Volume: 11 Issue: IV Month of publication: April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50500

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Symptoms Diagnosis Using Machine Learning Model Random Forest

Sumit Maurya[1], Vibhur Garg[2], Kaushlendra Sharma[3], Vedant Shukla[4], Deepika Tyagi[5]
[1, 2, 3, 4]*Student,* [5]*Assistant Professor, Dept. of Computer Science and Engineering, Inderprastha Engineering College, UP, India*

Abstract: Symptoms diagnosis is the system based on the prediction method of determining the diseases of the patient based on the symptoms provided by the user. Health is of utmost importance for every living being in this world. As such, we as living beings should do our best to keep ourselves healthy. However, if we suffer from early symptoms, our system analyzes the symptoms provided by the patient and the system will determine what type of disease the person is suffering from. Random Forest is being used as a prediction model to predict the disease. This model is possibly key in increasing the detection rate of diseases at early stages and helps patients take preventive measures at early stages.

I. INTRODUCTION

Machine Learning is the computer programming used for making a system intelligent to take a decision and provide the result according to its experience and data. The system Symptoms Diagnosis is such type of machine learning program, which uses a machine learning algorithm to predict the disease of the patient. According to recently conducted surveys, India’s doctor-to-patient ratio is quite broad and concerning.

As such, this intelligent system can help us overcome this issue. The system helps patients know the disease that might be inhibiting their bodies.

Random Forest is the supervised learning algorithm used that helps efficiently solve complex problems. Random forest algorithm helps in predicting the disease efficiently from the symptoms provided by the user. This system is very beneficial for the health industries as it helps to save the time of doctors as well as the patient. So, this system helps in predicting the disease from the symptoms which when unchecked for a long time can cause fatality also our System help users to find precaution for the diseases and information about the disease. Thus, with the help of machine learning techniques, we created this system.

II. LITERATURE REVIEW

Various studies have been considered for the prediction of various types of diseases. Some of them are given below:

1) Vincy Cherian designed a heart disease system using the data mining classification algorithm naïve Bayes and the smoothing technique Laplace smoothing. The system used attributes such as age, gender, cholesterol, blood pressure, etc. to predict heart disease. Their main concern was to save time and money for doctors and patients by predicting heart disease.
2) Dr. J Selvakumar designed a system using data mining, big data, and machine learning algorithm. According to him, heart disease is a major cause of fatality for humankind. The system helps predict heart disease and cardiovascular disease. The system helps the medical field discover cardiovascular disease before it manifests itself.
3) Sonam Nikhar conducted a study to create a system based on naïve Bayes and decision trees to predict heart diseases in human beings. According to him, a hybrid model of a decision tree and a naïve Bayes algorithm is the most efficient ML algorithm to predict the disease.
4) Dr. S. Vijayarani created a system to predict the liver disease of the patient. According to his paper, a data mining classification algorithm, naïve Bayes, and a support vector machine were used to develop this system. The research papers concluded that the support vector machine is better than naïve Bayes in comparison to accuracy and prediction time.
5) Pushkar Patil designed a system that is used to predict heart disease using various prediction algorithms and selected the best algorithm for their system. Various attributes such as chest pain, discomfort, and faster heartbeat are some of the symptoms related to heart disease.
6) Ali Bou Nassif created a system, Heart disease prediction using machine learning. Cardiovascular disease refers to a condition that impacts the Heart. The Paper demonstrates different algorithms such as support vector machine, naïve Bayes, and multilayer algorithm and according to it, vector machine performs best with a high rate of accuracy of 91.6%.

7) Sohel Rana proposed a system "Prediction of Hepatitis Disease using K-nearest neighbors, Naïve Bayes, Support Vector Machine, Multi-layer perceptron, and random Forest". Hepatitis is one of the diseases which cause lots of death. Various algorithms are used to predict hepatitis disease with calculation and accuracy. The paper concludes that the random forest algorithm has high accuracy than other algorithms.
8) Muhamad Huzaimi Bin Abdul Ghafar designed a system "Chronic Kidney Disease based on data mining method and support vector machine". The paper aimed to predict the stages suffered by chronic key disease patients. The early detection of the disease helps the patient to prevent disease. Classification-based models and feature selection are used to detect disease. Support Vector Machine achieves high accuracy among other models.
9) Vinayak Singh proposed a system "Performance Analysis of Machine Learning Algorithm for Prediction of Liver Disease". There are many causes for liver disease such as alcohol consumption, drugs, food, and other cause. The paper defines different machine-learning algorithms for the prediction of liver disease. The major goal of the system is to predict disease at right time and helps in the prevention of disease.

III. PROPOSED SYSTEM

After evaluation of various methods, we used the machine learning algorithm 'Random Forest' in the programming language Python and developed the UI in the same environment and named it "Symptoms Diagnosis". The working of the system is done through the collection of data (such as diseases and their symptoms). The collected data is then cleaned and processed before it is used for further purposes. The data is then divided into two segments; testing and training. Different algorithms were used to predict diseases and we concluded that 'Random Forest' fetched us the highest accuracy from other algorithms.

IV. SCOPE OF PROJECT

Disease has gotten great attention in the field of medical science. The diagnosis of disease plays a major role in health industries. If not diagnosed at right time can cause a major problem to the health of individuals. The diagnosis can be done based on symptoms and signs. Our project helps patients to diagnose their disease based on symptoms in their homes. This helps patients to know from which disease he/she may be suffering. The system targets the common people who can't afford fees for meeting doctors. This System also helps in saving time for individuals and doctors also.

V. METHODOLOGY

A. *Dataset Testing And Training*

Health industries generate lots of data every year. For building the system, we required some datasets to implement our model. This dataset is useful for making predictions based on historical data using a machine learning algorithm. As such, we used Python's library 'Pandas' to read our CSV files and made them readable for our machine learning algorithm. Pandas is a Python library used to read the dataset it helps to analyze big data. Pandas also help in cleaning and manipulating data to make them readable.

B. *Model Building*

After importing the dataset using Pandas, we implemented our Machine Learning algorithms such as naïve Bayes, random forest, and ADA boost, to predict the disease based on the user's input. The System takes user input and based on testing training data predicts the disease.

1) *Naive Bayes:* Naïve Bayes algorithm is a supervised learning ML technique. It is a simple and effective classification algorithm based on the Bayes theorem. The Naïve Bayes algorithm is used for classification problems. The Naïve Bayes algorithm is a probabilistic model, which predicts based on the probability of the object. Naive Bayes classifiers are intensively scalable, taking some parameters direct in the number of variables in a literacy problem [3]. As the Naïve Bayes algorithm is based on Bayes theorem, Bayes theorem can be defined by the given formula: $P(X|Y) = (P(Y|X) * P(X)) / P(Y)$
2) *Random Forest:* Random Forest is a supervised machine learning algorithm that is used for classification and regression problem-solving in machine learning. This algorithm is used to improve the model by using multiple classifiers to solve the complex problem. Random forest is useful because it takes less time compared to other algorithms and also it helps maintain the accuracy of our prediction model [7]. The accuracy of the algorithm depends upon the number of trees in it, a higher number of trees leads to higher accuracy.

3) *AdaBoost:* AdaBoost algorithm was proposed by Robert Schapire and Yoav Freund in 1996. AdaBoost stands for adaptive boosting. It is a boosting technique that combines multiple weak classifiers into a single strong classifier and gave its output. In layman's terms, it means that weak learners are converted into strong learners. The concept of this algorithm is to provide weights to the classifier and train the dataset in each iteration to produce the output of the unusual behavior.

C. Model Deployment

After the completion of dataset training, testing, and implementation of the model, we require an interface through which a user can interact with the system. We used a Python library 'Tkinter' to design our interface. Tkinter is a library used in Python to create a graphical user interface simply and easily. Tkinter consists of inbuild models to create the interface of applications.

D. Visual Studio

For the deployment of all the modules we have used visual studio as a tool. Visual Studio is a platform for the code editor, we can edit, debug, and build code. Visual Studio provides many features like completion tools, graphical design, and many more features to enhance the development process. Visual Studio helps to develop any type of application. Thus, the visual studio provides such type of platform where we can do our work and develop an application that can be further published.

E. Weka

Weka is a collection of machine learning algorithms used for solving real-world problems. Weka consists of a visualization tool used for analyzing data. It contains various tasks such as data preprocessing, clustering, etc. Weka provides a visualization tool to inspect the data. Weka helps in the quicker development of the machine learning model. Through the Weka tool, we can measure the performance of the model such as accuracy, confusion matrix, etc.

VI. SYSTEM REQUIREMENT

A. Hardware Requirements

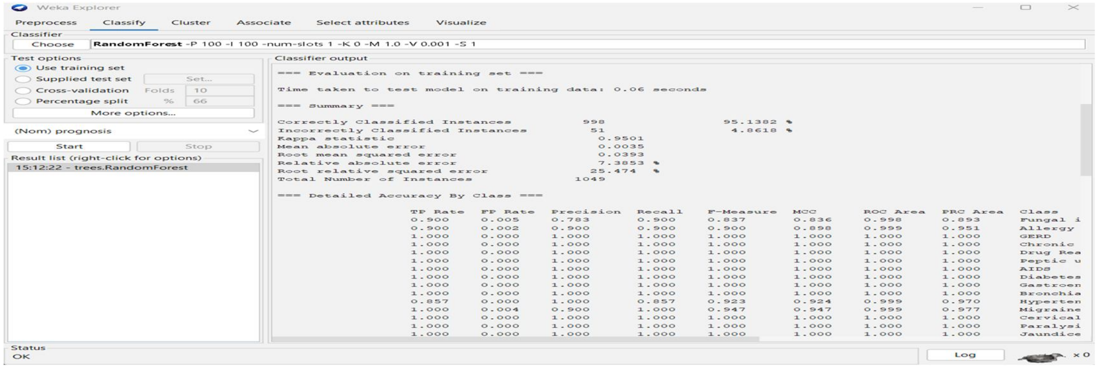
Processor: Any Update Processer
Ram: Min 4GB
Hard Disk: Min 100GB

B. Software Requirements

Operating System: Windows family
Technology: Python3.7
IDE: Jupiter notebook/visual studio

VII. RESULT

Our System "SYMPTOMS DIAGNOSIS" successfully predicts the disease according to the symptoms provided by the user. By comparing and checking the accuracy of different algorithms, we concluded that the random forest algorithm is the best algorithm for prediction with a high accuracy score of 95.13%. The following figure represents the result of the proposed system on WEKA 3.8 in terms of performance such as accuracy.



Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options:
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) prognosis

15:12:22 - trees.RandomForest

Status: OK

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances         998               95.1382 %
Incorrectly Classified Instances        51                4.8618 %
Kappa statistic                          0.9501
Mean absolute error                      0.0035
Root mean squared error                  0.0393
Relative absolute error                  7.3853 %
Root relative squared error             25.474  %
Total Number of Instances             1049

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.005    0.783      0.900    0.837      0.836    0.998     0.893     Fungal i
                 0.900    0.002    0.900      0.900    0.900      0.898    0.999     0.951     Allergy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     GERD
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Chronic
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Drug Rea
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Peptic u
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     AIDS
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Diabetes
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Gastroen
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Bronchia
                 0.857    0.000    1.000      0.857    0.923      0.924    0.999     0.970     Hyperten
                 1.000    0.004    0.900      1.000    0.947      0.947    0.999     0.977     Migraine
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Cervical
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Paralysi
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Jaundice
```

Fig. 1: Proposed Model Accuracy on Weka 3.8

Now, after evaluating the accuracy of different algorithms, the result is arranged in form of an accuracy table and graph of accuracy. Here, the following accuracy table gives the summary of the accuracy of different algorithms obtained from the instance of the dataset. This graph helps in clear visualization. The Y-axis of the graph has accuracy values and X-axis has names of different algorithms.

Algorithm	Accuracy
Random Forest	95.13%
Multinomial Naïve Bayes	93.94%
Logistic Regression	94.4%
Decision Tree	94.47%

Table 1: Accuracy Table

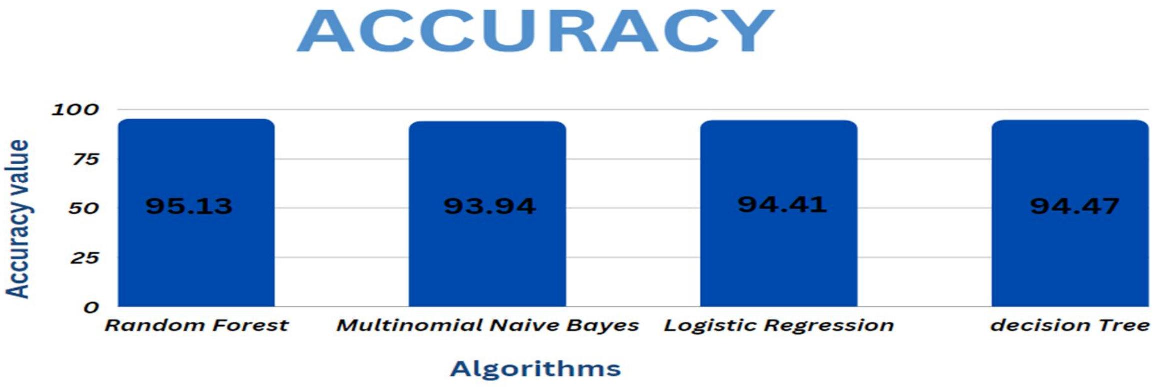
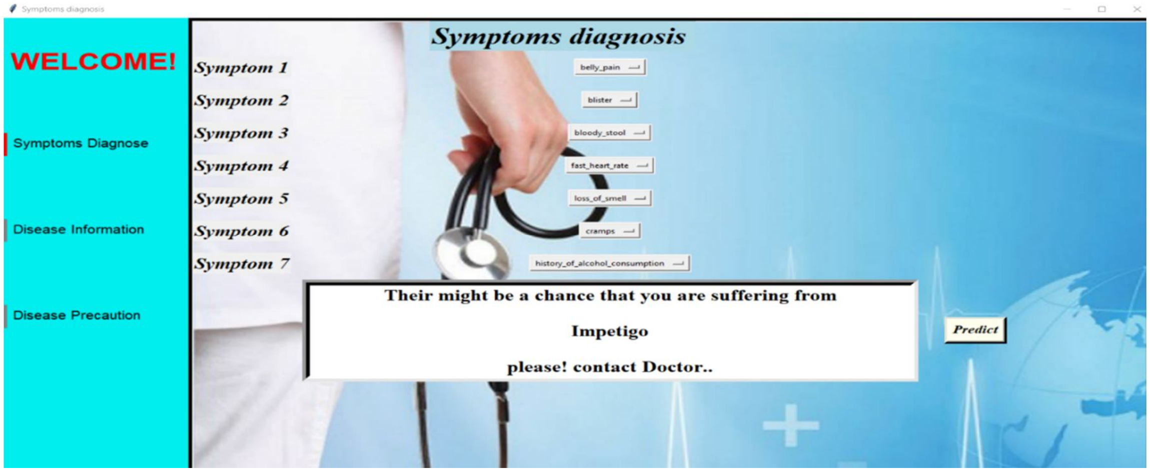


Fig. 2: Accuracy Graph


Symptoms diagnosis

WELCOME!

Symptoms diagnosis

Symptom 1 belly_pain

Symptom 2 blister

Symptom 3 bloody_stool

Symptom 4 fast_heart_rate

Symptom 5 loss_of_smell

Symptom 6 cramps

Symptom 7 history_of_alcohol_consumption

Their might be a chance that you are suffering from

Impetigo

please! contact Doctor..

Fig. 3: Implementation shot of Symptoms Diagnosis

VIII. CONCLUSION

In this paper, we have provided a system that analyzes and predict the disease(s) based on the symptoms provided by the patient. The main purpose of the system is to provide early context as to what sort of disease the patient might be inflicted with and potentially save lives in case of early detection of a fatal disease. Also, the system helps the user to get a description of the disease by getting disease information and disease precaution. As for our further goals with the project, we aim to implement this system on a cloud server. By providing this system on a cloud server, we will vastly improve the reachability of this system and help a lot more people in need.

REFERENCES

[1] Vincy Cherian and Bindu M.S, “Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique”, Journal of International Journal of Computer Science Trends and Technology (IJCST), Vol. 5, No. 2, Mar-Apr 2017.

[2] M Preethi and Dr. J Selvakumar, “A Survey of Predicting Heart Disease”, Journal of INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION, VOL. 4. NO. 2, 2020

[3] Sonam Nikhar and A.M. Karandikar, “Prediction of Heart Disease Using Machine Learning Algorithms”, Journal of International Journal of Advanced Engineering, Management and Science (IJAEMS), Vol-2, No. 6, June- 2016.

[4] Dr. S. Vijayarani, and Mr. S. Dhayanand, “Liver Disease Prediction using SVM and Naïve Bayes Algorithms”, Journal of International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, No. 4, April 2015.

[5] Mangesh Limbitote, Dnyaneshwari Mahajan, Pushkar Patil Pimpri, and Kedar Damkondwar, “A Survey on Prediction Techniques of Heart Disease using Machine Learning”, Journal of International Journal of Engineering Research & Technology (IJERT), Vol. 9, No. 06, June-2020.

[6] Chaimaa Boukhatem, Heba Yahia Youssef, and Ali Bou Nassif, “Heart Disease Prediction Using Machine Learning”, Institute of Electrical and Electronics Engineers (IEEE), march 2022.

[7] Sohel Rana, Md. Julker Nayeem, Farjana Alam, and Md. Ataur Rahman, “Prediction of Hepatitis Disease using K-Nearest neighbor, Naïve Bayes, Support Vector Machine, Multi-Layer perceptron, and Random Forest”, Institute of Electrical and Electronics Engineers, April 2021.

[8] Muhamad Huzaimi Bin Abdul Ghafar, Nurul Aleena Binti Abdullah, Abdul Hadi Abdul Razak, Megat Syahirul Bin Megat Ali, and Syed Abdul Mutalib Al-Junid, “Chronic Disease Prediction based on data mining method and Support vector machine”, Institute of Electrical and Electronics Engineers, 17 December 2022.

[9] Vinayak Singh, Mahendra Kumar Gaourisaria, and Himansu Das, “Performance Analysis of Machine Learning Algorithm for Prediction of Liver Disease”, Institute of Electrical and Electronics Engineers, September 2021.