Prediction of Heart Disease using RandomForest.

VISHAL TIWARI
INTERN AT SPINNAKAR ANALYTICS

Prediction of Heart Diseases using Random Forest

Abstract. The process of discovering or mining information from a huge volume of data is known as data mining technology. Today data mining has lots of application in every aspects of human life. Applications of data mining are wide and diverse. Among this health care is a major application of data mining. Medical field has get benefited more from data mining. Heart Disease is the most dangerous life-threatening chronic disease globally. The objective of the work is to predicts the occurrence of heart disease of a patient using random forest algorithm. The dataset was accessed from Kaggle site. The dataset contains 303 samples and 14 attributes are taken for features of the dataset. Then it was processed using python open access software in jupyter notebook. The datasets are classified and processed using machine learning algorithm Random forest. The outcomes of the dataset are expressed in terms of accuracy, sensitivity and specificity in percentage. Using random forest algorithm, we obtained accuracy of 83.6% for prediction of heart disease with sensitivity value 82%. From the receiver operating characteristics, we obtained the diagnosis rate for prediction of heart disease using random forest is 83.6%. The random forest algorithm has proven to be the most efficient algorithm for classification of heart disease and thereforeit is used in the proposed system.

1. Introduction

Data mining is also known as proficiency discovering from data. It attempts to withdraw hidden pattern and trends from huge data bases. Data mining also support automatic exploration of data. The main objective of data mining technique is to find the hidden data in the data base. It is also called as exploratory data analysis, data driven and deduction learning. It extracts meaningful information from database. When the database is very large i.e in terabyte to petabytes manual analysis of data is not possible. So, we need automatic data analysis. Data mining was introduced in 1990s.Various data mining technologies are as follows.

(i)**Statistics:**

Regression analysis, cluster analysis, standard deviation etc. are the foundation of data mining.

(ii)**Artificial Intelligence:**

It is the applying of human thoughts like processing

(iii)**Machine Learning**

It is the integration of statistics and AI technology. It is about learning by the software about data.

The world is filled with data such as pictures, video, music. Machine learning promise to derive a meaning for all the data. Arthur C. Clarke states that modern technology is filled with magic. There is lots of data in the world generated not only from people but also from mobile, computer and from another device. Automatic system can ascertain from data and can change the data. Machine learning has wide application in the field of speech processing, image processing, fraud detection. Also, in the field of medical science such as diabetes retina path, Skin cancer detection, heart disease. Using data isreferred to as for training and answer refer to as prediction. Training data refers to create a model and to predict. This predictive model can then use to serve predictions on previously unseen data and answer the questions. The paper is outlined as follows. Section 2 presents an idea about the related. work done by the paper. Section 3 gives an idea of the major cause of heart disease, symptoms, prevention of heart disease. Section 4 represents the result part of the experiment. Section 5 conclude the present paper.

2. Related Work

The proposed study gives a prediction method for classification of heart disease. The risk factor which can control and which cannot control was explained in this paper. The prediction of heart disease has been done by random forest machine learning algorithm.

Ref [1] proposed a user-friendly heart disease prediction system (HDPS). Authors have taken 13 clinical features for classifying heart disease using artificial neural network. Prediction accuracy obtained by the system is approximately 80%. HDPS system include clinical data section, ROC curve section, estimation display section.

Ref [2] authors have proposed a Diabetes disease prediction system that gives diabetes maladyanalysis. Two algorithms were applied namely Bayesian and K-NN for prediction of diabetes.

Ref [3] author has proposed a model for predicting heart disease by taking samples of 300 patient record using Naïve Bayes and decision trees. data was taken from UCI repository site Author used id3 algorithm for constructing decision tree. For small data set decision tree does not give accurate result but Naïve bayes gives more accurate result if the input data is cleaned.

Ref [4] author have proposed a data mining model to predict weather a patient has heart disease or not. Two types of data mining algorithm decision tree and naïve bayes were used for forecasting. These two algorithms were applied on the same data set. Decision tree show an accuracy of 91% and naïve bayes algorithm show an accuracy of 87%. So, in the paper decision tree gives better accuracy for predicting heart disease.

Ref [5] authors have proposed a data mining model for prediction of heart disease. Dataset was taken from UCI machine learning repository site. Four data mining algorithms such as Naïve bayes, random forest, Linear regression, Decision tree were applied by the authors to predict the heart disease. Among these algorithms random forest gives good accuracy of 90.16% compared to other algorithms.

Ref [6] authors have used knn, decision tree, linear regression, support vector machine algorithms for prediction of heart disease and compared their accuracy. All the datasets for prediction are accesses from UCI repository site. For implementation of the algorithm's python software is used. All the algorithms are processed in jupyter notebook. From the experimental result authors have obtained best accuracy of 87% by using k-nearest neighbor algorithm followed by support vector machine 83%, decision tree 79%and linear regression of 78% accuracy among all these algorithms for prediction of heart disease.

Ref [7] authors have proposed an application for prediction of heart disease for juveniles using multilayer perceptron algorithms. Authors used Cleveland dataset accessed from UCI library the dataset containing 76 parameters such as chest pain, CT scan, ECG etc. The data set was processed in python code using PyCharm tool. From the experimental result authors obtained precision, recall, support value for positive classes were 0.92,0.9,93and for negative classes 0.91,0.89,0.72 respectively.Ref [8] authors have proposed a model for prediction of cardiovascular disease using machine learningalgorithm hybrid random forest with linear mode. Authors obtained 88.7% accuracy for prediction of CVD using hybrid random forest with linear model. The dataset was collected from UCI repository site. Authors have chosen Cleveland dataset for this proposed study.

3. Heart Disease

The Heart is the most important organ of human body. If it does not function properly then it affects other organ of the body. According to a report 7,000,000 die from heart attacks each year. According to WHO report around 17.9 million people die due to CVDS in 2016. 31% of the death of people is due to Heart disease around the globe in every year. The pumping of blood to the human body is the vital function of heart which supply oxygen and nutrients to the human body and also remove other metabolic waste from the body. If there is deficiency of blood in human body then heart doesn't function properly and it stop working which causes the death of human being. Angina occurs when there is temporary loss of blood to the heart causing chest pain. Cardiovascular disease is of two types.

(1) Heart Attack-It occurs when the heart blood vessels are suddenly blocked.
(2) Heart Failure-It results from coronary heart disease, hypertension, cardiomyopathy. Heart failure is basically when the heart is unable to maintain a strong blood flow and this results in chronic tiredness, resist physical activities and shortness of breath. Heart failure can be divided into three types.1. right side heart failure 2. Left side heart failure 3. congestive heart failure.
Right sided heart failure usually causes left sided heart failure. In right sided heart failure blood backs up into other tissues such as liver and in the abdomen causing congestion in these areas. As a result of right sided heart failure, we can have Hepatomegaly and Anciles.
In left sided heart failure oxygenated blood cannot be pumped out from heart to the rest of the body. So, blood can back flow. Blood can accumulate in lung veins causing fluid accumulation in lungs causes shortness of breath and oedema.

Table 1. Major cause of heart disease [10].

Disease Type
Smoking
High Blood Pressure
High Cholesterol
Diabetes and Prediabetes
Being overweight
Physical inactivity
Metabolic syndrome

Risk factor that cannot control for heart disease
1. family history
2.55 years or older
3. History of preeclampsia
Symptoms of Heart attack
Nausea
(a)Dizziness
(b)Jaw pain
(c)Abdominal pain
Living a healthy life style can reduce the effect of heart disease. Drinking plenty of water, eating green vegetables, fat free food, doing exercises, regular check-up of heart, consulting with the doctor if there any family history of heart disease can reduce the effect of heart disease.

4. Methodology

For the proposed study dataset was taken from Kaggle site. Then it was downloaded in excel file using comma separated format. Data has processed by python programming using Jupiter notebook. The data set contains 303 sample instances as shown in table3. The dataset contains 14 clinical features as shown in table 2. Different types of python libraries such as pandas, Sklearn, NumPy, matplotlib are used for processing the algorithms. Using explorative data analysis technique data was analysed in jupyter notebook.10-fold cross validation technique is used for spitting the data set into training and testing data. Then using random forest algorithm dataset was processed.

description of the algorithms:

Machine learning is the ability of computer to learn automatically from the experience.

Machine can learn by three ways.

1.supervised learning

2.Unsupervised Learnig

3.Reinforcement learning

In supervised learning label data is given to the machine for prediction.

K-NN, Naïve Bayes, Support vector machine, Decision tree, Random forest algorithms are supervised machine learning algorithms.

In unsupervised learning algorithms label data is not given to the machine for prediction.

Clustering, c-means are the examples of unsupervised learning

In reinforcement learning machine learn by itself without any guidance. It learns from the environment and there is a reward for every action.

Q-learning is one of the examples of Reinforcement learning.

Random forest is a supervised machine learning algorithm that constructs several decision trees. The final decision is made based on the majority of decision tree. Decision tree suffer from low bias and high variance. Random forest converts high variance to low variance.

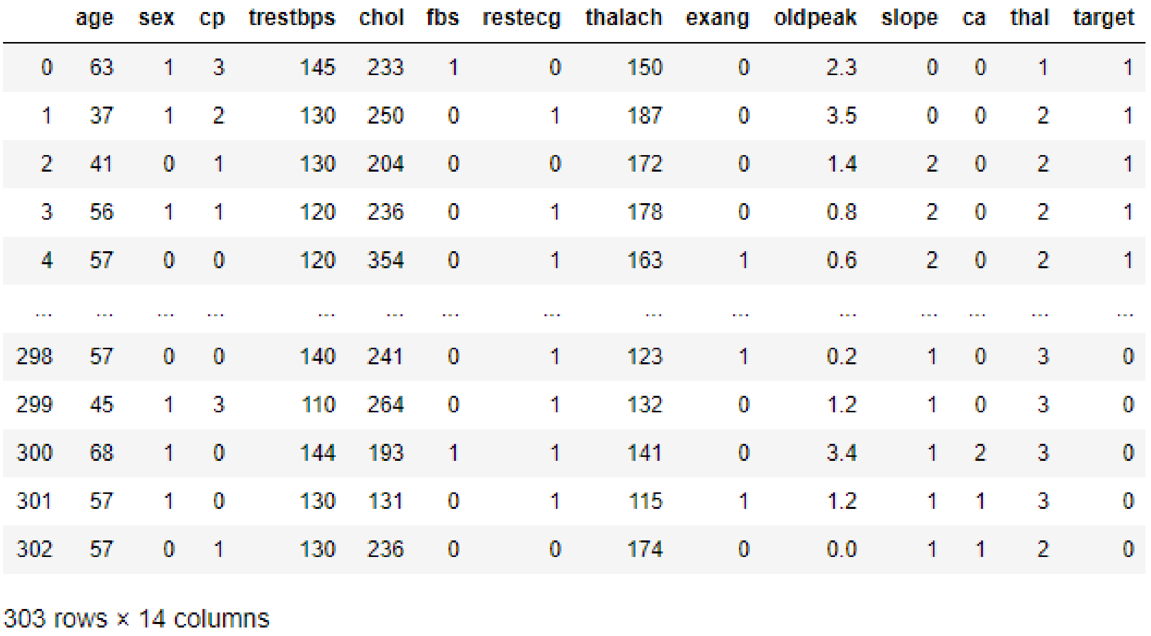
Table 2. Features for data prediction

Attribute	meaning
Age1	Age is continuous
Gender 1	1=male 0=female
Cp1	Chest pain
Trestbps	Resting blood pressure results during hospitalised: continuous(mmHg)
chol	cholesterol level in mg/dl
Fbs1	Fasting blood sugar 0:<=120mg/dl,1:>120mg/dl
restecg	electrocardiographic results during resting 1=true 0=false
thalach	Maximum heart rate achieved: continuous
exang	Exercise induced angina
oldpeak	ST depression
slope	ST segment slope
ca	Number of major vessels coloured by fluoroscopy: discrete(0,1,2,3)
thal	3: normal 6: fixed defect 7: reversible defect

5. Result and Discussion

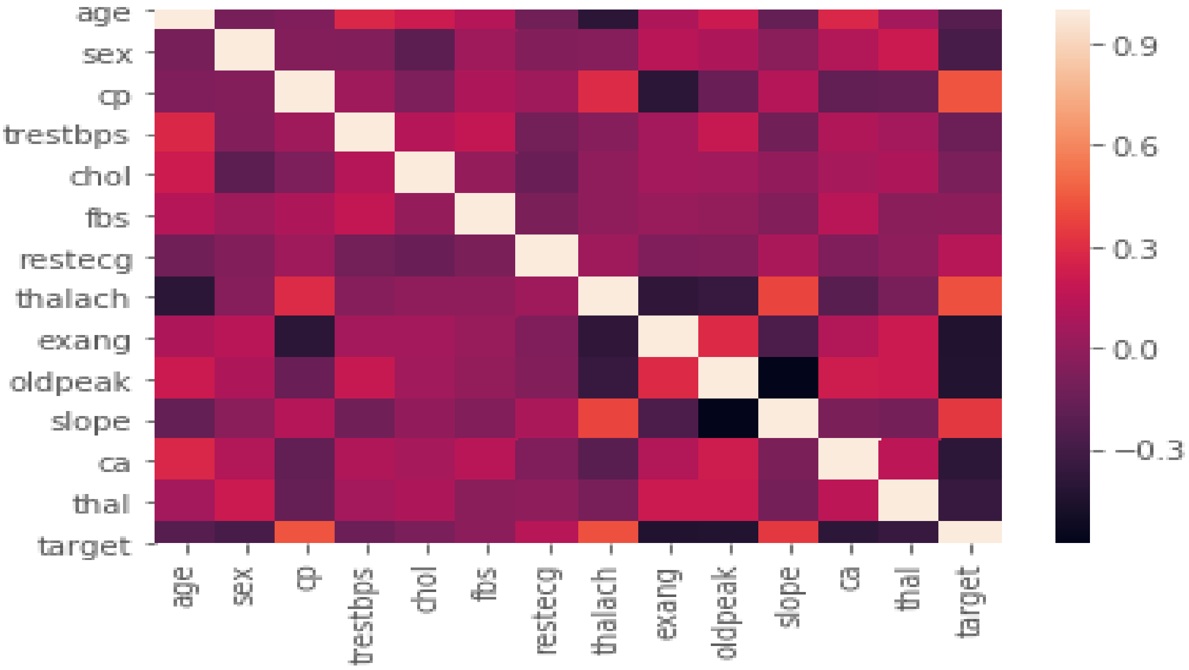
The present work predicts suffering rate of a patient from heart disease using random forest algorithm. Total 303 data samples of 14 clinical features have taken for prediction of heart disease.80% of the dataset has taken for training and 20% has taken for testing phase.

Table 3. Features for data prediction



	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

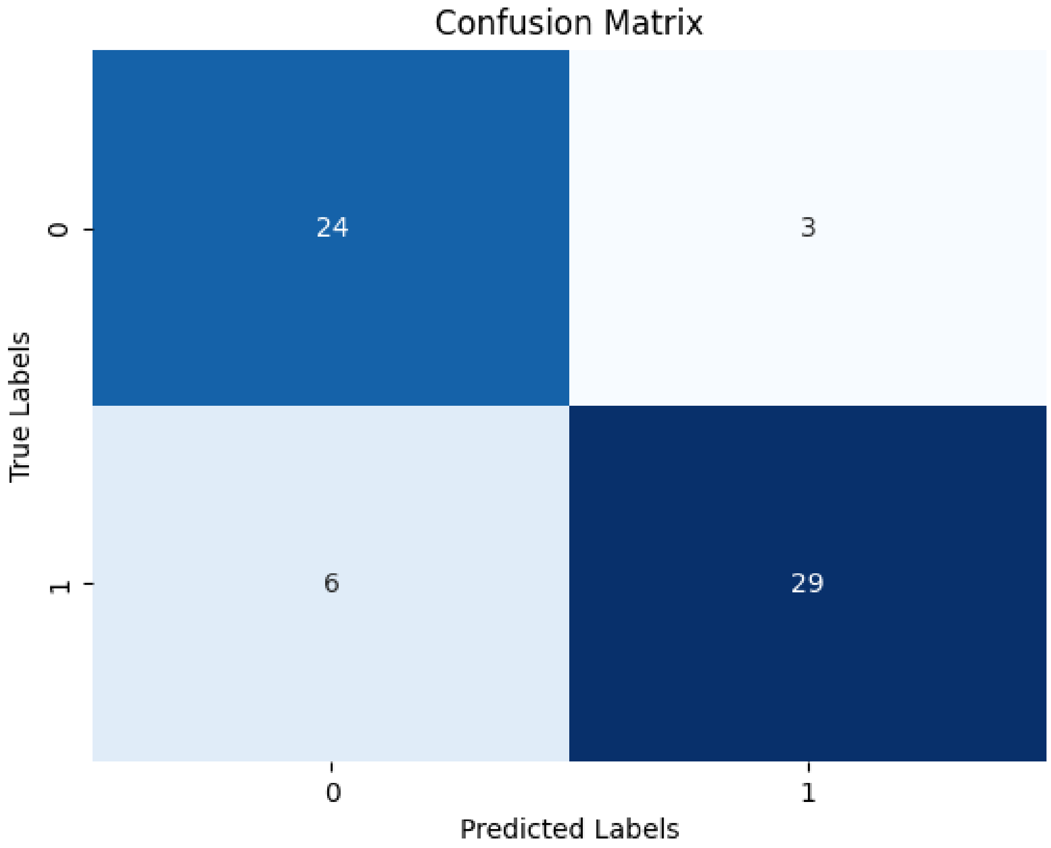


After doing explorative data analysis we obtained the correlation matrix which correlate the attributes of the data set.

We are applying random forest algorithm to the testing data set for creating a confusion matrix. From the confusion matrix we get more sophisticated metrics like sensitivity, specificity and AUC that can help us to make a decision in the classification process.

6.Conclusion and insights from model evaluation;

Our heart disease classification model demonstrates commendable performance, as evidenced by a thorough examination of key metrics and evaluation tools.



F1 Score and classification report;

The F1 score, a harmonic mean of precision and recall, stands at an impressive 0.86. This indicates a strong balance between correctly identifying patients with heart disease (sensitivity) and minimizing false positives. The classification report provides a detailed breakdown of precision, recall, and F1 score for both classes (0: absence of heart disease, 1: presence of heart disease). Notably, the model achieves high precision and recall for both classes, highlighting its ability to accurately predict both positive and negative cases

Confusion matrix;

Analyzing the confusion matrix further provides valuable insights into the model's performance:
True Positives (1): 24
False Positives (0): 3
True Negatives (0): 29
False Negatives (1): 6
These metrics elucidate the model's proficiency in correctly identifying individuals with heart disease (sensitivity) while minimizing false alarms.

ROC-AUC Curve Interpretation;

The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) curve is a pivotal tool for evaluating the model's ability to discriminate between positive and negative instances across various thresholds. With an AUC of 0.91 and a ROC AUC score of 0.92, the model showcases strong discriminatory power. The curve's proximity to the upper-left corner indicates high sensitivity and specificity.

Sensitivity and specificity;

Sensitivity (True Positive Rate): 0.80

Specificity (True Negative Rate): 0.91

A balance between sensitivity and specificity is crucial in a heart disease classification context. The model achieves a high true positive rate, ensuring that individuals with heart disease are accurately identified. Simultaneously, the high true negative rate emphasizes the model's ability to correctly identify individuals without heart disease.

In conclusion, our heart disease classification model excels in achieving a harmonious balance between sensitivity and specificity, as reflected in the robust F1 score, classification report, and the discriminatory power highlighted by the ROC-AUC curve. These findings instill confidence in the model's capability to assist in accurate heart disease predictions and guide healthcare interventions effectively.Theproposed system can also be used for prediction of other disease by applying with other machinelearning algorithm such as Naïve Bayes, decision tree, K-NN, Linear regression, fuzzy logic for betteraccuracy. Cloud computing technology can also be used for the proposed system to manage largevolume of patient data.

References

[1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: “Heart disease prediction system”. In 2011 Computing in Cardiology (pp. 557-560). IEEE.

[2] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining."In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017.

[3] Rajesh , T Maneesha, Shaik Hafeez, Hari Krishna“Prediction of Heart Disease Using MachineLearning Algorithms“May 2018International Journal of Engineering & Technology 7(2):363-366DOI: 10.14419/ijet. v7i2.32.15714

North-Holland/American Elsevier) p 517

[4] J. Krishnan Santana; S. Geetha “Prediction of Heart Disease Using Machine Learning Algorithms”. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

[5] Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam.” Heart Disease Prediction using Machine Learning” INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

[6] Singh, A., & Kumar, R. (2020). “Heart Disease Prediction Using Machine Learning Algorithms”. 2020 International Conference on Electrical and Electronics Engineering (ICE3). doi:10.1109/ice348803.2020.9122958

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”. IEEE Access, 1–1. doi:10.1109/access.2019.2923707

[8] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). “Heart disease prediction using data mining techniques”. In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE

[9] Al Essa, Ali Radhi, and Christian Bach. "Data Miningand Warehousing." American Society for EngineeringEducation (ASEE Zone 1) Journal (2014).

[10] National Health Council, ‘Heart HealthScreenings’,2017. [Online]Available: http://www.heart.org/HEARTORG/Conditions/HeartHealthScreenings_UCM_428687_Article.jsp#. WnsOAeeYPIV