An Efficient Machine Learning Based Methodology for Accurate Heart Disease Detection

PREM SINGH[1], PROF. SURAKSHA TIWARI[2]
[1, 2] *Shri Ram College of Engineering & Management Banmore*

Abstract- It is possible to examine the efficacy of medical treatments by using data mining, a multidisciplinary field of research originating in database statistics. Diabetics are at an increased risk of developing diabetes-related heart disease. When the pancreas stops producing enough insulin, or when the body doesn't utilise the insulin, it does generate correctly, diabetes sets in. Cardiovascular disease, or heart disease, refers to a group of illnesses that affect the heart or blood arteries. Many data mining classification methods exist for predicting heart disease, however there is insufficient data for predicting heart disease in diabetic individuals. Proposed decision tree-based method is achieving better accuracy than the existing classifier.

Indexed Terms- Data Mining, Machine Learning, Decision Tree, Naïve Bayes, Support Vector Machine, Accuracy, Classification, Prediction

I. INTRODUCTION

Analyzing and identifying patterns in large datasets is referred to as "data mining." Different applications employ these patterns to aid in decision-making and forecasting. Algorithms are used to make decisions and anticipate future outcomes. Both supervised and unsupervised learning may be accommodated by data mining methods. Data and class labels are both necessary for training in unsupervised learning; just the data is utilized for training in supervised learning. It is possible to improve classification performance by lowering the error factors inherent in the learning model with supervised learning [1].

These strategies are to be tested in the application of the predictions in this project. In clinical decision support systems, data mining techniques have been widely employed for the prediction and diagnosis of numerous illnesses with high accuracy. In creating clinical support systems, these strategies have been extremely successful because of their capacity to uncover patterns and linkages in medical data. The identification of heart disease, one of the main causes of mortality throughout the world, is a critical application for these systems [2].

Clinical datasets with complicated laboratory testing are used in almost every method for predicting cardiac disease. Age, family history, diabetes, high blood pressure, low HDL (the good cholesterol), smoking and alcohol use are all risk factors for heart disease. But no method can predict heart disease solely on these and other risk factors [3].

Heart disease patients [4] [5] have several of these readily identifiable risk factors, and they may be utilized to provide a quick and accurate identification of the patient's condition. Such a system would benefit medical professionals as well as patients by alerting them to the possibility of heart illness even before they visit a hospital or undergo expensive medical tests. In this method, neural networks and genetic algorithms are the two most effective data mining technologies. Using a hybrid system implementation, the evolutionary algorithm's global optimization advantage is used to initialize neural network weights. Faster, more reliable and more precise than back propagation learning.

II. RELATED WORK

When it comes to health care, it's common for it to be "information rich," but not all of that data is mined in order to find trends and make informed decisions. For medical research, advanced data mining techniques have been applied, most notably in the prediction of heart disease. Prediction methods for cardiac disease were examined by Chaitrali S. Dangare et al [6]. Sex, blood pressure, cholesterol and 13 other medical indicators are included in the algorithm to forecast the risk of acquiring heart disease. To far, there have been

13 factors that have been used to predict the future. Tobacco use as well as weight gain were included as brand-new traits in this investigation. Methods for classifying data from the Heart disease database include Decision Tree classifying, Nave Bayes and neural networks. The performance of various methods is evaluated based on accuracy. As seen by these numbers and graphs, the accuracy of neural networks, decision trees, and naive bayes models is a perfect 100%, 99.62 percent, and 90%, respectively. Analysis shows that Neural Networks are the best at predicting heart disease, compared to the other three models.

E-commerce, marketing, and retail are just a few of the industries where data mining has had a positive impact. One of these fields that is still in its infancy is healthcare. Despite the abundance of available data, there is a lack of expertise in the healthcare industry. There is an abundance of data in healthcare systems. Even yet, there aren't many tools for finding the underlying connections and trends in the data. Researchers JyotiSoni and co-authors [7] are attempting to provide a comprehensive assessment of current approaches to knowledge discovery in databases using data mining techniques, particularly in the field of heart disease prediction. When comparing the accuracy of several predictive data mining techniques on a single data set, several studies have shown that Decision Trees beat other approaches like KNN and Neural Networks, with Bayesian classification occasionally matching Decision Trees in terms of precision. When a genetic algorithm is used to reduce the real data amount, the accuracy of the Decision Tree and Bayesian Classification improves even more, resulting in a better ability to forecast cardiac disease.

An Intelligent System based on Naive Bayes data mining modeling is the fundamental objective of Shadab Adam Pattekari and AsmaParveen [8]. Predetermined questions are asked of the user using a web-based application. Data is extracted from the database and compared to a trained set of values. When it comes to heart illness, it can help healthcare professionals make educated clinical decisions that typical decision support systems can't. In addition, by providing efficient treatments, it helps to cut treatment costs.

Unfortunately, healthcare data is not mined to find hidden information that may lead to better decisions. Hidden patterns and connections are often overlooked. This problem can be solved using advanced data mining techniques. Using data mining techniques such as NaveBayes and WAC, N. AdityaSundar et al [9] have developed a prototype (weighted associative classifier). This system is capable of answering complex "what if" issues that ordinary decision-making systems are unable to. Based on characteristics such as age, gender and blood pressure, it can predict the risk of heart disease in individuals. For example, patterns and relationships between medical factors linked to heart disease can be established. Nurses and medical students may learn how to spot heart abnormalities in patients with the help of this simulation. In hospitals that have a data warehouse, this is a web-based solution that is easy to use. A range of performance metrics are currently being used to evaluate the success of the two categorization data mining methodologies.

Computer-aided data mining involves analyzing a large quantity of data and then extracting its meaning. Using data mining technology, firms may foresee future patterns and take preemptive, well-informed actions. When it comes to solving business problems, it has taken a long time, data mining techniques can help speed things up. It's impossible to process and assess the vast amounts of data gathered for heart disease prediction using normal methods since it's too intricate and voluminous. In order to make use of these mountains of data, data mining provides the skills and technology to do so. Health-related issues are easier to forecast using data mining techniques. Research by R. Thanigaivel et al [10] examines a number of papers in which data mining techniques were used to predict heart illness. Neural networks create outcomes that are nearly flawless. Thus, data mining gave accurate findings when used for prediction. Data mining techniques applied to treatment records for heart disease can yield results that are as reliable as those obtained from the detection of the illness itself.

Researchers at the University of California, Los Angeles (UCLA) have developed a method for classifying forum data to predict final marks in an undergraduate course. It is hoped that student participation in the course forum would be a good indicator of final course grades, and that the proposed classification by clustering approach will be as accurate as more traditional techniques. First-year university students provided the data for the experiments. Based on Moodle forum activity information, many clustering algorithms were compared to traditional classification algorithms in predicting whether students will pass or fail the course. When only a small number of distinct attributes are considered, the Expectation-Maximization (EM) clustering approach outperforms even the best classification algorithms, according to the findings. After all is said and done, the EM cluster centroids are provided to show how the two clusters are linked to the two groups of students they represent.

III. PROPOSED WORK

```
Input: D – Data Partition A – Attribute List GR – Gain Ratio
Output: A Decision Tree
Create a node N
If samples in N are of same class, C then
    return N as a leaf node and mark class C;
If A is empty then
    return N as a leaf node and mark with majority class;
else
    apply Gain Ratio (Dw, Aw)
    label root node N as f(A)
    for each outcome j of f(A)do
        subtree j =New Decision Tree (Dj,A)
        connect the root node N to subtree j
    endfor
endif
endif
Return N
For each instance I in D,
```

If there is an instance I which is not classified then remove the instance from the data set.

IV. RESULT ANALYSIS

To get started, we use the Cleveland data set [12]. Data from the Cleveland area has been pre-processed in order to remove any outliers and ensure consistency. The supplied data is clean and consistent after preparation. SVM, Nave Bayes, and Decision Tree C4.5 have all been trained on this data. There are algorithms that categorize the data provided to them. The data from the classification step is then sent into the prediction step as a source of training data. In this system, new patient data may be predicted based on learning data from the classes. In addition, it offers potential illness designations. A machine learning algorithm's accuracy may be seen in Figure 1.

The data model's ability to accurately classify input data is measured in terms of the percentage of data that is correctly categorized. Formula for measuring accuracy of the algorithm is provided below.

$$accuracy\ \% = \frac{total\ correctly\ classified\ data}{total\ input\ datasets} X100$$

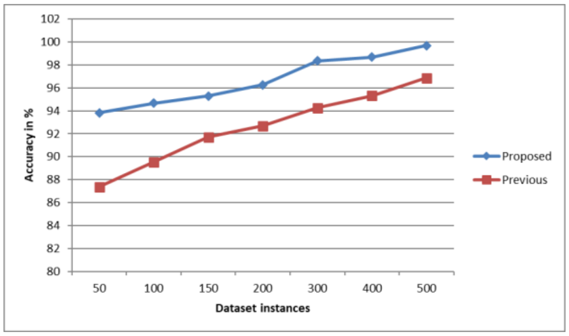

Figure 1: Accuracy of Heart Disease Classification using Machine Learning

CONCLUSION

When manually analyzing data is impractical, data mining techniques are used. The algorithms used in data mining are computer-based and are used to find patterns and relationships among large amounts of data. A decision tree-based classification method for heart disease categorization and prediction was given in this research. This approach has a higher degree of accuracy when it comes to classifying heart disease.

REFERENCES

[1] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_ trends.htm

[2] Chaitrali S. Dangare, Sulabha S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita 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

[4] Shadab Adam Pattekari and AsmaParveen, “Prediction System for Heart Disease Using Naive Bayes”, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[5] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, “Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base”, International Journal of Engineering Science & Advanced Technology, Volume-2, Issue-3, 470 – 478.

[6] Chaitrali S. Dangare, Sulabha S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita 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

[8] Shadab Adam Pattekari and Asma Parveen, “Prediction System for Heart Disease Using Naive Bayes”, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[9] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, “Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base”, International Journal of Engineering Science & Advanced Technology, Volume-2, Issue-3, 470 – 478.

[10] R. Thanigaivel, Dr. K. Ramesh Kumar, “Review on Heart Disease Prediction System using Data Mining Techniques”, Asian Journal of Computer Science and Technology (AJCST)Vol.3. No.1 2015 pp 68-74.

[11] M.I. López, J.M Luna, C. Romero, S. Ventura, “Classification via clustering for predicting final marks based on student participation in forums”, Proceedings of the 5th International Conference on Educational Data Mining.

[12] https://archive.ics.uci.edu/ml/datasets/heart+disease