International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

A Computational System for Disease Diagnosis and Prescription Generation

Nishant Bakhrey, Riya Bakhtiani, Jainam Soni, Dhananjay Kalbande

Abstract—Illnesses should be dealt with proper care and understanding and on time. If failed to be treated on time, they pave the way for numerous medical issues and these issues can lead to fatal deterioration of health, and in worst case death. These issues are winding up more awful in the villages because of the shortage of authorities, specialists and healthcare centres. To mention as a fact, the practitioners in Primary Health Care Centres (PHC) who provide 80% of outpatient care, don't have proper qualifications for it. Considering the scenario of rural India, more than 8% of 25,300 primary health centres in the country were operating without a proper doctor, 38% had absence of a laboratory technician, and 22% had no pharmacist. The reason for the shortages of educated doctors is that trained and city-bred doctors are unwilling to serve in rural areas. In this paper, with a goal to address such issues, we have made endeavors to plan and create a master framework for practitioners and patients to enter the symptoms and find out the relevant disease. For most users, however, just the recognition of the disease is not helpful. This paper introduces a novel concept to augment the result of the recognized disease with affinity analysis to perform market basket analysis to help users identify their symptoms more easily and ameliorate the set of symptoms-input for classification. Later, it takes into consideration any particular allergy of the user while generating prescription. This system possess enormous advantage in diagnosing diseases especially in rural areas and provide adequate and appropriate results and also makes reliable predictions to users. For achieving this, we use Decision Tree Classifier on the symptom disease dataset and a content based recommendation system for generating the appropriate prescription.

Index Terms—Disease Prediction, Prescription, Decision Tree Classifier, Market Basket Analysis

I. INTRODUCTION

Prescription Generator is a tool for doctors and practitioners for generating prescription for common diseases in villages where the availability of doctors is less as compared to urban areas. According to [1], people in rural areas prefer public healthcare facilities instead of private due to monetary issues and transportation cost to urban centres is expensive. Additionally, to reach the local centres and PHC's villagers have to travel long distances which further augments the transportation cost.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Nishant Bakhrey*, B.E, Department of Computer Engineering, Sardar Patel Institute of Technology
Riya Bakhtiani, B.E, Department of Computer Engineering, Sardar Patel Institute of Technology
Jainam Soni, B.E, Department of Computer Engineering, Sardar Patel Institute of Technology
Dr.Dhananjay Kalbande, Professor, Department of Computer Engineering, Sardar Patel Institute of Technology

Regardless of that, just 11% sub-centres, 13% Primary Health Centers (PHCs) and 16% Community Health Centers (CHCs) in India meet the Indian Public Health Standards (IPHS). Just one specialist is accessible for each 10,000 individuals and one state run emergency clinic is accessible for 90,000 individuals.

As mentioned in [2], Poor living condition, dangerous drinking water, absence of sanitation, utilization of biomass fuels increment the danger of various medical issues. Desai et al. [3] noticed a high frequency of minor sicknesses like cough, fever, loose bowels. (124 for every 1,000 people) among rural Indian populace. The minor ailments regardless of their short duration cause significant time loss from normal day-to-day activities. As mentioned in [4], one of the emerging technologies that can be utilized to face the specific challenges is to provide real-time patient data and aid with symptom based diagnosis which can help save doctor's time and permit them to counsel more patients. Machine Learning enables building models to rapidly break down information and convey results, utilizing both verifiable and continuous information [5]. With machine learning, doctors can settle on better choices on patient's treatment alternatives, which prompts the general improvement of medicinal services. Previously, it was laborious for experts to gather and examine the enormous volume of information for successful forecasts and medications since there were no innovations or instruments accessible. [5] The difference between traditional approach and the machine learning approach for disease prediction is the number of dependent variables to consider. In a traditional approach, very few factors are considered attributed to Body Mass Index (BMI), such as age, weight, height, gender, and more (because of computational restriction). On the other hand, machine learning methodologies allow considering extensive number of variables, which results in a better accuracy. Better results are obtained if the number of variables considered are more than 100.

Sometimes patients cannot identify their symptoms effectively. The practitioners find trouble in making a decision based on incomplete information about the symptoms provided by the patients as they do not possess skill in all the fields. In order to alleviate this problem, we proposed to use market basket analysis to detect the co-occurrence relationship among various symptoms. After the user inputs his symptoms, the system performs market basket analysis to find out if there are chances of any other symptoms which are being experienced by the user. This helps the user identify the symptoms more correctly and thus the model is able to predict the correct disease. While prediction,

we have calculated the probabilistic value of every class so that if the probabilistic value of two or more disease lie very close, we would predict multiple diseases and accordingly prescription will be generated. [6] posits that it is essential to perceive the symptoms of a medication drug, since they can be dangerous, however couple of unfriendly responses to drugs are really allergic in nature. Thus, we take into account if the user is allergic to a particular medicine or drug manufacturing company, and accordingly dosage and prescription is generated.

II. RELATED WORK

A. Medical Expert System

In [7], it is stated that an extensive number of expert systems are medicinal. The principle point of any medicinal expert tool is diagnosis and treatment of diseases. A medicinal expert framework is developed from programs and clinical knowledge base. The data acquired through the tool is like the data given by specialist or any doctor in that specific zone. Their medicinal system has thirty-two lung diseases in its learning base. The client or patient is requested to reply with YES or NO, If a specific symptoms shows up or not. At last, in light of users answers, the name of the infection is shown on the screen. A confinement of this expert system is that just symptoms entered by the software engineer in the knowledge base are accessible. It does not think and learn independent on itself. In this manner the knowledge base has to be updated whenever with new side effects and new illnesses. Medicinal Knowledge of particular specialist is imperative to the development of medical expert tool. This learning is done in 2 stages. In the principal stage, the medicinal states of infections are recorded amid the development of individual gathering with specialists and patients. In the second stage, a store of standards is shaped where each standard contains IF part that contains the side effects and THEN part that contains the infection that ought to be figured out.

B. A Neuro-fuzzy Mobile Application for Diagnosis and Treatment of Cardiovascular Diseases

In [8], they have considered the improvement of Mobile Neuro-Fuzzy System that uses the blend of the knowledge strategy of Artificial Neural Networks (ANN) and the humanlike thinking style of Fuzzy Logic to analyse and propose conceivable medications for cardiovascular illnesses through intelligence with client. Taking the upside of mobile innovation in term of telephones that are inexpensively accessible and with the far reaching availability to give real time healthcare service. The versatile framework will have an incredible potential in engaging patients and reshaping the desires for healthcare service delivery expectations consequently making it people-centered. In the proposed system the user interacts with the system and supply their basic information such as ailments and present symptoms, known as Personal Healthcare record. Then the symptom checker system will match the input symptoms to the respective rules, until one of the heart disease is confirmed. These results are then sent to Learning Neural Network Phase for further diagnosis. Since the module has effectively obtained enough learning of patient from the Health record database, it brilliantly associates the rest of the parts, checking to recognize the exact reason for heart illness and after that, delivers a last outcome which is gotten to by the client through the user interface.

III. DATA ACQUISITION AND PREPROCESSING

For the symptom-disease stage, Manual-Data available at [9] repository is used to predict the likelihood of a certain disease. The dataset has 133 symptoms and 41 unique diseases. The dataset consists of 4920 rows of medical records. The symptoms include itching, acidity, skin _rash, shivering, stomach pain to mention a few. Typhoid, Dengue, Migraine, Hepatitis, Tuberculosis are few of the prognosis included in the dataset. Decision Tree Classifier is trained on this dataset and the trained model is stored for real-time input. We cannot split our training dataset to find out the accuracy as the classifier cannot work on unseen data because it has never seen that disease before. Hence, we train the model on the entire dataset, and perform testing on Testing dataset provided by [9]. Testing dataset contains 42 records covering all the 41 diseases. For the market basket analysis, we took the column name of the corresponding symptom if it's value is equal to one and then prepared a new file containing the valid symptoms for the 4920 records. Tokenization is performed on each sentence wherein duplicate tokens are removed and tokens are sorted. For the prescription part, data is collected from a group of doctors and divided according to three age groups. There is a possibility that each age group has different drug, dosage and drug-usage instructions. The prescription part has been implemented for the following 10 diseases: Bronchial Asthma, Chicken pox, Heart attack (myocardial infarction), Impetigo,
Tuberculosis, Urinary Tract Infection, Peptic Ulcer , Dengue, Hypertension, Hyperthyroidism. Apart from the 10 disease, prescription is also collected for some of the basic symptoms like headache, vomiting, redness of eyes. The use of prescription for these common symptoms will be explained in subsequent sections.

IV. METHODOLOGIES IMPLEMENTED

A. Decision Trees

From [10], we know that Decision Tree involves representing all the possible solutions to a decision graphically. It starts with a root node which denotes the entire training dataset which is then split further into subsets. Each node is branched into subsets based on some conditional property of a feature which become the inputs to the child nodes. When further division is not possible the node is called as the leaf node. Any unwanted sub-tree can be eliminated in the process called as pruning.
According to [11], Gini coefficient score apprises us about how well the classifier has separated the classes. [11] proves that the knowledge discovered and the accuracy are better when the split is made using Gini coefficient rather than guess work. The redness of eyes is the top symptom that has the highest Gini impurity score of 0.9755 Fig. 1. Then comes internal itchiness with a score of 0.9749 and so on.

Basically this implies that the redness of eyes symptom has the potential to divide most samples into particular classes and hence it is selected as

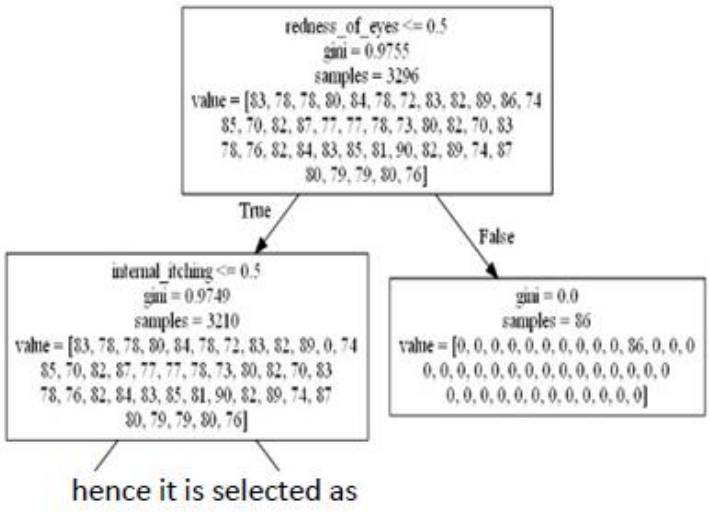

Fig. 1. A Part of Decision Tree

the root of the decision tree. From there we move down with decreasing order of Gini scores.

B. Market Basket Analysis

i. [12] posits that because of readability of the associative classifiers, they are especially fit in applications where the model can help domain experts in their diagnosis. Medical field being one good application, where predictive analysis helps identifying symptoms more extensively. [12] mentions three-step process for associative classification :

i. Produce the set of association rules from the training set with certain support and confidence thresholds as candidate rules.

ii. Pruning the set of discovered rules to remove those rules that can introduce over fitting

i. Classification Phase is to make a prediction for test data and measure the accuracy of the classifier. It is important to select interesting rules from the large set of possible rules. Confidence, Lift and Conviction are the three thresholds used that helped us identify best set of rules. Let X, Y refer to symptoms of our dataset, X => Y is an association rule and T refers to number of records in our database.

I. Support is a measure of how frequently the symptom will occur in the dataset. [Fig. 2]

ii. Confidence is an indication of how regularly that particular rule has been found to be true. [Fig. 5]

iii. Lift of the rule can be seen in Fig.3 .

iv. Conviction of a rule can be seen in Fig. 4 .

After tokenization, frequent itemsets are calculated using Apriori algorithm. The SUPPORT _THRESHOLD is set to 100. Time spent finding frequent itemsets is 0.30 seconds. To generate the rules, CONFIDENCE THRESHOLD=0.8, LIFT _ THRESHOLD=20.0 and CONVICTION THRESHOLD=5.0. Number of rules generated are 9791 which are used in displaying the possible symptoms on the user screen of android application. Time spent finding association rules is 0.06 second.

$$\mathbf{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Fig. 2

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

Fig.5.

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

Fig. 3.

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

Fig. 4

V. IMPLEMENTATION

We have divided the work flow in 2 stages:

A. Disease Prediction

The conceptual structure implemented is shown in Fig. 6 The trained model using manual dataset is hosted on the server. User enters their basic details like age, weight, height, blood group, and other relevant details. Once the user is logged in, he enters the symptoms experienced by him. Those 133 symptoms are stored in an array which helps the user in listing out their symptoms. After the user enters all his symptoms, rules that were generated using affinity analysis for market basket are used. They help the user in listing out more relevant symptoms that the user must have missed out on. Further, the set of symptoms are passed onto the server where the decision tree classifier is trained. The predict proba method of scikit-learn is used to predict the probability value of all the 41 classes(diseases). The obtained probability array is sorted and difference is calculated between the first three classes. If the difference is less than the set threshold value, then there could be a slight chance that the disease predicted may not be correct. In that case, we pass multiple possible diseases to the user and the prescription for such case will be explained in next section. If the predicted class/disease has

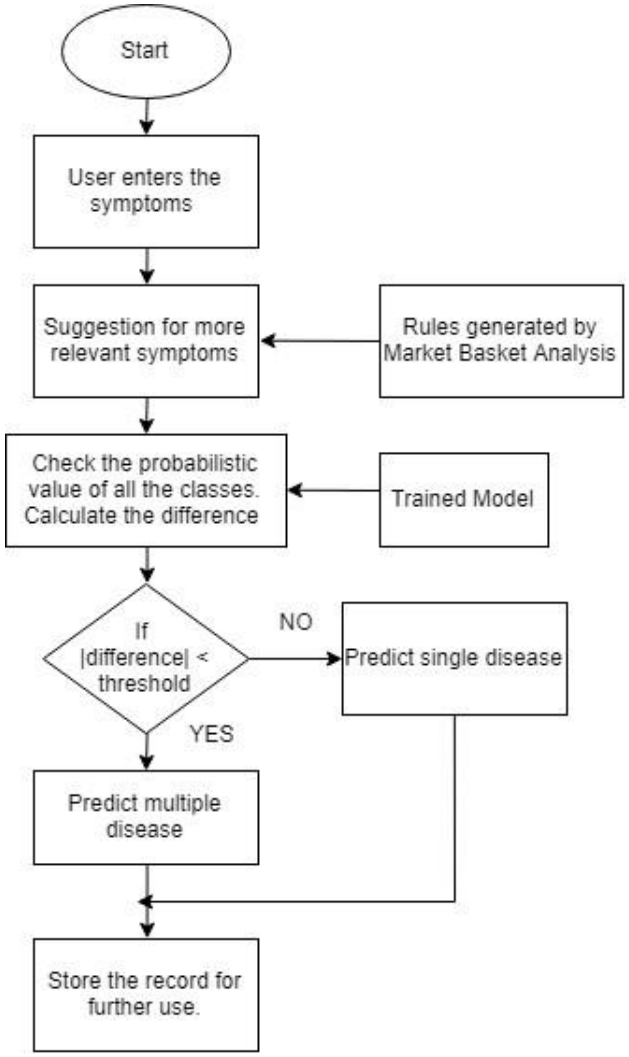

Fig. 6. Flowchart for Disease Prediction

comparatively higher probability than other classes, then that disease is displayed as the result. This particular user case is stored as a record in our database.

B. Prescription Generation

We have implemented content based recommendation system for generating a suitable prescription for the recognized disease as obtained from the result of step 1. This has been done for two kinds of cases :

1. Single Disease Prediction

The prescription generated for this case takes into consideration the various attributes of that user like weight, age, allergies and further sub-types of the recognized disease based on some more specific questions for that disease. For each use-case, a different prescription or quantity of drug is generated. The system also produces the instructions of dosage and frequency of consumption of each drug. For example, for Tuberculosis, there is a single tablet having all the 4 or 5 drugs together, and the dose of the tablet has to be given on the basis of drug Isoniazid and Body Weight. Eg: 60kg patient needs 300mg Isoniazid hence 1 combo tablet of 300mg Isoniazid will have sufficient of the other drugs; but 80kg patient needs 400mg Isoniazid which will be covered in 2 combo tablets and will have sufficient of the other drugs. Furthermore for the case of tuberculosis, the prescriptions are different based on two more factors

i. Phase : Intensive phase or Continuation phase ii. Case : New case or previously treated and discontinued cases.

2. Multiple Disease Prediction

Considering the case where the system recognizes multiple diseases and the difference of the probabilities of the top 3 recognized diseases is less than the threshold value, an intersection of the symptoms of these diseases is extracted. This set is compared with the symptoms initially inserted by the user. If these two sets are identical or if the cardinality of the set inserted by user is more than that generated after taking the intersection, the drugs required to subdue the effect of these symptoms are prescribed using the database that contains prescriptions for common symptoms and like the previous case, the attributes of the user are taken into consideration to provide content based prescription. Now, if the cardinality is less than that of the set generated after taking the intersection, the newly observed symptoms are suggested to the user as they are most likely to be observed. If the user agrees to the presence of these symptoms, the model is rerun for the new set of symptoms.

VI. RESULTS

The purpose of applying five classifiers is to get an idea of which model gives the most efficient results and in future which of the following classifiers can be combined together. The weighted average of these combined classifiers can help in devising a model that would accurately predict the results. With the available dataset, the highest classification accuracy is 97.91% for Decision Tree Classifier which can be inferred from Table. I. It is observed that they also have the highest Sensitivity and Specificity values.

For the apriori algorithm in market basket analysis, the candidate set values are $|C1|$ =132, $|C2|$ =944 and $|C3|$ =3506. And the set values after pruning are $|L1|$ =131, $|L2|$ = 904 and $|L3|$ =3413. With the constraint values as specified earlier, 9791 rules are generated from the symptoms of the manual dataset records. Some of the rules generated are :

{swollen extremeties} ->{*dizziness,fatigue*},*conf* = 0.85,*lift* =41.18,*conv* =6.42

$\{yellow\_urine\} \mapsto \{itching, lethargy\}, conf$ $0.89, lift = 20.37, conv = 8.92$

TABLE I
COMPARISON OF THE CLASSIFIER RESULTS - STAGE 1

Sr.No.	Classifier	Accuracy
1	KNN	89.91%
2	SVC	86.47%
3	MLP	79.71%
4	DecisionTree	97.91%
5	GaussianNB	95.26%

VII. FUTURE WORK

The implementation is not meant to substitute the real doctors, rather acts as a bridge between the patient and the doctor. The algorithms and processing techniques are currently implemented on the manual data-set hosted on [9]. For better real time results,

the symptoms and disease of different age group and ethnicity should be collected from various hospitals around. With a larger data-set and wider range of patients, the parameter of various classifiers can be accordingly tweaked. For stage 2, only 10 diseases have been covered and their corresponding prescriptions have been noted. A system generating prescription for more number of diseases can be developed, which consists of prescription collected and verified by a group of doctors. Also, the predicted disease and the generated prescription can be verified by a doctor for every case over the internet, which enables the user to build more trust over the system.

VIII. CONCLUSION

In this paper, we depict a Decision Tree Classifier to predict the disease based on the user-entered symptoms. Furthermore, to ameliorate the symptoms-input process, affinity analysis is conducted to perform market basket analysis and the generated rules are used for prediction of related user symptoms. If the probabilistic value of the classes lie really close, we predict multiple diseases. For the prescription of multiple diseases, we take the intersection of the symptoms and generate the corresponding prescription. We also encourage the user to visit the doctor at the earliest and have the medical check-up done for the predicted disease. Such system can play a major role where doctors are unavailable, or not that skilled. Further, it spreads health-care awareness among the people who have the habit of ignoring crucial symptoms, which then turn out to be of great danger to their health.

REFERENCES

1. "Smilefoundation,"https://www.smilefoundationindia.org/Media/rural healthcare.html.
2. D. Barik and A. Thorat, "Issues of unequal access to public health in india," *Frontiers in public health*, vol. 3, p. 245, 2015.
3. V. L. Narayanan, "Human development in india: Challenges for a society in transition," 2011.
4. "Reimagining the possible in the indian healthcare ecosystem with emerging-technologies "https://www.pwc.in/assets/pdfs/publications/2018/reimaginingthe-possible-in-the-indian-healthcare-ecosystem-with-emergingtechnologies.pdf.
5. "Disease prediction, machine learning and healthcare", https://dzone.com/articles/disease-prediction-machine-learning-application-fo.
6. "Drug allergy," https://www.emedicinehealth.com/drug allergy/article em .htm#what _is a drug _allergy.
7. J. Singla, "The diagnosis of some lung diseases in a prolog expert system," 2013.
8. J. S. Folasade O Isinkaye and O. P. Oluwafemi, "A mobile-based neurofuzzy system for diagnosing and treating cardiovascular diseases," 2017.
9. "ManualData",https://github.com/AniruddhaTapas/PredictingDiseases-From-Symptoms/tree/master/Manual-Data .
10. S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.
11. S. S. Sundhari, "A knowledge discovery using decision tree by gini coefficient," in *2011 International Conference on Business, Engineering and Industrial Applications*. IEEE, 2011, pp. 232–235.
12. S. Soni and O. Vyas, "Using associative classifiers for predictive analysis in health care data mining," *International Journal of Computer Applications*, vol. 4, no. 5, pp. 33–37, 2010.

AUTHORS PROFILE

Jainam Soni, MS in Computer Science, SBU BE in Computer Science, SPIT Publication - "Two-Stage Approach for Detection of Invasive and Cervical Intra- Epithelial Neoplasia" Undergraduate Teaching Assistant for Object- Oriented Programming and Data Warehouse and Mining

Riya Bakhtiani, BE in CS, SPIT Publication - Big Data Strategies A Review and Survey Undergraduate Teaching Assistant for Object-Oriented Programming Co - Curator: TEDxSPIT 2018 and TEDxSPIT 2019

Dr.Dhananjay Kalbande Dean (Industry Relations), Professor and Head of Department (Computer Engineering), SPIT
Post-Doctorate (TISS), Ph.D.,M.E.(I.T.), B.E.(Comp.) Senior Research Fellow (NCW-TISS Project,T.I.S.S., Mumbai

Nishant Bakhrey, BE in CS, SPIT Undergraduate Teaching Assistant for Object-Oriented Programming Curator and Licensee: TEDxSPIT 2018 and TEDxSPIT 2019