iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** III **Month of publication:** March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33518

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

Heart Disease Prediction in Medical Industries using Data Mining Techniques

Divya Pokkunuri[1], Tanishka Dodiya[2], Vishwa Babariya[3]
[1, 2, 3]*Computer Engineering Department, NMIMS University*

Abstract: ***Data mining plays a very important role in Health- care industry. Be it in predicting diseases based on symptoms or predicting the stage / level of severity of any disease, data mining has proved to be very helpful. Healthcare industries collect huge amounts of data and thus, the use of machine learning saves time and guarantees performance. In this paper we analysed the various data mining techniques which can be used in the healthcare industry for heart disease prediction and proposed a system using artificial neural networks for the same. The Proposed System uses 8 medical attributes such as sex, thallium test results, chest pain type, exang, age, etc., the accuracy and key influencers for the proposed system have been discussed too. The most preferred supervised learning techniques are decision trees, naïve Bayes and random forest and the analysis for the same has been done.***

Keywords: Data mining, HealthCare and predictive analysis, Disease analysis, Artificial Neural Network, Heart Disease, Python programming, Machine Learning.

I. INTRODUCTION

Data mining plays a very important role in Health-care industry. Be it predicting diseases based on symptoms or predicting the stage/level of severity of anyone type of disease, data mining has proved to be very helpful. Detection of any disease by making a patient go through several tests can be time taking and does not always guarantee positive results. Many times, we see the diseases are detected very late (Sometimes it is hard to detect unless they reach final stages like cancer) or sometimes not detected. Therefore, there is a need for machine learning which saves time and guarantees performance. Emergency clinics, centres, and other medical services associations all around the globe are working with software companies to create administrative systems that are growingly digitized and mechanized. Even more critically, researchers and scientists are utilizing machine learning (ML) to produce various savvy arrangements that can at last help in diagnosing and treating a disease. Patients are set to profit the most as the innovation can improve their result by breaking down the best types of treatment for them. ML can do more precisely recognizing an illness at a prior stage, assisting with diminishing the quantity of re-admissions in medical clinics and centres. In this paper we analysed the various data mining techniques which can be used in the healthcare industry. We focused mainly on how attributes that are used for predicting cardiovascular diseases are related to each other, the key influencers and how attributes can be subtracted or added to achieve greater accuracy than previously achieved by other researchers.

II. LITERATURE REVIEW

The very famous database used by researchers for their analysis and experiments for predicting cardiovascular diseases is the Cleveland database which contains 13 attributes. The techniques used by other researchers were naïve bayes, CART classifier, Decision trees, ID3 etc. The highest accuracy obtained by others is 79% using naïve Bayes algorithm, 72.93% using ID3 and 83.49% using CART classifier, respectively. Recent developments in this domain include the addition of two more attributes, smoking and obesity. The two mentioned attributes were introduced to increase the accuracy, and the model built on the same system was the neural networks, respectively. We have focused on finding the top attributes influencing the heart disease prediction and the attributes which are least correlated to the target. This analysis helped to eliminate 5 attributes which have not shown much correlation with the target. These insights were gained using visualizations in python and report made using reporting tool PowerBI. Further, we discussed in brief about the heart disease and the attributes' description of the database used.

A. Heart Disease

Heart disease is one of the leading causes of anguish and loss of life and due to this very reason predicting cardiovascular diseases is crucial in the section of clinical data analysis for healthcare sector.

In our study we implemented Decision Tree, Naive Bayes, and Random Forest. Factors causing heart disease are:

1) Smoking
2) High Cholesterol in blood
3) Blood Vessel inflammation
4) High Blood pressure
5) High amount of some type of fats in body

B. Attribute Description of Database Used

1) *Age:* Age of the person in years
2) *Sex:* This is a binary attribute suggesting the gender of the person. State 1 indicates the person is male, and 0 indicates person is female.
3) *CP (chest pain):* nothing but discomfort provoked by various reasons. It takes 4 values: asymptomatic, typical angina, non-angina, and atypical angina, respectively.
4) *THALACH:* This is the maximum heart rate recorded in thallium stress test. This attribute is very intuitive as it tells how heart rate of a person suffering from heart disease changes with respect to a person who is heart disease free.
5) *EXANG:* Specifies Exercise induced angina. This is the discomfort encountered after exercise or emotional stress. Angina is generally caused when heart does not get adequate blood through the arteries to pump. This can be common after exercise and can be relieved with rest. This is a binary attribute as well where state 1 indicates the person encounters angina during or after exercise; 0 indicates absence of any such discomfort during exercises.
6) *THAL:* This is the result of Thallium stress test. Thallium stress test's objective is to know how well blood flows into our heart. Based on its value, it can be classified as normal (value=3), fixed defect (value=6), reversible defect (value=7), etc.
7) *Old Speak:* This attribute has a numeric value based on the ST depression found on an electrocardiogram which is generated by exercise corresponding to rest.CA: This attribute is the number of major vessels that are coloured by fluoroscopy. The values 0 to 3 indicate the vessels.

C. Attributes removed from the Cleveland Database

1) *TRESTBPS:* resting blood pressure. It is the force with which the blood is hitting our artery walls. If resting blood pressure is high, several heart related issues can arise.
2) *CHOL:* serum cholesterol in mg/dl
3) *FBS:* Fasting blood sugar. It is a binary attribute as well. State 1 indicates the fasting blood sugar value is less than 120 mg/dl or greater than 120 mg/dl. If your fasting blood sugar is less than 100 mg/dl and greater than 70 mg/dl, then it can be said normal.
4) *RESTECG:* resting electrocardiographic results (normal, ST-T wave abnormality, or left ventricular hypertrophy)
5) *Slope:* The ST/heart rate slope is the relative shift in ST segments to exercise induced increments in the heart rate. Slope consists of three values where 1 indicates upsloping, 2 indicates flat, and 3 indicates down sloping.

III.DATA MINING FOR ANALYSIS

Data Mining is the technique of catching huge arrangements of information to distinguish the experiences and vision of that information. These days, the interest of data industry is quickly developing which has additionally expanded the requests for Data experts and Data researchers. Since with this procedure, we analyse the information and afterward convert that information into significant data. This causes the business to take exact and better choices in an association. Data mining assists with creating brilliant market choice, run precise missions, forecasts are taken and some more. There are different methods for data mining.

A. Naïve Bayes

Naïve Bayes is one of the simplest and most effective methods for classification that utilizes Bayes rule with a strong assumption that attributes are conditionally independent. This method is in the context of Bayesian Networks which is a probabilistic graphical model representing a set of random variables and their conditional probabilities. Conditional probability is the likelihood of an event A occurring based on the occurrence of a previous event B, where a dependence relation exists between both events A and B. We can represent this probability as: $P(A|B) = (P(B|A). P(A)) / P(B)$

B. Decision Tree

Decision Tree Algorithm is based on conditional probability. Decision tree constructs classification or regression models as a tree structure. A decision tree generates rules unlike Naïve Bayes. A rule is a conditional statement that can be easily understood by humans and can be used easily within a database to identify a set of records. A method that combines a lot of decision tree methods.

1) Step one is to pick random K data points from the training set. Then build a decision tree associated with those data points. The catch here is that instead of building a decision tree based on everything (all data points), you build a decision tree based on some (subset) of the data points that you have.
2) Next choose the number of decision trees you want to build.
3) Repeat steps 1 and 2 and then once you have all those trees and you have a new data point, you make each one of your entry trees predict the category to which the new data point belongs to.

And then assign the new data point to the category that wins the majority vote. In this case, the class labels or categories are nothing but the diseases that we are predicting, i.e., heart disease, breast cancer, and diabetes. If the prediction is based on a certain set of symptoms, then the class labels can be said "Yes" or "No" specifying whether the set of symptoms result to Heart disease or not.

C. Random Forest

The Healthcare industry focuses on performance as well as precision and in such cases, decision trees and random forest prove to be of great help. You go for decision trees when you want to focus more on data interpretation or performance while random forests come with precision too. Random forests follow Ensemble learning. You take multiple machines learning algorithms and put them together to create one bigger machine learning algorithm. Now this machine learning algorithm, the final one, is leveraging many different other machine learning algorithms making it highly efficient. The Random Forest algorithm is one of the most popular and powerful supervised machines learning algorithm which can perform both classification and regression tasks. This algorithm makes the forest with a various decision tree. The more trees in a forest the more vigorous the forecast, and in this manner higher the precision. In random forest we develop various trees instead of a single tree. To classify a new object based on the attribute, each individual tree gives a class prediction. The forest chooses the classification that has the most votes of all the trees in the forest, and that becomes the model's prediction. In the case of regression, it takes the average of the output of the different trees.

TABLE I
Performance for Classifier

Evaluation Criteria	Random Forest	Decision Tree	Naïve Bayes
Correctly classified instances	259	239	247
Incorrectly classified instances	44	64	56
Accuracy (in %)	85.52%	78.94%	81.57%

D. Confusion Matrix

A confusion matrix is a correlation between the actual class labels of the data points and the predictions of a model. Below is the confusion matrix for various classifiers on our test set (76 rows out of 303 rows). No Heart Disease: 0, Possible heart disease: 1.

TABLE III
Confusion Matrix

Classifier	No Heart Disease 0	Possible Heart Disease 1	Class	No. Of right predictions
Random forest	27	6	0	65/76
	5	38	1	
Decision tree	25	8	0	60/76
	8	35	1	
Naïve Bayes	23	10	0	62/76
	4	39	1	

Based on the above table and confusion matrix rules, we can define some very important ratios which are TNR (True Negative Rate), TPR (True Positive Rate), FPR (False Negative Rate), and FNR (False Positive Rate) respectively. For a smart model, TPR and TNR should be high because True positives and True negatives are the correct predictions made by the model. While FPR and FNR are the false/wrong predictions and hence should be less.

Now one may say that it is impossible to take care of all the ratios equally as no model can be perfect, which is true and hence, evaluation also depends on the domain. There can be certain domains which demand to keep one ratio as the main priority, while other ratios being poor. In healthcare domain, where we are predicting some dangerous diseases, where we cannot afford to miss any positive patients, TPR should be as high as possible. Even if we predict any healthy patient as diagnosed, it is still okay as he/she can go for further check-ups.

Random forests classifier, in our case has the highest accuracy. Hence, let us continue this analysis considering only this classifier. In random forests classifier, as seen, we have 5 false negatives and the false negative rate for the same is 15.62% which can be dangerous as previously discussed. This can be dangerous because heart disease patients are being predicted as non-heart disease patients. This can be more dangerous for serious diseases like cancer. Hence, though a model's accuracy is high, it can be proved dangerous in medical industries. So other evaluation parameters should also be used to check the validity of a model. For example, the table below explains how false positives can be reduced by changing the threshold.

TABLE IIIII
True and False Rates for Classifier

Classifier	TPR	TNR	FPR	FNR
Random Forest	86.36%	84.37%	13.63%	15.62%
Decision tree	81.3%	75.75%	18.60%	24.24%
Naïve Bayes	85.1%	79.59%	14.81%	20.40%

The numbers in bold signify false negatives which ideally should be zero, especially in this domain. Hence, our main goal is to improve the false negatives. By default, threshold is 0.5. As seen from the table, false positives have increased from 6 (when default threshold 0.5 was used) to 10 (threshold=0.3) which also decreases the accuracy from 85.52 to 84.21% respectively. Though accuracy has decreased, we can say this classifier is safer than the original one as false negative has decreased. In healthcare sector, only accuracy cannot be taken as a parameter to predict diseases. Hence after confusion matrix analysis, we plot ROC curve to evaluate the performance of our model.

E. ROC Curve

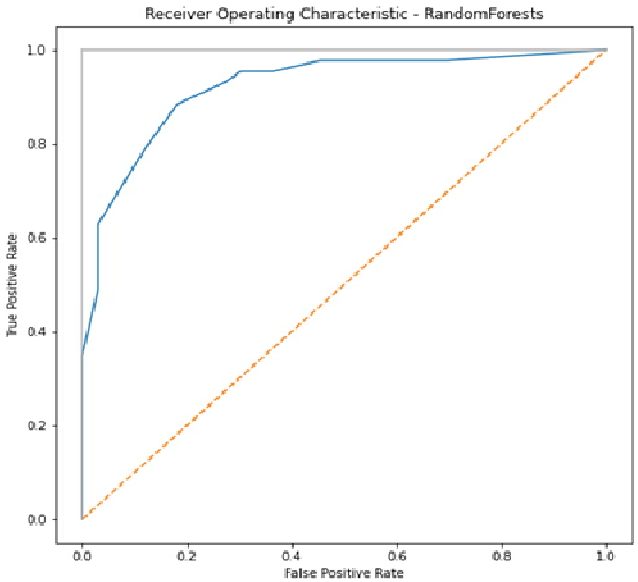


Fig. 1. ROC Curve - Random Forest

The roc_auc_score for the above plot is 0.920 which vali- dates the performance of our classifier. Hence, threshold can be reduced to 0.3 for a good and safe model which has less FNR. Though, as seen from the plot, the current model's performance is satisfiable too.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

IV. FINDINGS AND OBSERVATIONS

A. *Correlation Matrix*

Correlation coefficient in statistics is used to find how strong two variables are related to each other. We use the same coefficient here to see how closely our features are related. When we apply correlation coefficient formula to the data, the result is between -1 to 1.

1) 1 is perfect positive correlation (strong relationship between the variables). This means when there is an increase in one variable, there is a constant or proportional increase in another variable.

2) -1 is perfect negative correlation (strong negative relationship between the variables). This means when there is an increase in one variable, there is a constant or proportional decrease in another variable.

3) 0 means no correlation. This means when there is an increase in one variable, the other variable does not change. The two features just are not related.

With the help of correlation matrix analysis, we can conclude that:

There is a good negative correlation between-

a) Target and exang (Exercise induced angina) (correlation coefficient is -0.44) which means these two features are inversely related to each other.

b) Slope and Oldpeak (correlation coefficient are -0.58). Hence, we can say these two variables are highly correlated.

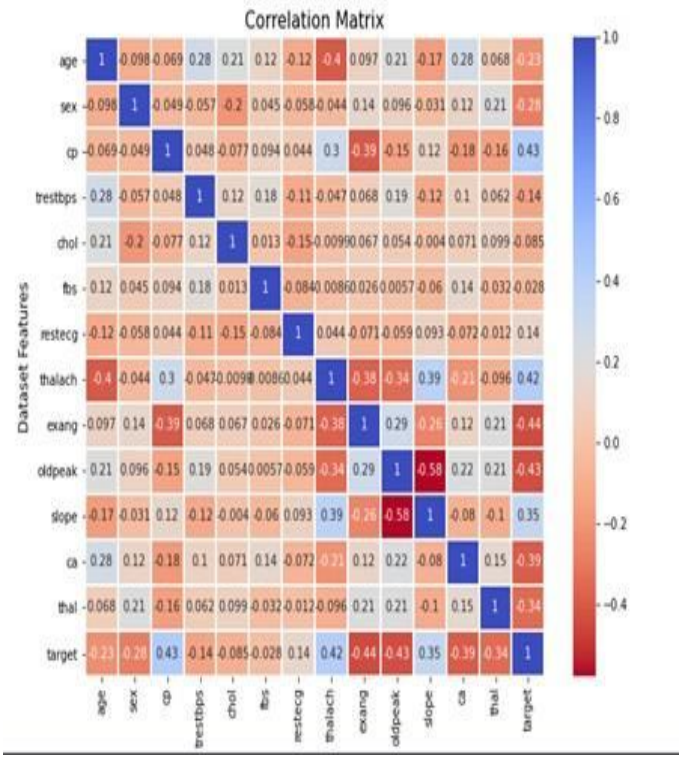


Fig. 2. Matrix displaying the relation between the attributes.

There is almost zero correlation or no correlation for following features w.r.t target:

- Chol (serum cholesterol in mg/dl)
- fbs (fasting blood sugar)
- restecg (resting electrocardiographic results)
- trestbps (resting blood pressure)

More number of attributes does not mean greater accuracy in machine learning. Reducing feature set is important and our proposed system identifies the attributes which affect the heart disease prediction the most. Based on the correlation coefficient obtained for the above attributes, we can safely say that one can remove them for the dataset for greater accuracy. There is a good positive correlation between: Target and cp (Chest pain type).

B. Key Influencers of Cleveland Database

From the visualizations in figure 3, the following relations between the attributes and disease detected can be perceived.

1) When Exercise induced angina is '0', target is 3.00 times more likely to be 'Disease detected' compared to other values of this attribute.
2) When Chest Pain Type increases (by 1.03), the likelihood of Target is 'Disease detected' also increases (by 2.33x).
3) A fall of 1.16 in (ST depression induced by exercise relative to rest) value leads to a 2.39x growth in the likelihood.
4) As maximum heart rate increases (by 22.87), the likelihood of Target is 'Disease detected' also increases (by 1.44x growth).
5) One segment (100% likely of target to be 'Disease detected') found in 18.6% of data is: Age is less than or equal to 54, Chest pain type greater than 0, maximum heart rate is greater than 152, and is less than or equal to 2, respectively.

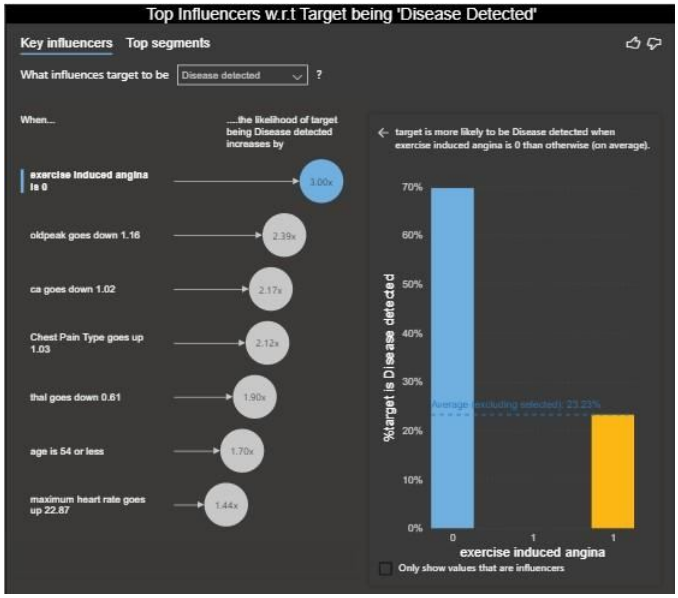


Fig. 3. Top influences with respect to Target

From figure 4 below, we can see the data distribution according to attributes.

a) It is observed that when the age is 54 or less, the target is more likely to be Disease detected. The likelihood of target to be 'Disease detected' increases by 1.70x.
b) According to the data, Males have higher chances of being detected positive for heart diseases than Females.

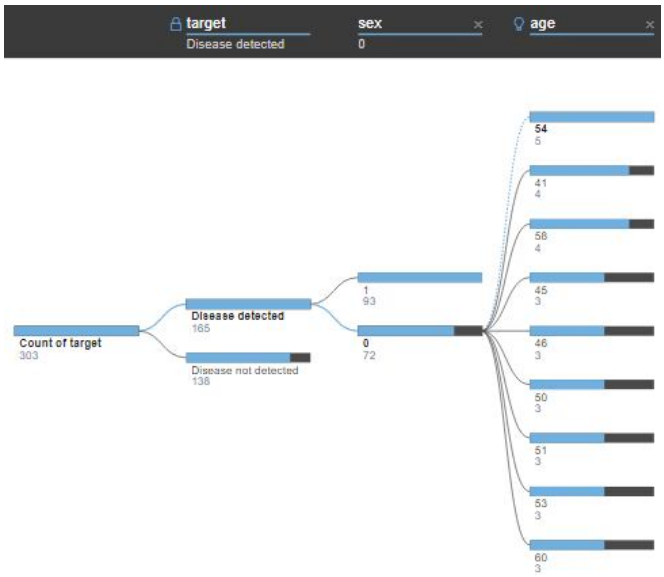


Fig. 4. Target by Sex and Age

C. Chest Pain vs Heart Disease

Experiencing discomfort or pain in the chest can be very alarming as we feel something is wrong with our heart. Chest pain usually is of 4 types which include: Asymptomatic, Atypical Angina, Typical angina, and Non-anginal pain. The type of chest pain encountered matters a lot when predicting whether a person has heart disease. The below plotted table gives a clear idea about the same.

As shown in table 4, maximum people with no heart disease lie in the category of chest pain type as 'Typical angina'. 75% of people who are 'Heart Disease not detected' are having chest pain of type 0 or typical angina. While people with chest pain type as 'Non-anginal pain' and 'Atypical Angina' have high likelihood to suffer from any heart disease. 42% of people having chest pain type as Non-anginal pain are categorized as 'possibly heart disease people'. The stacked chart below shows the same.

TABLE IV
Prediction of Chest Pain vs Heart Disease

CP target	Asymptomatic	Atypical Angina	Non-anginal pain	Typical angina
No heart disease	7	9	18	104
Possible heart disease	16	41	69	39

It is safe to conclude that as Chest Pain Type increases (0: typical angina to 3: Asymptomatic), the likelihood of a person having any heart related disease also increases.

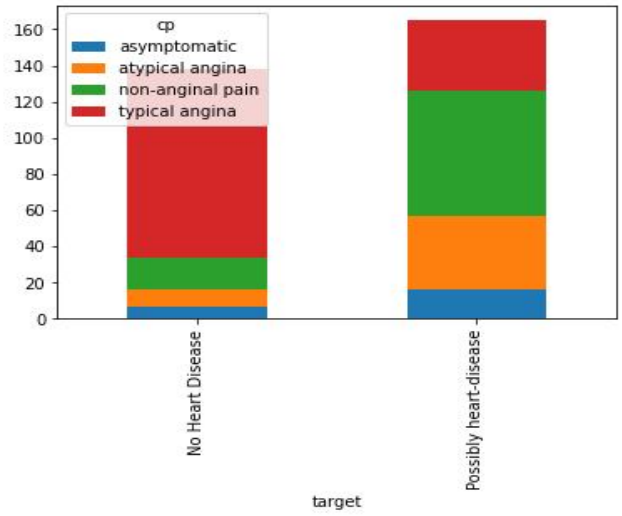


Fig. 5. Plot of likelihood of chest pain being a symptom of heart disease.

D. Resting-ECG vs Heart Disease

When the resting electrocardiographic result is obtained as ST-T Wave abnormality, the person is more likely to suffer from heart disease. While if resting electrocardiographic result is normal, chances are more than one is free from any heart related disease.

Table V
Prediction of Resting-ECG vs Heart Disease

Rest-ecg Target	ST-T wave Abnormality	Normal	Probable or definite Left ventricular hypertrophy
No heart disease	56	79	3
Possible heart disease	96	68	1

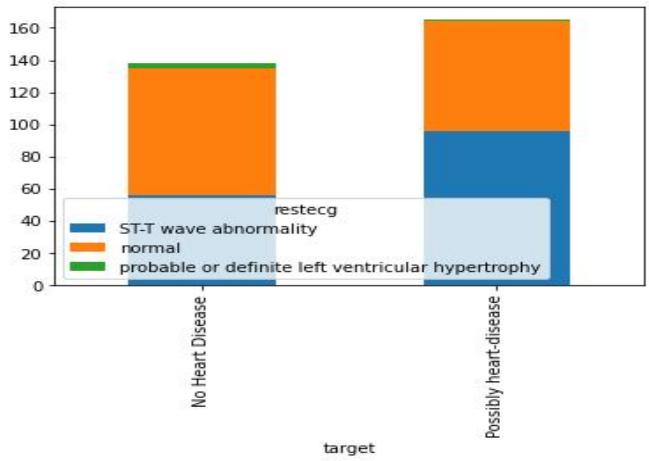


Fig. 6. Plot of likelihood of abnormality in resting-ECG being a chance of heart disease.

E. Max. heart rate achieved during thallium stress test vs Heart Disease.

People with heart disease have higher heart rate while people without any heart disease are having comparatively lower heart rate.

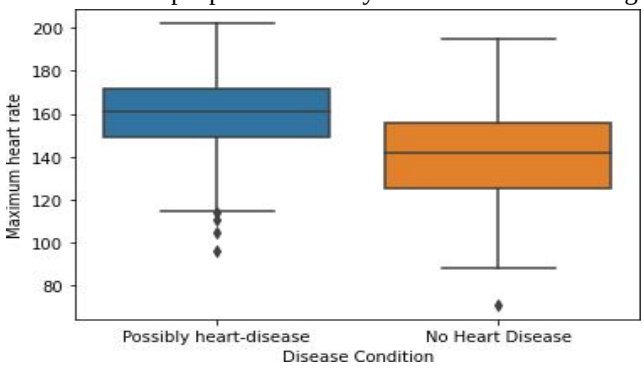


Fig. 7. Heart disease condition relating with the heart rate.

F. Serum Cholesterol

Insights gained from the box plot made by the study of serum cholesterol are.

1) From figure 7, we discovered that when Serum cholesterol is in the range 200-300, there is not much difference in the number of people who possibly have heart disease and people who were tested negative for the same. The same was seen in correlation matrix (coefficient -0.085).

2) Above 400 value, chances are more of target being 'Disease Detected' but as shown in the chart, serum cholesterol does not have much impact in predicting the target value.

3) In consequence to serum cholesterol not having much influence in predicting the target value, the attribute can be removed or not taken into consideration unlike other attributes having high correlation with the target.

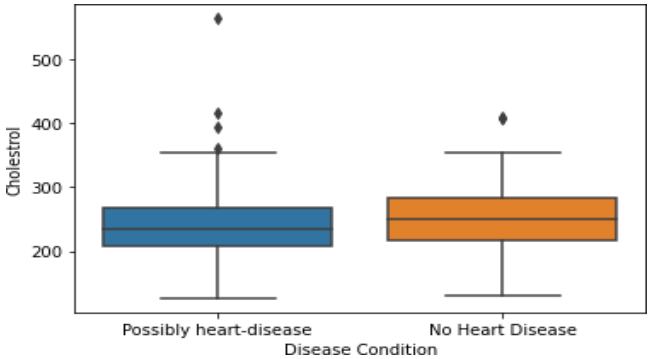


Fig. 8. Heart disease condition relating with Serum Cholesterol.

V. FUTURE SCOPE: ARTIFICAL NEURAL NETWORK(ANN)

Artificial Neural Network is biological intelligence model based on neural network structure of brain. It is an endeavour to simulate the network of neurons that act and behave as a human brain so that the computers will learn things and make choices as humans do. It is since neurons in our brain are linked to each other in various networking layers and this is how information is passed and processed.

The human brain consists of millions of neurons. It confers and measure signs as electrical and chemical signs. Human neural network comprises of 4 significant parts which include Dendrites, Cell Nucleus, Synapse, and Axons, respectively. Cell Nucleus just like in any cell resides inside the cell body of the neuron and is the main part of the neuron. Dendrites are the short branches at the top of cell body and Axons are the long tail following the cell body. These act as receivers and transmitters, respectively. The signal from the dendrite of first neuron travels down through its cell body to axon and then is passed on to the dendrite of the second neuron. These neurons are associated with a special design known as synapses. Synapses are the gaps between two neurons, which connect them and hence help pass signals. They grant neurons to pass signals from gigantic amounts of recreated neurons neural networks structures.

ANNs follow the same structure. Instead of Dendrites, Cell Nucleus, Synapse, and Axons, artificial neural networks have Inputs, nodes, weight, and output, respectively. They are ordinarily coordinated in layers. In place of neurons, we have nodes and millions of nodes are structured in these layers. There are three main layers in ANNs, which are input layer, hidden layer, and output layer. All layers are comprised of nodes which act as neurons. It is an association between nodes' input and output. The input signal and output signal are like dendrites and axons. The model receives the input signal which are our independent variables from external sources and then these are calculatedly assigned by computer to the succeeding hidden layers. Finally, the output generated is nothing but the dependent variable.

1) *Input layer:* The reason behind this layer is to get information input estimations of the explanatory attributes for each perception through an external source.
2) *Hidden layer:* The Hidden layers apply offered changes to the input values inside the network. It changes the input received in some form which is useful for the output layer.
3) *Output layer:* Output layer gets associations from hidden layers. It returns output value that looks at to the forecast of the reaction variable.

A detailed study of deep neural networks on the proposed data set can be expected in the future.

VI. CONCLUSIONS

The heart disease prediction system used by all researchers is the Cleveland database which has 13 attributes (excluding the target attribute).

The techniques used by other researchers were naïve bayes, CART classifier, Decision trees, etc. The above methods were used to provide the best classifier. Later, 2 more attributes that is smoking, and obesity were introduced. And it resulted in a 15-attribute system and increase in accuracy, respectively.

Our proposed system contains 8 attributes which gives an accuracy of 79% with decision trees model, 81.5% with naïve, and 85.5% with random forests model, respectively. Attribute subset selection and important insights on the same were gained using the famous reporting tool PowerBI and python visualizations.

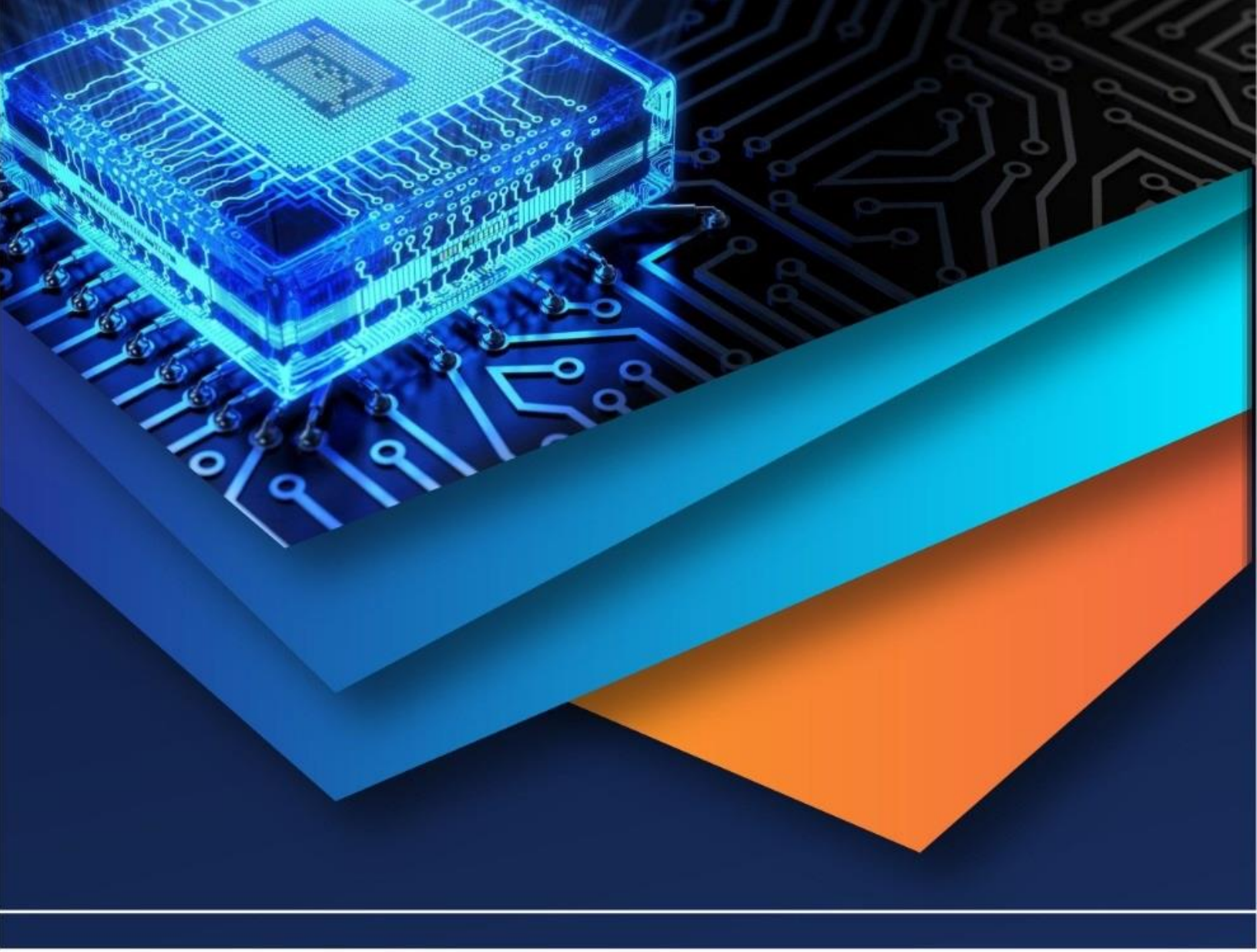
VII. ACKNOWLEDGMENT

We wish to acknowledge our university and the faculties for providing an inspirational and encouraging support.

REFERENCES

[1] K. Hafeeza, R Mohanraj, "Classification of Multi Disease Diagnosing and Treatment Analysis Based on Hybrid Mining Technique", March 2014, Volume 06, Issue No. 03, Pages 108-116
[2] Nidhi Bhatla, Kiran Jyoti, "An Analysis of heart disease prediction using different data mining techniques" 2012, Vol.1 issue 8.
[3] Manlik Kwong, Heather L. Gardner, Neil, Virginia, "Optimization of Electronic Medical Records for Data Mining Using a Common Data Model", Research Article, December 2019Volume 37, Publisher: ScienceDirect.
[4] K. Gomathi Kamaraj, D. Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", Year 2016, Conference Paper, Publisher: ResearchGate.
[5] Robert Nisbet, Gary Miner, Jhon Elder, "Handbook of Statistical Analysis and Data Mining Applications 1st Edition", May 2009, Page Count 864.

[6] Chaurasia V, Pal S (2013) Early prediction of heart diseases using data mining techniques. Carib J Sci 1:208–217.
[7] Chadha, R. Mayank, S. Prediction of heart disease using data mining techniques. CSIT 4, 193–198 (2016).
[8] https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm
[9] https://www.javatpoint.com/artificial-neural-network
[10] https://www.kdnuggets.com/2020/09/performance-machine-learning-model.html
[11] https://www.kaggle.com/ronitf/heart-disease-uci
[12] https://www.investopedia.com/terms/a/artificial-neural-networks-ann.asp#:~:text=An%20artificial%20neural%20network%20(ANN)%20is%20the%20piece%20of%20a,by%20human%20or%20statistical%20standards.
[13] https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY