Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations

Chittampalli Sai Prakash
Bachelor Student,Computer Science and Engineering
Institute of Aeronautical Engineering, Hyderabad, Telangana,India
MAILID:ch.saiprakash22@gmail.com

Myneni Madhu Bala
Professor, Computer Science and Engineering
Institute of Aeronautical Engineering, Hyderabad, Telangana,India
Mail ID:baladandamudi@gmail.com

Attluri Rudra
Bachelor Student,Computer Science and Engineering
Institute of Aeronautical Engineering, Hyderabad, Telangana,India
Mail ID:attulurirudra369@gmail.com

***Abstract*— Heart disease refers to a condition where the blood vessels are blocked and the heart stops functioning. Many of the researches are concluded that this disease has become number one cause of death cases. It is alarmed that abnormalities can only be detected and recognized in its last stages. However it is curable if the person detects the disease earlier. The goal of this paper is to develop a data science framework which addresses the how to discover the chances of existence of heart disease by applying different classification algorithms, influence and distribution of various parameters are playing major role in disease prediction along with visualizations on Cleveland cardiovascular medical records. To minimize the diagnostic error caused by the complexity of visual and subjective interpretation. This work majorly aims to find the optimal classification algorithm on the heart disease affected health records and majorly influencing parameters. This can be used for predicting the heart disease on the classification reports. This experimental work focuses on the performance of the system was tested and classified by 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; Prediction Model; healthcare; Visualization.

I. INTRODUCTION

Data science framework is used for making accurate predictions using past data by considering various data insight features which are used to make predictions and finding the co-relation between these features, consequently by combining observed symptoms and then make predictions. In general these are used to extract knowledge or insights from large amounts of structured or unstructured data. Today, the impact of data science using bio medical records and clinical features has far-reaching inferences in many fields. Forecasting of disease progress has become one of the challenge. Due to the advancements of technologies like image processing and recognition with the help of Magnetic resonance imaging(MRI), X-ray, Mammography, etc are used for Data acquisition and processing.

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

Heart disease symptoms depends on the sex such for Men are more like to get chest pain and Women also have chest pain and difficulty in breathing and fatigue. This disease kills a large number of people each year .The prediction of the existence of heart disease is an important role to prevent from heart attack. “A comparative analysis 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.

II. LITERATURE SURVEY

After going through the preliminary research papers, conclusion is drawn from the papers. Anjan Nikhil Repaka et al. [1] evaluated the data and establishing Smart Heart Disease Prediction that uses the Naive Bayesian approach and Classification and Regression with an accuracy of 89.77%. Aditi Gavhane et al. [2], “have experimented the Multi layer perceptron technique to build the predictive model. Aakash Chauhan et al [3] The dataset used in this work is Heart disease Cleveland Database. In this paper data mining techniquesis used for heart disease prediction with a success rate of 60%. Sowmiya et al [4] analyzed several classification techniques of data mining that can dig out important features and detect heart disease. R Latha et al. [5]. The experiment is regarding heart disease using POMDP which gives the condition of the patient. Gaurav Meena et al. [6] existing literature work to find out major information in data mining HDD and data mining techniques in Bioinformatics. Rashmi G Saboji et al [7] Evaluated the data that uses Random forest approach on Spark framework with an accuracy of 98%. Sarath Babu et al.[8] The focus of the paper is using different algorithms and finding out information in data mining. Out of several attributes only few attributes were used so that the accuracy of predicting heart disease increased. Heart disease prediction using optimization techniques are implemented [9]. Seyedamin Pouriyeh et al.[10] They discussed about the available tools for classification and processing of data. It was applied on 7 different techniques out of which SVM outperformed other techniques. Cincy Raju et al.[11] have done a survey on heart disease prediction using data mining techniques. Data science framework for prediction models are used in various domains[14]. It gives an effective way of step wise representation. Hence in this work the proposed framework is chosen from literature[15,16] for prediction of heart disease. The authors suggested a finely tuned VGG19 model for static gesture recognition implementation. The proposed model is evaluated on the ASL dataset and the resulting recognition rate is 94.8%[17].

III. METHODOLOGY

The data science framework for heart disease predictions is shown in figure 1.

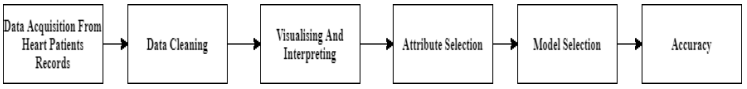

Fig.1.Data Science framework for heart disease prediction

A. Data Acquisition

Medical repository is used for prediction of heart disease which is obtained from medical reports which are collected from Cleveland through UCI repository[12]. The database contains 300 records with 76 various attributes but only 14 of these attributes is 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, cholesterol, fasting blood sugar (fbs > 120mg/dl), rest electro cardiograph, thalach (maximum heart rate), angiographic status.

B. Data Cleaning

This step includes the removal of noisy data, identification of not available and not applicable data items and treatment of that data segments are required to be performed. In case the data contains unfiltered and irrelevant segments, then the results of the analysis will not unkind anything. Hence one of the crucial steps as removing dummy values from the dataset has performed. In the Heart disease dataset, there are many missing values in each attribute. Hence the missing values are replaced by the median values of the data set[13]. The Heart disease dataset Target variable consists of 'Yes' and 'No' as target labels. Converting strings to numbers, Assigning Yes as 1 and No as 0 respectively. The Heart disease UCI data set is already data wrangled. So while processing the data the algorithms can work smoothly.

C. Data Explore

Checking the relation between variables is one the crucial task. This shows how one variable is affecting the other variable. Data visualization plays a key role in this analysis. Without using orthodox methods we visualize data to comprehend. In this paper the major visualizations are done bar, box, heat map and ROC plots are used.

D. Attribute Selection

Selecting the required features is one of the most important tasks to get the best results and less time to train the model. Hence, to decrease the training time and evaluation time. And increase the accuracy of prediction. Hence to find the best features Information gain is used to find the features which plays major role in predicting the disease. After applying on Heart disease dataset the predicted output is almost equal to combining all the features together. The top featured were age, sex, fasting blood sugar>120,type of chest pain, target.

E. Model selection

a) Logistic Regression(LR): Logistic regression is widely used in machine learning to solve classification problem. Where sigmoid function is used to determine the predicted value with the help of threshold value. It is used to apply on categorical variables that are on the variables which can be classified.

b) Support Vector Machine (SVM): Vector Support Machine is a supervised classification system used primarily for classification of knowledge into various classes.It uses hyperplanes as a decisionmaking limit between different classes.SVM analyse etiquette learning information and then classifies data about what it has learned throughout the training.After plotting the data a line is drawn to separate the classes.

c) Naive bayes(NB): Naive bayes is based on theorem bayes with independent characteristics. The relationship between the likelihood of the hypothesis before the evidence is obtained and the probability after data learning is described in this theorem.

d) Random Forest(RF): Logistic Random forest is an ensemble grouping of decision models. Ensemble models incorporate the effects of various models.The most widely used predictive modelling and machine learning software is a flexible algorithm that can perform both regression and classification.The more number if decision trees in the forest the more the accuracy.

1) XG-Boost(XG): XG-Boost is a structured or tabular information algorithm that is used. The gradient-boosted engines are implemented for speed and performance. The speed of execution of XG Boost is very quick compared to other gradient boosting implementations. It manages structured data sets on problems of predictive modeling classification and regression. The next section is discussed on results.

F. Model testing with Accuracy Parameters

1) Precision: Precision is quantity of units correctly predicted as faulty. It is shown in table[2] for each classifier.

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

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

2) Recall: Recall is the ability of a classification model to identify all relevant instances. It is in table[2] for each classifier.

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

4) Reciever Operating Characteristics: This shows the true positive vs. the false positive in the model where threshold indicates positive identification .

IV. Results And Discussions

This work is implemented in Anaconda distributions. Anaconda is an open-source distribution of python for data science and machine learning applications. The comparative analysis between classification algorithms is done using performance measures like precision, accuracy, f1 score.

In this experiment, the classification algorithms LR, SVM, NB, RF, XG is applied on 14 features of the dataset on the following terms:

True Positive(TP): Predicted heart disease and actual result is same.
False Positive(TN): Predicted heart but they don't have heart disease.
*True Negative(FP):*Predicted No heart disease but they don't actually have the heart disease.
*False Negative(FN):*Predicted No heart disease but actually they have heart disease.

A. Distribution graphs of all attributes:

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

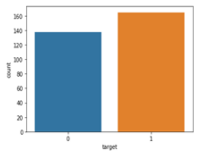

Fig. 2. Disease observations

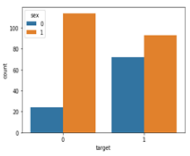

Fig. 3. Plot of Gender based observations

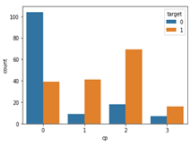

Fig. 4. Plot of Gender vs Chest pain

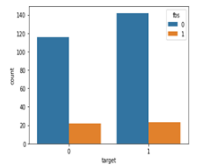

Fig. 5. Plot of Blood sugar impact

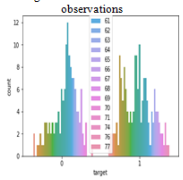

Fig. 6. Plot of Cholesterol impact

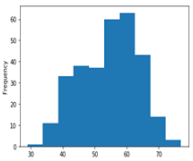

Fig 7. Area plot of age influence

1) Heat Map analysis: Logistic The analysis is done based on the correlation. The reciprocal relationship between all characteristics is used for evaluation. So the association between features can be determined using describe function. This correlation is called heat map. The heat map in figure 8 is observed using scale where the intensities of the scale shows how well the features are co-related.

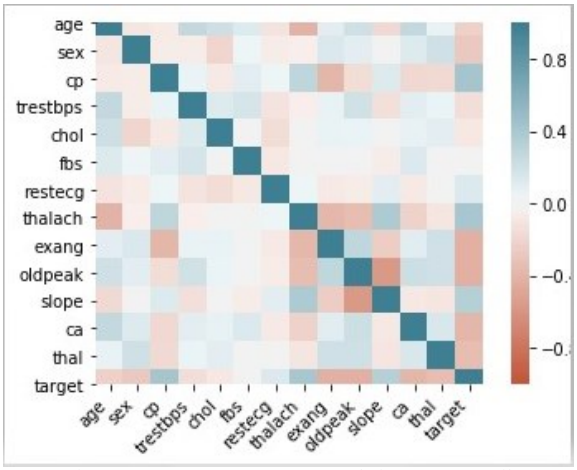

Fig. 8. Heat map to represent correlation among attributes

2) Box Plot: Shape of the distribution of age over chest pain among all levels is observed using this plot in figure 9.

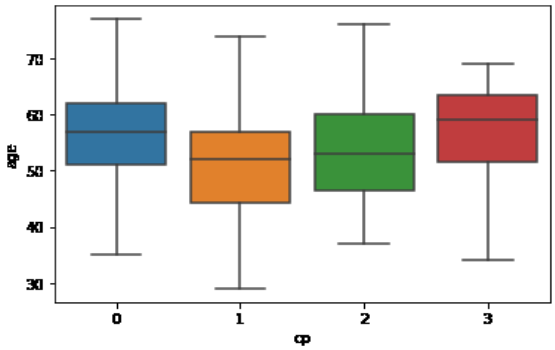

Fig. 9. Box plot on distribution of chest pain among various age groups

B. Confusion Matrix:

Confusion matrix displays correctly predicted and incorrectly predicted values by a classifier. In Table(1) it shows the relationship between features for various classification models.

TABLE(1).CONFUSION MATRIX OF VARIOUS 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]]**

C. Accuracy:

Accuracy is defined as disorderly predicted values and the total amount of times present in the dataset models.

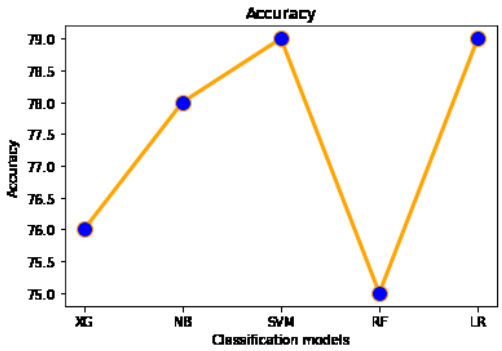

Fig. 10. Accuracy distributions among various classification models

From the figure 10 it can be concluded that after applying classification algorithms on the dataset. Highest accuracies are obtained for SVM(79%) and LR(79%) compares to XG(76%), NB(78%) and RF(75%).

1) Precision : Precision is quantity of units correctly predicted as faulty. It is the capability of a classification model to identify all pertinent instances. It is shown in table[2] for each classifier.

2) Recall : Recall is the ability of a classification model to identify all relevant instances. It is in table[2] for each classifier.

3) F1-score : It is a single bar that fuses recall and precision using harmonic mean.

Predictive accuracy of heart disease model comparison of various classifiers is shown in table 2.

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

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

D. Reciever 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.

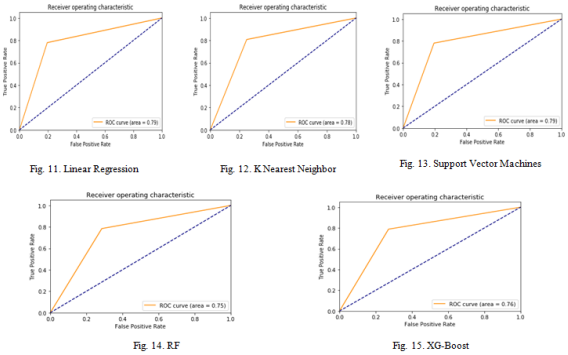

Fig. 11. Linear Regression

Fig. 12. K Nearest Neighbor

Fig. 13. Support Vector Machines

Fig. 14. RF

Fig. 15. XG-Boost

Fig. 11-15. ROC curves generated by different classifiers

V. CONCLUSION

Heart disease is now one of society's greatest fears. The chances of predicting heart disease manually on risk factors are difficult to assess. Machine learning techniques are however useful for estimation of heart disease from current data. The goal of this thesis is to observe the use of algorithms for analysis and prediction of heart disease in data mining classification. I conducted five classifications of heart disease prediction. The experiment of method has demonstrated that Support vector machine and logistic regression has produced predominant performance in terms of classification for heart disease dataset.

For the future, I am working on enhancement of the performance of prediction system accuracy by associating different classifier algorithms.

ACKNOWLEDGMENT

This research contribution is a part of undergraduate research and content development at institute of aeronautical engineering.

REFERENCES

[1] 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).

[2] 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)

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] Chaitanya Suvarnam, Abhishek Sali, Sakina Salmani, "Efficient heart disease prediction system using optimization technique", 2017 International Conference on Computing Methodologies and Communication (ICCMC).

[10] Seyedamin Pouriyeh,Sara Vahid,Giovanna Sannino,Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez,"A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease",2017 IEEE Symposium on Computers and Communications (ISCC).

[11] Cincy Raju, E Philipsy, Siji Chacko, L Padma Suresh, S Deepa Rajan,"A Survey on Predicting Heart Disease using Data Mining Techniques", 2018 Conference on Emerging Devices and Smart Systems (ICEDSS).

[12] Data set: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[13] Myneni, Madhu Bala, Y. Srividya, and Akhil Dandamudi. "Correlated Cluster-Based Imputation for Treatment of Missing Values." Proceedings of the First International Conference on Computational Intelligence and Informatics. Springer, Singapore, 2017.

[14] Madhu Bala Myneni, Rohit Dandamudi, "Harvesting railway passenger opinions on multi themes by using social graph clustering", Journal of Rail Transport Planning & Management,2019.

[15] Bennett, Tellen D., et al. "Data Science for Child Health." The Journal of pediatrics 208 (2019): 12-22.

[16] Sharma, Prerna, et al. "Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning." Artificial Intelligence in Medicine 102 (2020): 101752.

[17] Khari, M., Garg, A. K., Crespo, R. G., & Verdú, E. (2019). Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks. International Journal of Interactive Multimedia & Artificial Intelligence, 5(7).