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

A Study on Data Mining's Machine Learning-Based Disease Detection

[1]Prem Kumar Rathaor, [2]Manish Kumar Soni,
[1]M. Tech (Computer Science), Bansal Institute Engineering and Technology, Lucknow
[2]Assistant Professor, Department of Computer Science, Bansal Institute Engineering and Technology, Lucknow

Abstract: Health care informatics has benefited from the proliferation of new applications made possible by advances in information technology. Massive amounts of information are being produced through health care informatics. Data mining methods applied to these datasets will allow for disease forecasting. Data mining is the practice of systematically acquiring new information from existing data by means of automated analysis and the presentation of the resulting knowledge. Association, grouping, classification, and prediction are all examples of data mining methods. Health care data and illness prediction are examined using several data mining technologies, and their results are compared. Data mining is an interdisciplinary field that incorporates techniques from a wide range of other fields, such as data visualization, machine learning, database administration, statistics, and many more. It is possible to have these methods cooperate with one another in order to solve more complicated issues. In general, software or systems designed for data mining will make use of one or more of these approaches in order to cope with the various data needs, kinds of data, application areas, and mining jobs.

***Index Terms* - Data Mining, Fundamentals of Data Mining, Data Mining Techniques, Healthcare**

1.1 Introduction

An emerging method, known as machine learning, may assist in the diagnosis of illnesses by making use of either data from models or historical knowledge. Training and testing are the two iterations that make up the machine learning algorithm. Machine learning technology has been striving for many years to develop a method that can forecast diseases by utilizing a patient's symptoms and medical history. The technology of machine learning creates a suitable environment in the clinical setting, which enables a problem with medical treatment to be resolved effectively. We are keeping all of the hospital data up to date via the use of machine learning. Technology based on machine learning that enables the construction of models that can quickly assess information and transmit conclusions more quickly. Because clinicians are able to make sound decisions on patient diagnoses and treatment choices because to the application of machine learning technology, patient medical care management has been shown to significantly improve. The use of machine learning provides an adequate foundation in the area of medicine, which enables healthcare challenges to be effectively addressed. Our project's primary objective is to utilize the Django Python Web Framework to create the user interface as an integral element of the front end and to integrate it with the database modeling components. We will make two different prognoses for the condition, one that is generic and one that is tailored to a particular ailment. It is necessary to conduct research and develop a system that will make it easy for an end user to foresee the continuous illnesses without going to a specialist or a doctor for the treatment, and that will also forecast the kind of physicians they should go to see. Machine learning has a significant capacity for analyzing and coping with a wide variety of ailments, which enables more accurate disease prediction and lower overall treatment costs. Computers not only provide us with information but also keep our attention and aid us in a variety of other ways. A software designed to simulate intelligent human conversation using either text or voice is known as a Chatbot. However, the focus of this study is entirely on the text. These systems are capable of teaching themselves new information and regaining lost knowledge with the help of either humans or online resources. Due to the fact that information is saved in advance, this application is of the utmost significance. In order to provide responses to questions posed by end users, the application for the system implements a question-and-answer format in the form of a chatbot. Because it is not feasible for users to go to physicians or other specialists as soon as they feel the need arises, this system is being created in order to cut down on the time and money spent on healthcare by users [7]. A answer will be provided to the inquiry depending on the user's query as well as the knowledge base. The sentence and the response to the sentence are used to extract the important keywords. If the match is found or the important response is found, the answer will be provided, or answers that are similar to the answer will be presented.

1.2 Literature Review

Kai Hwang et al. , [1] “Disease Prediction by Machine Learning over Big Data from Healthcare Communities” Using medical records, the authors of this research present a novel multimodal disease risk prediction method based on convolutional neural networks. Predictions of whether or not a person is experiencing cerebral infarction (a chronic condition) have been made using machine learning algorithms. The prediction model was evaluated using 2013–2015 central China hospital data. Naive Bayesian, K-Nearest Neighbor, and Decision Tree algorithms classified structured data, whereas Convolutional Neural Network Algorithm classified unstructured data. Their approach predicts 94.8 percent, far higher than existing methods.

Jun Li ,Yongxin Zhai et al. , [2] “Application of Data Mining Methods in Diabetes Prediction ” The authors of this research argue that diabetes prognosis may be improved with the use of data mining techniques. They have investigated the potential of GMM, SVM, Logistic regression, ELM, and ANN Algorithms, together with four other data mining techniques, for the early prediction of diabetes. Compared to other data mining approaches, the experimental findings show that ANN (Artificial Neural Network) delivers the greatest accuracy (0.89). Due to the richness and diversity of the data set, Logistic regression and SVM are less effective in obtaining an anticipated result when compared to other approaches.

Pradeep Kumar Sharmaet al.,[3] “Heart Disease Prediction System using Data Mining Techniques: A study” The value of data mining methods in the healthcare industry has been emphasized in this research. They claim that it is possible to predict which people may develop heart disease by using Data Mining methods to healthcare databases. The authors of the publication provide an in-depth analysis of data mining methods. This paper's approach consisted of a systematic review of medical-related periodicals and publications. It was also determined via this study that a preprocessed and normalized dataset is required for optimal system performance. Some feature selection methods may be used to enhance classification precision. The importance of data mining technologies in evaluating healthcare data troves for illness prognosis and forecasting is also highlighted.

Veena Vijayan V, Anjali C. , [4] “Prediction and Diagnosis of Diabetes Mellitus –a Machine Learning Approach” Using the AdaBoost algorithm and Decision Stump as the basis classifier, A decision support system is proposed by the authors of this research. Predictions concerning diabetes were made using the AdaBoost algorithm in this research. The Support Vector Machine, Naive Bayes, and Decision Tree algorithms were used to verify AdaBoost's correctness. The dataset used for training was collected from the UCI machine learning library, and it had 768 instances and 9 features. the validation dataset was built locally in Kerala. When compared to other popular classifiers like Support Vector Machine and Naive Bayes, the suggested classifier scored 80.72 percent accuracy.

Jitendra Agrawal et al. , [5] “Machine Learning Techniques for Data Mining: A Survey” Machine learning techniques such as Decision Tree, Bayes' algorithm, Support Vector Machine, and Nearest Neighbor are compared in this study. The primary use of these algorithms is in classification. Data instances may be categorized with their help. In doing so, they compare and contrast distinct algorithm implementations. Data mining involves sifting through a big database in search of hidden yet predicted information. Decision Tree, Bayes' method, Support Vector Machine (SVM), and Nearest Neighbor are only few of the machine learning algorithms they've investigated. These calculations are often employed together. They are used in scenarios when it is necessary to foresee the recruitment of a large group. They provide a comparison of several numerical methods. In information mining, the useful but obscured information in a large database is omitted.

Jyoti Soni et al. , [6] “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction” proposed three classifiers based on Decision Tree, Naïve Bayes, and Clustering to identify heart disease in patients. They tested WEKA 3.6.0 on 909 records with 13 characteristics. For simplicity, characteristics were categorized and discrepancies handled. Genetic search improved classifier prediction. After feature subset selection, the Decision Tree outperformed the other two data mining methods, however model creation took longer.

1.3 Fundamentals of Data Mining

Data mining is the practice of retrieving previously undiscovered information from vast stores of data via the use of computer methods. It is highly crucial to extract valuable information from vast data sets and provide decision-making outcomes for the purpose of diagnosing and treating illnesses. Data mining is a technique that may be used to get information by studying and making predictions about a variety of illnesses. Data mining in the health care industry has a significant potential to unearth previously concealed patterns within the data sets of the medical sector [8]. There are many different data mining methods accessible, and the health care data will determine which ones are appropriate to use. Applications of data mining in health care have the potential to be very successful and have a great deal of promise. It does this by using automation to search through massive amounts of data for predicted insights. The ability to accurately anticipate diseases is a crucial aspect of data mining. In order to diagnose a patient with an illness, it is necessary to conduct a variety of diagnostic tests on them. However, reducing the number of tests required may be accomplished via the application of data mining tools. This much smaller test set has a major impact on both performance and timing[9]. The mining of data in the field of health care is an essential work since it enables medical professionals to determine which characteristics are more significant for diagnosis, such as age, weight, symptoms, and so on. The condition may then be diagnosed with greater accuracy by the medical professionals.

1.4 Data Mining Techniques

The term "data mining" refers to the practice of gleaning useful information from large data sets via the use of various algorithmic processes and a variety of other approaches [10]. The following categories of data mining methods are among the most common:

- Association rules, the process of looking for connections between factors, market segmentation, or market basket analysis. By attempting to link seemingly unrelated pieces of data, the dataset itself becomes more valuable. Association rules evaluate a company's sales data to find which products are often purchased together, which may help with planning, marketing, and forecasting.
- Classification uses predefined object classes. These classes define item traits or data points' similarities. This data mining method categorizes and summarizes related properties or product lines.
- Clustering is similar to classification. Clustering, on the other hand, finds commonalities across things and organizes them in terms of their differences. Classification can divide personal care products into categories like "shampoo," "conditioner," "soap," and "toothpaste," while cluster analysis may separate them into categories like "hair care" and "dental health."
- Decision trees utilize a set of criteria or judgments to categorize or anticipate a result. A decision tree uses cascading questions to sort the dataset. Decision trees, which look like trees, let users guide data analysis.
- K-Nearest neighbor (KNN) is a classification method that uses proximity to previous data to determine how to label new data. KNN assumes that nearby data points are more similar than distant ones. This supervised, non-parametric technique infers group features from individual data.
- Neural networks node-process data. Nodes have inputs, weights, and outputs. Supervised learning connects data like the brain. This model may provide threshold values for model correctness.
- Predictive analysis builds predictive models using previous data. This method, similar to regression analysis, supports an unknown figure in the future using existing data.

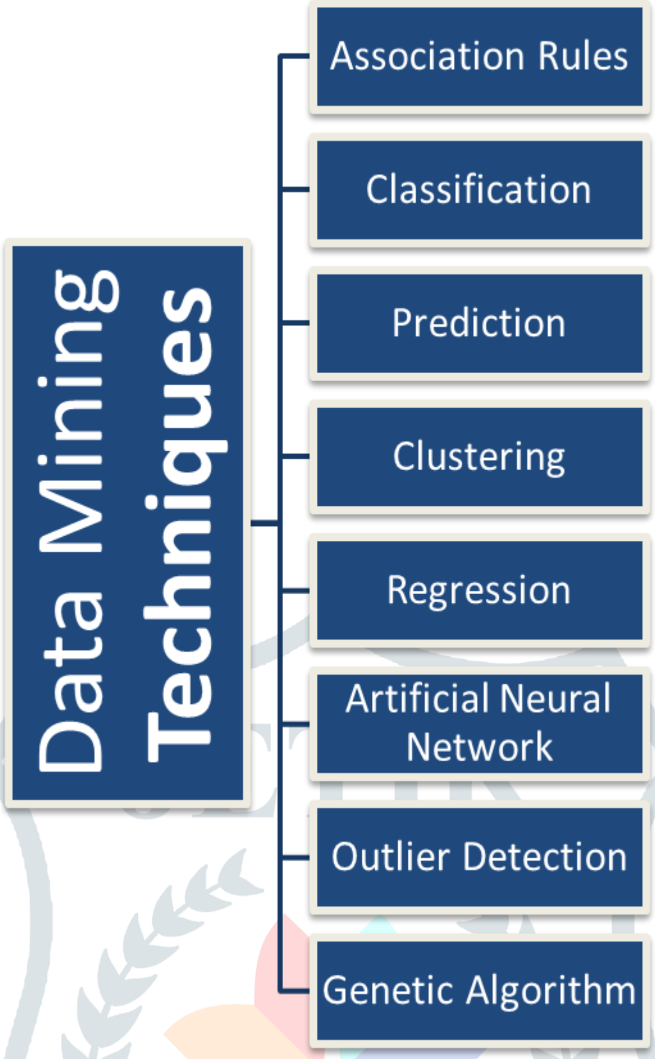

Fig 1.1 Data Mining Techniques

1.5 Conclusion

This article made an effort to get an understanding of the many methods and strategies that are available for predicting the risk of diseases in humans by anticipating the risk variables. The review of the publications provides a concise explanation of the data mining methods that were used on the datasets that were selected. The forecast makes up a significant portion of the core of the work. The ascribes have already been specified, and there is a possibility that in the future, data mining algorithms will be used in order to mine additional characteristics that have the potential to be persuasive in the detection of cardiovascular failures in patients. In comparison to traditional mining processes, data mining algorithms will be of assistance in considering a greater number of features or health issues when developing predictive models. This will be possible because of their ability to analyze large amounts of data.

References

1. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”2169-3536 (c) 2016 IEEE.
2. Messan Komi, Jun Ki, Y ongxin Zhai, Xianguo Zhang, “Application of Data Mining Methods in Diabetes Prediction” 978-1- 5090-6238-6/17/$31.00 ©20 17 IEEE

3. Syed Immamul Ansarullah, Pradeep Kumar Sharma, Abdul Wahid, Mudasir M Kirmani, "Heart Disease Prediction System using Data Mining Techniques: A study", © 2016, IRJET
4. Veena Vijayan V, Anjali C., "Prediction and Diagnosis of Diabetes Mellitus -AMachine Learning Approach" 978-1-4673- 6670-0/15/$31.00 ©2015 IEEE
5. Seema Sharma , Jitendra Agrawal, Shikha Agarwal., "Machine Learning Techniques for Data Mining: A Survey"978-1-4799- 1597-2/13/$31.00 ©2013 IEEE
6. Jyoti Soni et al., "Predictive Data Mining Diagnosis: An Overview of Heart Disease Prediction; International Journal of Computer Applications (0975-8887) Volume 17-No. 8, March 2011
7. Liu Jiquan Deng Wenliang Xudong Lu, Liu Jiquan Deng Wenliang Xudong Lu Huilong Duan, Design and evaluation of clinical Decision support system on Alzheimer disease diagnosis, IEEE. P 1-4, 2009
8. Lionel Brunie, Maryvonne Miquel, Jean-Marc Pierson, and Anne Tchounikine, "Information grids: managing and mining semantic data in a grid infrastructure; open issues and application to geno-medical data. 2003, 14th International workshop on Database and Expert Systems Applications.
9. N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik: Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey
10. Ogundele, I. O., Popoola, O. L., Oyesola, O. O., &Orija, K. T. (2018). A Review on Data Mining in Healthcare. International Journal of Advanced Research in Computer Engineering and Technology (IJARCET).