Disease Prediction using Machine Learning

G. Vasu Sena, K. Rajinikanth, Mohammed Khaja Faizan* and D. Rohit Rajan

Student, Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology
*Corresponding author: fabulousfaizan1234@gmail.com

Abstract. Predicting disease at an early stage becomes critical and the most difficult challenge is to predict it correctly along with the sickness. The prediction happens on the basis of the symptoms of an individual. The model presented can work like a digital doctor for the disease prediction which helps to diagnose the disease timely and can be efficient for the person to take immediate measures. The model is much more accurate in prediction of potential ailments. The work is tested with four machine learning algorithms and got the best accuracy with Random Forest.

Keywords: Machine Learning, Random Forest, Disease prediction, Naïve Bayes.

INTRODUCTION

The main goal of our project is to provide the disease name by taking the symptoms from the user or patients. Nowadays everything is available on the internet so we thought of predicting the disease based on the symptoms which are given by the customer through online. It’s an interacting system that takes symptoms from the customer. The customer has to provide a minimum of 2 symptoms which he is suffering from.

The system responds effectively Graphical user interface to make (GUI) it look like or it feels like its live interaction. You can create this type of disease prediction using an machine learning algorithms and as well as artificial algorithms to enquiry, identify, and respond to the costumer.

Random Forest Algorithm

1. Random forest selects k number of records randomly from data having m records.
2. A separate decision tree is created for each sample.
3. Output produced from every decision tree.
4. The result is taken based averaging for Classification and regression respectively. Random forest is considered as one of the effective algorithms used in classification.

Decision Tree Algorithm

Decision trees are commonly employed for classification. A decision tree is a classifier with a tree structure in which features being represented by internal nodes and branch of tree represents decision rules. The decision tree has two nodes. The judgment or test is made based on the dataset’s properties.

Naive Bayes Algorithm

The algorithm which is used in classification for binary and multiclass is Naive bayes algorithm. Naive bayes algorithm is very simple and easy to understand and moreover the naive bayes algorithm provide good output for wide range of output. P(class1 | data1) = P(data1 | class1) * P(class1) / P(data1) with the help of naives bayes algorithm the we can calculate the probability of a piece of data belonging to a given class.

K-Nearest Neighbour (KNN)

A pattern has been found to link the data and results which helps in improving the recognition with each iteration. It involves the following steps

1. we need to load the required data
2. we need to calculate distance between points which is called as Euclidean distance
3. we have top k top distances.

Python was chosen for a variety of reasons. It is dependent on your perspective and background. It’s made for programmers. One of the most well-known programming languages is Python. Python is one of the easiest programming languages to learn. It’s quite simple and we can use
the grammar language in it as a syntax. Python is one of the high-level languages which has an inbuilt garbage collector which is used to free up the memory from the elements which are not used in the code.

Bits and Pieces together

This approach can utilize the already done work by utilizing it as a starting point. All the information of accomplished work can be combined together.

Equations

The equations should be inserted in editable format from the equation editor.

\[ f(x) = a_0 + \sum_{n=1}^{\infty} \left( b_n \cos \frac{n\pi x}{s} + c_n \sin \frac{n\pi x}{s} \right) \]

PROPOSED SYSTEM

In this model we (GUI) takes the symptoms from the user and it predicts the disease he is suffering from. The interface responds immediately in fraction of time with accurate accuracy.

- The user has to fill the details like his name
- The user has to enter the symptoms his suffering at least 2 symptoms
- The System will store the data like his name and the disease he is suffering from so that next time it will be easy and fast to cure him

METHODOLOGY

A methodology is a representation of a system’s structure, behaviour, and other features. A system architecture is made up of system components and subsystems that interact to form the total. Individuals use an architecture diagram to abstract the overall structure of a software system and define constraints, linkages, and boundaries between components. The methodology of the work is shown in Figure 1.

Python has many applications. Some of them are:

- **Web Development:**
  Many of web development projects uses python because python has introduced a lot of framework which makes work easier and simple and attractive

- **Data Science:**
  Data science itself involves so many stages like data mining, data sorting, data processing etc. So python provides inbuilt function which makes work easier and simpler to work

RESULTS

This dataset was acquired from a Kaggle reference. Here in the data set we have 5000 rows of data which helps in training model very efficiently. (shown in Figure 2)

The testing data has nearly 45 rows which helps in calculating accuracy. (shown in Figure 3)

Figure 4 is showing interface of the disease prediction scenario and Figure 5 is the final result achieved after providing symptoms in the interface.

The work is being done with 4 Machine Learning Algorithms i.e. Decision Trees, Random Forest, KNN, and Naïve Bayes. Best result was achieved with Random Forest algorithm. Comparison of all classifiers is shown in Figure 6.
Figure 2. Training set.

Figure 3. Test set.

Figure 4. Interface for prediction of disease.
CONCLUSION

After completing the work we can conclude that random forest predicts the disease with high accuracy after random forest it is decision tree that gives one of the best accuracy. We have created a system that can decrease the rush at hospitals and medical areas and the it also helps in reducing the workload on the medical staff. So our system is helpful for both the patients and as well as medical fields. By building such type of systems we can save time and as well as money spent by the patients to undergo tests or scanning to know what they are suffering from. On an average our system achieved on accuracy of 93% on editing diseases with the symptoms given by user with random forest algorithm. Creating this system we also added a way to store the data entered by the user in the database which can be used in future to help in creating a better version of such a system.

REFERENCES