Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v30i18.3193

PREDICTION FOR HEART DISEASE USING DIVERSE MACHINE LEARNING APPROACHES AND TECHNIQUES

Naveed Ahmed[1*], Dr. Amina Arif[2], Intakhab Alam Qadri[3], Dr. Syed Raffay Ali Gillani[4], Shahid Iqbal Rai[5], Asma Asif[6], Husnain Saleem[7], Samra Sana[8]

[1]*Shaikh Ayaz University, Shikarpur Department of Computer Science.
[2]Faculty of Science and Technology, Department of Biochemistry, University of Central Punjab, Lahore.
[3]Shenzhen University, China, College of Computer Science and Software Engineering.
[4]Assistant Professor, Allama Iqbal medical College, Jinnah Hospital Lahore, Cardiovascular Department.
[5]Lecturer, Department of Computer Science, The University of Lahore.
[6]Comsats University Islamabad, Department of Computer Science.
[7]Lecturer, Department of Computer Science, Sub Campus Gomal University Tank, KPK, Pakistan.
[8]MS Mathematics, Islamabad, Air University Islamabad

***Corresponding Author:** Naveed Ahmed
*Shaikh Ayaz University, Shikarpur Department of Computer Science.

Abstract
The heart is the most important organ of the human body. There are two main functions of the heart firstly, to collect blood from tissues of the body and pump it to the lungs, and second, to collect blood from the lungs and pump it to all tissues of the body[1].Many people have died because of heart disease. Therefore, it is important to predict that disease at the right time. By using machine learning and data mining techniques diseases can easily be predicted and diagnosed. Wearable sensor devices also can be used in the Internet of Things, and streaming systems[2]. The main objective of this research is to analyze core machine learning algorithms for heart disease prediction, for instance, SVM (Support Vector Machine), and Logistic Regression. K-Nearest Neighbors Algorithm, Decision, and Random Forest. Our Trained model for Logistic Regression showed 83% accuracy prediction result whereas the Decision Treealgorithm showed only70%which is 13% less than Logistic Regression. The result of the K-Nearest Neighbors Algorithmis 84% whereas SVM showed90% accuracy prediction result which is quite better than previously used algorithms. Then Random Forest showed 91% result which is a better result than all previously used algorithms i., eDT (Decision Tree), RF (Random Forest) and K-Nearest Neighbors Algorithm, Python Programming jupyter Notebook which is excellent in code and data. Hlaudi Daniel Masethe*et al.*

Keywords: DT (Decision Tree),K-Nearest Neighbors Algorithm, Rf(Random Forest), SVM (Support Vector Machine).

1 Introduction

The heart is located at the center of the chest. It collects deoxygenated blood from the whole body and takes this deoxygenated blood to affect the lungs then it is converted into the form of

oxygenated blood. The heart does not only circulate blood but also performs some extra major functions like nutrients are digestions transferred to all cells of the body. The main function of the heart comprises the interstitial fluid pumping from the blood into the extracellular space. There haveused different algorithms *like (LR) Logistic Regression, (SVM),(DT) Decision Tree,and (KNN) etc*., sum up the accuracy performance of the mentioned algorithms, and based based onprediction try to achieve one is best. This means the heart plays a very important role in our body. Some factors that lead to heart disease are given as under:

i) Coronary Artery Disease: refers to the formation of cholesterol plaque which causes the toughening or lessening of the coronary artery

ii) Cardiomyopathy: refers to illness of the heart muscle forsome reasons

iii)Angina: in which chest pain is caused due to having low blood flow to a part of the heart muscle.

iv)Cerebrovascular Disease: refers to the disease of the blood that resource blood to the brain.

v) Heart attack: it was cut off supply when permanent damage to the part of the heart muscle.

2 Machine Learning (ML)

Machine Learning is a most effective technology that consists of important basic terms i.e. Test and Train. The system can yield data for training and testing. These two terms which are training and testing must be functional to unlike types of algorithms as per requirements. Big data analytics can also be accomplished with ML, and this will mechanize the analytical model structure. The environment of ML allows the system to distinguish unseen methods from big data by applying algorithms iteratively with no need for explicit programs. Gopi Battinenia*et al.,* Machine learning is considered the best approach for identifying, predicting, and analyzing given problems without humans[7]. Salam Ismaeel *et al.* machine learning will be a suitable choice to create a model for them[8].

The above diseases are very dangerous and affect both women and men. There are many symptoms of heart disease Immensity, pressure, worry or pressing pain in the chest, Weakness and fatigue, Littleness of breath, Swelling or irregular pulse and Swelling, Anxiety, Cough, Aching, Burning, etc, after understanding heart disease and its symptoms it is necessary to control various factor that causes heart diseases.

Avoidable heart disease is Smoking.

- Less sleep.
- High BP.
- level of High cholesterol
- Diabetes.
- Less physical activity.
- Unhealthy diet habits.
- Mental stress and depression

3 Related Work

So far much excellent research has been carried out in terms of predicting heart disease by using ML and other techniques like data mining. Different datasets have been used to observe the results. All over the world, designing the model to estimate CVD diagnosis has been mobile research for the past few years. Detection of heart disease is very important because it is a very important organ of a human being. Treating at an early stage is very crucial for better treatment. So many approaches have been used to predict heart diseases for Cardiovascular diseases are one of them [1]. It is always essential to analyze heart-associated things for any diagnosis or prediction in contrast to heart disease. There are numerous domains like AI, ML, and data mining that contributed to this work[3]. The suggested model was not operative for main tumors due to unrelated then terminated attributes available in the datasets. The outcome uncovered that the main cross-over rate for the inherent algorithm is (60%) KNN, the learning displays increasing in accuracy to predict heart

disease[1], and the dataset has been created from the application of the UCI ML storehouse. Two datasets have been gained one comprising(1026) instances and fourteen (14) attributes and the other has 303 instances and fourteen (14)kinds of attributes. The resultant datasets hold 1329 instances and fourteen (14) attributes. There is a need for careful conduct or else it leads to death condition. The strictness of heart problems are categorized based on methods like the KNN algorithm, and (DT) Decision Tree [3]. The initial method for the prediction system is collecting data and determining to train and test data. And also, used 73% for training the datasets and 37% for testing datasets[3]. Classification algorithms and LR, DT, Random Forest (SVM), and Adaptive Advancing were used with the particular feature and a comparison between their prediction accuracy was calculated using Train and Test split method [4]. The data desired to be web-scraped from the UCI storehouse database and stored locally in a format that would be accessible [5]. Yamala Sandhya, For This type of research technique to expect the heart problems of people that yield the attributes like sex type, age type, blood pressure type, chest pain type, and sugar levels of the people. Above mentioned datasets Gender denotes 1 shows the Male (M) and 0 means the Female (F). Chest pain denotes 1 shows "typical of angina", 2 shows "typical of angina", 3 shows "non-angina pain" and it shows 4 "asymptomatic". In prediction worth 0 for "No" and 1 for "yes". Resulting in 70% for the training data process and30% data for testing purposes[6].

4 Methodology

This section describes step by step and implements all data in the latest Anaconda jupyter Notebook version 6.4.6.It gives complete detail about the implementation of different machine learning algorithms that give a complete picture of data science and data mining. After collecting data and passing the data for the preprocessing phase and able to excerpt the data that is needed for the prediction process. Because it reduces the size of data removes all unnecessary data and performs normally on given datasets.

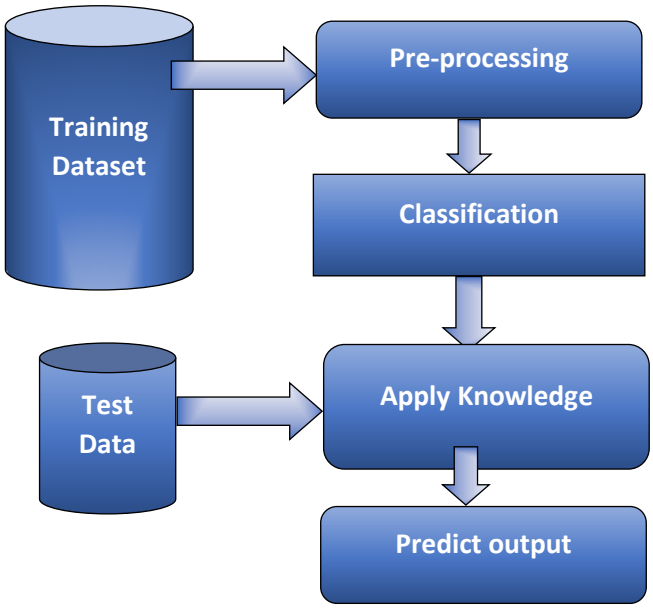


Fig 1: *Process of heart disease*

Data Collection

We have taken data from Kaggle source which is an online community of data scientists and machine learning practitioners and the best resource for data collection. And have implemented the best libraries in the jupyter not book to get results. All collected datasets are in CSV form. The following screenshots are given to show the starting and ending data frames of our datasets by using the train. head(). And have used thousands of rows and 15 columns

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Data Visualization

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Data visualization is a very useful method because a huge amount of text or numbers can not be understood by the humanbrain properly as charts and graphs can easily understand properly. And it is also a quick way of understanding. You can easily make changes smoothly. And therefore, datasets have been visualized so that we can identify each trend, etc. and it helps us to create a better model and distribute scaling.Following are the features that are used in the dataset.

Using the correlation feature this is done to find out if all the features are positively correlated or negatively correlated. Also used the Seaborn Library.

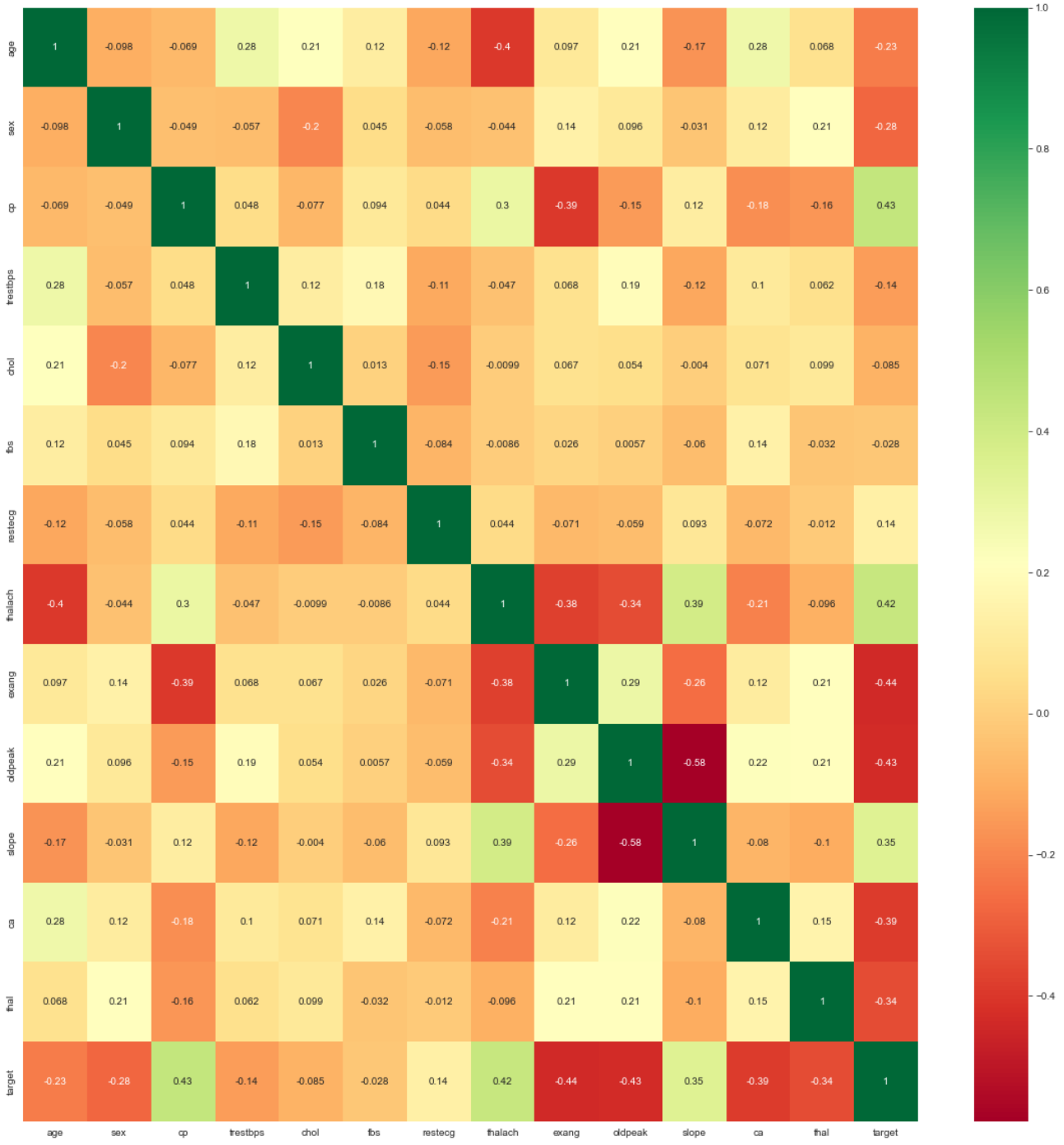


Fig 2: *histogram of correlated features*

Here target output is something cp has a positive correlation concerning the target output 0.43which is a symptom of heart disease, similarly, you have different features. Some have a negative correlation.

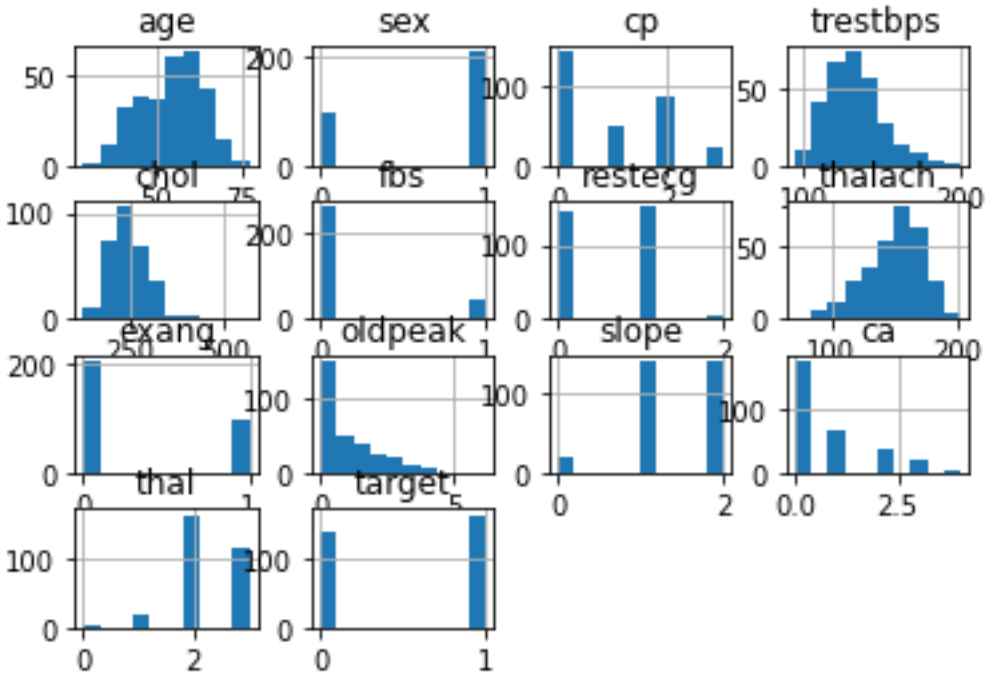


Fig3: *Histogram of each figure values*

Figure 3, represents a small look at every feature and how it is distributed and most of the features are standard in a normal distribution.

Data Preprocessing

The main part of understanding and finding out whether our datasets are valid or not taking a target value and some style mechanism and it is important to remember target output is that person having heart disease or not so that value is 0 or 1, zero means a person having not heart disease some around 140 and person having heart disease value 1 means 100. And it looks like totally balanced datasets.

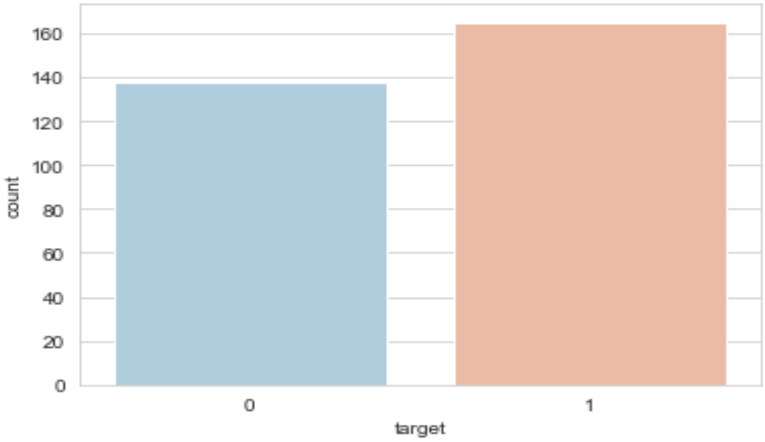


Fig4: *Target value of 0 and 1 by using counter plot*

4.1 Training Models

In this section, different types of machine learning algorithms have been highlighted that are widely used to predict many health diseases not only heart disease. It is categorized each algorithm properly and we have implemented every mentioned algorithm separately in jupyter notebook by taking two kinds of datasets and those datasets imported and got different performance results according to the training algorithm. The following are the algorithms that we have implemented throughout the research.

- **Decision Tree Algorithm**

A Decision Tree is known as (DT) and it is a supervised learning technique and is being used for classification and regressionproblems. It works on a tree-like structure. We used 14 columns and 302 rows.

0.7029702970297029

The following figure shows the data frame in the form of CSV after being imported intojupyter notebook used in the decision tree algorithm during implementation time.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig5: *data frame of heart disease prediction*

- **Logistic Regression**

It is considered a fundamental classification technique. It is a very fast algorithm comparatively unsophisticated, it is also considered a supervised classification algorithm means binary classification. It works on discrete values for given sets. For the betterment of the result, we put an accuracy score. We used alpha which is in the range of – 4 to 4 differentiated by 1 and ** (double star) showing exponential function. After that, we trained our data. Also a classifier, then compared y predict and y CV and stored in the score variable. By applying the append score and got theresults that are shown in Fig6a. The maximum up to 84.3% behind got till now out of 100%. And 88% is the highest accuracy recorded in the world.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
alpha = [10**x for x in range(-4,4,1)]
acc = []
for c in alpha:
  clf = LogisticRegression(penalty='l1',tol=0.1, C=c,max_iter=1000)
  clf.fit(X_trainS,y_train)
  y_pred = clf.predict(X_cvS)
  score = accuracy_score(y_pred,y_cv)
  acc.append(score)
  print(score)
```

```
0.8341880341880342
0.8341880341880342
0.8341880341880342
0.8427350427350427
0.8358974358974359
0.8358974358974359
0.8358974358974359
0.8376068376068376
```

Fig6a: *LR results*

First, predicted the results by using l1 now we predicted by using l2 which was the better result.

```
clf = LogisticRegression(penalty='l2',tol=0.0001, C=optimalC,max_iter=1000)
clf.fit(X_trainS,y_train)
y_pred = clf.predict(X_testS)
score = accuracy_score(y_pred,y_test)
print(score)
```

```
0.8415300546448088
```

Figure 6b: *LR results*

In the following datasets 16 attributes have been used which are shown in Figure 6 like male, age, education,etc.,there are only shown starting 5 rows of the dataset by mentioning the method named train. head ()

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Ten'
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	

Fig7: *starting 5 rows of dataset*

- **K-Nearest Neighbors Algorithm (KNN)**

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.0
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.3
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.6
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.0
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.0
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.0
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.0

Fig 8: *mean, median and min of every features of KNN*

It is a kind of supervised machine learning algorithm that can be implemented for both classifications as well as regression analytical problems. This algorithm is very easy to understand and its accuracy is very high. In this algorithm we

Have used different features like age, cp, treetops, etc. then using df. describe the () method and also found some mean, median, std, etc. that are shown in Figure 7.It shows only statistical details of our data frame. And got the result following after cross-validation when our K Neighbors = 12 and cv=10 means cross-validation.

0.8448387096774195

- **Support Vector Machine**

Support Vector Machine is considered a supervised machine learning algorithm that implements classification by discovering a hyperplane that distinguishes the classes of the plotted data points on an n-dimensional space [5]. Implemented the same datasets that have been implemented in KNN algorithms. By Separating the Feature and Target Matrix target feature is taken as axis=1. Then Split the dataset into a training set and a test set after splitting the 70% training and 30% test we got the following accuracy test score of SVM.

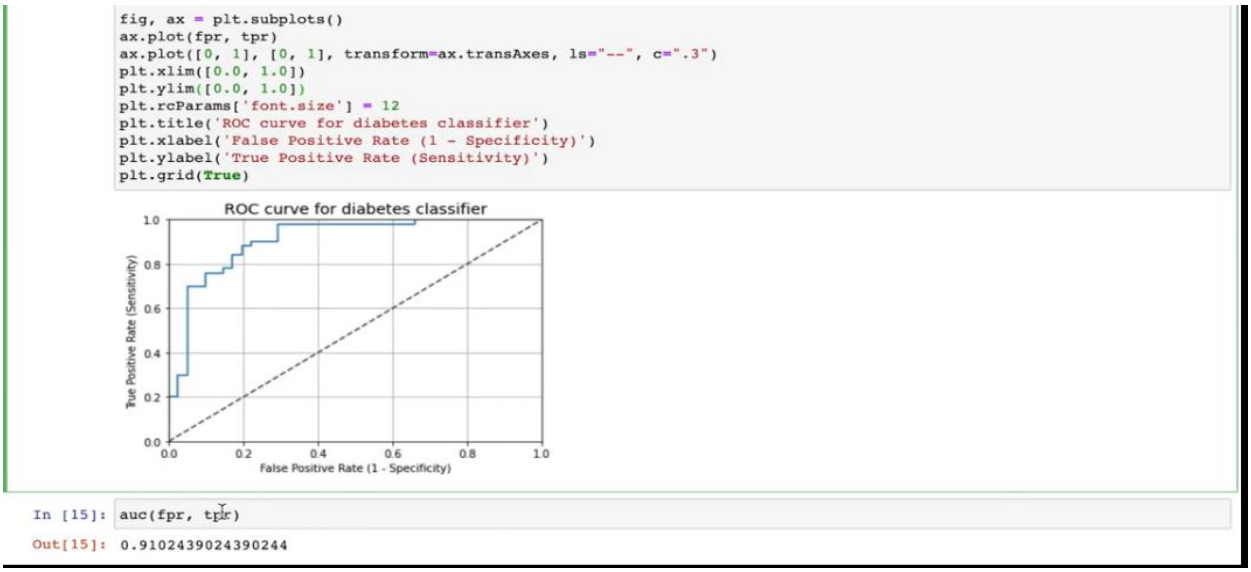
0.9010989010989011

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 9: *data frame of SVM*

- **Random Forest**

Random Forest is being used for classification and regression. An important feature of FR is that manage the dataset that consists of continuing variable in case of regression. Following figure 9 shows data frames. We used a discrete value that means 0 and 1. We spilled data for training and testing the test size is 30% and the random state 42% then trained the RF model by calling X train and y train in the end we predicted the accuracy score as we created an ROC curve that was used



For predictive, the model is performing very well with an accuracy score is 0.91.

Fig 10: *result of RF*

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 11: *data frame of RF*

5 Results

Now following table 1 shows a comparison of each algorithm. We used two kinds of datasets of heart disease and all datasets were imported into jupyter notebook by using Anaconda IDE. We have downloaded all datasets from the Kaggle website which is an excellent site for downloading any kind of dataset. First of all, we talk about the Decision Tree algorithm that works on the tree structure. By importing x_ train. shape method of 201 rows and 13 columns in x_ val. shape method we used 101 rows and 13 columns of our trained model decision tree classifier by

mentioning random state 10 after that printed the trained model resulting and got 70% accuracy score which isa little bit good. After getting the result of the DT file CSV file was imported and then the trained method for getting 16 attributes. In Logistic Regression, binary classification was implemented. For better scores, imported an accuracy score package and many machine learning hyperparameters to improve our output like radio volume return. Finally, got 83% for Logistic Regression quite better than DT. We imported the same dataset as we imported in the LR algorithm after importing packages and libraries and the by using KNN classifier that n_ neighbor =12 after that we used a cross_ val score that was x, y, cv=10 function resulting we got 84% accuracy score that little better than DT and LR algorithms. After getting the accuracy score of three algorithms decision tree, logistic regression, and KNN. imported the dataset for the SVM algorithm by separating the feature and target matrix we split the dataset into a training set and a test set means 70% for training and 30% for testing after both training and testing we created an SVM classifier for the fit and train model. In the end for accuracy, And got results from a trained model that predicted 90% which was an excellent result. A. Sankari Karthiga et *al.*, *the* decision is a very prominent classifier that requires no domain knowledge and it can handle high dimensional data [13].

Table 1. Accuracy Comparison of each algorithm

Decision Tree	Logistic Regression	K-Nearest Neighbors Algorithm (KNN)	Support Vector Machine	Random Forest
70%	**84%**	**84%**	**90%**	**91%**

6 Conclusion and future work

We have implemented four types of algorithms which are all mentioned in an earlier section. So far, we collected data from the most prominent website named Kaggle of heart disease parents and we understood important factors about data accuracy and how it can be preprocessed. By using our implementation, we got the initial result of Decision Tree which was 70% which was a little bit good but when we saw the accuracy score of Logistic Regression that 13% higher than DT. KNN algorithm is also an excellent algorithm that performs better than Decision Tree and Logistic Regression. Resulting KNN gave a result of 84%. Then we implemented SVM and RF same datasets were imported in both algorithms but RF gave tremendous accuracy results of all algorithms. Resulting in SVM at 90% and RF at 91%. In the future, more accuracy can be improved by using more efficient methods. We can take more heavy datasets and get more accuracy by using different algorithms and techniques. Hlaudi Daniel Mase the et *al*,. Some better algorithms can be used to predict heart attack diseases like J48, Bayes Net, and Naïve Bayes [9]. Jyoti Soni *et al.*, by focusing on different types of algorithms and a combination of different types of attributes for good intelligence you can predict heart disease using data mining techniques [10]. Nidhi Bhatla et *al.*, you can use large data mining process techniques by using the Weka tool [11].

References:

1. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;1:1-6.
2. Ed-Daoudy A, Maalmi K. Real-time machine learning for early detection of heart disease using big data approach. In2019 international conference on wireless technologies, embedded and intelligent systems (WITS) 2019 Apr 3 (pp. 1-5). IEEE.
3. Singh A, Kumar R. Heart disease prediction using machine learning algorithms. In2020 international conference on electrical and electronics engineering (ICE3) 2020 Feb 14 (pp. 452-457). IEEE.
4. Kohli PS, Arora S. Application of machine learning in disease prediction. In2018 4th International conference on computing communication and automation (ICCCA) 2018 Dec 14 (pp. 1-4). IEEE.

5. Rairikar A, Kulkarni V, Sabale V, Kale H, Lamgunde A. Heart disease prediction using data mining techniques. In2017 International conference on intelligent computing and control (I2C2) 2017 Jun 23 (pp. 1-8). IEEE.
6. Sandhya Y. Prediction of Heart Diseases using Support Vector Machine. International Journal for Research in Applied Science & Engineering Technology (IJRASET)(ISSN: 2321-9653) Volume. 2020 Feb;8.
7. Battineni G, Chintalapudi N, Amenta F. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). Informatics in Medicine Unlocked. 2019 Jan 1;16:100200.
8. Ismaeel S, Miri A, Chourishi D. Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015) 2015 May 31 (pp. 1-3). IEEE.
9. Masethe HD, Masethe MA. Prediction of heart disease using classification algorithms. InProceedings of the world Congress on Engineering and computer Science 2014 Oct 22 (Vol. 2, No. 1, pp. 25-29).
10. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications. 2011 Mar 8;17(8):43-8.
11. Bhatla N, Jyoti K. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering. 2012 Oct;1(8):1-4.
12. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;1:1-6.
13. Karthiga AS, Mary MS, Yogasini M. Early prediction of heart disease using decision tree algorithm. International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST). 2017 Mar;3(3):1-7.