https://doi.org/10.36375/prepare_u.iei.a194

A Machine Learning Approach for Predicting Heart Disease using Efficient Algorithm

Anuradha T (anuradhat@pdaengg.com)
Computer Science and Engineering Department
(PDA College of Engineering, Kalaburagi, 585102)

{Priyanka Vutkur: priyankavutkur@gmail.com}
Computer Science and Engineering Department
(SLN College of Engineering, Raichur, 584135)

Abstract: In our country, there is a high increase in death due to heart disease cases. It's very important to predict such cases beforehand. It is a primary concern as its necessity to safeguard the people to live their lives in secured way so as to take precautions. In modern medical research they have concluded that the over rate of heart disease is because of extreme levels of exercise, i.e peoples mindset is that doing jogging, bodybuilding, running, and exercises leads to good fitness but medical researchers have proved that they may lead to the sudden death of a person at the very early age. It's difficult to predict manually whether a person is undergoing heart disease or not. So to overcome this difficulty we proposed a heart disease prediction system based on the medical history of patients. In this paper, we mainly focus on proving how best the machine learning models can be built to predict even at the worst case to overcome the early death of a patient. In this paper, a novel method that identifies the disease in heart the proposed algorithm by name, Gradient Boosting Classifier algorithm. Our proposed Gradient Boosting Classifier is used to corporate with machine learning models so has to improve the performance of Prediction system by reducing the complexity. The simulated results of the proposed algorithm are compared with that of KNN algorithm. It is observed that the proposed Gradient Boosting Algorithm shows better performance as compared to KNN in terms of Learning rate and decision tree model.

Keywords:- Heart Disease, Confusion matrix, Machine Learning, KNN, Gradient Boosting Classifier.

INTRODUCTION

Data science is a study that combines domain knowledge, programming skills, mathematical and statistical skills to extract meaningful information from data, data can be either in the table format, text, image, audio or video. Machine learning algorithms can be applied to all these types of data to produce artificial intelligence systems to perform a given task. In turn, these systems analyse the task into tangible business values. A heart attack occurs when one or more of your arteries become blocked. India has the highest rate of heart disease worldwide.

A. LITERATURE SURVEY

There are few researchers who have worked on prediction of disease in heart. In [1], the authors have studied that yearly the number of deaths in India has been a rise from 3 million in 1990 to 5 million in 2020. In the rural population it has ranged from 1.6% to 7.4% of Indian population is discussed in [2]. Abdominal obesity, hypertension, tobacco, lack of physical activity, and higher diabetes, are the risk factors of causing CVD, at very young age, than among other ethnic groups [3]. The rates of CVD risk factors have been rapidly rising in Indian urban communities over 25 years is studied in [4]. In [5], authors have analyzed heart disease prediction using more input attributes. Till now the researchers have used 13 input attributes for predicting heart diseases, in this paper, the authors have two added attributes i.e smoking and obesity. Using this dataset they have applied a Decision tree algorithm and KNN achieved the highest accuracy of 87.5%. In [6], they built an intelligent classifier that predicts the heart disease problem this diagnosis component is integrated with the mobile applications with the real-time monitoring component and raises an alarm whenever an emergency occurs. Results of their proposed methods have achieved the highest accuracy of 85% for SVC and Naive Bayes. In [7], researchers used 13 input attributes and 1 output attribute and applied Naive Bayes machine learning algorithms, achieving 86% accuracy. In [8], the authors have reduced 2 input attributes from 13 attributes. The authors used three classifiers like Naive Bayes, J48 Decision Tree, and Bagging algorithm, which has achieved 85% for the bagging classifier algorithm. In [9] they used Naive Bayes classifier for pre-existing data and achieved 74% accuracy. In [10], in this paper experiments have been performed using UIC machine learning dataset and evaluated for all the methodologies, they have used Naive Bayes accuracy of 81% and SVM accuracy of 90 %. It is observed that the authors in [11], achieved 80% using KNN, SVM as 82% and ANN as 84% KNN. It is studied in [12], the authors have achieved 90% using KNN and 88.1% using Navie Bayes.

Table 1. Prediction of disease in heart with different Algorithms in terms of Accuracy

Authors	Accuracy	Techniques
Danger Chaitrali S [5]	87.5 %	KNN
Otoom AF[6]	84.5 % 84.5 %	SVM Naive Bayes
Vembandas amyK[7]	86.4 %	Naive Bayes
ChaurasiaV[8]	84 % 85.03 %	J48 Bagging
Parthiban G[9]	74 %	Naive Bayes
Deepika K[10]	81 % 90%	Naive Bayes SVM
Dwivedi AK [11]	80% 82% 84%	KNN SVM ANN
Devansh Shah1 [12]	90% 88.1 %	KNN Naive Bayes
Proposed models	91% 92.9%	KNN GradientBoostingClassifier

The analysis is carried out using publicly available data for heart disease in kegel. The dataset holds 303 instants with 14 attributes from [13], such as thalach, slope, age, thal, thena, gender, cp, trestbps, chol, fbs, restecg, exang, oldpeak, ca, target, dataset is analyzed with visualization tools, preprocessing, and Machine Learning Algorithms.

It is observed from Table 2.,represents the attribute name and its description for each attribute taken from [13]. The data classification is shown from Fig.1 we can classify the data into ‘Qualitative’ is again classified as Nominal and Ordinal data where as the ‘Quantitative’ data classified as Discrete and Continuous data.

Table 2. List of Various Attributes Used [13]

SL NO	Observations of Attributes	Description	Values
1.	Age	Age in years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	<, or >120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST Segment	Up/Flat/Down

12.	Ca	Gives the number of Major vessels Coloured by Fluroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/ Normal
14.	Num(Disorder)	Heart Disease	Not present/ present in the four major types

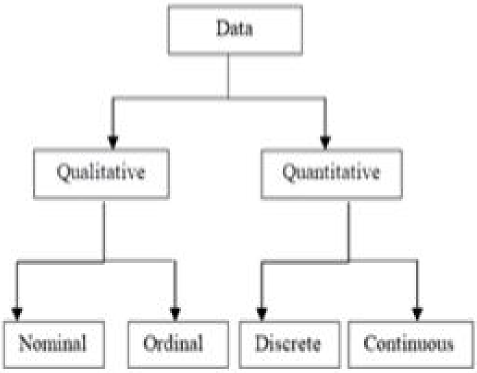


Fig: 1 Data Classification

B. PROPOSED METHODOLOGY

The proposed methodology employs Gradient Boosting Classifier Algorithm for prediction of heart disease in terms of Learning Rate and Decision Trees. In this paper we will discuss two supervised machine learning algorithms, namely, ‘K Nearest Neighbor Classifier’ (K-NN) and Gradient Boosting Classifier.

1. ‘K’ Nearest Neighbour Classifier

This algorithm is quite a simple and very powerful Machine Learning model, representation of the field is done using the training dataset. The predictions of the output values are calculated by checking the complete dataset for K data nodes with the same values, and using the Euclidean number to determine the resulting values. Such a dataset can require lots of space to store and process the data, when there are multiple attributes and have to be constantly curated . However, they very accurately, and efficiently at finding the needed values in a large dataset. The K value indicates the count of the nearest neighbours.

We need to compute the distance between the trained and tested labels. There is no mathematically or statically pre-defined way to find the K value. Initialize a random K value and start computing. Choose the K value which has the minimum error rate. Before applying any machine learning algorithm, split the data into training and testing. Split the dataset into X and Y, where X will be all the attributes present in the dataset leaving the answer key attribute, Y will only answer key attribute. 10% of data is used for testing and the remaining 90% of data is used for training,

KNN Algorithm for different K values:

```
Step 1: Load the training and testing data set.
Step 2: Now, we need to fix a K value for calculating the distance for the whole dataset, K value must be integer.
Step 3: Perform the following steps for each point in the test data:
    3.1- using the euclidean approach, calculate the distance between the value of the testing data and each row of the training data.
    3.2- Sort the calculated values in ascending order based on the distance between them.
    3.3- choose the top K rows from the sorted array
    3.4- Assign a class to the test points based on the most commonly categorised rows.Step 4: Evaluate the KNN module using a confusion matrix.
Step 4: Evaluate the KNN module using Matrix of Perplexity
Step 5: END
```

2.Gradient Boosting Classifier

As the name implies, this method is one of the machine learning algorithms used for categorization. The goal of this approach, Gradient Boosting, is to boost the algorithm's strength by tweaking a poor learning algorithm. PAC is the foundation for this type of learning (Probability Approximately Correct Learning). Gradient Boosting Classifiers apply a similar notion in this learning method, which tells the machine learning problem how complicated it is. The logarithmic loss is frequently used in the gradient boosting process.

Gradient Boosting algorithms frequently use logarithmic loss; however, there is no need to develop a new loss function every time the method is implemented; instead, any differentiable loss function can be applied to the model. There are two parts to the boosting systems: an additive component and a weak learner. In gradient boosting, decision trees are used to deal with weak learning. The additive components of this technique are gradually introduced to the model, and as a result, previous trees cannot be changed and their values remain fixed. By taking the computed loss and applying gradient descent, the error of the provided parameters can be minimised; tree parameters are updated to reduce the loss. The new tree's output is added to the output of the preceding trees in the model. This process is repeated until the loss falls below a particular level. Gradient Boosting Classifier is a useful technique for determining whether or not a patient is sick. There are 303 instances and 14 attributes in the collection. When the k value is 5, the first iteration of the KNN algorithm achieves 83 percent accuracy. When the number of instances is increased from 303 to 918 in the second iteration, we achieve 91 percent accuracy for the same K value. And after 918 instances of experimenting with the Gradient Boosting Classifier method, we were able to reach a 93 percent accuracy.

Gradient Boosting Classifier Algorithm

Step1: Compute the average of the target attribute.

Step2: Compute the residuals.

Step3: Construct a decision tree.

Step 4: Using all of the trees existing in the ensemble, predict the goal value.

Step 5: Subtract the old residuals from the new residuals.

Step 6: Repeat steps 3–5 until the total number of estimators equals the total number of estimators.

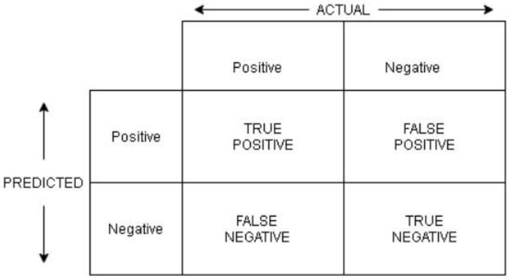
Step 7: Once the trees have been trained, make a final forecast using all of the trees in the ensemble.

C. EXPERIMENTAL RESULTS AND CONVERSATIONS

The suggested algorithm's simulation experiments are carried out with the Jupyter notebook python 3 simulators and the simulation parameters specified in the proposed technique. When compared to the K nearest neighbours classifier, the proposed Gradient Boosting Classifier produced better results. Before proceeding to the experimental results of machine learning algorithms, one of the most essential concepts in machine learning, the confusion matrix, should be discussed.

Matrix of Perplexity

The confusion matrix is also known as the 'error matrix' in machine learning and statistical classifications, and it is specifically in the form of a table arrangement that permits visualisation of an algorithm's performance as illustrated *in* below fig.6



		ACTUAL	
		Positive	Negative
PREDICTED	Positive	TRUE POSITIVE	FALSE POSITIVE
	Negative	FALSE NEGATIVE	TRUE NEGATIVE

Fig.6.Matrix of Perplexity

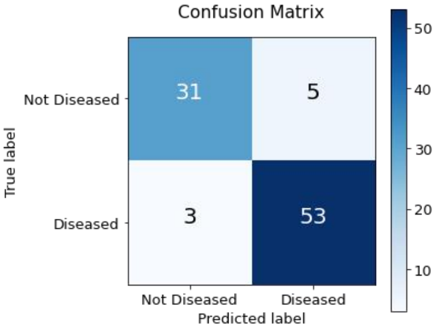
If we plot the predicted values against actual values we get the matrix as shown in the above table.

The matrix depicted in the above table is obtained by plotting anticipated values against actual values.
Let's define 'True positive,' 'False positive,' 'True negative,' and 'False negative,' respectively.
True Positive (TP) data points are those in which the projected values were positive and the actual values were positive as well.
True Negative (TN): These are data points with negative actual values and similarly negative forecasted values.
False Positive (FP): These are data points whose actual values were negative but which were wrongly identified by the algorithm.

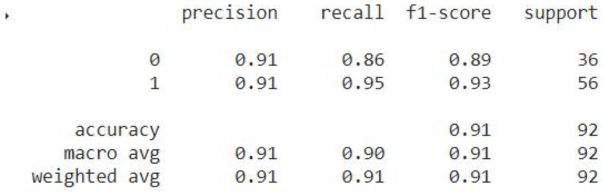
False Negative (FN): These are data points whose true values were positive but were misidentified as negative by the algorithm.
Now let's look at the findings of the algorithm and see which one is the best.

1. K Nearest Neighbour Classifier

Applying the KNN algorithm as mentioned in proposed method it has achieved an accuracy of 83% when the k value is 5 for 303 instances and 14 attributes. As it is less accurate, and tried to improve the better accuracy by increasing the number of instances from 303 to 918 and 14 attributes have drastically increased the accuracy to 91% for the same k value.



7.a. Matrix of Perplexity for KNN

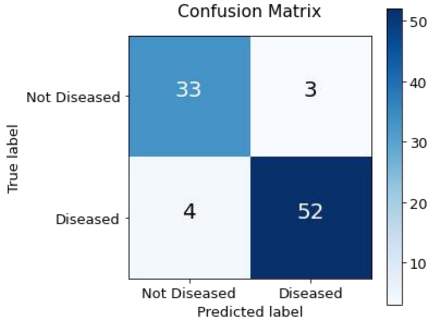


```
              precision    recall  f1-score   support

           0       0.91      0.86      0.89        36
           1       0.91      0.95      0.93        56

    accuracy                           0.91        92
   macro avg       0.91      0.90      0.91        92
weighted avg       0.91      0.91      0.91        92
```

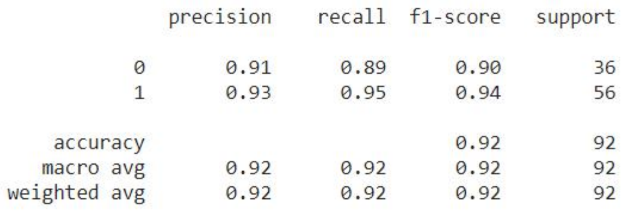
7.b.Classification report

Fig 7: a is a KNN confusion matrix for 918 instances and b is a classification report that explains precision, recall, and f1-score.

Applying Gradient Boosting Classifier as studied in previous sections, achieved the highest accuracy of 92.9% when compared with other algorithms for 918 instances and 14 attributes. The parameters of this algorithm estimators were set to 100, learning_rate = 0.1, and max_depth = 4.



8.a. Matrix of Perplexity (Confusion matrix) of Gradient Boosting Classifier



```
              precision    recall  f1-score   support

           0       0.91      0.89      0.90        36
           1       0.93      0.95      0.94        56

    accuracy                           0.92        92
   macro avg       0.92      0.92      0.92        92
weighted avg       0.92      0.92      0.92        92
```

8.b.Classification report for Gradient Boosting

Fig 8: a represents a confusion matrix of Gradient Boosting Classifier and Fig 8.b represents classification report which explains precision, recall, and f1-score.

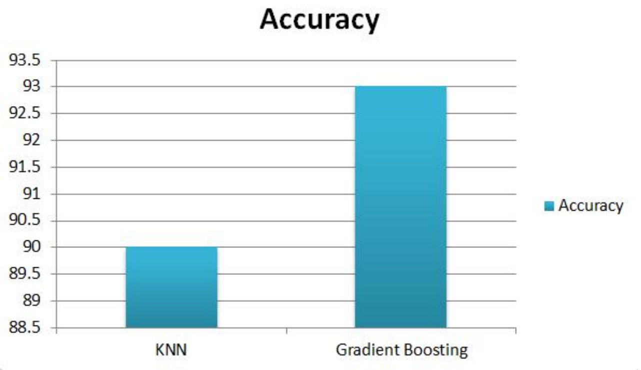


Fig.9 Comparison between KNN and Gradient Boosting classifier

It is observed from fig. 9 we can conclude that Gradient Boosting Classifier has achieved the highest accuracy compared to KNN algorithm. So we prefer the Gradient Boosting Classifier model in terms of Learning rate and Decision tree for prediction of heart disease that gives 93% of accuracy compared to KNN.

D. CONCLUSION

In this paper, we used Machine Learning algorithms to predict whether a person is suffering from heart disease. After importing the data, we analyzed it using histogram plot, barplot for target and age, barplot for grouped data i.e gender versus age concerning the target. We then generated dummy variables for nominal data and ordinal data features and scaled other features. We then split the data for training 90% and 10 % for testing. Then applied Machine Learning algorithms, K Neighbour Classifier for different K values, Gradient Boosting Classifier(parameters with estimators, learning rate, and depth), then varied dataset across each model to improve their scores. In the end, GradientBoosting Classifier achieved the highest score of 92.9% when the estimator value is 100 in terms of learning rate and decision trees.

References

[1] Murray CJ, Lopez AD, “Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *Lancet.* 1997;349:1, pp.498–504.

[2] Gupta R, Joshi P, Mohan V, Reddy KS, Yusuf S. “Epidemiology and causation of coronary heart disease and stroke in India”,.*Heart.* 2008;94:, pp.16–26.

[3] Yusuf S, Hawken S, Ounpuu S, et al. “Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): a case-control study”, *Lancet.* 2004, pp.37–52.

[4] World Health Organization (WHO) India [Accessed April 29, 2010]; National cardiovascular disease database., . Bhargava SK, Sachdev HS, Fall CH, et al. Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. N Engl J Med. 2004;350:8, pp.65–75.

[5] Dangare, Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[6] Ahmed Fawzi Otoom, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye, and Mohammad Ashour “Effective Diagnosis and Monitoring of Heart Disease”

[7] Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.

[8] Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Comput Sci Inf Technol (IJACSIT). 2014;2:56–66.

[9] G. Parthiban, S. Srivatsa. Applying machine learning methods in diagnosing heart disease for diabetic patients. Published 2012.

[10] Deepika K, Seema S. Predictive analytics to prevent and control chronic diseases. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE. p. 381–86.

[11] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.

[12] Devansh Shah1 · Samir Patel1 · Santosh Kumar Bharti1. Heart Disease Prediction using Machine Learning Techniques

[13] Kaggle.com for dataset.