

Heart Disease Prediction using KNN classification approach

Gagandeep Kaur^{1*}, Anshu Sharma², Anurag Sharma³

¹Department of Computer Science & Engineering, CTITR, Jalandhar, Punjab India

²Department of Computer Science & Engineering, CTITR, Jalandhar, Punjab India

³Department of ECE, CTITR, Jalandhar, Punjab India

DOI: <https://doi.org/10.26438/ijcse/v7i5.416420> | Available online at: www.ijcseonline.org

16/May/2019, Published: 31/May/2019

Abstract— In recent ten years, heart failure becomes the leading cause of death in whole world which is estimated by World Health Organization (WHO). Several types of heart diseases are expanding day by day because of way of life, genetic problem, blood pressure, cholesterol level, pulse rate etc. So the diagnose of disease plays important role for the prevention of heart related problems. Researchers received different methods to analyze it. These days the utilization of system innovation in the fields of medication zone, finding treatment of disease and patient activity has exceptionally expanded. The aim of this paper is to design a KNN based classification approach for prediction of the Heart failure which assists the doctors to identify disease easily. It is an intelligent classification approach because it provides accurate result. To accomplish the diagnosis process taken different risk factor, signs and symptoms from patients and experts. Classification approach consists of two algorithms such as KNN classification algorithm and Decision tree algorithm. The result of classification shows 86% accuracy by using n no. of neighbors in this approach.

Keywords—Classification, KNN, Decision Tree, cross validation.

I. INTRODUCTION

In these days Data Mining is getting to be noticeable famous in healthcare field. Just because nowadays there is need of proficient investigative techniques in healthcare industries. Data mining is a procedure of examining information from alternate point of view and turnout the knowledge from it. Data mining is one of the tool which can be used for many of disease prediction such as heart disease, breast cancer and lung cancer etc. The objective of this review paper is to analyze the different Data mining techniques for predicting the heart disease prediction [12].

Many of hazards factors for heart disease for example, age, sexual orientations, hypertension, smoking, family history etc. Heart disease involves various types of conditions that can influence the heart work in which types including heart strokes, heart failure, heart attacks, heart valve disease, cardiovascular disease which is leading cause of death over past few years. Clinical choices are frequently made of in views of specialist and experience rather than to on the knowledgeable data information covered up in database. Patients and doctors need genuine detailed information about the chance of growing heart disease.

Data Mining joins measurable examination, machine learning and database innovation to separate hidden paradigm and connections from huge database. They utilize distinctive methodologies for producing the grouping models, which expands the odds for finding a prediction display with high characterization exactness. Decision tree algorithm recursively separate perceptions in branches to develop a tree with the end goal of enhancing the expectation exactness. Neural Network is systematic method in which subjective framework and neurological elements of brain and capable predicting new observations from different perceptions in the wake of executing a procedure of existing information

II. RELATED WORK

This section intends to decompose the different Data Mining methods which are presented in late years for heart disease prediction. The main point of all these papers is to achieve better exactness and make the system more productive so that it can forecast the possibilities of heart disease. Different data mining techniques for analysis and accomplish various probabilities for various strategies have been studied by

using different classification, clustering and association techniques.

AH Chen et al (2011) develop a heart disease predict system that can assist medical professionals in predicting heart disease status based on the clinical data of patients. The approach includes three main steps in which first step involve the selection of 13 important clinical features. Step second was to develop ANN algorithm for classifying heart disease based on features and third step to develop user-friendly heart disease prediction system. The accuracy of prediction system was near 80% [2].

Nidhi Bhatla et al (2012) projected the study of different data mining techniques that can be employed in automated heart disease prediction systems. The analysis shows that neural network with 15 attributes has shown the highest accuracy. On the other hand, Decision tree has also performed well with 99.62% accuracy by using 15 attributes [10].

M.Akhil Jabbar et al (2013) introduced a lazy associative classification for prediction of Heart Disease in Andhra Pradesh and present some experimental results by using 7 data sets from UCI Repository. Researcher applied information centric attribute measure PCA to generate class association rule which will use to predict the occurrence of heart disease. Researchers design the system for Andhra Pradesh population because this state is in risk of most deaths due to Heart Disease [3].

R. Chitra et al. (2013) developed the computer aided heart disease prediction system that helps the physician as a tool for heart disease diagnosis. From the analysis it is concluded that neural network with offline training is good for disease prediction in early stage and good performance can be obtained by pre-processed and normalized dataset [11].

Theresa Princy. R et al (2016) focused on different classification techniques used for predicting the risk level of each person based on attribute age, gender, blood pressure, cholesterol level, and pulse rate. Data mining techniques like Naïve Bayes, KNN, Decision Tree, Neural Network used for the better accuracy of risk level. Using KNN and ID3 algorithm the risk rate of heart disease was detected and accuracy level also provided for different number of attributes [13].

Sarath Babu et al introduced database information which fed into several classifiers like KNN, naïve bayes, SVM and decision trees, each of them perform some specific tasks. These techniques are used to predict heart diseases at their early stage. It shows very effective performance in order to achieve the correct and perfect diagnose for the heart related

diseases. There are certain advantages of this approach such as the diseases can be predicted at their very initial stages and can be diagnosed correctly and properly on time. Therefore, the researcher concluded that, Decision tree has good efficiency while using 14 attributes and is enormous method for preventing heart related issues [17].

Monira Islam et al (2017) proposed a non-invasive technique to detect the heart rate from Photoplethysmography (PPG) signal. A suitable heart diseases detection mechanism is used with the help of PPG and extracted from human facial videos. This can overcome the expensive ECG machine for heart related disease detection. The FFT can detect the heart rate and re-check with the conventional heart rate using ECG machines. The contactless approach provides the solution and identifies the heart related disease it could be diagnosed by the doctor. Therefore, the result of this approach concludes that it ensures the safety by diagnosing the diseases by its early symptoms. It also concludes that green channel predicts more accurate heart rate than any other. It has been found 97.7% accurate results from green channel [18].

Tülay Karayölan et.al (2017) proposed back propagation algorithm for the prediction of heart diseases with the help of Artificial Neural Networks. 13 clinical attributes were used as input of Neural networks and is trained along with this proposed back propagation technique which can predict the heart related diseases with an accuracy of 95%. Therefore, the researcher draws the conclusion that the proposed approach has almost 100% accuracy in prediction heart related diseases at their early stages. It gives better results in comparison to the other techniques [19].

Tahira Mahboob et.al (2017) studied about various learning applications which assist the detection of numerous heart diseases. There are certain techniques were used like Hidden Markov Models, Support Vector Machine, Computational intelligent classifier, and data mining and so on. As the cardiac diseases treatment is very expensive and unaffordable to any normal individual so, these types of advanced technology are developed to overcome this problem. These techniques are also useful in early stage predictions. It avoids any other future sufferings by making slight changes in daily routine. Hence, the author concludes that the predicted approach has several advantages and is very useful [20].

Procheta Nag et.al (2017) builds up a very effective technique which is very useful in the prediction of heart diseases at the initial stage. 25 attributes related to symptoms of heart attack like chest pain, breathing problem, palpitation,

vomiting and continuous sweating. Researchers developed a prototype by integrating clinical data collected from patients admitted in different hospitals attacked by Acute Myocardial Infarction (AMI). Therefore, the researcher draw the conclusion that the advancement of computer technology in medical and health region provides useful aids and people are becoming more dependent on these technologies [4].

Syedamin Pouriyeh et.al (2017) recognizes gaps in research on machine learning method for the detection of heart diseases. The data related to the heart diseases consist of 303 instances for the training and development of the system. For the increment in the amount of data 10-Fold Cross-Validation has been used, otherwise there would be very limited amount of data available. Several approaches have used like; Decision Tree, naïve Bayes, Multilayer Perceptron, Support Vector Machine and so on. Researcher compares the proposed algorithm with the different machine learning algorithms and concludes that the SVM technique boosts and outperforms as compared to any other [5].

III. METHODOLOGY

This segment outlines the stages of proposed framework. We proposed a technique to predict the heart disease using classification approach. Following are the various steps:-

- 1. Pre-handling:** The pre-processing is the first stage of the heart disease prediction research process. In this phase, the dataset is taken as input and data set is cleared by discarding missing qualities.
- 2. Highlight Extraction:** In the second stage, the features of the input data are extracted for the classification and the connection between the different attributes is buildup for the recognizable proof of main attributes from the dataset. The strategy of decision tree classifier is applied in this stage for the feature extraction.
- 3. Grouping:** In the last stage, the relationship between the attributes is taken as input for this stage, the input data is divided into training set and test set. The algorithm of decision tree will be applied on the extracted features for the final classification. The decision tree will behave as base classifier for the classification, which is the classification approach used for prediction. The execution of proposed approach will be investigated as far as specific parameters like exactness, accuracy, review and f-measure.

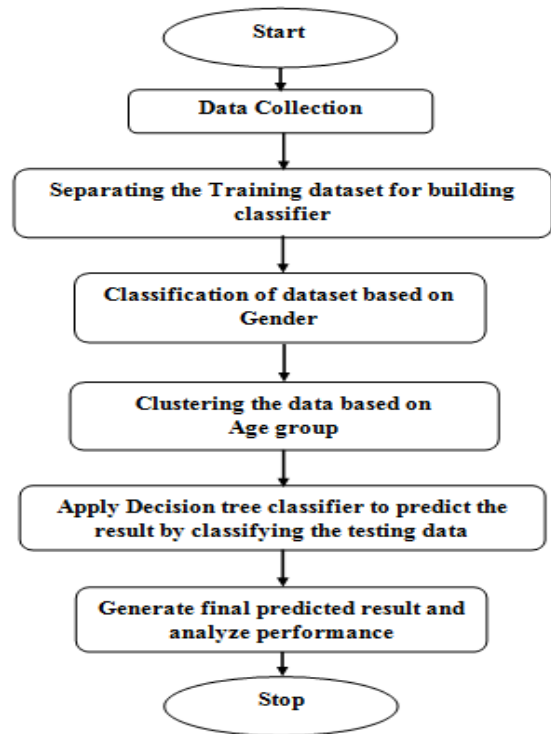


Fig 1. Proposed methodology

IV. RESULTS AND DISCUSSION

Cleveland Clinic Foundation dataset known as "Cleveland Clinic Foundation Heart Disease Dataset" utilized for the examination. Dataset include 14 attributes in which 303 samples as a patient's data were consider in it from which 13 attributes such as age, sex, chest pain, serum-cholesterol, fasting blood sugar, electrocardiographic, heart rate, induced angina, old peak, slope and that are taken as input parameters and last 1 attribute define the diagnosis of Heart disease. Specific data mining techniques were applied on the dataset by using cross validation method to validate the results. This technique classified the data into two parts such as training set and testing set from which the dataset with small number of samples were consider in testing dataset and data set with large number of sample were consider in training dataset. The benefit of contribution to weight transformation is that it decreases the quantity of correlations between the client esteems and the records of the dataset.

K-Nearest Neighbor and Decision tree were applied to the dataset to show the sensitivity, specificity and accuracy of these techniques.

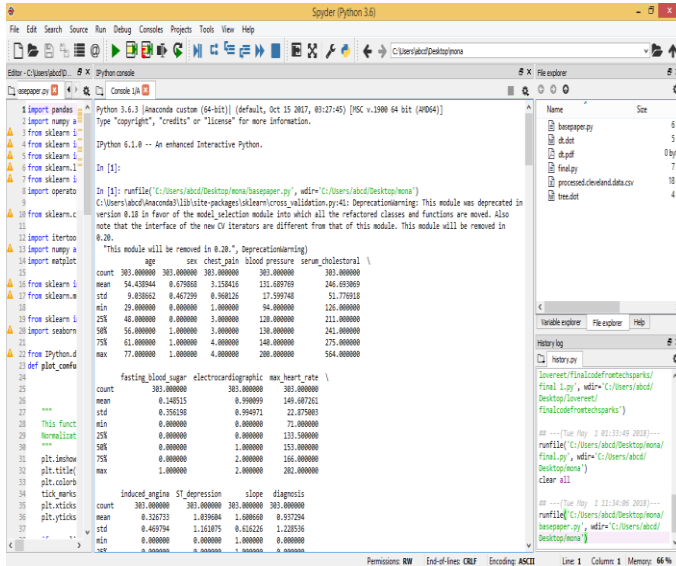


Fig 2: Data Description

As shown in figure 2, the dataset of heart disease is taken from the UCI repository. In this figure, dataset is described which mean value, standard deviation etc.

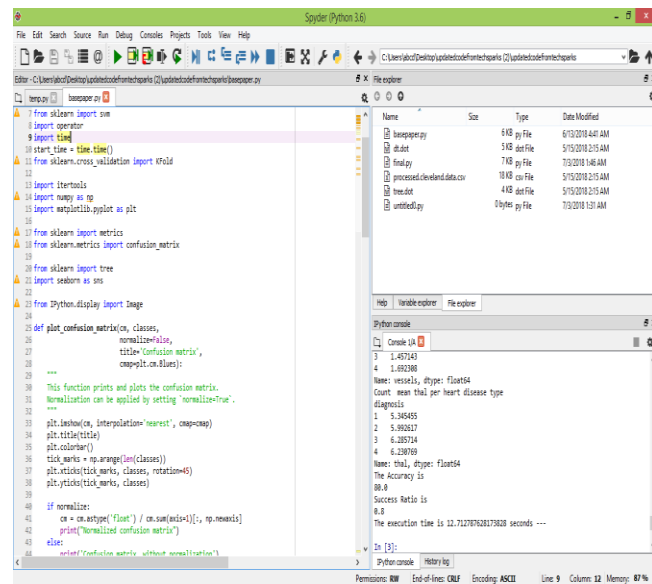


Fig 3: Apply Existing Approach

As shown in figure 3, the dataset of heart disease is taken from the UCI repository. The k-nearest neighbor algorithm is applied with Decision tree and its accuracy is 86 percent. The basis k-nn algorithm is applied which is not able to drive exact relationship between attributes and due to which accuracy is less.

In [1]: Number of records: 303
Number of variables: 14
['sex', 'chest_pain', 'fasting_blood_sugar', 'electrocardiographic', 'induced_angina', 'slope', 'no_of_vessels', 'thal', 'diagnosis']

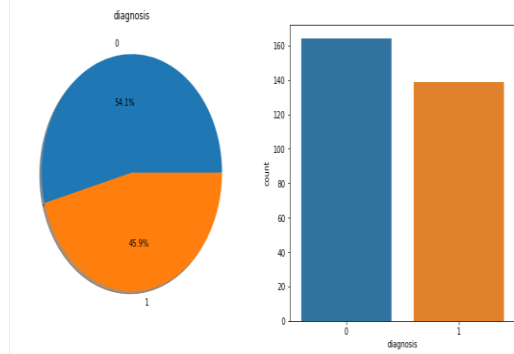


Fig 4: Comparison of diagnosis based on Gender

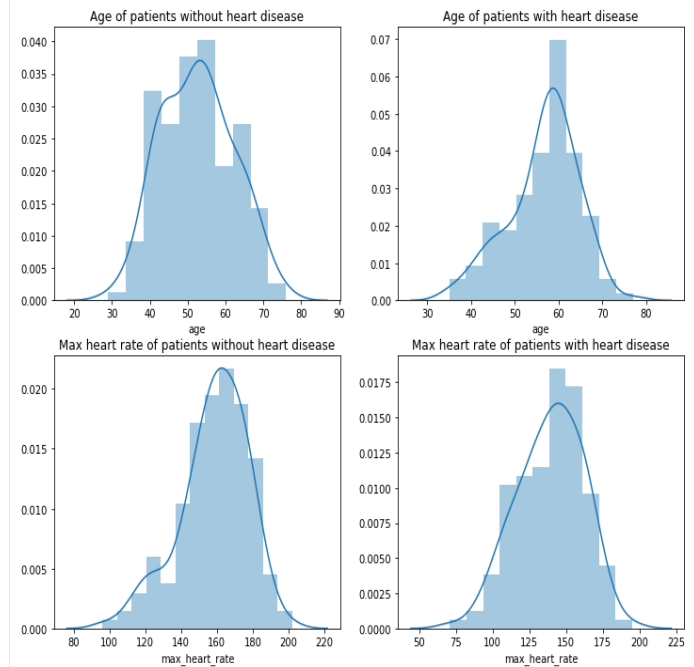


Fig 5: Comparison of patient's age and Max. heart rate based on diagnosis

The dataset is divide into training set and testing set by using ration of 70:30 respectively. In training dataset 212 samples of dataset are used and 91 samples are consider in testing dataset. So the overall accuracy achieved from training and testing datasets are 88.21 and 86.81 respectively. On the other hand by increase the n- neighbors the accuracy of both divided datasets are:

Value of n neighbors	Train Accuracy (%)	Test accuracy (%)
1	100	74
2	87.74	79.12
3	90.57	83.52
4	87.74	84.62
5	88.21	86.81
6	85.38	86.81
7	87.26	86.80
8	85.38	85.71
9	86.32	85.71

Fig. 6: Accuracy analysis of Training and Testing dataset

V. CONCLUSION AND FUTURE SCOPE

A most extreme expensive and predominant medical issue anguished by the present age is the diverse heart ailments. Consequently early foreseeing the sickness helps in turning away the future sufferings by slight changes in the way of life already. Then again finding of heart disease assumes a significant job to think about the status of heart patients. This work is identified with supervised learning for Heart disease prediction analysis. The technique K-nearest Neighbor is applied for prediction analysis and Decision tree is applied for classification which results the accuracy of 86% with KNN classifier. The accuracy analysis shows that by increasing the number of n- neighbours it gives slightly difference in the average. There is a huge extent of progress in KNN classifier which has higher problem because of which execution time is high for the prediction examine. So the framework can be improved by utilizing successful classifier to lower the problem.

REFERENCES

- [1] Umar S., Agarwal K., Rizwan Beg, "Genetic Neural Network based Data Mining in Prediction of Heart Disease using Risk Factors", IEEE, 2013.
- [2] Chen A.H, Huang S.Y, Hong P.S, Cheng C.H, E J Lin, "HDPS: Heart Disease Prediction System", IEEE, 2011.
- [3] Jabbar M.A., Deekshatulu B. L., Chandra P., "Heart Disease Prediction using Lazy Associative Classification", IEEE, 2013.
- [4] Nag P., Mondal S., Ahmed F., More A. and Raihan M., "A Simple Acute Myocardial Infarction (Heart Attack) Prediction System Using Clinical Data and Data Mining Techniques", IEEE, 2017.
- [5] Amutha J., Padmajavalli R, Prabhakar D., "A Novel approach for the Prediction of Treadmill Test in Cardiology using Data mining Algorithms implemented as a Mobile Application, ScienceDirect, 2018.
- [6] Pouriyeh S., Vahid S., Sannino G., Giuseppe De Pietro, Arabnia H, Gutierrez J., "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease", IEEE, 2017
- [7] Priyanga and Dr. Naveen, "Web Analytics Support System for Prediction of Heart Disease Using Naïve Bayes Weighted Approach (NBwa)", IEEE, 2017.
- [8] Priyanka N B.E., (M.Tech) and Dr.Pushpa RaviKumar B.E., M.Tech., Ph.D, "Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree", IEEE, 2017.
- [9] Srinivas K., Rao G.R., Govardhan A., "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", IEEE, 2010.
- [10] Rajkumar A, Sophia G., Reena, "Diagnosis Of Heart Disease Using Data mining Algorithm", Global Journal of Computer Science and Technology, volume 10, 2010.
- [11] Bhatla N., Jyoti K., "An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering Research & Technology, Volume 1, Issue 8, 2012.
- [12] Chitra R., Seenivasagam V., "Review of heart disease prediction system using data mining and hybrid intelligent techniques", ICTACT journal on soft computing, volume: 03, issue: 04, 2013.
- [13] Era S.K., Nishika," Prediction of Heart Disease using Data Mining Techniques", International Journal of Advance Research , Ideas and Innovations in Technology, Volume: 2, Issue-03, 2016.
- [14] Princy T. R, Thomas J., "Human Heart Disease Prediction System using Data Mining Techniques, IEEE, 2016.
- [15] Assari R., Azimi P., Taghva M.R," Heart Disease Diagnosis Using Data Mining Techniques", International Journal of Economics & Management Sciences, volume:6, Issue-03, 2017.
- [16] Andreeva P., "Data Modelling and Specific Rule Generation via Data Mining Techniques", International Conference on Computer Systems and Technologies, 2016.
- [17] Polat K., Sahan S., Gunes S., "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing", ScienceDirect, 2007.
- [18] Babu S., Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique", IEEE, 2017.
- [19] Monira I., Ashikuzzaman M.D., Tabassum T. and Yusuf S.U., "A Non-Invasive Technique of Early Heart Diseases Prediction from Photoplethysmography Signal", IEEE, 2017.
- [20] Karayölan T., Köllüç O., "Prediction of Heart Disease Using Neural Network", IEEE, 2017.
- [21] Mahboob T., Irfan R. , Ghaffar R., "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", IEEE, 2017.
- [22] Malav A., Kadam K., Kamat P., "Prediction of heart disease using k-means and Artificial Neural network as hybrid approach to improve accuracy" International Journal of Engineering and Technology, volume: 09, 2017.
- [23] [online] Available: <https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f>
- [24] [online] Available: https://www.python-course.eu/k_nearest_neighbor_classifier.php
- [25] [Online] Available: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>