Application of machine learning algorithms in predicting the heart disease in patients

Razia Sulthana A
Department of Computing and Mathematical Sciences
University of Greenwich, Old Royal Naval College
London, United Kingdom
razia.sulthana@greenwich.ac.uk

Jaithunbi A K
Department of Computer Science and Engineering
RMD Engineering College
Kavarapettai, India
akj.cse@rmd.ac.in

Supraja P
Department of Computer Science
SRM Institute of Science and Technology
Kattankulathur, India
suprajap@srmist.edu.in

***Abstract*—Healthcare services save the life of human beings by making timely effective decisions. The use of data mining tools is crucial for decision making, forecasting, and disease prediction. In this study, data mining algorithms are applied to predict heart disease. The dataset contains 14 attributes such as age, gender, blood pressure, blood fat, etc. These parameters are analyzed to predict the probability of patients prone to heart disease in future. Initially, the relationship between the parameters is analyzed. Following which Naïve Bayes, decision trees and Naïve Bayes with k-means clustering are applied over it for classification and prediction. These algorithms were employed to train the dataset and create a binary classification. The proposed system shows a better prediction of heart disease. The performance measures of the system are measured, and the obtained results illustrate the system can forecast the probability of developing the heart diseases.**

***Index Terms*—Heart disease, Data mining, Naïve Bayes, Decision Trees, prediction**

I. Introduction

Cardiovascular ailment is one of the most deadly diseases [1]. This sickness assaults a man immediately leaving no hope for a cure. So diagnosing patients accurately is challenging. Heart diseases can be because of a number of reasons: family history, cholesterol level, hypertension, diet, exercise, etc. Analyzing these factors needs the support of an automated analysis of these patterns over the datasets that can be utilized for clinical determination. Data mining is used to identify the best practices in diagnosing heart disease at an early stage. The steps taken to predict the disease in mining have resulted in the generation of diversified solution patterns. As disease prediction is so sensitive, it demands a huge dataset for study. In general, the datasets for disease prediction are huge and heterogeneous and need effective mining approaches to process them [2].

Existing studies apply statistical techniques for prediction. However, a detailed comparison has to be made between two or more algorithms to understand the distribution of the data and to identify the dependent factors causing heart disease. The paper presents a prediction system for detection heart disease at an early age. Data mining classification techniques are applied to extract hidden knowledge from the database. Naïve Bayes and decision trees are applied for classifying the heart disease of 303 patient records. Attributes in the dataset are analyzed for classifying the dataset and further prediction is made. The train-test operation is carried on the dataset and the accuracy of the model is measure. In addition, the paper proposed a prediction system that aid the doctors in analyzing the patient records using Naive Bayes, decision tree and Naïve Bayes with k-means clustering. The Naïve Bayes identifies the relationship between the attributes using conditional probability rules and decision trees uses entropy as a measure for classifying them. This bidirectional comparison shows significant results in analyzing the dataset. The organization of the paper is as follows. Section 2 describes the literature study, section 3 describes the algorithm used in the proposed system, section 4 the experimental results are given, and finally, the conclusion in section 5.

II. Methodology

The proposed method will incorporate a 3-step approach for both datasets.

A. Data cleaning

Cardiovascular infections are a major health disorder affecting 77% of the population in the world [3]. Numerous elements have to be analyzed to identify the symptoms of the disease. Typically, it needs specialized doctors for diagnosis. As it affects the majority of the population, an effective method at lessened cost could save the lives of all human beings. Several parameters like gender, age, pressure levels, etc. help in predicting the probability of patients getting affected with heart disorders in the future. A significant relation between these factors is studied and the multilayered neural network

with a backpropagation circuit is deployed to train the dataset in [4]. Though the system showed significant results, all the parameters were taken with equal weightage. The Table I details the data mining technique used in few of the reference papers.

Reference no	Data mining technique
[4]	MLP
[5]	Naive Bayes Neural network Decision tree
[6]	ANN
[7]	SOM
[8]	SVM
[9]	K-NN Decision list

TABLE I
THE LITERATURE PAPERS ON HEART DISEASE PREDICTION

Fuzzy techniques and KNN are applied in [10] for disease prediction and resulted in an accuracy of 87.0%. Neural networks and ensemble approach [11] predicts heart disorders with 89 percent accuracy and in [12] NN and genetic algorithm are applied for disease prediction. The experimental results in [13] show 77.0% accuracy in prediction using logistic regression and discriminant function. Decision trees namely C4.5 is applied in [14] to improve the attribute selection properties. On the other hand, few authors have discussed analyzing the textual descriptions written about the patients using rules and ontology [15], [16] and clustering them using the k-means algorithm. Another article [17] applied classification and regression tree (CART), logistic regression, and neural networks for prediction. Clustering being an unsupervised learning technique is applied to group the predicted heart disease patients with others [18]. This resulted in an accuracy of 78.0% over the Cleveland database. A number of researchers have worked on machine learning algorithm for prediction analysis [19]–[22]

III. METHODOLOGY

The Cleveland dataset from the UCI machine learning repository has been used. The inputs from the dataset are fed to the python code and the Naive Bayes classifier and decision tree classifier are applied over the dataset. The training and test split ratio applied for the aforementioned algorithms is 60:40 and 70:30 ratio. The decision tree classifier is implemented in two folds (5 and 10).

A. Dataset Description

The raw database obtained from the repository contains 76 attributes and 303 instances. For the implementation of this project, 14 attributes (numerically valued), and 303 instances have been used.

B. Data Model Specifications

In this paper, Naive Bayes and decision tree classifiers are used to create models for diagnosis. The data is split into train-test based on the splitting technique and cross-validation fold.

1. INPUT
Cleveland database inputted to the python program.
2. OUTPUT
A nearly accurate Naive Bayes classifier and decision tree to identity the patient suffering from heart disease.
3. PROCEDURE
Code the Naive Bayes classifier and decision tree classifier in the python software.
Obtain inputs from the Cleveland database.
Make the 60:40 and 70:30 split for Naive Bayes classification.
Extract the output
Make the 5 and 10 fold for decision tree classifier.
Extract the output
From the outputs received, analyze the results for patients likely to have heart disease or not.

The architecture of the proposed approach is given in Figure 1

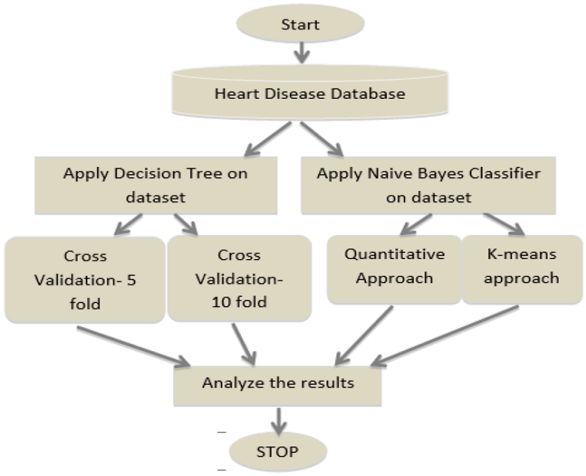

Fig. 1. The architecture of the proposed system

C. Naive Bayes Theorem

Naive Bayes theorem uses a probabilistic approach and assumes independence between the features. This algorithm is scalable and is one of the fastest classifier relying on probability factors. The Bayes procedure applied is given in Figure 2. Bayes algorithm follows a probabilistic grouping of similar data that are closer to each other. It is represented mathematically as

$$p(\frac{hypothesis}{class}) = \frac{p(\frac{class}{hypothesis})*p(hypothesis)}{p(class)}$$

Where $p\,(hypothesis)$ is the prior probability of the hypothesis, p(class) is the prior probability of the training data, $p(\frac{hypothesis}{class})$ is the conditional probability of a hypothesis being true for a given class and vice versa for $p(\frac{class}{hypothesis})$

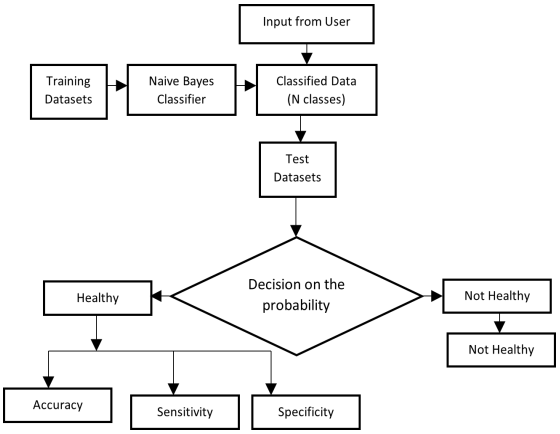

Fig. 2. Naive Bayes procedure applied

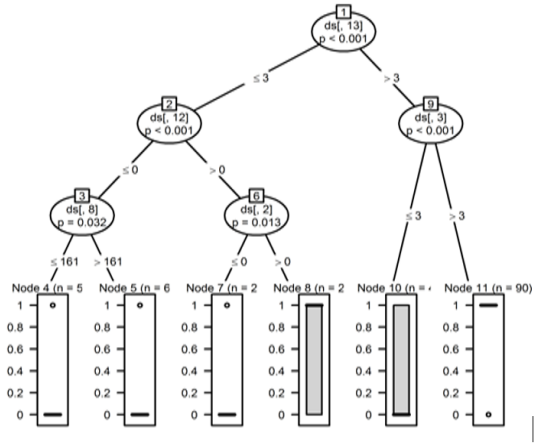

Fig. 3. Decision tree generated for the proposed model

D. Decision Trees

Decision trees are a supervised learning technique that derives a stream diagram after analyzing the dataset. The features contributing much to the decision analysis is opted to be in the top of the tree and the rest flowing in its branches. The tree follows a top-down approach in dividing the dataset into groups related to relevant features. The tree is navigated from root to leaf with intermediary nodes and branches. However, the split of a node into branches is decided by calculating the entropy or other approaches. Finally, the leaf nodes include the grouped tuples that are anticipated to have similar characteristics. The application of decision trees over the dataset in the proposed system identifies the similar history of patients using correlation and groups them together in the training phase. It learns the characteristics of people with heart diseases and without. The homogenous grouped characteristics or symptoms are further applied to predict the heart state of new patients reporting with the same symptoms. Thus predicting the survival probability.

The decision tree is constructed in python. Following which the decision rules are formed. The decision tree is shown in Figure 3. One of the important parameter in calculating the performance of decision tree is gain ratio and is given below. It relates entropy (information gain) and classified information.

$E = -\sum_{i=1}^{n} a_i log a_i$

Here n denotes the variable class count, a_i is the probability of occurrence of event among total events.

E. Results and Analysis

For better understanding, results for each data mining techniques have been shown separately in different tables. Naïve Bayes and decision tree classifiers are executed. From the results, its observed that the classifiers produce different results during different iterations. The statistics of the dataset is given in Table II.

The prediction results of the Naïve Bayes, decision tree, and Naïve Bayes with k-means clustering are given in Figure 4 and the predicted results of the male and female gender are shown in Table III.

Attributes	Age	Trestbps	Chol	Thalach
Average	54.43	131.68	246.69	149.6
Standard deviation	9.03	17.59	51.77	22.87
Minimum value	29	94	126	71
Median	56	130	241	153
Mode	58	120	197	162
Maximum value	77	200	564	202

TABLE II
DATASET'S ATTRIBUTE STATISTICS

It is observed that males are more prone to heart disease than females Figure 5. Out of all the algorithms tested the Naïve Bayes with k-means clustering have the highest performance with efficiency 88.06%. Thus we conclude for this particular data set Naïve Bayes with k-means clustering is the best option. On the dataset of size 303, it is observed in our experiment that Naïve Bayes with the k-means clustering approach took 0.02 seconds of model construction time and the decision tree classifier delivered the results in 0.09 seconds for the same. In the future, this system can be further enhanced

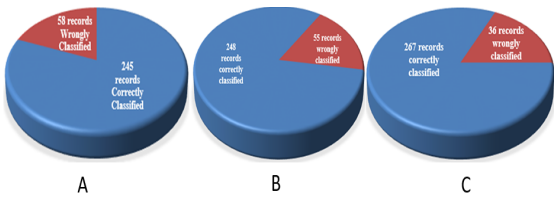

Fig. 4. Pie Graph results of the classifiers

Classifier Type	Accuracy	Model Construction Time (S)
Decision tree	85.85%	0.09
Naive Bayes- Quantitative Approach	81.84%	0.03
Naive Bayes- K-Means Clustering Approach	88.06%	0.02

TABLE III
EXECUTION RESULTS OF ALL THE CLASSIFIERS

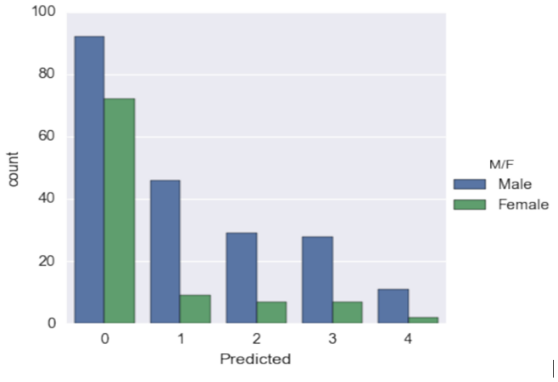

Fig. 5. Distribution of data by spending score

and expanded.

F. Conclusion

Human behavior and lifestyle decide on his health factors. Heart disease is one such illness that is caused by poor lifestyle. The medical industry faces a problem in diagnosing heart disease. Heart disease prediction system based on data mining can assist in determining the heart disease at very early stage. Heart disease risk can be reduced using the proposed prediction system. In the proposed system, the prediction is done with the UCI machine learning repository. Cleveland's heart disease dataset with 14 attributes and 303 instances is the training dataset used for analysis. The supervised algorithms namely: Naïve Bayes, decision tree classifiers and Naïve Bayes with k-means clustering are applied to classify the dataset. From the experimental results, we can conclude that Naïve Bayes with k-means clustering provides better results as compared to the decision tree classifier and Naïve Bayes because the combined classification possesses high precision and less error rate. One of the most appreciable features of using the Naïve Bayes classifier is that it consumes less time than the decision tree classifier. In the future, other ensemble learning algorithms can be applied to classify the heart disease dataset.

REFERENCES

[1] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.

[2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*. IEEE, 2008, pp. 108–115.

[3] K. Y. Le, M. R. Sohail, P. A. Friedman, D. Z. Uslan, S. S. Cha, D. L. Hayes, W. R. Wilson, J. M. Steckelberg, L. M. Baddour, M. C. I. S. Group *et al.*, "Impact of timing of device removal on mortality in patients with cardiovascular implantable electronic device infections," *Heart Rhythm*, vol. 8, no. 11, pp. 1678–1685, 2011.

[4] M. Durairaj, V. Revathi *et al.*, "Prediction of heart disease using back propagation mlp algorithm," *International Journal of Scientific & Technology Research*, vol. 4, no. 8, pp. 235–239, 2015.

[5] P. Andreeva, "Data modelling and specific rule generation via data mining techniques," in *International Conference on Computer Systems and Technologies-CompSysTech*, 2006.

[6] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International journal of nanomedicine*, vol. 13, no. T-NANO 2014 Abstracts, p. 121, 2018.

[7] T. Widiyaningtyas, I. A. E. Zaeni, and P. Y. Wahyuningrum, "Self-organizing map (som) for diagnosis coronary heart disease," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2019, pp. 286–289.

[8] S. Bhatia, P. Prakash, and G. Pillai, "Svm based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features," in *Proceedings of the world congress on engineering and computer science*, 2008, pp. 34–38.

[9] D. Chandna, "Diagnosis of heart disease using data mining algorithm," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 1678–1680, 2014.

[10] K. Polat, S. Şahan, and S. Güneş, "Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing," *Expert Systems with Applications*, vol. 32, no. 2, pp. 625–631, 2007.

[11] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert systems with applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[12] K. Kavitha, K. Ramakrishnan, and M. K. Singh, "Modeling and design of evolutionary neural network for heart disease detection," *International Journal of Computer Science Issues (IJCSI)*, vol. 7, no. 5, p. 272, 2010.

[13] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[14] Z. Yao, P. Liu, L. Lei, and J. Yin, "R-c4. 5 decision tree model and its applications to health care dataset," in *Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005.*, vol. 2. IEEE, 2005, pp. 1099–1103.

[15] A. R. Sulthana, A. Jaithunbi, and L. S. Ramesh, "Sentiment analysis in twitter data using data analytic techniques for predictive modelling," in *Journal of Physics: Conference Series*, vol. 1000, no. 1. IOP Publishing, 2018, p. 012130.

[16] R. Sulthana and S. Ramasamy, "Ontology based grouping of products using clustering and classification approaches," *J Adv Res Dyn Control Syst*, pp. 1032–1048, 2017.

[17] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert systems with applications*, vol. 34, no. 1, pp. 366–374, 2008.

[18] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial intelligence*, vol. 40, no. 1-3, pp. 11–61, 1989.

[19] F. Arif and U. N. Dulhare, "A machine learning based approach for opinion mining on social network data," in *Computer Communication, Networking and Internet Security*. Springer, 2017, pp. 135–147.

[20] R. Sulthana, A. Jaithunbi, H. Harikrishnan, and V. Varadarajan, "Sentiment analysis on movie reviews dataset using support vector machines and ensemble learning," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 17, no. 1, pp. 1–23, 2022.

[21] A. Mathur and R. Sultana, "A study of machine learning algorithms in speech recognition and language identification system," in *Innovations in Computer Science and Engineering*. Springer, 2021, pp. 503–513.

[22] A. Chauhan, A. Agarwal, and R. Sulthana, "Genetic algorithm and ensemble learning aided text classification using support vector machines," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.