Heart Disease prediction using machine learning Approaches

Eaishawarya Malviya[1], S Md. S Askari[2*], Bhaskar Jyoti Chutia[3], Satish Kr. Das[4]

[1,2,3,4] Department of Computer Science and Engineering,
Rajiv Gandhi University, Doimukh, Arunachal Pradesh, India-791112
[1]aishwarmalviya8@gmail.com, [2]sikdar.askari@rgu.ac.in, [3]bhaskar.chutia@rgu.ac.in, [4]satish.das@rgu.ac.in

Abstract: In simple terms heart disease can be described as the diseases that affect heart or blood vessels. According to WHO, the 6th most common cause of deaths is either heart or lung disease, excluding neonatal diseases. Heart disease are the leading cause of death globally. Each year 17.9 million death occurs, that is one death every second. One third of these deaths occur below the age of 70. In this research we try to predict weather a person has a heart disease or not (heart disease classification), based on different parameters in the data present in the "Cleveland dataset of heart disease ", which can be found in UCI machine learning repository. For this purpose, we are using machine learning (ML) techniques and different machine learning algorithms present in Sci-kit learn library. We consider this as a binary classification problem, where we try to predict the target variable based on the features available in the dataset.

Keywords: Machine learning; heart disease classification; binary classification;

1. Introduction

Heart disease is a critical health problem and accurate diagnosis of it can help in saving many lives throughout the world. Several different parameters and health conditions of patients come into play in the correct diagnosis of heart disease. Accurate and in time diagnosis of it, will not only save lives, but also save money and time of the patient, doctor and hospitals.

According to Mayo clinic, "heart disease describes a range of conditions that affects your heart. Heart disease include, blood vessel disease, heart rhythm problems, heart defects (congenital heart defects), heart valve disease, disease of the heart muscle, heart infection".[1]

Heart disease (HD) is the most important cause of mortality in developed as well as in developing countries. Therefore, improvements and rationalization of diagnostic procedures and treatment of HD are necessary. The usual procedure in HD diagnosis consists of four diagnostic levels, which contain evaluation of signs and symptoms of the disease and ECG at rest, sequential ECG testing during a controlled exercise and coronary angiography as a final test. Because suggestibility is possible, the results of each step are interpreted individually and only the results of the highest step are taken into consideration. The total amount of data available for each patient is too large to be efficiently and objectively evaluated by the clinicians.

The goal of a rational diagnostic algorithm is to establish the conclusive diagnosis of HD and to plan the most appropriate management of the disease using only the necessary diagnostic steps. This can be achieved by evaluating all the information collected by different diagnostic methods according to their importance and diagnostic value.

The performance of a diagnostic method is usually described as classification accuracy, precision and recall.

Here, accuracy is calculated by dividing the sum of true positives and false positives with total number of predictions made by the diagnostic method. Similarly, precision is calculated by dividing true positives with the sum of true positives and false positives and recall is calculated by dividing true positives with sum of true positives and false negatives. The true positives are all patients with the disease and positive test result, whereas the true negatives are all patients without the disease and negative test result. False positive stands for all the patients without disease but positive test result and false negatives are all the patients with disease but negative test result.

*Corresponding Author: S Md S Askari
Email: sikdar.askari@rgu.ac.in

The aim of this paper is to find a good diagnostic method for heart disease, based on the patient's clinical parameters. For this we are using machine learning. We are taking advantage of the classification algorithms of machine learning to predict the result, as this is a supervised classificationproblem.

We have labelled data with the predicted values which we are using here to predict the outcome, the goal here is to compare different machine learning algorithms and find an optimized algorithm that can accurately diagnose heart disease. With using machine learning we can take advantage of the computing power of modern computers and state of the art algorithms to help us with correct diagnosis of the disease.
Machine learning can be classified into two major types, which are supervised and unsupervised learning. In supervised learning we have structured, labelled data where we know the outcome of the samples in the training set. Whereas in unsupervised learning, training data can be structured or unstructured and also there's no label, which means that we don't know the outcome of samples.
Supervised learning can also be further classified into two types, regression and classification learning. In Regression, the output variable must be of continuous nature or real value.
In Classification, the output variable must be a discrete value. The task of classification algorithm is to map the input value with the discreet output value. In classification, we try to find decision boundary which can divide the dataset into different classes. Classification algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.

Classification problems can also be further divided into binary classification and multi-class classification. Binary classification happens when we have only two discreet classes as our output variable and multi-class classification is when we have multiple classes to predict or when we have multiple discreet values as our output variable. From multiple here we mean more than two.

Here, in this heart disease diagnosis problem, the output variable, at basic level, can take only two values, which are weather a patient or a person have heart disease or not. That's why we have considered this as a binary classification problem. There are many machine learning algorithms that can be used for binary classification. Here, based on the data we have used logistic regression, K-neighbor classifier and random forest classifier algorithm. We have compared the results obtained by these algorithms on accuracy, precision and recall parameter. For our problem we found that the logistic regression algorithm gave the most accurate results.

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common Logistic regression model can be a model with binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best intoa category. For problems related to cyber security, such as attack detection, logistic regression is a useful analytic technique.[2] The best way to think about logistic regression is that it is a linear regression but for classification problems. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1, whereas linear regression range is continuous [3]. Inaddition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables.
Here, we used logistic regression model and trained it with our data. We used various techniques to improve its result, based on how well this model performs on unseen testdata.

2. Related works

In a paper published by IEE, titled, 'comparative study of heart disease classification' [4], the authors have compared six machine learning algorithms by training them on heart disease data in Matlab© environment and WEKA©. In this paper, Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM, Decision Tree and Ensemble Subspace Discriminant machine learning approaches are used for classifying the heart disease.
In another paper titled, 'Heart disease classification using optimized fuzzy rule-based algorithm' [5], the authors have introduced a technique named as RBFL prediction algorithm. Here, the overall process of the RBFL prediction algorithm is divided into two main steps, such as 1) feature reduction using LPP algorithm, and 2) heart disease classification by means of rule based fuzzy classifier. Initially, LPP algorithm is employed to recognize the related attributes and then fuzzy rules are produced from the FFBAT algorithm. Next, the fuzzy system is designed with the help of designed fuzzy rules and membership functions so that classification can be carried out within the fuzzy system designed. At last, the experimentation is performed by means of publicly available UCI datasets, i.e., Cleveland, Hungarian, Swiderland datasets. The experimentation result proves that the RBFL prediction algorithm outperformed the existing approach by attaining the accuracy of 76.51%.

In another interesting paper tilted, 'Analysis and classification of heart disease using heart beat features and machine learning algorithms.' [6] The authors have proposed an ECG (Electrocardiogram) classification approach using machine learning based on several ECG features, for the classification of heart disease. This proposed approach is implemented using ML-libs and Scala language on Apache Spark framework; MLlib is Apache Spark's scalable machine learning library. They have proposed an efficient approach to classify ECG signals with high accuracy. Also , Sellapan et. al[13] developed an Intelligent Heart Disease Prediction System to predict the heart disease using three classifiers Decision Tree, Naïve Bayes and Neural Networks. Naïve Bayes performed with good prediction probability of 96.6%. Also, 13 attributes were used for prediction.

And, Carlos (2006) implemented efficient search for diagnosis of heart disease comparing association rules with decision trees.

In another paper titled, 'Heart disease diagnosis and prediction using machine learningand data mining techniques: a review' [7] , the authors have summarized some of the researches on predicting heart diseases using data mining techniques, they have analyzed various combinations of mining algorithms used and concluded which technique(s) are effective and efficient. Also, some future directions on prediction systems have been addressed in this paper.
In another paper, titled, 'Heart disease identification method using machine learning classification in E-healthcare' [8], the authors have proposed an efficient and accurate system to diagnose heart disease based on machine learning techniques. They have developed a system based on classification algorithms including Support vector machine, Logistic regression, artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree. In this paper, standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. They have also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are usedfor features selection to increase the classification accuracy and reduce the execution time of classification system. They have used, the leave one subject out cross-validation method for learning the best practices of model assessment and for hyper parameter tuning. Here, the performances of the classifiers have been checked on selected features as selected by features selection algorithms. Their experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. The suggested diagnosis system (FCMIM-SVM), here, achieved good accuracy as comparedto previously proposed methods.

Problem statement

Predicting heart disease in a patient, based on his/her health condition and clinical parameters, using binary classification techniques in machine learning.

3. Data

The dataset used here, is a subset of the data originally released by the UCI machine learning repository. The dataset [8] used have 13 feature variables and one target variable.

List of variables (attributes) present in the dataset

There are 14 columns in the dataset: age: age in years

sex: gender1 =
male

0 = female

cp: chest pain type Value 0:
typical angina Value 1: atypical
angina
Value 2: non-anginal painValue
3: asymptomatic
trestbps: resting blood pressure (in mm Hg on admission to the hospital)chol: serum
cholestoral in mg/dl
bs: (fasting blood sugar > 120 mg/dl)1 = true;

0 = false

restecg: resting electrocardiographic resultsValue 0:
normal
Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation ordepression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by establishedcriteria
thalach: maximum heart rate achievedexang:
exercise induced angina

1 = yes

0 = no

oldpeak = ST depression induced by exercise relative to restslope: the slope
of the peak exercise ST segment

Value 0: upsloping

Value 1: flat

Value 2: downsloping

ca: number of major vessels (0-3) colored by flourosopythal:
0 = error (in the original dataset 0 maps to NaN's)1 = fixed
defect

2 = normal

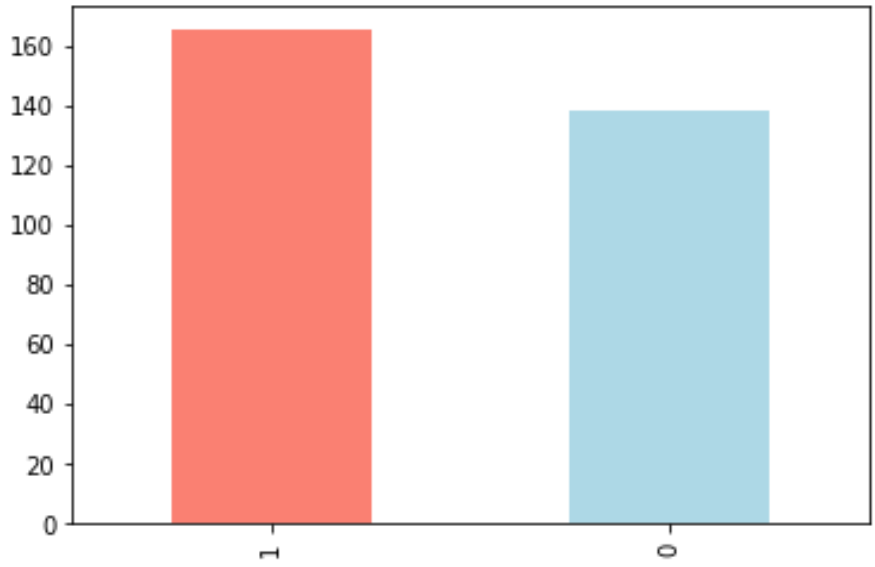
= reversable defect
target (the lable):
0 = no disease,
1=disease

These are all 13 features present in the dataset used. All the definitions are provided by the author of the dataset (subset of the data used) and it can be found on kaggle.com[9].

We also have one target variable (i.e. num.) that takes two values, 0 when heart disease is absent and 1 when it is present.
As this is a binary classification problem, we have used all the feature variables to predict our target variable.

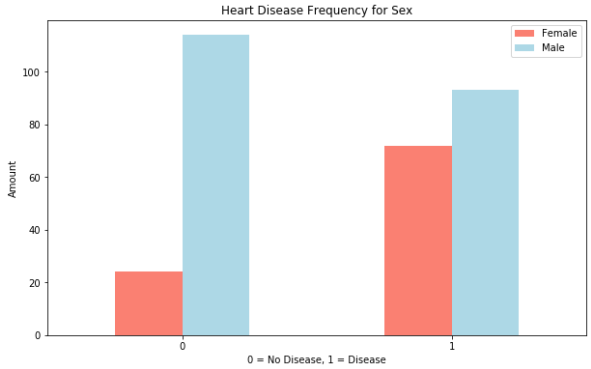
4. Exploring the data (EDA)

On close inspection, we found that there is 165 samples of patient data where heart disease is present and 138 samples where it is absent, in the dataset. This makes the total number of samples to be 303 (165+138).



Bar graph representing the value count of samples in the dataset based on heart disease diagnosis (1 stands for disease presence, 0 for absence)

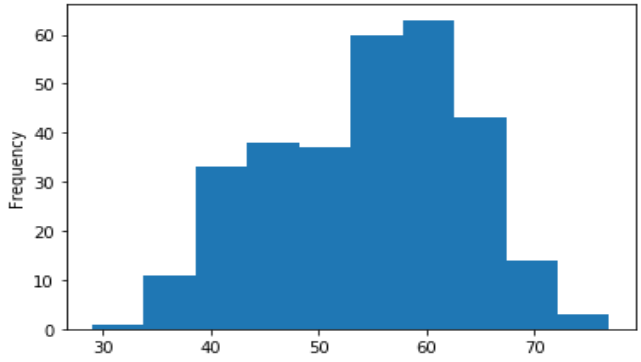
We found that there is 207 male samples and 96 females in our study. Among these about 75% of female present in the dataset had disease and about 44.9% of male presentare being diagnosed with disease. On combining age and heart rate parameter (maximum heart rate) and comparing it with our target variable we found that the younger patients had higher heart rate and older patients had it lower, in our dataset



Bar graph showing presence of disease based on the number of male and female samples in the dataset

Scafler plot comparing age and max heart rate, with heart disease presence

We then looked at the age distribution in the dataset.

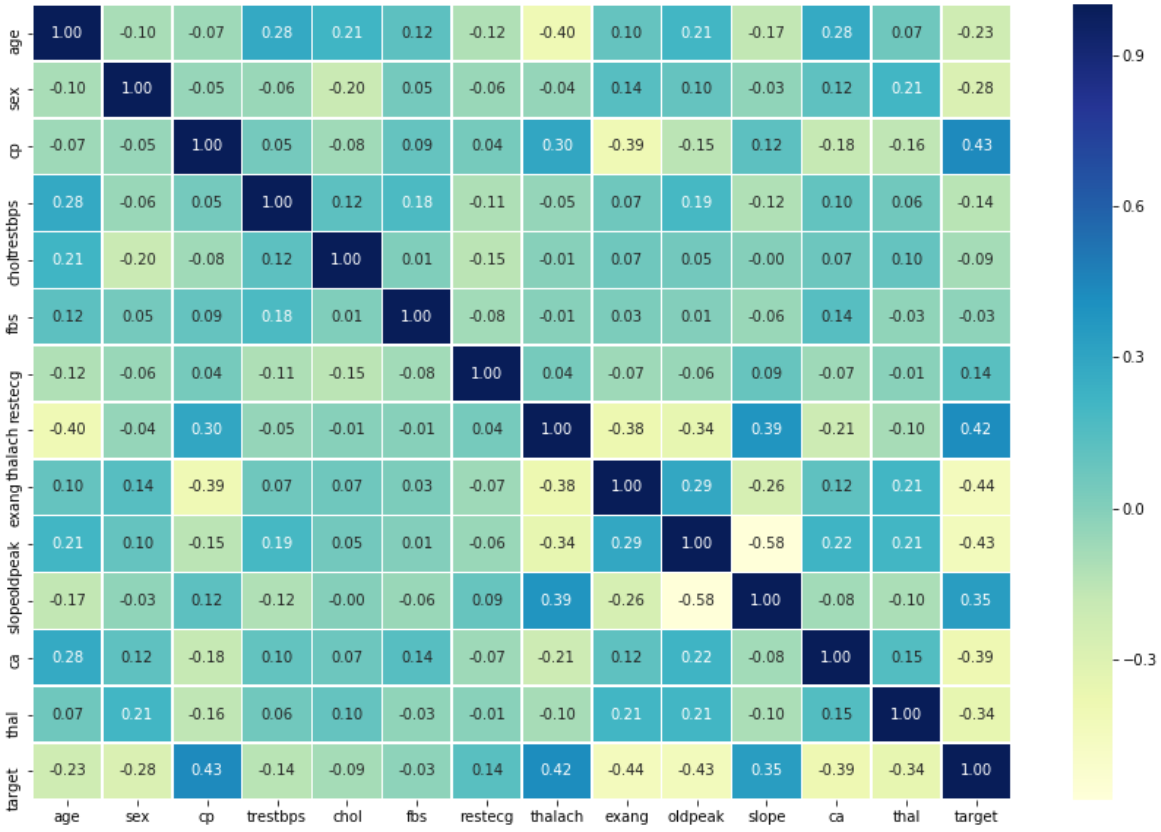


Histogram showing the age distribution in the dataset

Here we realized that the dataset had most number of samples of patients who were in 55-65 years age group.

Also there was no sample of patient under the age of 25 present in the dataset.

We further explored heart disease frequency, chest pain type and other variables. We plotted a correlation matrix to examine relationships and correlations among variables inthe dataset.



Correlation matrix representing linear relationships among variables

5. Methodology

We divided the dataset into two parts. We used 80% of the data for training the model and then we checked or tested the accuracy of the model on the remaining 20% of the dataset. We also divided the target variable from the remaining feature variables and stored it in a separate variable. After this we used this training data (first 80% of the dataset), to train the logistic regression , K neighbors classifier and Random forest classifier model. We used python's scikit learn library to directly implement or train these models in our work environment. After fitting these models with data, we evaluated their accuracy scores. Also we have used 5-fold cross validation to get unbiased predictions.

Then we tried to maximize correct predictions, by training the model with randomized search cv and grid search cv. Then, we compared accuracies obtained by these models to check which model performed better, on unseen test data.

6. Algorithms used

6.1 Logistic regression

Logistic regression is a supervised machine learning algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. Logistic regression is usually used for binary classification problem.[10]

Logistic regression equation is quite similar to linear regression model. Now, consider we have a model with one predictor "x" and one Bernoulli response variable "ŷ" andp is the probability of ŷ=1. Then linear equation can be written as:

$$p = b0+b1x \quad \text{--------> eq 1}$$

The right-hand side of the equation (b0+b1x) is a linear equation and can hold values that exceed the range (0, 1). But we know probability will always be in the range of (0, 1).

To overcome that, we predict odds instead of probability.

Odds: The ratio of the probability of an event occurring to the probability of an event not occurring.

Odds = p / (1-p).

The equation 1 can be re-written as: $P / (1-p) = b0+b1x$ --------> eq 2

Now, odds can only be a positive value, to tackle the negative numbers, we predict the logarithm of odds. Log of odds = ln(p/(1-p)).

Thus, the equation 2 can be re-written as: $\ln(p/(1-p)) = b0+b1x$ --------> eq 3

To recover p from equation 3, we apply exponential on both sides.

$$\exp(\ln(p/(1-p))) = \exp(b_0+b_1x)$$

$$e^{\ln(p/(1-p))} = e^{(b0+b1x)}$$

From the inverse rule of logarithms,

Now,

Dividing numerator and denominator by e(b0+b1x) on the right-hand side

$$p = 1 / (1 + e^{-(b0+b1x)})$$

Similarly, the equation for a logistic model with 'n' predictors is as below:

$$p = 1/ (1 + e^{-(b0+b1x1+b2x2+b3x3+ \ldots\ldots+bnxn)})$$

We get a sigmoid function; this sigmoid function helps to squeeze the output in the 0 to 1 range in logistic regression.

The sigmoid function is useful to map any predicted values of probabilities into another value between 0 and 1.

6.2 K nearest neighbors

The KNN algorithm assumes that similar things exist in close proximity. In otherwords, similar things are near to each other. [11]

Let's look at how KNN works:

KNN algorithm

Load the data

Initialize K to your chosen number of neighborsFor each

example in the data

Calculate the distance between the query example and the current example fromthe data.

Add the distance and the index of the example to an ordered collection

Sort the ordered collection of distances and indices from smallest to largest (inascending order) by the distances

Pick the first K entries from the sorted collectionGet the labels

of the selected K entries.

If regression, return the mean of the K labels

If classification, return the mode of the K labelsWe have

KNN for classification.

6.3 Random forest classifier

Random forest is a *Supervised Machine Learning Algorithm* that is *used widely in Classification and Regression problems*. It builds decision trees on different samples andtakes their majority vote for classification and average in case of regression.[12]

One of the most important features of the Random Forest Algorithm is that it canhandle the data set containing *continuous variables* as in the case of regression
and *categorical variables* as in the case of classification. It performs better results forclassification problems.

Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data sethaving k number of records.

Step 2: Individual decision trees are constructed for each sample.Step 3: Each

decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging*** forClassification and regression respectively.

A set or collection of decision trees is known as forest.

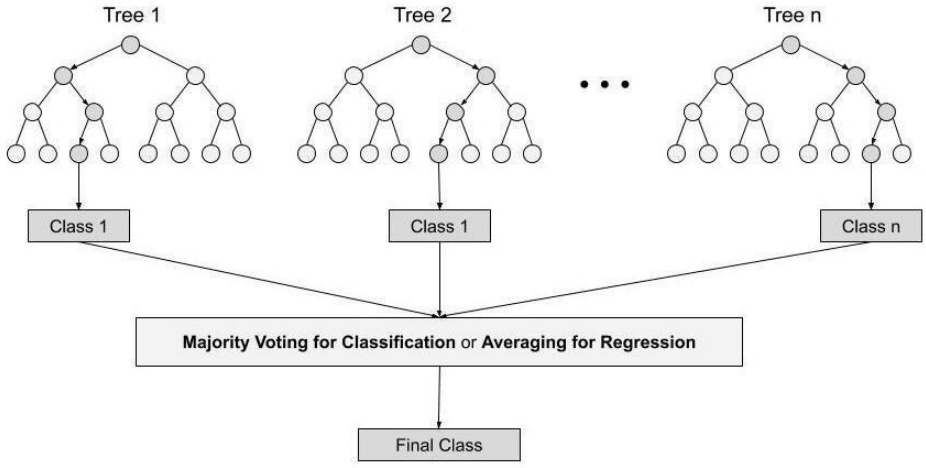
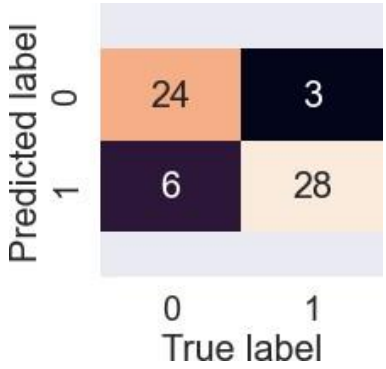


Figure representing working of random forest algorithm

Also I've used 5-fold cross validation. In 5-fold cross validation the data set is split into 5 folds (or parts). In the first iteration, the first fold is used to test the model and the restare used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated for all the 5 folds. With the help of cross validated training the model can generalize (understand) the databetter and predict the outcomes with less bias.

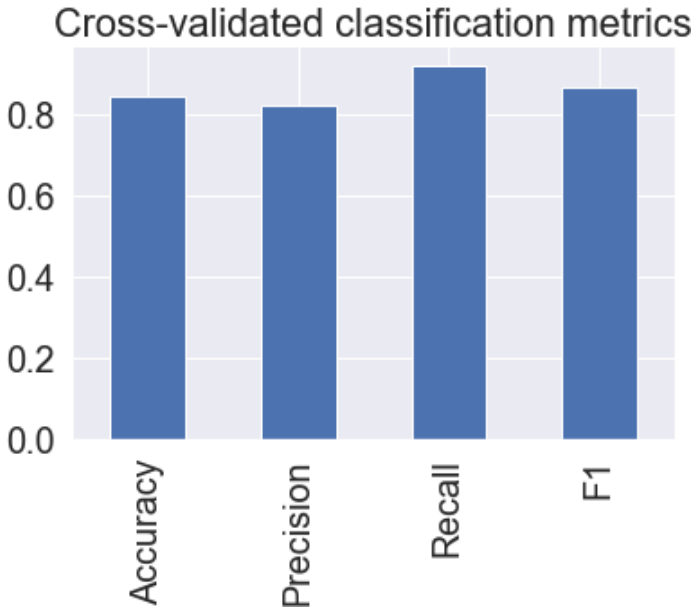
7. Result

We found that the model that was able to generalize the dataset and perform relativelybetter on this classification problem was logistic regression model. So we further experimented on logistic regression model and tried to boost its prediction accuracy.

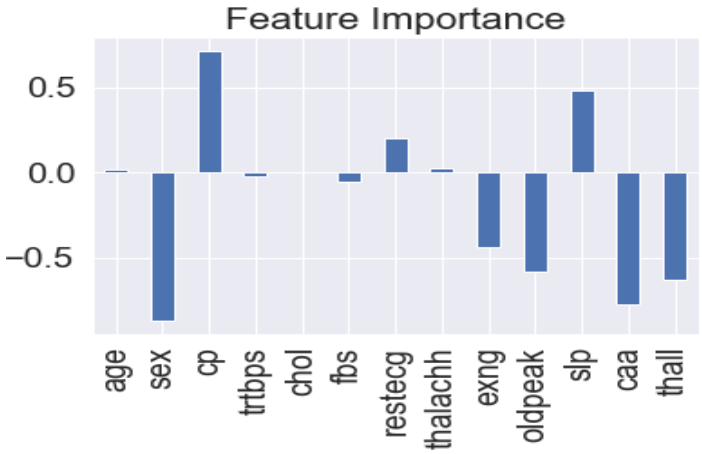


Confusion matrix representing correct predictions of target variables (true positive and false positive) by the logistic regression model

After cross validation, the precision, recall, accuracy, and f1score, of the model can beseen below.



We also looked at, to what extent a feature contributes in the predicted output made bythe model. A graph representing this can be seen below.



We can see that the chest pain positively impacted the predictions while the gender hadnegative impact on the outcome of the model. Based on this dataset serum cholesterol levels had negligible impact on heart disease predictions.

8. Conclusion

We were able to make heart disease predictions based on certain essential clinical parameters of the patient, using machine learning techniques like cross-validation andby using machine learning algorithm like logistic regression.

References

[1] Heart disease, Mayo clinic, published on 09 February 2021, accessed on 2 April 2022, (https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118)

[2] Thomas W. Edgar, David O. Manz, Chapter 4 - Exploratory Study, Editor(s): Thomas W. Edgar, David O. Manz, Research Methods for Cyber Security, Syngress, 2017, Pages 95-130, ISBN 9780128053492, https://doi.org/10.1016/B978-0-12-805349-2.00004-2.,(https://www.sciencedirect.com/science/article/pii/B9780128053492000042) ,10 April 2022

[3] Logistic regression vs linear regression, Javatpoint.com, Accessed on 2 May 2022, (https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning#:~:text=of%20independent%20variables.-,Logistic%20Regression%20is%20used%20to%20predict%20the%20categorical%20dependent%20variable,the%20value%20of%20continuous%20variables.)

[4] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2017, pp. 1-4, doi: 10.1109/EBBT.2017.7956761. (https://ieeexplore.ieee.org/abstract/document/7956761) 4 May 2022

[5] Gadekallu, Thippa & Khare, Neelu. (2018). Heart disease classification system using optimised fuzzyrule based algorithm. International Journal of Biomedical Engineering and Technology. 27. 183.10.1504/IJBET.2018.094122. (https://www.researchgate.net/publication/327105481_Heart_disease_classification_system_using_optimised_fuzzy_rule_based_algorithm) 5 May 2022

[6] Alarsan, Fajr Ibrahem, and Mamoon Younes. "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms." *Journal of big data* 6.1 (2019): 1-15. [7]https://www.researchgate.net/publication/319393368_Heart_Disease_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_Review,(https://journalobigdata.springeropen.com/articles/10.1186/s40537-019-0244-x#citeas) 6 May 2022

[8] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identificationmethod using machine learning classification in e-healthcare. *IEEE Access*, *8*, 107562-107582. (https://www.researchgate.net/publication/342058675_Heart_Disease_Identification_Method_Using_Machine_Learning_Classification_in_E-Healthcare) 7 May 2022

[9]Heart Disease dataset, Kaggle.com, A subset of UCI heart disease dataset taken from cleveland dataset, 2 April 2022, (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

[10]An Introduction to logistic regression, Analyticsvidhya.com, Harika Bonthu, 11 July 2021, accessed: 4 May 2022(https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/#:~:text=Logistic%20Regression%20is%20a%20%E2%80%9CSupervised,used%20for%20Binary%20classification%20problems.)

[11] Machine learning basics with k-nearest neighbor algorithm, Onel Harrison, 11 September 2018, towardsdatascinece.com,(https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761) 4 May 2022

[12]Understanding random forest, analyticsvidhya.com, Sruthi ER, June 17 2021(https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Steps%20involved%20in%20random%20forest,tree%20will%20generate%20an%20output) 6 May 2022

.[13] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS International Conference on Computer Systems and Applications*, Doha, Qatar, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.