A Project Final Report

on

HEART DISEASE PREDICTION BY MACHINE LEARNING

Submitted in partial fulfillment of the

requirement for the award ofthe degree of

B.Tech(CSE)

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Name of Supervisor :
Mr. Anandhan k
(Assistant Professor)

Submitted By

Name of Student/s
Enrollment/Admission No.
Yash Aggarwal 19SCSE1010037
Ayush Sharma19SCSE1010047

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DECEMBER,
2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"HEART DISEASE PREDICTION BY MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of f **JULY-2021 to DECEMBER-2021**, under the supervision of **Anandhan k Assistant Professor Department of Computer Science and Engineering**, , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Yash Aggarwal 19SCSE1010037

Ayush Sharma 19SCSE1010047

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Mr. Anandhan k
Assistant Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Yash Aggarwal 19SCSE1010037 Ayush Sharma19SCSE1010047 has been held on_____and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place:

Abstract

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease. This project presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor

Table of Contents

LIST OF TABLE

CHAPTER 1

Introduction

Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports .Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. But as time is passing, a lot of research data and patients records of hospitals are available. There are many open sources for accessing the patient's records and researches can be conducted so that various computer technologies could be used for doing the correct diagnosis of the patients and detect this disease to stop it from becoming fatal. Nowadays it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately. We can use different machine learning and deep learning models to diagnose the disease and classify or predict the results. A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge pandemic predictions and also medical records can be transformed and analyzed more deeply for better predictions .

CHAPTER 2
Literature Reviews

K. Polaraju et al, [7] proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

Marjia et al, [8] developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using kfold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease.

S. Seema et al,[9] focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Ashok Kumar Dwivedi et al, [10] recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithm

Spencer et al. [9] conducted experiments on four frequently used heart disease datasets using

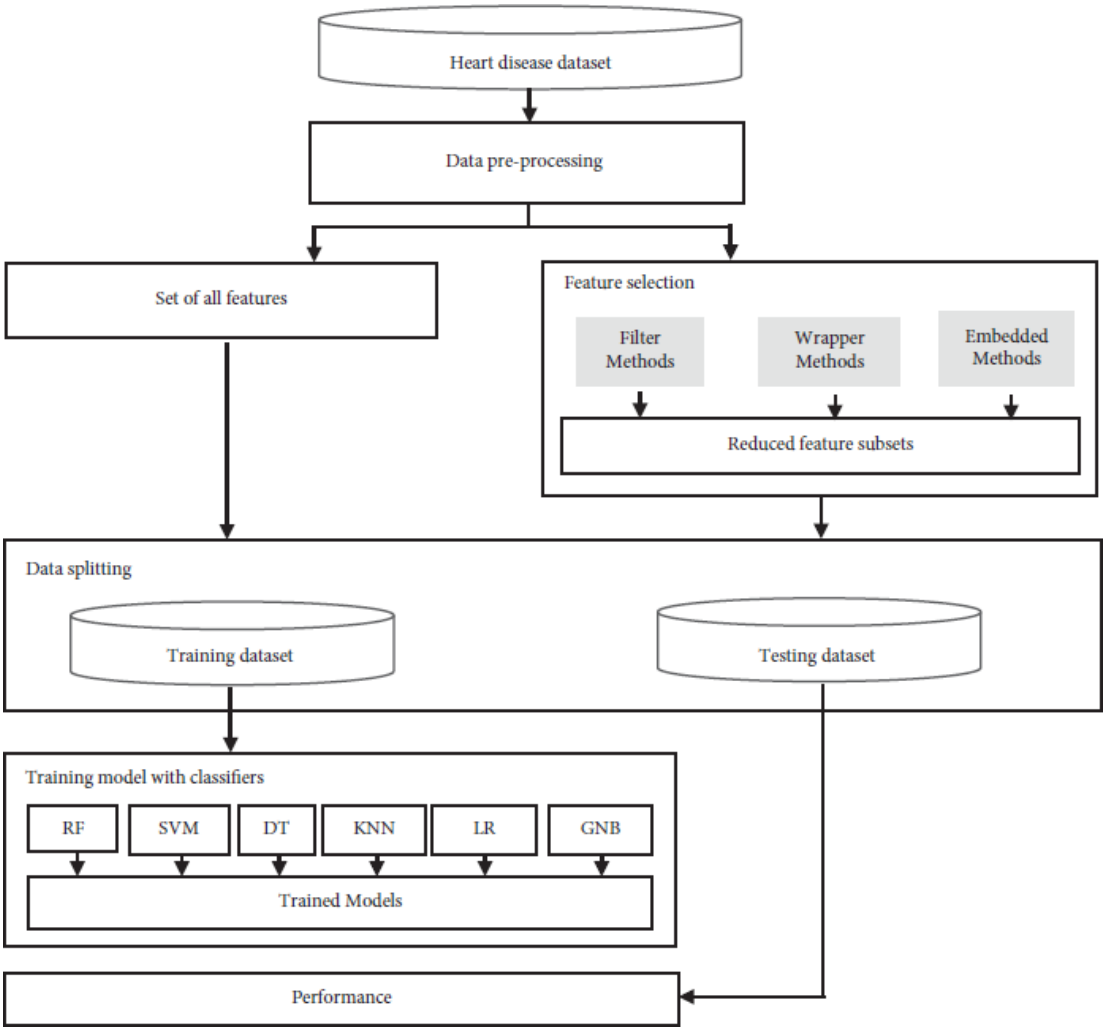
four different feature selection techniques: principal component analysis, Chi-squared testing, ReliefF, and symmetrical uncertainty. As noted by the authors, the benefits of feature selection differ depending on the machine learning approach employed for the cardiac datasets. For example, one of the most accurate models discovered had an accuracy of 85.0%, a precision of 84.73%, and a recall of 85.56% when Chi-squared feature selection was combined with the BayesNet classifier.

Semen et al. [10] constructed a diagnostic model for chronic kidney disease detection utilizing a dataset of 400 patients with 24 features. Recursive feature elimination (RFE) was used to select the most significant features. The k-nearest neighbors (KNN), support vector machine (SVM), decision tree, and random forest classification methods were used in this work. All classification methods achieved excellent performance. The random forest method beat all other algorithms, achieving 100% accuracy, recall, precision, and F1- score on all metrics.

Almansour et al. [11] used machine learning methods to assist in the early detection of chronic kidney disease by comparing two classifiers, SVM and ANN, improving their parameters using a random exhaustive search method. The 400-instance dataset from the UCI repository is preprocessed, and features are chosen using correlation coefficients. The classifiers' performance is compared against various best features (2, 3, 6, 12, and all) and the computing time required to train the model. Finally, the twelve best features are used to predict renal illness using SVM and ANN since ANN surpasses SVM with high accuracy.

CHAPTER 3
Proposed Model

The proposed study focuses on increasing classification accuracy by decreasing the number of features in a dataset of cardiac disease. The framework for classifying cardiac diseases is depicted in Figure . The key components of the framework include data collection, data preprocessing, feature selection, data splitting, model training with classifiers, and model evaluation. The following sections describe the building blocks of the suggested framework



Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features . Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below

SI NO.	Attribute Description	Distinct Value Of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0- Feamle, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering	0,1,2,3
4.	RestBP-It represents the patient‘s BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2

8.	Heartbeat- shows the max heart beat of patien	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	II,1
1II.	OldPeak- describes patient's depression level.	Multiple values between 0 to 0.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Uns1oping, Flat, Down sloping)	1,2,3
12.	CA- Result of fluoroscopy	0,1,2,3
13.	That- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	Target-lt is the final column of the dataset. It is class or label Comm. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0, l).ln class —01 represent there is less possibility of heart disease whereas —11 represent high chances of heart disease. The value —01 Or —11 depends on other 13 attribute.	0, l

Algorithm Used

In the Heart Disease Prediction we used different types Algorithm to trained our data so we can get the accurate and reliable result from our model .

- **Support Vector Machines (SVM)**

This classifier showed great empirical successes in classification tasks under supervised machine learning techniques. It separates data into different classes by a hyperplane or hyperplanes since it can handle multidimensional data. SVM separates data into hyperplanes with the help of the following mathematical formulas.

SVMs minimize the error in empirical classification and optimize the classification margin. SVM models can be categorized into four distinct categories, according to the error function method: Nu-SVM regression, C-SVM classification, Nu-SVM classification, and Epsilon-SVM regression. Support vector machine algorithm is deployed with kernel functions such as radial basis, polynomial, linear, and sigmoid kernels, which convert nonlinear data map to the linear form In this experiment, linear kernel function is used for classification.

- **Decision Tree (DT)**

A decision tree algorithm can be categorized under supervised learning, which is identical to a tree of nodes and edges ideal for classification problems. Every node in the tree represents the class of the problem, and every edge indicates the choice made based on the evaluated results. This classifier can be considered a predictive machine learning model that displays the correlation between dataset values and features. Each division in the decision tree indicates the potential value for a certain category. Dependent on entropy measurements of the dataset attributes, the nodes are identified. The maximum entropy value attribute is known to be the root node.

- **Random Forest (RF)**

One of the most widely used, technologically advanced, supervised ensemble classification methods is called random forest. It generates a large number of trees during the training stage and builds a forest of decision trees on multiple data subsets. At the testing stage, each tree in the forest assigns each piece of data a class label. When each tree predicts a class label, the ultimate determination for each set of test data is made by a majority vote. Whichever class label receives the greatest number of votes is assumed to be the right label for the test data to enhance its prediction accuracy. This procedure is performed for each piece of data in the dataset

- **K-Nearest Neighbor (KNN)**

K-nearest neighbor is a technique of supervised learning used to recognize the patterns in the dataset and determine the belonging class based on the points of existence with the nearest neighbors. It determines the class of instance by considering the distance from the nearest neighbor and consistency. KNN uses different distance functions such as Manhattan, Euclidean, Hamming, and Minkowski when identifying the

neighbor's points. This method is called a memory dependent approach because all the instance points are stored inside the memory .The weight is allocated for each point depending on the distance from each instance to improve the algorithm's performance. The following equation is used to calculate the Euclidean distance:

- **Logistic Regression (LR)**

Logistic regression is a widely used machine learning method that falls into the category of supervised learning. It is used to forecast a categorical dependent variable based on a set of independent variables. For example, it is used to extract significant statistical items from the model or forecast the tendency of data . The dependent variable in logistic regression is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). The logistic regression algorithm is based on the logistic function shown below:

$$f(z) = \frac{1}{1 + e^{-z}}$$

- **Gaussian Naive Bayes (GNB)**

This algorithm, based on the theorem of Bayes, expects independence between two variables. The construction of a Bayesian model is simple and particularly realistic, based on the massive datasets. Naive Bayes' simplicity and dominance over complex approaches are well established [2, 8].

Gaussian naive Byes is used when all of the data values of a dataset are numeric. The probability density function is defined using the mean and standard deviation. Next, it computes the mean and standard deviation for each of the dataset's features. After computing this, the probabilities for each test data pattern are calculated using the mean and standard deviation when any test data pattern occurs. Finally, it assigns an appropriate class label to the test data whose probability is near to 1. The following equation is used to determine the likelihood.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

where μy is the mean and σy is the standard deviation

CHAPTER 4
RESULT AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree, Naive Bayes and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

$$\text{F– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

• TP True positive: the patient has the disease and the test is positive.

• FP False positive: the patient does not have the disease but the test is positive.

• TN True negative: the patient does not have the disease and the test is negative.

• FN False negative: the patient has the disease but the test is negative.

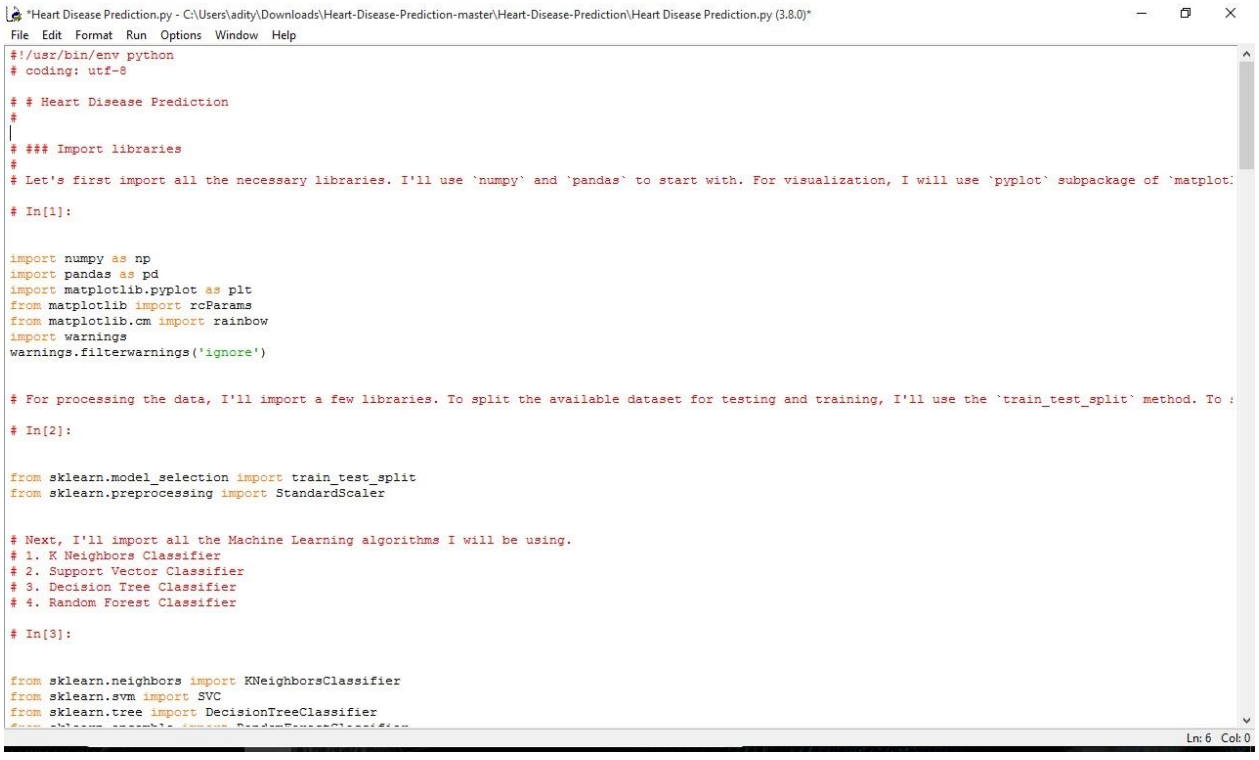
In the experiment the pre-processed dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 2. The accuracy score obtained for Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques[12] is shown below in Table

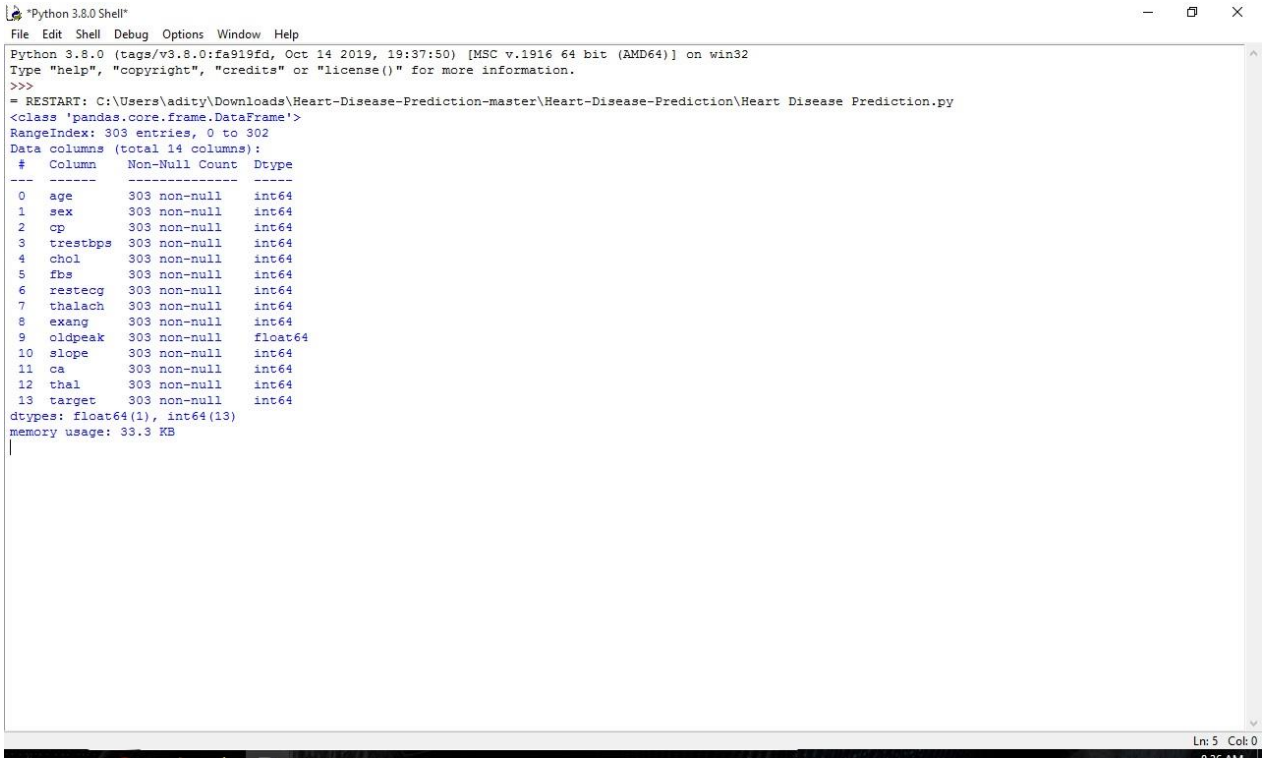
Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	22	5	4	30
Naive Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

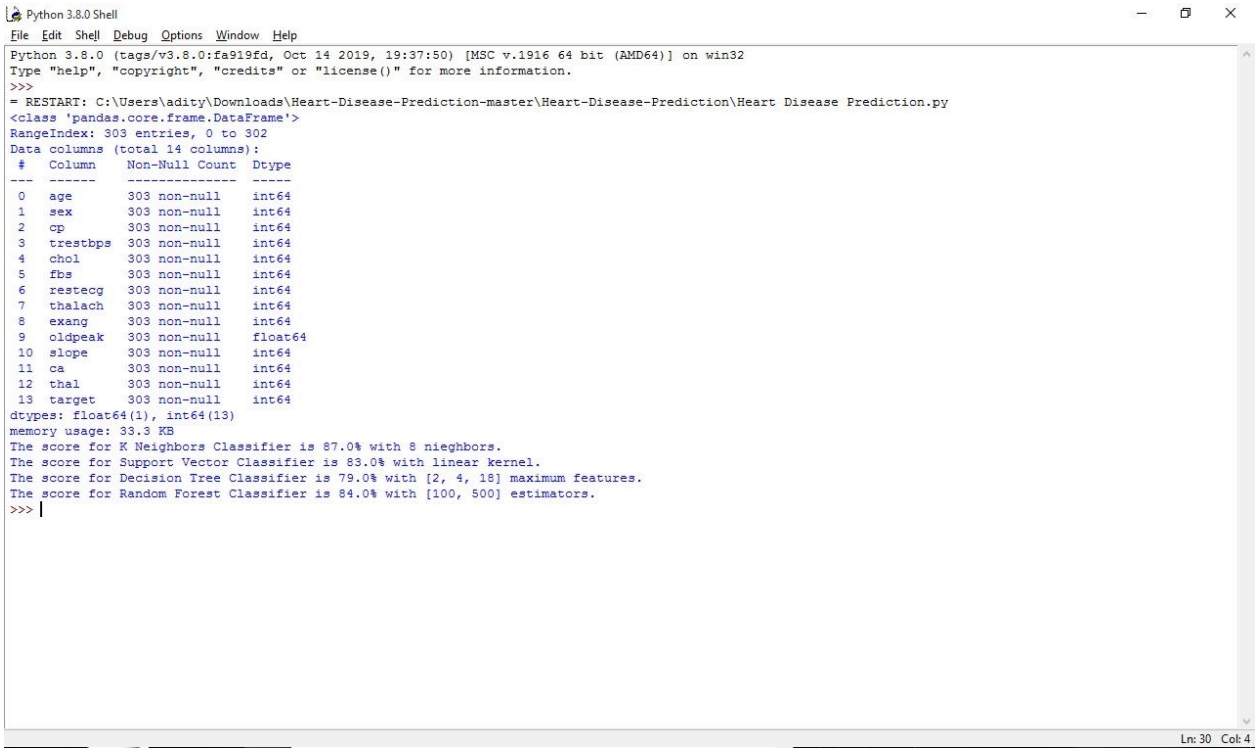
TABLE VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	Precision	Recall	F-measure	Accuracy
Decision Tree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
Random Forest	0.937	0.882	0.909	90.16%
Naive Bayes	0.837	0.911	0.873	85.25%

TABLE ANALYSIS OF MACHINE LEARNING ALGORITHM




Python 3.8.0 Shell
File Edit Shell Debug Options Window Help
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\adity\Downloads\Heart-Disease-Prediction-master\Heart-Disease-Prediction\Heart Disease Prediction.py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
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
Ln: 5 Col: 0


Python 3.8.0 Shell
File Edit Shell Debug Options Window Help
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\adity\Downloads\Heart-Disease-Prediction-master\Heart-Disease-Prediction\Heart Disease Prediction.py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
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
The score for K Neighbors Classifier is 87.0% with 8 nieghbors.
The score for Support Vector Classifier is 83.0% with linear kernel.
The score for Decision Tree Classifier is 79.0% with [2, 4, 18] maximum features.
The score for Random Forest Classifier is 84.0% with [100, 500] estimators.
>>>
Ln: 30 Col: 4

CHAPTER 5
Conclusion

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

Future work

There are many possible improvements that could beexplored to improve the scalability and accuracy of thisprediction system. Due to time limitation, the following research/ work needs to be performed in the future.

• Like to make use of testing different discretizationtechniques, multiple classifiers Voting technique anddifferent Decision tree types like information gain,gain ratio and Gini index. Eg. Experiment need to performon use of Equal Frequency Discretization GainRatio Decision Trees by applying nine Voting schemein order to enhance the accuracy and performance ofdiagnosis of heart disease.

• This paper proposes a framework using combinationsof support vector machines, logistic regression and decision trees to arrive at an accurate prediction ofheart disease. Further work involves development ofsystem using the mentioned methodology to be usefor checking the imbalance with other data miningmodels.

• Like to explore different rules such as Association, Clustering, K-means etc for better efficiency and easeof simplicity.

• To make use of Multivariate Decision Tree approachon smaller and larger amount of data.

REFERENCES

[1] Avinash Golande, Pavan Kumar T, ‖Heart Disease Prediction Using Effective Machine Learning Techniques‖, International Journal of Recent Technology and Engineering, Vol 8,pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy,‖ Heart Disease Prediction using Data Mining with Mapreduce Algorithm‖, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi,‖ Implementation of Machine Learning Model to Predict Heart Failure Disease‖, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6,2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, ‖Design And Implementation Heart Disease Prediction Using Naives Bayesian‖, International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477,2019.

[7] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[8] Sayali Ambekar, Rashmi Phalnikar,―Disease Risk Prediction by Using Convolutional Neural Network‖,2018 Fourth International Conference on Computing Communication Control and Automation.

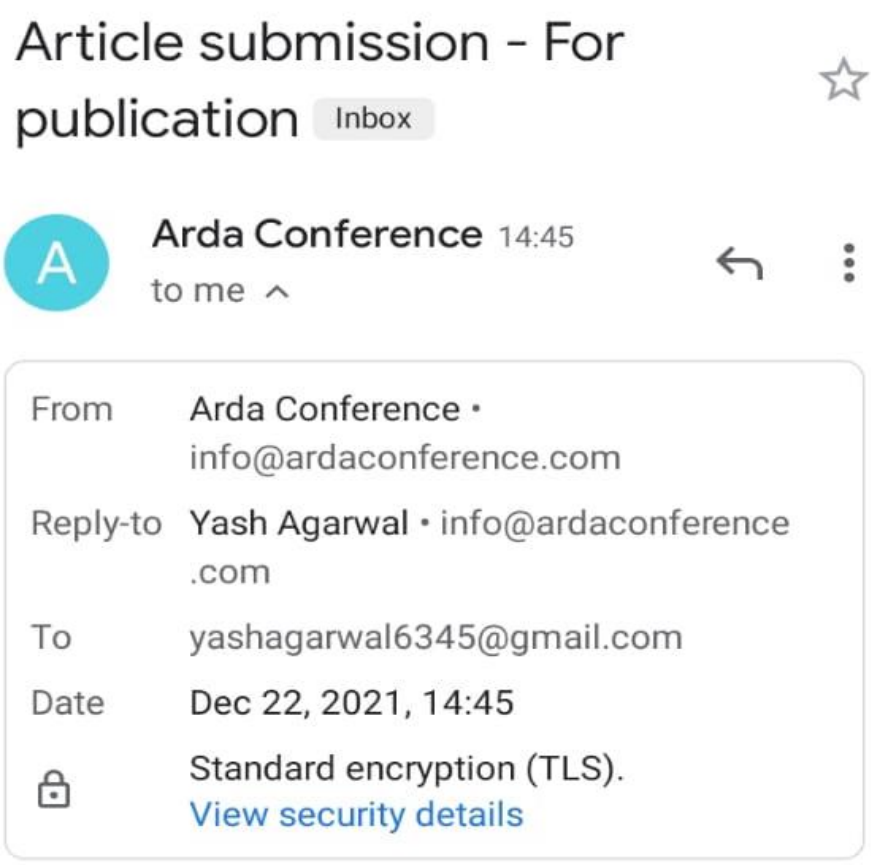
[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018.

[11] Fajr Ibrahem Alarsan., and Mamoon Younes _Analysis and classification of heart diseases using heartbeat features and machine learning algorithms‘,Journal Of Big Data,2019;6:81.

[12] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap

Publication/Communication Proof



Article submission - For publication Inbox

Arda Conference 14:45
to me

From Arda Conference • info@ardaconference.com
Reply-to Yash Agarwal • info@ardaconference.com
To yashagarwal6345@gmail.com
Date Dec 22, 2021, 14:45
Standard encryption (TLS).
View security details

Dear **Yash Agarwal**,
Thanks for submitting your article. We have received your article and sent the same for peer-review. Your paper ID is :ARDA_Publication_982064 and here after the same shall be used for all further communications regarding your article. You will get notified about the status of your article within 24-48 hours. Thanks for your patience.
Regards,
Team ARDA CONFERENCE.