HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Soniya Tiwari
Assistant Professor, Ajay Kumar Garg Engineering College, Ghaziabad, UP, India
tiwarisoniya@akgec.ac.in

Abstract: **Because of their effects on public health, heart diseases have gotten a lot of interest in medical science. Heart disease is a leading cause of death in the United States, affecting both men and women of all ages. The healthcare industry gathers a lot of data, but not all of it is mined, which is essential for finding secret trends and making informed decisions. Hidden trends and connections are often overlooked.In medical informatics, data mining has become a critical method for computer applications. Several data mining algorithms have greatly aided in the more obvious recognition of medical data. For heart disease prediction, we suggest an effective genetic algorithm based on the back propagation technique. supervised machine learning algorithms, such as Decision Trees and Neural Networks, are used in this study. To forecast heart conditions, SVM, KNN, and Naive Bayes are used. The R programming language is used to execute the machine learning algorithms. The algorithms' precision is used to assess their success. When a large number of attributes are used, the danger level's accuracy is very high.The algorithms' functionality was investigated, and the results were discussed. This study looked at predictive mechanisms for heart disease that had a larger number of input attributes. Gender, blood pressure, blood sugar, cholesterol, and other diagnostic terminology are used by the system to estimate the risk of a patient developing heart disease.**

Keywords: **Data Mining, Big Data, ML, Prediction**

I. INTRODUCTION

Heart disease is a terrible disorder that affects people all over the world, and the fatality rate from it is exceedingly high. In 2010, cardiovascular diseases are estimated to account for 23 million of all deaths worldwide. Cardiovascular disorders, in particular, will be the leading cause of death worldwide, responsible for more than a third of all deaths. Heart disease is a chronic ailment that affects individuals It is found all over the world, and it poses a significant danger of death. Cardiovascular disorders were estimated to be responsible for 23 million deaths worldwide in 2010. Cardiovascular diseases are the leading It is the leading cause of death worldwide, accounting for more than a third of all deaths.The key explanation for this is that, in their best efforts to take the necessary steps to restore their patients' fitness, doctors will not be able to do so in a timely manner. It is important for doctors to be able to forecast their patients' potential wellbeing so that they can take sufficient steps ahead of time to avoid any unfavorable consequences that might occur in the future. Fortunately, doctors would be able to foresee the future using data processing techniques. Data processing tools and regression models are in charge of determining when a patient is at risk for heart disease the future.Based on previous trends, prediction algorithms create assumptions for patients. If we can anticipate who will have heart disease in the future, the doctor will be willing to take the required precautions to help the patient. Which has the potential to greatly minimize, if not entirely prevent, hospital deaths. Researchers would benefit greatly from this big data prediction study,It may also be utilized in real life if the models are upgraded.

The number of electronic health reports obtained by healthcare providers has increased dramatically. When it comes to health treatment, accuracy is crucial, and computerizing this massive volume of data increases the overall system's consistency. So how do healthcare providers sift through all of this data in a timely manner? This is an area where data mining has proven to be incredibly useful. Data mining is a process that combines quantitative analysis, machine learning, and database technology to discover hidden patterns and relationships from large datasets [3]. Data mining is an interdisciplinary subject of computer science and analytics that aims to extract knowledge from a data collection and turn it into a comprehensible framework for future use.As a result, the aim of this study is to use data mining techniques to forecast results in health-care data. Cardiovascular disorders (CVDs) have now surpassed cancer as India's leading cause of death. Heart disease and stroke are the leading causes of CVD deaths, accounting for more than 80% of all deaths [1]. Healthcare organizations are concerned about the availability of high-quality facilities at reasonable pricing. Patients must be correctly diagnosed and treated in order to receive quality treatment. Procedures. Clinical decisions are frequently made based on clinicians' opinions and practices rather than knowledge-rich information from databases. Uninvited biases, blunders, and extravagant medical costs are all exposed as

a result of this practice, all of which have an impact on the level of treatment delivered to patients [2]. Unsupervised learning looks for hidden patterns or underlying components in incoming data, whereas supervised learning uses current input and output data to train a model and predict probable outcomes. The goal of this project is to predict cardiac disease using supervised machine learning algorithms. The aim of supervised approaches is to figure out how input attributes relate to a target attribute. A model is a system that represents the relationship that has been discovered. In supervised learning, the two primary methods are the classification model and the regression model. This paper focuses on a classification model.

Rather than evaluating constant numbers, classification involves assigning observations to distinct groups. This study compares the efficiency of various classification algorithms such as SVM, Nave Bayes, and KNN in predicting heart diseases.

"Data mining" [9] is defined as "the non-trivial retrieval of previously hidden, tacit, and potentially useful data information." The healthcare sector produces a huge volume of data about patients, disorders, and illness detection, among other things. To monitor their healthcare or patient records, the majority of hospitals now use hospital information management systems [11]. These structures typically generate huge amounts of data in the form of figures, tables, text, and photos. Unfortunately, these data are rarely used by doctors to help them make treatment decisions. K-nearest neighbor (KNN), decision trees like CART, C4.5, CHAID, J48, and the ID3 algorithm are just a few of the categorization approaches accessible in data mining. All of these classifiers, however, are useless and require bagging and boosting approaches to increase their performance.

II. DATA MINING TECHNIQUES IN HEALTH CARE

Several data frameworks for healing centres handle continuous charging, stock control, and the development of basic data. A few clinics offer choice support framework, although they are frequently ineffective. They'll answer simple questions like "What is the average age of heart attack patients?" "How many surgeries resulted in more than ten-day stays in the healing centre?" and "How many surgeries resulted in more than ten-day stays in the healing centre?" and "Uncover the female cancer sufferers who are unmarried and above 30 years old." They can't, however, answer difficult queries like "Distinguish the important preoperative signs that make strides the length of clinic stay," or "Distinguish the significant preoperative indicators that make strides the length of clinic remain." "Should treatment include only of chemotherapy, radiation, or both chemotherapy and radiation?" and "Should chemotherapy, radiation, or both be used as part of the treatment?""Should chemotherapy, radiation, or both chemotherapy and radiation be used as part of the treatment?" and "Should chemotherapy, radiation, or both chemotherapy and radiation be used as part of the treatment?" "Foresee the probability of patients experiencing heart disappointment based on clinic data," and "Anticipate the risk of patients experiencing heart disappointment based on clinic information." Clinical decisions are frequently made on the basis of data.

- Decision Trees, Nave Bayes, and Neural Networks are three data mining simulation methodologies used to achieve the main purpose. It will answer tough questions regarding heart disease diagnosis, allowing healthcare practitioners to make better treatment decisions than traditional decision support systems.
- It also attempts to save money on healthcare by providing proper treatment. To facilitate analysis and comprehension, the data is presented in both tabular and graphical representations.
- Integration of clinical decision support with computer-based health data, we hypothesized, may minimize medical mistakes, increase patient safety, reduce unnecessary practice inconsistency, and improve patient outcomes.

III. METHODLOGY BEHIND PREDICTION

This section explains the actions that were taken and the technologies that were employed at each stage. It covers the whole big data analysis and data mining life cycle. Data Gathering The data had to be scraped from the UCI repository database and saved locally in a format that could be pre-processed and used to generate our prediction models quickly. The parsed dataset was broken down into its separate properties and saved as a csv file. We chose to store the data locally rather than using a multi-node cluster because the dataset was modest. From this point on, the data was accessible locally for faster and more efficient access.

Data Visualizations

Because of the way the human brain processes information, using maps or graphs to visualise large amounts of abstract data is easier than poring through spreadsheets. Data visualisation is a quick and easy technique to explain universal ideas, and you may experiment with different scenarios while making modest adjustments.

Data Preprocessing

The process of cleaning a dataset and eliminating undesired components is known as pre-processing. The values in the dataset were converted from string to numerical format as the initial stage in data processing.Training Models

1. Decision Tree
2. Random Forest

3. Logistic Regression
4. Support Vector Machines
5. MLP
6. Naive Bayes

This proposes a prediction approach based on the KNN and ID3 algorithms. It is made up of two modules. The classifier module is in the first module, and the predictor module is in the second. Data is trained and graded in the Classifier module using the KNN algorithm.

IV CONCLUSION

Different classifiers and the impacts of data processing methods are explored, and experiments are conducted to establish the best classifier for forecasting heart disease patients and to understand the value of data processing in improving accuracy. In the future, this study might be expanded to incorporate more machine learning algorithms and pre-processing approaches. More data from other places should also be supplied to help explain the differences in attribute values. The technique might be generalised to a distributed setup if the dataset processing becomes unsustainable on a single node machine. Map-Reduce, Apache Mahout, HBase, and more frameworksand others may be used at this stage. The key goal is to provide information about how to use data mining methods to detect the danger of heart failure. The probability rate of heart attack was detected using the KNN and ID3 algorithms, and the accuracy rating was also given for a variety of attributes. Other algorithms may be used in the future to reduce the number of attributes while increasing precision.

REFERENCES

[1] Global Burden of Disease. 2004 update (2008). World Health Organization.

[2] N. Balasupramanian , Ben , and Imad Salim, "Using Big Data Analytics and Self-Organizing Maps, we can detect and prevent online fraud based on user patterns." International Conference on Intelligent Computing, Instrumentation and Control Technologies" (ICICICT) 2017 .

[3] Manyika, Chui, Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, "Big data: The Next Frontier for Productivity, Innovation, and Competition", Mckinsey Global Institute, May 2011.

[4] http://www.sas.com/enid/insights/analytics/machine-learning/machine-learning.html on recovered on February 20, 2017.

[5] Retrieved on February 20, 2017 from http://a16z.com/2015/01/22/machine-learning.

[6] Dong Hyun Jeong, Caroline Ziemkiewicz, William Ribarsky, and Remco Chang are all big-data/ researchers. "Using a Visual Analytics Tool to Understand Principal Component Analysis", Charlotte Visualization Center, UNC Charlotte fdhjeong, caziemki, ribarsky, 2009.

[7] "Visualization of labelled data using linear transformations," by Y. Koren and L. Carmel.." InfoVis, 00:16, 2003.

[8] E. A. "Rundensteiner," S. Huang, M. O. Ward, and M. O. Ward. The use of dimensionality reduction for text visualisation is being investigated. In CMV '05:Proceedings of the Exploratory Visualization Workshop on Coordinated and Multiple Views", IEEE Computer Society. pp 63-74, Washington, DC, USA, 2005.

[9] "Principal Component Analysis," Springer, second edition, ISBN 978-0-387-95442-4, 2002. I. T. Jolli e, "Principal Component Analysis," Springer, second edition, ISBN 978-0-387-95442-4, 2002.

ABOUT THE AUTHOR

Soniya is an Assistant Professor in Ajay Kumar Garg Engineering College. Currently she completed her Master's degree from CSE Department, JSS Academy of Technical Education NOIDA, U.P., India. Her research domain is Blockchain, AI,ML and Data Science.