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

ADVANCED TECHNIQUE FOR HEART DISEASE PREDICTIONS USING DATA MINING

Tanmayee Parbat, Rohan Benhal, Honey Jain

B.E I.T

Abstract: Data Mining is a methodology that use a variety of ways to uncover patterns or extract information from databases for use in decision-making and forecasting. In this study, an intelligent and effective method for predicting cardiac illness is examined utilising the Naive Bayes modelling technique. For the web-based application, the user must fill in the relevant values for the attributes. The data is retrieved from a database and is used to link training data to the value entered by the user. Traditional approaches cannot reliably detect cardiac illness, but this research can help clinicians make the best judgments possible. To diagnose heart illness, Naive Bayes is utilised for classification, and this method divides output data into no, low, average, high, and extremely high categories. As a result, two basic functions, categorization and prediction, are carried out. The accuracy of the system is determined by the method and database employed, and the Naive Bayes data categorization technique achieves a 98 percent accuracy.

I. Introduction

Nowadays, heart disease is the leading cause of death. The main causes of heart disease include high blood pressure, cholesterol, and a fast pulse rate. There are also some non-modifiable factors. Drinking, like smoking, contributes to heart disease. Our human body's operating system is the heart. If the heart's function is not performed effectively, it will have an impact on other parts of the human body. Family history, high blood pressure, cholesterol, age, poor diet, and smoking are all risk factors for heart disease. The risk level of blood vessels is enhanced when they are stretched too much. As a result, blood pressure rises. Blood pressure is usually measured via systolic and diastolic measurements. The pressure in the arteries while the heart muscle contracts is referred to as systolic, whereas the pressure in the arteries when the heart muscle is at rest is referred to as diastolic. Heart disease is caused by an increase in lipids or fats in the blood. As a result of the lipids in the arteries, the arteries narrow and blood flow becomes slow. Age is a non-modifiable risk factor for heart disease that is also a cause. Smoking is responsible for 40% of all heart disease deaths. Because it reduces the amount of oxygen in the circulation, it damages and constricts the blood vessels. To forecast the risk of heart disease, many data mining approaches such as Naive Bayes, KNN algorithm, Decision tree, and Neural Network are utilised [1].

The KNN method finds the values of heart disease factors by using the K user defined value. The classification report for heart disease is generated using the decision tree technique. The probability-based Naive Bayes approach is used to forecast cardiac disease. The Neural Network predicts cardiac illness with the least amount of error. Patient records are continuously categorised and forecasted in all of the above-mentioned approaches. The patient's activity is continuously monitored, and if any changes occur, the patient and doctor are informed of the disease's risk level. Because of machine learning algorithms and computer technology, doctors are able to predict heart illnesses at an earlier stage. This study discusses the KNN data mining technology, which is used to forecast cardiac disorders.

II. Related work

Various types of studies have been conducted in order to predict the onset of heart disease. For diagnosis, a variety of datamining approaches are utilised, each with a distinct level of accuracy.

N. Repaka et al. [1] The following processes are included in the proposed approach: dataset collection, user registration and login (application based), classification using Navies Bayesian, prediction, and secure data transfer using AES (Advanced Encryption Standard). Following that, a result is generated. Using data mining approaches for heart disease prediction, the research elaborates and presents a variety of knowledge abstraction strategies. The results show that the existing diagnostic approach is successful in predicting risk factors for cardiac disorders.

M. J. A. Junaid et al. [2] This study takes into account a variety of factors that have been linked to heart disease, including heredity, physical activity, total fat consumption, stress, and working conditions. This study provides a fresh perspective on the research area in which these issues arise.

M. T. Islam et al. [3] Because this is a heuristic strategy, this form of clustering can get caught in local optima. To avoid this issue, we employed the Hybrid Genetic Algorithm (HGA) for data clustering.

A. Gavhane et al. [4] propose developing an app that can forecast the vulnerability of a cardiac illness based on fundamental symptoms such as age, sex, pulse rate, and so on. The suggested system uses the machine learning algorithm neural networks, which has proven to be the most accurate and trustworthy algorithm.

S. Ambekar et al. [5] To build on this work, we propose leveraging structured data to forecast illness risk. We deploy a unimodel disease risk prediction technique based on convolutional neural networks. The CNN-UDRP algorithm has a prediction accuracy of above 65 percent. Furthermore, this system provides answers to questions about diseases that people experience in their daily lives.

Hybrid model by M. Kavitha et al [6] (Hybrid of random forest and decision tree). The heart disease prediction model with the hybrid model has an accuracy level of 88.7%, according to experimental results. The interface is meant to obtain the user's input parameter in order to predict heart disease, and we utilised a hybrid Decision Tree and Random Forest model to achieve so.

A. Chauhan et al., [7] Weighted Relationship Rule is a form of data mining approach that is used to automate manual tasks while also extracting data from electronic records. This will help to reduce the cost of services while also assisting in the saving of lives. In this study, we will look at a rule that can be used to forecast a patient's risk of coronary artery disease. The vast majority of the criteria help in the best prognosis of cardiovascular sickness, according to test results.

A. Lakshmana Rao et al [8] employed sampling strategies to locate the best features in an unbalanced dataset, and feature selection techniques to find the best features. Following that, multiple classifier models were used, with the ensemble classifier achieving high accuracy. Experiments on two datasets revealed that the suggested model is successful at predicting cardiac disease. All of the implementations were done in Python.

In comparison to prior research, S. Bhoyar et al [9] found a 12-13 percent increase in accuracy. To test the prediction system, a simple web application tool is created using Python programming. This study aims to provide a user-friendly tool for both medical professionals and the general public.

III. Proposed methodology

There are three categories of attributes in the dataset. Attributes for input, key, and prediction Age, Gender, Blood Pressure, Pulse rate, and Cholesterol are examples of input attributes, with age and gender being the only non-modifiable qualities. Gender is rigid and constant in nature, whereas age is continuous and dynamic. The other parameter has two types [9] of values: continuous and random. Additional characteristics such as smoking and a history of heart disease were also included in the study in order to obtain more accurate results. The modifiable [10] features were smoking and heart disease. To forecast the risk rate of heart disease, constant values were assigned to smoking and heart disease. The patient id is a crucial attribute that is unique to each and every user. The patient and doctor can obtain records using this key characteristic. The application [11] ensures that the user's identity is protected. The disease's risk level was discovered using the Prediction Attribute. The risk level was divided into three categories: low risk, high risk, and normal risk, with low, high, and normal risk indicating less than 50%, greater than 50%, and 0 correspondingly [12].

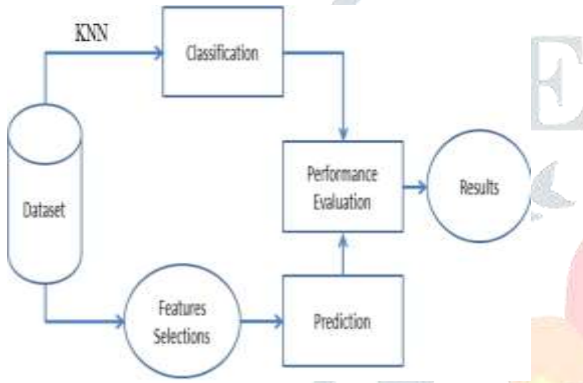

Figure 1: proposed approach flow chart

The suggested method [13] clusters the input dataset using the K-means clustering algorithm and uses the random forest algorithm to forecast type 2 diabetes based on the input parameters. K-Means is one of the most basic [14] unsupervised learning methods for solving the well-known clustering problem [15]. Clustering and other data mining problems can be readily solved with the K-Means approach. It is the most widely used approach for a variety of applications, including vector quantization, density estimation [16], and workload characterization. This clustering algorithm is extensively used and iterative [17] in nature, earning it the name Lloyd's algorithm. Other learning algorithms usually perform better than K-Means. This method searches for related data points and underlying patterns using the original set of cluster centres as a criterion to converge. The k-means algorithm's computing performance is determined [18] by the amount of data, cluster centre computation, and the number of iterations required to converge. K-means is a straightforward algorithm. shows the fitness function along with other factors including [19] the number of clusters, cases, and distance function. The K-Means Algorithm Process Flow is depicted.

Proposed algorithm

Step 1: At random, selects K centroids.

Step 2: Determine the closest centroid and assign data points to it.

Step 3: To determine the centroid, take the average of the cluster data points.

Step 4: Assign data points to the centroids that are closest.

Step 5: Repeat steps 3 and 4 until the results have been reassigned or the number of iterations has been reached.

Results analysis

The Diabetes Dataset is used from UCI repository. The dataset consists of features [20]. These nine features are used for diagnosis of type 2 diabetes. [21] contains nine features, out of these nine features the input features are first eight and the ninth feature is the output feature [22]. The presence of diabetes [23] is indicated by Sick value equal to 1 (one) and the absence of diabetes is indicated [24] by Sick value 0 (zero). The K-means algorithm is executer to cluster and the prediction [25] is done using random forest algorithm. The prediction accuracy of random forest is depicted with other clustering algorithms such as hierarchical clustering and Bayesian network.

The performance of the K-means algorithm in terms of accuracy is given in this section. The performance evaluation of proposed Heart Disease Predictions is evaluated using implementation of coefficient correlation analysis. The accuracy of the system can be given using the following formula.

Accuracy= (Total Correctly Classified Samples)/ (Total Samples Available) X100

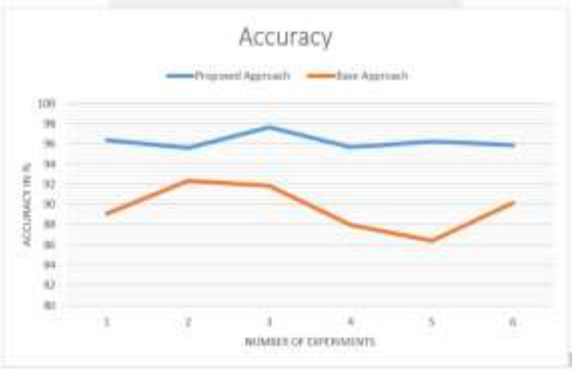

Figure 2: Comparative analysis proposed approach and existing approach in term of accuracy

IV. Conclusion and future work

The major goal of this paper is to provide information on how to use data mining techniques to discover heart disease risk rates. Many papers describe various data mining approaches and classifiers that are utilised for efficient and effective heart disease diagnosis. According to the analysis mode, many authors utilise a variety of technologies and a different number of attributes in their research. As a result, depending on a variety of factors, different methods provide varying degrees of precision. The risk rate of heart disease was discovered using the KNN and ID3 algorithms, and the accuracy level was also supplied for a variety of variables. Other algorithms could be used in the future to reduce the number of attributes while increasing accuracy.

Reference

[1]. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design and Implementing Heart Disease Prediction Using Navies Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, Doi: 10.1109/ICOEI.2019.8862604.

[2]. M. J. A. Junaid and R. Kumar, "Data Science and Its Application in Heart Disease Prediction," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 396-400, Doi: 10.1109/ICIEM48762.2020.9160056.

[3]. M. T. Islam, S. R. Rafa and M. G. Kibria, "Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means," 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392655.

[4]. A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[5]. S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697423.

[6]. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[7]. A. Chauhan, A. Jain, P. Sharma and V. Deep, "Heart Disease Prediction using Evolutionary Rule Learning," 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), 2018, pp. 1-4, doi: 10.1109/CIACT.2018.8480271.

[8]. A. Lakshmana Rao, A. Srisaila and T. S. R. Kiran, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 994-998, doi: 10.1109/ICICV50876.2021.9388482.

[9]. S. Bhoyar, N. Wagholikar, K. Bakshi and S. Chaudhari, "Real-time Heart Disease Prediction System using Multilayer Perceptron," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-4, doi: 10.1109/INCET51464.2021.9456389.

[10]. F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 338-341,

doi: 10.1109/ICREST51555.2021.9331158.

[11]. L. P. Koyi, T. Borra and G. L. V. Prasad, "A Research Survey on State-of-the-art Heart Disease Prediction Systems," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 799-806, doi: 10.1109/ICAIS50930.2021.9395785.

[12]. A. S. Rajawat and A. R. Upadhyay, "Web Personalization Model Using Modified S3VM Algorithm For developing Recommendation Process," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170701.

[13]. C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[14]. Rajawat A.S., Rawat R., Barhanpurkar K., Shaw R.N., Ghosh A. (2021) Blockchain-Based Model for Expanding IoT Device Data Security. In: Bansal J.C., Fung L.C.C., Simic M., Ghosh A. (eds) Advances in Applications of Data-Driven Computing. Advances in Intelligent Systems and Computing, vol 1319. Springer, Singapore. https://doi.org/10.1007/978-981-33-6919-1_5

[15]. R. Latha and P. Vetrivelan, "Blood Viscosity based Heart Disease Risk Prediction Model in Edge/Fog Computing," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 2019, pp. 833-837, doi: 10.1109/COMSNETS.2019.8711358.

[16]. Rajawat A.S., Rawat R., Shaw R.N., Ghosh A. (2021) Cyber Physical System Fraud Analysis by Mobile Robot. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) Machine Learning for Robotics Applications. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_4

[17]. B. Gnaneswar and M. R. E. Jebarani, "A review on prediction and diagnosis of heart failure," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-3, doi: 10.1109/ICIIECS.2017.8276033.

[18]. C. -C. Peng, C. -W. Huang and Y. -C. Lai, "Heart Disease Prediction Using Artificial Neural Networks: A Survey," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020, pp. 147-150, doi: 10.1109/ECBIOS50299.2020.9203604.

[19]. A. P and V. Kalyani David, "Feature selection using ModifiedBoostARoota and prediction of heart diseases using Gradient Boosting algorithms," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 19-23, doi: 10.1109/ICCCIS51004.2021.9397154.

[20]. P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-7, doi: 10.1109/INOCON50539.2020.9298354.

[21]. Rajawat, Anand Singh, et al. “Fusion Protocol for Improving Coverage and Connectivity WSNs.” IET Wireless Sensor Systems, vol. 11, no. 4, 16 Mar. 2021, pp. 161–168, 10.1049/wss2.12018. Accessed 21 Aug. 2021.

[22]. J. S. Sonawane and D. R. Patil, "Prediction of heart disease using learning vector quantization algorithm," 2014 Conference on IT in Business, Industry and Government (CSIBIG), 2014, pp. 1-5, doi: 10.1109/CSIBIG.2014.7056973.

[23]. G. Meena, P. S. Chauhan and R. R. Choudhary, "Empirical Study on

Classification of Heart Disease Dataset-its Prediction and Mining," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), 2017, pp. 1041-1043, doi: 10.1109/CTCEEC.2017.8455127.

[24]. Rajawat A.S., Barhanpurkar K., Goyal S.B., Bedi P., Shaw R.N., Ghosh A. (2022) Efficient Deep Learning for Reforming Authentic Content Searching on Big Data. In: Bianchini M., Piuri V., Das S., Shaw R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore. https://doi.org/10.1007/978-981-16-2164-2_26

[25]. R. G. Saboji, "A scalable solution for heart disease prediction using classification mining technique," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 1780-1785, doi: 10.1109/ICECDS.2017.8389755.

[26]. P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.