Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 9262 - 9266
Received 25 April 2021; Accepted 08 May 2021.

Disease Prediction Using Machine Learning

Dr. Saumya Chaturvedi
School of Computing Science and Engineering
Galgotias University
Greater Noida , India
Email: saumyanmishra5@gmail.com

Rishabh Kumar Pandey
School of Computing Science and Engineering
Galgotias University
Greater Noida , India
Email: rishabhpandey288@yahoo.com

Nandan Kumar
School of Computing Science and Engineering
Galgotias University
Greater Noida , India
Email: nandankumar123n@gmail.com

Prince Kumar
School of Computing Science and Engineering
Galgotias University
Greater Noida , India
Email: prince15aug1999@gmail.com

I.ABSTRACT-

The predictive modeling-based "Disease Prediction" framework predicts the user's disease based on the symptoms that the user offers as feedback to the system. The machine takes the user's symptoms as feedback and calculates the probability of the disease.

To predict disease, the Nave Bayes Classifier, Random Forest, and Decision Tree are used. The Nave Bayes Classifier is used to measure the disease risk. As a result, the average accuracy probability of prediction is 60%.

***Index Terms*—**Predictive Modelling, Na¨ıve Bayes Classifier,Random Forest, Decision tree.

II. INTRODUCTION

When anyone is suffering from an illness, they must see a doctor, which is both time consuming and expensive. Since the disease cannot be identified, it would be impossible for the patient if they were outside of the jurisdiction of physicians and hospitals. It would be simple if the above procedure can be done using an automated programme that saves both time and money.
to the patient, which may make the procedure go more smoothly
. Other programmes that predict heart disease use data-processing approaches to determine the patient's risk level. Illness Predictor is a web-based programme that uses the user's symptoms to predict the user's illness.

The Disease Prediction system has gathered data sets from a variety of health-related websites. The consumer will be able to assess the probability of a disease using Disease Predictor based on the symptoms presented.

People are always interested in learning new things, particularly as the internet's use expands. When an issue occurs, people often seek solutions on the internet. Hospitals and physicians have less access to the internet than the general public. When a person is diagnosed with a disease, there are few choices available to them right away. As a result, since people have access to the internet 24 hours a day, this strategy would be effective.

III. RELATED WORK

The study for the simplest diagnosis mining technique was performed by K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman [1]. For this study, the author compared Nave Bayes to five other classifiers, including Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN), and a simple rule-based algorithm (ZeroR).
15 real-world medical problems were chosen from the uci machine learning repository [2] to assess the efficiency of all algorithms. In the experiment, it was discovered that NB outperforms the

opposing algorithms in 8 of the 15 data sets, leading to the conclusion that the predictive accuracy of Nave Bayes is superior to other techniques.

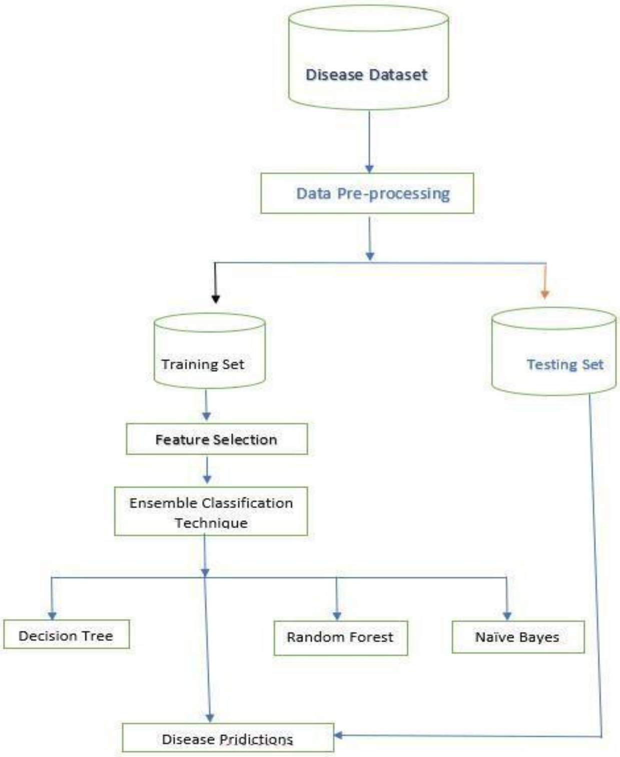


Fig 1: Methodlogy

This research paper was written by Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni [3] to provide a survey of existing techniques of information discovery in databases using data mining techniques that are used in today's medical research, specifically in Heart Disease Prediction.

Indrakumari et al, proposed the heart disease prediction by exploring dataset with the help of tableau [4]. A number of experiments have been carried out to compare the performance of predictive data processing techniques on an analogous dataset, and the results show that call Tree outperforms and that Bayesian classification has comparable accuracy to decision tree on a few occasions, but that other predictive approaches such as KNN, Neural Networks, and classification supported clustering do not perform well.

IV.PROBLEM FORMULATION

There are numerous tools for disease prediction. However, heart-related diseases are examined more and a risk level is calculated. However, there are no such methods that can be used to predict the onset of general diseases.

V. OBJECTIVE

General Objective: To incorporate a Nave Bayes Classifier that classifies diseases based on the user's input. To validate the result with two algorithms for high accuracy, Random Forest and Decision Tree will be used.

Specific Purpose: Build a web-based platform for disease prediction.

Different approaches, such as neural networks, decision trees, and the Nave Byes algorithm, have already been used to predict disease. The majority of research is focused on heart disease. According to the findings, Nave Bayes is more precise than other methods. As a result, Disease Predictor also employs Nave Bayes for disease prediction. For high accuracy, we used Random Forest and Decision Tree in addition to Nave Bayes.

VI. METHODOLOGY

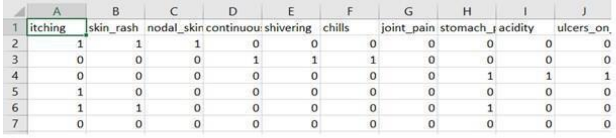
Different methods, such as neural networks, decision trees, and the Nave Byes algorithm, have already been used to predict disease. The majority of research is focused on heart disease. According to the findings, Nave Bayes is more reliable than other techniques. As a result, Disease Predictor employs Nave Bayes for disease prediction. [5] .

For high precision, we used Random Forest and Decision Tree in addition to Nave Bayes. used to foresee the advent of common diseases As a consequence, Disease Predictor helps with disease forecasting in general.

VII. DATA COLLECTION

Data was collected from the internet in order to classify the disease. There were no dummy values used; only the true symptoms of the disease were collected. The signs of the disease were collected from a variety of health-related websites. We begin by downloading a disease dataset from the Kaggle website, which is in the form of a disease list with symptoms.

After that, the dataset is cleaned, which involves removing commas, punctuation, and white spaces. As a consequence, the dataset is used for training. After that, a feature was extracted and selected. Methods such as Decision Tree, Naive Bayes, and Random Forest are then used to classify the data. Machine learning allows us to reliably predict disease.



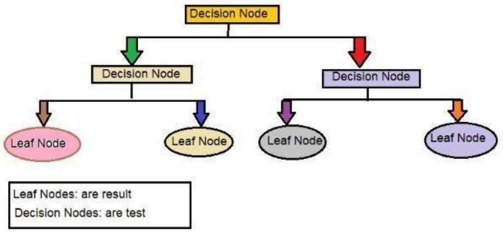
	A	B	C	D	E	F	G	H	I	J
1	itching	skin_rash	nodal_skin	continuou	shivering	chills	joint_pain	stomach_	acidity	ulcers_on_
2	1	1	1	0	0	0	0	0	0	0
3	0	0	0	1	1	1	0	0	0	0
4	0	0	0	0	0	0	0	1	1	1
5	1	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0

Fig 2: Sample Dataset

VIII. MACHINE LEARNING ALGORITHMS

The supervised learning algorithms family includes the Decision Tree algorithm. Unlike other supervised learning algorithms, the Decision tree algorithm is commonly used to solve regression and classification problems. The aim of using a Decision Tree is to build a training model that will predict the target's category or value using basic decision rules inferred from previous data.

When using Decision Trees to predict a record's class mark, we start at the top of the tree. The value of the base attribute is compared to the value of the attribute on the record. Based on the comparison, we follow the branch that corresponds to that value and leap to the next node. Figure depicts the general form of a decision tree.



Decision Tree Diagram

Naïve Bayes:The aim of Nave Bayes is to find the category of observation (data point) based on the values of features in a supervised learning algorithm for classification. A naive bayes classifier is used to measure the likelihood of a class given a set of feature values. (For example, p(yi — x1, x2,..., xn)). In Bayes' theorem, p(x1, x2,..., xn — yi) denotes the likelihood of a particular combination of features given a class name. To do so, we'd like to provide a probability distribution estimate for all possible combinations of feature values in extremely large datasets.

The naive bayes algorithm works around this problem by assuming that each function is independent of the others. Furthermore, since the denominator (p(x1,x2,..., xn)) just normalises the worth of conditional likelihood of a group given an observation (p(yi — x1,x2,..., xn)), it is often omitted to simplify the equation.

Calculating the likelihood of a group (p(yi)) is extremely simple: p(x1, x2,..., xn — yi) can be written as: The conditional likelihood for one feature provided the category mark (i.e. p(x1 — yi)) is often calculated more easily from the knowledge under the assumption of features being independent.

Every class's feature probability distributions must be tracked separately by the algorithm. There must be 50 separate probability distributions stored if there are five groups and ten features, for example. The form of distribution is determined by the characteristics of the distributions. For binary features (Y/N, True/False, 0/1) the binomial distribution is used.

- Multinomial distribution for discrete features (e.g., word counts)
- Gaussian (Normal) distribution for continuous features The distribution of features is commonly referred to as naive bayes (i.e. Gaussian naive bayes classifier)[5]. A special type of distribution may be required for mixed type datasets for various features.

Random Forest

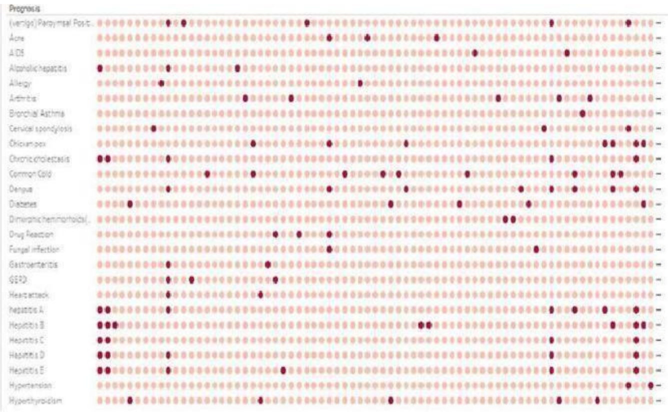
: Random Forest is a supervised learning algorithm. It constructs a "forest" out of a set of decision trees that are typically trained using the "bagging" process. [6]. The basic principle of the bagging approach is that integrating various learning models increases the overall outcome. Random forest has a range of benefits, including the fact that it is widely used for both classification and regression problems, which account for the vast majority of machine learning systems today.

Let's look at random forest in classification since classification is generally considered the building block of machine learning. A random forest's hyperparameters are identical to those of a decision tree or a bagging classifier. Instead of combining a decision tree and a bagging classifier, you can use the random forest classifier-class. You can create a random forest with random forest.

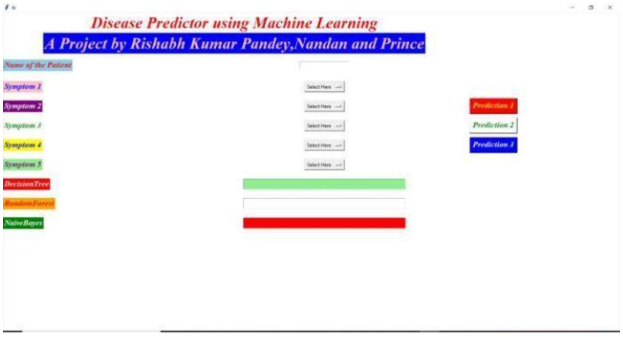
IX. RESULTS

Disease Prediction using Exploratory Data Analysis. Dark Red Circles denotes the symptoms for the corresponding diseases.

SYMPTOMS VS DISEASE GRAPH



X. CONCLUSION FUTURE SCOPE



The aim of this study is to use symptoms to predict disease. The project is set up such that the machine takes the user's symptoms as input and generates an output, which is disease prediction. The average accuracy probability of prediction is found to be 55%.

There were no drug suggestions for the consumer in this project. As a consequence, drug recommendations may be included in the project. A record of a user's disease history can be maintained, and drug instructions can be followed.

The aim of this project is to develop an online platform that can predict disease occurrences based on a variety of symptoms. The consumer will choose from a wide range of symptoms and diagnose diseases based on probabilistic estimates.

REFERENCES

1. Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2012). Medical data classification with Naive Bayes approach. *Information Technology Journal*, *11*(9), 1166.
2. A. Asuncion and D. J. Newman, 2007. “UCI Machine Learning Repository,” http://www.ics.uci.edu/~mlearn/ MLRepository.html
3. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.
4. Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart Disease Prediction using Exploratory Data Analysis. *Procedia Computer Science*, *173*, 130-139.
5. Huang, F., Wang, S., & Chan, C. C. (2012, August). Predicting disease by using data mining based on healthcare information system. In *2012 IEEE International Conference on granular computing* (pp. 191-194). IEEE.
6. Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60).
7. Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.