

Heart Disease Prediction Using Machine Learning Techniques

G. Saranya, J. Frankiline Joan, R. Dineshkumar, S. Elakkiyaselvan
 Department of Biomedical Engineering Muthayammal Engineering College
 Namakkal, Tamil Nadu, India

Abstract –Heart disease is one of the most significant problem that is arising in the world today. Cardiovascular disease prediction is a critical challenge in the area of clinical data analysis. Hybrid Machine learning (ML) has been showing an effective assistance in making decisions and predictions from the large quantity of data produced by the healthcare industries and hospitals. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse in predicting heart disease with ML techniques. In this paper, we propose a narrative method that aims at finding significant features by applying machine learning techniques that results in improving the accuracy in the prediction of cardiovascular disease. The prediction model is proposed with combinations of different features and several classification techniques. We produce an enhanced performance level with an accuracy level of 92% through the prediction model for heart disease with the hybrid random forest with a linear model.

Key words – Cardiovascular Disease Prediction, Machine Learning Techniques, Random forest linear model.

I. INTRODUCTION

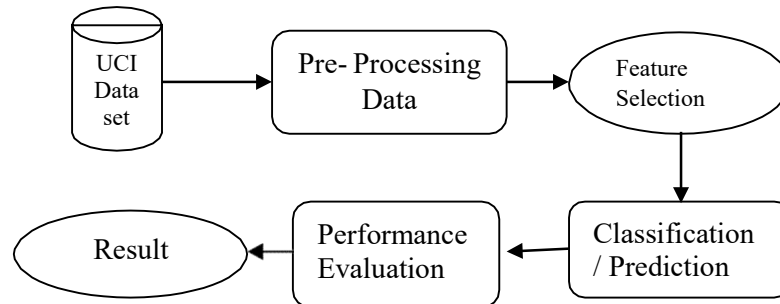
Now a days, heart disease prediction has been a major concept in recent world that is impacting the society towards health. The main concept is to identify the age group and heart rate using the Random forest algorithm. Our project tells how the heart rate and condition is estimated based on the inputs such as blood pressure and many more being provided by the user to a system. This is being much better way when it comes with others algorithms the implementation of RFA gives the better experience and provide accurate result. This helps in early prediction of the disease and is used in many ways, where as it is being provided with the input, in order to find the heart rate based on the health condition.

II. EXISTING SYSTEM

In this system, the input details are obtained from the patient. Then from the user inputs, using ML techniques heart disease is analyzed. Now, the obtained results are compared with the results of existing models within the same domain and found to be improved. The data of heart disease 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 for performance and accuracy with these algorithms. The proposed hybrid method returns results of 87% for F-measure, competing with the other existing methods.

III. PROPOSED SYSTEM

After evaluating the results from the existing methodologies, we have used python and pandas operations to perform heart disease classification for the data obtained from the UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection based on data cleaning, classification of modelling performance evaluation. Random forest technique is used to improve the accuracy of the result.



Feature Selection and Reduction

AGE	Numeric[29 to 77;unique=41;mean=54.4;median=56]
SEX	Numeric [0 to 1;unique=2;mean=0.68;median=1]
CP	Numeric [1 to 4;unique=4;mean=3.16;median=3]
TESTBPS	Numeric [94 to 200;unique=50;mean=131.69;median130]
CHOL	Numeric[126to564;unique=4152;mean=246.69;median=24]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=099;median=1]
THALACH	Numeric [71 to 202;unique=91;mean=149.61;median=153]
EXANG	Numeric [0 to 1;unique=2;mean=0.33;median=0.00]
OLDPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [0 to 3;unique=3;mean=1.60;median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.00]

Among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. Classification Modelling. The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error.

IV. DECISION TREES

For training samples of data D , the trees are reconstructed based on entropy inputs. These trees are simply constructed in a top down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D . Entropy = $-\sum_{j=1}^m p_{ij} \log_2 p_{ij}$

Algorithm for Decision Tree-Based Partition Require:

Input: D dataset – features with a target class for

forall features do

for Each sample

do Execute the Decision Tree algorithm

end for Identify the feature space f_1, f_2, \dots, f_x of dataset UCI.

end for Obtain the total number of leaf nodes l_1, l_2, l_3, \dots ,

In with its constraints Split the dataset D into $d_1, d_2, d_3, \dots, d_n$

based on the leaf nodes constraints. Output: Partition datasets d_1, d_2, d_3 ,

```

from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train,T, y_train,T)

acc = dtc.score(x_test,T, y_test,T)*100
accuracies['Decision Tree'] = acc
print("Decision Tree Test Accuracy (%.2f)%" % format(acc))

```

Decision Tree Test Accuracy 88.33%

LANGUAGE MODEL

For given input features (x_i , y_i) with input vector x_i of data D the linear form of solution $f(x) = mx+b$ equation is solved by subsequent parameters: $m = P$

```

accuracies = {}

lr = LogisticRegression()
lr.fit(x_train.T,y_train.T)
acc = lr.score(x_test.T,y_test.T)*100

accuracies['Logistic Regression'] = acc
print("Test Accuracy {:.2f}%".format(acc))

```

Test Accuracy 86.89%

/opt/conda/lib/python3.6/site-packages/sklearn/linear_model/logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
FutureWarning)

RANDOM FOREST

This ensemble classifier builds several decision trees and incorporates them to get the best result.

For tree learning, it mainly applies bootstrap aggregating or bagging.

For a given data, $X = \{x_1, x_2, x_3, \dots, x_n\}$ with responses $Y = \{y_1, y_2, y_3, \dots, y_n\}$ which repeats the bagging from $b = 1$ to B

Generating the input using python and random forest classification

```

C:\Windows\System32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

E:\Python_19-20_Projects\Heart_Diseases\Module - 1>python preprocess.py
Cleveland data. Size=(302, 14)
Number of missing values
age      0
sex      0
cp       0
trestbps 0
chol     0
fbs      0
restecg  0
thalach  0
exang    0
oldpeak  0
slope    0
ca       4
thal     2
num      0
dtype: int64

Hungary data. Size=(293, 14)
Number of missing values
age      0

```

SUPPORT VECTOR MACHINE

Let the training samples having dataset $Data = \{y_i, x_i\}$; $i = 1, 2, \dots, n$ where $x_i \in R^n$ represent the i th vector and $y_i \in R^n$ represent the target item. The linear SVM finds the optimal hyperplane of the form $f(x) = w^T x + b$ where w is a dimensional coefficient vector and b is a offset. This is done by solving the subsequent optimization problem: $\text{Min}_{w,b, \xi_i} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$ s.t. $y_i, w^T x_i + b \geq 1 - \xi_i, \xi_i \geq 0, \forall i \in \{1, 2, \dots, n\}$

```

svm = SVC(random_state = 1)
svm.fit(x_train.T, y_train.T)

acc = svm.score(x_test.T,y_test.T)*100
accuracies['SVM'] = acc
print("Test Accuracy of SVM Algorithm: {:.2f}%".format(acc))

```

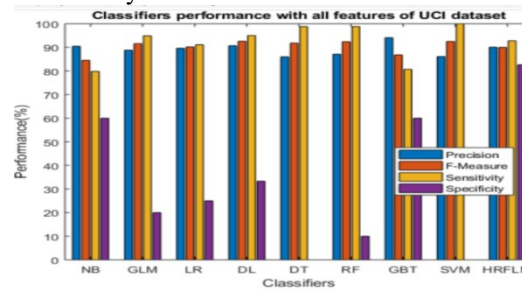
Test Accuracy of SVM Algorithm: 86.89%

/opt/conda/lib/python3.6/site-packages/sklearn/svm/base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.
"avoid this warning.", FutureWarning)

Comparing the obtained input results using Support Vector Machine and Language model classifier:

5. Performance Measures

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficiency of this model.



6. Performance comparison with various models.

V. CONCLUSION

In this paper, we proposed a method for heart disease prediction using machine learning techniques, these results showed a great accuracy standard for producing a better estimation result. By introducing new proposed Random forest classification, we find the problem of prediction rate without equipment and propose an approach to estimate the heart rate and condition. Sample results of heartrate are to be taken at different stages of the same subjects, we find the information from the above input via ML Techniques. Firstly, we introduced a support vector classifier based on datasets.

REFERENCES

- [1] A. S. Abdullah and R. R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.
- [2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, "Using PSO algorithm for producing best rules indiagnosis of heart disease," in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306–311.
- [3] N. Al-milli, "Backpropogation neural network for prediction of heart disease," J. Theor. Appl. Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.
- [4] 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.
- [5] P. K. Anooj, "Clinical decision support system: Risklevel prediction of heart disease using weighted fuzzy rules," J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012. doi:10.1016/j.jksuci.2011.09.002.
- [6] 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.
- [7] C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid 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.
- [8] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," in Proc. IEEE 4th Int. Conf. Knowl. Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.
- [9] F. Dammak, L. 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. (FUZZ- IEEE), vol. 9, Aug. 2015, pp. 1–8.
- [10] 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.
- [11]