IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 9, May 2023

Disease Prediction using A Machine Learning

Harshita Awasthi, Malvika Singh, Ms. Ritu Agarwal

Raj Kumar Goel Institute of Technology, Ghaziabad, India

Abstract: *Research on the application of computer systems to prediction, recommendation, and decision-making has grown in popularity during the past ten years. Medical science discoveries can be connected to recent developments in computer technology. Predicting medical behaviour is still a challenging undertaking, and it can only be accomplished with a medical professional's help. Every disease has a pattern in its occurrence that is based on its symptoms. This study's main objective is to propose a technique for using these patterns to predict the related diseases and the possible length of time required to treat them. This was based on the core notion that each disease symptom has a unique impact on the intensity and length of recovery. Using our system, we attempt to quantify this. Prediction is the act of anticipating the occurrence of an event based on a mathematical calculation. For this forecast, we need a recommender system. A program called a recommender system analyses input and, depending on the dataset used to train the program, finds patterns. Based on the pattern, the algorithm chooses a remedy for the problem. It might be unwise to create a database of every conceivable disease and its symptoms and base disease predictions on it. This method's primary drawback is how sluggish and inefficient it is, as well as how large the dataset it uses is. By combining patient ratings and symptoms, our method forecasts potential illnesses and their potential time to cure. Our approach is unique and superior because it predicts diseases based on the severity of the patient's symptoms and cure timeframes based on data from actual patients. Machine learning uses historical data to generate predictions. The process by which a computer program learns from data and experience is referred to as "machine learning." Testing and training are the two stages of the machine learning algorithm. Machine learning technology is still working through issues from decades ago when attempting to anticipate the disease from a patient's symptoms and from their past. Healthcare issues can be successfully solved with machine learning technologies. To keep track of patient health, we employ all machine learning methods that are currently available. Because it forecasts diseases based on the intensity of the patient's symptoms and cure times based on data from actual patients, our method is distinct and superior. Machine learning makes predictions using past data. Machine learning is the process through which a computer program learns from data and experience. The machine learning algorithm has two stages: testing and training. When attempting to predict the disease from a patient's symptoms and from their past, machine learning technology is still working through problems from decades ago. Machine learning technologies can be used to successfully resolve healthcare challenges. We use every machine learning technique currently available to monitor patient health.*

Keywords: Machine learning

I. INTRODUCTION

The method referred to as "disease prediction using machine learning" predicts diseases based on symptoms supplied by patients or other users. The program analyses the user's symptoms and outputs the likelihood that the ailment will manifest in the user. The Naive Bayes classifier, a supervised machine learning technique, is used to forecast the disease. The likelihood of the disease is calculated using the Naive Bayes method. As the volume of biological and healthcare data increases, accurate analysis of medical data aids in early illness detection and patient treatment. Using decision trees and linear regression, we are able to predict the occurrence of diseases such as diabetes, malaria, jaundice, dengue fever, and tuberculosis.

DOI: 10.48175/IJARSCT-10369

II. EXISTING SYSTEM

According to the existing system, chronic diseases specific to a given demographic and region are predicted. Only particular diseases can be predicted using this method. This method uses big data and CNN algorithms to predict the risk of diseases. For S-type data, the system employs Knearest Neighbours, Decision Trees, and Naive Bayesian techniques. The accuracy rate of the current method is 94.8%. Machine learning approaches are simplified in the recently released study for the precise forecasting of chronic disease outbreaks in populations with a high frequency of sickness. They use information from actual hospitals in central China to test the updated prediction models. They propose a convolutional neural network-based multimodal disease risk prediction (CNN-MDRP) system using structured and unstructured data from the hospital.

III. PROPOSED SYSTEM

Patients and other end users use this system to enter all the symptoms they are experiencing, which the machine learning model then uses to predict the disease. The best accuracy is then determined by using the appropriate algorithms. Our method can anticipate the majority of chronic diseases because it enables the machine learning model to take structured data as input. The decision tree is used to divide the enormous dataset into manageable portions, the Naive Bayes method is used to predict the disease using symptoms, the KNN algorithm is used for classification, the features with the highest impact value are retrieved using logistic regression, and the Naive Bayes algorithm is used to predict the disease using symptoms. The system's final result will be the disease predicted by the model.

IV. METHODOLOGY

We are using 3 algorithms: -

KNN (K Nearest Neighbour)

The K Nearest Neighbour (KNN) machine learning algorithm is incredibly straightforward, understandable, and flexible. The user of the healthcare system will forecast the disease. With this method, the user may forecast whether the illness will be discovered or not. The suggested method divides diseases into a number of categories that predict which disease will manifest depending on symptoms. The KNN is used for both classification and regression. The feature similarity approach forms the foundation of the KNN algorithm. When it comes to some classification-related tasks, it is the best choice. The K-nearest neighbour classifier algorithm forecasts the target label of a new instance by defining the nearest neighbour class. To find the nearest class, distance measurements like Euclidean distance will be employed.

The instance is only put into the category of its nearest neighbour if K is 1.

The optimal choice is determined by the date when the user enters a value for 'k'. A higher 'k' number lowers classification noise. The feature class that is closest to the most recent occurrence is chosen when the new feature, in this example the symptom, needs to be categorized. The Hamming distance must be used when working with categorical data. When the dataset includes both numerical and categorical variables, it also raises the challenge of standardizing numerical variables between zero and one.

Random Forest

Random Forest uses many decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy. The random forest uses forecasts from all of the trees, rather than relying on just one, to predict the result based on which predictions earned the most votes.

The supervised learning strategy includes the well-known machine learning algorithm, Random Forest. It can be used to employ ML Classification and Regression to solve issues. Its theoretical foundation is the idea of ensemble learning, which is the process of integrating various classifiers to address a challenging issue and enhance the performance of the model.

Support Vector Machine

Support Vector Machine, sometimes known as SVM[5], is a supervised technique for classifying and predicting data. Machine Learning Classification problems are its main usage. In order to categorize fresh data points in the future, the

SVM method seeks to create a decision boundary that can divide n-dimensional space into classes. A hyperplane is the ideal decision boundary. SVM chooses the extreme points that contribute to the formation of the hyperplane.

Support vectors, which are used to represent these extreme examples, are what give the Support Vector Machine technique its name.

V. BACKGROUND AND RELATED WORKS

Wasan et al [5] recommended employing a number of data mining techniques as diagnostic tools to detect trends in medical data. They have pointed out the potential application of such approaches to knowledge discovery in hospital management systems. Data mining makes it much simpler to find patterns in vast amounts of data.

Medical diagnosis is one area where pattern discovery can be very beneficial. The uses of such strategies were shown by Scales et al. in [6]. Durairaj and Ranjani conducted a comparison of data mining techniques and tools for diverse disorders [7]. Using pre-existing information, they have also looked at the success rate of medical procedures. They have concluded that combining several data mining approaches may lead to better results in the medical industry. A technique for assessing sickness risk through feature selection has been published by Yang et al. [8]. On various UCI datasets, they have applied SVM and random forest methods for this.

Meisamshabanpoor and Mehregan Mahdavi [9] proposed a method for forecasting diseases and their time to cure based on their symptoms. According to age and BMI, their system classifies diseases into different groups. They have developed a method of collaborative filtering that considers neighbourhood selection when making predictions.

Their methodology ignores different factors or weights for different disease symptoms. According to S Sudha and S Vijiyarani [10], data mining techniques should be utilized to predict diseases, mostly of three categories. Breast cancer, diabetes, and heart disease received more attention. They used multiple algorithms to predict different diseases.

A single disease is the subject of many studies on the subject of disease prediction using data mining techniques. For instance, a hybrid technique was used by the authors of [11] to predict asthma Disease. They merged the usage of naive Bayes and neural networks. The fuzzy K-NN technique was recommended by Krishnaiah et al. in [12] as a way to forecast cardiac disease.

The authors in [13] used the WEKA approach with 10-fold cross-validation to forecast dengue illness. Their dataset contains features derived from both clinical symptoms and the patient's current condition. The experiments described in the aforementioned articles are quite helpful and yield encouraging results, however, they are only applicable to one disease. In the proposed study, we make an effort to overcome this difficulty by using a generalized approach to disease prediction.

The suggested approach is based on reinforcement learning according to Barto [14]. This strategy promotes desired outcomes with high incentives and penalizes unwanted outcomes with low incentives. This ensures that when desired outcomes are obtained, they are frequently acknowledged correctly.

VI. RESULTS

In 30 studies, it was discovered that the SVM (Support Vector Machine) algorithm was used the most frequently and with the best accuracy. On the other hand, the KNN (K Nearest Neighbour) approach had the least accurate results, followed by the Random Forest method, which showed pretty decent accuracy.

The study that came before it[9] offers a strategy for predicting diseases and the time it will take for them to be cured based on symptoms. They propose an approach that gives each symptom the same weight. Our forecast performs better than their prediction because in our study, we rate the seriousness of each symptom. S. Vijiyarani and S. Sudha[10] offered data mining methodologies for the prediction of specific diseases, but our method is based on all disorders.

To the best of our knowledge, our implementation will result in improved outcomes when compared to existing efforts. The results of our execution were much in line with the findings. The dataset we used allowed us to get fairly good accuracy.

We were able to predict the beginning of numerous diseases and the length of their cure with high accuracy. Predicting an illness and when it will be cured is a difficult task because it depends on a number of variables, including the patient's immune system.

To the best of our knowledge, we can claim that our method outperforms all earlier studies since it simulates the doctor-patient connection by basing disease prognosis and treatment timeframes on the symptoms that the patient cites.

VII. CONCLUSION AND FUTURE WORK

This disease prediction system's primary goal is to forecast diseases based on their symptoms. Based on the user's symptoms, this approach determines the likelihood that they will recover from their illness. With a 100% chance, the average forecast accuracy is attained. The Grails framework was successfully used to construct the disease predictor. This system's user interface is simple and welcoming.

Because it is a web application, the user can use the system at any time and from any place. In conclusion, the variety of hospital data has an impact on how accurately risk is predicted for disease risk modeling.

This systematic review's objective is to assess the usefulness, potential, and potential future applications of software in the healthcare sector.

The results might help with personalized patient care and give future disease predictability software developers knowledge. The computer program predicts patient ailments. For disease prediction, user symbols are employed.

The data format is processed by the system using the machine learning methods Random Forest, Support Vector Machine, and K-Nearest Neighbour (KNN). The accuracy of the system is 98.3%. Machine learning methods are designed to predict epidemics with accuracy. We address the previously discussed problem of calculating disease duration and corresponding cure periods based on symptoms. The main focus was on classifying symptoms into groups based on their importance and seriousness, then using this data to compute a number value to identify diseases.

The method is particularly accurate when used in a small region, but it can also be used in larger-scale scenarios. In addition, we estimated the length of time needed to heal a condition using the experiences of prior patients. We also rate the severity of the present condition in relation to other users who have the same symptoms. Future studies might focus on diagnosing diseases based on a person's current symptoms and medical history.

The test results for various medical conditions can be used to further boost the system's dependability. The distinction between genuine and fraudulent interactions is essential because the results depend on what previous users are aware of.

REFERENCES

[1] Gediminas Adomavicius and YoungOk Kwon. New recommendation techniques for multicriteria rating systems. Intelligent Systems, IEEE, 22(3):48–55, 2007.

[2] Paul Resnick and Hal R Varian. Recommender systems. Communications of the ACM, 40(3):56–58, 1997.

[3] Wikipedia. List of medical symptoms, 2015. [Online; accessed 22- January-2015].

[4] WebMD. Disease symptoms, 2015. [Online; accessed 22-January2015].

[5] Siri Krishan Wasan, Vasudha Bhatnagar, and Harleen Kaur. The impact of data mining techniques on medical diagnostics. Data Science Journal, 5(19):119–126, 2006.

[6] Roshawnna Scales and Mark Embrechts. Computational intelligence techniques for medical diagnostics. In Proceedings of Walter Lincoln Hawkins, Graduate Research Conference from the World Wide Web: http://www. cs. rpi. edu/˜ bivenj/MRC/proceedings/papers/researchpaper. pdf, 2002.

[7] M Durairaj and V Ranjani. Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.

[8] Jing Yang, Dengju Yao, Xiaojuan Zhan, and Xiaorong Zhan. Predicting disease risks using feature selection based on random forest and support vector machine. In Bioinformatics Research and Applications, pages 1– 11. Springer, 2014.

[9] Meisamshabanpoor and Mehregan Mahdavi. Implementation of a recommender system on medical recognition and treatment. IJEEEE, 2(4):315–318, 2012.

[10] S Sudha. Disease prediction in data mining technique–a survey. IJCAIT, 2(1):17–21, 2013.

[11] Saloni Aneja and Sangeeta Lal. Effective asthma disease prediction using naive bayesneural network fusion technique. In Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on, pages 137–140. IEEE, 2014.

[12] V Krishnaiah, G Narsimha, and N Subhash Chandra. Heart disease prediction system using data mining technique by fuzzy k-nn approach. In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1, pages 371–384. Springer, 2015.

[13] Kashish Ara Shakil, Shadma Anis, and Mansaf Alam. Dengue disease prediction using weka data mining tool. arXiv preprint arXiv:1502.05167, 2015.

[14] Andrew G Barto. Reinforcement learning: An introduction. MIT press, 1998. [15] Geoffrey I Webb Claude Sammut. Encyclopedia of machine learning. Springer Science+Business Media, 2011.

DOI: 10.48175/IJARSCT-10369