A Cardiovascular Disease Prediction using Machine Learning Algorithms

Rubini P E, Deeksha G S, B Varshaa Shree, Deepa N, Abhinav Srivastava

Abstract*: Heart Diseases have shown a tremendous hit in this modern age. As doctors deal with precious human life, it is very important for them to be right their results. Thus an application was developed which can predict the vulnerability of heart disease, given basic symptoms like age, gender, pulse rate, resting blood pressure, cholesterol ,fasting blood sugar, resting electrocardiographic results, exercise induced angina,ST depression ST segment the slope at peak exercise, number of major vessels coloured by fluoroscopy and maximum heart rate achieved .This can be used by doctors to re heck and confirm on their patients condition. In the existing surveys they have considered only 10 features for prediction, but in this proposed research work 14 necessary features were taken into consideration. Also, this paper presents a comparative analysis of machine learning techniques like Random Forest (RF), Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes in the classification of cardiovascular disease. By the comparative analysis, machine learning algorithm Random Forest has proven to be the most accurate and reliable algorithm and hence used in the proposed system. This system also provides the relation between diabetes and how much it influences heart disease.

***Index Terms*: Heart disease; Machine learning algorithms; Random Forest; Logistic regression; Support Vector Machine; Naïve Bayes; Diabetes Influence**

I. INTRODUCTION

Coronary illness has the biggest level of passing on the planet. In 2012, around 17.5 million individuals kicked the bucket from coronary illness, implying that it comprises of the 31% of every single worldwide passing. Besides, coronary illness loss of life rises each year. It is relied upon to develop more than 23.6 million by 2030. The exploration from the January 2017 demonstrated that the main source of death worldwide is cardiovascular infections. The cardiovascular malady is considered as a world's biggest killer and is currently
taking the top position in the record of ten reasons for passing in the previous 15 years and in 2015 was numeration for fifteen million passing. Various human lives could be spared by diagnosing on schedule. Along these lines, diagnosing the syndrome is significant and an exceptionally muddled undertaking. Mechanizing this procedure would conquer the issues with the diagnosis. The utilization of AI in ailment arrangement is normal and researchers are especially fascinated in the advancement of such frameworks for simpler following and analysis of cardiovascular diseases. Since ML permits PC projects to ponder from information, building up a model to perceive ordinary examples and having the option to settle on choices dependent on assembled data, it doesn't have hitches with the deficiency of utilized medicinal database. The proposed model is to amass significant information relating all components identified with coronary illness and parameters impacting it, train the information according to the proposed calculation of AI and foresee how solid is there a probability for a patient to get a coronary illness. The relationship with the diabetes related credits are considered to set up the impact. [2]

II. METHODOLOGY

The methodology for predicting cardiovascular disease was done by using following four algorithms and the results are compared.Fig.1 describes the architecture diagram for predicting cardio vascular disease.

1. Random Forest
2. Logistic Regression
3. Naive Bayes algorithm
4. Support Vector Machines

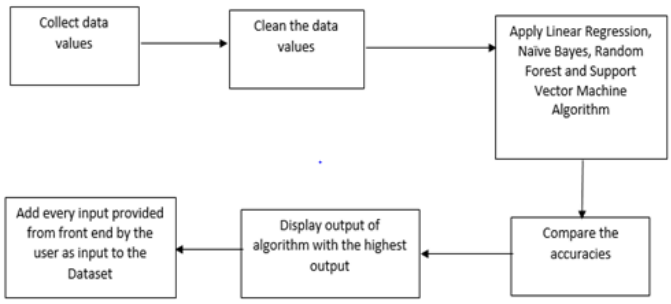

Figure 1:Methodology to predict heart disease

A. Random Forest algorithm

The Random Forest Algorithm is understood as a forest comprised of trees. Firstly, it creates call trees on every which way chosen knowledge samples from the dataset. It then gets the prediction from each tree and selects the most effective resolution through means voting. It is an enhancement from decision

Revised Manuscript Received on June 19, 2019

Mrs Rubini P E, Assistant Professor, Department of Computer Science Engineering, CMR Institute of Technology, Bangalore, India.
Deeksha G S, Department of Computer Science Engineering, CMR Institute of Technology, Bangalore, India.
B Varshaa Shree, Department of Computer Science Engineering, CMR Institute of Technology, Bangalore, India.
Deepa N, Department of Computer Science Engineering, CMR Institute of Technology, Bangalore, India.
Abhinav Srivastava, Department of Computer Science Engineering, CMR Institute of Technology, Bangalore, India.

trees [3]. Some of its applications are image classification, recommendation engines and feature selection. This algorithmic rule is considered as an extremely correct and strong methodology as a result of the number of trees collaborating within the method. One amongst its many advantages is that it does not suffer from the over fitting problem. Finally, it takes the average of all the predictions from every tree, which cancels out the biases.

1) Dataset collection and pre-processing

The dataset which were used for analysis are "Framingham" obtained from kaggle. Heart disease dataset with 14 features are obtained from UCI Machine Learning Repository [19]. Data is cleaned by replacing all the non-available values with the median of values in that column. Categorical data are assigned with numerical values.

2)Implementation

The implementation of random forest works as follows:

a) Load the heart disease dataset.

b) After Preprocess, Split the heart disease dataset into train and test data with the proportion of 60:40 using Random Forest Classifier function.

c) K-Fold Cross Validation is wherever a given knowledge set is split into a K range of sections/folds wherever every fold is employed as a testing set at some purpose.

d) Train the model using train set.

e) Make predictions on the test fold.

f) Map predictions to outcomes (only possible outcomes are 1 and 0).

g) Calculate the accuracy.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FP)} * 100$$

Where,

TP- True Positive (prediction is yes, and they do have the disease.

TN-True Negative (prediction is no, and they don't have the disease.)

FP-False Positive (We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

FN-False Negative (We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

The accuracy obtained by using random forest algorithm is 84.81%

```
predictors=["age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal"]
alg=RandomForestClassifier(n_estimators=75,min_samples_split=40,min_samples_leaf=1)
kf=KFold(heart.shape[0],n_folds=16, random_state=1)
predictions = []
for train, test in kf:
    # The predictors we're using the train the algorithm.  Note how we only take the rows in the train folds.
    train_predictors = (heart[predictors].iloc[train,:])
    #print(train_predictors)
    # The target we're using to train the algorithm.
    train_target = heart["heartpred"].iloc[train]
    #print(train_target)
    # Training the algorithm using the predictors and target.
    alg.fit(train_predictors, train_target)
    # We can now make predictions on the test fold
    test_predictions = alg.predict(heart[predictors].iloc[test,:])
    predictions.append(test_predictions)
# The predictions are in three separate numpy arrays.  Concatenate them into one.
# We concatenate them on axis 0, as they only have one axis.
predictions = np.concatenate(predictions, axis=0)

# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
```

Figure 2:Sample Code of Random Forest

```
# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
i=0
count=0
for each in heart["heartpred"]:
    if each==predictions[i]:
        count+=1
    i+=1
accuracy=count/i
print("Random Forest Result:-")
print("Accuracy = ")
print(accuracy*100)
```

```
Random Forest Result:-
Accuracy =
84.81848184818482
```

Fig3. Accuracy result of Random Forest algorithm

B. Support Vector Machine

1) Introduction

Support Vector Machines is a classification technique which separates data values by the creation of hyper planes. Hyper planes can be of different shapes based on the spread of data, but only those points which help in differentiating between the classes are considered for classification.

Kernel Functions

If data points are in nonlinear fashion, the kernel function make them towards linear decision surface.

Some Kernel functions are as follows:

a) Linear Function: In these kinds of kernel the hyper plane is a straight line. Linear Kernel functions can provide best results for classifiers which have exactly two target classes.

b) Polynomial Function: In such kinds of kernel functions the hyper plane is generally a polynomial like parabola, hyperbola.

c) Radial Basis Function: Radial Basis Function is put in use when points cannot be separated in a linear fashion. The function works to bring points into a shape mostly radial/circular fashion to perform further actions.

2)Implementation

The implementation of Support Vector Machine described as follows:

a)

Load the data sets and clean values, in case of no value for a particular

feature in a row replace with the median value the row from the dataset.

b) Split the data set into train and test in 60:40 ratio respectively.

c) Choosing the Kernel Function as Linear Kernel Function or Radial Basis Function.

d) Applying SVM by first creating a hyper plane with the help of test data set.

i. The train data is taken and both Kernel function namely Linear Kernel Function or Radial Basis Function is applied.

ii. Apply test data set on the trained model.

iii. The model uses hyper plane and finds closest proximity to either class that is having heart disease (yes/1) or not having heart disease (no/0).

e) Calculate the accuracy using

$$\text{Accuracy} = \frac{\text{(Number of data items predicted=actual value in test data set)}}{\text{Total Number of values in test data set}}$$

TABLE 1- Comparison of SVM accuracies with Kernel Functions

Kernel Functions	**Accuracy (%)**
Linear Kernel Function	74.05
Radial Basis Function (RBF)	58.577

In Table 1 the calculation accuracies for both SVM Models with RBF and Linear Function as Kernels are examined.Linear Kernel Function provides higher accuracy than RBF. This is because the problem is a two-class classifier problem. Hence a hyper plane in the form of a line would be the best way to classify such values. In comparison RBF uses a circle as hyper plane thus producing lower accuracy. The hyper plane plot for SVM for predicting heart disease is shown in Fig.4.In this the yellow plot represents patients having heart disease and purple dots represents the patients not having heart disease.

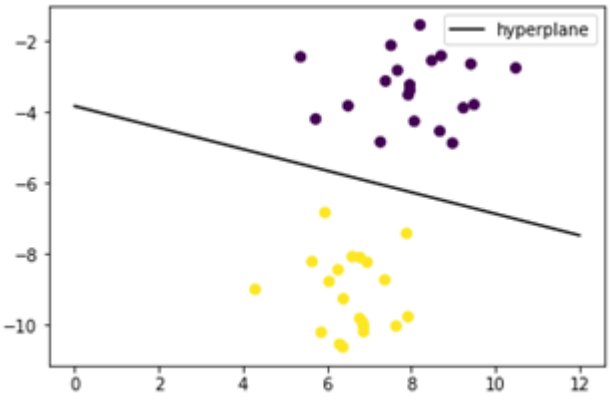

Fig4.Hyper plane and distribution of data points on either side of hyper plane for Heart Disease Prediction

C.Naïve Bayes Classification algorithm

Naive Bayes classifier is based on probability which is mostly used in the training phase. This algorithm is used for removing the redundant data from the datasets.

1) Implementation

The implementation of Naive Bayes is as follows:

a) Extract the dataset.
b) Apply cleaning on the dataset to remove unwanted values.
c) In case any values are missing then find the median value of the column and fill the missing value.
d) Find the deterministic probability with occurrence of heart disease with respect to 14parameters.
e) Then find the conditional probability of non-occurrence of heart disease with respect to 14 parameters.
f) Train the model using this probability formula given below

$$P(W|Q) = \frac{P(Q|W)P(W)}{P(Q)} = \frac{P(Q|W)P(W)}{P(Q|W)P(W)+P(Q|M)P(M)} \quad (1)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

$$P(y|x1,\ldots.x14) = \frac{P(x1|y)P(x2|y)\ldots\ldots P(x14|y)P(y)}{P(x1)P(x2)\ldots\ldots P(x14)} \quad (3)$$

$$P(y|x1,\ldots\ldots.x14) = \frac{P(y)\ \pi P(xi|y)}{P(x1)P(x2)\ldots\ldots P(x14)} \quad (4)$$

$$P(y|x1,\ldots\ldots,x14)\alpha P(y)\pi P(xi|y) \quad (5)$$

Where, x1- age; x2- sex; x3-cp; x4 - restbp; x5- chol; x6- fbs; x7- restecg; x8- thalach; x9 - exang; x10-oldpeak; x11-slope; x12-ca; x13-thal; x14-pulse rate

g) As soon as the model is trained, then apply the test data set.
h) Remove the last column of the test data set which determines the person will have heart attack or not.

i) Apply the model on the test data set and extract the values.

j) Compare the result between the last column and the predicted values.

k)Calculate the accuracy.

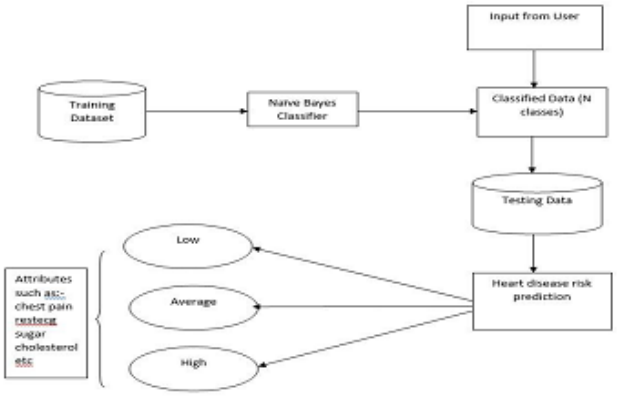

Fig 5. Working of Naïve Bayes

D.Logistic Regression

Logistic regression is a machine learning algorithm used for classification. It is based on the concept of probability. Logistic regression is used to assign observations to a discrete class. Transforming output is done using the sigmoid logic function. The logistic regression hypothesis tends to limit the cost function in range between 0 and 1. Therefore, linear functions can not represent as it can have a value >1 or <=0, which is not possible according to the regression hypothesis

1) Implementation

The implementation steps for logistic regression are given a follow:

a) Obtain the probabilities

Mapping predicted values to probabilities, using the Sigmoid function.

$1/(1+e^{-y})$(6)

Where, y is input to the function and e is the base of natural log

Obtain the probabilities by following equations:

$P = e^y/ 1 + e^y$(7)

here P is the probability of success. The eqn (7) is the Logic Function

q is the probability of failure written as:

$q = 1 – P= 1 – (e^y/ 1 + e^y)$ (8)

where q is the probability of failure

On dividing, (7) / (8), we get

$\frac{p}{1-p} = e^y$(9)

On taking log on both sides,

$$\log \frac{p}{1-p} = y \quad (10)$$

Here (p/1-p) is the odd ratio. When the ‘y’ is positive, the probability success is more than 50%.

b) Decision Boundary-Mapping probabilities to classes

Prediction function returns probability score between 0 and 1. To assign to a discrete class, a threshold value is selected above which it is classified as class 1 or else class 2. For example, if our threshold was 0.5 and our function value was 7, it is classified as positive. For say .3, classification is negative. Logistic regression can also have multiple classes where the highest probability predicted class is considered.

2) Analysis of result

The result can be analysed in following ways.

(a)Using Confusion Matrix: Accuracy is calculated by formula

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FP)} * 100$$

Where,

TP- True Positive

TN-True Negative

FP-False Positive

FN-False Negative

(b)ROC curve: The receiver operating characteristic summarizes the performance when evaluating the compensations between the sensitivity and the 1-specificity. To plot ROC, assume p> 0.5. The area under the curve, indicated as an index of precision or concordance index, is a performance metric for curve. The larger the area under the curve, the better the predictive power of the model.

```
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_yes[:,1])
plt.plot(fpr,tpr)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.title('ROC curve for Heart disease classifier')
plt.xlabel('False positive rate (1-Specificity)')
plt.ylabel('True positive rate (Sensitivity)')
plt.grid(True)
```

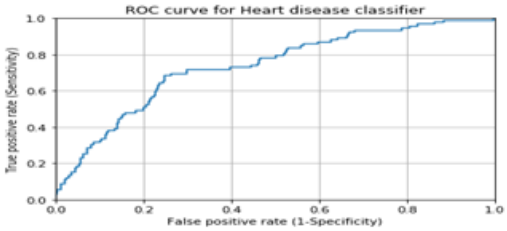

Fig7. ROC Curve - Logistic Regression

```
# Map predictions to outcomes (only possible outcomes are 1 and 0
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
i=0
count=0
for each in heart["heartpred"]:
    if each==predictions[i]:
        count+=1
    i+=1
accuracy=count/i
print("Logistic Regression Result:-")
print("Accuracy = ")
print(accuracy*100)
```

```
Logistic Regression Result:-
Accuracy =
83.82030203820383
```

Fig8. Accuracy result of Logistic Regression

III. RESULT

Results from Random Forest, Support Vector Machine, Logistic Regression and naïve Bayes are analysed, and Random Forest Algorithm has given the highest accuracy. Hence Random Forest has been implemented in the proposed system.

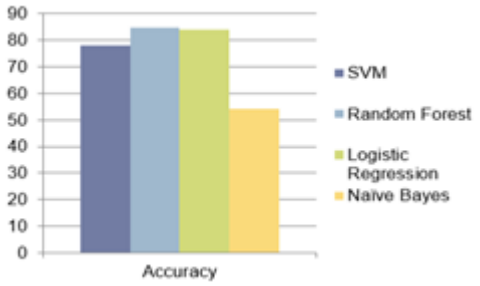

Fig9. Graphical Representation of Accuracy

TABLE II Comparison of Accuracies

ALGORITHM	ACCURACY (%)
RANDOM FOREST	84.81
LINEAR REGRSSION	83.828
SUPPORT VECTOR MACHINE (Using Linear Kernel Function)	74.05
SUPPORT VECTOR MACHINE (Using Radial Basis Kernel Function)	58.577
Naïve Bayes	54.08401

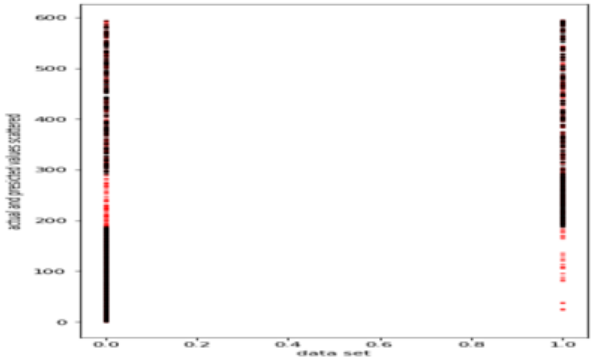

Fig6.Predicted versus Actual Results

IV. CONCLUSION AND FUTURE SCOPE

Heart disease prediction which uses Machine learning algorithm provides users a prediction result if the user has heart disease. Recent advancements in technology made machine learning algorithms to evolve. In this proposed method Random Forest Algorithm was used because of its efficiency and accuracy. This algorithm is also used to find the heart disease prediction percentage by knowing the correlation details between diabetes and heart diseases.The similar prediction systems can be built by calculating correlation between heart diseases and other diseases. Also new algorithms can be used to achieve increased accuracy. Better performance is obtained with more parameter used in these algorithms.

REFERENCES

1. Jaymin Patel, Prof.TejalUpadhyay, Dr.Samir Patel “Heart disease prediction using Machine learning and Data Mining Technique" Volume 7.Number1 Sept 2015-March 2016.
2. (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)
3. Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. International Journal of Engineering Research & General Science, Vol 2(6), pp 6-11, Oct 2014
4. Igor Kononenko” Machine learning for medical diagnosis: history, state of art& perspective" Elsevier -Artificial intelligence in Medicine, Volume23, Aug 2001
5. Gregory F. Cooper, *, Constantin F. Alfieris ”, Richard Ambrosino, John Aronisb, Bruce G. Buchanan, Richard Caruana', Michael J. Fine, Clark Glymour”, Geoffrey Gordon”, Barbara H. Hanusad, Janine E. Janoskyf, Christopher Meek”, Tom Mitchell”, Thomas Richardson”, Peter Spirtes” An evaluation of machine-learning methods for predicting of pneumonia mortality”-Elsevier Feb 1997
6. Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease comparing with different algorithms": Computing, Communication & Automation (ICCCA), 2015InternationalConference.
7. B.DhomseKanchan, M.Mahale Kishore “Study of Machine learning algorithms for special disease predictions using the principal of component analysis” Global Trends in Signal Processing, Information Computing and Communication(ICGTSPICC),2016InternationalConference
8. MatjazKuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, JureFettich" Analysing and improving the diagnosis of ischaemic heart disease with machine learning" Elsevier -Artificial intelligence in Medicine, Volume23, May 1999
9. Geert Meyfroidt, FabianGuiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"-Best practice &ReasearchClinicalAnaesthesiology, ElsevierVolume 23 (1) – Mar 1, 2009
10. Gregory F.Cooper, ConstantinF.Aliferis, Richard Ambrosino”An evaluation of Machine learning methods for predicting pneumonia mortality”-Elsevier, 1997
11. Sanjay Kumar Sen” Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms”- International Journal of Engineering And Computer Science ISSN:2319-7242Volume6Issue 6 June 2017
12. AbhishekTaneja” Heart Disease Prediction SystemUsing Data Mining Techniques”-Vol.6, No(4) December 2013.
13. AnimeshHazra, Subrata Kumar Mandal, AmitGupta,Arkomita Mukherjee and Asmita Mukherjee” Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”- Advances in Computational Sciences and Technology ISSN 0973-6107, Volume10, Number7(2017).
14. BeantKaur, Williamjeet Singh” Review on Heart Diseases Prediction System using different Data Mining Techniques”- International Journal on Recent and Innovation Trends in Computing and Communication Volume:2 Issue:10, October 2014.Transll. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982]. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
15. SonamNikhar, A.M. Karandikar" Prediction of Heart Disease Using different Machine Learning Algorithms"- Vol-2 Issue-6, June 2016.
16. S. U. Ghumbre and A. A. Ghatol, “Heart Disease Diagnosis Using Machine Learning Algorithm,” Advances in Intelligent and Soft Computing Proceedings of the International Conference on Information Systems Design and Intelligent Applications.
17. Machine learning based decision support systems (DSS) for heart diseaseDiagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01
18. DataSetURL-https://archive.ics.uci.edu/ml/machielearnindatabses/heartdiseas