Smart Health Prediction System Using Data Mining

Sachin Pundir

BTECHCSE, GRAPHIC ERA HILL UNIVERSITY

Abstract- The digital technology era demands the world to provide an excellent health system, to ensure the citizen and community to be alive and healthy. This work proposes the application of data mining algorithm for health prediction that can eventually shape a suitable health prediction system for patients. Although health care is available to everyone in the world, there is still no healthcare system that is completely reliable and accurate to carefully diagnose a patient on their current health issues. Even though some hospitals are well equipped to provide the best healthcare services to its citizens, some of the hospitals are still lacking in certain qualities. Consequently, patients are doubtful and uncertain when it comes to picking which hospital suits them. Numerous issues are faced by patients pertinent to hospitals such as being unable to provide medical services, insufficient number of qualified medical staffs, poor communication between doctors and patients, and unorganized health records and data. Patients need treatment and diagnosis that are accurate and precise for them to be able to recover back for their proper health and medical staffs are required to be well equipped in their clinical knowledge and communication skills to carefully assess their patients to ensure good health. Therefore, application of data mining in health prediction is considered in this work as the best practice to facilitate better healthcare system.

***Index Terms*- :** Healthcare, classification, Disease, Prediction , Prescription, Clinical prediction, Data mining, Smart health system, Medical service, Health prediction, Electronic health records.

I.INTRODUCTION

IT might have happened so many times that we or someone ours need doctors help immediately, but they are not available due to some reason. The Health Prediction system is an end user support and online consultation project for correct prediction of illness based on patients input. Here propose system allows users to get instant guidance on their health issues through an intelligent health care system online. The smart health prediction system is fed with various symptoms and the disease/illness associated with those systems. The system allows user to share their symptoms and issues then system processes patients symptoms to check for various illness that could be associated with it. Here some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's symptoms. If the system is unable to provide suitable results, it informs the user about the type of disease or disorder it feels user's symptoms are associated with. If patients symptoms do not exactly match any disease in our database, is shows the diseases user could probably have judging by his/her symptoms. Disease prediction using patient input symptoms history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Necessity: Sometimes we need the help of doctors immediately, but due to some reasons they unavailable.

In project proposed system is user favourable to get guidance on health issues instantly through online health care system. The System is helpful in emergency of patients by suggesting the doctors and immediate prescriptions on their disease. patient can get help from anywhere at any time. In medical fields, the foreign students have solved some medical issues that are laborious to be settled in classic statistics by classification of Bayesian. Without an extra information, classification rules are generated by the samples trained by themselves.

I. LITERATURE SURVEY

2.1 Research work

Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in liver, kidney and heart disease prediction. Classification algorithm is the most common way in several automatic medical health diagnoses. Many of them show a good classification accuracy listed below

[1]. The paper “Smart E-Health Prediction System Using Data Mining” applies the data mining process to predict hypertension from patient medical records with eight other diseases.
[2]. The reference paper “A Health Prediction Using Data Mining” Predict Liver Disorder with methods C4.5, NBC in 2016. The result shows NBC algorithm has the highest accuracy .
[3]. It analyzed the Liver Disorder Using Data Mining Algorithm with Naïve Bayes algorithms, FT tree algorithm , and K Star in 2010. A novel approach for Liver disorder Classification using Data Mining Techniques” with methods Fuzzy, K-means classification in 2015.
[4].Fuzzy based classification gives better performance and achieved above 94% accuracy for each type of liver disorder analyze the performance of Classifier Over Liver Diseases ”.
[5] It have propose Prediction of Liver Disease (Biliary Cirrhosis) Using Data Mining Technique with method FT Tree algorithm in 2015. The classification accuracy is found to be better using FT Tree algorithm Liver Disease Prediction using Data Mining Technique with method Naïve Bayes SVM in 2015. The SVM classifier is considered as a best classification algorithm .
[6]. The paper “An approach to devise an Interactive software solution for smart health prediction using data mining”
[7] aims in developing a computerized system to check and maintain your health by knowing the symptoms. The another research have diagnosis the Chronic Kidney Disease Using Machine Learning Algorithms” with methods Random Forest ,Back Propagation ,Radial Basis Function in 2016.The results indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection
[8].This is the most effective model to predict patients with heart disease proposed Classification of Heart Disease Using K-Nearest Neighbour and Genetic Algorithm with methods KNN, Genetic algorithm in 2013. KNN is more accurate than genetic algorithm
[9].The Intelligent Heart Disease Prediction System Using Data Mining Techniques with methods Decision Trees, Naïve Bayes and Neural Network in 2008
[10].It is user friendly, web based, scalable, reliable and expandable Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm with methods Genetic Algorithm in 2008
[11]. Medical Knowledge Acquisition through Data Mining with method KNN in 2008. This paper presents a model of medical knowledge acquisition through data mining .

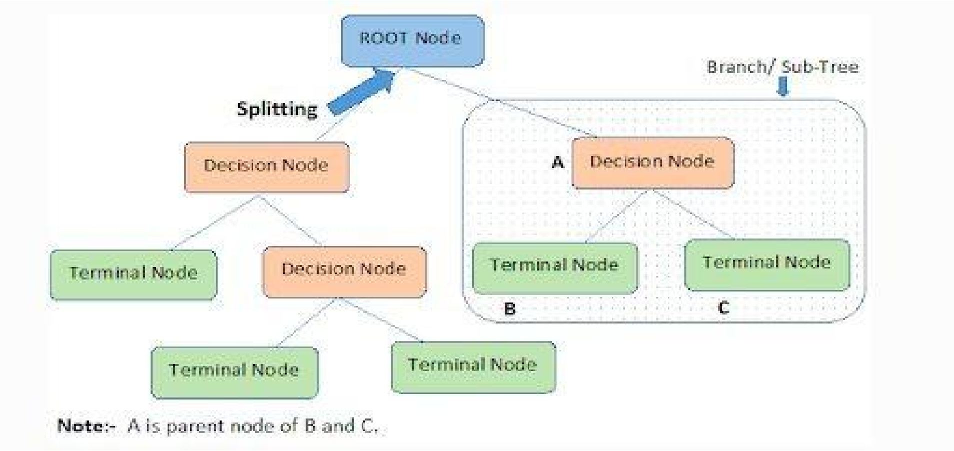
II.DESIGN METHODOLOGY AND ALGORITHM

A. DECISION TREES

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Important Terminology

- **Root Node:** It represents the entire population or sample and this further gets dividedinto two or more homogeneous sets.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called thedecision node.
- **Leaf / Terminal Node:** Nodes that do not split are called Leaf or Terminal nodes.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning.You can say the opposite process of splitting.
- **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.
- **Parent and Child Node:** A node ,which is divided into sub-nodes is called a parent node of sub-nodes where as sub nodes are the child of a parent node.



If the dataset consists of N attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criteria like Entropy, Information gain, Gini index, Gain Ratio, Reduction in Variance, Chi-Square. These criteria will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e. the attribute with a high value (in case of information gain) is placed at the root. While using Information Gain as a criterion, we assume attributes to be categorical, and for theGini index, attributes are assumed to be continuous.

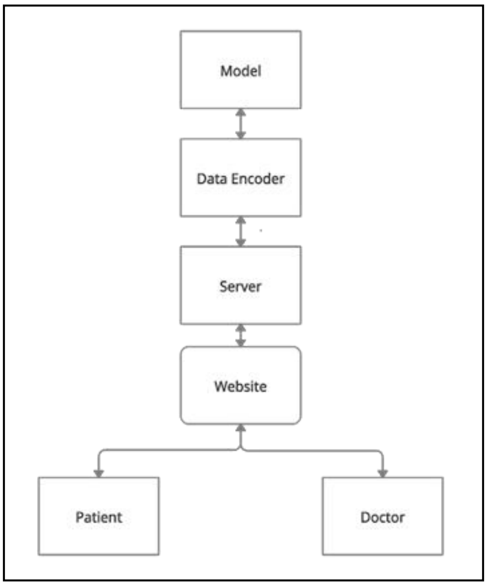
1.1 Technology Stack

Data Science and Machine Learning libraries based on python were used in this project. The main tool used for coding was Jupyter Notebooks. Pandas and NumPy were used for data analysis. Streamlit was used to create a user interface for the web application. Finally Machine Learning models available in the Scikit Learn library were used for model training and evaluation primarily including decision tree and random forest algorithm.

IMPLEMENTATION

2.1 Software Architecture

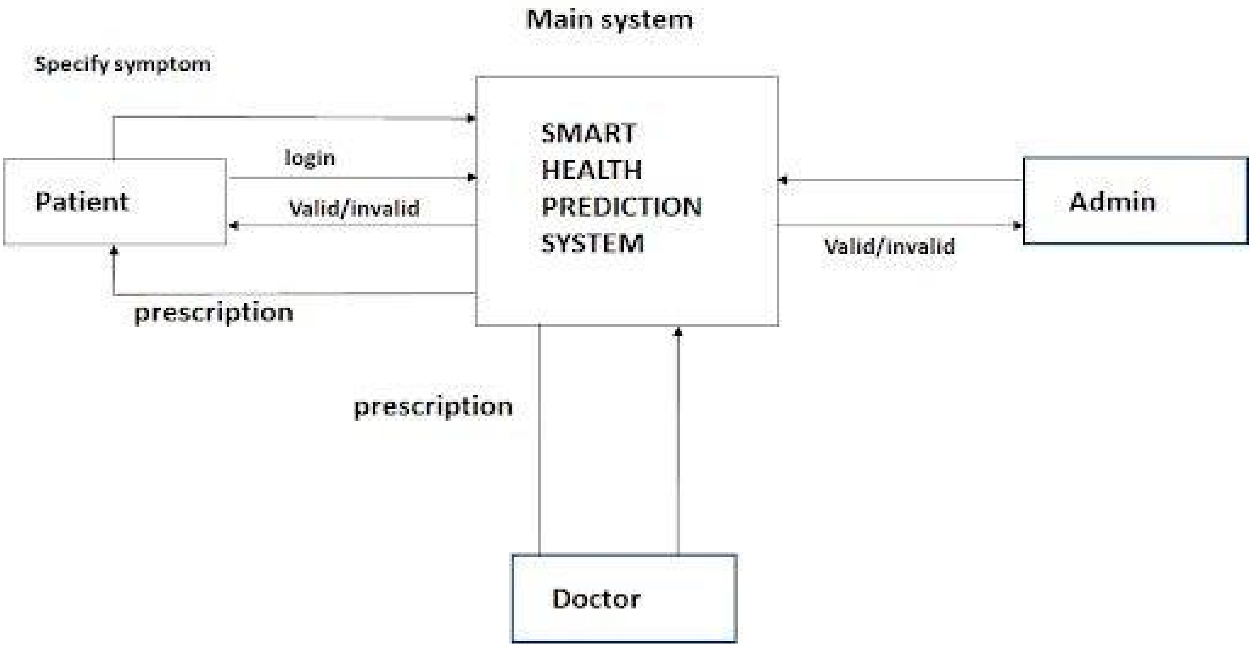
The software architecture follows an encoded path from model to hosted website on the user side, where the user has access to patient portal information and the doctors can see the database for all the previously evaluated symptoms and predictions, this in turn gives the doctor administrator rights to check and consult with the patient depending on the severity.



2.1 Design Architecture

It is an end user support and online consultation project. Here we propose a system that allows users to get instant guidance on their health issues through an online intelligent health care system. The system is fed with various symptoms and the disease/illness associated with those symptoms. The system will allow the users to share their symptoms and issues. It then processes the user's symptoms to check for various illnesses that could be associated with it. Here we use some intelligent data mining techniques to guess the most accurate illness that could be associated with a patient's symptoms. In the doctor module when a doctor logs in to the system doctor can view his patient details and the report.

Doctors can view details about the patient search and what patient searched for according to their prediction. Doctor can view his personal details. Admin can add new disease details by specifying the type and symptoms of the disease into the database. Based on the name of the disease and symptom the data mining algorithm works. Admin can view various diseases and symptoms stored in the database. This system will provide proper guidance when the user specifies the symptoms of his illness. In our project we propose a system that is user favorable to get guidance on health issues instantly through an online health care system.



B.DATA MINING TECHNIQUES

Data mining can be described as a process of searching patterns or correlations from large data sets to valuable information that can solve problems and predict outcomes. It involves analyzing certain amount of information to locate certain patterns of occurrence to predict future tendencies, using several processes of effective data collection, warehousing and computer processing. With this functionality, therefore, it serves a great purpose when it comes to predicting people’s health diseases especially on finding the correlation between the health information that has been given by both the medical staff and the patient. These finding may provide a beneficial advantage in the healthcare industry as it may be used to manage patients on their current health issues and for the doctor to alleviate them from their jobs.

There are currently a lot of health institutions that have been developed such as hospitals and medical centres which are crucial to maintain and improve the health of the community around us. It is a prime establishment of giving proper health care especially for every one of us who have ever lived. For every illness and diseases that people may face today and sometime in the future, it is because of these medical institutions and all the doctors who worked at these places that have made our lives physically better and healthy. Although hospitals now are well-equipped with their staffs working, there are still known issues that persist that cause the staffs to make the poor clinical decision that affects a patient’s health such as the lack of qualified doctors, unorganized health information and poor communications between doctors and patients.

Data mining techniques such as association, classification and clustering are used by healthcare organization to increase their capability for building appropriate predictions regarding patient health information from large data. This encompasses a number of technical approaches like clustering, data summarization, classification.

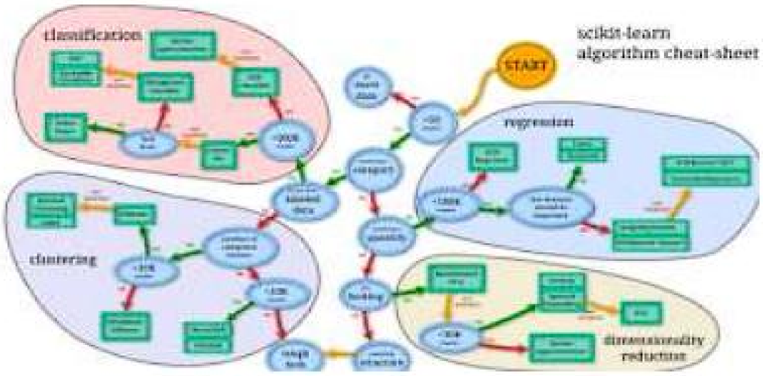


Figure 4.1 Algorithm Analysis

Classification:

Classification comprises of two steps: -

1) Training and 2) Testing.

Training builds a classification model on the basis of training data collected for generating classification rules. The IF-THEN prediction rule is popular in data mining; they signify facts at a high level of abstraction. The accuracy of classification model based on the degree to which classifying rules are true which is estimated by test data.

Prediction:

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction in data mining is to identify data points purely on the description of another related data value

IV.CONCLUSION

Machine learning enabled individuals in tackling challenges and shortcomings that else would have been burdensome in a prudent manner. Machine learning tools aid to bring out insights on data to analyze patterns and to build models to make predictions. Having Machine learning methodologies being enforced in the health domain benefits for processing immense amounts of data beyond the scope of human ability, vivid predictions to be formulated with machine learning models and assistance for physicians for diagnosis in an efficacious manner. All those tedious and time consuming processes can be hastened to save both time and labor. Our project entitled in the name ‘The Health Prediction system’ provides assistance to determine the possible disease in reference to symptoms. However the challenges are still unsolved. Models are prone to overfitting that may end up in wrong predictions. Diagnosis cannot be done merely in light of symptoms, there exist various factors concerned about the patient that can lead to diseases. They include lifestyle, gender, hereditary etc. Advancements have to be brought in models to predict the disease based on factors other than symptoms which aids doctors to rely on these models for efficient disease predictions.

V. REFERENCES

1.Karleshia, Y.Nagesh and M.VeeraKrishna, ”Performance comparison of three data mining techniques for predicting kidney disease survivability”, International Journal ofAdvances in Engineering & Technology, Mar. 2014.

2.D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in healthcare:using analytics to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

3.Disease prediction methodology: D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd InternationalConference on Computing Methodologies and Communication (ICCMC), Erode, India,2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782

4.Boshra Brahmi, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Diseaseby Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.

5.Streamlit official documentation: https://docs.streamlit.io/en/stable/

6.Decision tree and random forest exploration: https://www.khanacademy.org/computing/computer-science/informationtheory/info-theory/pi/decision-tree-exploration