ISSN No:-2456-2165

Detection of Heart Failure Using Different Machine Learning Algorithms

Raghav Sharma
Mayuri Mukewar
Anurag Navale
Asmita Manna

Abstract:- Heart is the key organ of our body as blood circulation towards other organs depends upon efficient working of the heart . Nowadays, Coronary artery diseases diminish the working ability of hearts to a large extent, resulting in failure of hearts in many cases. A survey conducted by WHO reveals that around 29.20% of the world's population i.e 17 million people die due to various heart diseases each year. For identifying various heart diseases, several pathological procedures and medical investigations are being done by doctors. With the use of data mining and machine learning techniques, better insights can be provided from the existing test results and the number of pathological procedures can be reduced. A system created using Data Mining and Machine Learning algorithms, can overcome the dearth of examining tools for classifying the data and predicting the Risk state of Cardiac patients. In this paper, a comparative survey of such approaches for investigation of Cardiac diseases using Data Mining techniques is presented. These comparative study results would be really helpful for researchers in this domain for channelizing their research in the appropriate direction.

Keywords:- *Comparative Study, Machine Learning, Investigation, Naive Bayes, K-Nearest Neighbor, Random Forest, Decision Table, K-Means.*

I. INTRODUCTION

Heart Diseases are increasing day-by-day. As per WHO survey, approx. 17.9 million people from all over the globe die due to various health diseases, and approx 80% of these deaths are due to coronary heart diseases (eg heart attack) and cardiovascular diseases (eg strokes).Due to this development of most of the countries gets affected to some extent .Predicting heart diseases beforehand and changing lifestyle accordingly - would help to reduce the number of deaths. Machine learning and data mining ae are used now a days in solving the problems related to many health diseases. Prediction is one of the important problems where machine learning techniques are widely utilized. In this work, we have done a comparative study of existing data mining and ML algorithms which helped us to predict heart diseases by processing existing heart patients' data using machine learning algorithms. Heart diseases are not just coronary diseases but they vary for more inner parts which are connected to the heart.

Data gathering about such diseases has been part of the study for a long time. We are considering cholesterol,lung-function test, treadmill check, etc.. as parameters to predict the risk rate of the heart.

II. LITERATURE REVIEW

The paper "Predictive Analysis of Heart Disease using K-Means and Apriori Algorithms" by Hadia Admin[2019] and his team ,proposed a technique for detecting heart failure in patients using K- Means and Apriori. Their approach showed that Apriori and K-mean algorithms when applied together, the infection can be anticipated much superior than before and it helps the doctor to make necessary decisions of diagnosing the patients These algorithms effectively predicted the cardiac risk stage (low, medium, or high) and assisted clinicians in accurately understanding the patient's condition and providing appropriate diagnosis. These algorithms also aided in the hospital's storage and maintenance of the patient's record.

The paper "Congestive heart failure detection using random forest classifier" by Zerina Masetic and his colleagues demonstrates the outstanding performance of the Random Forest algorithm, demonstrating that it is useful in determining the defeat of jammed heart and may be a treasure in conveying information that will be convenient in therapy.

The paper "Heart Disease detection using Naive Bayes Algorithm" by K. Vembandasamy [2015] and his team stated that data mining techniques have acted as one of the most important and known solutions for many health-related issues. Among this techniques Naive bayes algorithm has given appropriate results regarding the prediction of heart disease. The results thus obtained shows that Naive Bayes algorithm provides accuracy of 86% with minimum execution time.

The Decision Tree technique, according to Dilip Kumar Choubey[2020], relies on the top randomness of input samples. It's a Divide and Conquer (DAC) method in which the trees are constructed from the top down method. The data is first preprocessed by splitting it into training and test data in this algorithm.

ISSN No:-2456-2165

Edward Choi [2017] found that Deep learning modules are adapted to secure relations that have seemed to build on execution of models for investigation of event HF with a short monitoring period of 12-18 months.

Vishal Dineshkumar Soni [2020],proposed that Comparison of distinct ways to estimate cardiac illnesses utilising data mining methodologies, studying the numerous contrasts of professional Data Mining algorithms, and assessing those methodologies are significant and fortunate. The most often used Data-Mining techniques for predicting risk are Naive Bayes, Random Forest, and Decision Tree.

Rajesh N, T Maneesha [2018] stated that KNN is a controlled classifier and employs inspection from inside a trail station to identify categorization tags. KNN is one of the most often used classification algorithms for heart disease research. When implementing KNN, a few assumptions are made, the most common of which are a dataset with low disturbance, labelled, and containing relevant characteristics. Processing big datasets with KNN takes a lengthy time. This algorithm achieves a 63.4 % accuracy rate.

Logesh Kannan N[2020] along with his colleagues stated in their paper "Heart Disease Detection using Machine Learning" that machine learning and Data Mining methodology have been proven to be one of the important and appropriate solutions for predicting the risk state of the heart of cardiac patients. This paper along with machine learning techniques uses python programming to predict the risk state of heart.

Author	Title	Method	Pros	Cons	Remark
Rajesh N ,T Maneesha , Shaik Hafeez , Hari Krishna (2018)	Machine Learning Algorithms for Heart Disease Prediction	Naive Bayes	We may utilise a combination of ML methods such as Naive Bayes and K-means to achieve appropriate accuracy using a fusion of ML techniques such as Naive Bayes and K-means.	In some cases, Naïve Bayes will not give exact outputs so we need to consider the outputs of different ML techniques.	● If the input data is clean and well kept, Nave Bayes produces a more accurate answer.
Hadia Amin,Abita Devi,Nida UI Amin (2019)	K-means and the Apriori Algorithm are used to predict heart disease.	K-Means and Apriori	The K-Mean Algorithm produces a more perfect and significant result than the Apriori algorithm's weighted correlation criteria.	Alone Apriori can't give effective results.	● The results of the experiments suggest that the best approach to predict heart disease is to combine Apriori and K-mean algorithms. ● It helps the doctor to make necessary decisions to diagnose the patients
K.Vembandasamy,R.Sasipriya *PP*and E.Deepa,(2015)	Heart Diseases Detection Using Naive Bayes Algorithm	Naive Bayes	A Naive Bayes technique is easy to build, having no complex dull variable evaluation which enables it to be helpful particularly in the area for determining the risk rate of heart patients.	Decision trees give less accurate results while synthesizing small datasets in some cases.	● The outcomes acquired show that the Naive Bayes technique provides 86.4198% of correctness with least time.
Zerina Masetic,Abdul Hamit Subasi(2016)	Congestive heart failure detection using random forest classifier	Random Forest	Normal and congestive heart failure are both treated with machine learning approaches (CHF). The random forest algorithm detects CHF with 100% accuracy.		● The results of different trials were analysed using a variety of statistical metrics (sensitivity, specificity, accuracy, F-measure, and ROC curve), and it was discovered that the random forest method provides 100% accuracy.

Chithambaram T,Logesh Kannan N,Gowsalya M (2019)	Heart Disease Detection Using Machine Learning	K-Nearest Neighbor, Random Classifier, Correlation, SVM	For minimising the occurrence of heart disease, the author employed a decision tree method. The author has implemented Gini index method using hyperplane and decision tree in this SVN algorithm, which exhibits the maximum gain in characteristics and displays strong representation of the decision tree technique.		● The major aim is to determine which algorithm provides the most accuracy in predicting future issues that the illness may bring, as well as which algorithm provides the most accuracy in determining whether a person has heart disease or not.
Keshav Srivastava, Dilip Kumar Choubey(2020)	Heart Disease Prediction using Machine Learning and Data Mining	Decision Trees, KNN,Naïve Bayes, Random Forest, SVM.	A web app is built using flask, and these packages are used to make predictions based on the data supplied by the user. Future researchers can enhance their accuracy by employing data mining techniques to retrieve hidden information from samples..		● In their experiment, they used the Cleveland Heart Disease dataset from the UCI repository to pre-process data with missing values and used algorithms such as Decision Tree, K-Nearest Neighbour, Support Vector Machines, and Random Forest to achieve accuracy of 79 %, 87%, and 83 %, respectively. ● The AUC for Decision Tree, K-Nearest Neighbour, Support Vector Machines, and Random Forest is 71.6 %, 88.5 %, 90.4 %, and 90.8 %, respectively, according to the ROC curve.

Algorithmic Survey

Here are some of the algorithms that we have identified for getting more accurate results regarding investigation of Heart Disease.

Parameters	K Means	Decision tree	Random Forest
Definition	● **The K Means method is a recursive technique that attempts to split a dataset into K separate clusters, each of which contains just one data point.** ● Data Mining projects are done by mostly using K means Quality of clusters remains the same throughout the execution process for showing the accurate output.	● Decision Trees are a supervised Machine Learning method in which data is continually divided according to a set of rules. ● Decision nodes and leaves are two procedures that may be used to describe the tree. ● The ultimate results are represented by the leaves, while the decision nodes are the points where the data is split.	● Random Forest is an extractor that holds several decision trees on distinct subsets of a dataset and averages them to increase the dataset's prediction accuracy. ● It is superior than a single decision tree because it reduces over-fitting by averaging the results.

Algorithmic steps	1. K is the number of clusters to specify. 2. Initialize the centroids by shuffling the dataset and then picking K data points at random for the centroids without replacing them. 3. Continue iterating until the centroids do not change. i.e. the clustering of data points does not change. ● Calculate the total of all data points' squared distances from all centroids. ● Assign each data point to the cluster that is closest to it (centroid). ● Calculate the cluster centroids by averaging all of the data points that correspond to each cluster.	1. S begins the chapter with the root node, which includes the whole dataset. 2. Using the Attribute Selection Measure, find the best attribute in the dataset (ASM). 3. Subdivide the S into subsets that include the best attribute's potential values. 4. Create a node in the decision tree that holds the best attribute. 5. Create new decision trees in a recursive manner using the subsets of the dataset produced in step -3. 6. Continue this procedure until you can no longer categorise the nodes any further and refer to the last node as a leaf node.	1. Choose K data points at random from the training set. Create decision trees for the data points you've chosen (Subsets). 2. Choose N for the number of decision trees you wish to create. 3. Steps 1 and 2 should be repeated. 4. Find the forecasts of each decision tree for new data points, then allocate the new data points to the category with the most votes.
Accuracy	Gives less accurate results	Gives less accurate results.	Gives more accurate results.
Dataset	Can handle massive amount of data and noisy data	Cannot handle large data and noisy data	Can handle enormous amounts of data as well as noisy data
Speed	Faster	Faster	Faster than Decision tree and K means
Pros	1. Execution is rather simple. 2. It can handle huge data sets. 3. Convergence is guaranteed. 4. It's possible to warm up the locations of centroids. 5. Adapts quickly to new situations. 6. Generalizes to other forms and sizes of clusters, such as elliptical clusters.	● Decision trees need less work for data preparation during pre-processing than other methods. ● A decision tree does not need data normalisation. ● A decision tree does not need data scalability. ● In addition, missing values in the data have no significant impact on the decision tree-building process. ● A decision tree model is simple to understand and convey to technical teams and stakeholders.	● Random Forest can handle both classification and regression problems. ● It can handle big datasets with a lot of dimensionality. ● It improves the model's accuracy and eliminates the problem of overfitting.
Cons	● It necessitates determining the number of clusters (k) ahead of time. ● It can't deal with noisy data or outliers. ● Clusters having non-convex forms are not appropriate for detection.	● A little change in the data can result in a huge change in the decision tree's structure, resulting in instability. ● When compared to other algorithms, a decision tree's calculation might get rather complicated at times. ● The training period for decision trees is frequently longer.	● Despite the fact that random forest may be used for both classification and regression tasks, it is not better suited to regression tasks..

		● Because of the intricacy and time required, Tree of Decisions training is relatively costly. ● When it comes to using regression and predicting continuous values, the Decision Tree method falls short.	

Source for database:
https://data.world/informatics-edu/heart-disease-prediction

III. CONCLUSION

After doing the literature survey, we found that among all the techniques available in machine learning and data mining, Random Forest with accuracy 100%,K nearest neighbor with accuracy 88.5% and Naive Bayes algorithm with accuracy of 86% have proven to be the most important and efficient algorithms to determine the Risk state of heart of cardiac patients.

REFERENCES

[1]. Hadia Amin,Abita Devi,Nida UI Amin, "Predictive Analysis of Heart Disease using K-Means and Apriori Algorithms",Journal of Applied science and Computations, ISSN NO:1076-5131,Aug 2019
[2]. Zerina Masetic, AbdulHamit Subasi,"Congestive heart failure detection using random forest classifier", Computer Methods and Programs in Biomedicine, Volume 130, July 2016.
[3]. K.Vembandasamy,R.Sasipriya*PP*and E.Deepa, "Heart Disease detection using Naive Bayes Algorithm", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015.
[4]. Dilip Kumar Choubey,Keshav Srivastava, "Detection of Heart Disease using Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020
[5]. Edward Choi ,Andy Schuetz, Walter F Stewart, Jimeng Sun "Using recurrent neural network models for early detection of heart failure onset", Journal of the American Medical Informatics Association, Volume 24, Issue 2, March 2017
[6]. Vishal Dineshkumar Soni, "Detection Of Heart Disease Using Machine Learning Techniques", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 08, AUGUST 2020 ,ISSN 2277-8616
[7]. Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna,"Prediction of Heart Disease using Machine Learning Algorithm" , *International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366*
[8]. Chithambaram T,Logesh Kannan N,Gowsalya M, "Heart Disease Detection using Machine Learning" ,DOI: https://doi.org/10.21203/rs.3.rs-97004/v1