International Journal for Modern Trends in Science and Technology, 8(S03):215-224, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST08S0345
Available online at: https://www.ijmtst.com/vol8si03.html

Predicting the Risk Level of Heart Disease using Machine Learning

Sravani.Ch.V.K, Dr.Sujatha.B, Dr.Leelavathy.N

Department of Computer Science and Engineering, Godavari Institute of Engineering and Technology (A), JNTUK, Kakinada.

To Cite this Article
Sravani.Ch.V.K, Dr.Sujatha.B and Dr.Leelavathy.N. Predicting the Risk Level of Heart Disease using Machine Learning. International Journal for Modern Trends in Science and Technology 2022, 8(S03), pp. 215-224. https://doi.org/10.46501/IJMTST08S0345

Article Info
Received: 26 April 2022; Accepted: 24 May 2022; Published: 30 May 2022.

ABSTRACT

Predicting heart disease can be named as most difficult tasks in medical field. In present world, among millions of people exactly one to two people expires per two minutes. Now-a-days machine learning playing vital role in healthcare filed. Predicting heart disease is difficult task so, we need identify it in early stage and warn patient ahead of time which needs automation of predicting process. So, we identify patient risk level based on some risk factors , classify dataset using machine learning algorithms(classification models) and also presents performance comparison between different classification models and will choose the best model based on its performance on how it classifies the data. The classification models we are going to use are "KNN", "Random Forest", "SVM", "Decision Tree". The data set we considered is "Cleveland dataset ", "Z-Alizadeh Sani" .We find the relationship between attributes of dataset, find best feature associated to target.

Keywords: KNN, SVM, Random Forest, Classification models, Decision Tree.

1.INTRODUCTION

Heart disease may occur due to smoking, based on family history, alcohol consumption and an unhealthy life style. Most of the deaths in this world are occurred due to **Heart Diseases** like Heart Failure, Heart Attack, Cardio Vascular Disease etc. These diseases occur due to the changes in lifestyle, consumption of beverages like alcohol, wine, beer etc smoking, increase in blood cholesterol due the intake of fat containing foods. According to "WHO", people around all countries passed away due to heart diseases are exceeding more than 10 million. It becomes easy to deal with heart disease if it identified in early stage. And we also need effective methods to deal with so, by using machine learning algorithms we can analyze the situation of patient condition by acquiring their data records. It is difficult to find patterns of heart disease in patients through data records so, we find patterns using Python Programming by plotting graphs which will be helpful to know the main cause of heart disease. The data records are available in different websites used to identify and retrieve patterns in the form of graphs. Machine learning not used in medical field but also in other fields like banking, business etc . The importance of this paper to show how crucial role machine learning plays in medical field and assurance to heart disease patients. This paper presents performance analysis of ML algorithms like Decision Tree, SVM, Random Forest, KNN after classifying the data.

Related Job

Many people have worked on datasets of "UCI Machine Learing" related to heart diseases prediction using

various machine learning algorithms some of them are mentioned below.

H.Benjamin's paper 3 shows the differentiation between Classification models like "Random Forest", "Decision Trees", "Clustering" & "Naive Bayes". And he concluded that "Random Forest" done best job in order to predict the heart disease than other used algorithms like decision tree, clustering etc.

Theresa Princy. R 4 provided conclusion on different classification algorithms used for predicting heart disease i.e., which is best. The classification models that are include were "Naive Bayes", "KNN (KNearest Neighbour)"," Decision Trees", "Neural Networks". Finally, the accuracy of classifiers are analyzed based on the various attributes of datasets.

M.A.Jabbar's 5 model uses random forest algorithm, decision tree algorithm as classification algorithm. Feature selection, chi square and genetic algorithms are used as measures to predict risk level of heart disease. To construct a classifier dataset divided into 75% of training set and 25% testing set. The highestaccuracy obtained for decision tree with 63.3%.

Fahd Saleh Alotaibi 6 developed a "Machine Learning Model " by differentiating 5 types of classification algorithms. Tools like "Rapid Miner tool ", "Matlab","Weka tool " were used for accuracy retrieval in which Rapid Minor tool given an higher accuracy when compared to other tools. In this survey the accuracy of "Decision Trees", "Logistic Regression", "Random forest", "Naive Bayes" and "Support Vector Machine" classification algorithms were also compared and found that decision tree algorithm had the highest accuracy.

Anjan Nikhil Repaka 7 given another ml model that uses "Naïve Bayesian" technique for classification of dataset and "Advanced Encryption Standard" algorithm to provide security while data transferring in order to predict the disease.

Nagaraj M Lutimath 8 done survey on the heart disease(HD) prediction using "Naive Bayes" and "SVM (Support Vector Machine)". The performance measures used in this analysis are "Mean Absolute Error"," Sum of Squared Error" and "Root Mean Squared Error". Finally, it was known that SVM given better accuracy.

S.Kiruthika Devi, S. Krishnapriya and Dristipona Kalita's 9 paper is as follows. This paper is for perfect analysis of heart disease. The output of each algorithm is combined. After combining, the output will be compared. Different algorithms are used such as "Decision Tree", "KNN", "Nave Bayes "and "SVM" algorithms are used.

Vishal Jadhav, Devendra Ratnaparakhi, Tusdhar Mahajan's 10 paper gives us details about Heart Disease Prediction System which is a web application. This web application fetches the data from stored database and compares them with the stored dataset.

Mohammed Abdul Khaleel 11 implemented various data mining techniques to know risk level of heart disease. In his research he included various measures that are responsible for heart disease like lungs related issues too. Using "Information data mining" technique he found the patterns of patient situation i.e., risk level of heart disease used to warn every individual in advance. Here the classification model like "Naive Bayes" algorithm is used. Basically, "Naïve Bayes" algorithm uses Bayes theorem to perform tasks. Finally, in his paper it was concluded that "Naive Bayes" done best job to predict heart disease very accurately. Sources for datasets were taken from one of the leading institute i.e., "Diabetic Research Institutes of Chennai, Tamilnadu". The dataset comprises data records of more than five hundred and data split was done like 70% for train set and 30% for test set. Final accuracy obtained by "Naive Bayes" is 86.419%.

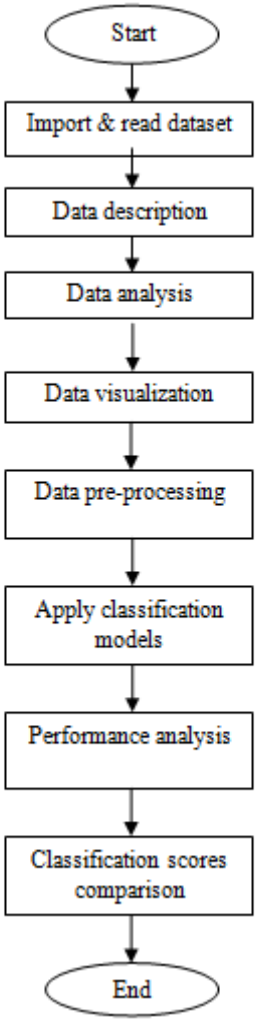
Costas Sideris, Nabil Alshurafa, Haik Kalantarian &Mo-hammad Pourhomayoun 12 done survey on "Remote Health Monitoring " Outcome Success prediction using First Month and Baseline Intervention Data. **RHS** systems are effective in saving costs and reducing illness. In this research, they produced a upgraded **RHM** framework, Wanda- CVD which is a type of mobile helps people in giving instructions to help other people.

Contribution

Our proposed model is to find risk from acquired dataset and inform the patient in advance which is different from other papers. The risk is calculated from the features(cholesterol, chest pain etc) of a dataset . We apply classification models on dataset and also perform feature selection, construct a decision tree which helps us to find the best feature which is relatable to target.

2. METHODOLOGY

Methodology involves step like data description means describing the data i.e., to know the shape of the data, finding missing values, knowing type of data whether it is categorical , nominal etc. We import dataset "heart_dataset_aliza.csv", "heart_dataset_kaggle.csv". In data analysis step we need to find the relationship between attributes or features with respective to target and the best relatable feature towards target is selected. In data visualization we visualize data in the form of bar graphs using countplot, distplot line graph etc. In data pre-processing step we add custom columns to dataset like risk and plot graphs between risk and other features to know if the features are relatable to risk or not. Finally we apply classification models on dataset and construct a decision tree to know the best split attribute. Classification is used to know whether patient is suffering from heart disease or not



Data Description

Dataset is taken from 2. Cleveland dataset as shown in below table

TABLE 1. Cleveland Dataset

Attributes	Attribute Description
1. age	Age in years values between 29-71
2. sex	(female) 0 & (male) 1
3. cp	Represents the severity of chest pain in patients of values 0,1,2,3
4. trestbps	Represents patients BP ranges 94-200
5. chol	*Shows cholesterol level of patients ranges 126-564*
6. fbs	Fast blood sugar either 0 or 1
7. restecg	Resting ECG , values 0,1,2
8. thalach	Max heart rate achieved ranges 71-202
9. exang	Exercise induced angina 0= No, 1=Yes
10. oldpeak	Describes patients depression level ranges 0-6.2
11. ca	Number of major vessels (0-3) colored by fluoroscopy
12. slope	Says patient condition during peak exercise. It is divided into 3 segments (Unsloping , Flat, Down Sloping)- values 1,2,3
13. thal	Normal, fixed, reversible, non-reversible (0,1,2,3)
14. target	(no HD) 0 and (HD) 1

The presented dataset in this paper is compression of 2 .Dataset is collected from 2. is combination of "seventy six" attributes and 1025 entries and this dataset is known to be "Cleveland" dataset. Only 14 out of 76 features was chosen which are mentioned in above table.

The second dataset is taken from 1. which contains 56 attributes and summarized to 20 attributes and one class variable i.e., target variable shown in TABLE 2.The summarized dataset contains 303 rows.

In TABLE 2 there are variety of attributes that explains the risk level of HD in patients.Some of the features that can be considered as risk factors are like chest pain, cholesterol, blood pressure. The risk of chest pain can be increased if LDL (low density lip-protein) from TABLE 2 is high. Heart disease can also occur through inheritance which is given as FH (family history) from TABLE 2. Due to high level of LDL may also increase in BMI which leads to obesity. In TABLE 2 it was also shown that whether the patient is current or ex-smoker which is also a reason for chest pain.

Using these features we can identify the risk level of patients in advance all the features used are numeric, binary types.

Another dataset is taken from 1. Z. Alizadeh sani dataset

TABLE 2. Z.Alizadeh Sani Dataset

Attributes	Attribute Description
1. Age	Age in years
2. Weight	Weight in kg
3. Height	Height in cm
4. Sex	(female) 0 & (male) 1
5. BMI	*Body mass index*
6. Obesity	Either TRUE or FALSE
7. DM	Diabetes Mellitus (0 or 1)
8. FH	Family history (0 or 1)
9. C-smoker	Current smoker (0 or 1)
10. EX-Smoker	Ex-smoker(0 or 1)
11. BP	Blood pressure
12. PR	Pulse rate (given per minute)
13. Cp	Chest pain values -0,1,2,3,4,5
14. HB	Hemoglobin
15. Depression	Either 0 or 1
16. FBS	Fasting blood sugar(mg/dl)
17. TG	Triglycerides (mg/dl)
18. LDL	Low density lipo-protein(mg/dl)
19. HDL	High density lipo-protein(mg/dl)
20. Total chol	Total cholesterol(mg/dl)
21. Target	(No HD) 0 or (HD) 1

Data Analysis

Data analysis is a significant step used to know the relationship between attributes i.e., features with respective to target feature. Data analysis is done in 3 steps.

Univariate Selection

Statistical tests are done to identify best feature that hav e the good relationship with performance variable i.e., ou tcome. TABLE 3 showed as the example of univariate sel ection for TABLE 1 from 2.

TABLE 3 describes the scores of features in sorted order i. e., from high to low.

TABLE 3. Univariate Analysis using Cleveland Dataset

Index(as per dataset)	Feature names	Scores(In sorted order)
4	thalach	650.608493
6	oldpeak	253.653461
8	ca	210.625919
5	exang	130.470927
1	chol	110.723364
0	trestbps	45.974069
7	slope	33.673948
9	thal	19.373465
3	restecg	9.739343
2	fbs	1.499550

The above table shows the top 10 features of dataset which have best relationship with target variable.

Feature Importance

To know the importance of features individually of taken dataset using "Model Characteristics" method. "Feature value" produces the score of every function of our produced results, the increase in score the most important or appropriate the performance variable is i.e., result. The below figure shows the feature importance for TABLE 2 from 1.

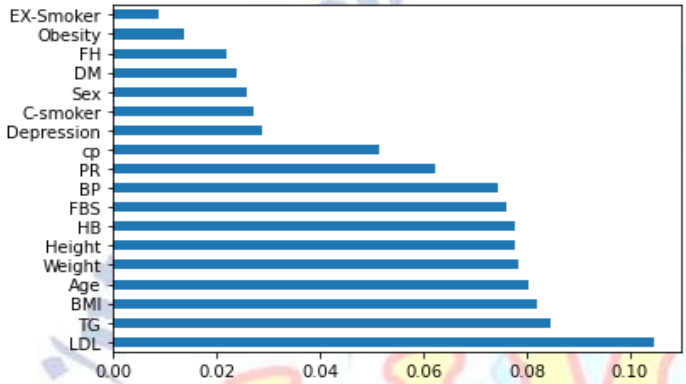


FIGURE 1. Feature Importance using Dataset Alizadeh Sani

Correlation Analysis

Correlation displays either the attributes of dataset are interconnected to each other or to the target variable. Correlation can be +ve means increase in one value, the value of the objective variable increases or -ve means increase in one value, the value of the target variable decreased.

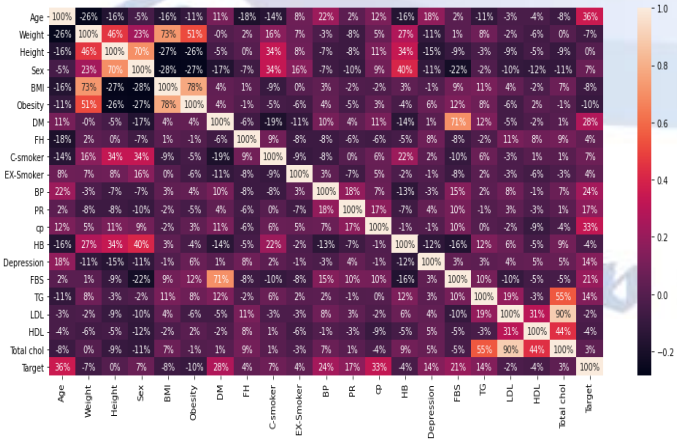


FIGURE 2. Correlation Matrix using Dataset Alizadeh Sani

TABLE 4. Data Analysis Comparison for Both Datasets

Datasets	Univariate selection	Feature importance	Best split
For alizadeh sani	TG	LDL	Cp
For Cleveland	Thalach	Ca	Cp

Data Visualization

In this part we create or plot graphs like count plots, dist plots.

For example :- At what age does the heart disease would be more can be calculated using TABLE 2.

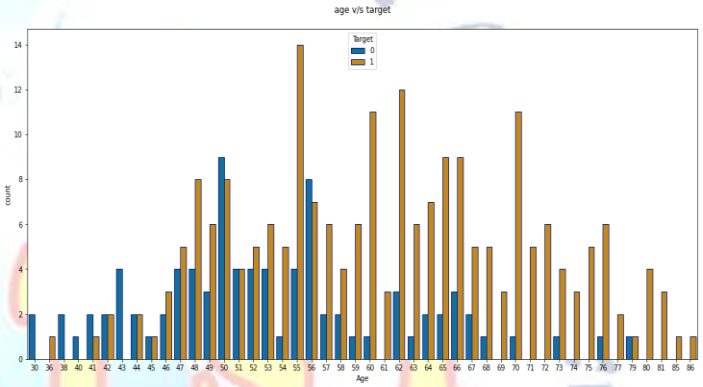


FIGURE 3. At What Age does the Heart Disease would be more for Alizadeh Sani Dataset

So, from FIGURE 3 we can say that occurrence of heart disease is more for 55 aged and no occurrence of heart disease is high for 50 aged.

Data Pre-processing

In this step we create risk a custom column to find patient risk level

For Alizadeh sani dataset we calculate risk as below

$$Risk=(Total\ chol)/HDL \quad (1)$$

For Cleveland dataset we calculate risk level using risk factors cp, chol, frestbp. We also create custom columns for Cleveland dataset namely risk_factor_chol, risk_factor_cp, risk_factor_bloodpressure . The below figure shows bar plot between Target and risk_factor_chol.

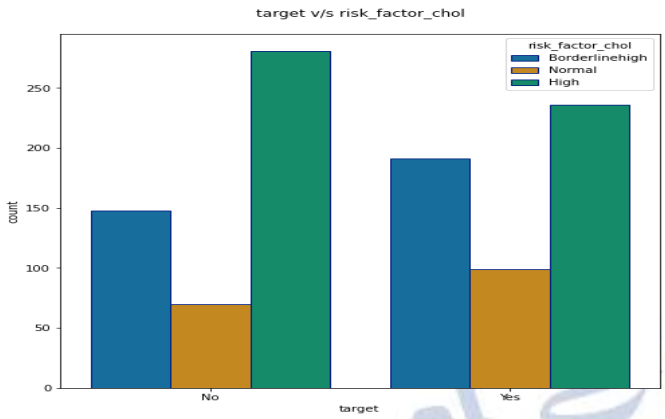


FIGURE 4. Bar Plot for Target v/s risk_factor_chol for Cleveland Dataset

The risk for male should be less than 5.0 and for female it must be less than 4.4. The risk for female greater than 4.4 is obtained in the form of excel sheet as a list "female_risk.xlsx" , for male as "male_risk.xlsx"

Apply Classification models

The classification models applied in our research are SVM, Decision tree, Random forest and KNN is explained clearly in Classification section(3)

Performance Analysis

This is done using Performance metrics like accuracy score, precision, recall, f-measure shown in section(4).

Classification Score Comparison

We compare classification scores of different classification models for values or estimators or functions as shown FIGURE 9 and FIGURE 10.

3. CLASSIFICATION MODELS

The classification models we used is KNN, Random Forest, SVM, Decision tree. The input dataset is split into 70% training set and 30% test set. Training dataset is the dataset which is used to train the model. Testing dataset is used to verify the performance of trained model.

K-Nearest Neighbor

KNN is a primitive level of classification technique can be applied deliberately in many studies, especially when there is only less or no details(not well decrypted) about the data distribution . It is a unsupervised algorithm, means that KNN does not make presumptions about distribution of data used in analysis. It is unsupervised learning algorithm. The main need of KNN algorithms is for example, if there is a new point to be placed in either category 'A' or category 'B' then we check for the nearest neighbor to the new point and the point which has less distance with new point is selected i.e., the point to be placed in that category. The distance can be calculated using Euclidean distance.KNN also called a lazy algorithm, or it only uses quick training phase. Euclidean distance is calculated in KNN classifier to know the similarities between training and test data. We got highest classification score at k=5 of 71.4% which means we have 5 nearest neighbors to a particular new point and the new point is placed in a particular class by calculating Euclidean distance between the new point and five neighbors. The reason we achieved highest accuracy for k =5 is due to similar cluster values in our dataset.

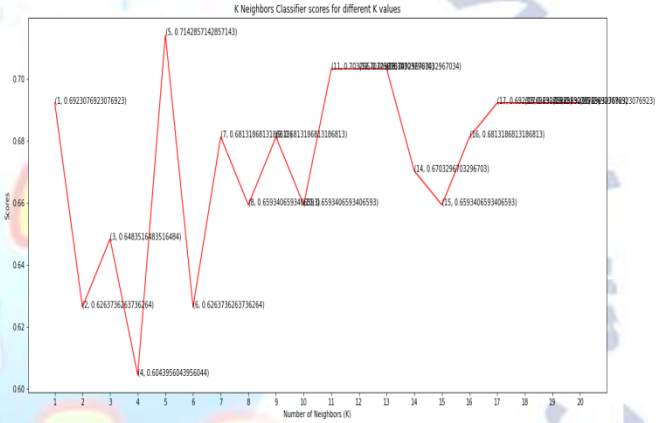
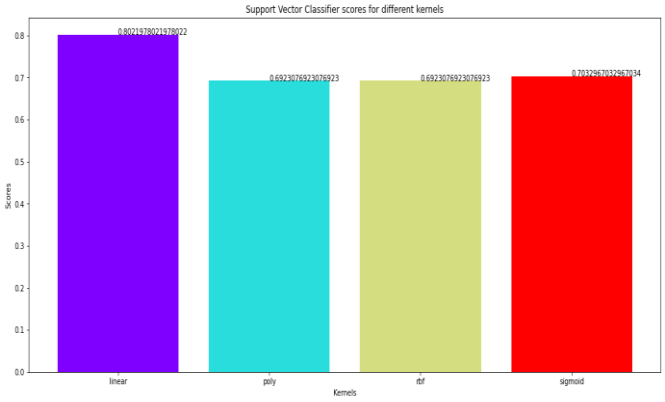


FIGURE 5. Accuracy Scores for each k-value using Cleveland Dataset

Support Vector Machine

A support vector machine(SVM) is a classification model that is used to analyze and find patterns in data during classification and regression analysis. Support vector machine (SVM) is only used for the data which doesn't exceeds more than two classes. An SVM classifies dataset by finding the best "hyper plane " that deviates all data points of one class from those of the other class. If there is higher margin between the 2 classes or clusters then it would be stated as better model. There must be a gap between margin and interior region i.e., no points must be there. SVM is based on mathematical functions and can be applied on various complex models and real-world problems. SVM can be used on datasets which contains more attributes i.e., atleast 10. Support Vector Machines(SVM) map the "training data" into " kernel space". There are many differently used kernel spaces like "linear (uses dot product)",

"quadratic, polynomial", "Radial Basis Function kernel". We got highest classification score linear function of 80.2%



FUGURE 6. Accuracy Scores for Dataset for each Function using Cleveland Dataset

Random Forest

The random forest combines multiple classifiers to solve complex problem, solves both classification and regression problems. It is capable of handling large datasets with high dimensionality. Random forest is supervised learning technique uses both classification and regression properties. The name itself suggests that it contains number of decision trees constructed on subsets of dataset to improve accuracy. In random forest we select attributes randomly in order to find best split at each node. The more number of trees solves the problem overfitting. This model is different from other tree models i.e., finds the best split at each node based on taken attributes. The predictions obtained from each tree must be at low correlation. Further, new values are predicted by combining the predictions of many constructed decision trees. We achieved highest classification score at estimator 100 of 79.12%.

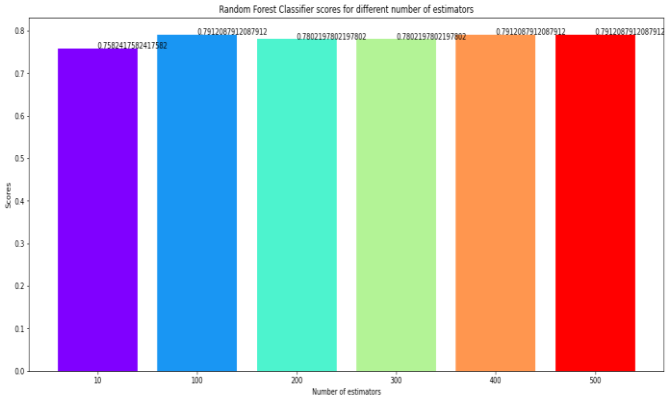


FIGURE 7. Accuracy Scores for each Estimators using Cleveland Dataset

Decision Tree

A decision tree is also a supervised learning algorithm which also solves both classification and regression problems but mostly used for solving classification algorithms. It is tree structured classifier , where internal nodes are considered as features of dataset and branches represents decision rules and leaf node as outcome. In order to built a tree we use CART algorithm. It is used to solve decision related problems.We achieved highest classification score at value 17 of 79.12%

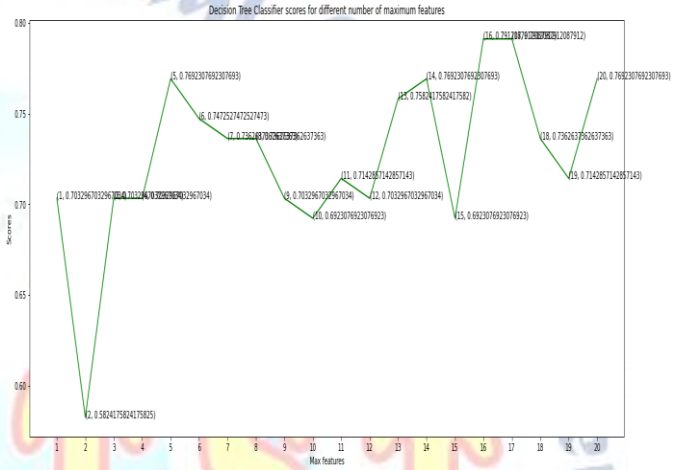


FIGURE 8. Accuracy Scores with Respective to Max Features using Cleveland Dataset

4. RESULTS

The outcomes retrieved by performing "Random Forest", "Decision Tree", "KNN" and "SVM "are shown in this section. The metrics used to perform performance analysis on the algorithms are Accuracy score, Precision (P), Recall (R) and F-measure.

Precision

Gives the correctly identified as positives out of all predicted as positives. The formula is shown below.

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

Recall

Gives the correctly identified positive out of all positives. The formula is shown below.

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

F-Measure

Harmonic mean of model's precision and recall. Harmonic mean means if either one of the parameters

i.e., precision and recall is 0&1 or1&0 then the precision or recall no longer valid for f-measure. The formula is shown below.

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

Confusion Matrix

The above named performance metrics are obtained by calculating confusion matrix . Confusion Matrix describes the performance of the model. The structure of confusion matrix would be as follows.

	Positive(1)	**Negative(0)**
Positive(1)	TP	FN
Negative(0)	FP	TN

TP (True positive):-The patient has the disease and the algorithm predicted correctly.

FP (False positive):-The patient does not have the disease but the algorithm predicted it wrongly.

TN (True negative):-The patient does not have the disease and the algorithm predicted it correctly.

FN (False negative):-The patient has the disease but the algorithm predicted it wrongly.

The below tables show the values obtained by constructing confusion matrix for different algorithms:

TABLE 5. Confusion Matrix for Train Data for TABLE 2

Algorithm	**True Positive(TP)**	**False Positive (FP)**	**True Negative(TN)**	**False Negative (FN)**
Decision tree	153	0	59	0
KNN	151	46	13	2
SVM	145	58	1	8
Random forest	153	0	59	0

TABLE 6. Confusion Matrix for Test Data for TABLE 2

Algorithm	**True Positive(TP)**	**False Positive (FP)**	**True Negative(TN)**	**False Negative (FN)**
Decision tree	53	12	16	10
KNN	63	28	0	0
SVM	62	26	2	1
Random forest	57	13	15	6

TABLE 7. Confusion Matrix for Train Data for TABLE 1

Algorithm	**True Positive(TP)**	**False Positive (FP)**	**True Negative(TN)**	**False Negative(FN)**
Decision tree	379	0	338	0
KNN	290	90	248	89
SVM	225	181	157	154
Random forest	379	0	338	0

TABLE 8. Confusion Matrix for Test Data for TABLE 1

Algorithm	**True Positive(TP)**	**False Positive (FP)**	**True Negative(TN)**	**False Negative(FN)**
Decision tree	143	1	160	4
KNN	107	48	113	40
SVM	99	99	62	48
Random forest	147	4	157	0

Performance Analysis

TABLE 9. Performance Analysis for Z.Alizadeh Sani Dataset

Metrics	**Precision(%)**		**Recall(%)**		**F-measure(%)**		**Accuracy score(%)**	
Algorithms	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Decision tree	100	83.60	100	80.95	100	82.25	100	75.82
KNN	76.49	69.23	98.69	100	86.28	81.81	77.35	69.23
SVM	71.42	70.45	94.77	98.41	81.46	82.11	68.86	70.32
Random forest	100	81.42	100	90.47	100	85.71	100	79.12

TABLE 10. Performance Analysis for Cleveland Dataset

Metrics	Precision(%)		Recall(%)		F-measure(%)		Accuracy score(%)	
Algorithms	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Decision tree	100	99.30	100	97.27	100	98.28	100	98.37
KNN	76.31	69.03	76.51	70.86	76.41	72.78	75.03	71.42
SVM	55.41	50.00	57.32	67.34	59.36	57.39	53.27	52.27
Random forest	100	97.35	100	100	100	98.65	100	98.72

TABLE 11. Highest Classification Scores Obtained by using Machine Learning Algorithm

Models	Decision tree	KNN	SVM	Random Forest
Classification score for Alizadeh sani dataset	79.12%	71.42%	80.21%	79.12%
Classification score for Cleveland dataset	99.68%	98.05%	81.49%	98.70%

Classification Score Comparison

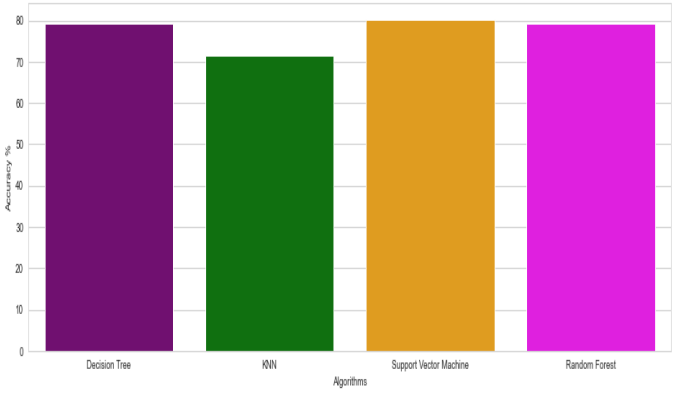


FIGURE 9. Classification Score Comparison using Alizadeh Sani Dataset

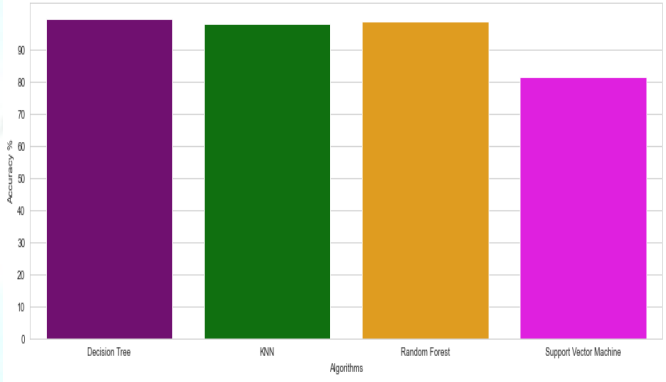


FIGURE 10. Classification Score Comparison using Cleveland Dataset

5. CONCLUSION

With the raise in count of demises because of "heart diseases", it is compulsory to develop system to identify heart disease effectively and accurately .This study perform performance analysis on different classification models like Decision tree , Random forest , SVM, KNN . Basically to select a model accuracy score metric is used when true positives and true negative are more important and f-measure is used when target is predicted falsely i.e., false negative , false positive. In other researches classification models are selected based on highest classification score but in our research we conclude that it is more important to know how falsely predicted than correctly. So, based on f-measure we select the model and the result of study says that for dataset "Z.Alizadeh sani " the f-measure is more for **" Random forest"** with 100% for train set and 85.7% for test set, classification score is 79.12%.For dataset "Cleveland " f-measure is more again for **"Random forest"** with 100% for train set and 98.6% for test set, classification score is 98.70%

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

[2] https://www.kaggle.com/ronitf/heart-disease-uci

[3] H. Benjamin Fredrick David and S. Antony Belcy, "Heart disease prediction using data mining techniques" ICTACT journal ,October 2018

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

[5] M.A.Jabbar, B.L.Deekshatulu and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing ISSN 2160-2174 Volume 4 (2016)

[6] 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

[7] 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)

[8] 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

[9] S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita, "Prediction of Heart Disease using Data Mining Techniques", Department of CSE, SRM University

[10] Vishal Jadhav, Devendra Ratnaparakhi, Tusdhar Mahajan, "iDiagnosis -The Intelligent Medical Diagnostic System", (July 2019)

[11] Deeanna Kelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal

[12] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Out-come Success Prediction using Baseline and First Month Interven-tion Data" in IEEE Journal of Biomedical and Health Informatics