JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENHANCING HEART DISEASE PROGNOSIS WITH MACHINE LEARNING TECHNIQUES

[1]Sali Nikitha, [2]Mr.P.Yugundhar Reddy, Dr.K.Chaitanya[3]

[1]MTech Student, [2]Assistant Professor, [3]Assistant Professor
[1]Department of Computer Science of Engineering,
[1]ANU college of Engineering and Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh

Abstract: In the present day, there is a growing incidence of health issues, primarily attributed to lifestyle factors and hereditary factors. Notably, heart disease has become increasingly prevalent, posing a significant risk to people's lives. This document presents a survey of various classification methods employed to assess the risk level of individuals based on factors like age, gender, blood pressure, cholesterol, and pulse rate. The "Disease Prediction" system relies on predictive modeling to anticipate a user's health condition based on the symptoms provided as input to the system. The system evaluates the user's input symptoms and produces the likelihood of a specific disease as an output. Disease prediction is accomplished through the implementation of three techniques: Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree Algorithms. These methods compute the probability of the disease, resulting in an average prediction accuracy of approximately 81%.

Keywords: Decision Tree, Naive Bayes, KNN, Heart Disease Prediction**.**

1. Introduction

In our daily lives, various factors can impact the human heart, and we are witnessing a rapid increase in heart-related issues, with new cardiac conditions continually emerging. In today's stress-filled world, the heart, a vital organ responsible for pumping blood throughout the body, plays a crucial role in maintaining our overall health. The well-being of the human heart is intricately linked to an individual's life experiences and is heavily influenced by their personal and professional behaviors. Additionally, there may be genetic factors contributing to the inheritance of specific heart diseases across generations.

Classification algorithms play a pivotal role in predicting an individual's susceptibility to heart disease. Health care is an indispensable service sector and ranks as the second-largest industry in the 21st century. This research endeavors to identify the most effective classification algorithm for assessing the likelihood of heart disease in a patient. This undertaking is substantiated by conducting a comprehensive comparative analysis using well-established classification algorithms, including Naive Bayes, Decision Tree, K-Nearest Neighbors (KNN), across various assessment levels.

Although these machine learning algorithms are commonly used, predicting heart disease is a critical task necessitating the highest achievable accuracy. This research aims to provide valuable insights to researchers and medical practitioners, aiding them in better understanding the problem and facilitating the identification of the optimal method for heart disease prediction.

2. Related Work:

According to Ordonez [4], heart disease can be predicted using a few basic characteristics about the patient. In their work, they have developed a system that incorporates a person's characteristics based on a total of 13 basic characteristics, such as blood pressure, cholesterol, and sex, to predict the likelihood that a patient will develop heart disease. The research dataset has been expanded and two new attributes—fat and

smoking behavior—have been introduced. Predictions are made using data mining classification techniques including Decision Tree, Naive Bayes, and Neural Network, and the outcomes are examined on the Heart disease database.

Yilmaz, [5] has presented a technique for classifying cardiotocograms using a binary decision tree and least squares support vector machine (LS-SVM) to determine the patient's state.

Fifty-three individuals who had experienced cardiac arrest were included in a study by Duff et al. [6] and their data was used to analyze the likelihood of developing heart disease. They generally used Bayesian networks for both traditional statistical analysis and data mining analysis.

The prediction of survival for coronary heart disease (CHD) is a difficult study topic for the medical community. Frawley et al. [7] have worked on this. For the goal of comparing the three prediction models' unbiased estimates for performance, they also employed 10-fold cross-validation techniques.

In order to further explore and analyze the multi-parametric characteristic of Heart Rate Variability as well as its linear and nonlinear properties for the diagnosis of cardiovascular disease, Lee et al. [8] developed a unique technique. To estimate different classifiers, including Bayesian classifiers, CMAR, C4.5, and SVM, they have conducted a number of tests on both linear and non-linear data. SVM fared better than the other classifiers, according to their experiments.

An associative classifier built on the effective FP-growth approach is the classification strategy proposed by Noh et al. [9]. They provided a method to gauge coherence since the number of patterns might be enormous and varied, which in turn enabled a difficult decision to prune patterns throughout the pattern-generating process.

In a recent study, Parthiban et al. [10] suggested a Coactive Neuro-Fuzzy Inference System (CANFIS) for the identification and prediction of cardiac disease. Their methodology uses fuzzy logic in conjunction with genetic algorithms and the collective adaptive capacities of neural networks to identify disease onset. Training results and classification accuracy were used to assess how well the suggested CANFIS model performed. Ultimately, their findings indicate that there is a lot of potential for the suggested CANFIS model to predict cardiac illness.

A computational model based on a three-layer multilayer perceptron has been suggested by Guru et al. [12] and is used to expand a decision support system for the identification of five main cardiac disorders. A back propagation algorithm enhanced with the momentum term, adaptive learning rate, and forgetting mechanics is used to train the suggested decision support system.

Palaniappan, et al. [13] conducted research and developed a model called the Intelligent Heart Disease Prediction System (IHDPS) utilizing a variety of data mining approaches, including Neural Networks, Decision Trees, and Naïve Bayes.

A study by Shantakumar et al. [14] used a multi-layer perceptron with back-propagation to create an intelligent and successful heart attack prediction system. As a result, using the retrieved data, the MAFIA algorithm mines the frequent patterns of heart disease.

By evaluating HRV (Heart Rate Variability) from ECG, Yanwei et al. [15] developed a classification approach based on the origin of multi parametric characteristics. The data is pre-processed, and a heart disease prediction model is constructed to classify a patient's heart illness.

Peter et al. [16] talked about a new feature selection method algorithm which is the hybrid method which combined CFS and Bayes theorem (CFS+Filter Subset Eval) and evaluated accuracy 85.5%.

Shouman [17] presented work by integrating k-means clustering with Naive Bayes using different initial centroid selection to improve the Naive Bayes accuracy for diagnosing heart disease patients and accuracy was 84.5%.

Rupali et al. [18] decision support in Heart Disease Prediction

System (HDPS) is developed by using both Naive Bayesian Classification and Jelinek-Mercer smoothing technique. This Laplace smoothing is used to make an approximating function which attempts to capture important patterns in the data to avoid noise & accuracy is 86%.

Elma et al. [19] proposed a classifier with the distance-based algorithm K-nearest neighbor and statistical based Navie bayes classifier and achieved the accuracy 85.92% for heart disease dataset.

3. Proposed Methodology:

This section outlines the methodology and analysis employed in this research. The process begins with data collection and the selection of pertinent attributes as the initial steps. Subsequently, the collected data is preprocessed to meet the necessary format requirements. The data is further divided into two distinct categories: training and testing datasets. Algorithms are then applied, and the model is trained using the provided data. Model accuracy is assessed using the testing data. The study's procedures are facilitated by

employing various modules, including data collection, attribute selection, data preprocessing, data balancing, and disease prediction.

The detailed description of each stage of the proposed system is described as follows:

3.1. Data Collection

In this article, the dataset has been sourced from the UCI repository, a widely recognized resource in research analysis, as referenced by numerous authors [20,21]. The initial phase involves organizing this dataset from the UCI repository for the purpose of heart disease prediction. The dataset is subsequently split into two sections: a training set and a testing set. It's noteworthy that in this article, 80% of the data has been allocated for the training dataset, while the remaining 20% of the dataset is reserved for testing purposes.

3.2 Dataset and Attributes

Attributes within a dataset are essential properties that hold significance when it comes to analyzing and making predictions concerning the subject of interest. In this context, a variety of patient attributes, including gender, chest pain, serum cholesterol, fasting blood pressure, and exang, among others, are taken into account for disease prediction. Moreover, the correlation matrix serves as a valuable tool for attribute selection when constructing a predictive model.

Sl. No.	Attributes	Description	Values
1.	Age	Patients age in years	Continuous
2.	Sex	Sex of subject (male-0, female-1)	Male/Female
3.	CP	Chest pain type	Four types
4.	Trestbps	Resting blood pressure	Continuous
5.	Chol	Serum cholesterol in mg/dl	Continuous
6.	FBS	Fasting blood pressure	< 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 introduced by exer.	Continuous
11.	slope	Slope of Peak Exercise ST segment	up/flat/down
12.	Ca	Number of major vessels	0-3
13.	thal	Defect type	Reversible/Fixed/Normal
14.	Targets	Heart disease	1 (disease), 0 (no disease)

Table 1: Attributes used are listed.

3.3 Preprocessing of Data

Data cleaning is an essential step to ensure the accuracy and reliability of the results. This process involves the removal of missing or noisy values from the dataset. In Python 3.8, you can utilize standard techniques for filling in missing data and handling noise, as detailed in reference [22]. Following data cleaning, the dataset should be subjected to transformation, including normalization, smoothing, generalization, and aggregation.

Integration is a critical phase in data preprocessing, where various issues related to data integration are addressed. Sometimes, datasets can be complex or challenging to comprehend. In such instances, it's beneficial to reduce the dataset into a more manageable and structured format to achieve more reliable results.

3.4 Balancing Data

Balancing the dataset is essential for enhancing the performance of machine learning algorithms. A balanced dataset ensures an equal number of input samples for each output class or target class. To rectify an imbalanced dataset, two primary methods can be employed: under-sampling and over-sampling. Under-sampling involves reducing the number of samples in the majority class, while over-sampling entails increasing the number of samples in the minority class. These techniques help address the issue of class imbalance, ultimately improving the effectiveness of machine learning model training and predictions.

This article employs five distinct machine learning algorithms for classification and conducts a comparative analysis of their performance. The objective is to identify the machine learning algorithm that achieves the highest accuracy in predicting heart disease, as illustrated in Figure 1

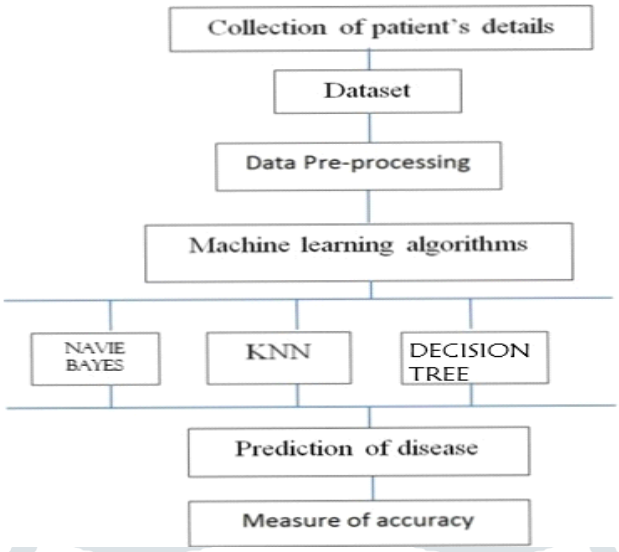

Figure 1. ***Architecture of prediction models.***

3.5 Classification Techniques

A data analysis technique called machine learning automates the development of analytical models. In this observation, five different algorithms are studied to obtain the accuracy for finding the best one.

Navie Bayes: Naive Bayes is a classification algorithm commonly used for both binary (two-class) and multi-class classification problems. It is particularly straightforward to comprehend when applied to binary or categorical input data. The name "naive" Bayes or "idiot" Bayes arises because it simplifies the calculation of probabilities for each hypothesis to make them computationally feasible. Instead of trying to calculate the values of each attribute conditionally on the hypothesis, they are assumed to be independent given the target value and computed as P(d1|h) * P(d2|h), and so forth. This assumption, which implies that the attributes are conditionally independent, is often unrealistic in real-world data. However, Naive Bayes performs surprisingly well even when this assumption doesn't hold. The Maximum A Posteriori (MAP) probability is used to determine the most likely hypothesis

$$MAP(h) = max(P(d|h) * P(h))$$

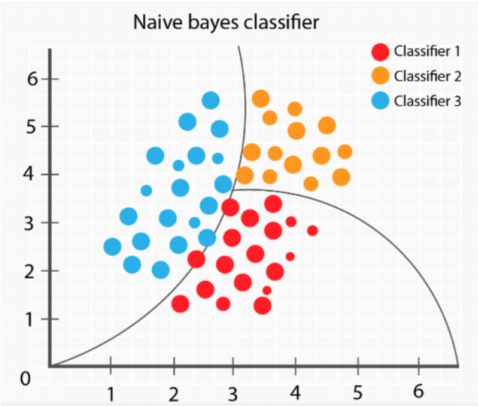

Figure 2. Navie Bayes.

K-Nearest Neighbour: K-NN is the most straightforward classification algorithm based on supervised learning techniques. However, the K-NN algorithm can also be used for regression but is mostly used for classification [25]. A new data point is classified by using the K-NN algorithm depending on how similar the existing data is stored. It indicates that the K-NN algorithm can quickly classify new data when it appears in a suitable category, see Figure 3. Here, the horizontal x-axis and vertical y-axis are independent and dependent variables of a function, respectively. Figure 3 is a simple example of the K-NN classification algorithm.

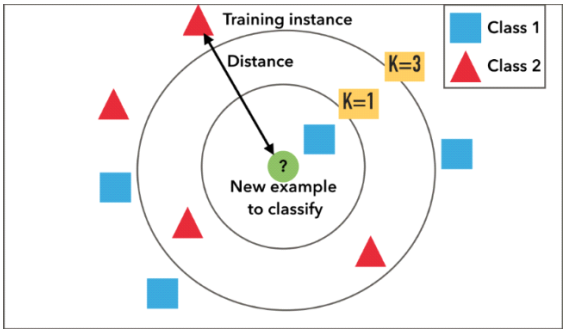

Figure 3. ***K-Nearest Neighbour.***

Decision Tree: Decision Tree is one of the supervised learning algorithms. In this the data is continuously split based on a certain parameter after which we end up getting the decision nodes and the leaves. What makes it different from the other supervised algorithms is that is can also solve the regression and classification problems easily. The main aim is to create a system that can predict the results that we desire just by learning the decision rules from the prior data, i.e., the training set[27].

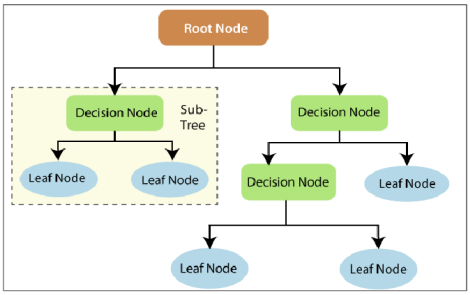

Figure 4. ***Decision Tree.***

Experimental Results

After performing the machine learning algorithms for training and testing the dataset, we can find the better algorithm by considering the accuracy rate. The rate of accuracy is calculated with the support of a confusion matrix. As shown in Table 2, the Navie Bayes algorithm gives us the best accuracy to compare with other ML algorithms.

Algorithms	Accuracy
K-Nearest-Neighbors (KNN)	0.60
Decision Tree (DT)	0.77
Navies Bayes	0.81

Table 2: Accuracy comparision of algorithms.

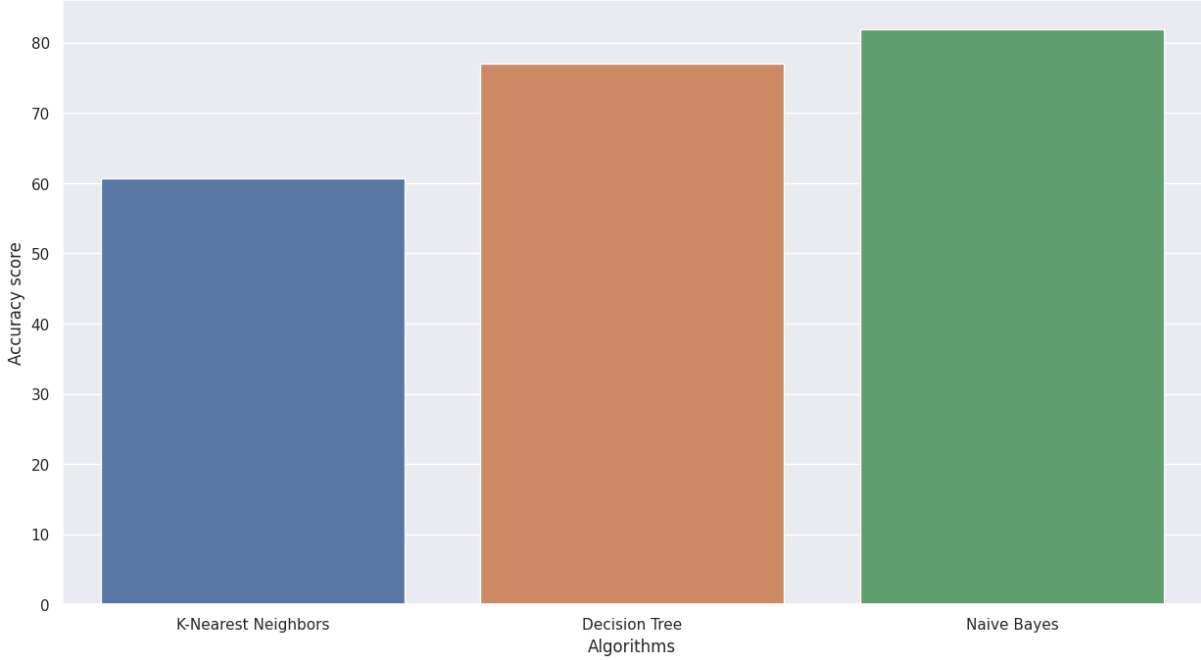

Figure 5. Accuracy comparison of machine learning algorithms by bar diagram.

Acknowledgment

This research was partially supported by Department of Computer Science and Engineering, Acharya Nagarjuna University.

References

1. M. K. Awang and F. Siraj, "Utilization of an artificial neural network in the prediction of heart disease," *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 4, pp. 159–165, 2013.
2. P. Selvakumar and S. P. Rajagopalan, "A survey on neural network models for heart disease prediction," *J. Theor. Appl. Inf. Technol.*, vol. 67, no. 2, pp. 485–497, 2014.
3. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," vol. 6956, no. October, pp. 53–59, 2014.
4. Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules," *Tech. Semin. Present. Univ. Tokyo*, 2004.
5. M. C. and P. M. Franck Le Duff, CristianMunteanb, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method," *Stud. Health Technol. Inform.*, vol. Vol. 107, no. 2, p. No. 2, pp. 1256–1259, 2004.
6. W. J. F. and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview," *AI Mag.*, vol. Vol. 13, N, no. 3, pp. 57–70, 1996.
7. K. Y. N. and K. H. R. Heon Gyu Lee, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," *Proc. Int. Conf. Emerg. Technol. Knowl. Discov. Data Min.*, p. pp. 56–66, 2007.
8. B. J. L. and K. H. R. Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, "Associative Classification Approach for Diagnosing Cardiovascular Disease," *Intell. Comput. Signal Process. Pattern Recognit.*, vol. 345, pp. 721–727, 2006.
9. L. P. and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," *Int. J. Biol. Biomed. Med. Sci.*, vol. Vol. 3, no. No.3, pp. 1-8, 2008.
10. D. and N. R. Niti Guru, "Decision Support System for Heart Disease Diagnosis using Neural Network ," *Delhi Bus. Rev.*, vol. Vol. 8, no. 1, pp.1–6.
11. S. P. and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. Vol. 8, no. No. 8, p. pp. 1–6, 2008.
12. Shantakumar B. Patil and Y.S. Kumaraswamy, "'Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network'," *Eur. J. Sci. Res.*, vol. Vol. 31, no. No. 4, p. pp. 642-656, 2009.
13. X. Y. et Al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease," *Proc. Int. Conf. Converg. Inf. Technol.*, pp. 868–872, 2007.

14. E. Y. and C. Kilikcier, "Determination of Patient State from Cardiotocogram using LS- SVM with Particle Swarm Optimization and Binary Decision Tree," *Master Thesis, Dep. Electr. Electron. Eng.*, no. Uludag, 2013.
15. N. S. and D. Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time," *Dep. Comput. Sci. Eng. Barkatullah Univ. Inst. Technol.*, no. Ph.D Dissertation, 2012.
16. Peter et al. talked about a new feature selection method algorithm which is the hybrid method which combined CFS and Bayes theorem (CFS+Filter Subset Eval) and evaluated accuracy 85.5%.
17. Shouman presented work by integrating k-means clustering with Naive Bayes using different initial centroid selection to improve the Naive Bayes accuracy for diagnosing heart disease patients and accuracy was 84.5%.
18. Rupali et al. decision support in Heart Disease Prediction System (HDPS) is developed by using both Naive Bayesian Classification and Jelinek-Mercer smoothing technique. This Laplace smoothing is used to make an approximating function which attempts to capture important patterns in the data to avoid noise & accuracy is 86%.
19. Elma et al. proposed a classifier with the distance-based algorithm K-nearest neighbor and statistical based Naïve Bayes classifier (cNK) and achieved the accuracy 85.92% for heart disease dataset.
20. Rohit Bharti, Aditya Khamparia, Mohammed Shabaz, Gaurav Dhiman, Sagar pande, and Parneet Singh. Prediction of Heart Disease Using a combination of Machine Learning and Deep learning. Hindawi Computational Intelligence and Neuroscience, Volume 2021, Article ID 8387680, 11 pages. https://doi.org/10.1155/2021/8387680
21. Mai Shouman, Tim Turner, and Rob Stocker. Applying k- Nearest Neighbour in diagnosis heart disease patients.. International Journal of Information and Education Technology, vol. 2, No. 3, June 2012.
22. Wikipedia contributors. (2022, June 21). Logistic regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:36, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=1094256072.
23. Wikipedia contributors. (2022, June 1). Linear regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:39, June 26, 2022, from
24. https://en.wikipedia.org/w/index.php?title=Linear_regression &oldid=1091044459.
25. U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," in *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, 2017.
26. Wikipedia contributors. (2022, June 4). K-nearest neighbors algorithm. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:40, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=K- nearest_neighbors_algorithm&oldid=1091525121.
27. Wikipedia contributors. (2022, June 20). Random forest. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:41, June 26, 2022, from
28. https://en.wikipedia.org/w/index.php?title=Random_forest HYPERLINK "https://en.wikipedia.org /w/index.php?title=Random_forest&ol"& HYPERLINK "https://en.wikipedia.org /w/index.php?title =Random_forest&ol"ol did=1094130824.
29. Q. J., "Induction of decision trees.," *Machine Learning*, vol. 1, p. 81—106, 1986.