Efficient and early prediction of heart diseases using Random Forest Classifier and other model refining techniques.

Prof. Sonali Lunawat, Sachin Patil, Pratik Manjare, Shubham Raut And Gaurav Dalal

Pimpri Chinchwad College of Engineering and Research, Ravet
Department of Computer Engineering

Abstract

Heart is a complex organ and heart disease symptoms are very common which can be easily mistaken by gastric problems or simple chest pain. Also assessing one's heart condition without an expert health professional is not possible. In today's technical advancements in fields of computer science and artificial intelligence, more specifically machine learning if clubbed with the health sector can produce a smart heart disease prediction system. Many attempts have already been made to work on such systems but at the end of the day what matters for these systems is their accuracy. In this research article we are focusing on bettering the accuracy of a heart disease prediction system using appropriate machine learning algorithms and further assessing and improving created model using newly emerged technologies provided by python libraries such model explanation using SHAP values, Eli5 etc. Also considering feature selection for less impure data to be provided to the model we have suggested data visualization techniques.

Keywords:-Random Forest Classifier, model explanation, SHAP values, Eli5, Heart Disease Prediction System

I. INTRODUCTION

Heart is an internal organ and hence people mostly tend to ignore heart health. The causes of heart diseases or cardiac arrest are the common factors that in our day to day life like unhealthy or excessive eating, improper sleep schedule, work stress, lack of exercise etc.People are not very well aware of heart disease until the damage is serious or it is very much late to save a person's life. Another problem is the symptoms of heart diseases are also common factors that can be seen in other minor diseases such as acidity, chest pain etc. so when we go to the doctor he/she cannot have a full overview of your health. During cardiac arrest or a heart attack each second is valuable to save a patient's life.[1]According to an article by WHO, Cardiovascular diseases are no. 1 cause of deaths globally, recording around 17.9 million deaths each year. It further states that individuals at risk of CVD may demonstrate raised blood pressure, glucose, and lipids as well as overweight and obesity. These factors can be predicted or determined using machine learning models, in particular random forest classifier. [2]Machine learning is an artificial intelligence technique whose goal is to repetitively learn from data, gain experiences and ultimately generate predictions based on their experience with minimal human intervention.
We are interested in Supervised learning algorithms. In Supervised machine learning algorithms, the machine(computer program) learns from the labelled data and on the basis of that it predicts the output of input data. In our case we will be providing a Cleveland dataset

to the machine learning model and gain predictions on the basis of its previous experiences.(This will be discussed in detail in upcoming sections).Also in order to classify and predict data we are proposing Random forest classifier.But we'll also see other tree based classification algorithms as proofs to why Random forest classifier suits best for a heart disease prediction system.

II. Proposed Algorithm

2.1 Dataset description:

Dataset which we are going to be using for our model is named as Cleveland dataset and it is publicly available on [3]UCI machine learning repository and it is contributed by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D.,

This dataset contains 76 attributes but we are going to refer to 14 attributes as mentioned below

Sr.no	Feature Name	Feature Description
1	age	Age of Person
2	sex	Gender of person
3	cp	Chest Pain Type
4	trestbps	Resting Blood Pressure
5	chol	Cholesterol Levels
6	fbs	Fasting Blood Sugar
7	restecg	Resting ECG
8	thalach	Maximum Heart Rate Achieved
9	exang	Exercise Induced Angina
10	oldpeak	ST Depression Induced By Exercise Relative to Rest
11	slope	The Slope Of The Peak Exercise ST Segment
12	ca	Number Of Major Vessels (0-3) Colored By Fluoroscopy
13	thal	Thalassemia Defect
14	num	The Predicted Attribute

These attributes are the critical factors which can be considered for a patient's heart health, and are also referred to by [4] in their research and are considered to be giving near to accurate results on various machine learning algorithms.

2.2 Brief introduction of Decision tree based machine learning algorithms:

2.2.1 CART:

The Classification and Regression Trees algorithm is based on the functioning of the basic boolean properties. The decisions made by the algorithm are a result of a comparison operation followed by a boolean operation. This helps the system to segregate the information which further makes it easier to analyse the processed data to generate the output. It uses the Ginn ratio in decision making modules. The algorithm uses a recursive model to segregate the data until no more separation can be done.There are 3 stages in this operation:

- Greedy splitting
- Stopping criterion
- Pruning the tree

i.Greedy splitting. The Recursive binary splitting technique is used to divide the input space. The cost is calculated before every division from where the minimum cost is considered for splitting. The evaluation is done in a greedy manner to get the optimum result. The Ginn ratio is used for the scrutiny of the 'purity' of the nodes. This index then is used for further classification.

ii.Stopping criterion.The recursive binary operation needs to stop at a certain point during the execution of the program. Like every other recursive algorithm the CART algo also has a base condition which when occurred prevents further recursion and switches the control out of the module.

iii.Pruning the tree.The efficiency of the decision tree is determined by the number of splits it makes. The more complex the tree the more splits it can make. The pruning technique used in CART is by checking the leaves of the tree to see if they can be further classified or not. The cost complexity phenomenon of the decision trees plays an important role in pruning of the tree. These leaves are only removed if they are increasing the overall cost of the tree.

CART when implemented on the dataset gave an AUC score of 0.87. Also other quality measures are as mentioned below:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	27
1	0.91	0.85	0.88	34
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

Fig 1. Classification report for CART

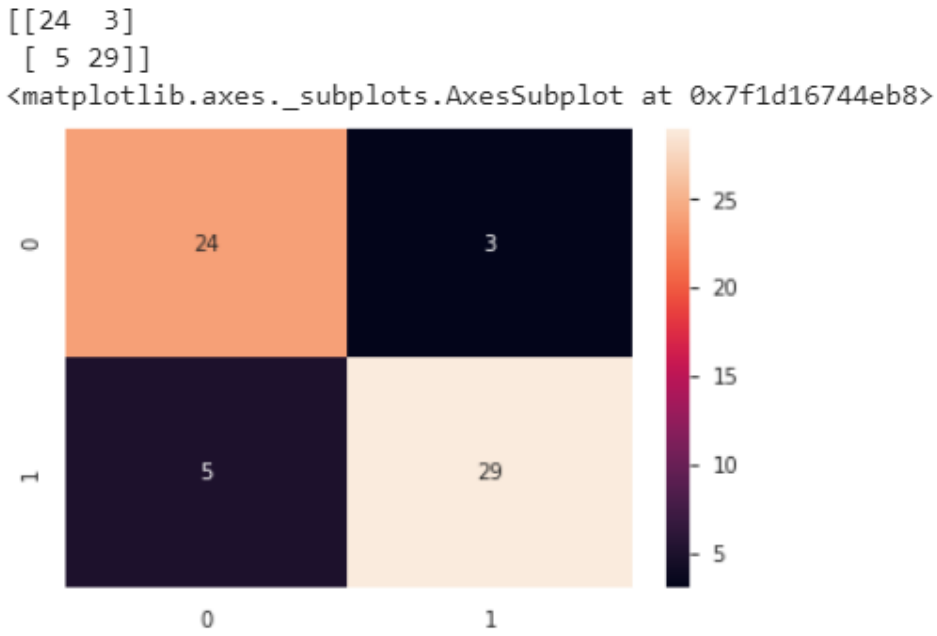

Fig 2.Confusion matrix for CART

Where,

- True Negative (TN) : when prediction is negative, and case is also negative.
- True Positive (TP) : when prediction is positive and case is also positive.
- False Negative(FN) : when a case is positive but predicted negative.
- False Positive(FP) : when a case is negative but predicted positive.

Precision - Accuracy of Positive predictions

Precision = TP / (TP + FP)

Accuracy = TN +TP / (TN + TP + FN + FP)

2.2.2 Random Forest Classifier:

The Random Forest Classifier is a decision making algorithm which uses several decision trees to produce the result.The Random forest classifier is based on supervised learning techniques. It can also be used for regression purposes along with classifying the data. The algorithm is basically a bunch of standard classifiers integrated to work together and provide almost accurate results given the context. It not only classifies the given data into trees but also predicts the results using a voting mechanism among the results of individual trees.

The classifier works in two phases accordingly,

- Collection and Analysis of data
- Choosing the optimum result

i.**Collection and Analysis of data.** In this phase the classifier collects the labelled data to further analyse using the decision tree algorithm.

In our case the data includes all the medical parameters entered by the user and the analysis is done by using a comparison model. This includes the implementation of the basic decision

tree model where, the given parameters are compared to the ideal parameters of a healthy human body with respect to their sex.

ii.**Choosing the optimum result.**The second as well the final phase of the classifier's primary job is to select the most suitable result for the given scenario. It does this by collecting the results of every decision tree.

In prediction of cardiovascular disease it also refers to the medical history of the patient if necessary.

Mathematical models:

i.Regression.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where N is number of data points,
fi is the value returned by the model and
yi is the actual value for data point i.

ii.Classifier.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

The algorithm vastly uses the Gini coefficient to decide the homogeneity of the nodes and leaves whenever a new element is added. It uses probability and class to determine originality and purity of the newly added node. In our case this helps to get a more detailed classification of the disease depending on the symptoms.

Also, Random Forest Classifier when implemented on the dataset gave the AUC score of 0.91 and other quality measures are as given below.

	precision	recall	f1-score	support
0	0.86	0.89	0.87	27
1	0.91	0.88	0.90	34
accuracy			0.89	61
macro avg	0.88	0.89	0.88	61
weighted avg	0.89	0.89	0.89	61

Fig. 3 Classification report for Random Forest Classifier

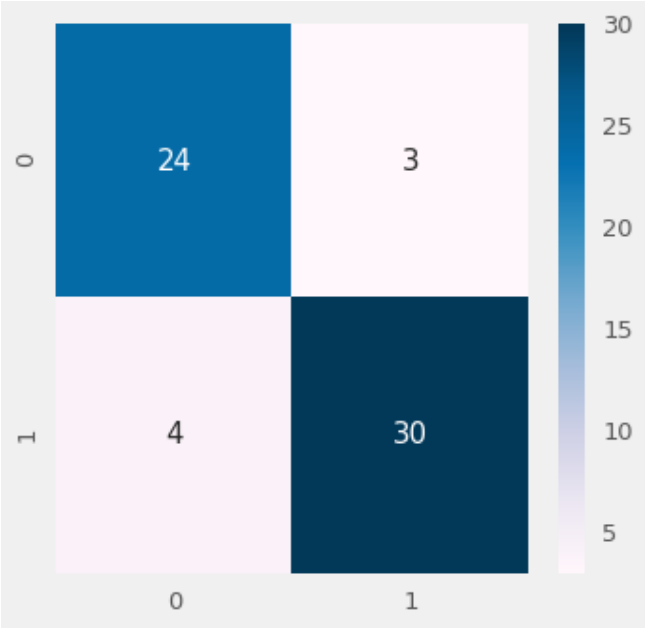

Fig. 4 Confusion matrix for Random Forest Classifier

Where,

- True Negative (TN) : when prediction is negative, and case is also negative.
- True Positive (TP) : when prediction is positive and case is also positive.
- False Negative(FN) : when a case is positive but predicted negative.
- False Positive(FP) : when a case is negative but predicted positive.

Precision - Accuracy of Positive predictions

Precision = TP / (TP + FP)

Accuracy = TN +TP / (TN + TP + FN + FP)

From the above analysis we've established one thing that Random forest Classifier predicts more accurately than other tree based machine learning classification algorithms.

2.3 Algorithm:

The proposed system is a machine learning model and the heart of the system will be Random forest classifier which will predict which factors are critical and should be tended to first by the health professional based on the readings given by the patient. To further make this model accessible to developers and understandable to the users we will be using the [5] ELI5(Explain me Like I'm 5) library provided which can be used in python.It is model explanation library provided by python which is helpful debug and assess a ML model. We have implemented a permutation importance operation on the model and then used the ELI5.show_weights which shows the weight of a particular feature selected by the algorithm. In other words it shows the impact on a model's prediction due to change in the value of a particular feature. This way we can find the dependencies of a model being biased to a particular feature and it can be normalized based on that. Next we have implemented SHAP values which stands for SHapely Additive Explanations.This will help us generate the pretty outputs for the users so that they can classify which feature has or may have a major impact for a patient's heart condition. One such sample is explained in the results section.

III.Result

The results are generated in format of a initjs SHAP values horizontal bar graph which basically shows the impact of a particular health factor responsible for the patient's heart condition so that he/she can be tended by the doctor in time before the damage is not major.Following is the representation of an output generated for a patient.

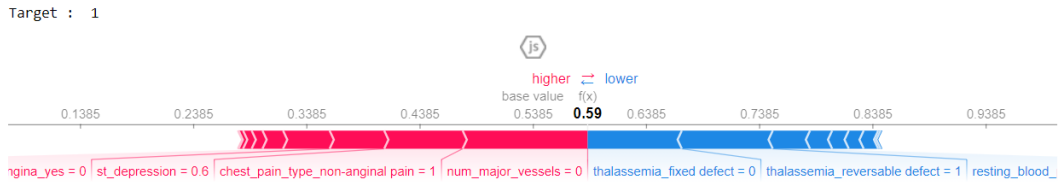

Fig. 5 Output for patient with heart disease

As we can see the model predicted the person is having a heart disease or has a high chance of having a heart disease and the SHAP vertical bar graph explains which health factors have high values and should be considered by the health professionals.

Similarly a result for the person with prediction of not having a heart disease is given below:

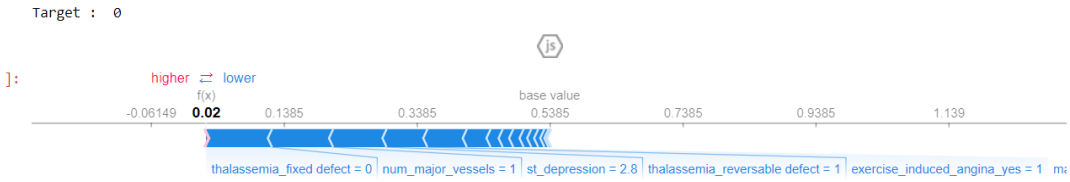

Fig.6 Output for patient without heart disease

IV. CONCLUTION

Hence we have implemented an efficient heart disease analysis model using random forest classifier and also attained an accuracy of 91%. i.e. the results predicted by the model are 91 times accurate in each 100 cases.Also we have implemented technologies like SHAP values and ELI5 to make the model simple to understand for developers as well as users. This model has wide applications for eg. It can be implemented on wearable smart watches to track the heart data and analyse it on the go. A mobile application can also be developed for the same. This patient data can also help doctors to better understand the condition of the patient.

References

1. WHO:**https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)**
2. El Naqa I., Murphy M.J. (2015) **What Is Machine Learning?.** In: El Naqa I., Li R., Murphy M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1
3. Cleveland dataset: **https://archive.ics.uci.edu/ml/datasets/heart+disease**
4. **Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques** C. Beulah Christalin Latha, S. Carolin Jeeva
5. ELI5: **https://eli5.readthedocs.io/en/latest/**