Disease Prediction using Machine Learning based on User Symptoms

Ananya Ubarhande, Kalpit Vyas, Sahil Sapar, Rutuja Jadhav ,
Prof. Rupali Waghmode
Department of Computer Engineering ,Zeal college of engineering and research Pune, India

Abstract

Nowadays, we live in a world where people suffer from various diseases due to conditions like environmental changes and various dangerous experiments like Covid-19. So the prediction of disease at an earlier stage became the most important task and correct prediction of disease is the most challenging task. Here we proposed a system that can identify diseases based on symptoms. There are many applications like online consultation with a doctor but there is less number of application that predict diseases. There are systems where only one disease can be predicted by using the symptoms but in our system, we can predict forty-one common diseases. For prediction purpose, we required some symptom's dataset. With the help of the disease symptom dataset from the Kaggle platform, we find a huge amount of information to process in this system for accurate prediction. For the prediction purpose, we use three algorithms of machine learning which are Naïve Bayes, Random Forest, and Decision Trees. After prediction of disease, the system will provide detail information about diseases and provide the option to book an appointment for a required doctor which is also a part of our system where patient selects doctor for his/her treatment and book their appointments accordingly.

Keywords— Machine Learning, *Diseases*, Decision Tree classifier, Random forest classifier, Naive Bayes classifier, Symptoms

I. INTRODUCTION

The healthcare and medical sector are more in need of data mining today. When certain processing methods are used in the right way, valuable information is often extracted from a large database which may help the medical practitioner to wish for early decision and improve health services. Machine Learning helps in prediction by its emerging approach for the diagnosis of a disease. This paper depicts the prediction of diseases by using patient symptoms. Machine Learning algorithms like Naive Bayes, Decision Tree, and Random Forest are used for predicting diseases. Its implementation is completed through the python programming language.

Diseases and health-related problems like malaria, Impetigo, Diabetes, Migraine, Jaundice, Chickenpox, etc., cause a significant effect on one's health and sometimes might also lead to death if ignored. The healthcare industry can make a decent deciding making by, "mining" the large database they possess i.e. by extracting the hidden patterns and relationships within the database. Data processing algorithms like Decision Tree, Random Forest, and Naive Bayes algorithms can provide a remedy to the current situation. Hence, we have developed an automatic system that can discover and extract hidden knowledge associated with the diseases from a historical (diseases-symptoms) database according to the rule set of the algorithms.

II. LITERATURE SURVEY

In [1]: Sneha Grampurohit and Chetan Sagarnal have conducted research and implementation using Naïve Bayes Algorithm, Random Forest Algorithm, and Decision Tree algorithm to predict the common diseases where the user provides the information which is compared with a trained set of values. So from this research, patients were ready to provide their basic information which is compared with the information and therefore the relevant disease is predicted. But they were facing variety of the issues in result which are mentioned by them during this paper.

In [2]: The overall day to day health of someone is important for the efficient functioning of the flesh. Taking certain prominent symptoms and their diseases to create a Machine learning model to predict common diseases supported real symptoms is that the target of this research with the dataset of the foremost most typically exhibited diseases. With the assistance of text processing input is combined with various algorithms to check the similarities to urge output

In [3]: Machine learning when implemented in healthcare can results in increased patient satisfaction. During paper, we try to implement functionalities of machine learning in healthcare in a very single system. Rather than diagnosis, when a disease prediction is implemented using certain machine learning predictive algorithms then healthcare will be made smart. Some cases can occur when early diagnosis of a disease isn't handy. Hence disease prediction can be effectively implemented. As widely said “Prevention is best than cure”, prediction of diseases and epidemic outbreak would result in an early prevention of an incident of a disease.

III.DATA SUMMARY AND PROCESSING

1. *Dataset Preparation*

The data we used was gathered from Kaggle websites. In this, we discover some dataset with diseases and its symptoms. For the most testing and dealing model, we use a dataset with 41 diseases and their symptoms because we want as many as possible diseases. This data proved to be more accurate for prediction since it had been to collect an inventory of the foremost appropriate words which uniquely predict disease or condition. We test this data with three main algorithms which are Decision Trees, Random Forest, and Naïve Bayes.

2. *Text Processing*

The model is straightforward therein it ignores the order of words and relations rather than concentrate on the occurrence of words in the dataset. For all three algorithms, we use some parameter which helps in keyword extraction to assist in faster computation. The Term Frequency and Inverse Document Frequency plays an important role in prediction and understanding the dispersion of symptoms. This shows how the dataset has keywords that modify in occurrences supported a specific disease.

IV. METHODS

We test the information with 3 main classification algorithms Decision Trees, Random Forest and Naïve Bayes. We make use of the sci-kit learn library and pandas data frames of the Python linguistic communication to process the information and implement the above algorithms

A. Decision Trees

With the assistance of the Decision Tree classifier, we can derive the pairs of diseases and generate a tree-supported Gini index. The primary step is to classify the information to suit the model of decision trees for the given dataset. We split the information into test data and training data for the model. Construct a node table to assign the various classifiers and Gini for splitting the nodes. Classify the model using Decision Tree Classifier. The worth of the target variable has to be predicted using simple decision rules created using within the dataset. The sole extra advantage of using this is often to use it for both numerical and categorical dataset classification. This helps in reducing data cleaning supported what form of data is employed for processing for every of the algorithm. Statistical testing is additionally easy compared to other methods. It uses three main criteria for determining the right split. Gini Index, Information gain and Entropy. The Gini Index is subtracted sum of the squared probabilities of every class from one. Information Gain specifies the all-time low entropy for every split and accordingly produces each node for the lowest entropy calculated for every split.

B. Random Forest

This is often a more enhanced version of decision trees where we pick N random records from the dataset which contains symptoms. It's a supervised algorithm within which multiple decision trees are built with the assistance of the bagging method. It does not depend upon the foremost important features rather it uses a random subset of features is taken into consideration while splitting a node. There is very low or no bias since it relies on the flexibility of "the crowd". It follows 4 main steps:

1) Pick random samples from the dataset.

2) Produce a decision tree using each sample and acquire a predicted result from each tree.

3) For every result, gain a vote to predict the result.

4) The last word prediction with the most overall votes gained is going to be the result.

C. Naïve Bayes

This is a contingent probability-based algorithm. It's the foremost widely used and fastest algorithm since it uses less training data and powerful independence assumptions. In our case, we use an in-built function called Multinomial Naïve Bayes which is mainly used for discrete features like text classification. It requires a feature count parameter which helps in determining each class while fitting the sample with appropriate weights. In-text classification, the most aim is to search out the most effective class for the given document. It has 2 main functions. The primary function is used to train the multinomial Naïve Bayes model supported by the feature extraction and count vector. Each of the counts of words and Document will be done worn-out single experience this training data. The probability of Vector in each case is returned to assign a score to every term collected as bag-of-words. The ultimate score is obtained because of the cumulative score for the given document.

V. PREDICTION MODEL

A. Input from the user (Symptoms)

While designing the model we've assumed that the user contains a clear idea about the symptoms he's experiencing. The Prediction developed considers 95 symptoms amidst which the user can give the symptoms his processing because of the input.

B. Data preprocessing

The data mining technique that transforms the information or encodes the information to a form that may be easily interpreted by the algorithm is named data preprocessing. The preprocessing techniques utilized in the presented work are x Data Cleaning: Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies within the data. X Data Reduction: We have also replaced missing value with actual value and we also remove orthographical errors from dataset.

C. Models selected

This system aims to predict the diseases with the help of three algorithms:

- Disease Tree Classifier
- Random Forest Classifier
- Naïve Bayes Classifier

A study is presented to inspect the performance of every algorithm of the considered database.

D. Output (diseases)

Once the system is instructed with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model, those symptoms are processed according to the rule set developed, thus making classifications and predicting the foremost likely disease.

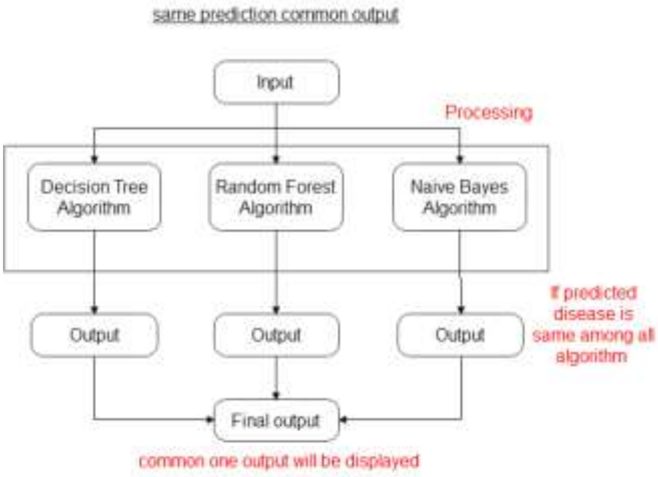
same prediction common output
Input
Processing
Decision Tree Algorithm
Random Forest Algorithm
Naive Bayes Algorithm
Output
Output
Output
If predicted disease is same among all algorithm
Final output
common one output will be displayed

Fig. prediction model

VI. CONCLUSION

The knowledge germination concerning algorithms and algorithms utilization in the medical domain it can be noted that methods and methodologies have developed that have allowed complicated data reports by mild and honest use of algorithms like disease tree classifier, random forest classifier, naive Bayes classifier. This paper introduces a comprehensive similar study of three algorithms completion on a medical record the execution analyzed is done with an accuracy score artificial intelligence will perform an even more significant role in data interpretation in the future due to the availability of enormous data produced and deposited by advanced technology.

REFERENCE

1. "Disease Prediction using Machine Learning Algorithms" Sneha Grampurohit, Chetan Sagarnal 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020

2. "Symptom Based Health Prediction using Data Mining" Vijaya Shetty S, Karthik G A, M Ashwin Proceedings of the Fourth International Conference on Communication and Electronics Systems (ICCES 2019) IEEE Conference Record # 45898; IEEE Xplore ISBN: 978-1-7281-1261-9

3. Y.Deepthi,K. Pavan Kalyan,Mukul Vyas,K. Radhika,D. Kishore Babu,
N. V. Krishna Rao "Disease Prediction Using Machine Learning".

4. Shadab Adam et.al "Prediction system for Heart Disease using Naïve Bayes", International Journal of advanced Computer and Mathematical Sciences, ISSN 2230- 9624, Vol 3,Issue 3,2012,pp 290- 294[Accepted- 12/06/2012]

5. Sohail M.N., Jiadong R., Uba M.M., Irshad M. (2019) A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews. In: Patnaik S., Jain V. (eds) Recent Developments in Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing, vol 752. Springer, Singapore.