Effective Heart Disease Prediction using Distinct Machine Learning Techniques

N. Suganthi[1], R. Abinavi[2], S. Deva Dharshini[3], V. Haritha[4]

[1]*Assistant Professor, Department of Information Technology, Jeppiaar SRR Engineering College, Chennai*
[2,3,4]*Final Year Student, Department of Information Technology, Jeppiaar SRR Engineering College, Chennai*

---***---

Abstract - Heart disease is one among the foremost significant causes of mortality within the world today. Prediction of disorder may be a critical challenge within the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the massive quantity of knowledge produced by the healthcare industry. We have also seen ML techniques getting used in recent developments in several areas of the web of Things (IoT). Various studies give only a glimpse into predicting heart condition with ML techniques. In this paper, we propose a completely unique method that aims at ending significant features by applying machine learning techniques leading to improving the accuracy within the prediction of disorder . The prediction model is introduced with different combinations of features and a number of other known classification techniques. We produce an enhanced performance level through the prediction model for heart condition with the hybrid random forest with a linear model (HRFLM) with an accuracy level of 88.7%.

Key Words: **Machine Learning (ML), Hybrid Random Forest with a Linear Model (HRFLM), K-Nearest Neighbour Algorithm (KNN)**

1. INTRODUCTION

It is difficult to spot heart condition due to several contributory risk factors like diabetes, high vital sign , high cholesterol, abnormal pulse and lots of other factors. Various techniques in data mining and neural networks have been employed to out the severity of heart disease among humans. The severity of the disease is assessed supported various methods like K-Nearest Neighbor Algorithm (KNN), , Genetic algorithm (GA),Decision Trees (DT) and Naïve Bayes (NB). The nature of heart condition is complex and hence, the disease must be handled carefully. Not doing so may affect the guts or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a signifcant role in the prediction of heart disease and data investigation. We have also seen decision trees be used in predicting the accuracy of events related to heart disease [1]. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease.

In this work, numerous readings are administered to supply a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly referred to as hybrid methods. We introduce neural networks using pulse statistic. This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial _utter (AFL), Premature Ventricular Contraction (PVC)), and Second degree block (BII) to and out the exact condition of the patient in relation to heart disease. The dataset with a radial basis function network (RBFN) is employed for classification, where 70% of the info is employed for training and therefore the remaining 30% is employed for classification [4]. We also introduce Computer Aided Decision Support System in the field of medicine and research.

In previous work, the usage of data mining techniques in the healthcare industry has been shown to take less time for the prediction of disease with more accurate results. We propose the diagnosis of heart condition using the GA. This method uses effective association rules inferred with the GA for tournament selection, crossover and the mutation which results in the new proposed fitness function. For experimental verification, we use the well-known Cleveland dataset which is collected from a UCI machine learning repository. We will see later on how our results prove to be prominent when compared to some of the known supervised learning techniques [5]. The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) is introduced and some rules are set up for heart disease. The rules are applied randomly with encoding techniques which end in improvement of the accuracy overall [2]. Heart disease is predicted supported symptoms namely, pulse , sex, age, and lots of others**.**

2. RELATED WORKS

Vijayashree, N.Ch. SrimanNarayanaIyengar [1] ,Heart Disease Prediction System Using Data Mining and Hybrid intelligent technique which is published in the year of 2016, used data mining techniques to predict heart diseases where they have found that neural networks use supervised and unsupervised learning, decision tree algorthm uses ID3 algorithm, Naïve Bayesian classifier uses probability and genetic algorithm, Heart disease is one of the main sources of demise around the world and it is crucial to predict the disease at a premature phase. The computer supported systems help the doctor as a tool for predicting and diagnosing heart disease. The objective is to widespread

about Heart related cardiovascular disease and to explain about existing decision support systems for the prediction and diagnosis of heart disease supported by data mining and hybrid intelligent techniques

Ibrahim Umar Said1 , Abdullahi Haruna Adam2, Dr. Ahmed [2 Avinash]:Association rule mining on medical data to predict heart disease, which is published in the year of 2016 ,this paper describes our experience on discovering association rules in medical data to predict heart disease. Heart disease is the causes of mortality accounting for 32% of all death, a rate is high as in Canada (35%) and USA.Association rule mining a computational intelligence approach is employed to spot the heart condition and Uci Cleveland data set, a biological data base is used along with the Apriori which is called as rule generation algorithm. Analyzing the information available on sick and healthy individuals and taking faith as indicator. Males are seen to have more chance of getting coronary heart disease than females. Without showing or producing no symptoms of chest pain and the presence of exercise- induced angina indicate the likely of existence of heart disease for both men and women. On the other hand, the result showed that when exercise induced angina (chest pain) was true, it was a bad indicator of a person being unhealthy irrespective of gender. This research has demonstrated and performed with the use of rule mining to determine interesting knowledge.

V. Krishnaiah, G. Narsimha [3]: Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach , which is published in the year of 2016.The Healthcare trade usually clinical diagnosis is ended typically by doctor's knowledge and practice. Computer Aided Decision network plays a serious task in medical field. Data mining provides the methodology and approach to change these mounds of knowledge into useful information for determining . By using data processing techniques it takes less time for the prediction of the disease with more accuracy. Among the increasing research on heart condition predicting system, it's happened to significant to categories the research outcomes and provides readers with an overview of the prevailing heart disease prediction techniques in each category. Data mining tools can answer trade questions that conventionally in use much time overriding to make a decision . In this paper we study different papers in which one or more algorithms of data mining used for the prediction of heart disease. As of the study it is observed that Fuzzy Intelligent Techniques increase the accuracy of the heart disease prediction system. The generally used techniques for Heart Disease Prediction and their complexities are summarized in this paper.

Golande, Pavan Kumar T [4] : Heart Disease Prediction Using Effective Machine Learning Techniques which is published in the year of 2019. In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and feminine category and this ratio may vary consistent with the region also this

ratio is taken into account for the people aged group 25-69. This doesn't indicate that the people with other age bracket won't be suffering from heart diseases. This problem may start in early age bracket also and predict the cause and disease may be a major challenge nowadays. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases.

Animesh Hazra, Subrata Kumar Mandal, Arkomita Mukherjee , Amit Gupta, and Asmita Mukherjee [5] - Heart condition Diagnosis and Prediction Using Machine Learning and data processing Techniques A Review , which is published in the year of 2017.A popular saying goes that we are living in an "information age". Terabytes of data are produced every day. Data mining is that the process which turns a set of knowledge into knowledge. The health care industry generates an enormous amount of knowledge daily. However, most of it is not effectively used. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes aren't much prevalent. The aim is to summarize a number of research on predicting heart diseases using data processing techniques, analyse the diverse combinations of mining algorithms used and conclude which technique(s) are effectual and efficient. Also, some future directions on prediction systems are addressed.

Reddy Prasad, Pidaparthi Anjali, S.Adil, N.Deepa [6] : Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning which was proposed in the year of 2019. We are during a period of "Information Age" where the normal industry can pressure the rapid shift to the economic revolution for industrialization, supported economy of data technology Terabytes of knowledge are produced and stored day-to day life due to fast growth in „Information Technology". Terabytes of knowledge are produced and stored day-to day life due to fast growth in „Information Technology". The data which is converted into knowledge is done by data analysis by using various combinations of algorithms. For example: the massive amount of the data regarding the patients is set up by the hospitals like x-ray results , lungs results ,heart paining results, pain results , personal health records(PHRs) ., etc. There is no effective use of the info which is generated from the hospitals. Some certain tools are used to extract the information from the database for the detection of heart diseases and other functions are not accepted. The main theme of the paper is the prediction of heart diseases using machine learning techniques by summarizing the few current researches. In this paper the logistic regression algorithms is employed and therefore the health care data which classifies the patients whether or not they are having heart diseases or not consistent with the information in the record.

3. PROPOSED SYSTEM

We have used python and pandas operations to perform heart condition classification of the Cleveland UCI

repository. It contributes an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection supported data cleaning, classification of modeling performance evaluation, and therefore the results with improved accuracy. n this system we are implementing effective heart attack prediction system using Naïve Bayes algorithm. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is Naïve Bayes. After accessing data set the operation is performed and effective attack level is produced. The proposed system will add some more parameters significant to attack with their weight, age and therefore the refore the priority levels are by consulting expertise doctors and the doctors . The heart disease prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

4. PROPOSED ARCHITECTURE

The data from the UCI laboratory is selected and it is preprocessed (i.e). removal of unwanted information. Then only the features that are required are taken into consideration, and are then classified and prediction is done using the algorithms. After that process is over, the performance is calculated and the results are displayed to the users.
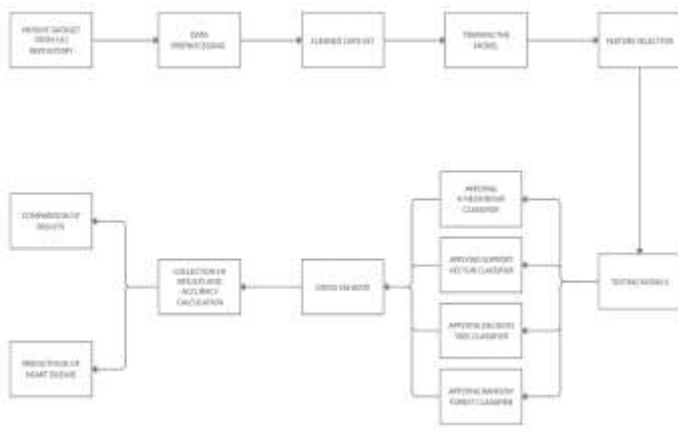

Table 1 Explanation of 13 Input Attributes used for model formation and validation

#	Attributes	Description	Values
1	Age	Patient's age in years	Continuous Value
2	Sex	Sex of Patient	1 = Male 0 = Female
3	Cp	Chest pain	Value 1: typical angina Value 2: atypical angina Value 3: non-angina pain Value 4: asymptomatic
4	Trestbps	Resting blood pressure	Continuous value in mm/Hg
5	Chol	Serum cholesterol in mg/dl	Continuous value in mg/dl
6	Fbs	Fasting blood sugar	$1 \geq 120$ mg/dl $0 \leq 120$ mg/dl
7	Restcg	Resting electrocardiographic results	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value
9	Exang	Exercise induced angina	1: yes 0: no
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	Slope	the slope of the peak exercise ST segment	1: upsloping 2: flat 3: down sloping
12	Ca	number of major vessels colored by fluoroscopy	0-3 value
13	Thal	defect type	3 = normal 6 = fixed defect 7 = reversible defect
14	num	diagnosis of heart disease	no_heart_disease have_heart_disease

5. METHODOLOGY

5.1 K-NEAREST NEIGHBOURS

KNN are often used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. It is also easy to evaluate for calculating time, ease to interpret output, for predictive power. KNN algorithm fairs across all parameters of considerations. It is commonly used for its low calculation time and easy of interpretation.

5.2 RANDOM FOREST METHOD

Random forest is like bootstrapping algorithm with Decision tree (CART) model. If we have 1000 observation in the complete population with 10 variables. Random forest tries to create multiple CART model with different sample and different initial variables. For illustrate, it'll take a random sample of 100 observation and 5 randomly chosen initial variables to create a CART model. It will repeat the process 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction.

5.3 NAIVE BAYES ALGORITHM

Naive Bayes model is easy to erect and particularly useful for very large data sets. Naive Bayes is known to outperform even highly sophisticated classification methods along with simplicity too. Bayes theorem provides a way of calculating posterior probability.

5.4 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for either regression or classification challenges. Yet, it is mostly used in classification problems. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a bound which best segregates the two classes like line and hyperplane.

6. MODULE DESCRIPTION

6.1 DATA PRE-PROCESSING

Heart disease data is pre-processed after collection of distinct records. The dataset contains a complete of 303 patient records. The 6 records have been removed from the dataset where 6 records are with some missing values and the remaining 297 patient records are used in pre-processing.

The data set for this research was taken from UCI data repository.1 Data accessed from the UCI Machine Learning Repository is freely available. In precise, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing. The UCI Machine Learning Repository may be a collection of data generators ,databases and domain theories that are engaged by the machine learning algorithms for the empirical analysis. The overall objective of our work is to predict the presence of heart condition more accurately . It is integer valued from 0 to 4. Experiments with the Cleveland database have concentrated on seeking to distinguish presence absence (value 0) of presence (values 1,2,3,4). Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more exactly.

6.2 Feature Selection and Reduction

From the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered essential as they contain vital clinical records. Clinical records are learning the severity of heart disease and vital to diagnosis**.**

6.3 Classification Modeling

The clustering of datasets is done on the support of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are tested to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error. The overall concept is to build a tree that provides balance of flexibility & accuracy.

- Decision Trees Classifier
- Support Vector Classifier
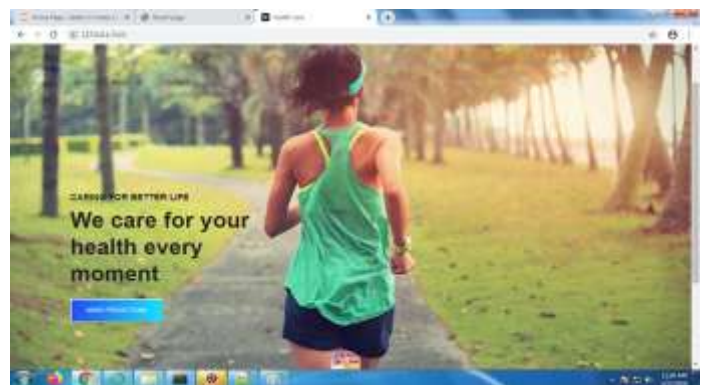- Random Forest Classifier
- K- Nearest Neighbor

7. PERFORMANCE MEASURES

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model.
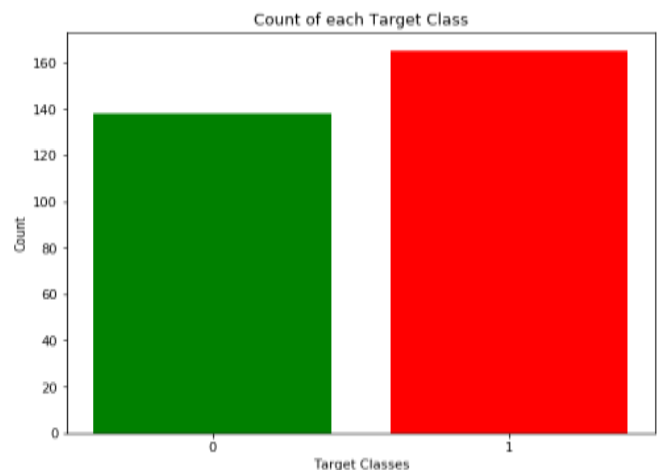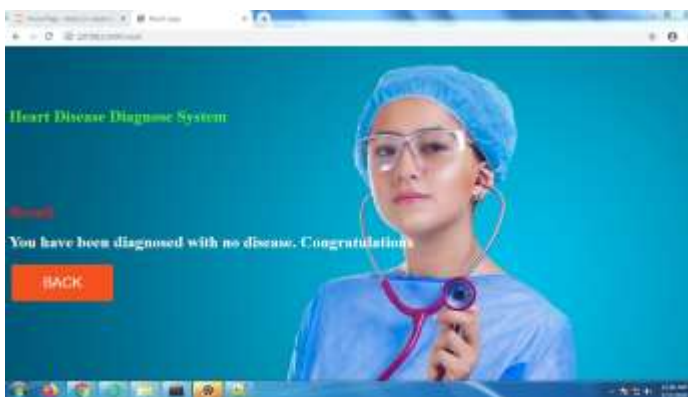

8. RESULTS

Our webpage

This is our webpage asking for the requirements from the user. Once the user enters the data, the system analyses.

When the user is diagnosed with no heart disease.

The webpage when the user has a heart disease.

9. CONCLUSION

Heart disease is one of the main sources of demise all around the world and it is imperative to predict the disease at a budding phase. Machine Learning can play an essential role in predicting presence/absence of Heart diseases and more. Such information, if predicted beforehand, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis. Our project involved analysis of the heart disease patient data-set with proper data processing. Then these four models were trained and tested.

10. FUTURE ENHANCEMENTS

We have used python and pandas operations to perform heart condition classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment along with building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection supported data cleaning, classification of modeling performance evaluation, and therefore the results with improved accuracy. Instead of UCI data set one can choose a different data set which can include any data set from the hospitals all around the world. We also introduce Computer Aided Decision Support System in the field of medicine and research. In previous work, the usage of data mining techniques in the healthcare industry has been shown to take less time for the prediction of disease with more accurate results. We propose the diagnosis of heart condition using the GA. This method uses effective association rules inferred with the GA for tournament selection, crossover and the mutation which results in the new proposed fitness function. In our project we have used the well-known Cleveland dataset which is collected from a UCI machine learning repository.

11. REFERENCES

[1] A. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predicting the coronary heart disease using random forest classi_er,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22_25.

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ``Using PSO algorithm for producing best rules in diagnosis of heart disease,'' in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306_ 311.

[3] N. Al-milli, ``Backpropogation neural network for prediction of heart disease,'' J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131_135, 2013.

[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of neural networks based heart disease prediction system,'' in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233_239.

[5] P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,'' J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27_40, Jan. 2012.

[6] L. Baccour, ``Amended fused TOPSIS-VIKOR for classi_cation (ATOVIC) applied to some UCI data sets,'' Expert Syst. Appl., vol. 99, pp. 115_125, Jun. 2018.

[7] C.-A. Cheng and H.-W. Chiu, ``An arti_cial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database,'' in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566_2569.

[8] H. A. Esfahani and M. Ghazanfari, ``Cardiovascular disease detection using a new ensemble classi_er,'' in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011_1014.

[9] F. Dammak, L. Baccour, and A. M. Alimi, ``The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,'' in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1_8.

[10] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart disease through neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4, pp. 7675_7680, May 2009.

[11] M. Durairaj and V. Revathi, ``Prediction of heart disease using back propagationMLPalgorithm,'' Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235_239, 2015.

[12] M. Gandhi and S. N. Singh, ``Predictions in heart disease using techniques of data mining,'' in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE), Feb. 2015, pp. 520_525.

[13] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ``Prediction of heart disease using machine learning,'' in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275_1278.

[14] B. S. S. Rathnayakc and G. U. Ganegoda, ``Heart diseases prediction with data mining and neural network techniques,'' in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp. 1_6.

[15] N. K. S. Banu and S. Swamy, ``Prediction of heart disease at early stage using data mining and big data analytics: A survey,'' in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256_261. 81