International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Improving the Accuracy of the Heart Disease Prediction Using Hybrid Machine Learning

Sruthi S[1], Kanaga Priya K[2], S Rama[3]

[1,2] Under Graduate, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India
[3] Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

---

Abstract - *Heart Disease is the main threat for death. This paper used Machine Learning Techniques, as it proved to show a beneficial result in making decision from large quantities of dataset like blood pressure level, cholesterol, chest pain types and many more biomarkers, which have been collected from hospitals. After which data pre-processing was applied to the dataset and there are 13 variables in which 2 variables concerning to age and gender are used as a major criteria. Below 10 years, heart disease probability is less. Then the classification modelling to evaluate the performance was done and the optimal models were recognized from the low rate of error, Decision Tree , Support Vector Machine, Random Forest methods were utilized. In addition, we have built a web application. Using these algorithms, we have got a better accuracy. The predictive model of heart disease with the hybrid combination of K-nearest neighbor and fuzzy K-nearest neighbour performed with 94% accuracy.*

***Key Words*: Heart disease, Classification modelling, Machine Learning, Biomarkers, web application.**

1. INTRODUCTION

Heart disease has been a major threat in recent days. Age is one of the main factors to be considered for heart disease prevalent among people mostly above 50[1]. In some exceptional cases, people below this age group also suffers from this disease and in other cases it directly leads to incurable complications. Women suffers a lot than men from heart diseases[2]. Age, gender, chest pain types, cholesterol, blood sugar level, angina induced, blood pressure rate while resting, maximum rate of heart are taken as the main risk factors to predict the accuracy to find the optimal algorithm among all. The dataset used in this paper are supervised dataset where we have labelled variables and they were collected from websites.

We have also been seeing many other ways to solve the issue and from the related works, Machine Learning has shown an effective way to get a better accuracy. Generally to predict a cardiovascular disease, it is time consuming and costlier in hospitals and scan centres. In order to reduce them, many systems and predicting models have shown up and from one of those existing system we are improvising the accuracy rate of heart disease prediction using random forest as the preliminary stage and then using the hybrid of KNN and fuzzy K-nearest neighbour algorithm to get more accuracy than the existing one. To predict heart disease needed bulk data. Set. The dataset consisted of variables like age, gender, cholesterol(chol), blood sugar level(fbs), chest pain types(cp), Maximum rate of heart(thali), level of blood pressure at resting mode(trestbps). The datasets were collected from UCI repository and few from scan centre. The modules which we used in our paper was pre- processing, feature selection, classification modelling and performance measures.

The raw dataset is often inconsistent with many errors, so it is pre-processed where all the null values and missing values will be removed and later will be refilled with proper values. A separate test dataset is used to test the accuracy of the predictive models and hence the test dataset is cleaned and pre-processed the same way as the training dataset.

After which feature selection process is carried on where it removes irrelevant variables and improves accuracy and reduces training time. There are four classifiers, K-neighbour, Support vector machine, Random forest classifier, Decision Tree. The classifiers processed each dataset and verified to generate the estimated accuracy of the prediction. The classifier which had more accuracy was considered the best. Kernel is a type of algorithm used for pattern analysis where to study the relations like correlations, clustering, classifications. It is a mathematical method of using linear classifications to classify all the non-linear points. In Machine Learning, certain problems will have more weighting functions and to avoid all complications kernels are used.[3] We have used cross validation to select the right kernel. Four kernels such as linear, poly, sigmoid and radial basis function have been used. We have also used the Fuzzy K-nearest neighbour algorithm to get more accuracy rate then the random forest classifier.

In this work, the hybrid of Fuzzy KNN and KNN has been used. In the proposed system we have improved the accuracy using fuzzy k-nearest neighbor algorithm. The

experimental results have shown an efficient way to analyse and predict the heart disease.

2. RELATED WORK

[4]Researchers have worked on diagnosis of heart disease prediction, which was quite hard to achieve. V. Krishnaiah, G. Narasimha, N. Subhash Chandra (2016) had collected several datasets and worked on them. The problem was that, all health issues related to heart are usually diagnosed by doctors and their guidance. In order to reduce them, Computer Aided Decision Support System has been introduced to play a vital role in diagnosing Cardio disease. Data mining solves and provides the techniques to convert raw data into valuable information for decision making. Using these techniques less time is taken for the prediction of heart disease with precision. Data mining tools can easily solve the issues related to the heart and can replace conventional methods which use much time and provide an accuracy to decide. It is observed that the Fuzzy Intelligent Technique increases the accuracy of heart disease prediction system.

[5]H. Benjamin Fredrick David, S. Antony Belcy (2018) provided a new technique in this paper. Clustering, Classification, Association and Prediction Techniques were used. The authors have used Fuzzy Intelligent Technique to acquire more accuracy in the prediction system. Three classification algorithms namely: Random Forest, Naïve Bayes and Decision trees, were used to determine the prediction and to analyse the probability of patients who suffers from heart disease. The main goal of this project is to spot and discover the ideal and optimal algorithm which will be used to distinguish the patients and normal person. The performance evaluation is made to test the level of the prediction with the help of dataset that are collected from UCI repository. The current experimental procedure has been made to evaluate the performance of algorithms and have found that Random Forest method performs with 81% accuracy than other algorithms for cardiovascular disease prediction.

[6]Another research of heart disease prediction includes this study. This study by Indu Yekkala, Vardhaman (2018) aimed to obtain more accuracy. In this paper, the data has been produced from the medical fields. Collecting such data is relatively hard where the research requires scan reports, scripts and many more on a large scale. The complicated nature and volume of data, poses the need for techniques, that can extract relevant information from the available data in an expeditious and an efficient way.

This has enabled them not only to predict the disease but also to prevent by taking the necessary medical precautions for the same. Cardio disease is one of the root causes for death all over the world. Application of single data mining technique did not yield precise results. Further research about combining two or more algorithms to get a better accuracy, has been proven . The authors have also worked on the stalog dataset, which was taken from the UCI repository and used the data by applying the random forest and rough sets algorithms to find the accuracy to predict the cardio disease.

[7]Ibrahim Umar Sais1, Abdullahi Haruna Adam, Dr. Ahmed (2015) conducted a research on Heart disease prediction. It holds a major percentage for high death rate, with 32% as in Canada(35%) and united states of America. The factors and variables which were identified using association rule mining. These factors contribute to heart disease and have used UCI Cleveland dataset, which provides all medical database along with this, the rule generation method was also used. All the information which was gathered on the basis of healthy person and sick patients were separated for training and test data and later compared with the health records. From the comparison, men seem to have more probability to get coronary heart disease than women. Factors like angina induced and many chest pain types, blood sugar level are seen both in men and women. By testing with the algorithms, the evaluation showed that individuals who have tested 'false' in exercise induced angina, are more likely to be free from coronary heart disease. In this paper, authors have used 'the mining' to determine interesting knowledge.

3. PROPOSED METHOD

In the proposed system, we have used four algorithms, decision trees, support vector machine, random forest algorithms and K- neighbour classifier algorithm. The dataset was separated as train data and test data (0.33). Only K-nearest neighbor algorithm will be considered where the other three algorithm are only to prove the accuracy gained, will not be feasible to predict. Machine learning process initialized from data pre-processing stage which was then continued by correlation matrix with heatmap used for feature selection, classification of modelling and performance evaluation and tried to get a better accuracy. The datasets that we have collected are blood pressure level, chest pain types and biomarkers, where the patient's characteristic were taken into consideration.

3. DATASET DESCRIPTION

The dataset for Heart Disease prediction was taken from Kaggle website, a data repository for data analyst and few data was collected from scan centre. The variables present in dataset are age, gender, chest pain types, cholesterol, blood sugar level, resting blood sugar. Dataset like scan images were not needed in this paper and only numerical data was used in predictions.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	num
67	1	4	160	286	0	2	108	1	2
67	1	4	120	229	0	2	129	1	1
37	1	3	130	250	0	0	187	0	0
41	0	2	130	204	0	2	172	0	0
56	1	2	120	236	0	0	178	0	0
62	0	4	140	268	0	2	160	0	3
57	0	4	120	354	0	0	163	1	0
63	1	4	130	254	0	2	147	0	2
53	1	4	140	203	1	2	155	1	1
57	1	4	140	192	0	0	148	0	0
56	0	2	140	294	0	2	153	0	0
56	1	3	130	256	1	2	142	1	2
44	1	2	120	263	0	0	173	0	0
52	1	3	172	199	1	0	162	0	0
57	1	3	150	168	0	0	174	0	0
48	1	2	110	229	0	0	168	0	1
54	1	4	140	239	0	0	160	0	0
48	0	3	130	275	0	0	139	0	0
49	1	2	130	266	0	0	171	0	0
64	1	1	110	211	0	2	144	1	0
58	0	1	150	283	1	2	162	0	0
58	1	2	120	284	0	2	160	0	1
58	1	3	132	224	0	2	173	0	3
60	1	4	130	206	0	2	132	1	4
50	0	3	120	219	0	0	158	0	0
58	0	3	120	340	0	0	172	0	0
66	0	1	150	226	0	0	114	0	0
43	1	4	150	247	0	0	171	0	0

Fig-1: Heart Disease Dataset

4. METHODOLOGY

The collected datasets were pre-processed, where the records with missing value or unnecessary value were removed from the dataset and it was filled with proper values. After which, the training dataset was sent to the four classifiers to evaluate and find the estimated accuracy. Among the gained accuracies, the suitable algorithm was taken and it was tested with the test data.

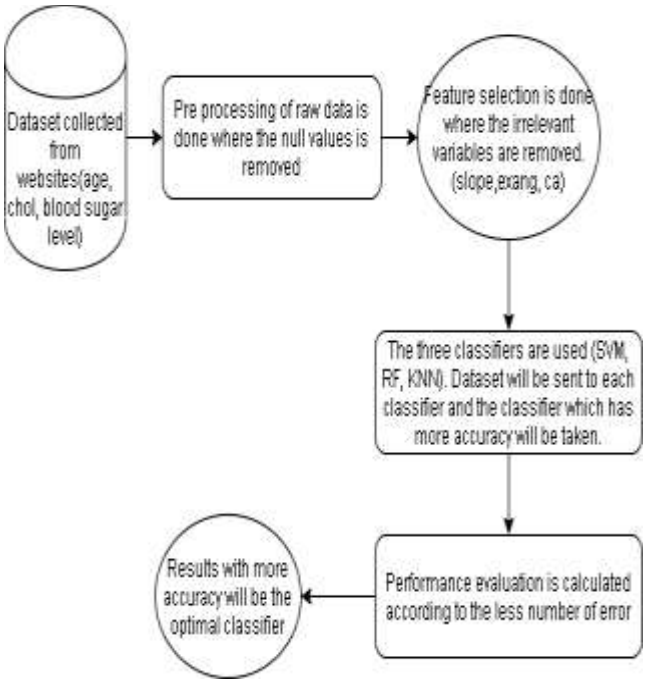

Fig-2: Heart disease system design

5.1 Data pre-processing

The data collected from the patient about the characteristic of their behaviour (age, gender, cholesterol, blood sugar level, angina induced) is pre-processed and classified according to their characteristic. Data pre-processing is a technique where the raw datum is transformed into a proper format where the unnecessary values are removed. [8]Pre-processing is the initial step in conducting the analysis. The datasets are read as .csv file and the variables are processed into numerical and categorical variables for our study. This module provides elimination of unwanted values (null values) and attributes (old peak), slope(fluoroscopy coloured major vessels)) as well as cleaning the dataset. A dataset can be loaded with Not Applicable (NA) values or even empty values. These values need to be removed by removing the whole row which contains such a value.

5.2 Feature Selection

Among the variables of the dataset, two attributes pertaining to age and gender are used to identify the personal information of the patient. The variables (age, gender, blood sugar level, cholesterol, chest pain types, exang) are considered important as they contain vital clinical records where feature importance method have been used to univariate selection.

Feature selection which is based on Correlation matrix has been used where the heat map will be easier to find the most related features to the variable. The data will be imported after which the correlation matrix with .corr. was calculated. The matrix is not numeric because the h-type column is not present in the matrix creating heat map in seaborn.

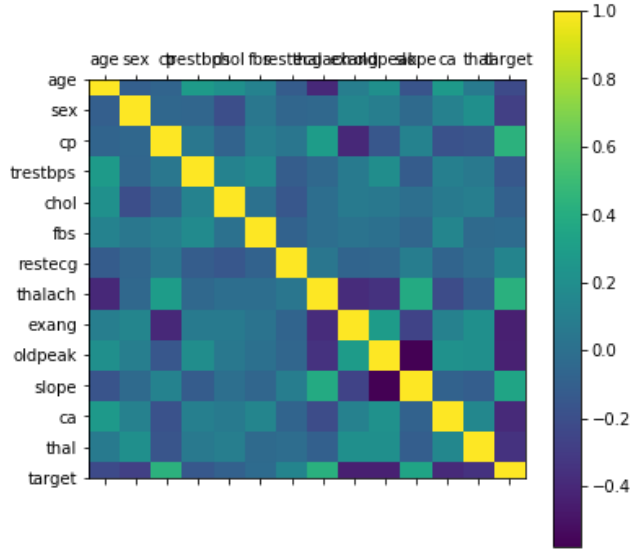

Fig-3: Correlation heat map

5.3 Classification Modelling

The classification models used in this paper include Random forest, support vector machine, decision tree, k nearest neighbour and fuzzy k nearest neighbour. The accuracy of these models are checked individually and the model which gets more accuracy was chosen for constructing heart disease prediction web application. Cross Validation technique is used here to separate the test data and training data.

Some of the python packages needed for this classification models are pandas, numpy, sklearn, matplotlib. Matplotlib is a 2-dimentional python plotting library which produces a better quality plot in every format. From this, the K-nearest neighbor classifier gets more accuracy with 87% and followed by random forest method with 84% accuracy and decision tree method with 79% accuracy. included K-fold in k nearest neighbor and that resulted in hybrid combination of algorithms.

In the existing system, they have used many Machine learning algorithms where the overall optimal accuracy was 87%, but here we have included the hybrid version of fuzzy k-nearest neighbour and the accuracy gained by this method is 94%. So we will only consider this fuzzy k nearest neighbour algorithm to develop heart disease prediction web application for each individual to check by themselves. Fig-4 diagram is shown for comparing the four algorithms where the Fuzzy Knn algorithm has not been included in the bar plot.

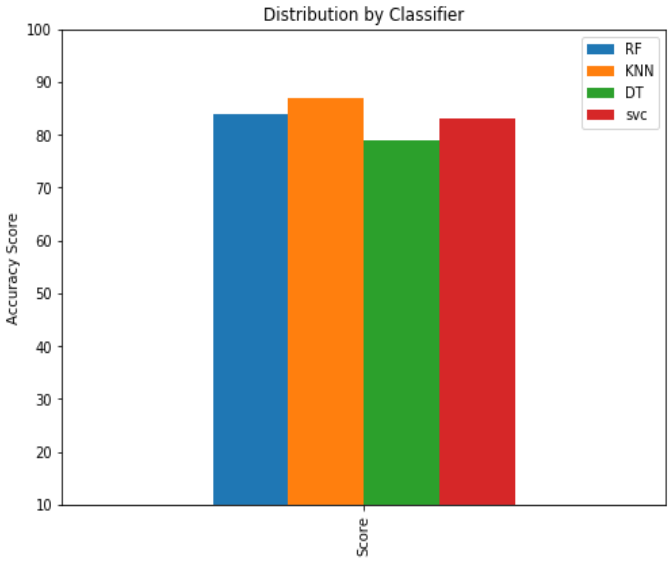

Fig-4: Comparison bar plot

6. RESULTS AND ANALYSIS

The accuracy which was gained by the four prediction model is done by seperating the test data and training dataset. We have given a separate code for the test and train sample to be identified by the system. The problem people faced with visiting scan centers to check if they have heart disease or not are actually cost efficient and time consuming too. In order to reduce them, we have improvised an existing system with different algorithms and hybrid of fuzzy knn and knn algorithms. The accuracy we have gained by these algorithms are much higher than the existing ones.

The framework that have been used in this project application is flask framework in python. This helps in building a proper light weight application. Fuzzy knn classifier predicts in the heart disease prediction application.

7. CONCLUSION

The implementation of the data analysis with different classifiers was done successfully with a web application which can clearly specify whether he/she has the probability of getting heart disease or not. It has been proven that the users can reach out for this prediction and the application can help them predict the disease easily. The integration of fuzzy knn and knn helps the user to understand the each plot through this application. The convolutions of the project can be increased by also adding the scans and images as inputs in further developments and trying to predict the disease by using other algorithms.

REFERENCES

[1] "What is Heart Disease? Age group of people getting affected"Mayoclinic,2018.[Online].Available:https://www.mayoclinic.org/diseasesconditions/heartdisease/symptoms-causes/syc-20353118.

[2] "Who has a higher risk of heart attack- Men or Women" ClevelandClinic.[Online}.Available:https://health.clevelandclinic.org/women-men-higher-risk-heart-attack/.

[3] "Support vector machine kernel trick" Towards data science.[Online]-(2018)

[4] V. Krishnaiah, G. Narasimha, N. Subhash Chandra, "Heart Disease Prediction system using Data Mining Techniques and Intelligent Fuzzy Approach", International Journal of Computer Applications- (2016).

[5]] H. Benjamin Fredrick David, S. Antony Belcy ,"Heart disease prediction by using Data mining Techniques", ICTACT Journal on Soft Computing-(2018).

[6] Indu Yekkala, Vardhaman, "Heart disease prediction using Rough set and Feature Selection", International journal of big data and Analytics in Health care-(2018).

[7] Ibrahim Umar Sais1, Abdullahi Haruna Adam, Dr. Ahmed," Association rule mining on medical data to predict Heart Disease" International journal of Science Technology and Management-(2015).

[8] Data pre-processing technique:[Online].Available

https://towardsdatascience.com/data-pre-processing-techniques-you-should-know-8954662716d6.