error function in neural networks. By doing all This it gives the output. If the output is wrong, the weights on the neural network are changed. By doing so on up to the resemblance output is given by the neural network. Then we stop adding the weights and we consider the weights and we check it for another dataset. The weights we consider in next data in the dataset goes wrong then the process repeats.

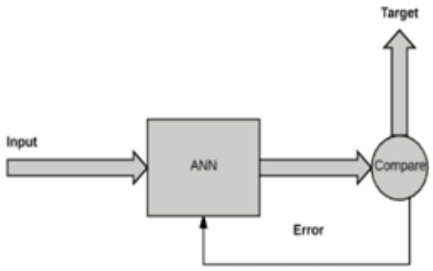

DIAGRAM-2[4]

Three layers are present in Back propagation neural network:

- Input layer
- Hidden layer
- Output layer

In this back propagation neural network comprises of 32 inputs and also having 185 neurons in the hidden layer. Hidden layered neural networks are calculated by the sigmoid function. It is having one neuron in output layer. All these neurons are interrelated with each layer. In Back propagation neural networks input signals are captured by input neurons and output signals are captured by output neuron. So, it is called as acyclic nature. When neural network is constructed then we have to adjust the weights of the input according to the outputs. Among them it is the effective approach which is driving from input layer to output layer.

2.4 RANDOM FORESTS:
Random forest algorithm is a supervised algorithm used for classification. As the title indicates, with a number of trees, this algorithm produces the trees. The more trees in the forest, in general, the more robust the forest is. Similarly the higher the number of trees in the forests in the random forest category results in high accuracy. For both classification and regression function, the same random forest algorithm or random forest classifier can be used. The missing values will be handled by the random forest classifier. When we have more trees in the forest, the template won't be over fit by random forest classifier. It can also design for categorical values the random forest classifier The explanation why the random forest model works so well in data science communication is: a large number of fairly uncorrelated models (trees) working as a committee would outperform any of the individual component models. The primary thing is the model's weak correlation. Just as assets with low correlations (such as stocks and bonds) form a portfolio greater than the sum of their parts, uncorrelated models can generate predictions that are set more accurately than any of the predictions.

2.5 SUPPORT VECTOR MACHINE: Support vector machine (SVMs), a way for each linear and nonlinear data to be categorized. AN SVM is a formula that works as follows in an excessively container. This uses a nonlinear mapping to rework the initial training data in a better dimension. This new dimension seeks the best linear separating hyperplane at intervals (i.e. a "decision boundary" separating 1 category tuples from another category). A hyperplane must isolate data from 2 groups with a suitable nonlinear mapping to a sufficiently high dimension The SVM finds this hyperplane victimization support vectors ("essential" coaching tuples) and margins (defined by the support vectors).We will take away additional into these new ideas later. "I found that late SVMs attracted a lot of attention. What's the reason?"the primary paper on support vector machines was conferred in 1992 by Vladimir Vapnik and colleagues Bernhard Boser and Isabelle Guy on, though the groundwork for SVMs has been around since the Sixties (including early work by Vapnik and Alexei Chervonenkison applied math learning theory). Although the training time of even the fastest SVMs will be extremely slow, due to their ability to model complicated nonlinear call boundaries, they are extremely accurate. These are overwhelmingly less vulnerable than various strategies to overfitting. Additionally, the found support vectors provide a compact description of the learned model. In addition, SVMs will be used for identification of numerical prediction. Together with written digit recognition, object recognition, and classification, they need to be applied to a variety of areas as well as benchmark time series prediction studies. Support Vector Machines area unit assisted hyperplanes thinking defining boundaries. There's a hyperplane. Support Vector Machines (SVMs) area unit assisted hyperplanes thought defining boundaries. A hyperplane is one that separates a group of entities with completely different memberships in the category.

3.RESULTS:

Decision tee:

```
> confMat

    0  1
0 46  3
1  3 39
> accuracy <- sum(diag(confMat))/sum(confMat)
> print(accuracy)
[1] 0.9340659
```

Fig 3.1 *Accuracy of Decision tree*

Naïve Bayes:

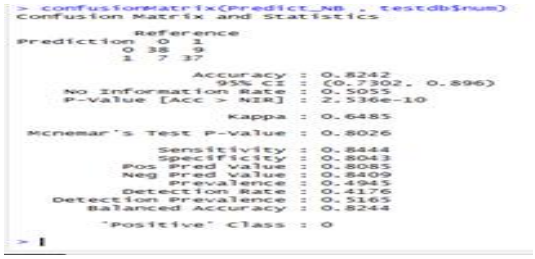

Figure 3.2 *Accuracy of Naïve Bayes.*

Random Forest:

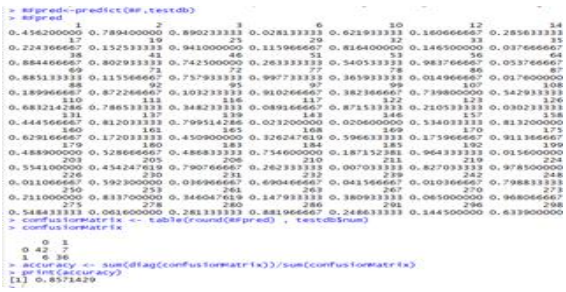

Fig 3.3 *Accuracy of Random Forest*

ARITIFICIAL NEURAL NETWORKS:

```
> roundedresults<-sapply(results,round,digits=0)
> roundedresultsdf=data.frame(roundedresults)
> cc <- table(actual,prediction)
> cc
      prediction
actual  0  1
     0 27  4
     1 11 21
> accuracy <- sum(diag(cc))/sum(cc)
> print(accuracy)
[1] 0.7619048
```

Fig 3.4 *Accuracy of ANN*

SUPPORT VECTOR MACHINE:

```
> confusionMatrix(test_pred, testing_svm$num )
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 44 12
         1  1 33

               Accuracy : 0.8556
                 95% CI : (0.7657, 0.9208)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.588e-12

                  Kappa : 0.7111

 Mcnemar's Test P-Value : 0.005546

            Sensitivity : 0.9778
            Specificity : 0.7333
         Pos Pred Value : 0.7857
         Neg Pred Value : 0.9706
             Prevalence : 0.5000
         Detection Rate : 0.4889
   Detection Prevalence : 0.6222
      Balanced Accuracy : 0.8556
```

Fig 3.5 *Accuracy of SVM*

Methods	*Accuracy*
Decision tree	*93.40*
Naive Bayes	*82.42*
Neural Network	*76.19*
SVM	*85.56*
Random Forest	*85.71*

4. CONCLUSION:

Prediction of cardiac disease using techniques of machine learning we have taken a dataset with 14 attributes and 303 rows. With the statistics we completed unique strategies like Decision tree, Naive Bayes, Neural network, Support vector machine and Random Forest. We have got better results at decision of accuracy 93.4% of prediction of heart diseases and then we got at Random forest and in SVM and Naïve Bayes. We got less accuracy on Neural Networks.

5.REFERENCES:

[1].Maryam I. Al-Janabi 1, Mahmoud H. Qutqut 1 2 *, Mohammad Hijjawi 1 Machine learning classification techniques for heart disease prediction: a review 7 (4) (2018) 5373-5379 IJET

[2].M. Nikhil Kumar et al. Int J S Res CSE & IT. 2018 Mar-Apr;3(3) : 44-51 Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools

[3].Meng hsuen hsieh1, li-Min sun2, Cheng-li lin3,4 Meng-Ju hsieh5 Chung-Y hsu6 Chiahung Kao6–8 Hsieh et al Dove Medical Press Limited Cancer Management and Research 2018:10 6317–6324 Cancer Management and Research Dovepress © 2018

[4].Tariq N (201) Breast Cancer Detection using Artiicial Neural Networks. J Mol Biomark Diagn 9: 371. doi: 10.4172/2155-9929.1000371Alaá Rateb Mahmoud Al-shamasneh, PhD Unaizah Hanum Binti Obaidellah, PhD University

[5].Sonam Nikhar1; A.M. Karandikar2," Prediction of Heart Disease Using Machine Learning Algorithms", Vol-2, Issue-6, June- 2016, ISSN : 2454-1311

[6]. Ashok Kumar Dwivedi1," Performance evaluation of different machine learning techniques for prediction of heart disease" 6 September 2016

[7]. Chaitrali S. Dangare, Sulabha S. Apte" Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques "Volume 47– No.10, June 2012.

[8]. 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.

[9]. M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017.

[10]. s Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN:2347-2200

[11]. Bommadevara, H. S. A., Sowmya, Y., & Pradeepini, G. (2019). Heart disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(5), 270-272.

[12]. Kousar Nikhath, A., & Subrahmanyam, K. (2019). Feature selection, optimization and clustering strategies of text documents. International Journal of Electrical and Computer Engineering, 9(2), 1313-1320. Doi

[13]. Rupa Sri, G., Supriya, A. L., Reddy, E. R. A., & Mandhala, V. N. (2019). An efficient test case prioritization using hierarchical clustering for enhancing regression testing. International Journal of Innovative Technology and Exploring Engineering, 8(5), 914-917. Retrieved from www.scopus.com

[14]. Balaji, G. N., Subashini, T. S., & Chidambaram, N. (2016). Detection and diagnosis of dilated cardiomyopathy and hypertrophic cardiomyopathy using image processing techniques. Engineering Science

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

[16]. Bolisetty, P. K., & Yalla, P. (2016). An efficient component based software architecture model using hybrid PSO - CS algorithm. International Journal of Intelligent Engineering and Systems, 9(3), 46-52. doi:10.22266/ijies2016.0930.05

[17]. Chandana, K., Prasanth, Y., & Prabhu Das, J. (2016). A decision support system for predicting diabetic retinopathy using neural networks. Journal of Theoretical and Applied Information Technology, 88(3), 598-606.

[18]. Kishor Kumar Reddy, C., & Babu, V. (2016). ISLGAS: Improved supervised learning in quest using gain ratio as attribute selection measure to nowcast snow/no-snow. International Journal of Control Theory and Applications, 9(24), 277-289.

[19]. Razia, S., & Narasingarao, M. R. (2017). A neuro computing frame work for thyroid disease diagnosis using machine learning techniques. Journal of Theoretical and Applied Information Technology, 95(9), 1996-2005.