Detection Of Heart Disease Using Machine Learning Techniques

Vishal Dineshkumar Soni

Abstract: We live in a 'information age,' a popular saying says. Data of terabytes are generated daily. Data mining is the method that turns data processing into information. The health industry creates huge volumes of data every day. But most of it is not used effectively. Efficient methods to obtain information from such repositories are not widespread for clinical disease diagnosis or other purposes. This paper aims at comparing specific approaches for forecasting cardiac diseases using data mining techniques, examining the numerous variations of mining algorithms employed, and assessing the techniques are efficient and successful. In fact, several potential paths have been discussed on prediction systems. Naïve Bayes, SMO, Random Forest, Decision table is one such method of data mining that can be used to diagnose patients with cardiac diseases. This paper analyzes few parameters and predicts heart disease, suggesting a prediction system based entirely on data mining approaches.

Index Terms: Algorithms ,Dataset, Heart disease, Machine learning, Naïve Bayes, Weka

1. INTRODUCTION

The heart disease is considered one of the world's most complex and life-threatening human diseases. In this state, the heart normally cannot pump the required volume of blood to certain areas of the body to reach the regular functionalities and, as a consequence, heart failure occurred.[1] Coronary disorder indications involve breathlessness, muscular body fatigue, swelling feet and tiredness with associated indicators, for instance elevated jocular blood pressure and p. The performance classification of various machine learning algorithms in the Cleveland cardiovascular dataset was reported in the literature review. The report for Cleveland heart attack is accessible online from multiple authorities on the University of California Irvine Data Mining Repository (UCI)[3]. We may use various types of techniques and algorithms to continue with the research. Machine learning methods are used in this paper to increase the accuracy performance. We may use the following algorithm in machine learning

1. Decision Table
2. Naïve Bayes
3. SMO
4. Lazy Kstar

The detailed description of the proposed is give below

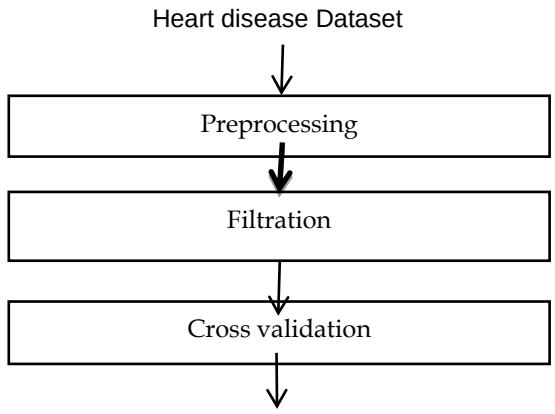
Step 1: Take the Dataset from UCI Library.
Step 2: Apply filtering on the Heart patient dataset.
Step 3: Remove missing values Data through filtering.
Step 4: Cross-validation of Range 5k to 10k.
Step 5: Apply the various algorithms for achieving highest accuracy.
Step 6: Visualize various parameters like Accuracy , ROC , Precision and Recall.

2 LITERATURE REVIEW

Within the following study, we can see different data mining methods used to identify cardiovascular diseases. Research conducted by [4] in 2000 indicated that since human beings cannot arrange data when it is large, we should use the data mining techniques available to find different patterns from the huge database available and that they can then be reused for clinical research and perform various operations. [5], using neural networks, decision tree and naive Bayes algorithms in their research on "Improved Heart Disease Projection System Using Data-Mining Classification Techniques." For the three models, the neural network model forecasts the most reliable heart attack. A predictive system is developed using the neural network for predicting heart level risk using "Effective cardiac prediction system using data mining techniques." The findings revealed that the test device developed would accurately predict the likelihood of heart attack. [7] carried out work involving 533 people who had suffered a heart attack who were included in the study of the risks of cardiac disease. Classical statistical analysis and data mining research were performed primarily through the Bayesian networks. The classification system focused on the sources of multiparameter characteristics has been generated by analyzing the ECG's HRV (Heart Rate Vary), and the data has been preprocessed and a cardiovascular prediction model is built to diagnose a patient's cardiac condition. [8]

3 PROPOSED METHOD

The main purpose of this exercise is to forecast cardiac disease in the other data set attributes. This is a problem of classification. Weka 3.6 is the app to use. The system proposed was developed to classify people with heart disease and healthy individuals. The efficiency of various predictive models for the diagnosis of cardiac disease have been evaluated on complete and selected apps. The commonly used computer modules generate an detailed report using a powerful predictor algorithm. The main goals of the present framework are to evaluate and test patients with condition results and new patient diseases in order to evaluate the potential for a particular person to develop cardiac disorder.
The flow map for the proposed form is seen in the diagram. 1



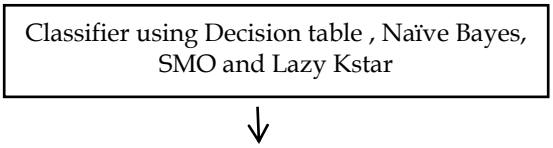


Fig 1: *The flow chart of proposed method*

A. PREPROCESSING

Pre-processing data is required in order to train and successfully check an effective data and machine learning classification. Preprocessing methods such as elimination of missed values, normal Scalar and MinMax Scalar have been used for successful usage in classification schemes. The standard scalar ensures that the mean 0 and variance 1 of each function have the same coefficient. The data is also moved in MinMax Scalar to ensure that all the features range from 0 to 1. Only data is removed from the feature row missing values. All these preprocessing techniques have been used in this work.

1. Missing Value Filter

Missing principles are normal and a plan is required to cope with them. A missed meaning may signify a variety of items in your results. Maybe the data was not accessible or could not be utilized or the incident was not present. The person who entered the data could not have understood or neglected the correct meaning. Data processing methods vary in the way they handle missing values. Missing values are actually overridden, missing values are removed, missing values are substituted by mean values and missing values are extracted from existing values. Initially, the 'ReplaceMissingValue' feature was used to remove missing info. This feature combines all missed details with modes for each nominal and numeric attribute. A further feature called Randomize was used that does not affect the overall output significantly by replacing the missed area.

B. Machine Learning Classifiers

Machine learning recognition algorithms are used to identify people in the heart and stable persons. In this paper a short description is given to some common classification algorithm and their theoretical context.

1) Decision Table

A decision table illustrates the dependent reasoning when designing a set of activities to view corporate law. The Decision Table Tables may be used if a consistent number of conditions are established and certain actions are assessed and assigned when the conditions have finally been met. Decision tables are very similar to decision-making trees, but decision tables also have the same numbers of variables and actions to be tested only after validation of the legitimacy of the list of divisions to be examined[9].

2) Naïve Bayes

Rational Bayesian is effective in making decisions. The Naive Bayes interpretation is probability. It works on the Bayes probability theorem that the unknown data set class is predicted. For a Bayes model taught naively, a collection of likelihood files is protected. This includes:

• Class Likelihoods: The probabilities in the training dataset for each class.

• Conditional Likelihoods: The conditional probability of each input value given for each class value.

3) SMO

This Algorithm is based on Support Vector Machine for binary classification which was developed by [10]. The new algorithm for training supporting vector machines (SVMs) is sequential minimum optimization (SMO). John Platt is the algorithm of minimum sequential optimization (SMO) in 1998 [11] was a quick and simple method for SVM training. The underlying theory is for the dual quadratic optimization to be solved by optimizing the total sub-set of two components. The benefit of SMO is that the presentation is quick and logical. Introducing. A broad issue of quadratic programmatization is the learning of a vector supporting computer. This central question of quadratic programming is separated by SMO in a variety of small potentials. Such a limited square problem system, which prevents the usage of time-consuming quadratic numerical programming as an internal loop, has been analytically resolved. SMO has a linear memory function that allows SMO to perform extensive workouts. Since it is avoided to measure the matrix, the regular SMO scales for various research problems like linear and quadratic chunking SVM algorithm scales like linear and cubic. SMO time is controlled by SVM assessment; SMO is also the fastest for linear and sparse data sets.

4) Lazy Kstar

K * is a classified instance, that is, the class of an instance is determined by a similarity function based on the class of these training instances. The assumption that similar cases will have similar classes is the underlying assumption of instance-based classifiers like K *, IB1, PEBLS etc . It was developed by John G. Cleary in 1995 [12]

5 MATERIALS AND PERFORMANCE ANALYSIS

The data collection was derived from the UCI Repository. The proposed method was simulated using the Weka 3.6 software in an Intel center i7 CPU 1.80 GHz with 8GB RAM PC. The test criteria are accuracy, precision and Recall which are measured to demonstrate the efficacy of the proposed approach in contrast to related algorithms. Accuracy quantifies the efficiency of machine learning algorithm. It is calculated by

$$\text{Accuracy} = \frac{TP+FN}{TP+TN+FP+FN} * 100 \quad (1)$$

where True Positive represents the number of samples that confirm the presence of Heart disease from the proposed algorithm decision and the ground truth label; True Negative represents the number of samples where both the proposed algorithm decision and the ground truth label confirm the absence of tumor; False Positive and False Negative are the number of samples where the decisions mismatch.

6 EXPERMENTAL RESULTS

The data we have should be classified according to the patient's heart characteristics in different structured data. We need to create a model using the logistic method to forecast the illness of the patient based on available data. First, you have to import the data sets. The data will contain specific variables such as age , race , sex, cp, slope, target. To check

the facts, the data should be analyzed. Create a temporary variable and construct a machine-learning algorithm pattern. All the accuracy values for different algorithms are displayed as bars in various colors in Figure 9(a). All the Accuracy values for different algorithms are displayed as bars in various colors in Figure 9(b). All the Precision values for different algorithms are displayed as bars in various colors in Figure 9(c). All the Recall values for different algorithms are displayed as bars in various colors in Figure 9(d).

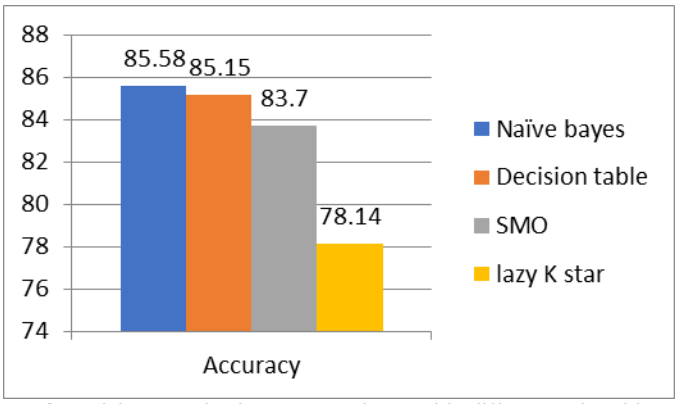


Fig. 9*(a) Quantitative comparison with different algorithms using Accuracy metric.*

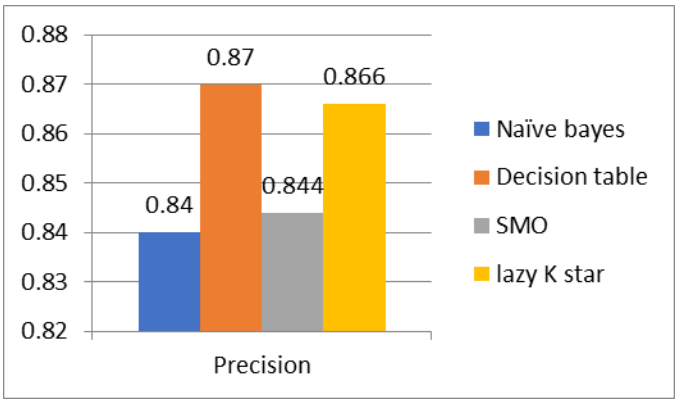


Fig. 9 (*b)Quantitative comparison with different algorithms using Precision metric.*

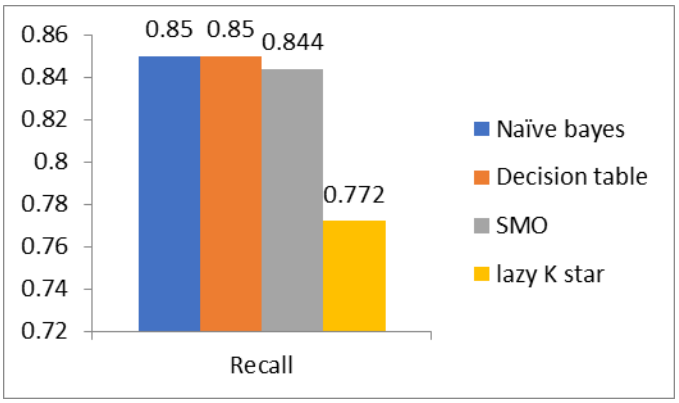


Fig. 9 *(c)Quantitative comparison with different algorithms using Recall metric.*

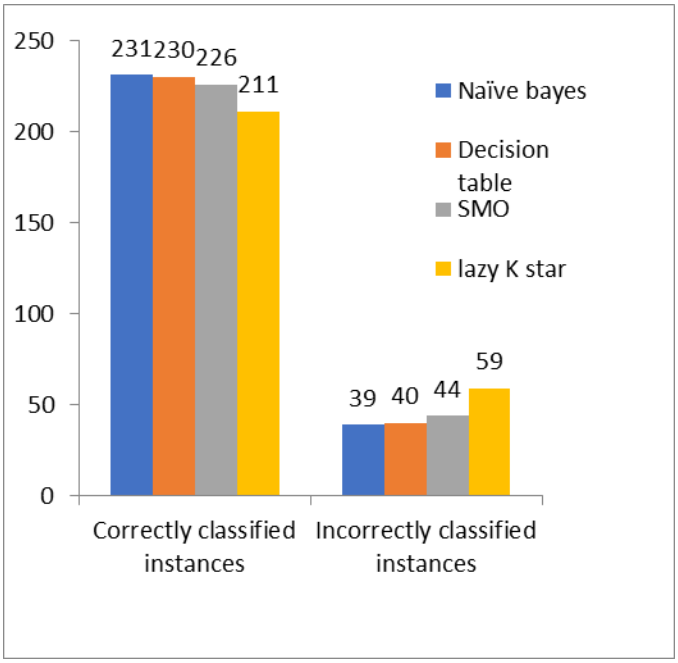


Fig. 9 (*d)Quantitative comparison with different algorithms on correctly classified instances.*

In addition, this approach involves no parameter configuration or prior domain awareness. The grading steps are quick and simple. Both these properties led to a strong degree of accuracy. The only restrictions underlined by this technique are that it does not work well with non-numerical data. In comparison, for limited training data sets, the classification error rate decreases.

7 CONCLUSION

Artificial intelligence can make a computer to act and think like humans[13]. The number of heart diseases can go beyond the control line and reach the peak. Heart disorders are difficult, and a number of people suffer each year from this condition While utilizing these methods one of the key drawbacks of these works is relying exclusively on the specification, with all this researching different data cleaning and mining technologies, of classifications strategies and algorithms to forecast heart disease. So that I can use this machine learning algorithms by predicting whether or not a patient has heart disease in various machine learning algorithms. Any non-medical personnel may use this process to forecast heart failure to reduce doctors' time-complexity. It shows the efficiency of the proposed procedure for classifying Dataset with correct results.

ACKNOWLEDGMENT

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, “Epidemiology and risk profile of heart failure,” Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, 2011

[2] M. Durairaj and N. Ramasamy, “A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,” International Journal of

Control @eory and Applications, vol. 9, pp. 256–260, 2016.
[3] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” Expert Systems with Applications, vol. 68, pp. 163–172, 2017.
[4] VikasChaurasia, Saurabh Pal, “Early Prediction of Heart disease using Data mining Techniques”, Caribbean journal of Science and Technology,2013
[5] Chaitrali S. Dangare et.al, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications,Vol.47,No.10,pg.no:44 – 48, 2012.
[6] Poornima Singh et.al, “Effective heart disease prediction system using data mining techniques”, International Journal of Nano medicine, pg.no:121- 124, 2018.
[7] W.J. Frawley and G. Piatetsky-Shapiro, “Knowledge Discovery in Databases: An Overview”, AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.
[8] X. Yanwei et al., “Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease”, Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.
[9] Shiffman, R. N., & Greenes, R. A. (1991). Use of augmented decision tables to convert probabilistic data into clinical algorithms for the diagnosis of appendicitis. Proceedings. Symposium on Computer Applications in Medical Care, 686–690.
[10] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121{167, 1998. URL citeseer.nj.nec.com/burges98tutorial.html.
[11] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. Burges, and A. Smola, editors, Advances in kernel methods: support vector learning. MIT Press, 1998.
[12] John, G. Cleary and Leonard, E. Trigg (1995) "K*: An Instance- based Learner Using an Entropic Distance Measure", Proceedings of the 12th International Conference on Machine learning, pp. 108-114.
[13] Soni, Vishal Dineshkumar, Challenges and Solution for Artificial Intelligence in Cybersecurity of the USA (June 10, 2020). Available at SSRN: https://ssrn.com/abstract=3624487 or http://dx.doi.org/10.2139/ssrn.3624487