Disease Prediction System Using Machine Learning Algorithms

Lavanya Prasad,
From Computer Science and Technology
Usha Mittal Institute and Technology
Mumbai-400049

Shriya Raina,
From Computer Science and Technology
Usha Mittal Institute and Technology
Mumbai-400049

Aishwarya Salian
From Computer Science and Technology
Usha Mittal Institute and Technology
Mumbai-400049

Professor Prajakta Gotarne
Computer Science and Technology Department
Usha Mittal Institute of Technology
Mumbai-400049, India

***Abstract*— The world is moving at a fast speed and in order to keep up with the whole world we tend to ignore the symptoms of disease which can affect our health to a large extent. Many working professional's get heart attacks, bad cholesterol, eye diseases and they are unable to treat it at the right time as they are busy coping up with progressive world.**

In order to have an early and accurate prediction of disease with correct symptoms and avoid multiple diagnosis provided earlier, we have developed a disease prediction system with the help of machine learning algorithms

We have used Machine Learning algorithms of decision tree, random forest and Naive Bayes which take into account the symptoms felt by a person and according to those symptoms it predicts the disease as a given output which the person can be suffering from. It saves time as well as makes it easy to get a warning about your health before it's too late.

Keywords: Machine Learning, Decision Tree, Random Forest, Naïve Bayes

I. INTRODUCTION

It is both time consuming and costly to see a doctor when you have an illness. It can be difficult for the patient to identify the illness if they are not near doctors and hospitals. If the above procedure can be done using an automated software that saves time and money, it would be better for the patient. Data mining methods can be used to analyze the patient's risk level. Disease Predictor is a system that predicts a user's disease based on their symptoms. There are data sets from various health related websites.

Disease Predictor will allow the consumer to determine the likelihood of a disease based on symptoms. As the use of the internet grows, people are always curious to learn new things. People want to look at the issue on the internet. Hospitals and physicians don't have as much access to the internet as the general public. People with an illness don't have many options. The system can be beneficial to people. Many chronic diseases can't be cured but can be managed with daily treatments, which is why chronic illness is a disease that lasts a long time. India is undergoing significant social and economic shifts which are causing a rise in the number of chronic diseases. Machine learning is the process of programming computers to improve their output. Training and Testing are part of the machine learning algorithm. Predicting a disease based on the signs and medical history of the patient has been a stumbling block for decades. The medical sector has a strong forum for efficiently resolving healthcare issues.

II. RESEARCH OBJECTIVE

One-third of the population in each nation is affected by chronic disease. It is difficult for people who are sick to pay for chronic disease care. In the medical field, a huge number of chronic disease datasets are gathered and processed, and data mining aids in disease early detection. There is a lot of healthcare data that is not being mined in a more efficient and reliable way to uncover secret knowledge for successful decision-making. Data mining techniques can be used to detect chronic diseases early. End users will be able to predict chronic diseases without having to go to a doctor.

To identify diseases by observing the symptoms of patients. There is no proper way to handle text and structured data. The framework would consider both structured and unstructured data. The accuracy of predictions can be improved by machine learning and traditional key-point based methods. A filtering algorithm is used to prune the falsely matched regions and the key-point density is adjusted by an iterative improvement strategy.

III. PROPOSED SYSTEM

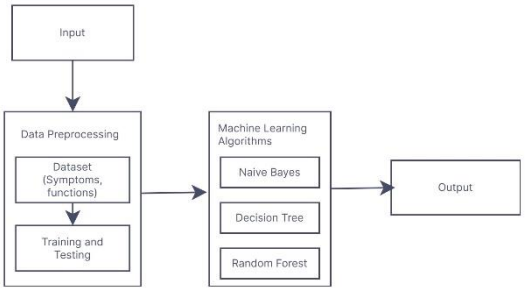

Fig.1. System Architecture

We are predicting a disease based on the symptoms a person is suffering from Random Forest, Decision Tree and Naive Bayes are used to evaluate five symptoms from the patient.

These are steps to model building:

1. The problem statement and requirements are analyzed.

Predicting the disease suffered by a patient is what we want to do by analyzing the symptoms

2. The data is collected and cleaned.

Gathering or recollecting the past data forms the foundation of the future learning, be it the raw data from excel, access, text files etc. The better the variety, density and volume of relevant data, the better the learning prospects for the machine becomes.

3. Preparing the data for an application.

The quality of the data used is an important factor in any analytical process. To fix issues such as missing data and treatment of outliers, one needs to spend time determining the quality of data. Exploratory analysis can be used to study the nuances of the data in detail.

4. A model is being trained.

The data is represented in the form of a model. The first part of the data is used for developing the model and the second part is used for testing. The test data is used as a reference.

5. The model is being evaluated.

The second part of the data is used to test the accuracy. The precision in the choice of the algorithm is determined by this step. Performance on data which was not used in the model build is a better test to check accuracy.

6. The improvement of performance.

This step might involve choosing a different model or introducing more variables. It takes a lot of time to collect and prepare data.

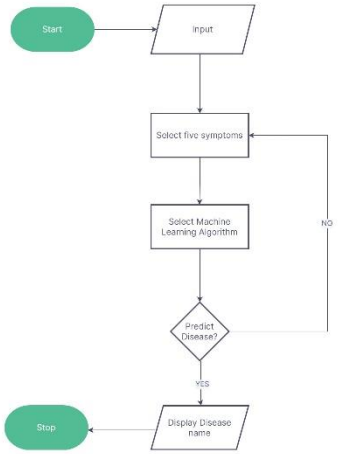

Fig.2. System Flowchart

IV. METHODOLOGY

We have used 3 ML algorithms for our predictive analysis

1. Decision trees
2. Naive Bayes algorithm
3. Random Forest algorithm

Decision Tree

It is a sort of supervised learning program that is used for classification issues. It works for categorical and continuous dependent variables. In this program, we usually split the population into 2 or more sets. The most vital attributes are supported to form distinct teams. A tree has several analogies in the real world, and it seems that it's influenced a large space of machine learning, covering each classification and regression. A choice tree is used in call analysis to represent selections and higher cognitive processes. It uses a model of decisions. It is used in both data mining and machine learning. We will use the trained model to determine whether the balance scale tip to the right or left, or be balanced.

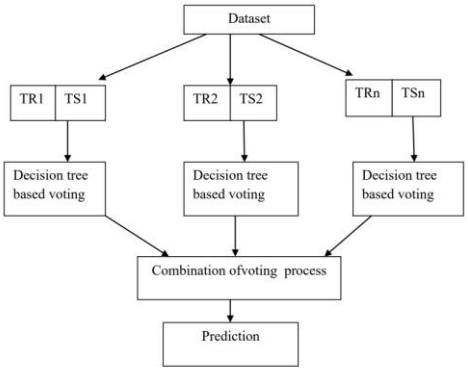

Fig.3. Decision Tree

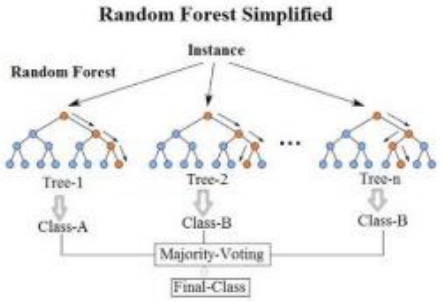

Fig.4. Random Forest

Naïve Bayes Algorithm

The probability of an object with certain features is learned by the Naive Bayes algorithm. If you are trying to identify a fruit based on its color, shape, and taste, then an orange colored, spherical, and tangy fruit would most likely be an orange. The fruit is known as "naive" because of the properties that contribute to the probability that it is an orange. The Nave Bayes Algorithm is named after Thomas Bayes, who is a statistician and philosopher. The following equation is stated by Bayes 'Theorem.

$$P(A/B) = (P(B/A) * P(A)) / P(B)$$

Random Forest Algorithm

Random Forest is a great algorithm to train early in the model development process. On prime of that, it provides a sensible indicator of the importance it assigns to your options. Random Forests are terribly onerous to ram down terms of performance. And on prime of that, they'll handle tons of various feature varieties, like binary, categorical and numerical. Overall, Random Forest may be a (mostly) quick, easy and versatile tool, though it's its limitations. Random Forest can be trained early in the model development process. It gives a good indicator of the importance it assigns to your options. Random forests are hard to measure performance against. They will handle a lot of different feature varieties, like categorical and numerical. Random Forest may be a quick, easy and versatile tool, but it's not perfect. Random forests are an efficient learning method for classification, regression and other tasks, that is operated by constructing a multitude of variant decision trees at training time and outputting the class that is the mode of the categories (classification) & mean prediction (regression) of the individual trees. Random call forests corrects for call trees and their habit of over fitting to their own training set.

V. RESULTS

We evaluated our model over the Accuracy Score and draw inference out of it.

ALGORITHM	Accuracy Score (%)
Decision Tree	0.9512
Random Forest	0.9595
Naïve Bayes	0.9435

TABLE I. Result of different Machine Learning Algorithms

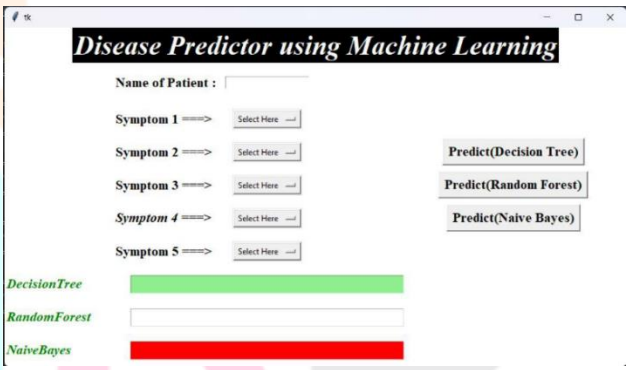

Fig.5. GUI interface with list of 5 symptoms and 3 algorithms

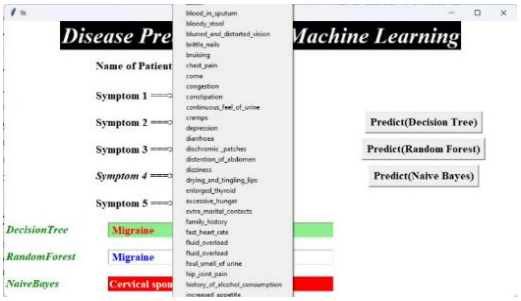

Fig.6. Selecting symptoms from the list

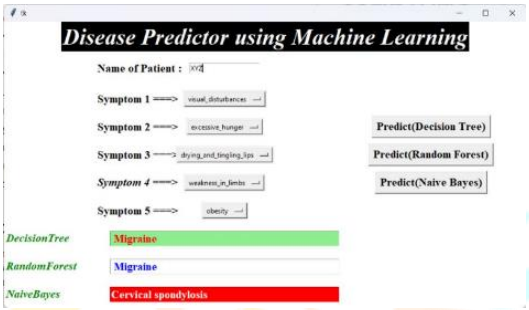

Fig.7. Display of Diseases

VI. CONCLUSION

Predicting disease based on symptoms is the aim of the project. The project is set up in such a way that the device takes the user's symptoms as input and creates an output, which is disease prediction in the future. Learning models should be adjusted more often for better performance. De-identification of personal patient data can be done in order to obtain larger datasets. To avoid overfitting and increase the accuracy of the models, the datasets should be expanded. The learning models should be improved with more relevant feature selection methods**.**

VII. ACKNOWLEDGMENT

It is indeed a great pleasure and proud opportunity for us to present this paper for final year degree at 'Usha Mittal Institute of Technology'. The success of this project has throughout depended upon a combination of hard work and an unending co- operation and guidance provided to us by our project guide. Ultimately no words could describe the deep sense of co- operation and ready nature to help us. We would like to thank, Dr. SHIKHA NEMA

(Principal), Prof. Kumud Wasnik (H.O.D. of CST Department), Prof. Prajakta Gotarne. (Project Guide), who made very valuable guidance and co- operation during our project. Further we are thankful to all the teaching and non-teaching staff of Computer Science and Technology Department for their co- operation during the project work.We are very grateful to those who in the form of books had conveyed guidance in this project work.

VIII. REFERENCES

[1]M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", ," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2]B. Qian, X. Wang, N. Cao, H. Li and Y.G. Jiang, "A relative similarity based method for interactive patient risk prediction," Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070- 1093, 2015.

[3]IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017. ISSN No: 0932-4747 Page No:25 Zeichen Journal Volume 6, Issue 5, 2020

[4]Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyber physical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5]L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016.