International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 394-397
Published Online May 2021 in IJEAST (http://www.ijeast.com)

SYMPTOM BASED DISEASE PREDICTION USING DECISION TREE

Akanksha, Chitranshi Varshney, Anamika Awasthi, Anjali Negi
Department of Computer Science And Engineering
IMS Engineering College, Ghaziabad, Uttar Pradesh, India

Abstract— **The world is moving with a fast speed and in order to be in pace with the world we invest our entire time working. This leaves us with no time to look after our body. We tend to ignore the symptoms we get as looking after ourselves is not in our to do list. Even if we do notice we barely have time to go to the hospital to seek doctors suggestion. So, we have built a disease prediction system using the symptoms provided as input to give the people some idea of what they are dealing with and take action accordingly owing to the seriousness of the disease.**

Keywords - Decision Tree, Python, Information Gain, Gini Index

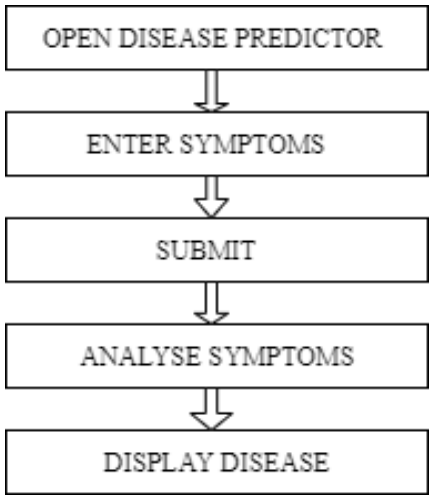
I. INTRODUCTION

We have made a virtual health hub which has four modules admin, doctor, patient and disease prediction. Any user could use this system to get an idea of the disease he might be having to consult the doctor. Firstly, we have cleaned the dataset by removing all the null values, checking if the dataset is balanced so that we do not get biased result. We also checked the correlation between the various symptoms and checked if it satisfies the null hypothesis. After cleaning and pre-processing the dataset we have applied decision tree and we have used gini-impurity as our criteria for splitting the nodes. Along with the disease prediction we have also added doctor, patient and doctor module to make it a full virtual health hub.
Our project predicts diseases according to the symptoms entered by the patient. We have used python as a platform to run our machine learning algorithms. The first step to the analysis is to decide the problem we want to solve. Then gathering the dataset to work on. We then, visualized our data with the help of scatter plots, heatmap, etc with the help of seaborn library, matplotlib library, etc to remove outliers, missing values etc from our dataset.

II. OBJECTIVES

1. Provides online medical services to everyone hardly matters whether the people live in metro or a remotely located village.
2. The system helps to automate all the activities.
3. Users can connect through their home internet to get services

III. METHODOLOGY



Flow Chart of Disease Prediction

3.1 Steps of model building:

STEP 1 Objective
To predict the disease suffered by the patient depending upon the symptoms.

STEP 2 Collecting data
Be it the raw data from excel, access, text files etc., gathering the past data forms the foundation of the future learning. Better the variety, density and volume of relevant data, better are the learning prospects for the machine.

STEP 3 Preparing the data
Any analytical process flourishes with the nature of the data utilized. One needs to invest time deciding the quality of information and afterward making strides for fixing issues like missing information and treatment of outliners. Exploratory data analysis is perhaps one method to study the nuances of the data.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 394-397
Published Online May 2021 in IJEAST (http://www.ijeast.com)

STEP 4 Training the model
This step includes picking the proper algorithm and portrayal of data as the model. The cleaned data is split into two parts – training data and testing data (proportion depending on the prerequisites); the first part (training data) is used for developing the model. This second part (test data), is utilized as a perspective.

STEP 5 Evaluating the model
In order to test the accuracy, the test data is used. This step decides the accuracy in the decision of the algorithm dependent on the result. A better test to check the accuracy of the model, is to see its performance on the data which was not used at all during model build.

STEP 6 Accuracy and cross validation
Checking the accuracy of the model in order to check whether we have to further improve our model.

STEP 7 Remove the overfitting
As the model was overfitting so in order to remove overfitting we used hyperparameter tuning.

3.2 DECISION TREE ALGORITHM

3.2.1 How does Decision Tree work?
A decision tree is drawn upside down with root at the top. At each step or node of a decision tree we try to form a condition based on the features to separate all the labels or classes contained in the dataset to the fullest purity. We can use either the entropy or gini-impurity as the criteria for the split. We then calculate the information gain and the split which gives the maximum information gain is chosen.

3.2.2 Execution
We have calculated the gini impurity of each symptom and each class of the symptoms and then we calculated the information gain from it. The feature with highest information gain is chosen as the criteria for split at each level. The splitting continues until we reach the leaf node or a desired threshold depth has been reached.

3.2.3 Recursive Part
In the recursive part, we repeat the above approach with increasing tree-level in order to construct the tree. We set the present node as a leaf node when there's no doubt to ask if the output is published for the symptoms given.

Algorithm:
TreeGrowing (S,A,y)
Where:
S - Training Set
A - Input Feature Set
y - Target Feature
Create a replacement tree T with one root node.
IF one among the Stopping Criteria is fulfilled THEN
Mark the basis node in T as a leaf with the foremost common value of y in S as a label.
ELSE
Find a discrete function f(A) of the input attributes values such splitting S consistent with f(A)'s outcomes (v1,...,vn) gains the simplest splitting metric.
IF best splitting metric > treshold THEN
Label t with f(A)
FOR each outcome vi of f(A):
Set Subtreei= TreeGrowing (σf(A)=viS,A,y).
Connect the basis node of tT to Subtreei with a foothold that's labelled as vi
END FOR
ELSE
Mark the basis node in T as a leaf with the foremost common value of y in S as a label. END IF
END IF
RETURN T
TreePruning (S,T,y)
Where:
S - Training Set
y - Target Feature
T - The tree to be pruned
DO
Select a node t in T such pruning it maximally improve some evaluation criteria
IF t6=Ø THEN
T=pruned(T,t)
UNTIL t=Ø
RETURN T

3.2.4 Prediction
After fitting the model we checked the accuracy and it turned out that the training accuracy was 100% and the decision boundary was weird looking. The model was overfitting.

3.2.5 Handling overfitting
There are various methods to handle overfitting like pruning, adding more dataset, bagging, Ensembling, Hyper-parameter tuning, etc. We removed it by hyperparameter tuning.The various hyperparameters in decison tree are min_samples_split,min_weight_fraction_leaf,max_features,random_state,min_impurity_decrease,class_weight,max_depth. We have used min_samples_split and max_depth then the accuracy of the model came around to be 94.21%.

IV. HARDWARE REQUIREMENT

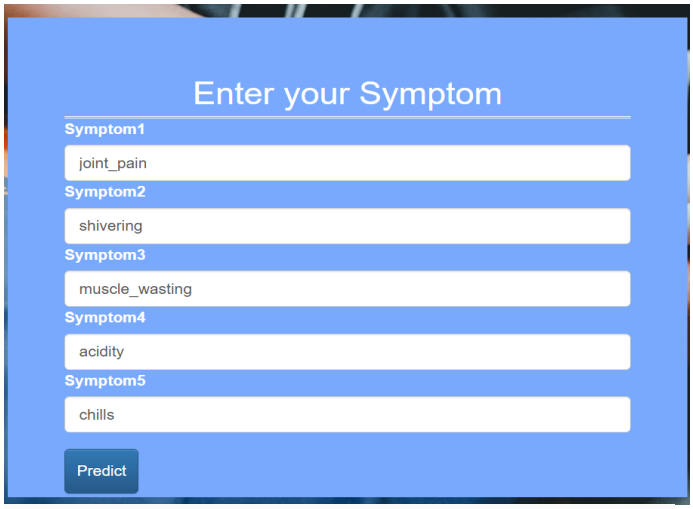
- Desktop with Internet Connectivity
- i3 Processor-Based Computer or higher
- Memory: 4GB RAM or more
- Hard Drive: 50 GB

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 394-397
Published Online May 2021 in IJEAST (http://www.ijeast.com)

V. SOFTWARE REQUIREMENT

- Visual studio
- Any Browser
- Windows 7 or higher

VI. RESULTS



Disease Predicted: Allergy
Accuracy: 94.21

VII. LIBRARIES

- SKLEARN
- NUMPY
- PANDAS
- SEABORN
- MATPLOTLIB

VIII. ADVANTAGES

i. Data security and retrieving availability.
ii. Improve patient care.
iii. Make diagnosis and treatment better.
iv. Paperless operation and easy-to-understand.
v. Track Financials better

IX. DISADVANTAGES

i. Over dependency on technology

ii. As everything is online there would be lack of employment

iii. Technical bandwidth

X. APPLICATION

i. This application is used by all patients who need help in emergency.

ii. This system ensures that people get to know their diseases in the comfort of their homes.

iii. It predicts disease accurately based on symptoms.

XI. FEATURES

1. Improve patients' experience during any interactions with the hospital
2. Prediction is easy
3. A cordial and helpful UI/UX configuration
4. Consistence with all information security guidelines and conventions

XII. CONCLUSION

A definitive objective is to work with a facilitated and very much educated medical system for guaranteeing the most extreme patient fulfilment. In developing nations, predictive analytics is the next big idea in medicine –the next evolution in statistics – and roles will change as a result. This project has a lot of value in everybody's everyday life and it is fundamentally more significant for the medical services area since they are the ones that day by day utilizes these frameworks to anticipate the sicknesses of the patient's dependent on their overall data and their symptoms that they are experienced.
Now a day's health industry plays a major role in curing the diseases of the patients so this is also some kind of help for the health industry to tell the user and also it is useful for the user in case he/she doesn't want to go to the hospital or any other clinics, so just by entering the symptoms and all other useful information the user can get to know the disease he/she is suffering from.
If the health industry adopts this project then the work of the doctors can be reduced and they can easily predict the disease of the patient. The Disease prediction is to provide a prediction for the various and generally occurring diseases that when unchecked and sometimes ignored can turn into fatal diseases and cause a lot of problems to the patient and as well as their family members.

XIII. ACKNOWLEDGEMENT

It offers us a brilliant feel of delight to provide the research paper of the B.Tech Project undertaken for the duration of B.Tech Final Year. We would like to extend our sincere gratitude to the Head of the Department, Computer Science and

Engineering, **Dr. Avdhesh Gupta**, for his commendable support and encouragement for the completion of our project. We convey our deep gratitude and we are very much indebted to our versatile project guide **Mr. Mukesh Kumar Singh**, Assistant Professor for his valuable suggestions and spontaneous guidance to complete our project. We would also like to thank our faculty members for their kind assistance and cooperation in the course of the improvement and completion of our project. Last but not the least, we would also like to acknowledge our parents and friends for their constant support throughout the project

XIV. REFERENCES

- https://www.researchgate.net/publication/32195647 Plagiarism
- https://developers.google.com/speed
- https://www.coursehero.com/file/73531061/Disease Prediction-System-2pdf/