International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Disease Prediction using Machine Learning

Raj H. Chauhan[1], Daksh N. Naik[2], Rinal A. Halpati[3], Sagarkumar J. Patel[4], Mr. A.D.Prajapati[5]

[1-4]Student, Dept. of Computer Engineering, R.N.G.Patel Institute of Technology, Gujarat, India
[5]Professor, Dept. of Computer Engineering, R.N.G.Patel Institute of Technology, Gujarat, India

---

ABSTRACT

Disease Prediction system is based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the Decision tree Classifier. Decision tree Classifier calculates the probability of the disease. With big data growth in biomedical and health care communities, accurate analysis of medical data benefits early disease detection, patient care.

Key Words: Machine Learning, Symptoms based disease prediction, Python.

1. INTRODUCTION

Machine learning is programming computers to optimize a performance using example data or past data. Machine learning is the study of computer systems that learn from data and experience. Machine learning algorithm has two tracks: Training, Testing. Prediction of a disease by using patient's symptoms and history machine learning technology is striving from past decades. Machine Learning technology gives an immeasurable platform in the medical field so that healthcare issues can be resolved efficiently.

We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyze data and deliver results faster, with the use of machine learning technology doctors can make a good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare services. Healthcare is the most prime example of how machine learning is used in the medical field.

To improve the accuracy from massive data, the existing work will be done on unstructured and textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm. The order of reference in the running text should match with the list of references at the end of the paper.

2. OBJECTIVE

There is a need to study and make a system which will make it easy for an end-user to predict the permanent diseases without visiting a physician or doctor for a diagnosis. To detect the Various Diseases through the examining Symptoms of patient's using various methods of Machine Learning Models. To Manage Text data and Structured data is no Proper method. The Recommended system will examine both structure and unstructured data. The Predictions Accuracy will Improve using Machine Learning.

3. EXISTING SYSTEM

Since the arrival of advanced computing, the doctors' still requires the technology in various possible ways like surgical representation process and x-ray photography, but the technology perceptually stayed behind. The method still
requires the doctor's knowledge and experience due to alternative factors starting from medical records to weather conditions, atmosphere, blood pressure and numerous alternative factors. The huge numbers of variables are granted as entire variables that are required to understand the complete working process itself, nevertheless, no model has analyzed successfully. To tackle this drawback, Medical decision support systems must be used. This system can assist the doctors to make the correct decision.

We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyze data and deliver results faster, with the use of machine learning technology doctors can make a big decision for patient diagnoses and treatment choices, which leads to enhancement of patient healthcare services. Healthcare is the most prime example of how machine learning is used in the medical field.

4. PROPOSED SYSTEM

This system is used to predict disease according to symptoms. This system uses decision tree classifier for evaluating the model. This system is used by end-users. The system will predict disease based on symptoms. This system uses Machine Learning Technology. For predicting diseases, the decision tree classifier algorithm is used.

We have named this system as 'AI THERAPIST'. This system is for those people who are always fretting about their health, for this reason, we provide some features which acknowledge them and enhance their mood too. So, there is a feature for the awareness of health 'Disease Predictor', which recognize disease according to symptoms.

5. DATASET AND MODEL DESCRIPTION

This dataset is a knowledge database of disease-symptom associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted during 2004. The first column shows the disease, the second the number of discharge summaries containing a positive and current mention of the disease, and the associated symptom. Associations for the 150 most frequent diseases based on these notes were computed and the symptoms are shown ranked based on the strength of association. The method used the MedLEE natural language processing system to obtain UMLS codes for diseases and symptoms from the notes; then statistical methods based on frequencies and co-occurrences were used to obtain the associations. A more detailed description of the automated method can be found in Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical reports. AMIA Annu Symp Proc. 2008. p. 783-7. PMCID: PMC2656103.

Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall

[Fig 5.1. Dataset]

6. SYSTEM ARCHITECTURE

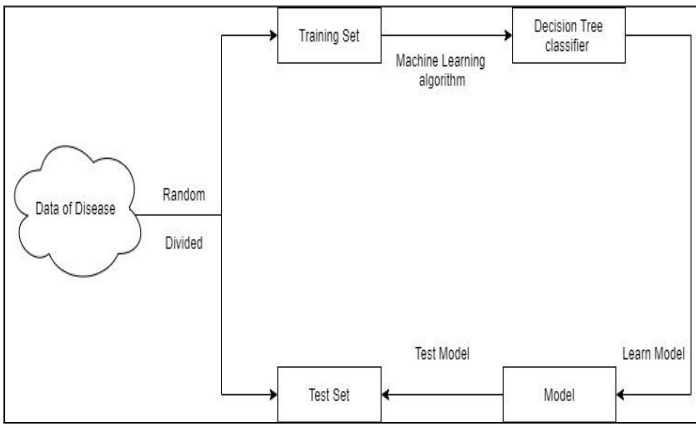


Fig -1: System Architecture

7. ALGORITHM

7.1 DECISION TREE

The decision tree type used in this research is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, therefore maximizing the information gain. Information gain is the contrast between the original information content and the amount of information required. The features are ranked by the information gains, and then the top-ranked features are chosen as the potential attributes used in the classifier. To distinguish the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula:

$$E = \sum_{i=1}^{K} P_i \ log_2 \ P_i$$

Where k is the number of classes of the target attributes Pi is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring). To reduce the effect of bias resulting from the use of information gain, a variant is known as gain-ratio was introduced by the Australian academic Ross Quinlan. The information gain measure is biased toward tests with many consequences. That is, it favours to select attributes having a large number of values. Gain Ratio regulates the information gain for each attribute to allow for the breadth and uniformity of the attribute value.

Decision Trees are supervised learning method used for regression and classification. It learns the simple decision rules after inferring the data features and hence predicts target variable value.

There are various decision tree algorithms like ID3, C4.5, C5.0 and CART. CART is the most recent and enhanced version and hence the same has been used in our model.

(a) Gini impurity It is used by the CART algorithm for classification trees. It is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

(b) Information gain it is used by the ID3, C4.5 and C5.0 tree generation algorithms. It is based on the concept of entropy and information content from information theory. It is used to decide which feature to split on at each step in building the tree.

8. EVALUATING THE MODEL& RESULTS

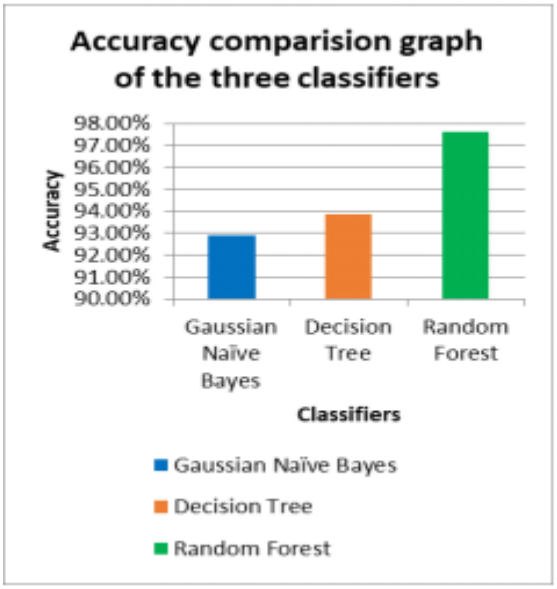
The results obtained from our model are summarized in the following table:

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%

[Table 3: Accuracy comparison table of the three algorithms]

It can be observed that the preprocessing technique, discretization has improved the performance of all the three algorithms. Though Naïve Bayes saw significant increase in accuracy due to discretization, Random Forest gave the highest accuracy for our dataset.

The below bar chart shows that the Random Forest classifier outperforms the other two classifiers and hence is best suited for our dataset:-



[Fig 6: Accuracy comparison graph of the three classifiers]

9. CONCLUSIONS

To conclude, our system is helpful to those people who are always worrying about their health and they need to know what happens with their body. Our main motto to develop this system is to know them for their health. Especially, people who are suffering from mental illness like depression, anxiety. They can come out of these problems and can live their daily lives easily.

Besides, our system provides better accuracy of disease prediction according to symptoms of the user, and also it will provide motivational thoughts and images. In the end, we can say that our system has no boundary of the user because everyone can use this system.

REFERENCES

[1] Pingale, Kedar, et al. "Disease Prediction using Machine Learning." (2019).Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[2] Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease Prediction using Machine Learning.

[3] Aiyesha Sadiya, Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning(2019)

[4] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.

[5] Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom based disease prediction in medical system by using K-means algorithm." International Journal of Advances in Computer Science and Technology 3.

[6] Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data mining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5.4 (2015).