International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

PATIENT'S SICKNESS PREDICTION SYSTEM

Author's: Mohd Nazir [1], Mohammad Shahid [2], Mohammad Arbaz Khan [3], Dushyant Sharma [4], Kedarnath Singh [5], Sana Anjum [6]

[1]B.Tech (Scholar) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India
[2]B.Tech (Scholar) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India
[3]B.Tech (Scholar) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India
[4]B.Tech (Scholar) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India
[5]B.Tech (Mentor) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India
[6]B.Tech (Mentor) Department of Computer Science & Engineering, NIET, Greater Noida, Uttar Pradesh, India

Abstract: *As we have seen in the recent years how people have suffered across the globe in the pandemic where corona virus has affected almost every part of the world. Similarly, we too have faced the pandemic and people almost of all age groups were affected because of the restrictions and virus. As the regular OPDs were closed and the beds were occupied it became difficult for the patients specially those who were having their routine checkup at regular intervals so overcome that we got the idea of patient sickness prediction system using multiple Machine learning algorithms. Here in this work on the basis of symptoms, age, and gender of an individual, it gives the output as sickness from which patient might be suffering from early and accurately allows for effective and efficient treatment, which will provide them relief. This paper presents a brief analysis of patient sickness using Naive Bayes technique, Decision tree and Random Forest methodology.*

Keywords: Sickness Prediction and Machine learning

1. INTRODUCTION

The economy and human survival depend heavily on medicine and healthcare. Our current environment has drastically changed from the one we were in couple of year ago. The world has become chaotic and unpleasant. In this virtual world, Doctors, nurses, and other medical professionals are risking their own lives to save as many lives as they can.

There are some remote communities without access to medical care. Physicians with board certification who choose to do their online consultations via the phone and video, albeit this is not viable in an emergency. Since there is no room for human error. Typically, individuals believe that machines are superior to them in terms of speed and accuracy. A disease predictor, sometimes referred to as a practical doctor, can forecast the sickness of a patient with accuracy without the assistance of a person. Since a disease predictor may identify a person's illness without any physical contact, it can be useful in situations like COVID-19 and EBOLA.

Although there are some virtual doctor models, they are not accurate enough since not all the requirements are taken into account. The main purpose was to create a variety of modelsin order to figuring out which one produced the most precise projections. Machine learning efforts range in size and complexity, but the fundamental structure is constant. Numerous rule-based techniques were adapted from machine learning to recall the development and application of the forecasting model. To be able to categorize the raw data into groups according to gender, age, and symptom types, many models were developed using a range of machine learning (ML) approaches.

Fine, Medium, and Coarse Decision trees, Random Forest and naive bayes were used to process the data. The accuracy varies according to ML models. The input parameters data-set was provided to each model during data processing, and the illness was obtained as an output with varying levels of precision. The model with the greatest accuracy was chosen.

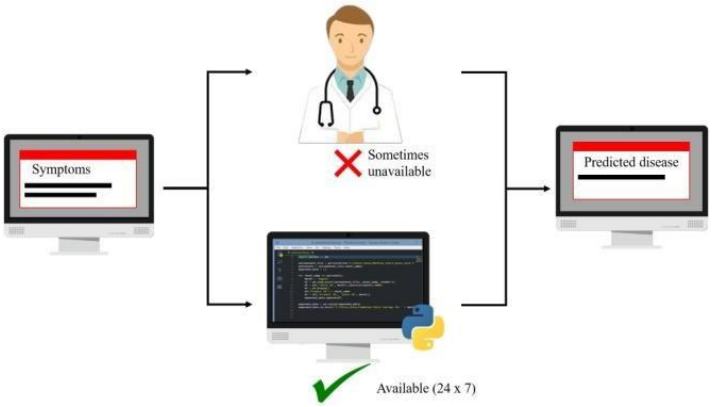

Fig 1.1: Working of Decision Tree

2. LITERATURE SURVEY

To predict diseases based on a person's symptoms, several research have been done using machine learning algorithms.

Whether or if a patient had influenza Among them were 3744 unvaccinated adults and adolescents with fever and at least two additional symptoms were diagnosed with influenza.

Influenza virus 2470 people out of 3744 were confirmed tohave laboratory diagnosis of influenza.
Based on these results, their model has an accuracy of 79%. Sreevalli et al. [7] forecasted the sickness based on the symptoms using the random forest machine-learning method. The method had a little time and expense impact when it came to predicting illnesses. An accuracy of 84.2 percent is achieved by the algorithm. Created a number of techniquesto identify Alzheimer's disease. For the goal of training the ML algorithm, data from 29 people was employed. They built categorization models to find trustworthy absolute changes in scores using the Decision Tree and Random Forest methods.

A hybrid random forest and linear model (HRFLM), which employed machine learning techniques to categories diabetes, was shown to have an accuracy of 88.4% for the prediction of heart disease. Utilizing logistic regression, the risk factors for diabetes sickness were found (LR). The overall accuracy of the machine learning-based approach was 90%.

Dahiwade offered a method of sickness prediction based on machine learning. The UCI ML repository included the symptoms dataset, which comprised signs and symptoms of many illnesses. [2]. CNN and KNN were utilized in the system. As classification approaches for predicting various illnesses Furthermore, the recommended solution was enhanced.

Providing additional details about the people's life habits patient was evaluated, which was beneficial in comprehending the degree of risk associated with the illness to be anticipated Dahiwade and colleagues compared the KNN and CNN algorithms' findings.

In terms of processing speed and precision. CNN processing time was 84.5 percent and 11.1 seconds, respectively. Statistics show that the KNN method is underutilized.

This study's results supported Chen's conclusions that CNN performed better than conventional supervised algorithms like KNN, NB, and DT. The ability of the proposed model to distinguish intricate nonlinear interactions in the feature space led the authors to the conclusion that it was more accurate.

Additionally, CNN unearths highly valuable components that enhance the description of illnesses, enabling it to accurately foresee diseases of vast complexity. Empirical data and statistical analysis provide solid evidence for this view.

However, the models provided lacked variables like network size, architectural type, learning rate, and back propagation mechanism, among other aspects of neural networks.

The published results are also called into doubt because the performance analysis is only assessed for correctness. The bias problem that the tested algorithms face was also not taken into consideration by the authors. For example, adding more feature variables to underperforming algorithms might greatly improve their performance measures.

3. METHODOLOGY AND PLANNING OF WORKING

In this paper the statistical method namely Naive Bayes, Decision tree and Random Forest are taken up for the prediction of the sickness of the patient.

3.1. NAÏVE BAYES

Using the Bayes probability theory as a foundation, it is a machine learning technique for categorization issues. For text categorization, which needs high-dimensional training data sets, this is mostly employed. In our work, we employed the Bayes theorem, which is known as:

$$P\left(\frac{h}{d}\right) = p\left(\frac{d}{h}\right) * \frac{p(h)}{p(d)} \quad (1)$$

Equation 3.1.1: Probability of hypothesis to data.

The probability that hypothesis h will be true given the data d is denoted by P(h/d). This is what is meant by the posterior probability. The probability that data d would be present if assumption h were true is represented by the ratio p(d/h). The likelihood that h is true is represented by p(h) (regardless of the data). The prior probability of h is used to describe this. The data's likelihood is expressed as p(d) (regardless of the hypothesis). This model is providing the best accuracy among all.

The Naive Bayes algorithm, which may be described as follows, combines the two concepts Naive and Bayes:

Naive: It implies that the presence of one feature has nothing to do with the occurrence of another, which is why it is termed naive. An apple is a red, spherical, sweet fruit if the fruit's color, shape, and flavors are used to identify it. Because of this, each trait may be used to identify the apple without depending on the others.

Theorem of Baye:

The Bayes' theorem is a mathematical technique for estimating a hypothesis' likelihood given prior knowledge, sometimes referred to as the Bayes' Rule or Bayes' rule. Determined by the conditional probability.

Bayes' theorem has the following formula:

$$P\left(\frac{A}{B}\right) = P\left(\frac{B}{A}\right) * \frac{P(A)}{P(B)} \quad (2)$$

Equation 3.1.2: Baye's theorem

$P\left(\frac{A}{B}\right)$ indicates the likelihood that hypothesis A will be true given the observed occurrence B (posterior probability).
$P\left(\frac{B}{A}\right)$ stands for the probability that the evidence presented indicates that a hypothesis' chance of being true is high.

$P(A)$ Prior Probability before seeing the data, is the likelihood of a theory.
$P(B)$ indicates Marginal Probability for Probability of Evidence.

The Nave Bayes Classifier in Action:

The following diagram explains how the Nave Bayes' Classifier works.

Consider the following scenario: We have a dataset with the aim variable "Play" and the weather conditions. Therefore, utilizing this dataset, we must decide whether or not to play on a specific day dependent on the weather. The actions below must be taken in order to solve this issue:

Step 1: Using the given dataset, create frequency tables.
Step 2: To generate a table of likelihood, determine the probability of the given features.
Step 3: Apply Bayes' theory to calculate the posterior probability.

3.2. RANDOM FOREST

As suggested by the name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of focusing just on one DT, the RF gathers forecasts from all of the trees. As the number of trees in the forest rises, so does accuracy, and the probability of the RFA overfitting.

Some decision trees may be able to accurately predict the output, while others may not, because the random forest mixes a variety of trees to predict the class of the dataset. When all of the trees are connected together, the expected result is anticipated, nevertheless. The next two assumptions are for an enhanced RF classifier:

In order for the classifier to correctly predict results as opposed to making educated guesses, the feature variable in the dataset must have some real values.

There must be very little connection between the predictions from each tree.

Why should you utilize Random Forest?

The following are some justifications:

i. In comparison to other algorithms, it needs less training time.
ii. It runs rapidly even with a large dataset and makes accurate output predictions.
iii. It is still possible for it to be accurate even when a sizable portion of the data is missing.

The random forest is constructed in two steps: the first is the combination of N decision trees, and the second is the prediction of each tree established in the first step.

The steps of the working procedure are as follows:

Step 1: From generated training set regardless select K data points.
Step 2: For the selected data points, build DTs (Subsets).
Step 3: Choose N for the DTs you want to build.
Steps 1–4 are repeated.
Step 5: Find each DTs predictions for brand-new data points, then provide the brand-new data points to the category that has received the most votes.

3.3. DECISION TREE

Decision tree is used for regression and classification. In the decision tree, for prediction, it uses the method of tree diagram at the peak. It has a root node before being divided in the dominant input feature, which is followed by another split. These pro cases carry on until the very last node, which contains the weights and categorizes the input according on those weights, once all input has been added.

In a coarse tree, each node can only split out a maximum of four times. In a medium tree, there can be up to 20 splits from a single node in contrast. A fine tree allows for a maximum of 100 splits from each node.

Why should you use Decision Trees?

Since there are several methods used in machine learning, this might be challenging. The Decision Tree should be used for the two reasons listed below:
Decision trees are made to be easily understood since they are created to mimic how people think while making decisions.
The decision tree's justification is easy to understand since it takes the shape of a tree.

Consider the following scenario: After receiving a job offer, an offer of employment must be accepted or rejected by the candidate after it has been made. Consequently, in order to solve this challenge, The root node of the decision tree is where it all begins. The decision node eventually divides into two leaf nodes (Accepted letters and Declined letters).

Starting with the decision tree's root node, the process for the determination of a particular dataset's class is carried out. The method co-relates the values of the record attribute to the values of the root property, based on the outcome of the comparison, moves to the next node after following the branch.
The algorithm moves on to the next node after evaluating the attribute value in comparison to the previous sub-nodes. It continues the procedure all the way up to the tree's leaf node. By employing the algorithm below, you may better know the whole approach:
Step 1: S indicates that you should start from the root node because it contains the entire dataset.
Step 2: With the help of Attribute Selection Measure, find out the most crucial attribute in the dataset (ASM).
Step 3: Subsets or Decision Trees of S that contain all the potential values for the best attribute should be generated.
Step 4: Establish the optimum aspect decision tree node.
Step 5: Using the selections from the dataset obtained in step 3, construct more DTs iteratively.

Follow this procedure until all nodes are categorized as leaf nodes and there are no more nodes that can be categorized.

Consider the following scenario: After receiving a job offer, the applicant must determine whether to accept it. The decision tree now starts from the root node to tackle this situation (Salary attribute by ASM). According to the labels that match, the root node divides into one leaf node and the following decision node (distance from the workplace). One decision node (Cab facility) and one leaf node make up the decision node that follows. The decision node divides into two leaf nodes (Accepted letters and Declined letters).

3.4. WORKING ANALYSIS

By providing symptoms by the patient from the generated database, preprocessing of query is done on the provided dataset by the means of some sort of Machine Learning algorithms i.e., Naïve Bayes, Random Forest, Decision Tree. Thus, result in prediction of disease.

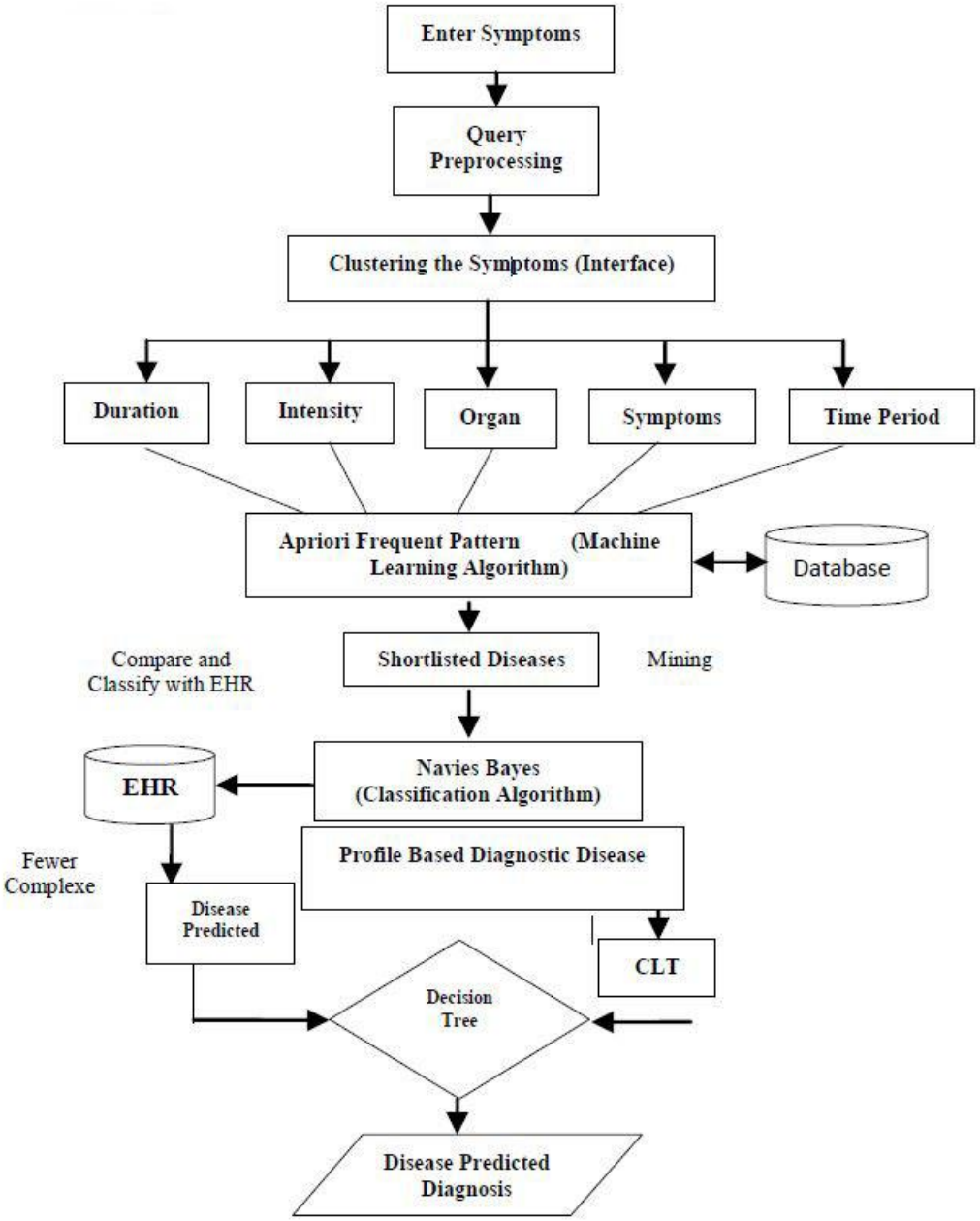

Fig 3.4.1. Working Flow Chart

3.5. DATA COLLECTION

We have generated two dataset files for symptoms and diseases in MS Excel database. Symptoms dataset contains 100 of different body symptoms which helps in choosing the exact indication by the patients that's are occurring in their body. These generated datasets are stored in the local system named DESKTOP-BQKU24Q and location of dataset is E:\Final Project\PROJECT\dataset.

4. RESULTS

This project is presenting the way of predicting the sickness of a patient on the basis of symptoms possessed by him/her. On seeing the result, it is known that Naïve bayes algorithm is providing the highest accuracy of 72.4% in comparison with other two algorithm that is random forest and decision tree. Almost all these models have good accuracy. As few times they were dependent on some parameters lead to which they were unable to predict and the accuracy percentage fall of.

Once the sickness of the patient is predicted one can take necessary actions accordingly for the treatment. This would aid in reducing the expense involved in treating the illness and aid in their recovery.

5. CONCLUSION

With the help of information technology, the health care industry plays a significant role. It is difficult to diagnose an illness using machine learning because of the high risk, vast amounts of information, and noisy and irrelevant data. In this framework input symptoms are stored locally or on server to retrieve the useful information from generated database. The GUI we built is useful for both doctors and regular people. The GUI forecasts the disease stage after receiving the data. The GUI will be turned into a mobile application as part of our future work. Instead of only classifying the condition into five categories, we are attempting to increase accuracy and add a few more levels. Making the primary treatment fast and simply via the use of technology is the overarching goal of system development. As it is often said, prevention is preferable to treatment, thus the system will assist the patient in informing them of their symptoms until a physician responds. In the future, for greater convenience, paramedical recommendations and online doctor consultations can be made.

6. REFERENCES

[1] A. Mir, S.N.Dhage, in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (IEEE).
[2] Y. Khourdifi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12(1), 242 (2019)
[3] S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and naive bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2015)
[4] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access **7**, 81542 (2019)
[5] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar,T.P.R. Vital, Intelligent parkinson disease prediction us- ing machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3), 1568 (2013)

[6] J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine 160(21),3243 (2000)
[7] R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm
[8] D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen,M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length, Clinical genetics 65(4), 267 (2004).
[9] 9.K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I.Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structuralbiotechnology journal 13, 8 (2015)
[10] T. Karayılan, Ö. Kılıç, in 2017 International Conference on Computer Science and Engineering (UBMK) (IEEE,2017), pp. 719–723
[11] 11.M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Dis- ease prediction by machine learning over big data from healthcare communities, Ieee Access 5, 8869 (2017)
[12] 12.S. Chae, S. Kwon, D. Lee, predicting infectious disease using deep learning and big data, international journal of environmental research and public health 15(8), 1596 (2018)
[13] 13.A.U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart diseaseusing machine learning algorithms, Mobile Information Systems 2018 (2018)
[14] 14.M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7(2020).