March- April 2020
ISSN: 0193-4120 Page No. 3701 - 3707

Machine Learning Based Predictive Analysis of Heart Disease

S. Sathya Bama, S. Soundarya, A. Balamurugan, Sruthi M.S, V.J Aishwarya Devi

[1]Assistant Professor, Sri Krishna College of Technology, PH-9500341062. *E-mail: sathyabama.s @skct.edu.in*
[2]Assistant Professor, Sri Krishna College of Technology, PH-87542414232. *E-mail: soundarya.s @skct.edu.in*
[3]Professor, Sri Krishna College of Technology, PH-6383436117. E-mail: *a.balamurugan @skct.edu.in*
[4]Assistant Professor, Sri Krishna College of Technology, PH-87542414232. *E-mail: soundarya.s @skct.edu.in*
[5]Assistant Professor, Sri Krishna College of Technology, PH-9489666044. *E-mail: aishwarya.vj@skct.edu.in*

Article Info
Volume 83
Page Number: 3701 - 3707
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 23 March 2020*

Abstract
In current toiling world cardiovascular diseases is becoming a main cause that affects human survival. Machine learning algorithms are becoming more popular in the domain of health care. Heart disease contributes to high mortality rates in India for the past years. This study focus on estimating the efficiency of several machine learning models by providing a predictive analysis model for heart disease. Heart disease UCI dataset with 13 different attributes of 303 patients has been utilized and upon which several supervised machine learning algorithms has been applied and their accuracy has been determined. It is concluded that K-Nearest Neighbor and Random forest algorithm shows better accuracy compared to other algorithms.
Keywords:*Heart disease, Machine Learning, Classification, Random forest, predictive model, KNN, Healthcare*

I INTRODUCTION

Cardio vascular disease refers to various abnormal conditions such as plasma vessel disorders, heartbeat problems and injuries on heart. Cardiovascular disease commonly indicates conditions that include obstructed blood vessels can lead to a heart attack, chest pain or even stroke. Different heart infirmities, that influence tissues of heart, are also responsible for cardio vascular disorders. A study [21] shows that heart problems leads more than 2.1 million mortality in India irrespective of ages. There are several technological concepts related to machine learning concepts are being used in the area of medical science. Vijayakumariet.,al [19] surveyed several detection methods for extraction of nerve details

from the optical or retinal images. In predicting death or heart attack artificial Intelligence is going beyond human knowledge. Machine learning (ML) is creating an impact on the healthcare domain. Healthcare organizations throughout the globe are using the resources of Machine Learning. This study demonstrates a method of proof and shows how to mine the power of Machine Learning and utilize it to real-world problems. A predictive model that prophesize heart disease based on Machine Learning algorithms will be constructed[22]. This system will be useful in earlier identification of heart disease.

II RELATED WORKS

Alaka et, al[1] investigated the advantage of machine learning algorithms in the prediction of stroke impairment and concluded that the study based on the regression model provided better accuracy. Austin et al.,[2] predicted that the tree based methods are more powerful than the conventional methods of classification. Chavda,

Paras, et al., [3] introduced the system that uses machine learning and IoT techniques that predicts the heart problem earlier than they occur. Cikes, Maja, et al.[4] designed a system that predicts the heart problem in patients approximately based on integrating the clinical parameters with unsupervised machine learning techniques. Diller, Gerhard-Paul, et al.[5] enhanced ACHD diagnostic metrics by evaluating the prognosis using machine learning algorithms. These DL algorithms can be readily scaled to multi institutional datasets due to the mainly automated method engaged to further enhance precision and eventually serve as internet decision-making tools. A research introduces numerous healthcare-related problems and different machine learning algorithms that must be able to deliver the highest performance possible. These methods assist mine the appropriate and helpful quantity of data, form the medical dataset that enables the medical organizations to provide beneficial information. A extensive review of the literature to illuminate the prior job conducted in this sector has been summarized [6]. A research paper [13] utilized the image processing technology and algorithms such as –means to perform predictive analysis of brain tumors using the MRI images. Sreeja et.[18] Al in the year 2015 optimized the pattern matching algorithm and concluded that the evaluation time of pattern matching is very effective for pattern matching based classification. A precise version of WoLF-IGA assured to adhere in all general-sum matches to Nash Equilibrium strategies.

III DATASET DESCRIPTION

The dataset used here has the data which have the features, based on which the existence of heart disease is predicted. The data set is loaded into the working environment with pre-installed libraries, which are helpful for analytics. The packages such as pandas seaborn are imported imported. After importing the required libraries the dataset is loaded and read.

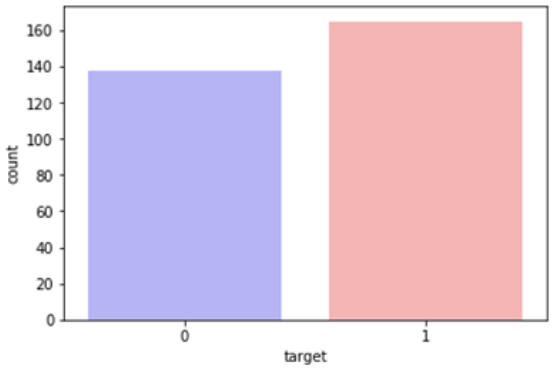

Figure 1. Overall Data Exploration

Figure 1 represent the data exploration of the entire dataset, which represents among 303 samples 165 patients have heart disease. The fig 1 acquired with the help of coutplot(). Then the percentage of patients having and not having the heart disease has been visualized as in fig.2

```
Percentage of Patients Haven't Heart Disease: 45.54%
Percentage of Patients Have Heart Disease: 54.46%
```

Figure 2. Patient with and without Heart Disease

The data has been again classified based on the attribute sex (male and female) and visualized as in fig 3, which depicts that male patients are having more probability of having heart disease than the female.

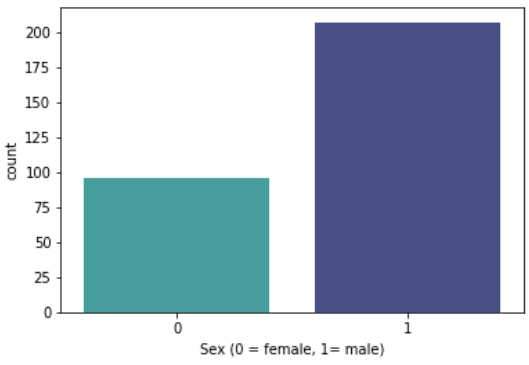

Figure 3. Gender based Exploration

Again the percentage of patients having the disease has been analyzed as shown in fig 4

March- April 2020
ISSN: 0193-4120 Page No. 3701 - 3707

```
Percentage of Female Patients: 31.68%
Percentage of Male Patients: 68.32%
```

Fig. 4. Visualization of gender based exploration

The attribute age has been taken into account and it has been analyzed that the frequency of occurrence of heart disease is more in the age group between 42 to 54 and it is visually represented in fig 5.

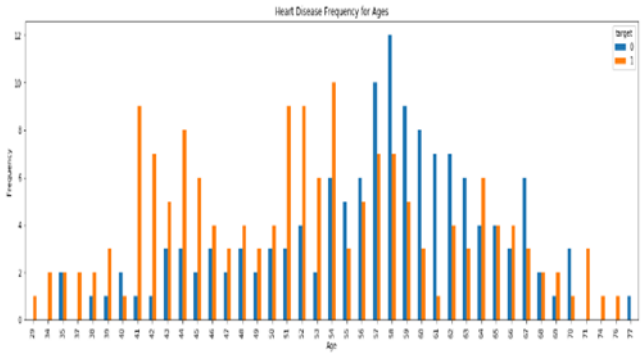

Figure 5. Chest pain based Exploration

Then the chest pain type attribute is taken into account and analyzed. It has been found that type2 pain leads to heart disease in most frequent cases. Fig 6 shows the analysis of heart disease occurrence based on 4 pain types.

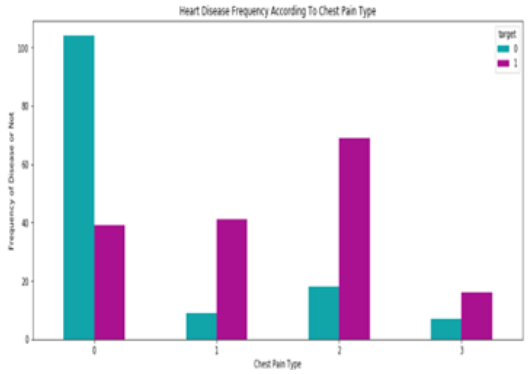

Figure 6. Frequency of heart disease based on chest pain type

IV PROPOSED WORK

Based on the Knowledge gained from the data set exploration the dataset is subjected to preprocessing. The preprocessed data is then subjected to several machine learning classification algorithms.

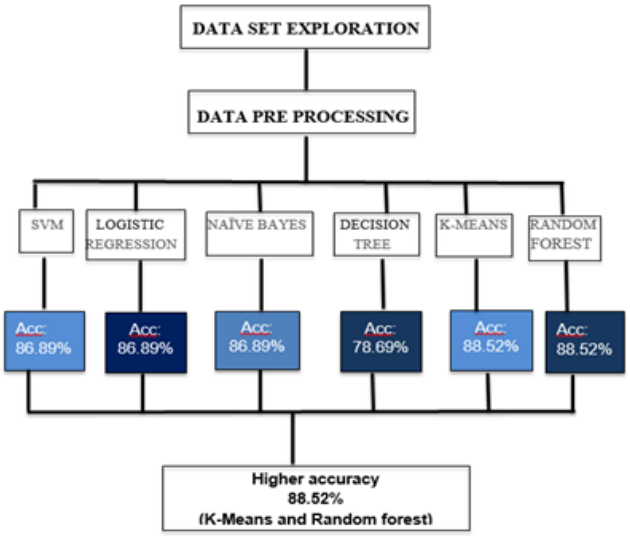

Figure 7. Overall Architecture

Fig.7 depicts the overall flow of the system, the data set has been explored and the unrelated attributes required to predict patients with heart disease has been removed by data preprocessing. Algorithms such as Support Vector Machine, Logistic regression, K-Means, Decision tree and Random forest has been imposed on the dataset and the accuracy of classification has been computed. It is concluded that K-Means and Random forest results in higher accuracy of 88.52%.

V EXPERIMENTAL RESULTS

5.1 Logistic Regression

Logistic regression is a binary classification algorithm that classifies the dataset into two classes. Binary classification implies two values either 0 or 1. Fig 8 represent the binary classified values ranging from 0 to 1. The function for the logistic function has been written as follows

Xchanged=X-Xmin/Xmax-Xmin.

The data is thensplited as 80% training data and 20% testing data. Logistic regression classified the data with 86.89% accuracy.

March- April 2020
ISSN: 0193-4120 Page No. 3701 - 3707

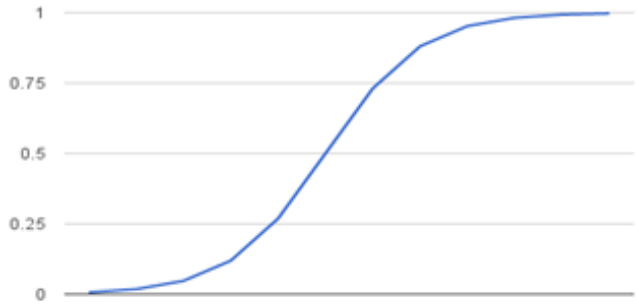

Figure 8. Logistic Regression

5.2 K-Means Classification

K-means algorithm divides the dataset into clusters. Each clusters has its data points and centroids. Based on the distance between the data points and the centroids the sub groups were identified. The main aim of k means classification is to reduce the error function. The heart disease dataset containing several attributes were divided into several clusters and the accuracy of classification is found to be 88.52% which is shown in fig 9. The algorithm classifies the dataset based on the patients having the heart disease and patients without heart disease. For the given dataset the algorithm first identifies the number of clusters by shuffling the dataset. After initializing the centroids the data points with less mean square distance will be identified, this step is repeated until it finds the better data point.

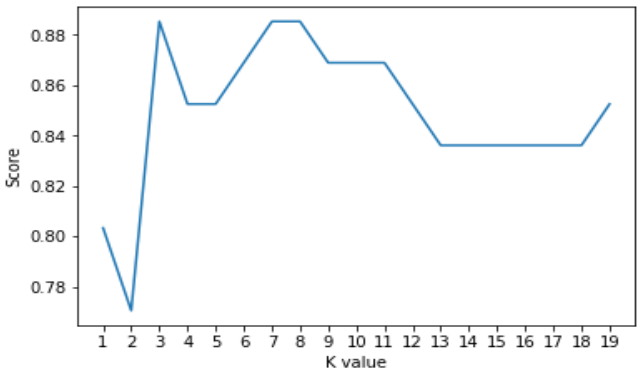

Figure 9. K means classification

5.3 SUPPORT VECTOR MACHINE

Support Vector Machine is mainly used for classification problems. The main objective of SVM is to identify the optimized hyperplane that classifies the data points resulting in two classes. In this system based on certain features the heart dataset is classified and its accuracy is found to be 86.89%.

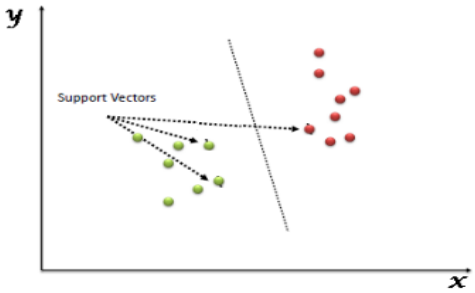

Figure 10. Support Vector Machine

Fig 10, depicts the support vectors and the hyper plane that classifies the heart disease dataset. The hyperplane is placed between two classes of data points, one represent the patients with heart disease and the other one represent the patients without heart disease.

5.4 NAÏVE BAYES ALGORITHM

Naïve Bayes classifier can be realized on high-dimensional datasets. Naïve Bayes classifier prophesies the possibility of each class based on the feature vector. Naïve Bayes algorithm classified the data with an accuracy of 86.89%.

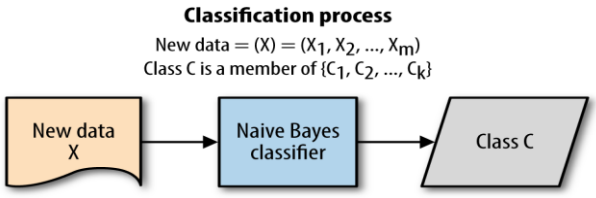

Figure 11. Naïve Bayes algorithm

5.5 DECISION TREE ALGORITHM

Decision tree comes under the category of supervised learning algorithm. This algorithm will

generate a decision tree for the given dataset. The dataset will contain nodes that are inter connected in a hierarchical manner. Based on the condition present in the parent node of the child nodes are generated. By employing decision tree algorithm on heart disease dataset, the accuracy of 78.69% is achieved. Fig 12 represents a view of decision tree that contains parent and child nodes.

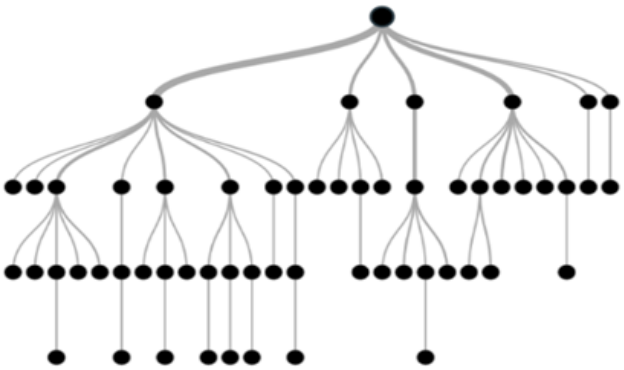

Fig. 12. Decision tree algorithm

5.6 RANDOM FOREST CLASSIFICATION

Random forest is a type of ensemble learning. Random forest algorithm is nothing but the collection of several decision trees. Each decision tree will offer its predictive value. By considering the outcome of all the decision trees the value with higher votes are identified and the one with higher vote is chosen. Fig 13 pictorially represent the overview of random forest algorithm. By applying Random Forest provides the accuracy of 88.52%.

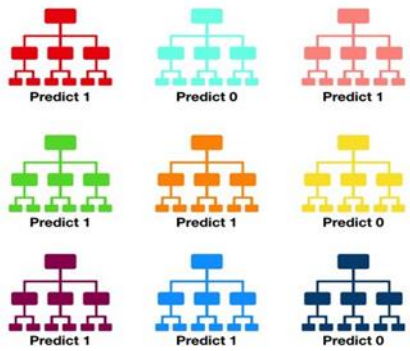

Fig. 13. Random Forest algorithm

5.7 ALGORITHM COMPARISON

Fig 14. Compares the algorithm accuracies. By comparing the six classification algorithms it has been found that random forest and KNN algorithms have the greater accuracy compared to other algorithms.

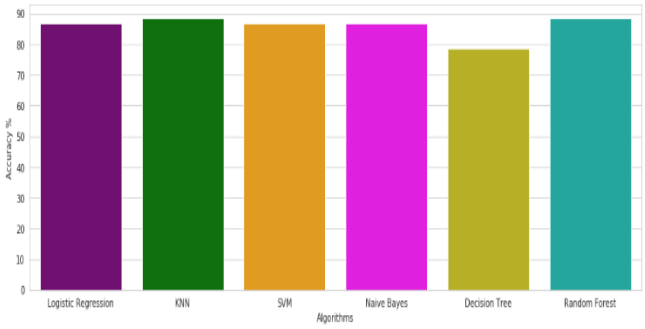

Fig. 14. Algorithm Comparison

VI CONCLUSION

Several classification algorithms has been analyzed and applied to classify the heart disease dataset. Among the six classification algorithms random forest and KNN algorithms results in higher accuracy of 88.5%. The predictive model can be designed using either any of the algorithms. The accuracies of algorithms may vary on different datasets. This application may be further extended by applying or hybridizing the existing algorithms to achieve higher accuracy rate.

VII REFERENCES

[1] Alaka, Shakiru A., et al. "Abstract WP182: Machine Learning Models are More Accurate Than Regression-based Models for Predicting Functional Impairment Risk in Acute Ischemic Stroke." Stroke 50.Suppl_1 (2019): AWP182-AWP182.

[2] Austin, Peter C., et al. "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes."

Journal of clinical epidemiology 66.4 (2013): 398-407.

[3] Chavda, Paras, et al. "Early Detection of Cardiac Disease Using Machine Learning." Available at SSRN 3370813 (2019).

[4] Cikes, Maja, et al. "Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy." European journal of heart failure21.1 (2019): 74-85.

[5] Diller, Gerhard-Paul, et al. "Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients." European heart journal 40.13 (2019): 1069-1077.

[6] Kannan, R., and V. Vasanthi. "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease." Soft Computing and Medical Bioinformatics. Springer, Singapore, 2019. 63-72.

[7] Ke, C., Gupta, R., Xavier, D., Prabhakaran, D., Mathur, P., Kalkonde, Y.V., Kolpak, P., Suraweera, W., Jha, P., Allarakha, S. and Basavarajappa, D., 2018. Divergent trends in ischaemic heart disease and stroke mortality in India from 2000 to 2015: a nationally representative mortality study. The Lancet Global Health, 6(8), pp.e914-e923.

[8] Kukar, Matjaž, et al. "Analysing and improving the diagnosis of ischaemic heart disease with machine learning." Artificial intelligence in medicine 16.1 (1999): 25-50.

[9] Mazzotti, Diego R., et al. "0832 Evaluating Supervised Machine Learning Models for Cardiovascular Disease Prediction Using Conventional Risk Factors, Apnea-Hypopnea Index and Epworth Sleepiness Scale." Sleep42.Supplement_1 (2019): A334-A334.

[10] Melero-Alegria, Jose Ignacio, et al. "SALMANTICOR study. Rationale and design of a population-based study to identify structural heart disease abnormalities: a spatial and machine learning analysis." BMJ open 9.2 (2019): e024605.

[11] Nuti, Sudhakar V., et al. "Characterizing Subgroups of High-Need, High-Cost Patients Based on Their Clinical Conditions: a Machine Learning-Based Analysis of Medicaid Claims Data." Journal of general internal medicine (2019): 1-3.

[12] Palaniappan, Sellappan, and RafiahAwang. "Intelligent heart disease prediction system using data mining techniques." 2008 IEEE/ACS international conference on computer systems and applications. IEEE, 2008.

[13] Perumal, TamijeSelvy&Purusothaman, T. (2011). Performance Analysis of Clustering Algorithms in Brain Tumor Detection of MR Images. Eur J Sci Res. 62.

[14] Pimentel, Angela, et al. "Coronary heart disease prognosis using machine-learning techniques on patients with type 2 Diabetes Mellitus." Chronic Illness and Long-Term Care: Breakthroughs in Research and Practice. IGI Global, 2019. 198-217.

[15] Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." Global journal of computer science and technology 10.10 (2010): 38-43.

[16] Ravindran, R., K. Manonmani, and R. Narayanasamy. "An analysis of void coalescence in AL 5052 alloy sheets annealed at different temperatures formed under different stress conditions." Materials Science and Engineering: A 507, no. 1-2 (2009): 252-267.

[17] Reddy, N. Satish Chandra, et al. "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction." International Journal of Innovative Computing 9.1 (2019).

[18] Sreeja, N.K. and Sankar, A., 2015. Pattern matching based classification using ant colony optimization based feature selection. Applied Soft Computing, 31, pp.91-102.

[19] Vijayakumari, V., and N. Suriyanarayanan. "Survey on the detection methods of blood vessel in retinal images." Eur. J. Sci. Res 68, no. 1 (2012): 83-92.

[20] Weng, Stephen F., et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." PloS one 12.4 (2017): e0174944.

[21] Wu, Jionglin, Jason Roy, and Walter F. Stewart. "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches." Medical care(2010): S106-S113.

[22] Radha , B.MeenaPreethi “Machine Learning Approaches For Disease Prediction From Radiology And Pathology Reports”, Journal of Green Engineering, Alpha Publishers, Vol.9,No.2,pp.149-166,2019

[23] Uhrmann, L.S., Nordli, H., Fekete, O.R. and Bonsaksen, T., 2017. Perceptions of a Norwegian clubhouse among its members: A psychometric evaluation of a user satisfaction tool. International Journal of Psychosocial Rehabilitation, 21(2).

[24] Das, B. and KJ, M., 2017. Disability In Schizophrenia and Bipolar Affective Disorder. International Journal of Psychosocial Rehabilitation, 21(2).

[25] Elsass, P., Rønnestad, M.H., Jensen, C.G. and Orlinsky, D., 2017. Warmth and Challenge as Common Factors among Eastern and Western Counselors? Buddhist Lamas' Responses to Western Questionnaires. International Journal of Psychosocial Rehabilitation, 21(2).

[26] Weston, S.M., Martin, E.D., Shippen, M.E., Kraska, M.F. and Curtis, R.S., 2017. Parents with Serious Mental Illness Served by Peer Support Specialists. International Journal of Psychosocial Rehabilitation, 21(2).