International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.8, August 2021 E-ISSN: 2347-2693

Study of Predicting Heart Diseases Using KNN, Decision Tree and Random Forest Methods

Devulapalli Sudheer[1*], Anupama Potti[2], N. Anjali devi[3], C. Chandana Reddy[4]

[1,2,3,4]Department of Computational Sciences and Engineering, Sree Dattha Institute of Engineering and Science, Sheriguda, Rangareddy District, Telangana, India.

[*]*Corresponding Author: 2498sudheer@gmail.com Tel.: 9966427152*

DOI: https://doi.org/10.26438/ijcse/v9i8.2729 | Available online at: www.ijcseonline.org

Received: 16/Aug/2021, Accepted: 20/Aug/2021, Published: 31/Aug/2021

Abstract— Healthcare is a sought after task in the human life. One in four deaths are due to heart disease in India alone. In order to reduce the number of deaths, there is a need to automate the prediction process and alert the patient well in advance. Healthcare industry contains a lot of medical data which aids machine learning algorithms in making decisions accurately in predicting the heart diseases. This project makes use of the heart disease dataset available in Cleveland database of UCI machine learning repository. This project has delved into different algorithms namely Decision tree, k-nearest neighbour algorithm (KNN), Random Forests. The database consists of 303 instances and 14 attributes. Using the decision tree algorithm, we will be able to identify those attributes which are the best one that will lead us to a better prediction of the datasets. Here each internal node of the tree represents an attribute, and each leaf node corresponds to a class label. Random Forests consists of multiple decision trees that operate as an Ensemble. Random Forests out perform as they are collection of large relatively uncorrelated models. KNN can easily identify and classify people with heart disease from healthy people. The proposed project compares the results using different performance measures, i.e. accuracy, precision, etc. This project delivers the prediction valued from no presence to likely presence. The proposed project's aim is to try and reduce the occurrences of heart diseases in patients and thus assist doctors in diagnose it effectively.

Key words — Health care, Prediction, Random Forest, Classification, Machine Learning.

I. INTRODUCTION

Heart disease (HD) is one of the most common diseases now a days ,due to number of contributing factors, such as high blood pressure, diabetes, cholesterol fluctuation, exhaustion and many others. An early diagnosis of such disease has been sought for many years, and many data analytics tools have been applied to help healthcare providers to identify some of the early signs of HD. Many tests can be performed on potential patients to take the extra precautions measures to reduce the effect of having such a disease, and reliable methods to predict early stages of HD, such as the methods proposed in this project, can be a crucial task for saving lives. Number of Machine Learning (ML) algorithms, such as, K- Nearest Neighbor (K-NN), Decision tree, Random Forest were applied for the purpose of classification and prediction of HD dataset, and many promising results were presented in the literature. This project makes use of the heart disease dataset available in Clevel and database of UCI machine learning repository. The database consists of 303 instances and 14 attributes. The proposed project compares the results using different performance measures, i.e. accuracy, precision, etc. This project delivers the prediction valued from no presence to likely presence. The proposed project's aim is to try and reduce the occurrences of heart diseases in patients and thus assist doctors in diagnose it effectively.

II. DATASET DISCRIPTION

The dataset consists of 303 individual data. There are 14 columns in the dataset. Our dataset contains the following attributes:

Age:
It displays the age of the individual.

Sex:
It displays the gender of the individual using the following format: 1 = male
0 = female

Chest-pain type:
It displays the type of chest-pain experienced by the individual using the following format:
1 = typical angina
2 = atypical angina
3=non –anginal
4=asymptotic

Angina is chest pain or discomfort caused when your heart muscles does not get enough oxygen-richblood.Butanginaisnotadisease.Itisasymptomofunderlyingheartproblem.

1. Typical angina: It is usually a dull pain or pressure sensation.
2. Atypical angina: When one experiences chest pain that does not meet the criteria for angina it is known as a typical angina.
3. Non – angina pain: It is also known as Non cardiac chest pain is the term that is used to describe pain in the chest that is not caused by heart disease or a heart attack. The patient feels a pressure or squeezing pain behind the breast bone. The pain can lost for a few minutes or for hours.
4. Asymptotic: A silent heart attack is a heart attack that has few, if any or has symptoms you don't recognize as assignofheartattack.Peoplewithcp1,2,3aremorelikelytohave heart disease than people with cp0

III. METHODOLOGY

The architecture of the proposed system is as displayed in the figure 1. The major components of the architecture are as follows: patient database, preprocessing, tokenization, training the model, test the model, design fitness function, application of genetic algorithm, results collection and prediction of heart disease. The proposed method takes the input from the dataset and applies validation process and finding the mean of samples to converting categorical to numerical. The pre processing steps involved finding the missing values, detecting and removing outliers. Applying the feature selection and reduction techniques like PCA to reduce the correlation between the features. To study the performance of classification process with the selected features and applied various classification techniques such as decision tree, random forest.

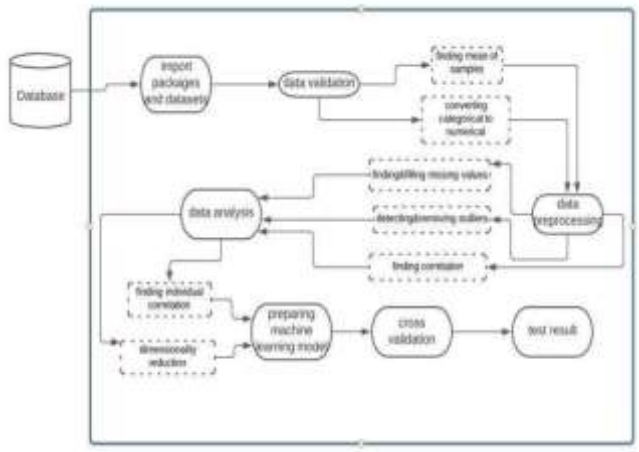

Figure 1. Architecture of the proposed system.

3.1 Pseudo code

Step 1: Import the required packages.
Step 2: Load the dataset.
Step 3: Summarizing the dataset
Step 4: Applying the data mining techniques like data preprocessing.
Step 5: Applying the feature selection and reduction process to normalize the data. Step 6: Visualizing the dataset (Correlation matrix)
Step 7: Implementing the algorithms mentioned- K Nearest Neighbors (KNN), Decision Trees, and Random Forest.
Step 8: Finding the accuracy using of algorithms.

1.2 Random Forest

Random Forest can be used for both classification and regression problems. Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach. The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions. Through the decision tree algorithm, you can generate the rules. You can then input the features of this movie and see whether it will be liked by your daughter.

The process of calculating these nodes and forming the rules is using information gain and Gini index calculations. The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the processes of finding the root node and splitting the feature nodes will run randomly. Over fitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't over fit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

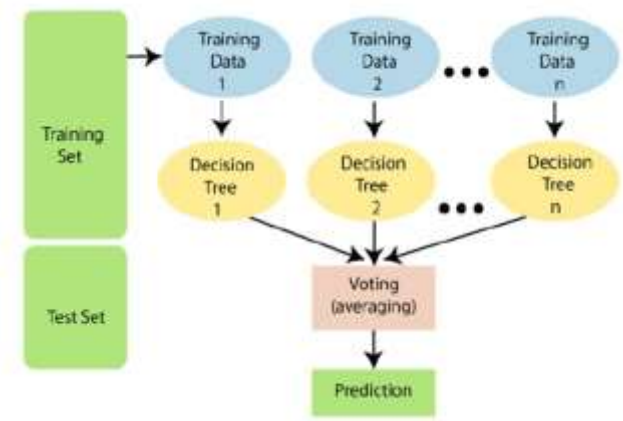

Figure 2. Random Forest block diagram.

IV. RESULTS AND DISCUSSION

The results show that the highly effected heart diseases in the age parameter wise. The confusion metrics will explain the false positive and true positive classifications results and their accuracies at individual categories in Figure 3.

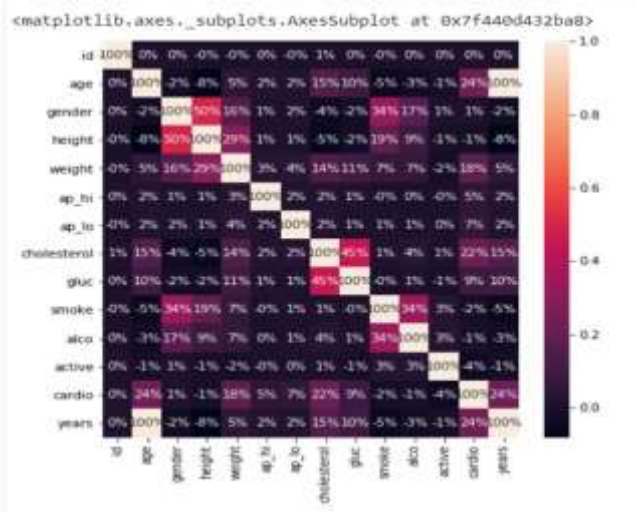

Figure 3. Confusion metrics for heart disease detection.

V. CONCLUSION AND FUTURE SCOPE

It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data. In this project, we proposed a method for heart disease prediction using machine learning techniques, these results showed a great accuracy standard for producing a better estimation result. We have concluded that by using K Nearest Neighbors we get accuracy of 91% which is highest among random forest (90.16%) and Decision Trees (81.97%) and it is highest among the existing system which is developed using Support Vector Machines and Naive Bayes. this what we found is during small datasets in some other cases most of time decision trees direct us to a solution which is not accurate, but when we look at KNN results we are getting more accurate results and random forest results with probabilities of all other possibilities but due to guidance to only one solution decision trees may miss lead.

REFERENCES

[1] Vembandasamy, K., R. Sasipriya, and E. Deepa. "Heart diseases detection using Naive Bayes algorithm." *International Journal of Innovative Science, Engineering & Technology* 2.9: **441-444, 2015.**

[2] Kumar, Priyan Malarvizhi, and Usha Devi Gandhi. "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases." *Computers & Electrical Engineering* 65: **222-235, 2018.**

[3] Alarsan, Fajr Ibrahem, and Mamoon Younes. "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms." *Journal of Big Data* 6.1: **1-15, 2019.**

[4] 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. **63-72, 2019.**

[5] Dhar, Sanchayita, et al. "A hybrid machine learning approach for prediction of heart diseases." *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, **2018.**

[6] Singh, Jagdeep, Amit Kamra, and Harbhag Singh. "Prediction of heart diseases using associative classification." *2016 5th International conference on wireless networks and embedded systems (WECON)*. IEEE, **2016.**

[7] Devulapalli, Sudheer, et al. "Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques." *Materials Today: Proceedings*, **2021.**

[8] Sudheer, D., R. SethuMadhavi, and P. Balakrishnan. "Edge and Texture Feature Extraction Using Canny and Haralick Textures on SPARK Cluster." *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*. Springer, Singapore, **2019.**

AUTHORS PROFILE

Mr.D.Sudheer currently working as Assistant Professor in Department of Computer Sceince and Engineering in Sree Dattha Enginneing And Science.He Completed Master's and Bachelor's degree from JNTU Kakinada in Compter Science and Engineering.Interested Research areas are Machine Learning,Deep Learning,Computer vision.

Mrs.P.Anupama currently working as Assistant Professor in Department of Computer Sceince and Engineering in Sree Dattha Enginneing And Science.She Completed Master's degree from JNTU Hyderabad and Bachelor Degree from Andhra University in Compter Science and Engineering.Interested Research areas are Machine Learning,Network Security. She has 7 years of teaching experience.

