International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952297

Generic Disease Prediction using Symptoms with Supervised Machine Learning

Ashish Kailash Pal*, Pritam Rawal, Rahil Ruwala, Prof. Vaibhavi Patel
Department of CSE, Parul University, Vadodara, Gujarat, India

ABSTRACT

Data Mining and Machine Learning plays most inspiring area of research that become most popular in health organization. It also plays a vital part to uncover new patterns in medicinal science and services association which thusly accommodating for all the parties associated with this field. This project intend to form a diagnostic model of the common diseases based on the symptoms by using data mining technique such as classification in health domain. In this project, we are going to use algorithms like Random forest, Naive Bayes which can be utilized for health care diagnosis. Performances of the classifiers are compared to each other to find out highest accuracy. This also helps us to find out persons who are affected by the infection. The test based on the outcomes of the diseases.

Keywords : Naive Bayes, Decision Tree, Random Forest

I. INTRODUCTION

Nowadays we humans are so busy in our work and professional life that we neglect the small illness that we are suffering from, but sometimes the small diseases lead us on big problems. With this project, we will apply classification algorithms for the given data to determine which algorithm is best and suitable to classify. So the main idea regarding this project is providing a personal health audit which will provide an up to the mark assistance regarding your current health status on the basis of a small survey of personal details and current health status. The user has to go through a simple survey of his/her current health status and through different techniques like Machine learning, Naïve Baye's, decision tree etc. Our system will predict the disease and will provide its remedies.

II. ALGORITHM USED

Naive Bayes: In data mining, naive Bayes classifiers are considered family of simple probabilistic classifiers based on applying Baye's hypothesis with solid autonomy presumptions between the components. It was introduced with a different name into the text rescue community in the early 1960's, and remains a standard method for text categorization, the problem of judging documents belonging to only one category or the other one with word frequencies of the elements. With suitable pre-processing, it became competitive in this domain with highly advanced methods including bolster vector machines. It likewise discovers application in programmed medicinal analysis. Another type which is credulous Baye's classifiers are highly scalable, requiring a number of parameters direct in the quantity of factors in a learning issue. Greatest probability preparing should be possible by assessing a shut frame expression, which takes straight time,

instead of by costly iterative guess as utilized for some different sorts of classifiers. In the measurements and software engineering writing, Naive Bayes models are known under an assortment of names, including straightforward Bayes and autonomy Bayes. Every one of these names reference the utilization of Bayes' hypothesis in the classifier's choice administer, yet innocent Bayes is not a Bayesian strategy[1].

Decision Tree: A decision tree is a flowchart structure in which each internal node denotes a test on a characteristic, where each branch signifies the result of the test, and each leaf node denotes a class label. The paths from the root to leaf denotes classification rules. In tree the interrelated diagram are used as the analytical, visual and decision support tool, where the apparent values are calculated. A decision tree contains three types of nodes: one is decision nodes – denoted by squares, Chance nodes denoted by circles, and the other is End nodes –denoted by triangles you begin a decision tree with a conclusion that need to make. Draw a small square to represent the tree towards the left of a great piece of paper. From this draw out lines in the direction of the right for each likely solution, and write the solution along the line. At the end of each line, reflect the results. If the result of decision is undefined, then draw a small circle. If the result is alternate decision then you need to make, draw a different square. Squares represent decisions, circles that denotes uncertain result. Starting from the new decision squares from the diagram, draw the lines denoting options that could be select. From the circles we can even draw lines that represent possible outcomes. Finally make a short-term note on the line by saying what it says[2].

Random Forest: Random forest is a bootstrapping algorithm with the cart model. It built multiple trees with different initial variables and consider a sample of 100 observation and 5 random samples chosen initial variable to build a cart model. It will repeat the same process 10 times and they make a final prediction for each observation. Final prediction is a function of each prediction. This final one can simply be a mean of each prediction. Basically this process is done in the Weka tool. Weka tool is the machine learning tool which contains a large number of data science algorithms which can be used for classification, prediction and to find the missing values[3].

It is a collective learning method for organization and other tasks that operate by building a gathering of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the single trees. The process for the random forests was shaped by Tin Kam Ho used the random subspace method, in Ho's formulation, it is a process to implement the "stochastic discrimination" method to classification planned by Eugene Kleinberg. The above process defines the original bagging algorithm for trees. Random forests change in one from this general arrangement: they had used a modified tree learning algorithm that selects, at each person in learning process, a random subset of the features. This process is sometimes called "feature bagging". The importance of this correlation is the multiple trees in a normal bootstrap sample: if one or more additional features are correct predictors for the response variable, these features will be sensibly chosen in many of the B trees, making them to be correlated[4].

A. Advantage of RF algorithm.

1) RF algorithm is most accurate ensemble learning algorithm.
2) RF runs efficiently for large data sets.
3) It can handle hundreds of input variables.
4) RF estimates the important variables in classification.
5) In training data, RF is less sensitive to outlier.
6) Parameters can be set easily in RF and eliminates the need for pruning.

7) Generated forests in this method can be saved for future reference[5].
8) In RF, accuracy and variables importance is automatically generated[6].

B. Module Description

Step 1: First did research on the problems of this project, then what kind of solutions we can approach and its outcomes regarding our project.

Step 2: Then in the second part which is data acquisition, we collected our data from UCI Repository.

Step 3: As we know collecting data invites many inconsistent and unwanted data, so it is mandatory to go for the i) data cleaning and ii) Data Transformation

Step 4: With the help of EDA, we defines and refines the selection of feature variables that will be used in the model development.

Step 5: We repeatedly apply the machine learning techniques to the data that best fits the requirement.

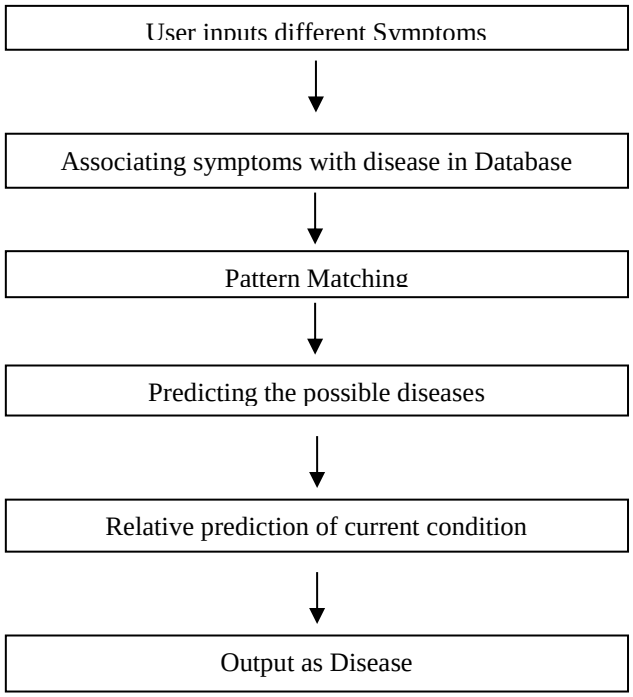

Fig 1 : Flowchart of Disease Predicted

III.LITERATURE RIVIEW

This section reviews literature used in this paper.

Heart Disease: Heart related diseases also known as Coronary Heart Diseases (CHD),which is problem of deposition of fats inside the blood tubes passing the blood to the Heart muscles. Heart diseases could occur as early as 18 years and they could be detected when the blockage exceeds about 70%.If these blockages remain undetected or not treated than could cause rupturing of the membrane covering the blockage because of the excess pressure of blood flow. The mixturing of the chemicals released from the membrane with the blood could lead to blood clot, and would excessively leads to various Heart diseases[7].

The reasons which increase blockage are called as risk factors. These risk factors are classified as modifiable and non-modifiable risk factors. Non modifiable risk factors are age, gender, and heredity. These risk factors can't be modified and they will always keep causing heart disease. Risk factors which can be changed by our efforts are called as modifiable risk factors. Some modifiable risk factors are 1) Food related 2) Habit related 3) Stress related 4) Bio chemical and miscellaneous risk factors. Atherosclerosis, coronary, congenital, rheumatic, myocarditis, arrhymia and angina are the different types of heart diseases[8]. Common symptoms of heart disease are listed in table 1[9].

Table 1 : Symptoms of heart disease[9].

Sl.no	Symptoms name
1	Chest pain
2	Strong compressing or flaming in the chest
3	Discomfort in chest area
4	Sweating
5	Light headedness
6	Dizziness

7	Shortness of breath
8	Pain spanning from the chest to arm and neck
9	Cough
10	Fluid retention

Table 2 : Risk factors of heart disease[10].

Sl.no	Risk factor
1	Diabetes
2	High blood pressure
3	High LDL
4	Low HDL
5	Not getting enough Physical activity
6	Obesity
7	Smoking

Effective decision support system should be developed to help in tackling the menace of heart disease.

IV. WORKING METHODOLOGY

We come through a wide range of different and major algorithms for predicting the monotonous diseases with comprehensible symptoms while working in the field of Supervised Machine Learning and its different models. A wide range of algorithms model are adopted through Sci-Kit Learn Library has provided number model of algorithm such as Decision tree, Naive Bayes and Random Forest in which we divided the trained data into train X and train Y as an output, following the comparison with the test X and test Y respectively. Naïve Bayes has three different models i.e. Gaussian, Multinomial, Bernoulli Naïve Bayes. We worked on Gaussian Naive Bayes because it is quite easy to understand and quite simpler comparatively .The application of data fitting is mostly same in all the three models. Each model has its own accuracy to predict the result of diseases. As it is commonly known that Random Forest results are more accurate than that of Decision tree. As shown in Result, in our project named "Idiosyncratic Health Audit", as an example we took at most five symptoms and minimum one to perform the prediction of Health diseases.

V. EXPERIMENTAL RESULTS

In the survey of our project, Diseases predicted by the algorithm and it's data by using the different data mining technique. This analyze the medical data in multiple ways, like that, multidimensional ways and view based collects that data and it escapes the hard risks then, prediction is easily completed. The data is classified in to two types namely, (i) Structured data, (ii) Unstructured data. The concept fulfill the existing system focused both types of data prediction in medical area, that is big data analytics. There are numerous researches from various domains are continuously working towards developing Achieving Disease Prediction. The aim of this survey was to Summarize the recent researches and its demerits towards achieve Disease Prediction. This paper gives the personal and easy way of prediction of the diseases and its symptoms. This paper concludes that there is no effective method discovers for Achieving Disease Prediction. So, further approaches should overcome all the above issues. Further implementation has to be done in order to Achieving High Disease Prediction using machine learning algorithm.

Table 3: Accuracy comparison for Disease Data set.

Sl.no	Approach	Accuracy
1	Naïve Bayes	78.56
2	Decision Tree	82.43
3	Random Forest	85.78
4	**Our Approach**	

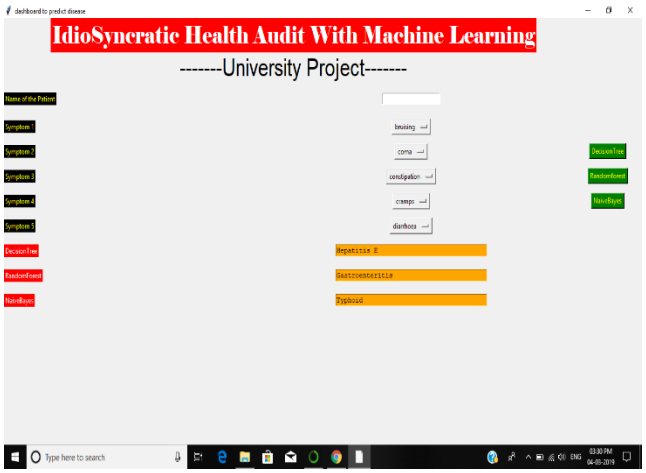

Fig 2 : Result of disease predicted

VI. CONCLUSION

In the proposed methodology, Disease prediction is based on Machine Learning. It intended to give better output result. A personal health audit which will provide an up to the mark assistance regarding your current health status on the basis of a small survey of personal details. Our system takes symptoms as input and gives output as disease, possible causes. Because of our system patients get their medicines. Patients won't have to wait for Doctors appointment, due to our system patients save their money and time. Our proposed approach (Random forest, Decision tree, Naïve Bayes) achieved an accuracy of 82.26% for our data set.

VII. REFERENCES

[1]. Onisko A, Druzdzel M.J and Wasyluk H, A Bayesian Network Model for Diagnosis of Liver Disorders. In Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering, 2, 1999, 842-846.

[2]. Lin R.H, an Intelligent Model for Liver Disease Diagnosis. Artificial Intelligence in Medicine, 47 (1), 2009, 53-62.

[3]. Rajeswari P and Reena G, Analysis of Liver Disorder using Data mining Algorithm. Global Journal of Computer Science and Technology, 10 (14), 2010, 48-52

[4]. Ramana B.V, Babu M.S.P and Venkateswarlu N.B, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. Global Journal of Database Management Systems, 3 (2), 2011, 101-114.

[5]. home.etf.rs/~vm/os/dmsw/Random%20Forest.pptx, last accessed 10/8/2015.

[6]. Jehad Ali et.al, "Random forest and decision trees", IJCSI,Vol 9,No 3,pp272-278(2012).

[7]. Saaol times, Monthly magazine" Modifiable risk factors of heart disease", pp 6-10, July (2015).

[8]. Khan MG, "Heart disease diagnosis and therapy", a practical approach, 2nd Edition Springer, pp544(2015).

[9]. Khan MG, "Heart disease diagnosis and therapy ", a practical approach,2nd Edition Springer,pp544(2015).

[10]. M.A. Jabbar, B L Deekshatulu, Priti chandra, "classification of heart disease using artificial neural network and feature subset selection", GJCST,Vol13, issue 3,2013.

Cite this article as :

Ashish Kailash Pal, Pritam Rawal, Rahil Ruwala, Prof. Vaibhavi Patel, "Generic Disease Prediction using Symptoms with Supervised Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1082-1086, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952297
Journal URL : http://ijsrcseit.com/CSEIT1952297