ISSN 2319 – 7595
Volume 12, No.6, November – December 2023
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs011262023.pdf
https://doi.org/10.30534/ijiscs/2023/011262023

# Heart Risk Assessment Analysis and Prediction using Various Machine Learning Algorithms

[1] **Sharon Susan Jacob**
Assistant Professor, School of Data Analytics
Mahatma Gandhi University, Kottayam, Kerala, India
[1] sharonsusanjacob@gmail.com

Received Date : October 27, 2023 Accepted Date : November 26, 2023 Published Date : December 07, 2023

**ABSTRACT**

In today's world artificial intelligence plays a major and crucial role in technological development. Machine Learning is an extension of Artificial Intelligence through which accuracy and prediction rate of model is made without the assist of any external programs. The evolution andopportunity of ML is expandingeach and every day. Many organizations have made ML as one the important assets of their company.Supervised, Unsupervised and Reinforcement learning and ensemble learning's are different types of ML through which datasets accuracy can be measured. Our current study focuses mainly on Utilizationof ML in medical industry. Based on current report of World Health Organization the most predominant cause of death is detecting most on cardiovascular disease. Myocardial infarction also said as heart attack is caused when flood flow to the heart muscles has been suddenly blocked. It should be considered as the severe one detection should be done in early stage so life can be prevented. This article presentsaexamination on machine learning for projection of heart disease. Distinct expert algorithms have been used for the evaluation and model deployment purpose.

**Key words:** Logistic-Regression, Decision-Tree, KNN, Naïve-Bayes, Neural-Networks, Random-Forest, Support-Vector-Machine, XG-Boost

## 1. INTRODUCTION

One of the most threatening and deadly global health conditions is heart disease through which much life's can be taken. The principal part is identifying and diagnosing the disease early as possible. The combination of Machine Learning into health care appliances can identify and prevent the heart disease. The experimentation has been regulated for the medicaments and methodology of cardiac arrest and upcoming research is under process. Finding and identifying illness is most crucial task. Many individuals can be saved if illness is treated at early stage. Proper guidance and accurate efficiency can be provided by the assistance of ML which can make a high contribution impact in the clinical field. Heart disorder which drastically developed can be rapidly identified by many emerging technologies. The living nature and way a person is measured by the health condition of an adult. The expansion of cardiovascular disease through decades consists of different fundamental genetics. The probability of men is higher comparing to women on terms of cardiac disorders. In this research, parameters such as gender, diabetes and BP are used for speculating disease diagnosis. There are numeroussources involved in prediction of disease but we are taking few important parameters. Establishment on the condition of heart the symptoms may differ. Blood vessel illness is one cause for heart disease. Congestive disease has been developed when there is no proper blood artery pumping to the various organs of our body. Heart injuries can be caused when flow of blood is reduced it is said as coronary heart failure. Drowsiness and death in today's lifestyle is happened because of cardiovascular disease. Silent systems monitors have been adopted by hospital to monitor welfare and patient health. In this article most popular and reputed ML techniques has been used to envision the ratio to obtain major accuracy by providing reliable and explicable speculative model is the aim of this article. The model consists of algorithm such as Logistic-Regression, Decision-Tree, KNN, Naïve-Bayes, Neural-Networks, Random-Forest, Support-Vector-Machine, XG-Boost. The structure of paper consists of literature review in section II. Proposed methodology on section III. Result introduction and discussion in section IV.

## 2. LITERATURE REVIEW

The appliances of Neural Network and Decision Tree were used in heart disease prediction by Gandhi and Singh. From the research they identified that while training data set many features have been affected so they decided to lower the quantity of data and they identified process timing has been reduced and notable to obtain proper accuracy [1]. Neural network, Decision tree and Naive Bayes were used by Thomas and princy to identify uncertainty for heart disease but the research were failure because they focused mainly on Data Mining algorithm [2]. Enhancement of swarn particle for the association and classification using appliances genetic algorithm and Artificial NN was done by bharti and singh. More accurate precision was obtained by mining law of association into predictive model. It is considered as successful methodology [3]. Electronically device could play an vital role in medical diagnosis was stated by sharmasaxena and purushottam there major aim

was to produce low cost efficient working framework. Based on the metrics of health set of instructions will be automatically created by the model for providing probability level for patient health. The instructions may be categorized based on user needs. This methodology was reliable for speculating heart disease risk rate [4]. Palaniyappan and awang created IHDPS intelligent heart illness projection system by combining the concepts of DT, Artificial NN and NB. By providing most beneficial care the price of health care is minimised. The most advancement techniques of data mining will help to decipher the questionnaires. To make others to understand the concepts data was perfectly tabulated with graphical representation [5]. For the better heart disease prediction Sharma and rizvi employed combinational method which group's deep learning, KNN, SVM and decision tree. By doing continuous cleaning and pre-processing the noisiness can be removed from data so that better result can be derived and they identified to obtained more reliable accuracy while using neural networks [6]. Clustering and grouping methods were introduced by mukherjee, gupta and mandal for identifying various reasons for cardiovascular disorder [7]. Different methodologies were performed by krishnaiah, narsimha used Data Mining factors for producing high accuracy [8]. The data set obtaining function should be moulded to obtain better outcome to achieve pre-processing is mandatory to remove repeated data set was stated by kaur[9]. Various multilayer categorization algorithms for dataset havebeen utilized to translate vast quantity of data easily from a repository and utilized in an efficient way suggest in the paper by vijayashree [10]. The cause of heart disease metrics such as stroke, diabetes, smoke, obesity, diet was identified to reduce heart illness by chamberlain [11]. Detecting silent heart attack is not an easy one but it was well handled by kumar by using recurrent NN, svm, CNN by producing equitable precision [12]. J48 algorithm was performed immediately by Deepak pre dataset available in library. For performing repository UCI RF and DT is used. [13]. Weng trained different concept by using gradient boosting and logistic regression DL algorithms by using technique of maintaining patient details in systematic documentation so that heart disease can be perfectly forecasted [14]. garibaldiproved that ML produced best result comparing to other models. CPRD link data was taken for the model. The dataset contains vast details such as medicine information, history of populations a statistical report [15]. Data mining was primarily used for heart disease detection by murugeshwarifor detecting cardiac disease. Dataset was extracted from UCI heart disease library. Data mining processing uses R program. In their process they came to conclusion that the effectiveness of algorithm was low when preparation with one algorithm but while using hybridization by training two or more algorithm better precision was obtained. Diagrammatic representations of processing, reduction dimensionality normalization was obtained for result analysis for classification algorithms. The algorithms used were naïve bayes, NN and DT while comparing to classifier NN performance was high in classifiers problem. Eight distinctML algorithms were utilized in this article and the performance was high in all task and application [16].

## 3. PROPOSED METHODOLOGY

The procedure of using medical record for predicting cardiovascular disease using various ML methods was discussed in this article. The Independent input has been segregatedfrom heart failure dataset. Inaccessible isolated data values from the column have been pre-processed and the cleaned and inspected using various machine learning algorithm. In supervised learning algorithm the data is segregated into two equal setsone is training and another one is testing. To verify whether previous process was correctly done cross validation is used. After the process of cross validation correctnessof the method is displayed. The system will stop the process after successful completion. In this research, different machine learning algorithms like Logistic Regression, Decision Tree, KNN, Naïve Bayes, Neural Networks, Random Forest, Support Vector Machine, XG-Boost, which are known as the best classifiers for providing more accuracy and results, have been used.The workflow of themodel is illustrated in Figure 1.

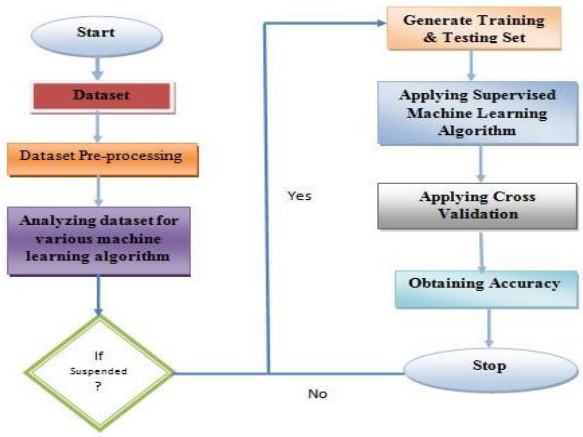


**Figure 1:**Workflow of Model

### 3.1 K-Nearest Neighbours (KNN)

The most useful non parametric model used for predictive model comparability between new and existing data points is called as the K –Nearest Neighbours algorithm and is versatile used for both regression & classification problems. Output of the model is obtained from the given input data theperception about the data is large so its prediction is highly rated. The model is segregated into two halves training and testing. The training is mostly used in development of model in training phase. The quantity of observation determines from the square root is known as the k value. The k value determines the set. Test data is extracted from the built foundation there are various universal distance dimension. Euclidean distance, Manhattan distance and minkowski measurements are different measurements together with constant variables. But probably used distance is Euclidean distance the formula is stated below.

On the built foundation, test data is now anticipated. There are certain distance dimensions that are universal. With constant variables, measures such as Minkowski distance, Manhattan distance and Euclidean distance can be utilized. The commonly used metric, though, is the distance from Euclidean. For Euclidian reach, the formula is as follows:

$$d = \sqrt{((x1 - x2)^2 + (y1 - y2)^2)} \quad (1)$$

The K-NN works based on following steps Training, prediction, voting and averaging.

**3.1.1Training**

The dataset first half is taken for training in which every point of data contains labelled class in classification task) and regression task consists of values in numbers.

**3.1.2 Prediction**

Euclidean distance is applied for measuring closeness. Closeness is a concept where new unlabelled points in data have been predicted by obtaining recentdata points which are close to the training data set of k data point.

**3.1.3 Classification (voting)**

For latest data points many numbers of k-nearest neighbours can be assigned for class labels in classification tasks. For example, k=10. Seven values of neighbours belong to class X and four belongs to class Y. Here according to higher value data point is classified as Y.

**3.1.4 Regression (Averaging)**

Here the data point has been predicted by taking averages of k nearest neighbours target values.

**3.2 Support Vector Machine (SVM)**

The mostly useful algorithm for classification tasks because of optimal hyper plane where high dimensional space is derived by separating different data points. Here in the instruction data will get separated into two points group with hyper planeto portray existence and non-existence of cardiac disease. SVM works based on hyper plane specification intensify interval between two various dimensions. Penalized SVM is used to bridge gap between discrimination of class. In ML a class variance issue may be arrived when trace of inequality obtained between positive and negative attributes. There may be a failure in class if the distinction of class not yet been addressed. A series of mathematical function is one type of kernel used in machine learning. In our approach fine tuning of hyper parameter is done and linear kernel has been added to obtain better result in SVM classifier.

$$K(x,x')=\exp((-||x-x'||2)/2\sigma 2) \quad (2)$$

To execute this process Grid search CV is used. Data can be transferred in form of separate c values. From group of given values, it automatically develops SVM models by finding best c value so that best potential of model is obtained.

**3.3 Naïve Bayes Algorithm (NB)**

The highly influential probabilistic machine learning algorithms which is mostly used in classification tasks. It involves Bayes theorem and its name naïve bayes is taken from naïve meaning features taken on basis of assumption from independence value. The deployment of classification is done when it has high input dimensionality. The naïve bayes classifier states difference between existences of one characteristic in a class to companionship of another character

$$P(Y/X) = P(X/Y)\ P(X) \quad (3)$$

It needs minimum training knowledge where preceding instance is X and dependent case is Y. it is easy and used mostly in binary classification problem.

**3.4 Decision Tree (DT)**

The supervised machine learning algorithm used for regression and classification tasks. It is structured like a tree through which decisions can be made on possible sequences it is the widely used model among all sectors. Data set consists of many values like weight, cholesterol, BP and nicotine. Selecting root node is the tedious one in decision tree root node element in data should be specially categorized on non-parametric implication of feature selection.

**3.5 Random Forest**

Random forest is one among classification methods. Here process output has been done through individual trees where large number of data has been processed through selection trees. In this method multiple decision tree generates output from the data's retrieved from input phase selection trees. Higher result performance has been obtained by reducing correlation on random decision tree and by increasing strength of result. Based on various information predictions are made by aggregate prediction methods. To verify the accuracy of parameters cross validation is done among models. Finally using precision, parameter and recall accuracy for model is obtained.

**3.6 Neural Networks**

It is one of the ML models designed in the form and functionality of human brain. It is mostly used in the concept of pattern reorganization weredecision making has been done automatically done by self-learned data. The artificial neurons or inter nodes is used for connecting neural networks. Input layer, hidden layer, output layer are three main layers of neural networks. The data has been collected buy input layer and processing of data has been done by hidden layer and finally output has been derived, verified and displayed buy output layer. The input has been processed and received using nodes called as neurons. It is one of the activity functions which process input and produce output.

$$Z= q(T.X+c) \quad (4)$$

Where Z is output of neuron, Input of neuron X , T weight of neuron , activating function q and bias as c.

**3.7 Logistic Regression**

The logistic regression is one the important model for binary classification problem mainly used in the

concept of statistical learning method. The value lies between 0 and 1. The data has been gathered and missing values has been identified and pre-processed by identifying regularized numerical features and encoding qualitative variables. In logistic regression model sigmoid function is used for calculating binary outcome values.

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}} \quad (5)$$

Where x1, x2…xk are called as features, β0, β1…, βn are coefficient and Y is called as binary outcome variables.

### 3.8 XG Boost

Extreme gradient boosting is called as XG boost algorithm it is one of the members of gradient boosting algorithms. The main advantage of this algorithm is efficiency and speed well suited for popular tabularised structured data mainly used for Ranking, classification and regression tasks. It is a type of ensemble method used for increasing functionality by making error corrections. It has two parts the first one training and best fit of model has been obtained using loss term and the complexity of model's performance has been completely controlled by regularization model. The unseen data has been easily identified and generalized by objective function. Data has been collected and initialization for max tree depth, learning rate, trees in ensemble hyper parameters has been done. Data set is trained and optimization of error is done so that error made by previous tree has been corrected by upcoming one and hyper parameters have been fine tuned to obtain better result.

## 4. RESULTS AND DISCUSSIONS

The accuracy and performance of model has been obtained by machine learning algorithms.Heart disease data file has been tested in proposed methodology. The retrieval of data is done from Kaggle. It is available to general people for free. Hungarian and Cleveland dataset is considered perfect model for developing models because they have minimal and a smaller number of missing values and outliers [19]. From the group of dataset only major parameters are taken for research. Cleveland database has been mostly used by machine learning research. To verify the working methodology of proposed work comparison metrics is done using various algorithms to verify effectiveness of proposed work. Here methods like Logistic-Regression, Decision-Tree, KNN, Naïve-Bayes, Neural-Networks, Random-Forest, Support-Vector-Machine, XG-Boost has been used. Cleaning and pre-processing are the first and main process in our model. The overall motto of the research is to recognize the current presence of cardiac disease. The more accurate result is obtained because of UCI data set. Heart disease existence is obtained using value 4 and no existence if value is 0. The foremost aim is to focus the differentiation between existence value (1, 2, 3, and 4) and non-existencevalue (0). Integer value is accepted by machine learning model. So categorical values are transferred into numeric values. If the variable has two or more categorical variables than duplicate variables were constructed. The scheduled work efficiency is displayed in Figure 2.

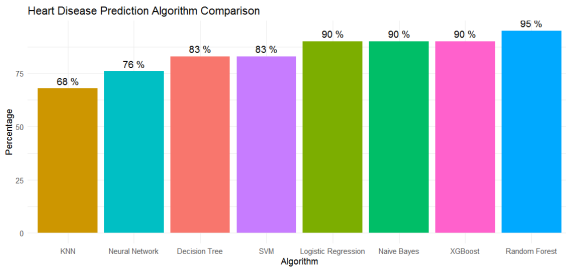


**Figure 2:**Result Prediction and Comparison

The Logistic Regression achieved 90% of accuracy, Decision-Tree achieved 83% of accuracy, KNN achieved 68% accuracy, Naïve-Bayes achieved 90% accuracy, Neural-Networks achieved 76% accuracy, Random-Forest achieved 95%, Support-Vector-Machine achieved 83% accuracy and XG-Boost achieved 90% accuracy. Random Forest outperformed other methods.

## 5. CONCLUSION

With expansion in growth of heart disease in current scenario it is necessary to build a system which has all the potentials to execute the application with proper efficiency forecasting and illness. The virtue of this examination is to create a potential ML model which predicts heart disease very accurately. The evaluation was based on Logistic-Regression, Decision-Tree, KNN, Naïve-Bayes, Neural-Networks, Random-Forest, Support-Vector-Machine, XG-Boost algorithms accuracy to predict heart disease. The heart attack metrics of patient were taken from the dataset such as chest pain, cholesterol and BP but there are some groups where family gene also plays role for heart attack. This Research could help victims at the beginning stage of heart disease and proper treatment can be given to them. Various Machine Learning methods isevolved to prevent from heart diseases heart disease estimation, Random Forest achieved higher accuracy of 95%. In the future, a convolutional neural network (CNN) can be used to improvise future findings.

## REFERENCES

[1] M. Gandhi and S. N. Singh. "**Predictions in heart disease using techniques of data mining**", 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE ,2015.

[2] J. Thomas and R. T. Princy. "**Human heart disease prediction system using data mining techniques**",2016, 10.1109/ICCPCT.2016.7530265.

[3] S. Bharti and S. N. Singh. "**India analytical study of heart disease prediction comparing with different algorithms**", May 2015.

[4] S. Purushottam, K. Saxena, and R. Sharma. "**Efficient heart disease prediction system using decisiontree**",2015, 10.1109/CCAA.2015.7148347

[5] S. Palaniyappan and R. Awang. "**Intelligent heart disease prediction using data mining techniques**", August 2008.

[6] H. Sharma and M. A. Rizvi. "**Prediction of heart disease using machine learning algorithms:A survey**", August 2017.

[7] A. Hazra, S. K. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee. "**Heart disease diagnosis and prediction using machine learning and data mining techniques**: A review", Advances in Computational Sciences and Technology, 2017.

[8] V. Krishnaiah, G. Narsimha, and N. S. Chandra. "**Heart disease prediction system using data mining techniques and intelligent fuzzy approach**: **A review**", International Journal of Computer Applications, February 2016.

[9] R. Kaur and P. Kaur. "**A review–Heart disease forecasting pattern using various data mining techniques**", International Journal of Computer Science and Mobile Computing, June 2016.

[10] J. Vijayashree and S. N. Iyengar. "**Heart disease prediction system using data mining and hybrid intelligent techniques**:" A review, 2016.

[11] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, and S. Cheng. "**Heart disease and stroke statistics 2018 At-a-Glance**", American Heart Association, 2018.

[12] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "**Heart attack prediction using deep learning**," Department of Computer Engineering., Army Institute of Technology, Pune, Maharashtra Professor, 2018.

[13] M. N. Kumar, K. V. S. Koushik, and K. Deepak. "**Prediction of heart diseases using data mining and machine learning algorithms and tools**", 2018, 10.13140/RG.2.2.28488.83203.

[14] A. Kaur and J. Arora. "**India heart disease prediction using data mining techniques**:" A survey, 2018.

[15] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi. "**Canmachine learning improve cardio-vascular risk prediction using routine clinical data**", 2017.

[16] A. S. Arthy and G. Murugeshwari. "**A survey on heart disease prediction using data mining techniques**",International Journal of Computer Sciences and Engineering,April 2018.

[17] A. Sudha, P. Gayathri, and N. Jaisankar. "**Effective analysis and prediction model for stroke disease using classification methods**", April 2012.

[18] D. Baroud, A. N. Hasan, and T. Shongwe. "**A study towards implementing various artificial neural networks for signal classification and noise detection in PFDM/PLC Channels**," 12th IEEE international symposium on communication systems, networks and digital signal processing (CSNDSP), 2020.

[19]**https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data**