International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.154

Efficient Prediction of Infectious and Non- Infectious Diseases using Decision Tree Classifier Algorithm

Dr. SV. Shri Bharathi
Department of Data Science and Business Systems, School of Computing
srmist-ktr
chennai,india
shribharathi01@gmail.com

KothaRaghavenderreddy*Department of Data Science and Business Systems, School of Computing*
srmist-ktr
chennai,india
kr2824@srmist.edu.in

Sanjay Therani
Department of Data Science and Business Systems, School of Computing
srmist-ktr
chennai,india
st8433@srmist.edu.in

***Abstract*—Disease prediction system is devolved to predict the diseases. This is done from the information/symptoms provided by the user; these symptoms are considered as input to the system. The proposed system analysis helps to discover the symptoms that are provided from the users and shows a probability of the disease as an output. This process of prediction of diseases is done using a decision tree classifier. With data present on these disease, the user can predict the disease bases on systems and also book an appointment based on the result from prediction.**

I. Introduction

The branch of AI and computer science called "machine learning" uses algorithms and information to simulate how people learn, progressively increasing the accuracy of the system. ML has two steps that ensure the accuracy of the output. Those two steps are: training and testing. By using machine learning, issues in the medial field can resolved efficiently.

By using machine learning we are able to securely maintain the records of the hospital with this information we can create a model that to get quickly analyse data and deliver results faster.

With this invention doctors can swiftly get results and also can accede the issue with more quickly from this action more lives can be saved.

For more accuracy from massive data the proposed system used linear, KNN, decision tree algorithm.

II. LITERATURE REVIEW

A. *Common Diseases (CD)*

Dahiwad et al. suggested an ML based technique for anticipating common diseases. The UCI ML repository of various prevalent ailments was used to import the symptoms dataset, which comprised symptoms. The system used the categorization techniques convolution neural network and k-nearest neighbor to forecast a variety of ailments. Furthermore, the proposed remedy included further data on the tested patient's lifestyle choices, which was useful in deciding the degree of risk associated with the anticipated illness. Dahiwad et al. examined the results of the k-nearest neighbor and convolution neural network algorithms in terms of processing speed and correctness. convolution neural network had a processing time of 83.7% and an accuracy of 12.1 seconds. Statistics revealed that the k-nearest neighbor approach outperformed the convolution neural network algorithm. Based on their findings, Che et al. concluded that convolution neural network outperformed conventional supervised algorithms such as k-nearest neighbor and decision tree. The suggested model's ability to recognize complicated nonlinear interactions in the feature space it states why it performed better than previous models. In addition, convolution neural network identifies significant characteristics and improves illness descriptions, allowing it to accurately forecast diseases of great complexity.

B. *Kidney Problems*

A comparison of classifiers' capacities to diagnose chronic kidney disease was performed using the Kidney Function Test dataset. The classifiers used in this study are k-nearest neighbor, Naïve bayes, and random forest, and their performance is measured in terms of F-measure, precision, and accuracy. While Naïve bayes produced superior precision, random forest performed better in terms of F-measure and accuracy, according to the results of the analysis. Considering this study, users attempted to diagnose renal diseases using support vector machine and Naïve bayes. The classifiers were used to differentiate four types of kidney disorders: Acute Nephritic Syndrome, Acute Renal Failure, Chronic Glomerulonephritis, and chronic kidney disease. The study also focused on determining the classification algorithm that performed the best in terms of accuracy and execution time. support vector machine outperformed Naïve bayes in terms of accuracy, making it the best algorithm, according to the data. Naïve bayes, on the other hand, classified data swiftly. Because it effectively manages semi-structured and unstructured data, Charleonna et al. and Kottur et al. determined that the support vector machine classifier is more suitable regarding renal illnesses. Many empirical studies are dedicated on identifying chronic kidney disease. Because of the support vector machine's scalability to larger feature areas, it was able to identify complex kidney disorders with high accuracy.

C. Heart Problems

Marimuth et al. attempted to forecast heart diseases using supervised ML techniques. The classified the attributes of the data as sex, age, aim, slope, and chest pain. The four applicable machine learning algorithms used are decision tree, k-nearest neighbor, logical regression, and Naïve bayes. When compared to the other algorithms, the logical regression algorithm performed the best, with an accuracy score of 87.79%. In 2017, Dwivedi attempted to improve the correctness of heart disease prediction by taking into account new parameters such as resting BP, SC in mg/dl. The dataset used had 119 positive heart disease tests and 148 negative heart disease tests; it was imported from the UCI ML laboratory. Dwived et al. analyzed the performance of Artificial Neural Networks, support vector machine, k-nearest neighbor, Naïve bayes, logical regression, and Classification Trees. The results of tenfold cross-validation demonstrated that logical regression has the highest classification accuracy and sensitivity, exhibiting remarkable dependability in detecting cardiac abnormalities. Polaraj and Vahi et al.'s results, in which LR outperformed other techniques such as Artificial Neural Networks, support vector machine, and AdaBost,

II. Objective

To visit a doctor first we need to book an appointment and pay consult fee but with this system we can save time and money at the same time. This can predict users' disease by analysing the symptoms that were provided as input to the system. As the time progress, accuracy of the system will be improved by ML.

III. Existing System

Although surgeons still need technology in a variety of ways, such as surgical representation and x-ray photography, it has perceptually lagged behind. Due to other aspects like weather, environment, blood pressure, and several other parameters, the approach still requires the doctor's knowledge and expertise. there are a great deal of factors that are acknowledged as being necessary to comprehend the total functioning process,no model has ever been able to examine them adequately. To overcome this issue, medical decision support technologies must be used. This method can help doctors make the best selection.

ML is being used to keep entire healthcare data. With the use of machine learning technology, doctors can make important decisions regarding patient diagnoses and treatment options, improving patient healthcare services. Machine learning technology enables constructing models to evaluate data rapidly and give answers faster. The use of machine learning in the medical industry is best illustrated by the example of healthcare.

IV. Proposed System

Using symptoms, this system is used to forecast illness. This system evaluates the model using a decision tree classifier. End users make use of this system. Based on symptoms, the system will be able to forecast illness. The technology used by this system is machine learning. The decision tree classifier method is used to forecast illnesses.

Using symptoms, this system can be used to forecast infectious and non-infectious diseases. proposed system evaluates using a decision tree classifier. users make use of this system. The technology used by this system is machine learning. The decision tree classifier method is used to forecast illnesses.

V. Dataset Model Description

This dataset used in this analyzes base of symptom connections created automatically using data from textual discharge summaries of patients hospitalized in many Hospitals. The condition is displayed in the first column, followed by the number of discharge summaries that mention it positively and recently, as well as any accompanying symptoms. Based on these notes, associations for the 148 most prevalent diseases were calculated, and the symptoms are displayed sorted according to the strength of link. The method involved extracting UMLS codes for ailments and symptoms from the notes using the MedLEE natural language processing system, and then determining relationships using statistical techniques based on frequency and co-occurrence.

The original information's substance and the quantity of information needed is known as information gain. The features are ordered according to information gains, and the top-ranked features are then picked as prospective classifier qualities. Evaluating the data gained for each feature and then selecting the one that maximizes the information gain will assist you in identifying the decision tree's splitting attribute. The IG for each characteristic is calculated using the following formula:

Evaluating the data gain for each feature and then selecting the one that maximizes the information gain will assist you in identifying the decision tree's splitting attribute. The IG for each characteristic is calculated using the following formula: Figure 1 shows the architecture diagram.

VI. System Architecture

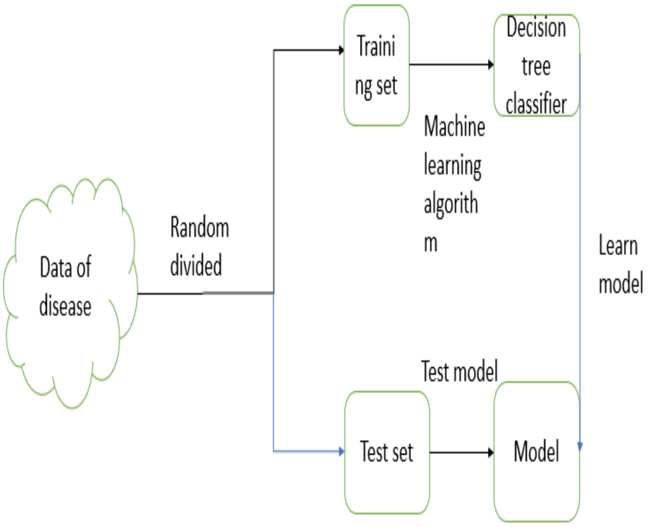


Figure 1: System Architecture Diagram

VII. ALGORITHM

Decision Tree (DT)

The gain ratio decision tree (GRDT) was the type of decision tree used in this investigation. The GRDT's strong foundation line is the entropy (IG) approach, which decides the splitting characteristic that minimises the range of entropy while maximising the IG. The distinction between measures favors tests with a diverse set of effects. In other words, it is like traits with a wide range of potential values. Gain Ratio regulates the IG for individual attributes to ensure consistency and breadth of attribute's value.

Here k is value of target attribute class. Pi is ratio of the value of cases of class1 to the total number of occurrences (i.e., the likelihood of I am happening). To reduce the influence of bias caused using information gain, Australian academic Ross Quinlan created a version known as gain-ratio. The information gain measure favors tests with a wide range of effects. In other words, it prefers to choose qualities with a lot of possible values. Gain Ratio controls IG for each attribute in order to ensure the consistency and breadth of the attribute's value.

For regression and classification, decision trees are a supervised learning technique. After deducing the data attributes, it learns the straightforward decision rules and predicts the target variable's value. ID4, C3.6, C4.9, and CART are examples of DTA. CART, the most recent and improved version, was used in our model.

i) The classification trees of the CART algorithm use Gini impurity. Based on how the labels are split across the subset, it assesses the probability that a randomly selected piece from the set would be erroneously identified if it were given a label at random.

ii) Information acquisition is utilized by the tree creation algorithms ID4, C3.6, and C4.9. It is based on the information theory concepts of entropy and information content. At each stage of the tree-building process, it is utilized to determine which characteristic to split on.

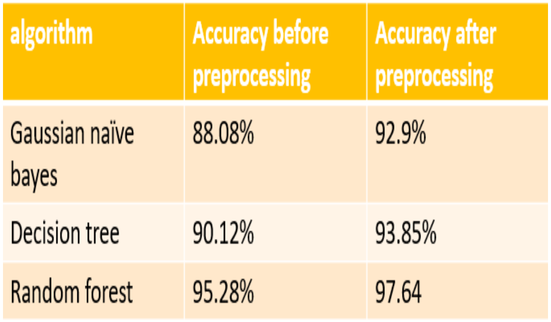
VIII. EVALUATING THE MODELS AND RESULTS

The following table provides an overview of the outcomes from our model:

It is evident that the performance of all three algorithms has been enhanced by the preprocessing method of discretization. Random Forest had the greatest accuracy for our dataset, despite the fact that Naive Bayes' accuracy experienced a large improvement as a result of discretization.

Table 1 shows the comparison chart between previous algorithms with the proposed one.

TABLE1: COMPARISON OF ALGORITHMS



algorithm	Accuracy before preprocessing	Accuracy after preprocessing
Gaussian naïve bayes	88.08%	92.9%
Decision tree	90.12%	93.85%
Random forest	95.28%	97.64

Figure 2 shows the bar graph below demonstrates that the Random Forest classifier performs better than the other two classifiers and is thus the most appropriate for our dataset: -

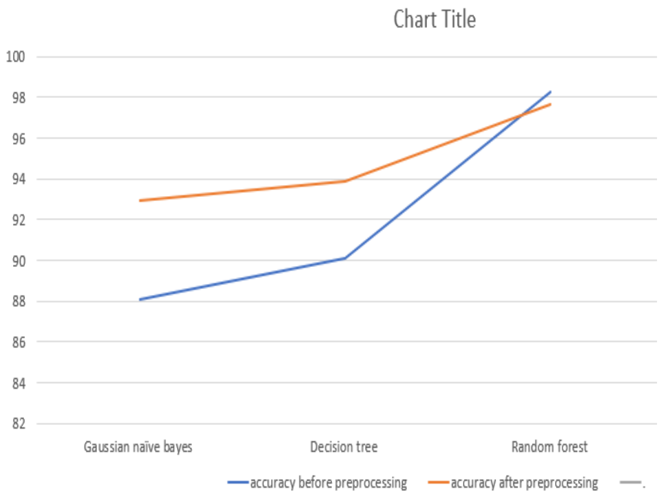


Fig. 2: Comparison graph

VIII. CONCLUSION

Future research will greatly benefit from the development of ML algorithms to enhance sickness detection devices. After the training period, methods of learning should be improved more often to potentially achieve higher performance. Furthermore, to prevent overfitting and enhance the accuracy of the models used, records on different characteristics should be expanded. After time progresses this model, accuracy might vary with good data or methods.

REFERENCES

1. Venugopalkadamba, article aims to implement a robust machine learning model that can efficiently predict the disease of a human, based on the symptoms that he/she possess.
2. Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.
3. E.R.Nika, MabialaBabelaJ.R., S.V.MissambouMandilou, and G. Moyen, "Study of 9 cases of tuberculosis pneumonia in children at Chu of Brazzaville, Congo," Global Pediatr Health, 2016, 3 2333794X16651512.
4. DivyaTomar, "A survey on Data Miningapproaches for Healthcare,"International Journal of Bio-Science and Bio-Technology, October 2013.
5. Rahul Isola, RebeckCarvalho and Amiya Kumar Tripathy, "Knowledge discovery in medical system by using Differential Diagnosis, AMSTAR and K-NN," IEEE Transaction on Information Technology in Biomedicine, vol.16, no.6, November 2012.

6. Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

7. H.C. Koh and G. Tan, “Data mining application in healthcare”, Journal of Healthcare Information Management.