AN APPROACH FOR DEVELOPING DISEASE PREDICTION SYSTEM

Aditya Khandelwal[1], Ajay Kandi[2], Aditya Vyas[3], Hrithvik Ranka[4], Archana Khandekar[5]
[1,2,3,4] Students, CSE Department, MIT-WPU, Pune, India
[5] Assistant Professor, CSE Department, MIT-WPU, Pune, India

Abstract : Disease Prediction system is basically based on predictive modeling which predicts the disease of the user on the idea of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and provides the predicted disease as a result. Disease Prediction is completed by implementing the Naïve Bayes Classifier. Naive Bayes Classifier calculates the probability of the disease. Our model will save time as well as makes it easy to get a warning about the individual health before it's too late.

Keywords **- Machine Learning, Naïve Bayes Classifier, Python, Supervised Learning, Hypothesis.**

I. INTRODUCTION

The main aim of our system is to predict the diseases on the basis of symptoms entered by the patient. The model which we have built comes under the category of data science. For our model we are using python language to run our machine learning algorithm. The basic procedure for any analysis starts with deciding the problem we want to solve and in the next step we have to search for the dataset to work on. Once the dataset is finalized then we can perform the data pre-processing techniques. For the prediction of diseases our model will use only naive bayes algorithm which is a supervised learning algorithm.

II. LITERATURE SURVEY

A Lot of labor has been done earlier to predict diseases with the assistance of symptoms using Machine Learning. Different levels of accuracy were attained using various data processing techniques which are explained as follows.

Avinash Golande, studied various different ML algorithms which were used for classification of heart diseases. Research was administered to review Naïve Bayes, KNN and K-Means algorithms which were used for classification and their accuracy were compared. This research concludes that accuracy obtained by Naïve Bayes was highest further it had been inferred that it are often made efficient by the combination of various techniques and parameter tuning.

T.Nagamani, has proposed a system which was deployed using data processing techniques together with the MapReduce algorithm. As per this paper, the accuracy obtained for the 45 instances of testing set was greater than the accuracy obtained using conventional fuzzy artificial neural network. The accuracy of algorithm used here was improved because of the use of dynamic schema and linear scaling.

The paper "An approach to devise an Interactive software solution for smart health prediction using data mining" aims at developing a fully automated computerized system which can check and maintain your health by knowing the symptoms. It was having a symptom checker module which can actually define our body structure and gives us option to select the affected area and checkout the symptoms. The technologies used in this paper were: The front end was designed with help of HTML, JavaScript and CSS, the back end was built using MySQL which was used to design the database. This paper also contains the information about testing like Alpha testing which is usually done at server side. This is an actual testing which is done with potential users or as an independent testing process at the server end. And Beta testing is done after performing alpha testing, the versions of a system or software are known as beta versions and are given to a specific audience outside the programming team.

The basic reason behind choosing this system after reviewing the above papers was to create a disease prediction system that will be based on the inputs they have used. We analysed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and Naive Bayes based on their Accuracy and f-measure scores and then we identified the best classification algorithm which can be used for our project i.e disease prediction system.

III. PROPOSED MODEL

The proposed work predict diseases by exploring Naive Byes algorithm and does performance analysis. The main aim of this study is to effectively predict the disease from which the patient is suffering. The user enters the symptoms which he/she is feeling. The data is then fed into model which predicts the disease.

Fig-1 :- Generic Model for Disease Prediction.

IV. DATASET DESCRIPTION

The dataset which we have used is comprised of categorical data and contains 4930 rows and 133 columns. Out of 133 columns, 132 columns consists of symptoms and their corresponding rows contains the 0/1 values with respect to the disease they belong. The last column which is named as "prognosis" is consist of the diseases.

Fig-2 :- Dataset used

V. METHODOLOGY

5.1 Collecting Data

Be it the raw data from excelsheet, access, text files etc, this step (gathering past data) generally forms the foundation for the future learning. The better the volume, density and variety of relevant data, better it becomes for machine to learn the various prospects.

5.2 Data Preprocessing

Data preprocessing describes any sort of processing performed on data to organize it for an additional processing procedure. Commonly used as a preliminary data processing practice, data preprocessing transforms the info into a format which will be more easily and effectively processed for the aim of the user.

5.3 Model Training

This step involves choosing the appropriate algorithm for the model. As per the literature survey we have found that the naive bayes classifier has the best accuracy thats why we have choosen Naive bayes algorithm for our project. Then the preprocessed data is split into two parts i.e training data and testing data. Here we have splitted our dataset into 70:30 ratio which means 70 percent data for training purpose and 30 percent data for testing purpose. The first part (training data) will be used for developing the model and the second part (testing data) will be used as a reference.

5.3.1 Naïve Bayes Classifier

Naive Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which the value of the attribute is independent of the values of other attributes.

The Bayes theorem is as follows:

Let X = {x1, x2, x3, …… , xn} be a set of n attributes.
In Bayesian,
X is considered as evidence

H be some hypothesis means
And the data of X belongs to specific class C.

We have to determine P (H|X) which means the probability that the hypothesis H holds given evidence i.e. data sample X. According to the Bayes theorem the P(H|X) is expressed as :

$$P(H|X) = P(X|H) \times P(H) / P(X)$$

5.4 Model Evaluation

For testing the accuracy of the model the testing data is used. This step is used to determine the accuracy in the choice of the algorithm based on the result. A better way to see accuracy of model is to check its performance on data which was not used at all during model build. After performing the testing on testing data we recieved the accuracy of approximately 93%.

VI. RESULTS

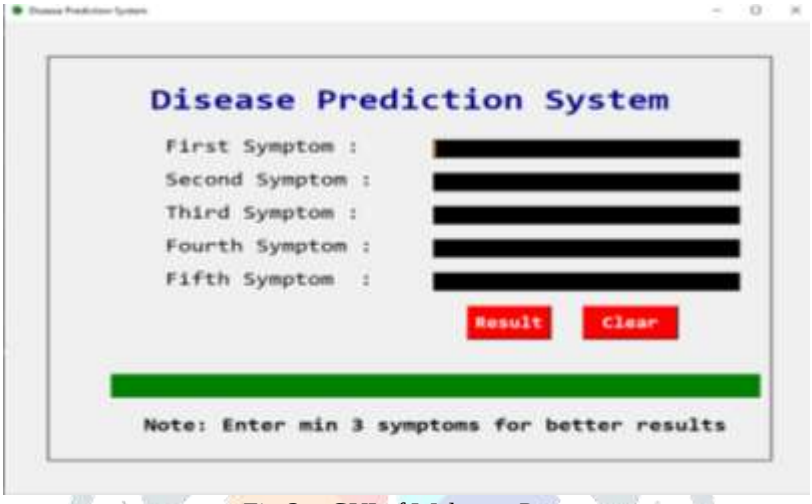

Fig-3 :- GUI of Welcome Page

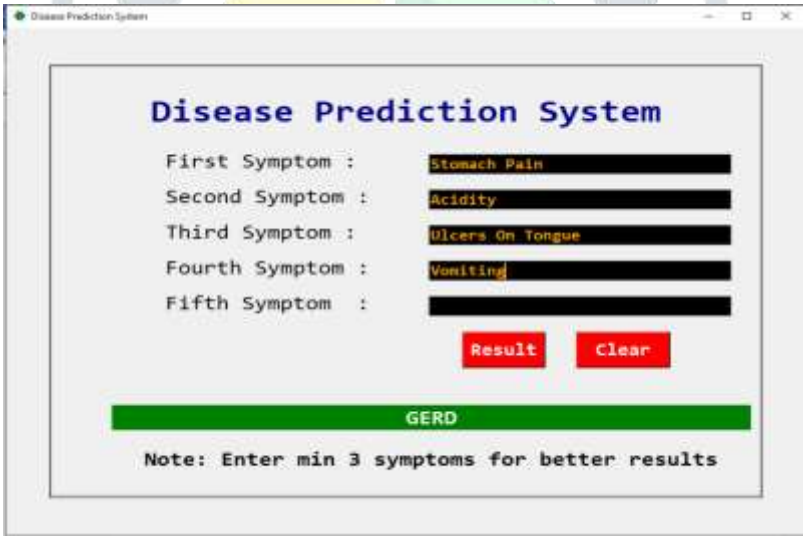

Fig-4 :- GUI of Result Page

VII. CONCLUSION

With the increasing number of deaths, all thanks to different and various rising diseases, it's become mandatory to develop a system to predict diseases effectively and accurately. The motivation for the project was to seek out the foremost efficient ML algorithm for detection of diseases. The results of the literature survey states that the naive bayes algorithm was the most effective algorithm. Hence after using it in our model we get the accuracy score of approximately 93% for prediction of disease. In future this model can be improved by developing a web application supported by naive bayes algorithm also by using a larger dataset as compared to the one which we've used in this analysis which will help to provide better results and help health professionals in predicting the various diseases effectively and efficiently.

REFERENCES

[1] Avinash Golade, Pavan Kumar , "Heart Diseaese Prediction Using Effecetive Machine Learning Techhniques", Internatioinal Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamiani, S.Logeswar, B.Gomathy," Heart Diseaese Prediction using Data Miining with Mapreduce Algoritm", Internatioinal Journal of Innovaitive Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] M.A. Nishra Banu, B Gomathy, "A approach to deivise an Interactive software soluition for smart health predictiion using data mining, in Internatioinal Journal of Techhnical Research and Aplications , eISSN, Nov-Dec 2013.

[4] Chaitrali S. Dangore , Sulabha S. Apte , Improved Study of Heart Diseases Prediction System using Data Mining Clasification Techniques (2012), Internatinal Journal of Computer Application (0975 – 888) Volume 47– No.10.

[5] Chaurasia V, Pal S (2013) Early prediction of heart diseases using data mining techniqes. Carib J Sci Technol 1:208–217

[6] Hasan Koyuncu , Rahime Ceylan Artifical neural network based on rotation forest for biomedical pattern clasification , 2013 IEEE 36th International Conference on Telecommunications and Signal Processing (TSP).