IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 11, November 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

A Survey on Heart Disease Prediction Framework using Various Data Science Techniques

Prajwal Arun Dudhe[1], Prathamesh Bajare[2]

B. E. Student, Department of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon BK, Pune, India[1]

Professor, Department of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon BK, Pune, India[2]

ABSTRACT:Heart disease is a condition where blood veins get blocked and hearts stop working. It is well known fact that heart disease is causing the most of deaths in the world. A person suffering from heart disease can recognize this disease only in the last stages. And this is the reason that most of the deaths occur. Scientist and Doctors reveals that it is not a sudden process and is the result of being on a particular lifestyle for long time and also results after giving some basic and common symptoms. So, scientists have concluded that heart disease can be detected in the early stages with the help of data science algorithms. Hence, the goal of this paper is to develop a data science framework which gives an idea to discover the chances of existence of heart disease by applying different classification algorithms, influence and distribution of various parameters that are playing major role in disease prediction along with visualizations on Cleveland cardiovascular medical records. There are many researches related to that problem but the accuracy of prediction is still needed to be improved. So, this paper focuses to find the optimal classification algorithm on the heart disease for showing the accuracy improvement. These algorithms can be used for predicting that the person has a heart disease or not, on the classification reports. This experimental work was done with the help of various algorithms such as Random Forest, Vector support, Logistic regression and XG-Boost for building the heart disease prediction model and evaluates the performance of the model.

KEYWORDS:Heart disease,Data science, symptoms, classification algorithms,framework,visualizations.

I. INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, algorithms and systems to extract knowledge from noisy, structured and unstructured data, and apply it across a broad range of application domains. Today, data science using bio medical records and clinical features is used to create good models for inference. Forecasting of disease progress has become one of the most faced challenge. Due to the improvement in technologies like image processing and recognition with the help of Magnetic resonance imaging(MRI), X-ray, mammography, etc, they are used for Data acquisition and processing.

Categorization means method of data processing in medical domain. It is used to assume the target variable for each data point in the dataset. The classification methods contains K - Nearest Neighbor, Random Forest, Support Vector Machine, etc

Heart disease symptoms depends on the sex. Men are more like to get chest pain and Women with chest pain can also have difficulty in breathing and fatigue. This disease kills a lot of people each year .With the rapid increase in the heart stroke rates at juvenile ages, the prediction of the existence of heart disease is an important role to prevent heart attack. In this paper, a comparison is done on different algorithms like XG-Boost (XGB), Logistic Regression , Naive Bayes (NB), Random Forest classifier to check the classification accuracy for diagnosing CKD and find the best algorithm.

II. LITERATURE SURVEY

After going through previous research papers, conclusion is drawn from the papers. Anjan Nikhil Repaka et al. [2] assessed the information and building up Smart Heart Disease Prediction that utilizes the Naive Bayesian methodology and Classification and Regression with an exactness of 89.77%. Aditi Gavhane et al. [3], have tested the multiple layer perceptron method to assemble the prescient model. Aakash Chauhan et al [4] The dataset utilized in this work is Heart Disease Cleveland Database. In this paper data mining procedure is utilized for coronary illness forecast with a triumph

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

pace of 60%. Sowmiya et al [5] dissected a few grouping classification techniques of data mining that can uncover significant elements and identify heart disease. R Latha et al. [6]. This experiment is in regards to heart diseasee utilizing POMDP which gives the state of the patient. Sarath Babu et al.[7] The focal point of the paper is utilizing various algorithms and discovering data in data mining. Out of all attributes just couple of attributes were utilized with the goal that the exactness of Heart disease expanded Gaurav Meena et al. [8] existing work to discover significant data in data mining HDD and data mining procedures in Bioinformatics. Rashmi G Saboji et al [9] Assessed the data that utilizes Random forest approach on Spark system with an exactness of 98%.

III. METHODOLOGY

Following fig 1 shows the process of Framework for Heart Disease Prediction.

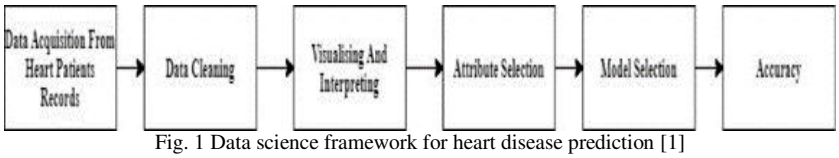

Fig. 1 Data science framework for heart disease prediction [1]

A. Data Acquisition

Data acquisition is the processes for gaining data for training model. Medical repository is used for prediction of heart disease which is obtained from medical reports which are gained from Cleveland through UCI repository [12]. There are 300 records with 76 various attributes in this database but only 14 of these attributes are used for heart disease prediction like Age (in years), Sex (Male or Female), chest pain type like typical angina, atypical angina, non-anginal pain, trestbps (the person's resting blood pressure), cholesterol, fasting blood sugar (fbs > 120mg/dl), rest electro cardiograph, thalach (maximum heart rate), angiographic status.

B. Data Cleaning

Data cleaning is the process that removes data that does not belong in specific dataset. This step includes the removal of noisy data and waste, identification of null and not applicable data items and treatment of that data parts are required to be performed. If data contains unfiltered data then the it will affect the end accuracy of model. Hence, one of the crucial steps as removing unfiltered and dummy values from the dataset is performed. In the dataset we have selected, there are many missing values in each attribute. Hence, we need to fix them. So, these values are replaced by the median values of the specific attributes in dataset [13]. The Heart disease dataset Target variable consists of "Yes" and "No" as target labels. Converting strings values to numerical values such as assigning Yes as 1 and No as 0 respectively. The data wrangling of Heart disease UCI data set is already performed. So, while processing the data the algorithms can work correctly.

C. Data Visualizing and interpreting

Data visualization is the process of presenting data into graphical presentation such as charts, graphs and other visuals to help analyse and interpretdata. Checking the correlation between variables is one the important task. It checks whether one attribute is dependent on another attribute. Neglecting orthodox ways, we visualize data to comprehend. Box plot, heat map and ROC plots are used for visualizing data. By these graphical representations one can analyse the data.

D. Attribute Selection

Attribute selection is also known as variable selection or feature selection. It focuses on decreasing the number of irrelevant attributes by some techniques. Selecting the required features is one of the crucial tasks to get the best results and less time to train the model. Attribute selection is used to decrease the training time, evaluation time and increase the accuracy of prediction. Therefore, to achieve all of this it is necessary to find the relevant feature or attributes. After finding relevant features the Heart disease dataset predicted output is almost equal to combining all the features together. The top features were age, sex, fasting blood sugar>120, type of chest pain, target.

E. Model selection

There are various data science algorithms based on which we can create models for Heart Disease Prediction. Following are the various Data Science algorithms used in this paper.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

a) Logistic Regression (LR): Logistic regression is a supervised classification model which is used to predict the odds in favour of a specific event which we want to predict. It is used in machine learning to solve classification problem many times. This algorithm works on sigmoid function which is used to determine the predicted value with the help of threshold value. It is used to apply on the variables which can be classified.

b) Support Vector Machine (SVM): Vector Support Machine is a very powerful supervised classification to maximize the margin among class variables. It uses hyperplanes as a limit between different classes for decision making. This large margin helps to deal with classification of positive and negative hyperplanes with adjustable bias-variance ratio. SVM is not prone to outliers as it only takes the points which re closest to decision boundary.

c) Naïve bayes: Naive Bayes is a simple algorithm for constructing classifiers. Bayes theorem is used in this algorithm. A Naive Bayes classifier assumes that there is no relation between presence of a specific feature in a class and the presence of another feature. Naive Bayes model is not hard to build and particularly useful for complex and big data sets. It can also perform better than highly sophisticated classification methods.

d) Random Forest (RF): Random forest is a decision treebased algorithm. The process of fitting number of decision trees on different subsample and then taking out the average to improve the performance of the model is called "Random Forest". It is an ensemble technique. This algorithm is used to solve both regression and classification problem.

e) XG-Boost (XG): XG-Boost is termed as Extreme Gradient Boosting Algorithm which is an ensemble method that works by boosting trees. It is a structured or tabular information algorithm. It uses the gradient descent algorithm which why it is called Gradient Boosting. In this algorithm gradient-boosted engines are implemented for speed and performance. The speed of execution of XG Boost is very quick as compared to other gradient boosting implementations.It is used on classification and regression predictive modelling problems.

F. Model testing with Accuracy Parameters

1) Precision: Precision is the measure to check correct predictions made over the amount of correct and incorrect predictions. Precision is used to determine if the cost of false positive is high. It is shown in table[2] for each classifier.

$$\text{precison} = \frac{TP}{TP+FP}$$

Where, TP is True positive,
FP is False positive.

2) Recall: Recall is the measure of correct predictions made by our model over the total amount of observations. It is the ability of a classification model to recognize all relevant instances. It is in table[2] for each classifier.

3) F1-measure: F1-measure is a measure which is used to find how good the model is predicting.

4) **Receiver Operating Characteristics (ROC):** It is a plot for continuously showing the trade-off between the true positive rate and false positive rate for a binary classifier at different classification thresholds.

The next section is discussed on results.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

IV. EXPERIMENTAL RESULTS

A. Distribution graphs of all attributes:

It is used to observe the data. And found relations between features are shown in figure[2-7] [1].

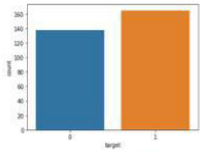

Fig. 2. Disease observations

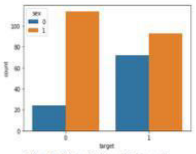

Fig. 3. Plot of Gender based observations

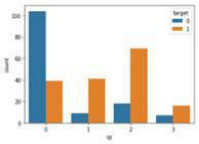

Fig. 4. Plot of Gender vs Chest pain

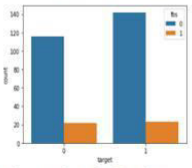

Fig. 5. Plot of Blood sugar impact

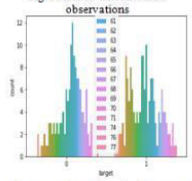

Fig. 6. Plot of Cholesterol impact

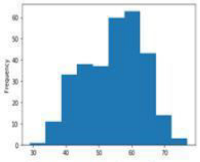

Fig 7. Area plot of age influence

1) Heat Map analysis:A heatmap could be a way of graphical presentation where discrete values of a matrix are represented as different colours. This analysis is used determine the correlation between attributes. The reciprocal relationship between all characteristics is employed for evaluation. So, the association between attributes are often determined using describe function. This correlation is termed heat map. Heat maps can help the user visualize easy or complicated information. the heat map in fig 8 is observed using scale where the intensities of the scales tells how well the features are co-related.

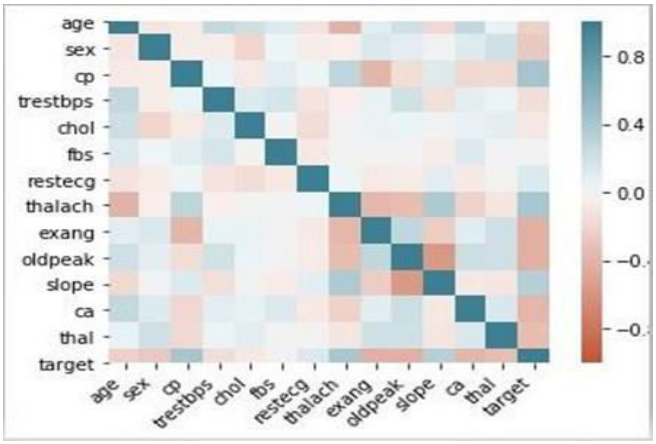

Fig. 8. Heat map to represent correlation among attributes [1]

2) Box Plot: Box plots visually display the distribution of numerical data and skewness via presenting the data quartiles and averages. Horizontal and vertical box plot can be drawn. Shape of the distribution of age with chest pain among all levels is observed using this plot in figure 9.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

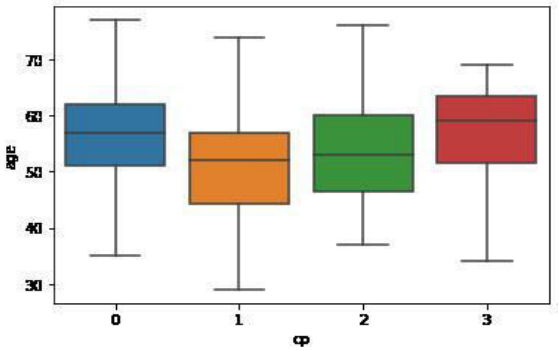

Fig. 9. Box plot on distribution of chest pain among various age groups [1]

B. Confusion Matrix:

The confusion matrix is a matrix used to evaluate the performance of the classification model for a given set of test data. It can only be determined if the true values for test data are known otherwise it's unable to construct a confusion matrix. It displays how many values are correctly predicted and how many values are incorrectly predicted by a model. In Table(1) it shows the confusion matrices for various classification models.

Model	Confusion Matrix
LR	[33 13] [8 46]
SVM	[33 13] [8 46]
XGB	[35 11] [13 41]
NB	[36 10] [12 42]
RF	[35 11] [14 40]

TABLE(1).CONFUSION MATRIX OF VARIOUS MODELS [1]

C. Accuracy:

The disorderly predicted values and the total amount of times present in the dataset models is known as accuracy.

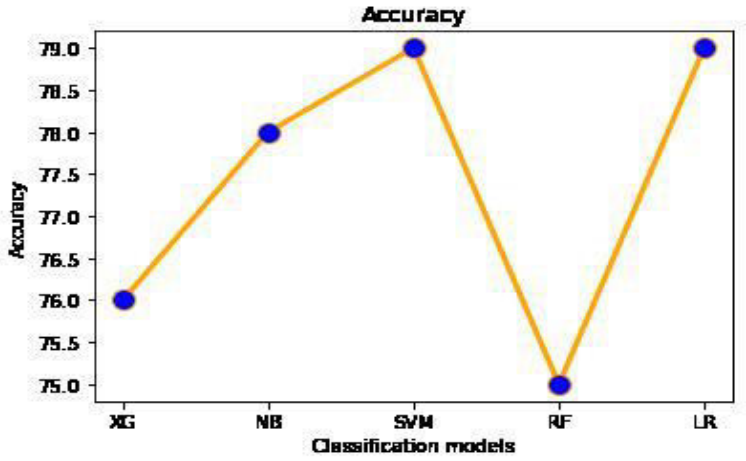

Fig. 10. Accuracy distributions among various classification models [1]

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

It can be concluded from Figure 10 that after applying classification algorithms to the dataset. SVM (79 %) and LR (79 % have the highest accuracies when compared to XG (76 %), NB (78 %), and RF (76 %) (75 %)

1) Precision: *Precision* is the measure to check correct predictions made over the amount of correct and incorrect predictions. It is used to determine if the cost of false positive is high. It is shown in table [2] for every classifier.

2) Recall: Recall is the measure of correct predictions made by our model over the total amount of observations. It is a classification model's ability to recognize all relevant occurrences. It is in table [2] for each classifier.

3) F1-score: F-measure is a measure which is used to find how good the model is predicting.

The comparison between predictive accuracy of different heart disease models is shown in table 2.

s.no	Classifier	Classification Report					
		Precision		Recall		F1-score	
		0	1	0	1	0	1
1	LR	0.80	0.78	0.72	0.85	0.76	0.81
2	SVM	0.80	0.78	0.72	0.85	0.76	0.81
3	XGB	0.73	0.79	0.76	0.76	0.74	0.77
4	NB	0.75	0.81	0.78	0.78	0.77	0.79
5	RF	0.71	0.78	0.76	0.74	0.74	0.76

TABLE(2). COMPARATIVE ANALYSIS ON PREDICTIVE ACCURACY OF HEART DISEASE [1]

D. Receiver operating characteristic(ROC) curve*:*

The following figure[11-15] illustrates the ROC curves of the classification model we have worked on, This shows the real positive levels of the false positives as a function of the threshold for classification.

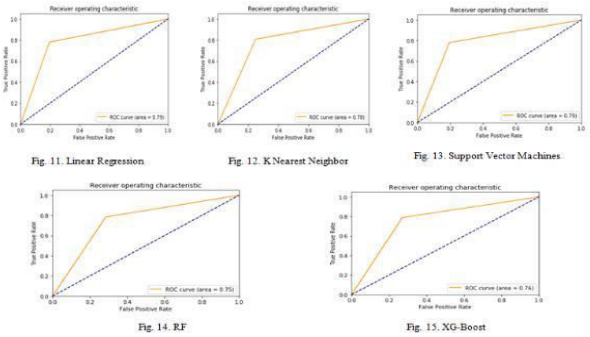

Fig. 11-15. ROC curves generated by different classifiers [1]

V. CONCLUSION

Heart disease is currently one in every of society' greatest fears. The probabilities of predicting heart disease manually on risk factors are troublesome to assess. Machine learning techniques are but helpful for estimation of heart disease from current data. The goal of this thesis is to watch the employment of algorithms for analysis and prediction of heart disease in data processing classification. In this paper, conduction of 5 classifications of heart disease prediction has been performed. The experiment of technique has incontestable that Support vector machine and logistic regression has created predominant performance in terms of classification for heart disease dataset.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011087|

REFERENCES

[1] Chittampalli Sai Prakash, Myneni Madhu Bala, Attluri Rudra, “Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations”, 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)

[2] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "
Design and Implementing Heart Disease Prediction Using Naives Bayesian", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)

[3] Aditi Gavhane, Gouthami Kokkula , Isha Pandya , Prof. Kailas Devadkar , "Prediction of Heart Disease Using Machine Learning",2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA).

[4] Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT)

[5] C. Sowmiya, P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques",2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS).

[6] R Latha, P Vetrivelan, "Blood Viscosity based Heart Disease Risk Prediction Model in Edge/Fog Computing",2019 11th International Conference on Communication Systems & Networks (COMSNETS)

[7] Sarath Babu, E M Vivek, K P Famina, K Fida, P Aswathi, M Shanid, M Hena, "Heart disease diagnosis using data mining technique", 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)

[8] Gaurav Meena, Pradeep Singh Chauhan, Ravi Raj Choudhary, "Empirical Study on Classification of Heart Disease Dataset-its Prediction and Mining",2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)

[9] Rashmi G Saboji, "A scalable solution for heart disease prediction using classification mining technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details