International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Heart Disease Prediction using Supervised Machine Learning Algorithms

Sri Sai Saran Reddy Yeturu, Vergin Raja Sarobin M, Jani Anbarasi, Mohith Krishna Gunapathi, Helen D

Abstract*: Generally, the most complicated task in the healthcare field is the diagnosis of the disease itself. The diagnosis phase in disease detection is usually the most time-consuming task and is prone to most of the errors. Such complications can be effectively handled if the disease detection process is well automated by incorporating effective machine learning algorithms trained with some benchmark datasets. It should also be noted that huge amounts of data that are acquired from Heart Specialization Hospitals are being wasted every year. In this paper, various classification algorithms have been used to train the machine to diagnose heart disease. By a comparative study of various learning models, we have identified the appropriate learning model for the heart disease dataset. Initially, the work will begin with an overview of various machine learning algorithms followed by the algorithmic comparison.

***Keywords*: Artificial Intelligence, Machine learning, Classification algorithms, Heart Disease.**

I. INTRODUCTION

Artificial Intelligence (AI) means making a computer system to learn, understand and take decisions by it. Artificial intelligence is used so that the computer gains experience in a particular field by analyzing the existing data or real-time data. Based on the learned data it will be capable of both predicting and taking decisions on its own. This learning part of the computer is dealt-by Machine Learning (ML) which is the part of AI. Machine Learning is now the most rapidly developing part of Artificial Intelligence (AI). The collection of a huge amount of data particularly in the healthcare sector is being possible by the digital revolution. This vast amount of data is being wasted without being used for analysis purpose. It is a tiresome task for the human to understand and diagnose the disease. However, the availability of huge data, lack of swiftness and inconsistency of human resources are the major reasons for preferring a machine to tackle the job rather than humans.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Sri Sai Saran Reddy Yeturu* ,CSE, Vellore Institute of Technology, Chennai, India. Email: yeturusri.saisaran2016@vitstudent.ac.in
Vergin Raja Sarobin M, CSE, Vellore Institute of Technology, **Chennai**, India. Email: verginraja.m@vit.ac.in
Jani Anbarasi, CSE, Vellore Institute of Technology, Chennai, India. Email: janianbarasi.l@vit.ac.in
Mohith Krishna Gunapathi, CSE, Vellore Institute of Technology, Chennai, India. Email: mohith.krishna2016@vitstudent.ac.in
Helen D, CSE, Amet University. Chennai, India, Email: helensaran15@gmail.com

Therefore, machines are hugely beneficial in saving lives because even amateurs and domain naive can also use machines for the diagnosis process.

Heart diseases are usually termed as cardiovascular diseases that form the major causes for the deaths of many people in various countries including India, United States, etc. So, if these life-threatening diseases are dealt efficiently and quickly we can save many precious lives. In this paper heart disease data set is specifically chosen for training purposes.

This paper starts with the overview and background study of various machine learning algorithms in section II. Section III gives the methodology followed in this work for classification. Section IV will explain the implementation details, results, and discussion. Section V discusses the conclusion

II. BACKGROUND

There are several prediction systems for different diseases that had been proposed and implemented.

Haitham and Alan have proposed automated recognition of sleep apnea using SVM classifier, in which they have used features from signals obtained from the thoracic and abdominal respiratory effort and the classification has been evaluated, the highest accuracy of this system is about 95% [1].

A system that classifies heart disease using SVM and multilayer perceptron neural network architecture has been proposed by Gudadhe et al, where, with SVM accuracy was 80.41% and with ANN accuracy was 97.5% [2].

Myocardial heart disease from ultrasonic images by optimizing fuzzy membership functions was proposed by Tsai and Watanabe, the texture features from the images are used along with the genetic algorithm-based fuzzy classifier are used for classification [3]..

III. METHODOLOGY

A comparison of various machine learning techniques to predict heart disease is the main task to be accomplished. This section has the insights of the techniques on which we have worked for the comparison purpose.

A. Naïve Bayes (NB)

Naïve Bayes Classification is based on the probability of occurrences of the output. It assumes that every feature is independent and calculates the relative occurrence of the output for given input based on the output dependency on the given independent attributes. This is one of the easiest techniques to understand and implement. [4-6]

B. Logistic Regression (LR)

Logistic regression is a supervised learning algorithm [7]. It is a go-to algorithm for binary classification problems. It is used when our data is not linearly separable and used to assign observations to a discrete set of classes in this algorithm; we use the sigmoid function for mapping our predictions to probabilities.

C. Decision Tree(DT)

In this technique, the attributes will be arranged in a multi-level tree-like structure, where the most significant attributes are placed at the roots and the least significant attributes are placed at the leaves. Here significance is calculated by the entropy of the attributes. The decision will be taken according to the traversal of the tree for the given input [8].

D. K-Nearest Neighbor(KNN)

KNN belongs to the class of supervised learning [9]. It is also known as a lazy learning classifier. This algorithm does not construct the model from the data. It classifies the input based on its nearest k training instances and decides its class based on the similarity of the k nearest neighbors it uses Euclidean distance to calculate the distance of an attribute from its neighbors.

E. Support Vector Machine(SVM)

Support vector machine is a supervised learning algorithm [10]. It is a discriminative classifier formally defined by a separating hyperplane. It is one of the most advanced and efficient algorithms which gives better accuracy when compared to the normal classification algorithms.

F. Random Forest(RF)

Random forest is an ensemble learning algorithm, it is based on decision tree algorithm, this classifier creates a set of decision trees from randomly selected subsets of training data set, it then undergoes a voting process from different decision trees and then it decides the final class for the test data set.

G. Data Preprocessing

Data preprocessing is the most important step in machine learning. Better preprocessing leads to better accuracy. Data preprocessing refers to steps applied to the dataset before applying the algorithm. Datasets can have many errors, missing values, duplicates, noises and many other problems that cause the data to be unsuitable for applying the algorithm. These problems can be solved by analyzing the datasets and then using suitable data preprocessing steps. Data preprocessing includes data cleaning, data transformation, feature selection, missing values imputation, data normalization, and many other steps depending on the dataset.

IV. IMPLEMENTATION

In this particular section, the overview of the dataset is given and which is followed by the results obtained by us by applying the above-mentioned classification algorithm on the dataset.

A. Overview of Dataset

The dataset being used is Heart Disease dataset [11] which is obtained from the UCI repository. It contains 14 features with 303 instances, out of which 165 are of Target: Yes-type and 138 are of Target: No-type. The features in the dataset are represented in Table I. The dataset has preprocessed accordingly so that the data fits into the algorithm easily. After the pre-processing phase, all the predictive algorithms are used to fetch the results accordingly.

Table- I: The Features present in the dataset and their description.

Attribute	Description
Age	Age in Years
Gender	Male or Female
Cp	Chest pain type(3 types in total)
trestbps	Resting blood pleasure in mmHg
chol	Serum cholesterol in mg/dl
Fbs	(Fasting blood sugar>120mg/dl)(1=true;0=false)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina(1=yes;0=no)
oldpeak	St depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
Ca	Number of major vessels coloured by flourosopy
thal	Thallium heart scan
target	Possessing disease or not.

Correlation is one of the important factors to be focused while using the classification techniques. The correlation between the attributes with the target attribute has been shown in Table II.

Table II. The Correlation between the attributes and the target attribute.

ATTRIBUTE	Correlation with target
cp	0.433798
thalach	0.421741
slope	0.345877
restecg	0.13723
fbs	-0.0280458
chol	-0.0852391
trestbps	-0.144931
age	-0.225439
sex	-0.280937
thal	-0.344029
ca	-0.391724
oldpeak	-0.430696
exang	-0.436757

As we can observe from the above Table II, the most correlated attribute with the target is ‘cp’ i.e. Chest Pain which infers that the chest pain the most important attribute to determine the target class. The ‘cp’ attribute is followed by the attribute ‘thalach’ which is the maximum heartbeat obtained.

B. Performance Evaluation Metrics

Performance evaluation metrics are used to evaluate machine learning algorithm. Some of the performance metrics which are used in the work are accuracy, precision, recall, and F-measure.

V. RESULTS AND DISCUSSION

This section is all about the results obtained by applying various classification algorithms like Logistic Regression, K nearest Neighbors Algorithm, Support Vector Machine, Naive Bayes Algorithm, Decision Tree and Random Forest Algorithm. The results are nothing but the performance measures which are obtained by splitting the dataset into 80% as the training data and 20% as the testing data. The various machine learning classification algorithms used in the performance comparison graphs are indexed as below.

a- Logistic Regression Algorithm.
b- KNN Algorithm (k=5) *
c- KNN Algorithm (k=7) *
d- KNN Algorithm (k=9) *
e- Support Vector Machine (Cost=0.1) #
f- Support Vector Machine (Cost=0.5) #
g- Support Vector Machine (Cost=1) #
h- Naive Bayes Algorithm
i- Decision Tree Algorithm
j- Random Forest Algorithm

The KNN algorithm is implemented for certain K values such as K= 4 to 10. Using the elbow method of KNN we have found the optimal number for K as 7. Just to show the difference of performance metrics, some sample observations with K as 5,7 and 9 are made and the results are shown in the figures below. Similarly, the SVM classifier was analyzed using various penalty parameters (0.1, 0.5, and 1) and it was found that 0.1 yielded the best results.

A. Accuracy

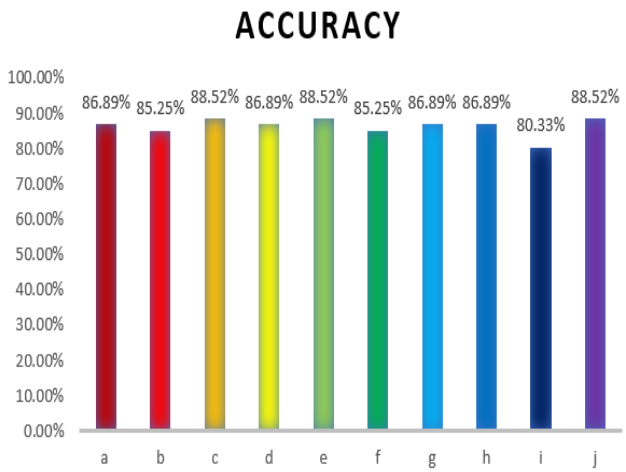

Fig. 1. Accuracy measures of different algorithms.

Accuracy is a measure of the number of correct predictions divided by the total number of predictions made as shown in equation (1).

$$Accuracy = \frac{correct\ predictions}{Total\ no\ of\ predictions} \quad (1)$$

From Fig.1 we can conclude that the KNN Algorithm (k=7), Support Vector Machine (Cost=0.1), Random Forest Algorithm, have the highest accuracy of 88.52% and Decision Tree algorithm with least accuracy of 80.33%.

B. Precision

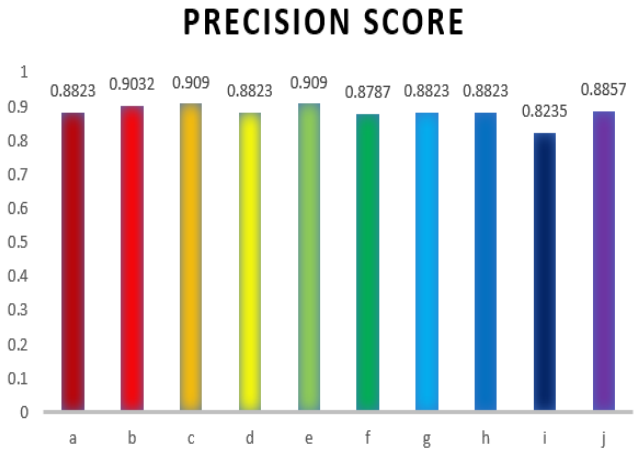

Fig. 2. Precision-Scores of different algorithms.

Precision-Score is the proportion of positive identifications that are actually correct as shown in equation (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

From Fig. 2, we can conclude that the algorithm with highest precision-score are the KNN Algorithm(k=7) and the Support Vector Machine (Cost=0.1) with precision score of 0.909. The algorithm with the least precision-score of 0.8235 is Decision Tree Algorithm.

C. Recall-Score

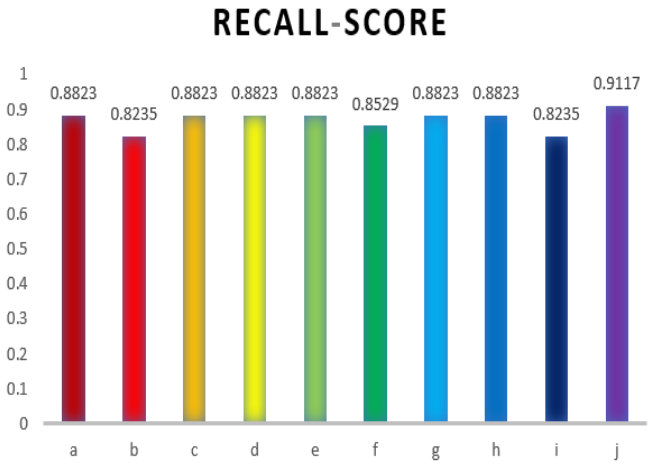

Fig. 3. Recall-Scores of different algorithms.

Recall-Score is the proportion of actual positives that were identified correctly as shown in equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

From Fig.3 the algorithm with the highest recall-score of 0.9117 is the Random Forest Algorithm and least performers with recall-score of 0.8235 are KNN Algorithm (k=5) and Decision Tree Algorithm.

D. F-1 Score

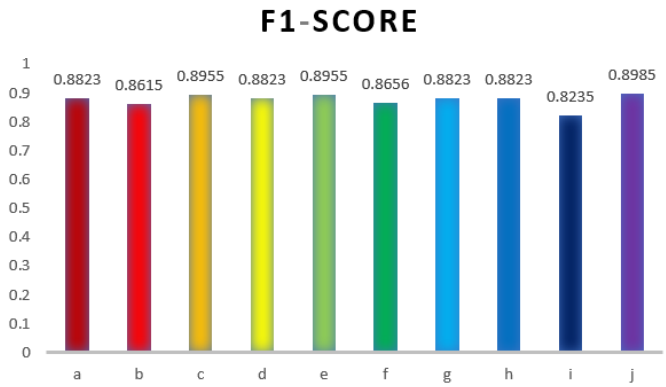

Fig. 4. F1-Scores of different algorithms.

F1-Score can be termed as weighted average of precision and recall as shown in equation (4).

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

From Fig.4 the algorithms with highest F1-score are KNN Algorithm (k=7), Support Vector Machine (Cost=0.1) and Random Forest Algorithm with F-score of 0.8985 and the lowest F1-score of 0.8235 obtained by Decision Tree Algorithm.

VI. CONCLUSION

Prediction of heart disease is an important task in the healthcare field as it can save a person's life. So, this prediction has a significant impact on treatment. In this work, we have experimented with heart disease prediction using various machine learning classification algorithms and the algorithmic performance is analyzed. In the case of healthcare-related predictions, the performance of the algorithm is mainly based on measures like precision score rather than accuracy alone. Considering the factors like Accuracy, Precision, Recall Score, F1 Score from the results section, it is well understood that Random Forest and SVM algorithms are found to be more accurate in terms of accuracy and least RMSE value. It is also found that after dataset pre-processing phase, SVM algorithm with cost as 0.1 could be considered as the best choice among the different SVM variants referred in this paper.

REFERENCES

1. Al-Angari, H.M. and Sahakian, A.V., 2012. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. IEEE Transactions on Information Technology in Biomedicine, 16(3), pp.463-468.
2. Gudadhe, M., Wankhade, K. and Dongre, S., 2010, September. Decision support system for heart disease based on support vector machine and artificial neural network. In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 741-745). IEEE.
3. Tsai, D.Y. and Watanabe, S., 1999. A method for optimization of fuzzy reasoning by genetic algorithms and its application to discrimination of myocardial heart disease. IEEE Transactions on Nuclear Science, 46(6), pp.2239-2246.
4. Jensen, F.V., 1996. An introduction to Bayesian networks (Vol. 210, pp. 1-178). London: UCL press.
5. Murali, M., Bhargava, M., Snehaa, G., Anand, A. and Haque, M.A. Vergin M, Data Analytics on IoT-based Health Monitoring System, International Journal of Recent Technology and Engineering, 8(1), May 2019, pp.220-223.
6. Castillo E (1997) Expert systems and probabilistic network models. Springer, Berlin.
7. Hosmer DW Jr, Lemeshow S (2004) Applied logistic regression. Wiley, New York.
8. Tu, P.L. and Chung, J.Y., 1992, November. A new decision-tree classification algorithm for machine learning. In Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI'92 (pp. 370-377). IEEE.
9. Aha DW (1997) Lazy learning. Kluwer academic publishers, Berlin.
10. Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar "Foundations of Machine Learning", MIT Press, 2012.
11. http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

AUTHORS PROFILE

Sri Sai Saran Reddy Yeturu is currently pursuing his bachelor of technology at Vellore Institute of Technology, Chennai. This paper is one of his beginning works in research area. His main interests include Machine Learning, Artificial Intelligence.

M. Vergin Raja Sarobin holds Ph.D. in Computer Science and Engineering and has been the Professor in the School of Computing Science and Engineering at Vellore Institute of Technology, Chennai. She has several research articles published in international journals. Her research interests include wireless sensor networks, IoT, computational intelligence, and smart grid.

Dr.L.Jani Anbarasi received B.E degree from Manonmanium Sundaranar University in Computer Science and Engineering in 2000, M.E Degree from Anna University in 2005 and Ph.D degree from Anna University in the year 2015. She has around 10 years of experience in various institutions and is currently working as Assistant Professor(Sr) in School of Computing Science and Engineering, VIT Chennai. Her area of expertise includes Cryptography, Image Processing and Soft Computing Techniques. She has published around 24 technical publications in various International Journals and Conferences. Her Professional membership includes India Society for Technical Education (Life Member) and ACM.

Mohith Krishna Gunapathi is currently pursuing his bachelor of technology at Vellore Institute of Technology, Chennai. This paper is one of his beginning works in research area. His main area of interest include IoT, networking.

Dr.D.Helen, working as Assistant Professor, Academy of Maritime Education and Training Deemed to be University (AMET University) Chennai, She obtained her doctoral degree from AMET University, Chennai. Her area of research interest is Wireless Networking, Routing Protocols, MANET and IoT. She has published number of articles in Scopus, UGC listed and high impact factor journals. She presented number of research articles in National, International conferences. She has received Best Paper Award and Outstanding Researcher Award in the International Conferences.