Journal
of the
Oriental Institute
M.S. University of Baroda
ISSN: 0030-5324

HEART DISEASE PREDICTION USING MACHINE ALGORITHM

Dr.D.Lakshmaiah[1],K.Padma[2] , K. Naveen[3], L. Rishitha Reddy[4], Md Haseena Begum[5], M Bhanu Prasad[6]

[1] Professor & Head of the Dept, Dept of ECE Sri Indu Institute of Engineering & Technology, Hyderabad

[2] Assistant Professor, Dept of ECE Sri Indu Institute of Engineering & Technology, Hyderabad

[3-6]Student, Dept. Of ECE, Sri Indu Institute Of Engineering & Technology, Hyderabad.

Abstract—

To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning[1-3] is the branch of Artificial Intelligence(AI), it provides prestigious support in predicting any kind of event which take training from natural events. In this paper, we calculate accuracy of machine learning algorithms for predicting heart disease, for this algorithms are k-nearest neighbor, decision tree, linear regression and support vector machine(SVM) by using UCI repository dataset for training and testing. For implementation of Python programming Anaconda(jupytor) notebook is best tool, which have many type of library, header file, that make the work more accurate and precise.

Keywords:Machine Learning, SVM,

I. INTRODUCTION

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases[4] is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized atlast stage due to lack of accuracy of instrument so thereis need to know about the more efficient algorithms for diseases prediction.

Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning[1-3] is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence.

As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used four algorithms which are decision tree, linear regression, k-neighbor, SVM[5]. In this paper, we calculate the accuracy of four different machine learning approaches and on the basis of calculation we conclude that which one is best among them.It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients

II .LITERATURE REVIEW

Heart is one of the core organ of human body, it play crucial role on blood pumping in human body which is as essential as the oxygen for human body so there is always need of protection of it, this is one of the big reasons for the researchers to work on this. So there are number of researchers working on it. There is always need of analysis of heart related things either diagnosis or prediction or you can say that protection of heart disease . There are various fields like artificial intelligence, machinelearning.

Data mining that contributed on this work . Performance of any algorithms depends on variance and biasness of dataset. As per research on the machine learning for prediction of heart diseases himanshu et al .naive bayes perform well with low variance and high biasness as compare to high variance and low biasness which is knn. With low biasness and high variance knn suffers from the problem of over fitting this is the reason why performance of knn get decreased. There are various advantages of using low variance and high biasness because as the dataset small it take less time for training as well as testing od algorithm but there also some disadvantages of using small size of dataset.

When the dataset size get increasing the asymptotic errors are get introduced and low biasness, low variance based algorithms play well in this type of cases. Decision tree is one of the nonparametric machine learning algorithm[1] but as we know it suffers from the problem over fitting but it cloud be solve by some over fitting removable techniques. Support vector machine is algebraic and statics background algorithm, it construct a linear separable ndimensional hyper plan for the classification of datasets. The nature of heart is complex, there is need of carefully handling of it otherwise it cause death of the person.

III.PROPOSED SYSTEM

Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling" model designed by applying neural network. This paper used the electronic health record (EHR) data from real world datasets related to congestive heart disease to perform the experiment and predict the heart

disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal importance of respecting the sequential nature of clinical.

K.Prasanna Lakshmi, Dr. C.R.K.Reddy (2015) designed "Fast Rule-Based Heart Disease Prediction[1-4] using Associative Classification Mining". In the proposed Stream Associative Classification Heart Disease Prediction (SACHDP), we used associative classification mining over landmark window of data streams. This paper contains two phases: one is generating rules from associative classification mining and next one is pruning the rules using chi-square testing and arranging the rules in an order to form a classifier. Using these phase to predict[1-4] the heart disease easily.

M.Satish, et al. (2015) used different Data Mining techniques like Rule based, Decision Tree, Navie Bayes, and Artifical Neural Network. An efficient approach called pruningclassification association rule (PCAR) was used to generate association rules from cardiovascular disease warehouse for prediction of Heart Disease. Heart attack data warehouse was used for pre-processing for mining. All the above discussed data mining technique were described.

Aakash Chauhan et al. (2018) presented "Heart Disease Prediction[1-4] using Evolutionary Rule Learning". This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient's dataset. This will facilitate (help) in decreasing the amount of services and shown that overwhelming majority of the rules helps within the best prediction of coronary .

Ashir Javeed, Shijie Zhou et al. (2017) designed "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". This paper uses random search algorithm (RSA) for factor selection and random model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program.

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposed by Senthilkumar Mohan, Chandrasegar Thirumalai et al. (2019) was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes were chosen from the preprocessed knowledge(data) set of cardiovascular disease[8-9] .After prep-processing , the hybrid techniques were applied and disgnosis the cardiovascular disease.

Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik (2015) "An

Intelligent Decision Support System for Cardiac Disease Detection", designed a cost efficient model by using genetic algorithm optimizer technique. The weights were optimized and fed as an input to the given network. The accuracy achieved was 90% by using the hybrid technique of GA[2] and neural networks.

"Prediction and Diagnosis of Heart Disease[9] by Data Mining Techniques" designed by Boshra Bahrami, Mirsaeid Hosseini Shirvani. This paper uses various classification methodology for diagnosing cardiovascular disease. Classifiers like KNN, SVO classifier and Decision Tree are used to divide the datasets. Once the classification and performance evaluation the Decision tree is examined as the best one for cardiovascular disease prediction[8-9] from the dataset.

IV.METHDOLOGY OF SYSTEM

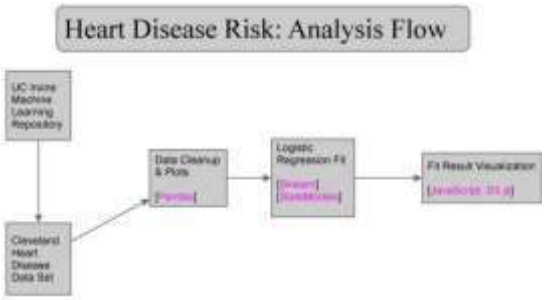
Processing of system start with the data collection for this we uses the UCI repository dataset which is well verified by number of researchers and authority of the UCI
. A. Data Collection First step for predication system is data collection and deciding about the training and testingdataset. In this project we have used 73% training dataset and 37% dataset used as testing dataset the system. B. Attribute Selection Attribute of dataset are property of dataset which are used for system and for heart many attributes are like heart bit rate of person, gender of the person, age of the person and many more shown in TABLE.1 for
predication system

The dataset used was the Heart disease[1-2] Dataset which is a combination f 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features . Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis.

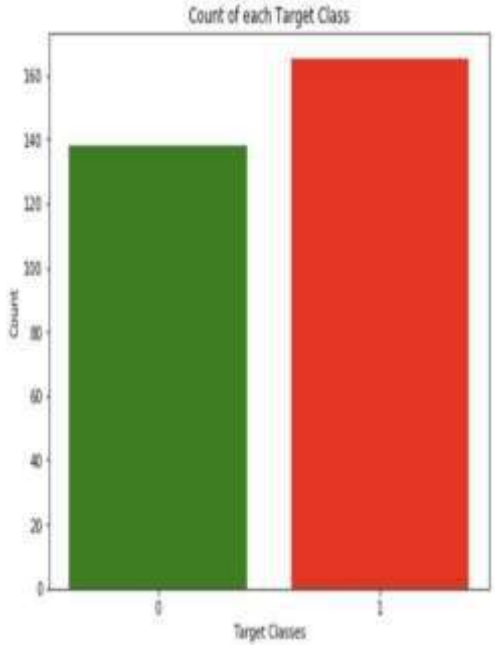
Attributes mentioned are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques . The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further.

In our case, we use a data set of people who have performed analyses and tests to detect heart disease. The data set is a matrix where the rows represent the patients and the columns represent the factors or attributes (features) to be tested. Preprocessing needed for achieving prestigious result from the machine learning algorithms[1]. For example Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data. For our project we have to convert some categorized value by dummy value means in the form of "0"and "1" by using following code.

.



FLOW CHART:



IMPLEMENTATION

1) Impute Missing Values By Knn: knn for missing values working by calculate the distance or similarity to find the most similar case in the dataset and change the missing value with it .by applying .

2) Min Max Normalization: This method is convert each numerical feature value into new value depending on the minimum and maximum values of the feature , by applying .

X = X − M in

Where Min is the smallest value in the selected feature, Max is the biggest value in the selected feature, *X* is a new select value after applying normalization, X is a selected value from a numerical feature.

3) Z-Score Standardization*:* This method is convert each numerical feature value into new value depending on the standard deviation and Mean of the feature , by applying

X= – *X*

.

4) One Hot Encoding: One Hot Encoding splits the categorical feature into a separate number of features depending on the number of the cases in the original categorical feature, and give 0 for absence and 1 for presence in each new feature.

5) **Ordinal Encoding***:* In this technique, each case in the categorical feature is converted into integer value. **6)Equal Width Discretization**: This is an easy method that sorting the values of numerical feature and split the range of sorting values into predefifined equal-width bins by applying*:*

Where W is the width of the bin, V Max is the maximum value in the selected numerical feature, V Min is the minimum in the selected numerical feature, i
= 1. . .
. . k-1.

7)Equal Frequency Discretization: In this method, fifirstly sorting the values in ascending order. Split the range of sorting values into predefifined number of equal frequency bins by applying N\K , each bin has the same number of values.

C. Data Balancing:

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where “0” represents with heart diseases patient and “1” represents no heart diseases patients.

For classification tasks, one may encounter situations where the target class label is un-equally distributed across various classes. Such conditions are termed as an Imbalanced target class. Modeling an imbalanced dataset is a major challenge faced by data scientists, as due to the presence of an imbalance in the data the model becomes biased towards the majority class prediction.

V.ALGORITHMS

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine learning algorithms[1-2] are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine

learning[1-2] use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

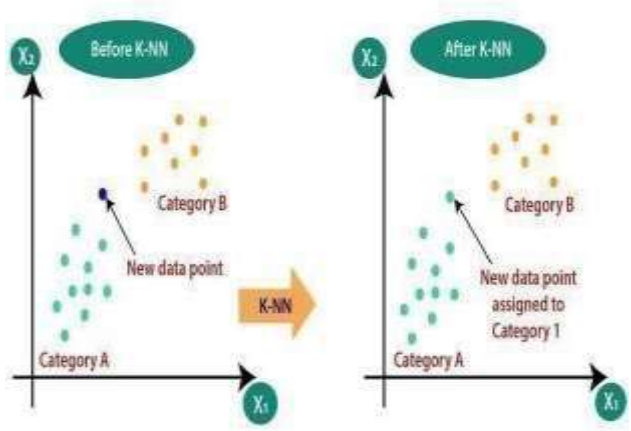
Decision tree :

On the other hand decision tree is the graphical representation of the data and it is also the kind of supervised machine learning algorithms. Fig.6 Decision tree For the tree construction we use entropyof the data attributes and on the basis of attribute rootand other nodes are drawn. Entropy= -∑ Pij log Pij (1)In the above equation of entropy (1) Pij is probability of the node and according to it the entropy of each node is calculated. The node which have highest entropy calculation is selected as the root node and this process is repeated until all the nodes of the tree are calculated or until the tree constructed. When thenumber of nodes are imbalanced then tree is create the over fitting problem which is not good for thecalculation and this is one of reason why decision tree have less accuracy as compare to linear regression.

Support Vector Machine:

It is one category of machine learning technique which work on the concept of hyperplan means it classify the data by creating hyper plan between them. Training sample dataset is (Yi, Xi) where i=1,2,3,…….n and Xi is the ith vector, Yi is the targetvector. Number of hyper plan decide the type of support vector such as example if a line is used as hyper plan then method is called linear support vector.

K-nearest Neighbor: It work on the basis of distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other and number of neighbor are decided by the user whichplay very crucial role in analysis of the dataset.



VI.FUTURE SCOPE

Heart is one of the essential and vital organs of human body and prediction about heart diseases[1-4] isalso important concern for the human beings so that the accuracy for algorithm is one of parameterfor analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find KNN is best one. For theFuture Scope more machine learning approach will be used for best analysis of the heart diseases[4] and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

VII.CONCLUSION

A cardiovascular disease[5] detection model has been developed using three ML classification modeling techniques. This project predicts people with cardiovascular disease by extracting the patientmedical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection[1-4] system assists a patient based on his/her clinical information of them beendiagnosed with a previous heart disease. The algorithms used in building the given model are support vector , and KNN . The accuracy of our model is 87.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not . By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying support vector machine[8] and KNN to get an accuracyof an average of 87% on our model which is better than the models having an accuracy of 85%. Also, it is concluded that accuracy of KNN is highest between four algorithms that we have used i.e. 87%

The computational time was also reduced which is helpful when deploying a model. It was also found out that the dataset should be normalized; otherwise, the training model gets overfitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real- world data problems which can vary drastically to the dataset on which the model was trained. It wasalso found out
that the statistical analysis is also important

when a dataset is analyzed and it should have a Gaussian distribution, and then the outlier's detection is also important and a technique known as Isolation Forest is used for handling this. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning[8-9] and ML as well. The algorithm applied byus in ANN architecture increased the accuracy which we compared with the different researchers.The dataset size can be increased and then deep learning with various other optimizations can be used and more promising results can be achieved. Machine learning and various other optimization techniques can also be used so that the evaluation results can again be
used and the results can be compared. And more ways could be found where we could integrate heart-disease- trained ML and DL models with certain multimedia for the ease of patients and doctors.

VIII.REFERENCES

1. Santana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms"ICIICT,.

2. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease PredictionUsing Hybrid Machine Learning Techniques” IEEE Access.

3.Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using MachineA Survey” InternationalJournal on Recent and Innovation

4.Trends in Computing and Communication Volume: 5 Issue: 8, IJRITCC August 2017.
M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Predictionof Heart Diseases Using Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining

5.Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

6.M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using

7.S. Kumra, R. Saxena, and S. Mehta, “An ExtensiveReview on Swarm Robotics,” pp.

8.Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “ A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” Advances in Computational Sciences and Technology, Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique” Journals of Computer Science & Electronics Chavan Patil, A.B. and Sonawane, P.“To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-

9.M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, “Prediction of risk score for heart disease using associative classification and hybrid feature subset selection,” Int. Conf.using Machine Learning”, Proceedings of the 2nd International conference on Electronics,Communication and Aerospace Technology (ICECA),