Automated Detection of Coronary Artery Disease using Machine Learning Algorithm

Dibakar Sinha1, Ashish Sharma2

1Department of Computer Science & Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore (MP) India,
2Department of Computer Engineering & Applications, GLA University, NH#2, Delhi Mathura Highway, Post Ajhai, Mathura (UP) India,

Ashishs.sharma@glau.ac.in

Abstract. Data mining, an excellent development technology for discovering and gathering essential knowledge from vast data collection that can help analyze and draw up trends for decision-making in the industry. Talking about the medical sphere, data mining can be used to uncover and withdraw useful data and trends that can be helpful in clinical diagnostic results. The research focuses on the diagnosis of heart disease, taking past evidence and information into account. To achieve this SHDP, non-linear SVC with RBF kernel algorithms is designed to perfect this SHDP (Smart Heart Disease Prediction). The final is a useful algorithm to look for the right combination of hyper parameters to increase the precision of the algorithm (C, α). The requisite data was arranged in a structured way. The following features are derived from medical profiles for the estimation of the risks of heart failure in a patient: BP, age, sex, cholesterol, blood sugar, etc. The collected characteristics serve as an input to the Navies Bayesian heart disease prediction classification. The data collection used is divided into two parts, 80% of the data are used for preparation, and 20% are used for research. The method suggested includes data collection, user authentication, and log in (based on application), classification through Navies Bayesian, prediction, and safe data transmission via the AES application (Advanced Encryption Standard). An average accuracy, specificity, sensitivity, precision, 93.53% f-score, 89.22%, 91.24% and 86.98%, respectively. This method is also possible in clinical settings to help clinicians predict cardiac arrest.

Keywords: Machine Learning, Classification, cardio vascular, Logistic Regression.

1. Introduction

As shown in [1], cardiovascular disorders include a variety of problems; many of them linked to atherosclerosis, a condition which develops when a material named plaque builds up the arteries in the walls. This mechanism cuts these arteries and induces blood pressure complications in the veins. Therefore, the possible risk of heart failure continues to increase as a result. The World Health Organization estimates that every year people worldwide die from cardiovascular diseases rather than any other cause. More than 80% of deaths are reported in low middle-income countries, and they are increasing [2]. Therefore, the biggest question is whether it can be expected before it occurs. The use of intelligent methods and algorithms (e.g., neural networks, fluid logic, and genetic algorithms) has
begun to play a decisive role in complicated and unpredictable medical functions such as disease diagnostic, as shown in [1, 3].

In the last decade, the medical literature has seen an immense volume of similar research [3, 4, 5, 6, 7]. As said in [6], computer-aided software and diagnostic and treatment tools tend to be of more recent concern. Also, medical professionals often use computerized technology to help with diagnostics and opinions when there is confusion in medical diagnosis [6]. On the other hand, fluctuating logic and neural networks are promising approaches for solving these uncertainties [1, 6]. As said in [1], both have some benefits over classical methods where ambiguous data or preliminary information is involved.

Data mining primarily uses many tools and algorithms to classify and collect data patterns stored [4]. The extraction of medical knowledge in the health sector is an essential and complex process that needs to be conducted with accuracy and quality. It aims to extract diagnostic knowledge to address real-world health issues for the diagnosis and treatment of diseases. In this article, we use the K-mean algorithms and the A-priori algorithms to forecast heart disease. This allows physicians to consider the state of the patient and find and diagnose the illness correctly. You can also store specific patient details and hold a patient record using patient information systems.

Many of these methods are useful in forecasting cardiac diseases. These methods are, however, sluggish to predict cardiovascular diseases. Also, specificity is not high, which is one of the essential requirements for any predictive process for heart disease.

A new scheme has been implemented in this paperwork for a correct analysis of coronary disorders that can mitigate complications. By MATLAB applications, the proposed diagnostic method was developed using the clustering K-means and Apriori algorithms. Both algorithms can be used to diagnose the disease process effectively.

This advice method is focused primarily on patients' cardiac problems, i.e., details from previous days. The efficiency of this new advice framework is measured in terms of accuracy, risk, and saving workload. The studies are carried out using cardiac data from many patients with heart disease, and the findings are compared with other classifications, including K-means and Apriori algorithms.

2. Related Work

Zou et al. (2018) [1] analyzed web-based analysis evidence for the utility of disease tracking multitasking and explored linear and non-linear models, in particular elastic networks extensions and Gaussian multitasking systems, and compared them with their single-task formulae.

Cousyn et al. (2019) [2] have proposed a new software programmed to automatically derive the effects of these diseases from scientific papers based on the Object Recognition (NER) algorithm (NER) for the numerical term reverse text frequency (TF-IDF).

Maji and Arora (2019) [3] suggested a hybridization approach by which decision-making trees and classifier artificial neural networks cross to help predict heart disease using WEKA. A 10-fold validation analysis on the cardiac patient data set in the UCI repository was conducted to check the performance of the proposed algorithm. Prince. Princy. R and Thomas (2016) [4] concentrated on the estimation of cardiac disease incidence for patients using the data mining strategies KNN, Naïve Bayes, Neural Network, and Decision Tree algorithms. The effect is precise as more attributes are used.

The Artificial Neural Network, in [11,12] Chaitrali et al. proved other data-mine methodologies, for example, Decision Tree and Naïve Bayes. In this research, the desired mechanism for heart disease was set up using 15 characteristics [11,12]. The work involved two added weights and smoking with the opportunity to finish coronary illness to build a persuasive system for coronary disease.

Vikas Chaurasia [14] Vikas et al. used in their forensic analysis, three traditional data mining systems: the CART, ID3, and Decision Table, which are separated from a Decision Breeze or Regulatory Classification, to create conjecture models using a more comprehensive dataset. Discernment illustrated that it was more comfortable to present the CART measurement is separated and two other intervention procedures.
Beant et al. [16] An analytical report has been circulated in the IJRITCC Heart Disease Review using Data Mining techniques. The producer cited an extensive range of experts and looked at different implementation and quality data mining techniques.

The above discussions show that a great deal of analysis is needed to improve the precision and find the best solution between analyzing cardiac data and finding the connection between cardiac data.

This advice framework in this paper is focused on patients' cardiac problems, i.e., knowledge from past days. The efficiency of this new advice framework is measured in terms of accuracy, risk, and saving workload. The studies are carried out using cardiac data from many patients with heart disease, and the findings are compared with other classifications, including K-means and Apriori algorithms.

In data analytics, a large amount of work has been completed. We began with basic linear regression, and now we are grappling with deep neural network or deep learning problems. This ability of algorithms for machine learning is used to predict cardiac arrest. The protocol practiced is shown in Fig. 1.

![Flow chart of the proposed Model](image)

The proposed model for coronary disease prediction is described below. The figure proves coronary heart disease prediction practices system.

The model imports the dataset and deal with the missing values about the preprocessing than the dataset is analyzed. It has been searched whether the dataset is suitable for which type of ML algorithm. If the dataset is suitable for supervised, then the dataset is divided into training and testing sets. The supervised learning algorithm is applied. Accuracy and other statistical report are generated. If the dataset is not suitable for supervised learning, then no further processing is done.

It involves the planning of the Dataset and consumer contribution. Weka knowledge mining devices with programming interfaces have been used to update the system for coronary disease forecasting.
The segments used are instances, common classifiers, and evaluation techniques. It uses the Administered Learning Technique. The preparation data was analyzed by a guided learning calculation and ability from the named preparation collection was obtained. For mapping new models, it appears to be used. The preparedness model is the planning data collected from the UCL repository coronary disease database.

The suggested algorithm changed calculation reveals a superior method for evaluating the underlying centroids and is difficult to perform anything but. This altered K-includes attempts to upgrade the k means the party algorithms without any of its drawbacks. K-implies was used to relate as it were to numerical knowledge. However, we have numerical and mixed knowledge figures. This equation needs no cluster number (k), as the details are seen below. Through choosing two initial centroids, two classes are first made that are divided the most remote in the datasets. It may form two categories with the information persons at the underlying development, who are the most distinct. 2. Input-D: n tuples with AI, A2, Am attributes. Both attributes are numeric. (where m = no of markers) 3. Performance- of n-tuple numbers proper to the number of clusters distributed appropriately. The technique of apprenticeship is used here. A guided study estimation analyses the preliminary information and derives a skill from the set called. It is also used for diagraming new iterations. The preparation information from the Ucl Coronary Disease Archive is the Model for preparation.

3. Results:

3.1 Datasets:
The first data collection (D1) has been extracted from the Calabrian register on dialysis of all varieties of interest from the IFC-CNR Institute of Clinical Physiology, in Reggio Calabria, Italy. The first file consisted of 522 data-free samples. There were 29 attributes, eight of which were binary (including the result), 18 discrete and continuous variables (including patient ID), and three categorical variables.

3.2 Data Pre-processing:
The Dataset of Cleveland Heart Disease [1] includes 74 attributes, 303 instances per point, but only 14 of them were dropped because the prediction of heart disease was found to rely on 13 features only. Any of these qualities were undesirable classified and a few other ignored values. As with the cardiac disease diagnosis attributable, category 0 was given for a blockade of less than 50% of the veins and classes 1, 2, 3, and 4 showing diverse types of heart disease with blockages of more than 50% in the backup of the streets, but the need was to predict only whether the patient had heart disease.

| Category | Features                              | Data Type       |
|----------|---------------------------------------|-----------------|
| 1        | Pain in Chest type                    | Categorical     |
| 2        | Age                                   | Integer         |
| 3        | Gender                                | Categorical     |
| 4        | BP                                    | Integer         |
| 5        | Cholesterol in serum                  | Integer         |
| 6        | ECG outcome                           | Categorical     |
| 7        | blood sugar in fasting                | Categorical     |
| 8        | HRmaximum                             | Integer         |
| 9        | Angina due to Angina                  | Categorical     |
| 10       | ST depression                         | Real            |
| 11       | Exercise ST segment of slope           | Categorical     |
| 12       | Thalassemia                           | Categorical     |
| 13       | Result due to fluoroscopy             | Categorical     |
| 14       | Cardiac arrest                        | Categorical     |
To choose the Model to use, we must first understand whether we have a non-linear problem or linear. We typically carry out visual analysis using scatter plots to calculate the degree of correlation (i.e., linear dependence) among many attributes. This method was not followed because of the substantial number of variables. Table 1 displays the accuracy of the different machine learning algorithms (linear and non-linear) used with the various generated datasets. They are separated by themselves for more excellent readability. For brevity's sake, in the following, we are only concentrating on the data collection for ischemia prediction and a single machine learning process, namely SVCR. This is attributed to two varied reasons: the different data sets yield identical results, and SVCR is very much efficient non-linear classification algorithms for sanitary use.

Various assessment matrices are available and can aid with these data sets. These measurement measures are called measurement recall metrics.

To learn the consistency and to note, you must grasp the Table below and all its names. Consider a binary grade. It will return either 0 or 1. If the real class is 1 for specific training data and the expected level is once again, it is considered a real positive. If the whole class is 0 and the predicted course is 1, this is incorrect. If the present study is 1, but the expected level is 0, it is referred to as false negative. If the current class, as well as the scheduled class, are 0, that is negative.

Classification Accuracy: defined as the number of correct predictions compared to the total sample number. Sensitivity (recall) is the number of positive (first) class instances that have been projected. Specificity is the number of negative (second class instances that have already been correctly predicted.

Precision: shows the willingness of the classifier not to mark a negative sample as positive. Recall shows us how fraudulent a percentage of all transactions initially dishonest. The F1 score is the overall consistency and reminder. But the average formula is different. The average daily recipe does not fit here. The following table shows statistical values achieved after applying Dataset into the Model.

| Parameter   | Values |
|-------------|--------|
| Accuracy    | 89%    |
| Sensitivity | 87%    |
| Specificity | 83%    |
| Precision   | 89%    |
| F-Score     | 82%    |

4. CONCLUSION AND FUTURE WORK
The essential purpose of our research is to estimate the presence of heart disease more specifically. With fewer attributes, data mining is a daunting job rather than a sequence of checks. There have been two techniques of data classification, namely modified K-means and SVM. A modified K algorithm is proposed in this paper, which tries to eliminate one of the critical shortcomings of the simple K-means algorithm requiring several clusters as an input. This framework can also be used in future jobs, for example. In addition to the above list, it may include other medical attributes. The Text Mining, accessible in the healthcare industry archive, can be used to mine vast amounts of unstructured data.

References
[1] Heart Disease Data Set of Cleveland, UCI Machine Learning Repository.http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[2] Dilip, R.C., Mridula, C., Samanta, R.K.: An artificial neural network model for neonatal disease diagnosis. Int. J. Artif. Intell. Expert Syst. (IJAE)2(3), 96–106 (2011)
[3] Milan, K., Sunila, G.: Comparative study of data mining classification methods in cardiovascular disease prediction. Int. J. Computer. Sci. Technol.2(2), 304–308 (2011)
[4] Niti, G., Anil, D., Navin, R.: Decision support system on heart disease diagnosis using neural network. Delhi Bus. Rev. 8(1), 99–101 (2007)

[5] Iqra, B., Ali, R.A.: A framework for classifying unstructured data of cardiac patients. Int. J. Adv. Computer. Sci. Appl. 7(2), 133–141 (2016)

[6] Michael, P.M., Biykem, B., Christie, M.B., Ximena, S., Nicolaos, C., Glennon, S., Vijay, N., Arunima, M., Craig, S.M., Jeffrey, L.E., Charles, C., John, T.M.: Cardiac Scorecard: a diagnostic multivariate index examine system for predicting a spectrum of cardiovascular disease. Expert Syst. Appl. 54, 136–147 (2016). Elsevier Ltd.

[7] Jabbar, M.A., Deekshatulu, B.L., Chandra, P.: Classification of heart disease using artificial neural network and feature subset choice. Global J. Computer. Sci. Technol. 13(3), 1–11 (2013). Ver. 1

[8] Luyang, C., Qi, C., Sihua, L., Xiao, J.: Predicting Heart Attacks. Stanford University, Stanford (2014)

[9] Alwan, A.: Global Status Report on Non-communicable Diseases 2010. World Health Organization, Geneva (2011)

[10] Yadav, D. P., et al. "Feature extraction-based machine learning for human burn diagnosis from burn images." IEEE Journal of Translational Engineering in Health and Medicine 7 (2019): 1-7.

[11] D. Sinha and A. Sharma, “Predicting Heart Disease Stage by tuning convergence and cross validation”, published in the 2nd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI’ 2020), 1-3 April 2020, Berlin, Germany, Scopus Indexed.

[12] A. Sharma, Chaudhary J Varshney, D P Yadav, “Sentiment Analysis using Ensemble Classification Technique”, in the Proceedings of 2020 IEEE Students’ Conference on Engineering & Systems (SCES), July 10-12, 2020, Prayagraj, India, Scopus Indexed. DOI: 10.1109/SCES50439.2020.9236754

[13] A. Sharma, V. Bhadana, “Comparative Study of Machine Learning Technique for Prediction of Kidney Disease”, accepted in the International Conference on Advancement in Interdisciplinary Research, July 24-26, 2020, Organized by Forum for Interdisciplinary Research in Mathematical Sciences and BM Jaipur Rajasthan University, Jaipur (Raj), Proceeding Springer, Scopus Indexed.

[14] A. Sharma, A. Mishra, Aashi. Bansal and AchintBansal, “Bone Fractured Detection Using Machine Learning and Digital Geometry”, presented in the 1st International Conference on Mobile Radio Communications & 5G Network (MRCN-2020) during 26th -28th March 2020, in University Institute of Engineering & Technology, Kurukshetra University, Kurukshetra, Haryana, India. Proceeding LNNS Springer, Scopus Indexed.

[15] Jyoti Soni, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887) Volume 17–No.8, March 2011.

[16] Nidhi Bhatla and Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181.

[17] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, Early Heart Disease Prediction Using Data Mining Echniques, Sundarapandian et al.

[18] Shimpy Goyal and Dr. Rajender Singh Chhillar , A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases, International Journal of Engineering Sciences Paradigms and Researches (IJESPR) (Vol. 20, Issue 01 ,may 2015)

[19] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, “Enhanced Prediction of Heart Disease with FeatureSubset Selection using GeneticAlgorithm”,International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376.