MACHINE LEARNING ALGORITHMS FOR PREDICTION OF DISEASES

[1] Dr G Revathy, [2]T.Preethi, [3]A.Benazir Begum, [4]S.PriyadarshiniAND [5]R.T.Subhalakshmi
1-Assistant Professor III, School of Computing, SASTRA Deemed to be University, Thanjavur.
2- Assistant Professor ,Computer science and Engineering,KPR Institute of Engineering and Technology, Coimbatore ,
3- Assistant Professor ,Computer science and Engineering, Hindusthan Institute of Technology, Coimbatore ,
4- Assistant Professor ,Computer science and Engineering,Hindusthan Institute of Technology, Coimbatore
5-Assistant Professor,Computer science and Engineering,Hindustan Institute of Technology, Coimbatore

ABSTRACT

The development and application of several well-known data mining techniques in a variety of real-world application areas (e.g., industry, healthcare, and bio science) has led to their use in machine learning environments to extract useful information from specified data in healthcare communities, biomedical fields, and other fields. Early illness prediction, patient treatment, and community services all benefit from precise medical database analysis. Machine learning techniques have been effectively used in a variety of applications, including disease prediction.The goal of constructing a classifier system utilising machine learning algorithms is to greatly assist physicians in predicting and diagnosing diseases at an early stage, which will greatly aid in the resolution of health-related difficulties. For study, a sample of 4920 patient records diagnosed with 41 disorders was chosen. There were 41 diseases in the dependent variable. We chose 95 out of 132 independent variables (symptoms) that are strongly associated to diseases and improved them.The illness prediction system constructed utilising Machine learning techniques such as Decision Tree classifier, Random forest classifier, Nave Bayes classifier, and K-NN classifier is demonstrated in this study work. The project gives a comparison of the results of the algorithms mentioned above.

Keywords: Machine Learning, Data Mining, Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, K-NN Classifier.

1. INTRODUCTION

Today, datamining is more needed in the healthcare and medical sectors. When certain data mining techniques are applied correctly, important information can be mined from vast databases, allowing medical practitioners to make more informed decisions and improve health care. The goal is to aid the physician by using the classification.Health-related information demands are also changing information-seeking behaviour, which can be seen all around the world. Many people experience difficulties when looking for health information online on ailments, diagnostics, and treatments.

It will save a lot of time if a recommendation system for doctors and medicine can be developed using review mining. Because the users are laypeople, they have difficulty grasping the diverse medical jargon in this type of system. Because there is so much medical information available on many channels, the user gets perplexed.The concept behind the recommender system is to respond to the unique needs of the health domain in terms of users. Malaria, dengue fever, impetigo, diabetes, migraine, jaundice, chickenpox, and other diseases and health issues have a substantial impact on one's health and can even lead to death if ignored.

The healthcare business can make better decisions by "mining" its massive database, that is, extracting hidden patterns and relationships in the data.Decision Tree, Random Forest, and Nave Bayes algorithms are examples of data mining techniques that can help in this case. As a result, according to the rule set of the relevant algorithms, we constructed an automated system that can find and extract hidden knowledge connected with diseases from a historical(diseases-symptoms) database.

2. LITERATURE SURVEY

Many subjects relating to data mining techniques, such as Naive Bayes and KDD, have been discussed by Nikita Kamble et al (Knowledge discovery in Database). Economic sociology and other subjects can benefit from Bayesian statistics. This checks the patients at a basic level and suggests possible disorders automatically.Similar subjects to the paper [2] have been explained by Krishna Kumar Tripathi et al. However, there is a full explanation of the system's core algorithms. The Naive Bayes technique can be used to create models for assigning class labels of various formats. The Naive Bayes algorithm is a collection of algorithms that share a common premise.

The majority of the subjects discussed by Pooja Redd et al. [3] are related to system architecture. The design characteristics of the system are the primary emphasis of this study. The author of this paper has provided a detailed framework for overcoming the disadvantages of the current system. The design aspects of the project are implemented using the smart health framework.Oana Frunza.et.al.[4] uses machine learning to develop a model that predicts healthcare data. It extracts diseases and therapies with the help of many papers and detects relationships between diseases and treatments.

According to L. Hunter et.al, [5] This technique use natural language processing to establish a link between diseases and treatments. In this case, the user types in the name of the ailment, and the NLP returns a solution that is saved in the database. Natural language processing approach and different machine learning techniques are employed by Pravin Shinde et al[6] to make the relationship between diseases and therapies based on a brief text. The user types in the name of the ailment, and the system uses natural language processing to provide a solution that is saved in the database. It also refers to human healthcare diagnosis, treatment, and prevention of illness or injury.

MarimuthuMuthuvel et. al. [7] used machine learning techniques to forecast heart disease. Here, information from the user is necessary, such as blood pressure, hypertension, diabetes, and other inputs. For numerous heart-related problems, Vidya Zopeet.al[8] used approaches including association rule mining, clustering, and classification algorithms like decision trees. K-means clustering algorithms can be used to increase disease prediction accuracy.

3.IMPLEMENTATION

Health-related information demands are also changing information-seeking behaviour, which can be seen all around the world. Many people experience difficulties when looking for health information online on ailments, diagnostics, and treatments. It will save a lot of time if a recommendation system for doctors and medicine can be developed using review mining. Because the users are laypeople, they have difficulty grasping the diverse medical jargon in this type of system. Because there is so much medical information available on many channels, the user gets perplexed.

The concept behind the recommender system is to respond to the unique needs of the health domain in terms of users. The new method is based on a machine learning algorithm. The dataset we looked at has 132 symptoms, which can be combined or permuted to produce 41 disorders. We hope to construct a prediction model based on the 4920 patient data that takes the user's symptoms and forecasts the ailment he is more likely to have.

3.1 Dataset Collection:

The data for this project came from a Columbia University study conducted at New York Presbyterian Hospital in 2004. The dataset's link is provided below.

http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html

3.2 DECISION TREE CLASSIFIER

There are two or more branches in a decision node. All symptoms are regarded as decision nodes in the work provided. The classification, or decision, of any branch is represented by the leaf node. The Diseases correlate to the leaf nodes in this diagram.

The root of the tree contains the attribute that has the greatest impact on the outcome, the leaf verifies the value of a certain attribute, and the leaf gives the tree's output.

The first prediction approach we employed in our project was the decision tree. It provides us with a 95% accuracy rate.

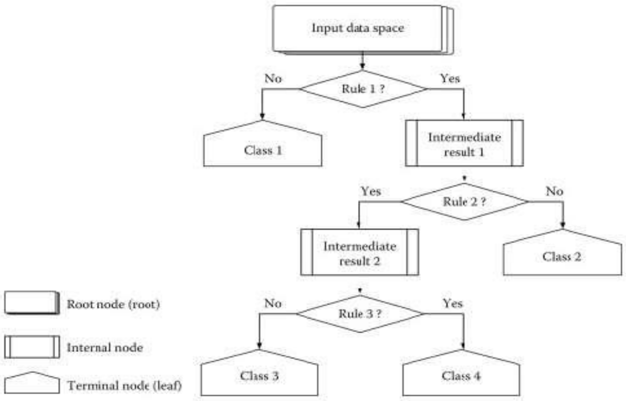

Figure 1 Decision Tree Classifier

3.3 RANDOM FOREST CLASSIFIER

Using different algorithms or the same algorithm many times is referred to as ensemble learning. A group of Decision trees is known as a Random forest. The more decision trees there are in a Random forest, the better the generalisation. Random forest works in the following way:

1. Randomly selects k symptoms from a dataset (medical record) containing m symptoms (where km). Then, based on those k symptoms, it constructs a decision tree.
2. Repeats n times, yielding n decision trees comprised of various random combinations of k symptoms (or a different random sample of the data, called bootstrap sample)
3. Passes a random variable through each of the n-built decision trees to forecast the Disease. The projected Disease is saved, resulting in a total of n Diseases predicted using n Decision trees.
4. Calculates the votes for each projected Disease and uses the mode (most often predicted Disease) as the random forest algorithm's final forecast.

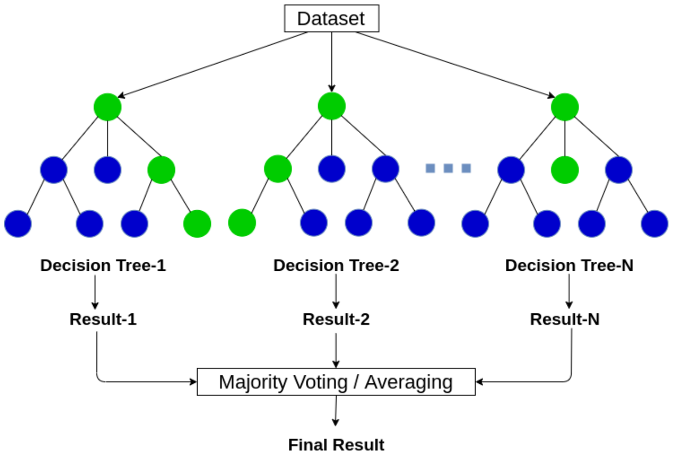

Figure 2 Random Forest Classifier

3.4 NAÏVE BAYES CLASSIFIER

The basic Nave Bayes assumption n is that each feature contributes an equal and independent contribution to the result. It has the advantage of working quickly even on large datasets because it takes less computer capacity.

3.5 K-NN CLASSIFIER

The K Nearest Neighbour algorithm is a supervised learning method. It's a simple yet crucial algorithm. It is widely used in pattern recognition and data mining. It works by identifying a pattern in data that connects data to outcomes, and it becomes better at pattern identification with each iteration. We were able to classify our dataset with 92 percent accuracy using K Nearest Neighbour.

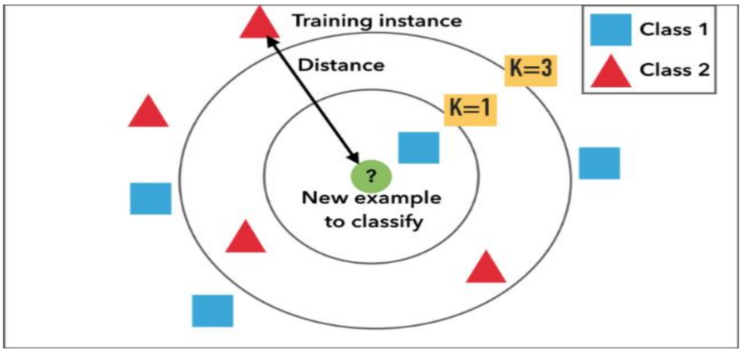

Figure 3 K-NN Classifier

3.6 ARCHITECTURE DIAGRAM

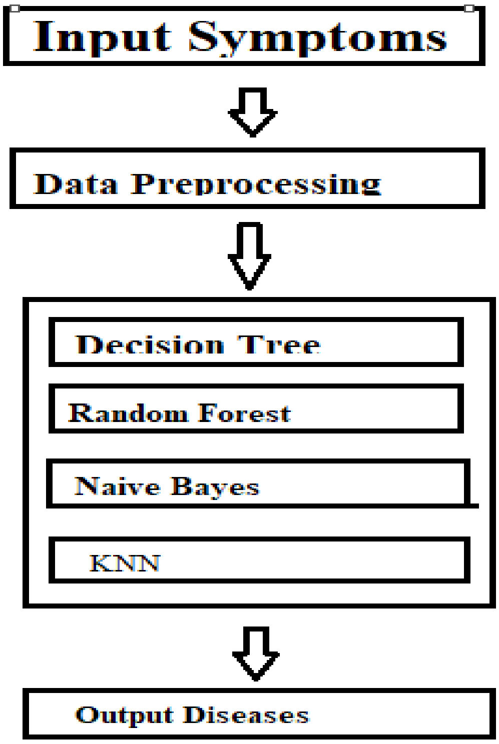

Figure 4 Overall Architecture Diagram

The symptoms are fed into the model, which then processes the data using the previously learned data. Following that, four distinct algorithms are used to predict the disease for a specific patient. Finally, after the algorithms have been implemented, an interface is utilised to display the results to the user, and the user's information is saved in a database file for future use.

3.7 Results Comparison

The system will next use the user's symptoms to predict the ailment, after which the model will deliver the results using four separate machine algorithms and present the results using the tkinterface GUI. Finally, the outputs of four algorithms are compared to determine which one provides the highest level of accuracy.

Algorithm	Accuracy
Decision Tree	0.932927
Random Forest	0.932927
K-NN	0.921331
Naïve Bayes	0.931179

Table 1 Algorithm Accuracy

4.CONCLUSION

Machine learning algorithms were employed to forecast the diseases of patients based on their symptoms. In the prediction procedure, we compared four distinct Machine Learning approaches. Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbour Classifier, and Nave Bayes Classifier are examples of these methods. It was discovered that the

Decision Tree and Random Forest Classifiers have higher accuracy than other classifiers in our area.For the majority and accuracy of the project, we used these four methods in machine learning. In the future, we can use cloud services to link with nearby hospitals and quickly offer patient information for the manufacturing of pharmaceuticals and other valuable items.

REFERENCES

[1]NikitaKamble, International Journal of Scientific Research in Computer Science Engineering and Information Technology, “Smart Health Prediction System Using Data Mining”, Vol. 2, Issue 5, 2017

[2]Prof. Krishna Kumar Tripathi, International Research Journal of Engineering and Technology (IRJET) , “ A Smart Health Prediction Using Data Mining” , Vol.5 Issue:4 , Apr-2018,.

[3] G.Poojareddy, International Journal of Innovative Technology and Exploring Engineering (IJITEE), “Smart E-Health Prediction System Using Data Mining” Vol-8 Issue-6, April 2019,.

[4] OanaFrunza, Member, IEEE “A Machine Learning Approach for Identifying Disease Treatment Relations in Short Texts” IEEE transactions on knowledge and data engineering, vol. 23, no. 6, june 2011.

[5] L. Hunter And K.B. Cohen, “Biomedical Language Processing:What’sBeyondPubmed?” Molecular Cell, Vol. 21-5, Pp. 589-594,2006.

[6] Pravin Shinde and Prof. Sanjay Jadhav, “Health [4] Analysis System Using Machine Learning”, International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014.

[7] MarimuthuMuthuvel and DeivaraniSivaraju, “Analysis of Heart Disease Prediction using Various Machine Learning Techniques", International conference on Artificial Intelligence.

[8] Vidya Zope1 ,Pooja Ghatge2, Aaron Cherian3, Piyush Mantri4 ,Kartik Jadhav, IJSRD - International Journal for Scientific Research & Development| “Smart Health Prediction using Machine Learning” Vol. 4, Issue 12, 2017,.

[9] Dr G Revathy et.al, “ Prediction Of Long Cancer Severity With Computational Intelligence In Covid'19 Pandemic” , Natural Volatiles & Essent. Oils, 2021; 8(5): 3701 - 3707

[10] Dr G Revathy et.al , “ Of Derm Sickness Using Naïve Bayes Classification With Bipolar Fuzzylogic Set” , LINGUISTICA ANTVERPIENSIA, 2021 Issue-3, May 2021.

[11] Mrs.G.Revathy and Dr.K.Selvakumar, “Channel assignment using tabu search in wireless mesh networks”,Wireless personal communication ISSN NO 09296212.

[12] Dr G Revathy et.al , “Revelation of Diabetics by Inadequate Balanced SVM”, Turkish Journal of computer and Mathematics Education, Vol 12, no 2, 2021.