2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.242

Cognitive Approach for Heart Disease Prediction using Machine Learning

[a] Pranav Motarwar, [a] Ankita Duraphe, [b]G Suganya* [c]M Premalatha

[a] UG student, Vellore Institute of Technology, Chennai, Tamilnadu, India
[b] Associate Professor, Vellore Institute of Technology, Chennai, Tamilnadu, India
[c] Assistant Professor(SG), Vellore Institute of Technology, Chennai, Tamilnadu, India
* suganya.g@vit.ac.in

Abstract— **Prediction of patterns to prevent and control diseases is a challenging and a prominent requirement in medical domain. In this paper, we propose a machine learning framework to predict the possibility of having heart disease using various algorithms. The framework is executed using five algorithms Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree (LMT). Cleveland dataset is used for training and testing the model. The dataset is preprocessed followed by feature selection to select most prominent features. The resultant dataset is then used for training the framework. The results are combined and show that Random forest gives maximum accuracy.**

Keywords— Heart Disease, Data Mining, Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, Logistic Model Tree(LMT).

I. Introduction

The Health sector can be modernized using the latest technologies which will increase the life expectancy of the overall population. Leading diseases like Heart Disease, Cancer constitutes major deaths worldwide. Death rate involving Cardiovascular Disease is increasing at an alarming rate every year. World Health Organization report for the year 2016 suggests that, 31% of the deaths worldwide involved cardiovascular disease in which 85% were due to heart attack and stroke. The increasing popularity of alcohol, tobacco in the developing and developed countries directly contributes to the risk of heart- related disease. Rising obesity rates in developed countries like United States, England, Canada, New Zealand leads to the ascending risk of heart-related problems.

Considering the impact of cardiovascular diseases on world population, machine learning model for early detection becomes extremely useful. Constant efforts using different technological advancements are made to deal with this rising gigantic problem. Different bioengineering techniques are developed in recent years to cope-up with the ever-growing health problems. The continued research in the area is proving a boon to the improve reduction rate.

Due to technological advancements, data collection and storage from various medicinal institutions and hospitals around the globe has become possible. The data collected is then analyzed using different machine learning algorithms that discover certain patterns, correlations, and resemblance among each attribute in the dataset.

The first step is to use different data visualization methods to represent the text-based data into a visual format for identifying undetected trends. Next second step is to use feature selection to reduce redundant and trivial data which improves the prediction rate significantly. The third step is to use classification techniques to train the model and predict on the testing dataset. The fourth step is to propose a method for boosting the prediction rate for technique.

II. Objective

1. To demonstrate algorithms like Support Vector Machine, Hoeffding Decision Tree, Logistic Model Tree(LMT), Naïve Bayes, Random Forest.
2. To demonstrate prediction boosting for each machine learning technique.
3. To present the best prediction rate for each model in the final statement.

III. Literature Survey

The Emergence of Artificial Intelligence in the field of health science has encouraged numerous research intended to reduce the death rate by applying different data mining techniques.

Efficient Heart Attack Prediction by extracting significant patterns from the dataset was proposed by Shantakumar B. Patil et al. [1]. K-means clustering algorithm was used. Weightage of each item was calculated using MAFIA algorithm. Based on the calculated weightage, patterns with greater value than the threshold were considered for prediction. Prediction of Heart Disease using a 15 attributes dataset with data mining techniques like ANN, Time Series, Clustering Rules, Association Rules were proposed by Jyoti Soni et al. [2]. Increasing the accuracy by reducing the data size by applying the genetic algorithm was proposed in the paper.

Computer-aided system for diagnosis and prediction was proposed by R. Chitra et al. [3]. Neural Network with preprocessed and normalized data with feature reduction

978-1-7281-4142-8/$31.00 ©2020 IEEE

was considered for heart disease classification in the paper. Research by experimenting with various algorithms like j48, SIMPLE CART and reptree was proposed by Hlaudi Daniel Masethe et al. [4]. The prediction rate is compared and the best method was proposed in the paper.

The widespread data mining classification techniques like ANN, fuzzy logic, Neural Networks, Decision trees, data mining genetic Algorithm, and Nearest Neighbor method were presented by G. Purusothaman et al. [5]. Applied hybrid data mining methods were proposed in the paper. Importance of Big Data Analytics for predicting, preventing and treating chronic diseases were discussed in the paper by Cheryl Ann Alexander et al. [6]. The idea of IoT, cloud computing technologies in the medicinal field was proposed in the paper.

Improving heart attack prediction using feature selection was proposed by Headey Takci et al. [7]. Twelve classification methods and four feature selection algorithms were used for the prediction. Model accuracy, processing time, and ROC analysis were used for comparison. An IoT based application that will work for prediction was proposed by Fizar Ahmed et al. [8]. An effective heart disease prediction system was proposed by Poornima Singh et al. [9]. Algorithms used were MLPNN with backpropogation (BP) algorithm.

IV. PROPOSED METHODOLOGY

The standard dataset, Cleveland is used to classify diseased patients from non-diseased patients. Out of 76 attributes, 13 attributes are used for prediction. The dataset consists of 303 patients. The target attribute is 0 for patients without disease and 1 for patient with disease.

Future work will be intended to collect local hospital dataset for testing the proposed model. The attributes of the dataset can also be changed and the proposed model can be analyzed further. The dataset is visualized with their correlation values and hence the attributes are selected for the process. The following algorithms are used for training the framework.

A. Naïve Bayes

Naive Bayes classifier is a type of probabilistic predictive graphical model. Hybrid method using probability and statistics popularly known as Bayes theorem. It assumes features to be conditionally independent. Mathematically, it is given by:

$$pdf(x, mean, sd) = \left(\frac{1}{sqrt(2*PI)*sd}\right) * exp(-\left(\frac{x-mean^2}{2*sd^2}\right)) \quad (1)$$

where, PI: numerical constant, exp: Euler's number. A special case is known as Gaussian Naive Bayes which is used when the data is in continuous form. It is based on Normal distribution.

B. Support Vector Machine

A classifier for maximizing the margin using hyperplane is known as support vector machine. It has its roots in the field of machine learning which fits both classification and regression.

The training set, taking binary classification into consideration is classified as:

$$T = (xi, yi), i = 1, \dots, N, xi \in \{1, -1\} \quad (2)$$

Where, xi: M dimension feature of ith case yi: class identifier

Maximum margin hyperplane equation:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon i \quad (3)$$

$$s.t\ yi(w^T xi + b) \geq 1 - \varepsilon i, i = 1, \dots., N, \varepsilon i > 0, i = 1, \dots., N$$

Where, w and b: parameters separating hyperplane,

εi: slack variable,
C: importance factor for maximizing the margin

C. Random Forest

A cluster of decision trees is considered as Random forest. Each individual decision tree has a different accuracy rate. The model with the highest prediction rate is selected. It is used to avoid over-fitting and to improve the stability and accuracy.

Random forest randomizes the algorithm instead of the training data and uses averaging which improves the stability and accuracy of the model and also controls over-fitting.

D. Hoeffding Tree

Hoeffding Tree is incremental decision trees that classify based on the concept of Hoeffding bound. Hoeffding bound checks for the number of instances for each attribute to gain a certain level of confidence. Hoeffding bound equation is as follows

$$\epsilon = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2n}} \quad (4)$$

Where, R: random variable,
n: number of examples

Based on the instances, weak classifiers are determined among all the classifiers.

E. Logistic Model Trees

A normal decision tree with logistic regression at each node is defined as the logistic model tree. The model probability is given by:

$$Pr(G = j | X = x) = \frac{e^{Fj(x)}}{\sum_{k=1}^{j} e^{Fk(x)}} \quad (5)$$

where,

$$Fj(x) = \alpha_o{}^j + \sum_{k=1}^{m} \alpha_a{}^j . a \quad (6)$$

α: Set of parameters; a=vk: Input vector

The decision tree and logistic regression are special cases or subset of logistic model trees.

V. ENHANCEMENT METHODS FOR EACH TECHNIQUE

A. Gaussian NB

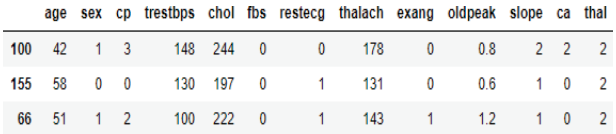
Gaussian NB is based on the concept of Bayes theorem.

$$P(xi|y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp(-(xi - \mu y)^2 / 2\sigma^2 y \qquad (7)$$

Gaussian NB develops a decision boundary using the above equation. Now if the proposed model can get more systematized plotting points(in our case, cp, thal, slope column divided into new columns with each column value ranging from 0 to 3), the decision boundary classifies more accurately.

So in our dataset, the proposed model is dividing columns having short- range of scattered values as a separate column. Values of cp lie in the scope of (0, 1, 2, 3).

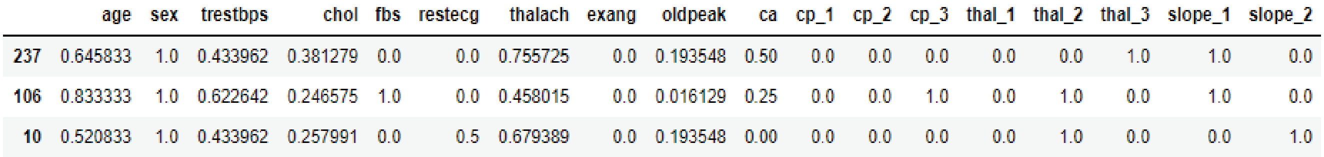
Table 1: INITIAL COLUMNS



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
100	42	1	3	148	244	0	0	178	0	0.8	2	2	2
155	58	0	0	130	197	0	1	131	0	0.6	1	0	2
66	51	1	2	100	222	0	1	143	1	1.2	1	0	2

The values of that lie in scope of (1, 2, 3). Values of slope lies in the scope of (0, 1, 2). Tables 1 and 2 represents the initial and final columns. get_dummies() is used to convert the cp, thal, scope values as a separate column. The prediction rate changes significantly with this technique. In this case, it is increased from 88% to 93%.

.Table 2: FINAL COLUMNS



	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	cp_1	cp_2	cp_3	thal_1	thal_2	thal_3	slope_1	slope_2
237	0.645833	1.0	0.433962	0.381279	0.0	0.0	0.755725	0.0	0.193548	0.50	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
106	0.833333	1.0	0.622642	0.246575	1.0	0.0	0.458015	0.0	0.016129	0.25	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
10	0.520833	1.0	0.433962	0.257991	0.0	0.5	0.679389	0.0	0.193548	0.00	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0

B. Support Vector Machine

The support vector machine (SVM) classifies by a separating hyperplane. The proposed model divides a single column into more specific columns, so that the model get more plotting points, which are perfectly systematized (either 0 or 1 or 2 or 3). As the model have systematized data, the hyperplane equation for maximizing margin is producing better results.

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon i \qquad (8)$$

Increasing the scatter points, the prediction rate of the support vector machine increases significantly. The proposed model divides the columns having short-range of scattered values as a separate column. get_dummies() method is used to convert the cp, thal, scope values as a separate column. The prediction rate changes significantly with this method. For the Support vector machine it increased from 86% to 90%.

C. Random forest

A cluster of decision trees is considered as Random forest. Each individual decision tree has a different accuracy rate. The model with the highest prediction rate is selected. Hyperparameter tuning is used to predict with the best accuracy. Hyperparameter can select the number of decision trees in the random forest or number of features for each tree while splitting a node. In our case, the proposed model used data visualization for finding data discrepancies and then selected the best features using feature selection. Now after this process, model can go for classification. The proposed model used the looping method for tuning the random state hyperparameter. The model came up with the best random state which increased the prediction rate from 89% to 95%.

D. Hoeffding Trees

Hoeffding tree is based on the concept of Hoeffding bound. Hoeffding bound checks for the number of instances for each attribute to achieve a certain level of confidence. This way weak classifiers are collected within the dataset. The proposed model then chooses the AdaBoost method to increase accuracy. AdaBoost divides the training dataset into a number of instances. These instances are then labeled with a weighted value. Error value and therefore stage value is calculated for each instance. So that weak instances are classified.

Now model can go for the Bagging technique. Dataset is divided into sub-samples creating different decision trees. The model is trained from each decision tree and for testing the dataset the average prediction of the decision tree is referred. Now the proposed model select the Blending technique. Dataset is trained using different algorithms. The testing dataset is then predicted on the basis of results for different algorithms trained on training dataset. The proposed model compared the results of all the above techniques and came up with a conclusion of selecting the Bagging technique which increased the accuracy from 79% to 81.96%.

E. Logistic Model Trees

The logistic model tree is derived on the concept of logistic regression and decision tree combined. At each node of the decision tree, the logistic regression model is applied. The proposed model chooses the AdaBoost method to increase accuracy. AdaBoost divides the training dataset into the number of instances. These instances are then labeled with a weighted value. Error value and therefore stage value is calculated for each instance. So that weak instances are classified.

Now model can go for the Bagging technique. Dataset is divided into sub-samples creating different decision trees. The model is trained from each decision tree and for the testing dataset, the average prediction of the decision tree is referred. Now model can select the Blending technique. Dataset is trained using different algorithms. The testing dataset is then predicted on the basis of results for different algorithms on the training dataset.

The proposed model compared the results of all the above techniques and came up with a conclusion of selecting the AdaBoost technique with 80.32% accuracy.

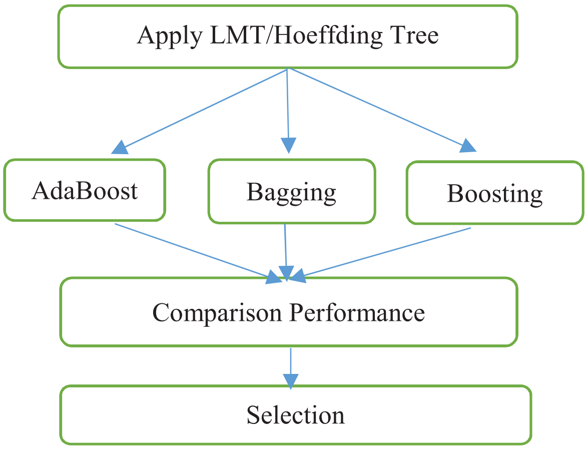


FIGURE 1: PROCEDURE FOR HEART DISEASE PREDICTION

VI. RESULT AND COMPARISON

Data visualization was used on the dataset to visualize correlation or dependency between any of the featured attributes in the following dataset. Feature selection was used to select the best attributes in the dataset. This process presents the highest data quality for performing classification algorithms. Further enhancement techniques are used to increase the basic accuracy of each algorithm. The dataset was trained with 80% data, 242 instances. The rest 20% data, 61 instances are predicted. The following table shows the increased percentage for each technique.

Table 3: Comparison table

Classification Algorithm	Initial Accuracy	Final Accuracy	Increase Percent
Gaussian NB	88.52%	93.44%	4.92%
Support Vector Machine	86.88%	90.16%	3.28%
Random Forest	88.52%	95.08%	6.56%
Hoeffding Tree	78.65%	81.24%	2.59%
Logistic Model Tree	79.12%	80.69%	1.57%

Table 4a: Confusion matrix for Gaussian NB

	C 1	C 0
R 1	29	3
R 0	1	28

Table 4b: Confusion matrix for SVM

	C 1	C 0
R 1	27	5
R 0	1	28

Table 4c: Confusion matrix for Random Forest

	Target 1	Target 0
R 1	27	0
C 0	3	31

Table 4d: Confusion matrix for Hoeffding Tree

	C 1	C 0
R 1	26	7
R 0	4	24

Table 4e: Confusion matrix for Logistic Model Tree

	C 1	C 0
R 1	25	7
R 0	4	28

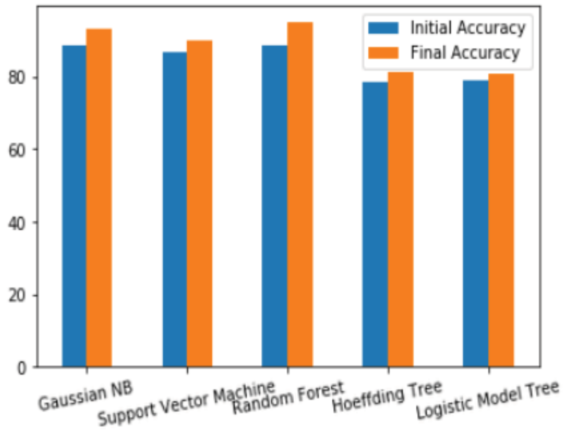


Figure 2: Prediction comparison using graphical representation

VII. CONCLUSION

The prime purpose of this paper is to predict possibilities of heart related disease more accurately. We used techniques like Gaussian NB, SVM, Random Forest, Hoeffding Tree, LMT for effective prediction and hence to increase and boost the accuracy. The performance of each algorithm was analyzed on the cleveland dataset and results were compared with respect to the accuracy. It is further found that Random forest suits better because of its nature to accommodate individual's interest. The future aspect of this paper will be to add more input attributes and analyze the results using proposed models.

REFERENCES

[1] Patil, Shantakumar B., and Y. S. Kumaraswamy. "Extraction of significant patterns from heart disease warehouses for heart attack prediction." *IJCSNS* 9.2 (2009): 228-235.

[2] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.

[3] Chitra, R., and V. Seenivasagam. "Review of heart disease prediction system using data mining and hybrid intelligent techniques." *ICTACT journal on soft computing* 3.04 (2013): 605-609.

[4] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. "Prediction of heart disease using classification algorithms." *Proceedings of the world Congress on Engineering and computer Science*. Vol. 2. 2014.

[5] Purusothaman, G., and P. Krishnakumari. "A survey of data mining techniques on risk prediction: Heart disease." *Indian Journal of Science and Technology* 8.12 (2015): 1.

[6] Alexander, Cheryl A., and Lidong Wang. "Big data analytics in heart attack prediction." *J Nurs Care* 6.393 (2017): 2167-1168.

[7] Takci, Hidayet. "Improvement of heart attack prediction by the feature selection methods." *Turkish Journal of Electrical Engineering & Computer Sciences* 26.1 (2018): 1-10.

[8] Fizar Ahmed,"An Internet of Things (IoT) Application for Predicting the Quantity of Future Heart Attack Patients " International Journal of Computer Applications (0975 8887) Volume 164 No 6, April 2017.

[9] Singh, Poornima, Sanjay Singh, and Gayatri S. Pandi-Jain. "Effective heart disease prediction system using data mining techniques." *International journal of nanomedicine* 13.T-NANO 2014 Abstracts (2018): 121.

[10] Heart Disease dataset David W. Aha (https://archive.ics.uci.edu/ml/datasets/Heart+Disease)