Disease Indicator

Name: Ibtasam Amjad
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs104@student.uet.edu.pk

Name: Sana Parveen
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs128@student.uet.edu.pk

Name: Muhammad Saddam
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs142@student.uet.edu.pk

Name: Ali Shahid
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs111@student.uet.edu.pk

Name: Nabeel Shafiq
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs140@student.uet.edu.pk

Name: Ibrar Hussain
Department: Computer Science
Organization: UET
City: Lahore
Email: 2018cs145@student.uet.edu.pk

***Abstract*—To fall in any disease, is a part of human life. Sometimes we get sick, is normal, and sometimes we need to visit the doctor to cure ourselves. In this fast and high-tech era, the power of computer also helps in this medical domain. This study focuses on how computers help in predicting the disease in some person with maximum accurate results. Also, researchers set up practical implementation and discuss the outcome of those experiments. The supervised machine learning algorithm is being used to predict the disease using user input symptoms.**

I. Introduction

When we observe that a person has to wait for its test report of disease diagnosis, have to go to hospital. So, it will be very effective and helpful for the people, just by providing correct symptoms they will get their disease indication staying at home in just a few clicks over the internet. The idea is to facilitate the disabled persons, so by visiting the website they can get to know the possible disease indication without standing in the queue and waiting for doctors.

So, we built a web application underlying the principles of Artificial Intelligence techniques that can predict if the person has this disease by taking symptoms from the user. Our model is first trained by giving a dataset of four thousand records and will detect the diseases by computations and then classify the disease using the dataset. Computer will ask the required symptoms from the user and the user will respond to it. We also give a description of the disease which was indicated by our application and some standard and verified disease precautionary measurements to the user.

II. Literature survey

A. Disease Diagnosis

In the market as well as most of the work had been done in the field of disease diagnosis as compared to disease indication. We can easily get to it by looking amount of work done in disease diagnosis. The reason for it is that Disease Diagnosis is considered more valuable, effective and practical. Because those diagnosis systems are used in hospitals or clinics so, the accuracy and effectiveness require a lot. By this, we can see a lot of work in that domain.

B. Disease Indication

Disease Indication is somewhat different from disease diagnosis. In disease indication, we do not have such (might be complex) inputs as we tackle in diagnosis. In diagnosis we have a variety of inputs of different kinds, might we need several results of many reports/tests (Blood test, X-ray, CGI, MRI and many other reports/test), Blood Pressure, Pulse rate, and several kinds of input we need in diagnosis. But in Disease Indication we do not or in some cases we system do not need these kinds of heavily inputs to indicate the disease the person is suffering.

By observing different disease indication, users are asked to input the symptoms they are experiencing. On these inputs, a supervised machine learning algorithm is used, which trained itself by using the standard symptoms of a specific disease. WebMD is a project which inhibits our idea.

III. Methodology

To implement this project, our group decided to use the K-Nearest Neighbor Supervised Machine learning algorithm.

A. DataSet

As Data is a key element in machine learning programs, for this project we had to use the already build a dataset of most popular 41 different diseases which indulge with a human [1].

Data is collected from a questionnaire by patients and the expert doctor in such a way that, doctors are asked to the symptoms of a specific disease and all the inputs of doctors are recorded.

Disease Name: Symptom # 1 Symptom # 2 … Symptom # N

Disease Name: Symptom # 1 Symptom # 2 … Symptom # N

Disease Name: Symptom # 1 Symptom # 2 … Symptom # N

B. PreProcessing

We heavily required to pre-process the data to get accurate results and feed that data to the machine learning model. For preprocessing we had done the following steps

- Place symptoms as column headers
- Put the disease name as last column (target column)
- Place 1 in those columns where that column symptom belongs to that disease and for all other columns place zero

'1' in columns: indicate that this symptom column belongs to that row. Like if we have a value of '1' at 'Dataset_Frame[34][7]' it means that the disease mention at last column of row ‘34’ has the symptom which is present at column no. ‘7’ of row no ‘34’.

C. Logic

The final shape of our dataset is such that we have symptoms on columns and each row represents the disease record. We have 133 columns, 132 are symptoms and column # 133 is a target disease column.

Symptom # 1	Symptom # 2	….	Symptom # N	Disease Name
1	0	….	0	Fever
1	0	….	1	Malaria
….	…..	….	…..	……
….	…..	….	…..	……
0	1	….	0	Allergy

Now, we can see that our problem lies in the Classification problems of Machine learning. In this case, all the symptoms are the feature attributes and last column (Disease Name) is the dependent/target/ground column of a dataset. As stated earlier we had selected the K-Nearest Neighbor (KNN) Supervised Machine learning algorithm. KNN is widely used in classification problems. We had written this algorithm from scratch. Its main working points are given below for each training data we do

- Find the distance between test and row of training data
- Sort them on the base of that distance in ascending order
- Select the ‘k’ shortest distance elements
- Find the dominating class of that test point
- Save the result in the output list

KNN algorithm works such that we select the dominating class from the ‘k’ nearest neighbors data points and finally we get the results of our input. For calculating the distance, we had used the ‘Euclidean distance’.

D. Working

To make it realistic and more useable, we had added some functionality in it. The first user is asked to enter the symptom and then we run the DFS algorithm to calculate the corresponding symptoms of user had entered first. For this, we had maintained the list where the related symptom of each symptom is stored and on it, we run DFS. After that, we got common symptom suggestions and the user is asked to select those symptoms which he/she is observing in his/her body.

After getting the input we create an auxiliary input array and set ‘1’ on those columns which the user has selected. And we pass that input array to our algorithm to compute the result and display the disease information to the user which model had predicted.

Accuracy in Machine learning algorithms matters a lot. In this case, accuracy is calculated in such a manner that how many column’s values of the predicted disease row is matched with the actual disease row columns. If predicted disease row has 90% columns matched with actual disease row’s columns then the accuracy of prediction is 90%.

IV. Experiments and discussion

This project backend and working are implemented in the Python programming language and for the web application, Django Python web framework is used. Communication between backend and frontend is done by Python.

A. Experiments:

After coding the algorithm, we had done many random tests of the model on both backend and Website inputs to the algorithm and fortunately, we got the results as expected. During the training of the model, the accuracy lies greater than the 97% when the value of ‘k’ is 5. Some of the test results are mentioned in the table 1.

Table 1 Experiments Results

Sr No.	**Sample inputs and results**		
	Input (Symptoms)	***Disease Name***	***Accuracy %***
1	Muscle Pain,Vomiting, Headache, Diarrhoea ,Nausea	Malaria	98.48
2	Blackheads, Skin Rash, Itching, Pus Filled Pimples	Acne	97.73
3	Breathlessness, Sweating, Chest Pain	Heart Attack	100
4	Vomiting, Red Spots Over Body, Skin Rash, Fatigue, Mild Fever, Loss of Appetite, Itching, Swelled Lymph Nodes, Lethargy	Chicken Pox	97.73

By giving the correct symptoms the algorithm has predicted the right disease as the dataset has provided the information. More user enters the precise and disease related symptoms the algorithm will return the more accurate disease with high accuracy.

V. Conclusion and Future work

Disease Indicator is web base application whose backend is supported by Python with the Supervised machine learning KNN algorithm. In this project, user gives the symptoms to the algorithm and then by using the power of Artificial Intelligence the computer predicts the disease by using the large amount of a past and experts observations. The final output is as accurate as user input is precise and relative. Over all our website useability is dependent how much we have accurate, efficient and precise dataset in backend, and how relatively and precisely user had given the input to the program.

A. Future Work

- As data is very important element in Machine learning, so by having much more accurate and efficient data in our database will increase the efficiency of program.
- Interface and mechanism of input/output to the computer/user will enhance the flow of working.
- More information inserted into the dataset about diseases.
- Algorithm for calculating the related symptoms on the bases of relative diseases to be improve.

References

[1] Pranay Patil, Disease Symptom Prediction, Version 1, Retrieved: 13 October 2020, from https://www.kaggle.com/itachi9604/disease-symptom-description-dataset