Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary | Peer Reviewed | Open Access
Volume: SI, Number: 01, Pages: 178- 182, Year: 2020

EARLY Heart Disease Prediction with Minimal Attributes using Machine Learning

Sameen Aziz, Zahid Aslam, Muhammad Rizwan, and Shoaib Nawaz

Khawaja Fareed university of Engineering and Information Technology, Rahim Yar Khan, Pakistan
Corresponding author: Sameen Aziz (e-mail: sameenaziz.sami@gmail.com)

A*BSTRACT-* Heart Disease is one of major problem in our medical science and also in society. Largest-ever study on death shows that Heart Diseases have become Number one killer disease in the world. About 25 percent of deaths in the age group of 25 to 69 years occur due to heart diseases. In Our medical society all doctors are not equally skilled and experienced so every doctor has different opinion about Patients but still cases are reported of wrong diagnosis and treatment. There is need an intelligent system that predict about heart disease more accurately than doctors. In this paper we developed an intelligent Model using Neural Network that predict early Heart Disease base on minimum but major attributes of patient's health with accuracy 90%. Finally, we got Impressive Results with small number of input variables of patients and the comparative study of the different Machine learning algorithms and tools.

Keywords: Heart Diseases; Machine Learning; Neural Network.

I. INTRODUCTION

Human Heart is the most important and complex working muscle in human body. It supplies blood to every cell of the body in addition heart muscle is the engine of human body. Heart disease is a major health problem and became the cause of death worldwide. It can be a cause of serious cardiovascular actions just like stroke and heart attack. It has been observed in that, the risk of heart failure occurrence in individual at the age of 40 years is 1 out of 5 [1]. Heart Diseases which also known Cardiac Diseases causes roughly 735,000 Heart Attack each year in the U.S. and kills more than 630,000 Americans each year. According to the American Heart Association, over 7 million Americans have suffered a heart attack in their lifetime.

Because heart diseases are so common and often is silent until it strikes, it is important to recognize the factors that put you at risk. So, here is need of an Intelligent Model which can earlier predict heart diseases in any patient so that we can reduce these heart disease ratios and also proper diagnose so in this way we also reduce cost of diagnose and provide quality diagnoses system. A big problem facing in healthcare organizations is to provide quality services at affordable costs [1].

Quality services mean that diagnosing patients correctly. Poor clinical decisions which are often made based on doctor's experiences and all the doctors have not same knowledge and experience so it led to a wrong treatment and the result is useless treatment and also the loss of heavy budget consuming [2].
Mostly hospitals using software system to store patient's data but unfortunately this data are rarely used to help us to make clinical decision. There is lot of hidden information in these patient data, if we use this data and extract information and hidden patterns of the data then we can predict and make clinical decision easily and save valuable time and cost. Here needs an intelligent model which extract hidden patterns from data and make prediction base on that data, hence Machine Learning come into. Machine Learning is sub field of AI in which we use collection of data feed the model to train it then new data gave to it and model classify it according previous pattern. In this paper we use Machine Learning algorithm SVM (Support Vector Machine) to develop an intelligent model to make intelligent clinical decisions that other systems not. It is almost second time in Pakistan (First the students of NUST)[3]that we introducing an Intelligent Model base on Machine Learning which use minimum number of attributes and also minimum number of observations to train the model and this model make prediction with high accuracy and our work introducing an web base application and also an android application so everyone can easily check his heart status [4].

This paper is divided as Section II describe the Related Work III is dataset, IV is about Methodology, V have Experiments & Analysis and finally have VI is conclusion.

II. RELATED WORK

It is in experience faced by peoples [5] that the doctors are diagnosis the wrongly and obviously wrongly treatment will be faced by patient this kind of problem makes the danger of life and loss of highly expanses. They experiment with dataset having 13 attributes and make them to 6 and using the machine learning techniques to predict that the patient is that having the heart disease or not to help doctors for diagnosis and treatment. The model's results are NB 96.5% and DT 99.2% and 88.3% Classification with Clustering[6]. We can also classify the patients with or without disease by using the data mining techniques like bootstrap aggregation and make it boosting and the random forest and the SVM helps us to classify the patients with or without disease in this study the focus is on the heart failure with preserved ejection fraction is the key feature in the study. The collection of patients from different country produced the results are the two groups one with having hypertension have more risk for heart failure and also they have average age more than others the results are for sample-1 31.6% and results from sample-2 38.5% [7]. The Weka tool is the Tool used for data science user the tool help to deploy the text mining and the Bhatla & Jyoit using the machine learning algorithms based on the 15 attributes and apply the Naive Bayes , Decision Trees and Neural Networks and the results are NB is 90.74% and DT is 99.62% and ANN 100% [8]. In India the age 25-69 the heart disease is 25% and the minimum is 12% the study shows that the patients having the heart failure survivability is predicted by using CART with 83.49% accuracy and ID3 is 72.93% and the DT 82.50% the results shows that the heart failure perdition by CART is more efficient than others in the south India user age 25-69 [9].

There is also an application CDSS for the doctors help for check the HF severity this tool having the background

processing with machine learning algorithms which are NN and SVM and RF and the system will us to predict the heart related subtypes disease and with the regular use of the application help to collect the database and based on the data the system performance shows us CART with cross-validation accuracy is 81.8% in severity and the prediction of heart sub-type with accuracy is87.6%[10]. Features extraction es also help us to make the models more perfect and increase the accuracy there are many method for feature extractions the study in Sudia Arabia use the ECG record and some features extracted by using AR and then make the classification using the 5 algorithms , DT,KNN,SVM,ANN,RF they shows that the RF having the 100% accuracy with their proposed model for the prediction of CHF [11], [12]. The Naive Bayes helps us to predict only that the heart disease patients or not for this purpose the experiments shows that based on the 14 attributes model and sows the results to 89.58% with 303 training records and 240 testing records and the number of correctly classified is 215 and incorrect are only 25 [13].

But its only works of disease detection not the HF [14]. Another simple but different approach is also used in study which is CANFIS model and combined with the neural network this approach helps us to classify the HD and subtypes and the mean square error is only 0.000842 which is so productive results [3]. The same approach of applying NB and they proposed the risk prediction model in which they get collection the data from the armed forces cardiology department in the form of reports in which some are structured and unstructured the model achieved the accuracy 86.7% [15]. There is another DS system regarding the heart disease prediction web application based on the machine learning algorithm which is NB and get the patient data through a patient form or patient questioner and get the input from the user and based on the training data system helps the nurses and doctors about the diagnosis of the patient HD [16]. The system with name DSHDPS which is developed as web application [4].

III. METHODOLOGY

Here we collected the dataset and then after exploring the dataset we reduced the attributes before preprocessing we have 13 attributes and after exploring and preprocessing, we find the only four attributes and remove the other 9 attributes and deploy the algorithms on three different Mmachine learning and data mining tools. We found that the anaconda python is best then others and second one is the Neural Network perform best results with 90% correct predict The Methodology is shows in Fig. 1.

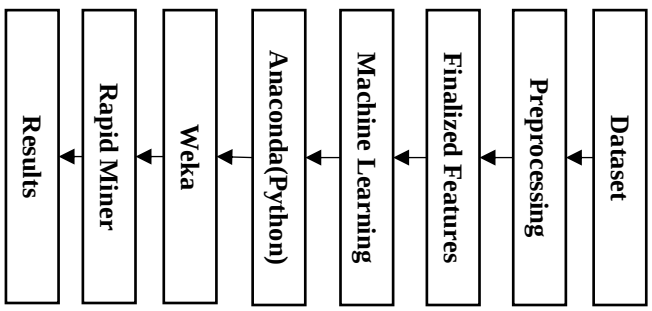

FIGURE 1: Proposed method methodology.

A). DESCRIPTION OF DATA SET

In this paper we want to detect either heart disease exists or not in patients, for this purpose as we know Supervised Machine Learning Models require training dataset because firstly, we train our model then our model is able to make prediction.

The Data we use to train our model is open source and publicly available [11] which have 270 observations and 4 attributes, details of these attributes are given in Tab.1.

TABLE 1: Description of Dataset

cp	thalach	exang	oldpeak	Pred_class
chest pain type: 1=typical angina; 2=atypical angine; 3=non-anginal pain; 4=asymptomatic	3=normal; 6=fixed defect; 7=reversable defect	exercise induced angina: 1=yes; 0=no	ST depression induced by exercise relative to rest	Prediction Class

This dataset contains almost those attributes which involves in heart diseases, we also discuss with the doctors of Sheikh Zayed Hospital Rahim Yar Khan about these attributes and they agreed that on these attributes we can make prediction with high accuracy.

Data set view with reduced attributes are given in Tab. 2.

TABLE 2: Dataset attributes

	cp	thalach	exang	oldpeak	Pred_class
0	3	150	0	2.3	1
1	2	187	0	3.5	1
2	1	172	0	1.4	1
3	1	178	0	0.8	1
4	0	163	1	0.6	1

Here is the description of these attributes the strength of the classification class or the target class in Fig. 2.

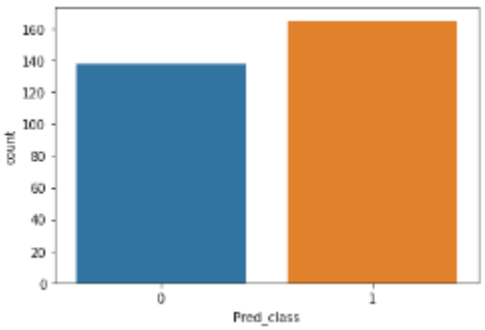

FIGURE 2: Strength of classification class.

B) EXPERIMENTS AND RESULTS:

Now we exploring the features and their values actually we first find the relationship or correlation between all variables and then based on the correlation values we exclude the low correlated attributed and then we selected the attributes which are highly correlated.

We selected the five attributes as shown in the figure 2.0 and these attributes gives us the accuracy at level same at eight attributes after the checking the correlation the results after and before excluding the attributes the actual attributes and after excluding attributes the dataset looks like this. One side is attributes name and other column is correlation values.

In the table the first column shows the names of the attributes and other one is correlation of the attributes so here we show that the before and after excluding the attributes we show here

the finalized attributes from 13 attributes to we stop at 4 attributes and predict the patient having weak heart or strong heart given in Tab. 3.

TABLE 3: Attributes of patient with weak/strong heart.

After excluding Attributed		Before Excluding Attributes	
Pred_class	**1**	**Pred_class**	**1**
exang	**0.436757**	**exang**	**0.436757**
cp	**0.433798**	**cp**	**0.433798**
oldpeak	**0.430696**	**oldpeak**	**0.430696**
thalach	**0.421741**	**thalach**	**0.421741**
Excluded attributes -→		**ca**	**0.391724**
		slope	**0.345877**
		thal	**0.344029**
		sex	**0.280937**
		age	**0.225439**
		trestbps	**0.144931**
		restecg	**0.13723**
		chol	**0.085239**
		fbs	**0.028046**

Now the feature exploration and the first one cp which is actual name is chest pain shown in Fig. 3.

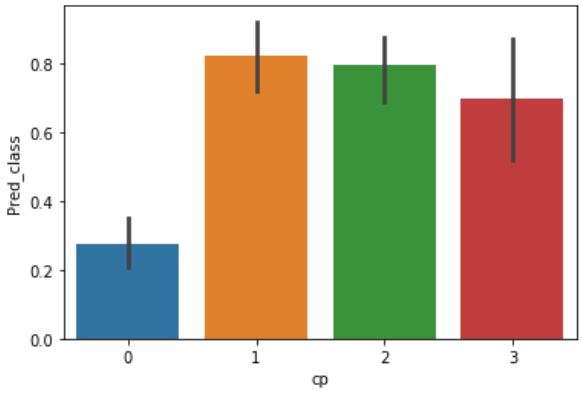

FIGURE 3: Feature exploration.

The second one is maximum heart rate achieved and which is *thalach* in the dataset shown in Fig. 4.

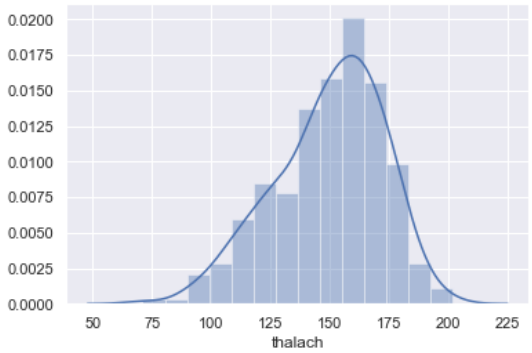

FIGURE 4: Maximum heart rate histogram.

Exercise including angina: 1 = Yes and 0 = No, as is seen in Fig. 5.

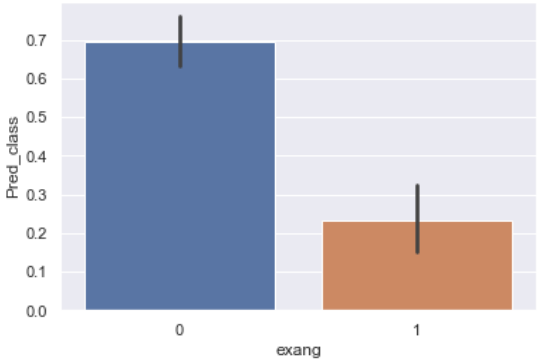

FIGURE 5: Strength of classification class with angina.

Another factor is depression which is most important and highly correlated within dataset as shown in Fig. 6.

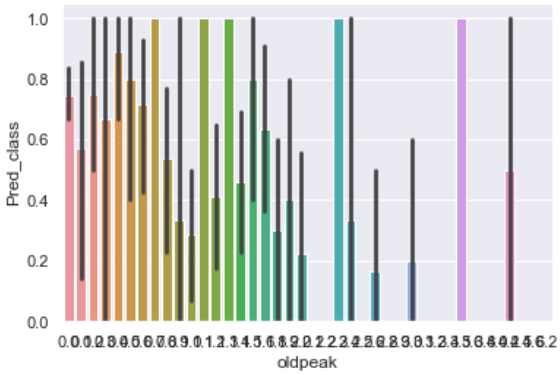

FIGURE 6: Strength of depression.

Now we ready data for machine learning we split the date into 80% for using in training and 20% for test our model accuracy as is given in Tab. 4.

TABLE 4: Machine learning data for accuracy.

Model	Accuracy	Classification Error	F1 Score
Naive Bayes	0.79	0.20	0.76
Generalized Linear Model	0.79	0.20	0.77
Logistic Regression	0.79	0.20	0.77
Fast Large Margin	0.80	0.19	0.76
Deep Learning	0.79	0.20	0.77
Decision Tree	0.80	0.20	0.79
Random Forest	0.77	0.23	0.76
Gradient Boosted Trees	0.76	0.24	0.73
Support Vector Machine	0.78	0.21	0.76

We use the nine machine learning algorithms and from these are deployed in rapid miner the results show in RapidMiner Tool on the same dataset with same values with same variables. Here we are showing the results received from the one Tool which is RapidMiner. Visualization of the results and comparison with error in classification as shown in the Tab. 4.

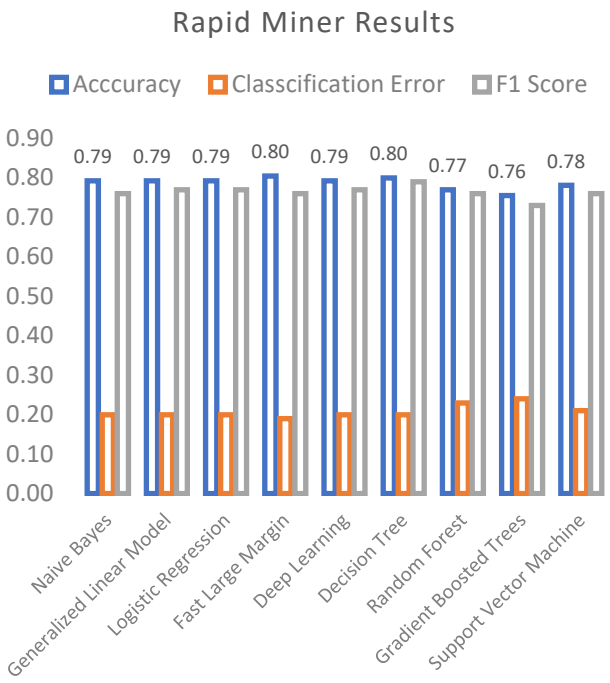

FIGURE 7: Decision tree.

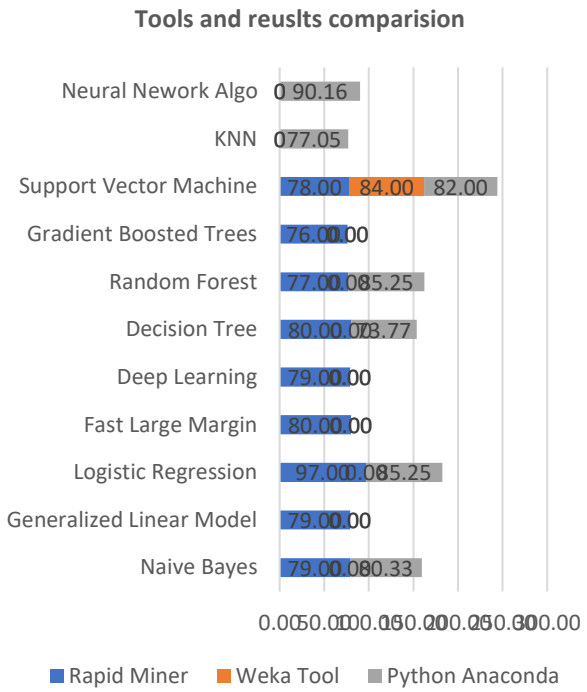

FIGURE 8: Strength of depression.

The best one is the Decision Tree at 80% Accuracy as shown in Fig. 7. Now we decided that deploy the algorithms on the same dataset in python the Neural Network algorithm gives us the best accuracy in Jupiter note book.

IV. RESULTS AND DISCUSSION:

We deploy the idea on three tools on the same dataset same attributes and same machine learning algorithms we found that the Weka Tool and python gives us almost same results and the Rapid miner gives us different results the comparison are in the table. We here received that results from the all tools and all algorithms. The python is best tool for data mining the second one is in our proposed problem we received results from the Neural networks which is 90% and the lowest results of the Decision Tree which is 73% shown in Fig. 8. We contribute our work to find the best results based on the minimum attributes and tools comparison and algorithms comparison on the same dataset with same settings and same values given in Tab. 5. For the results see the graph in the graph we show the all results and shows the tools comparison.

TABLE 4: Best dataset for accuracy.

Model	Rapid Miner	Weka Tool	Python Anaconda
Naive Bayes	79.00	nan	80.33
Generalized Linear Model	79.00	nan	nan
Logistic Regression	97.00	nan	85.25
Fast Large Margin	80.00	nan	nan
Deep Learning	79.00	nan	nan
Decision Tree	80.00	nan	73.77
Random Forest	77.00	nan	85.25
Gradient Boosted Trees	76.00	nan	nan
Support Vector Machine	78.00	84.00	82.00
KNN	nan	nan	77.05
Neural Nework Algo	nan	nan	90.16

Table 3.0

V. CONCLUSION

From our studies, we have managed to achieve our research objectives. The objective of our work is to predict more accurately the presence of heart disease with minimum number of attributes and less data for training. Algorithms are used to predict the Heart Diseases in patients with the same attributes and we also observed that Neural Network produce very high accuracy rather than others. The fact is that Intelligent Model is always work according to training it learns from training dataset, Intelligent model not analyze physically so there is chance of error but according to our research object to

introduce such an intelligent model which can make prediction about patient heart disease. So, here is a model with the accuracy of 90 percent it can use in hospitals to predict heart diseases. Instead of doctors which everyone has different skills, knowledge and experience, this model can make more accurate prediction than doctors and Hence doctors do proper treatment and saved lot of cost and useless treatments and this is purpose of our research.

REFERENCES

[1] M. H. Asmare, F. Woldehanna, L. Janssens, and B. Vanrumste, "Automated Rheumatic Heart Disease Detection from Phonocardiogram in Cardiology Ward," in *the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)*, 2020, vol. 5, pp. 839–844.

[2] R. Gómez-Gutiérrez *et al.*, "Early detection of and intervention for two newborns with critical congenital heart disease using a specialized device as part of a screening system," *SAGE Open Medical Case Reports*, vol. 8, p. 2050313X20926041, 2020.

[3] "Saqlain et al. - 2016 - Identification of heart failure by using unstructu.pdf." .

[4] P. Kora, A. Abraham, and K. Meenakshi, "Heart disease detection using hybrid of bacterial foraging and particle swarm optimization," *Evolving Systems*, vol. 11, no. 1, pp. 15–28, 2020.

[5] "Anbarasi et al. - 2010 - Enhanced prediction of heart disease with feature .pdf." .

[6] "Austin et al. - 2013 - Using methods from the data-mining and machine-lea.pdf." .

[7] "Bhatla and Jyoti - 2012 - An analysis of heart disease prediction using diff.pdf.".

[8] "Chaurasia and Pal - 2013 - Early prediction of heart diseases using data mini.pdf.".

[9] "Guidi et al. - 2014 - A machine learning system to improve heart failure.pdf.".

[10] "Masetic and Subasi - 2016 - Congestive heart failure detection using random fo.pdf.".

[11] L. Xiaoting, F. Zibo, T. Rong, M. Xiaona, and T. Yifeng, "Direct electrochemiluminescent immunosensing for an early indication of coronary heart disease using dual biomarkers," *Analytica Chimica Acta*, 2020.

[12] "Medhekar et al. - 2013 - Heart disease prediction system using naive Bayes.pdf.".

[13] N. E. I. Slitine *et al.*, "Pulse oximetry and congenital heart disease screening: Results of the first pilot study in Morocco," *International Journal of Neonatal Screening*, vol. 6, no. 3, p. 53, 2020.

[14] "Parthiban and Subramanian - 2008 - Intelligent heart disease prediction system using .pdf.".

[15] "Subbalakshmi et al. - 2011 - Decision support in heart disease prediction syste.pdf.".

[16] C. Yohannes, I. Nurtanio, and K. C. Halim, "Potential of Heart Disease Detection Based on Iridology," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 875, no. 1, p. 012034.

[17] https://www.dataschool.io/

[18] https://www.mldata.io/