International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Heart Disease Prediction Using Machine Learning Techniques

Galla Siva Sai Bindhika[1], Munaga Meghana[2], Manchuri Sathvika Reddy[3], Rajalakshmi[4]

[1,2,3]*Student, Department of Computer Science, R.M.D. Engineering College*
[4]*Assistant professor of Computer Science, R.M.D. Engineering College*

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.

1. 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.

2.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.

2.1 DISADVANTAGES

1. Prediction of cardiovascular disease results is not accurate.

2. Data mining techniques does not help to provide effective decision making.
3. Cannot handle enormous datasets for patient records.

3. 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.

3.1 ADVANTAGES

1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

4. APPROACH

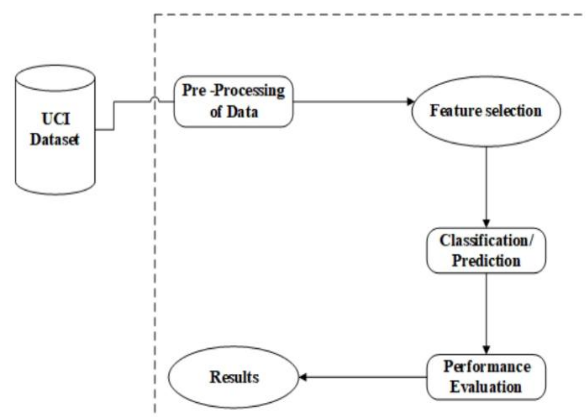
4.1 Data Pre-Processing

Heart disease data is pre-processed by using various collection of records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing.

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;median=130]
CHOL	Numeric [126 to 564;unique=152;mean=246.69;median=241]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=0.99;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]
OLPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [1 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]

4.2 Feature Selection and Reduction

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.



4.3 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.

4.3.1 DECISION TREES

For training samples of data D, the trees are constructed 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 = −Xm j=1 pij log2 pij

Algorithm for Decision Tree-Based Partition Require:

Input: D dataset – features with a target class
for ∀features do
for Each sample
do Execute the Decision Tree algorithm
end for Identify the feature space f1, f2, . . . , fx of dataset UCI.
end for Obtain the total number of leaf nodes l1,l2,l3,. . . ,
ln with its constraints Split the dataset D into d1, d2, d3,. . . , dn
based on the leaf nodes constraints.
Output: Partition datasets d1, d2, d3,.

```
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 80.33%
```

4.3.2 LANGUAGE MODEL

For given input features (xi, yi) with input vector xi 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: FutureWarn
ing: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence thi
s warning.
  FutureWarning)
```

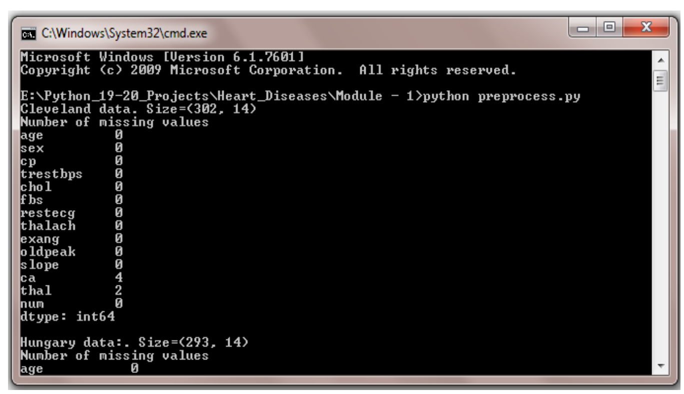
4.3.3 RANDOM FOREST

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 = {x1, x2, x3, ..., xn} with responses Y = {x1, x2, x3, ..., xn} 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
```

4.3.4 SUPPORT VECTOR MACHINE

Let the training samples having dataset Data = {yi, xi}; i = 1, 2, . . . , n where xi ∈ R n represent the i th vector and yi ∈ R n represent the target item. The linear SVM finds the optimal hyperplane of

```
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))
```

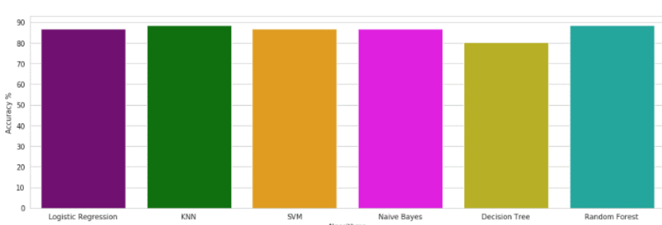
```
Test Accuracy of SVM Algorithm: 86.89%

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

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:

Minw,b, ξi 1 2 w 2 + C Xn i=1 ξi s.t. yi , w T xi + b ≥ 1 − ξi, ξi ≥ 0, ∀i ∈ {1, 2, . . . ,}

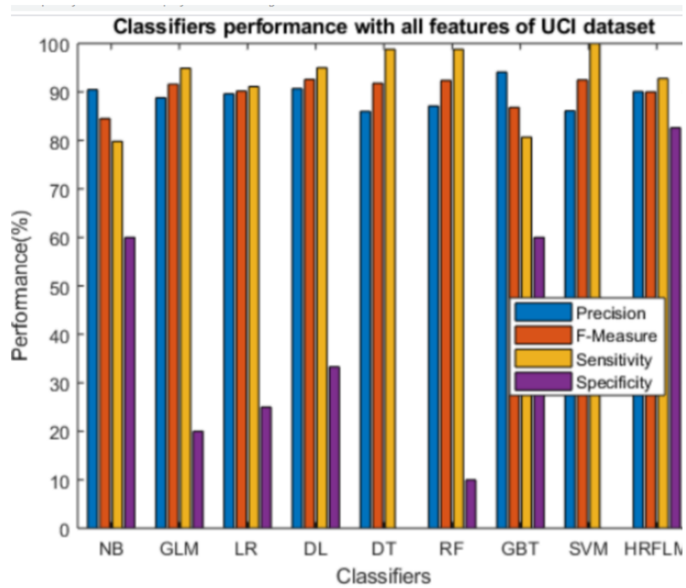
	A	B	C	D	E	F	G	H	I	J
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	num
2	67	1	4	160	286	0	2	108	1	2
3	67	1	4	120	229	0	2	129	1	1
4	37	1	3	130	250	0	0	187	0	0
5	41	0	2	130	204	0	2	172	0	0
6	56	1	2	120	236	0	0	178	0	0
7	62	0	4	140	268	0	2	160	0	3
8	57	0	4	120	354	0	0	163	1	0
9	63	1	4	130	254	0	2	147	0	2
10	53	1	4	140	203	1	2	155	1	1
11	57	1	4	140	192	0	0	148	0	0
12	56	0	2	140	294	0	2	153	0	0
13	56	1	3	130	256	1	2	142	1	2
14	44	1	2	120	263	0	0	173	0	0
15	52	1	3	172	199	1	0	162	0	0
16	57	1	3	150	168	0	0	174	0	0
17	48	1	2	110	229	0	0	168	0	1
18	54	1	4	140	239	0	0	160	0	0
19	48	0	3	130	275	0	0	139	0	0
20	49	1	2	130	266	0	0	171	0	0
21	64	1	1	110	211	0	2	144	1	0
22	58	0	1	150	283	1	2	162	0	0
23	58	1	2	120	284	0	2	160	0	1

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.

6. 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.

7. 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 in diagnosis 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: Risk level 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 ofcriterion 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.