Heart Disease Prediction Using Random Forest Algorithm

Kompella Sri Charan[1], Kolluru S S N S Mahendranath[2]

[1]K. Sri charan,4-177, Main Road, Amalapuram, Andhra Pradesh
[2]K.S S N S Mahendranath, Nellore, Andhra Pradesh
[3]Assistant Professor M.Thirunavukkarasu, Dept. of CSE Engineering, SCSVMV University, Tamilnadu, India

---

Abstract - *Heart disease is one of the most common and serious problems arising in the world today. nowadays there are many ways to solve this problem, one of the finest ways to solve this problem is through using machine learning. we know that in recent years machine Learning used in many platforms. Machine learning will help in predicting and making decisions from the large amount of data produced by healthcare industries and hospitals. We have also seen Machine Learning techniques are being used in many fields in different areas. In this paper, we discovered a new method that will help in finding significant features by applying machine learning techniques that results in improving the accuracy in the prediction of cardiovascular disease. The prediction model will contain different types of machine learning algorithms. By using random forest with a linear model, we get 92% of accuracy.*

Key Words: Cardiovascular Disease Prediction, Machine Learning Techniques, Random Forest Algorithm.

1. INTRODUCTION

Nowadays heart disease prediction is one of the major advantages of using Machine Learning. The main theme of the paper is to predict heart disease based on heart rate and blood pressure. The predicted model will take some inputs from the user and predict whether the person contains heart disease or not, by applying different kinds of machine learning algorithms. In this paper, we will choose the best model based on the other five models' performance and the best will predict the final output.

2. PROJECT DESCRIPTION

2.1 EXISTING SYSTEM

In the following system, inputs are collected from the patients, and using some machine learning techniques the results will be obtained. The data of patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared with other models. by using this we will get an accuracy of 86%.

DISADVANTAGES

1. Interpretation of Result.

2. Data Acquisition

2.2 PROPOSED SYSTEM

In this system, we take data from the patients and we will use different types of ml techniques to obtain results like data pre-processing, feature scaling, model building, one hot coding. In this we will use different types of python modules like pandas, NumPy to visualize the data. We use five different models and take the best one with high accuracy to obtain our final result.

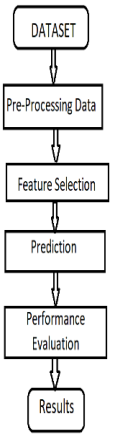
ADVANTAGES

1. Increased Accuracy

2. Low time and cost-friendly

3. PROJECT DESIGN AND ANALYSIS

3.1ARCHITECTURE



APPROACH

3.2 DATA PRE-PROCESSING

The data present in the data set will be pre-processed. The data set contains a total of 270 patient records and some records contain some missing values. Those records will be removed or replaced and the other records will undergo pre-processing

Data columns (total 14 columns):

#	Column	Non-Null Count	Datatype
0	age	270 non-null	int64
1	sex	270 non-null	int64
2	cp	270 non-null	int64
3	trestbps	270 non-null	int64
4	chol	270 non-null	int64
5	fbs	270 non-null	int64
6	restecg	270 non-null	int64
7	thalach	270 non-null	int64
8	exang	270 non-null	int64
9	oldpeak	270 non-null	float64
10	slope	270 non-null	int64
11	ca	270 non-null	int64
12	thal	270 non-null	int64
13	target	270 non-null	int64

3.3 FEATURE SELECTION AND REDUCTION

Among 13 attributes, the attributes which are used to identify the personal information are removed like age, sex and the remaining attributes are considered as they are important in finding the heart disease.

3.4 CLASSIFICATION MODELLING

Grounded on variables, clustering was done and criteria of Decision Tree features. Also, the classifiers are applied to each clustered dataset to estimate its performance. The best-performing models are linked from the below results grounded on their low rate of error

ALGORITHM

3.5 SUPPORT VECTOR MACHINE

SVM can be used for both regression and classification challenges. It is a supervised machine learning algorithm. It is widely used in classification problems. In this algorithm, we point to each data item as a point in n-dimensional space.

These are simply the coordinates of individual observation

SVC

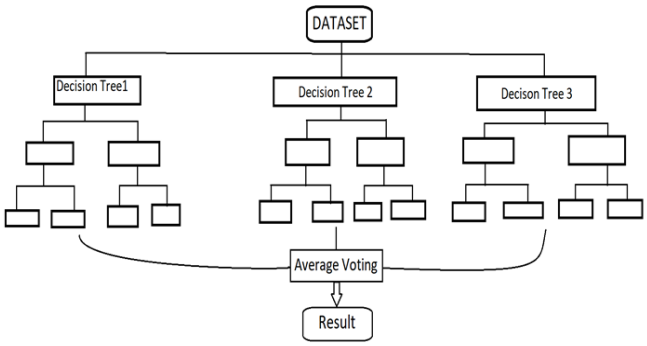
```
classification_report :
              precision    recall  f1-score   support

           1       0.71      0.90      0.79        30
           2       0.81      0.54      0.65        24

    accuracy                           0.74        54
   macro avg       0.76      0.72      0.72        54
weighted avg       0.76      0.74      0.73        54


confusion_matrix :
[[27  3]
 [11 13]]
```

3.6 RANDOM FOREST

In machine learning, random forest is the most powerful and widely used algorithm. It belongs to supervised machine learning. It is used for both classification and regression problems in Machine learning. The process of random forest is:

- It collects the information
- It builds decision trees on different samples
- It takes the average of the decision trees

It can handle the dataset containing categorical variables but compared to a single decision tree it is slower. Doesn't handle missing values.



GradientBoostingClassifier

```
classification_report :
              precision    recall  f1-score   support

           1       0.79      0.73      0.76        30
           2       0.69      0.75      0.72        24

    accuracy                           0.74        54
   macro avg       0.74      0.74      0.74        54
weighted avg       0.74      0.74      0.74        54


confusion_matrix :
[[22  8]
 [ 6 18]]
```

```
RandomForestClassifier

classification_report :
              precision    recall  f1-score   support

           1       0.76      0.73      0.75        30
           2       0.68      0.71      0.69        24

    accuracy                           0.72        54
   macro avg       0.72      0.72      0.72        54
weighted avg       0.72      0.72      0.72        54


confusion_matrix :
[[22  8]
 [ 7 17]]
```

3.7 ADA BOOST CLASSIFIER

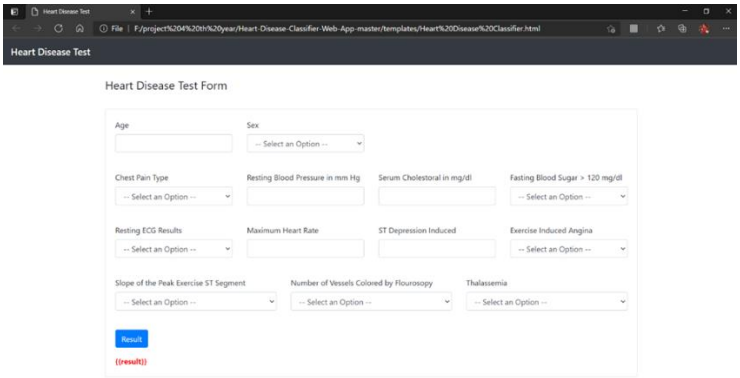
Ada boost classifier algorithm is one of the prominent algorithms used in machine learning.it is used as an ensemble method. The common method used with this algorithm is decision trees. the mechanism of this algorithm works like this, it begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset.

```
AdaBoostClassifier

classification_report :
              precision    recall  f1-score   support

           1       0.75      0.80      0.77        30
           2       0.73      0.67      0.70        24

    accuracy                           0.74        54
   macro avg       0.74      0.73      0.73        54
weighted avg       0.74      0.74      0.74        54


confusion_matrix :
[[24  6]
 [ 8 16]]
```

3.8 GRADIENT BOOSTING CLASSIFIER

Gradient boosting is a classifier algorithm in machine learning. It is a group of algorithms, it collects all the weak algorithms and makes a strong algorithm. In this algorithm mostly decision trees are used. This algorithm is a supervised learning algorithm. It is a combination of decision trees followed by a technique called boosting

3.4 OUTPUT SCREENSHOTS



4. CONCLUSION

With the adding number of deaths due to heart troubles, it has turned necessary to develop a system to read heart conditions effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for the detection of heart diseases. This study compares the accuracy score of Decision Tree, Random Forest, SVM, Ada boost, Gradient Boosting algorithms for predicting heart disease using the UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with an accuracy score of 92.16% for the prediction of heart disease. Random forest gives much more accurate predictions when compared to simple CART/CHAID or regression models in many scenarios. These cases generally have a high number of prophetic variables and a huge sampling size.

REFERENCES

[1]Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, Heart Disease Prediction Using a Combination of ML, Volume7,Publish Year: 2021.

[2]Mohd Faisal Ansari, Bhavya Alankar, HarleenKaur,Heart Disease Prediction Using a Combination of ML ,Volume -6,Publish Year: 2020.

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

[4] A. S. Abdullah and R.R.Rajalaxmi, ``A data mining model for predicting the coronary heartdisease using random forest classi_er,'' in Proc. Int. Conf. Recent Trends Comput. Methods,Commun. Controls, Apr. 2012, pp. 22_25.

[5] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ``Using PSO algorithmfor producing best rules in diagnosis of heart disease,'' in Proc. Int. Conf. Comput. Appl. (ICCA),Sep. 2017, pp. 306_ 311.

[6] N. Al-milli, ``Back Propagation neural network for prediction of heart disease,'' J. Theor.Appl.Inf. Technol., vol. 56, no. 1, pp. 131_135, 2013.

[7] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika,``Analysis of neural networks based heart disease prediction system,'' in Proc. 11th Int. Conf.Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233_239

[8] P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease usingweighted fuzzy rules,'' J. King Saud Univ.-Computer. Inf. Sci., vol. 24, no. 1, pp. 27_40, Jan. 2012.doi: 10.1016/j.jksuci.2011.09.002.

[9] L. Baccour, ``Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,'' Expert Syst. Appl., vol. 99, pp. 115_125, Jun. 2018. doi:10.1016/j.eswa.2018.01.025.

[10] C.-A. Cheng and H.-W. Chiu, ``An arti_cial neural network model for the evaluation ofcarotid artery stenting prognosis using a national-wide database,'' in Proc. 39th Annu. Int. Conf.. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566_2569.

[11] H. A. Esfahani and M. Ghazanfari, ``Cardiovascular disease detection using a new ensemble classifier,'' in Proc. IEEE 4th Int. Conf. Knowledge.- Based Eng. Innovation. (KBEI), Dec. 2017, pp.1011_1014.

[12] F. DammakL. Baccour, and A. M. Alimi, ``The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,'' in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZIEEE), vol. 9, Aug. 2015, pp. 1_8.

[13] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart disease through neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4, pp. 7675_7680, May 2009. doi:10.1016/j.eswa.2008.09.013.

[14] M. Durairaj and V. Revathi, ``Prediction of heart disease using back Propagation ML P algorithm, '' Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235_239, 2015.

[15] M. Gandhi and S. N. Singh, ``Predictions in heart disease using techniques of data mining, ''in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE), Feb. 2015, pp.520_525.

[16] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ``Prediction of heart disease usingmachine learning,'' in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar.2018, pp. 1275_1278.

[17] B. S. S. Rathnayakc and G. U. Ganegoda, ``Heart diseases prediction with data mining andneural network techniques,'' in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp.1_6.

[18] N. K. S. Banu and S. Swamy, ``Prediction of heart disease at early stage using data miningand big data analytics: A survey,'' in Proc. Int. Conf. Elect., Electron., Commun., Comput.Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256_261.81.