HEART DISEASE PREDICTION USING MACHINE LEARNING

NISHA GUPTA[1], GULBAKSHEE DHARMALE[2], DARSHANA PARMAR[3]

[1]UG Student, Dept. of Computer Engineering, Parul University, Gujarat, India.
[2, 3] Professor, Dept. of Computer Engineering, Parul University, Gujarat, India.
Computer Engineering,
Parul University, Gujarat, India.

ABSTRACT *:* Heart is the next major organ comparing to brain which has more priority in Human body. It pumps the blood and supplies to all organs of the whole body. Prediction of occurrences of heart diseases in medical field is significant work. Data analytics is useful for prediction from more information and it helps medical center to predict of various disease. Huge amount of patient related data is maintained on monthly basis. The stored data can be useful for source of predicting the occurrence of future disease. Some of the data mining and machine learning techniques are used to predict the heart disease, such as Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbor(KNN), Naïve Bayes and Support Vector Machine (SVM).

Index Terms **- Heart Disease Prediction, Disease prediction Techniques, Medical diagnosis using Machine Learning, Machine Learning**

I. INTRODUCTION

"Computers are able to see, hear and learn. Welcome to the future"- Dave Waters.

Artificial Intelligence and Machine Learning and Deep Learning are the concepts that have been around for quite a few decades now and have been implemented or thought to be implemented many times, to make the machines do possibly everything that the humans can do without being explicitly instructed. Haffner (2016) very appropriately wrote that Machine Learning is composed of algorithms that educate computers to carry out tasks that are performednaturally and effortlessly by humans on a daily basis. For example, reading, deciding and marking an email as spam; or simply looking at the weather and deciding if an umbrella would be rquired when going out; or merely recognizing the features of a given fruit and identifying whether it is an apple or an orange. Extending this thought, in her article, Priyadharshini (2017) mentioned that machine learning enables computers to find intuitive information by using algorithms that repeatedly learn from data instead of being explicitly programmed about where exactly to look for a piece of information.

II. LITERATURE REVIEW

III. 2015	Sharma Purushottam	Efficient Heart Disease Prediction System using Decision Tree.	Decision treeclassifier
2015	Boshra Brahmi	Prediction and Diagnosis of Heart Disease by Data Mining Techniques.	J48, Naïve Bayes, KNN,SMO
2015	Sairabi H. Mujawar	Prediction of Heart Disease using Modified K-means and by using Naïve Bayes.	Modified k-means algorithm, naive bayesalgorithm.
2015	Noura Ajam	Heart Disease Diagnoses using Artificial Neural Network.	ANN
2015	Sharma Purushottam	Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree.	C4.5 rules and Naive Bayesalgorithm
2016	Marjia	Prediction of Heart Disease using WEKA tool.	K Star
			J48
			SMO
			Bayes Net
			Multilayer Perception

2016	S. Seema	Chronic Disease Prediction by mining the data.	Naïve Bayes
			Decision Tree
			Support VectorMachine
2016	Ashok Kumar Dwivedi et al[10]	Evaluate the performance of different machine learning techniques for heart disease prediction.	Naïve Bayes
			KNN
			Logistic Regression
			Classification Tree
2016	K. Gomathi et al,[16]	Multi Disease Prediction using Data Mining Techniques.	Naïve Bayes
			J48
2016	Jayamin Patel et al, [37]	Heart Disease Prediction using Machine Learning and Data Mining Technique.	J48, Logistic model tree algorithm, Random forestalgorithm
2016	Ashwini Shetty A et al, [18]	Different Data Mining Approaches for Predicting Heart Disease.	WEKA tool, MATLAB. Neural Network
			Hybrid Systems
2016	Prajakta Ghadge et al, [22]	Intelligent Heart Disease Prediction System using Big Data.	Hadoop, Mahout,Naïve bayes.
2016	S. Prabhavathi et al, [23]	Analysis and Prediction of Various Heart Diseases using DNFS Techniques.	Decision tree, c4.5, SVM,naïve bayes.
2016	Sharan Monica. L et al,[25]	Analysis of CardioVasular Disease Prediction using Data Mining Techniques.	J48
			Naïve Bayes
			Simple CART
2017	Jayami Patel et al,[14]	Heart disease Prediction using Machine Learning and Data mining Technique.	LMT, UCI
2017	P. Sai Chandrasekhar Reddy et al, [17]	Heart disease prediction using ANN algorithm in data mining.	ANN
2018	Chala Bayen et al,[12]	Prediction and Analysis the occurrence of Heart Disease using data mining techniques.	J48, Naïve Bayes, Support Vector Machine.
2018	R. Sharmila et al, [13]	A conceptual method to enhance the prediction of heart diseases using the data techniques.	SVM in Parallel fashion

RESEARCH GAP

A major challenge confronting healthcare associations i.e. hospitals, medicinal focuses are the procurement of quality services at reasonable expenses. Quality services suggest diagnosing patients accurately and overseeing medicines that are more effective.

Poor clinical decisions can prompt to poor outcomes which are therefore unsatisfactory. Healthcare organizations can reduce costs by accomplishment of computer based data and/or decision support systems.

Healthcare services data is very huge as it incorporates patient records, resource management information and updated information. Human services associations must have capacity to break down information. Treatment records of many patients can be stored away in computerized way; furthermore data mining methods may help in finding out a few vital and basic inquiries related with healthcare organizations.

RESEARCH METHODOLOGY

One of the existing study applying neural network to self-applied questionnaire (SAQ) data to develop a heart disease prediction system. The validation of the work was provided by checkingagainst the result of the neural network with "Dundee Rank Factor Score" which is related to statistically 3 risk factors (blood pressure, smoking and blood cholesterol) together with sex andage to determine risk of having heart disease. In the study, they used multi-layered feedforward neural network which was trained with Backpropagation Algorithm:

The calculations included K Neighbors. Classifier, Support Vector Classifier, Decision Tree Classifier and random Forest Classifier. The dataset has been taken from Kaggle.

Import libraries

1) numpy: To work with exhibits
2) pandas: To work with csv records and data frames
3) matplotlib: To make diagrams utilizing pyplot, characterize parameters utilizing rcParams and shading them with cm.rainbow
4) warnings: To overlook all admonitions which may be appearing in the scratch pad due to past/future deterioration of an element.
5) train_test_split: To part the dataset into preparing and testing information
6) StandardScaler: To scale every one of the highlights, so the Machine Learning model better adjusts to the dataset

Import dataset

In the wake of downloading the dataset from Kaggle, I spared it to my working registry with the name dataset.csv. Next, I utilized read_csv() to peruse the dataset and spare it to the dataset variable. Before any examination, I simply needed to take a gander at the information. In this way, I utilized the information () technique. As should be obvious from the yield above, there are a sum of 13 highlights and 1 objective variable. Likewise, there are no missing qualities so we don't have to deal with any invalid qualities. Next, I utilized portray () strategy. The technique uncovered that the scope of every factor is unique. The most extreme estimation of age is 77 yet for chol it is 564. Along these lines, highlight scaling must be performed on the dataset

Understanding the data

Correlation Matrix

In any case, we should see the connection lattice of highlights and attempt to investigate it. The figure size is characterized to 12 x 8 by utilizing rcParams. At that point, I utilized pyplot to show the connection framework. Utilizing xticks and yticks, I've added names to the relationship network. colorbar() shows the colorbar for the framework.

Information Mining

In this undertaking, I took 4 calculations and differed their different parameters and analyzed the last models. I split the dataset into 67% preparing information and 33% testing information

Information Mining

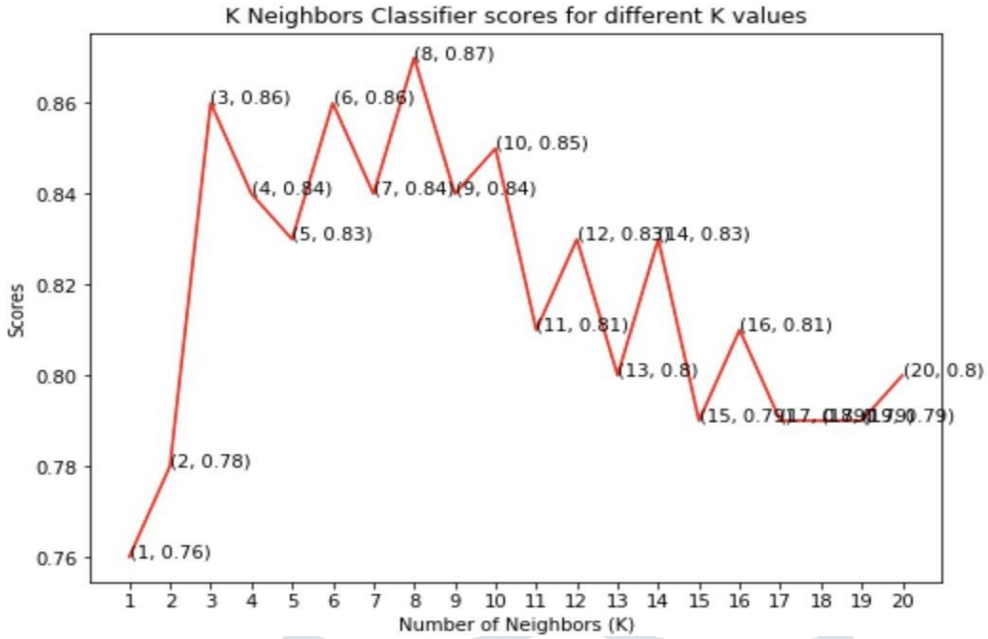
In this undertaking, I took 4 calculations and differed their different parameters and analyzedthe last models. I split the dataset into 67% preparing information and 33% testing information.

K Neighbors Classifier

This classifier searches for the classes of K closest neighbors of a given information point and dependent on the lion's share class, it allots a class to this information point. Be that as it may, the quantity of neighbors can be fluctuated. I changed them from 1 to 20 neighbors and determined the test score for each situation.

By then, I plot a line outline of the amount f neighbors and the test score achieved for everycircumstance.

Figure 4.1.1.4 K Neighbors Classifiers for different K values [37]



As should be obvious, we accomplished the most extreme score of 87% when the quantity of neighbors was picked to be 8.

Support Vector Classifier

This classifier targets framing a hyperplane that can isolate the classes however much as could reasonably be expected by changing the separation between the information focuses and the hyperplane. There are a few parts dependent on which the hyperplane is chosen. I attempted four pieces to be specific, direct, poly, rbf, and sigmoid. When I had the scores for every, I utilized the rainbow strategy to choose various hues for each bar and plot a visual diagram of the scores accomplished by each.

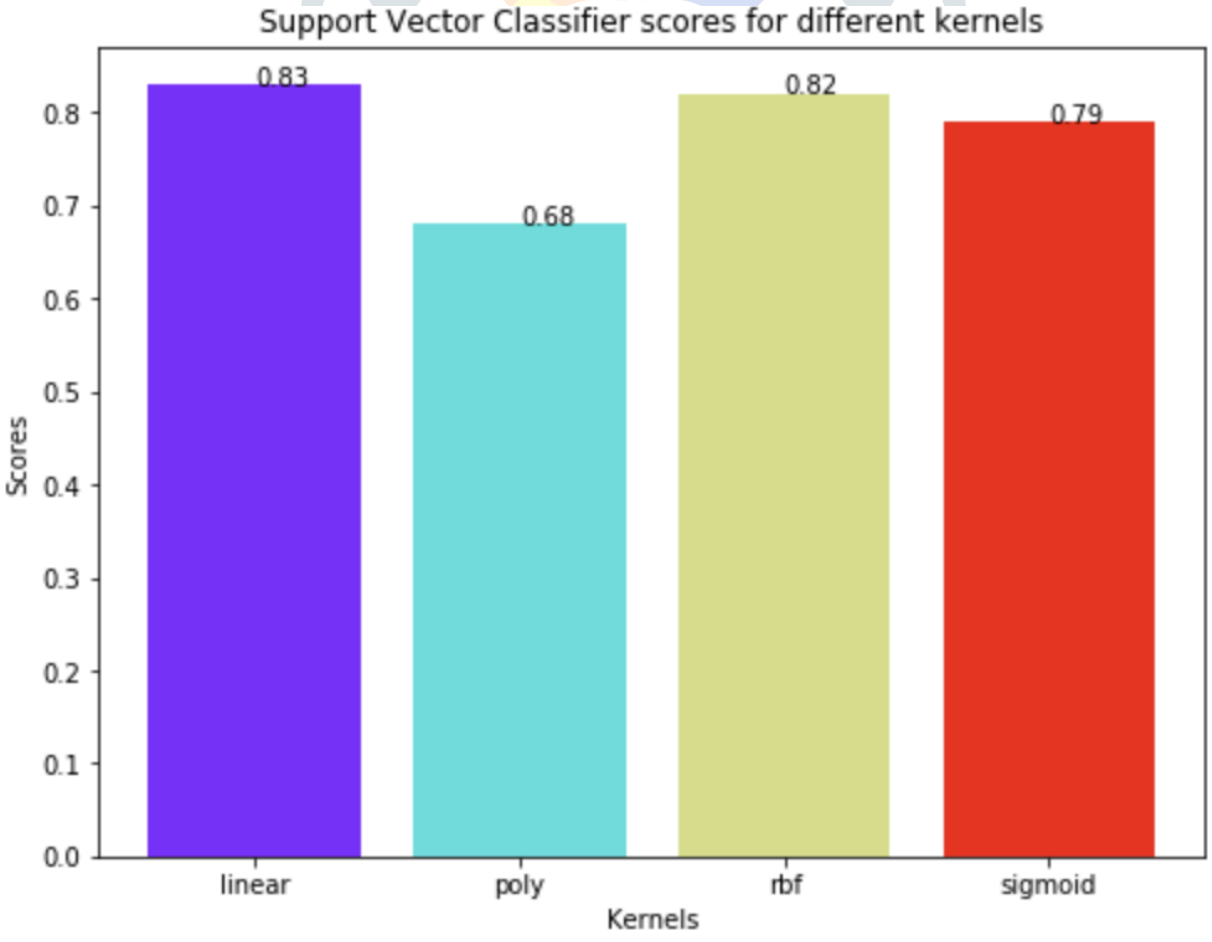


Figure 4.1.1.5 Support vector classifier scores for different kernels

As can be seen from the plot over, the straight piece played out the best for this dataset andaccomplished a score of 83%.

Decision Tree Classifier

This classifier makes a choice tree dependent on which, it allots the class esteems to every datum point. Here, we can shift the most extreme number of highlights to be considered while making the model. I extend highlights from 1 to 30 (the absolute highlights in the dataset after sham sections were included).

When we have the scores, we would then be able to plot a line chart and see the impact of the quantity of highlights on the model scores.

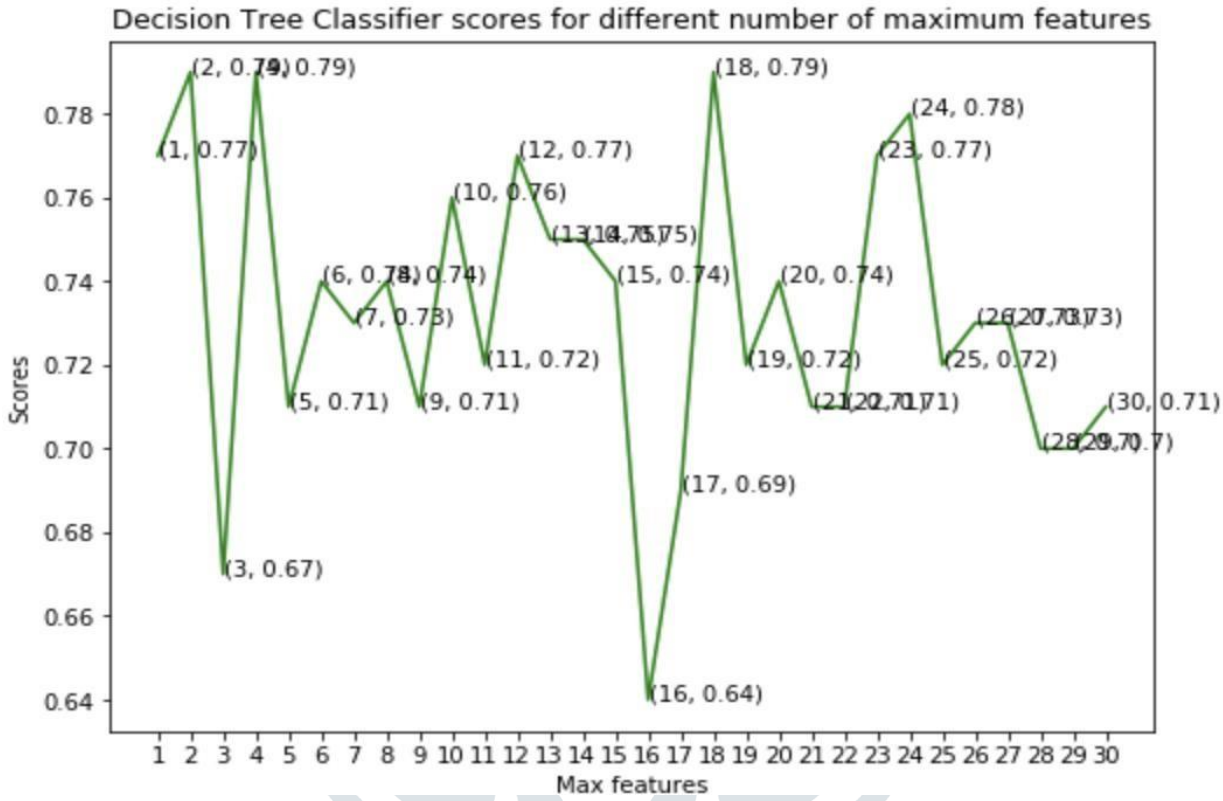


Figure 4.1.1.6 Decision Tree classifier scores for different number of maximum features

From the line diagram above, we can plainly observe that the most extreme score is 79% andis accomplished for greatest highlights being chosen to be either 2, 4 or 18.

Random Forest Classifier

This classifier takes the idea of choice trees to the following level. It makes a woods of treeswhere each tree is framed by an irregular determination of highlights from the complete highlights. Here, we can differ the quantity of trees that will be utilized to foresee the class.I compute test scores more than 10, 100, 200, 500 and 1000 trees.

Next, I plot these scores over a reference diagram to see which gave the best results. You may see that I didn't legitimately set the X esteems as the exhibit [10, 100, 200, 500, 1000].It will show a nonstop plot from 10 to 1000,which would be difficult to disentangle. Alongthese lines, to comprehend this issue, I initially utilized the X esteems as [1, 2, 3, 4, 5]. At that point, I renamed them utilizing xticks.

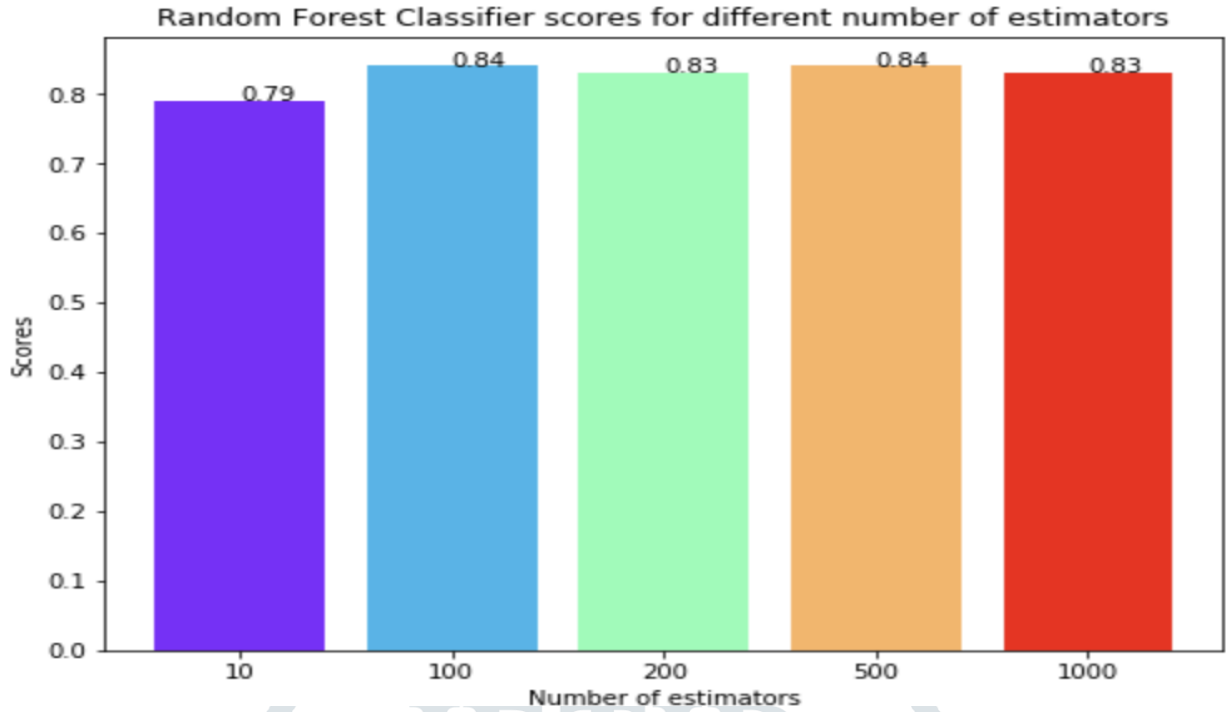


Figure 4.1.1.7 Random Forest classifier scores for different number of estimators [37]

Taking a glance at the reference diagram, we can see that the most extreme score of 84% was accomplished for both 100 and 500 trees.

Sr No.	Algorithm	Accuracy in %
1.	K Neighbors Classifier	87
2.	Support Vector Classifier:	83
3.	Decision Tree Classifier	79
4.	Random Forest Classifier	84

Table 4.1.1.1 Comparison of accuracy of different algorithms [37]

The work included investigation of the coronary illness understanding dataset with legitimate information preparing. At that point, 4 models were prepared and tried with greatest scores as pursues.

K Neighbors Classifier scored the best score of 87% with 8 neighbors

IV. REQUIREMENTANALYSIS

4.1 Data collection

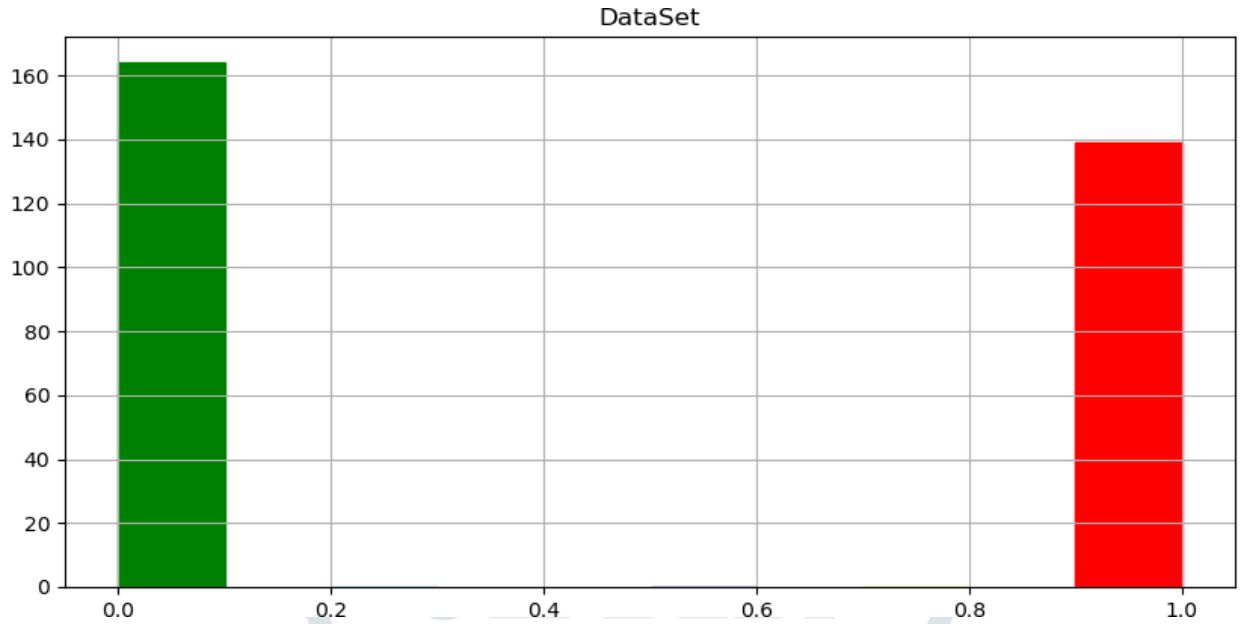
The data collection process involves the selection of quality data for analysis. Here we used Heartdisease dataset taken from uci.edu for machine learning implementation. The job of a data analystis to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques

4.2 Data visualization

The dataset collected with attributes age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slop, ca, thal, pred_attribute. A large amount of information represented in graphic form is easier to understand and analyze. Some

companies specify that a data analyst must know how to create slides, diagrams, charts, and templates. In our approach, the heart disease rates is shown as data visualization part.

Table 1 Dataset Descriptions



4.3 Data preprocessing

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

4.4 Dataset splitting

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

Training set. A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data. Test set. A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

4.5 Model training

After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

4.6 Model evaluation and testing

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to achieve an algorithm's best performance.

IV. ACKNOWLEDGMENT

I hereby avail this opportunity to express gratitude to a number of people who extended their valuable time, full support and cooperation to undertake the dissertation.
I am very grateful to Dr. Vipul Vekariya, Principal of Parul Institute of Engineering and Technology for providing facilities to achieve the desire milestone. I also extend my thanks to Head of Department and my dissertation guide Dr. Gulbakshee Dharmale & Prof. Darshana Parmar for their inspiration and continuous support by providing facilities for Dissertation work. I also express my gratitude to the entire faculty member of Computer Science & Engineering and Information Technology department.

V. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries. The heart disease dataset downloaded from uci.edu is considered for study. Machine learning algorithm is applied such as decision tree, and Random forest. We used these machine learning algorithm and indentified heart disease. To improve the work and novelty of the work, we implemented Decision Tree and Random Forest. The result shows that Heart disease detection is efficient using Random Forest algorithm. Random forest achieves 71.50% accuracy, Decision Tree achieves around 75% accuracy.The following table shows the accuracy arrived in our experimental study

Table 5.1: Experimental Results of proposed system

ALGORITHM	ACCURACY (%)
Decision Tree	75.00
Random Forest	71.05

VI. CONCLUSION

In this survey, major influencing factors for determining the heart disease and various research works in predicting the heart disease has been identified and reported. It is observed that not all attributes are taken into consideration by every researcher. Few attributes are eliminated to provide more accuracy by few researchers. We have carried out a detailed discussion about the key challenges of various research works for heart disease prediction that are not yet addressed. In future, the researchers should include all the factors for determining the heart disease using an effective algorithm.

In conclusion, as identified through the literature review, there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases. The research proposes a framework using combinations of support vector machines, ANN to arrive at an accurate prediction of heart disease. Using the Cleveland Heart Disease database, our research intents to provide guidelines to train and test the system and thus attain the most efficient model of the multiple rule based combinations. Further, our work proposes a comparative study of accuracy. In addition, the most effective and most weighed model can be found.

REFERENCES

1) Sharmila S et al., "Analysis of Heart Disease Prediction using Data Mining Techniques", International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05, Pages: 93-95 (2017).

2) Chaitrali S. Dangare et al., "Improved Study of heart disease Prediction system using Data Mining Classification Techniques", International Journal of Computer Applications, Volume: 47, Pages: 44-48 (2012).

3) Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159.

4) What is Predictive Data Mining?, https://www.techopedia.com/definition/30597/predictive- data-mining

5) http://www.heart.org/HEARTORG/Conditions/HeartAttack/WarningSignsofaHeartAttack/Warning-Signs-of-a-HeartAttack_UCM_002039_Article.jsp#.WNpKgPl97IU.

6) www.who.int/cardiovascular_diseases/en/.

7) http://food.ndtv.com/health/world-heart-day-2015-heart-disease-in-india-is-a- growing-concern- ansari-1224160.

8) B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

9) Sunil Ray, "Learn Naïve Bayes algorithm" and "Decision tree- Simplified", URL: www.analyticsvidhya.com, Retrieved on- 10.01.2018.

10) Kalaiselvi C, "Diagnosing of heart diseases using Average K- Nearest Neighbor Algorithm of Data Mining", 2016 InternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE, Pages: 3099-3103 (2016).

11) Hlaudi Daniel Masethe et al., "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science [WCECS], Volume: II, ISBN: 978-988-37253-7-4 ISSN: 2078-0958 (Print), ISSN: 2078-0966 (Online) (2014).

12) Theresa Princy R, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE (2016).

13) Vivekanandan T et al., "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease", www.elsevier.com/locate/compbiomed, https://doi.org/10.1016/j.compbiomed, Pages: 125-136 (2017).

14) Shamsher Bahadur Patel et al., "Predict the Diagnosis of Heart Disease patients using Classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR- JAVS), e-ISSN: 2337-2380, p-ISSN: 2337-2372, Volume: 4, Issue: 2, Pages: 61-64 (2013).

15) John Peter Tet al., "An Empirical study on Prediction of Heart Disease using Classification Data Mining Techniques", International Conference On Advances in Engineering, Science and Management (ICAESM -2012), IEEE, ISBN: 978-81-909042-2-3,