SYMPTOMS BASED DISEASE PREDICTOR

"First Step Towards Healing"

Veronica Naosekpam,

Bachelor of Engineering, Department of Computer Science,
Siddaganga Institute Of Technology , Tumakuru, Karnataka

Abstract : Day by day, in the field of medical science, there is a rise of lot of diseases in the world. It is difficult to analyze all kind of diseases and how to take the correct medication for all the diseases. This task is very difficult. The healthcare industry collects huge amount of data which are not mined unfortunately. If these data are mined, they will be helpful to discover hidden information for effective decision making. The data mining techniques are very useful for finding the medicinal decision for the appropriate diseases. Data mining applications in healthcare can have tremendous potential and usefulness. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. The motivation behind the project is to facilitate medical solicitation at the earliest stage to the user and providing free of cost system of diagnosis. Correct diagnosis is critical for effective treatment and of disease. Quality of service provided to the patients could be increased by eliminating unwanted bias, errors and excessive medical costs.

IndexTerms **- Data mining, data science, disease, symptoms, prediction, decision making, tree maps;**

INTRODUCTION

Background Study:

The computerization of various fields like Science and Engineering, Medicine, Business and Society has led to the explosive growth of available data. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. It can also be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns that is understandable to human. Data mining is not new it has been used intensively and extensively by Financial institutions; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling. The tasks in data mining can be divided into two categories that is, predictive and descriptive. In the predictive model, it makes a prediction about values of data using known results found from different data and its goal is to identify strong links between variables of a data table. Predictive model data mining tasks involves the classification, prediction, time series analysis and regression. Descriptive model identifies patterns or relationships in data. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. An important task in medical diagnosis is that it should be performed as accurately and efficiently as is possible. Unfortunately, all doctors are not equally skilled in every sub-speciality and they are a scarce resource in many places. A system for automated medical diagnosis would enhance medical care and reduce costs. A century ago, it was noted that a symptom was related to a single disease. But, now a symptom is related to a number of diseases. Hence, computer assisted information retrieval may help to support quality decision making and to avoid human error. As medicine plays a great role in human life, automated knowledge extraction from medical data sets has become an immense issue. There is fast growth in the research field on knowledge extraction from medical data.

Motivation:

Data mining is largely concerned with building models. A model is simply an algorithm or set of levels that connects a collection of inputs to a particular target or outcome. Data Mining has great potential for exploring the meaningful and hidden patterns in the data sets at the medical domain, these methods can be used for the medical and diseases diagnosis. Correct diagnosis is critical for effective treatment and prevention of disease. As a result, disease classification has become a key cornerstone of modern medicine. Using data mining as a convenient tool, an application to assist a non-medical person to detect or diagnose a disease that he/she has can be built based on the symptoms that he/she provides and also recommends the doctors and the tests and medications to undergo. Thus quality services can be provided at low cost as some diagnostic and laboratory procedures are costly and painful to patients. Medical history data comprises of a number of tests essential to diagnosis a particular disease. The data generated by the healthcare transactions are huge amounts and they are too complex and voluminous to be processed and analyzed by traditional methods. Decision making in data mining can be improved by discovering patterns and trends in large

amounts of complex data. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data.

A non-medical person can use data mining application to detect or diagnose a disease that he/she has and can be built based on the symptoms provided by the user and the system also recommends the doctors and the tests and medications to undergo after the detection of the probable disease. It saves time and costs. And it can also be used as a "First step towards healing" application. Data mining applications also can benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices.

Various factors are boosting data mining popularity. For instance, as a result of the Balanced Budget Act of 1997 (USA) [1], the Centres for Medicare Services must implement a prospective payment system based on classifying patients into case-mix groups, using empirical evidence that resource use within each case-mix group is relatively constant. CMS has used data mining to develop a prospective payment system for inpatient rehabilitation. The healthcare industry can benefit greatly from data mining applications. In addition to this, United Health Care has mined its treatment record data to explore ways to cut costs and deliver better medicine. It also has developed clinical profiles to give physicians information about their practice patterns and to compare these with those of other physicians and peer-reviewed industry standards. Similarly, data mining can help identify successful standardized treatments for specific diseases. According to the paper by Hian Chye Koh and Gerald Tan [5], launched the clinical best practices initiative with the goal of developing a standard path of care across all campuses, clinicians, and patient admissions. Blue Cross [1], has been implementing data mining initiatives to improve outcomes and reduce expenditures through better disease management. For instance, it uses emergency department and hospitalization claims data, pharmaceutical records, and physician interviews to identify unknown asthmatics and develop appropriate interventions.

PROPOSED WORK

In this era of technology boom, it has become quite essential for humanity to take steps further into the seemingly future technology. Data mining is such a tool which opens the portal for advancement. Data mining provides a set of techniques to discover hidden patterns from data. It is used to discover knowledge out of data and presenting it in a form that is easily understandable to humans. There are two primary goals of data mining, prediction and description.

A major challenge facing healthcare industry is quality of service. Quality of service implies diagnosing disease correctly and provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable. For detecting a disease, number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time, cost and performance. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. The proposed approach analyzes how data mining techniques are used for predicting different types of diseases. The system will predict the diseases based on the symptoms given by the patients. In this approach user is allowed to give symptoms which are considered as data set based on these data sets for total number of transaction the list of possible diseases is computed and hence, with the list of possible diseases one disease is predicted. Any number of diseases can be updated/ added by the admin.

The objectives of the proposed work is to collect the medical data as training sets, to develop a model for the disease prediction using Decision Making technique. For a given set of symptoms, prediction of probable disease will be done with an accuracy of 80-85%. Also, to recommend the doctors based on the relevant disease for medical queries. To recommend what probable tests to undergo for that predicted disease.

SOFTWARE DEVELOPMENT METHODOLOGY USED

Classical Waterfall model Waterfall model is a sequential design procedure, used in software development processes, in which progress is seen as owing steadily downwards through the different phases. Classical waterfall model divides the life cycle into the following phases as shown in the below figure.

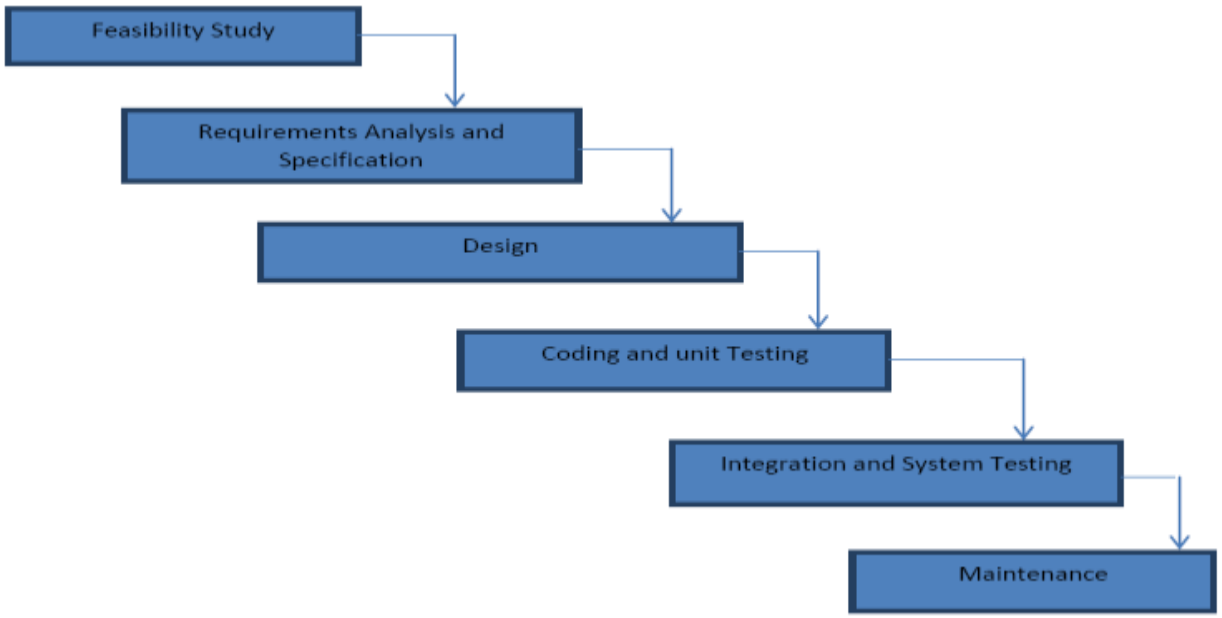

FIG: WATERFALL MODEL

Activities undertaken during requirements analysis and specification:
The aim of the requirements analysis and specification phase is to understand the exact requirements of the customer and to document them properly. This phase consists of two distinct activities, Requirements gathering and analysis and Requirements specification. The goal of the requirements gathering activity is to collect medical data sets such as symptoms and the corresponding diseases from various authorised healthcare centres. Also, collecting the information about doctors by conducting the discussions with residents of the particular locations and then this information is further used for recommending to the patients. Tests to be undertaken for the particular disease is listed. After all ambiguities, inconsistencies, and incompleteness have been resolved and all the requirements properly understood, the requirements specification activity can start.

Activities undertaken during design:
The goal of the design phase is to transform the requirements specified in the SRS document into a structure that is suitable for implementation in some programming language. The suitable decision making technique to be used for processing the so far collected medical data sets has to be chosen. Software and hardware components needed for developing the application will be decided and the overall model of the upcoming application will be designed.

Activities undertaken during coding and unit testing:
The purpose of the coding and unit testing phase (sometimes called the implementation phase) of software development is to translate the software design into source code. Each component of the design is implemented as a program module such as admin module, user module etc. During this phase, each module is unit tested to determine the correct working of all the individual modules. It involves testing each module in isolation as this is the most efficient way to debug the errors identified at this stage.

Activities undertaken during integration and system testing:
Integration of different modules is undertaken once they have been coded and unit tested. During the integration and system testing phase, the modules are integrated in a planned manner. The different modules making up a software product are almost never integrated in one shot. Integration is normally carried out incrementally over a number of steps. During each integration step, the partially integrated system is tested and a set of previously planned modules are added to it. Finally, when all the modules have been successfully integrated and tested, system testing is carried out. The goal of system testing is to ensure that the developed system conforms to its requirements laid out in the SRS document.

Activities undertaken during maintenance:
Maintenance of a typical software product requires much more than the effort necessary to develop the product itself. Many studies carried out in the past confirm this and indicate that the relative effort of development of a typical software product to its maintenance effort is roughly in the 40:60 ratios.
During maintenance we have to do one or more of the following activities. Correction errors that were not discovered during the product development phase were done in this phase. Also, it does the work of improving the implementation of the system, and enhancing the functionalities of the system according to the customers requirements.

DATA STRUCTURE AND ALGORITHM USED

Tree map is the main data structure being used in the algorithm. Treemap provides an efficient means of storing key/value pairs in sorted order and allows rapid retrieval. In this function Treemap stores key being number of symptoms being matched to the

particular disease and value being disease name. The disease having highest information gain i.e. value will be the probable disease. ArrayList is also used to store the result set obtained after executing the query on the symptoms stored in the database. Decision making algorithm using Treemaps is used. For each of the symptoms being entered by the user, first list will be populated with all the disease names containing that symptom. After that TreeMap will store key/value pairs as key being number of symptoms being matched to the particular disease and value being disease name.

CONCLUSIONS AND FUTURE SCOPES

The Project titled, "Symptoms Based Disease Prediction", is a datamining project, which in this era, tries to make use of the peta bytes of available data. Health Sector is a field where the data mining technique hasn't firmed its roots, that is to say that the use of data mining has been majorly confined to "Heart Diseases ". It was our humble attempt to bridge this gap through our project, "Symptoms Based Disease Prediction" and bring in the benefits of data mining to the greater population. In our project we have tackled the following objectives:

1. The user would be posed with a series of questions, with the current question being dependent on the answer of the previous question and this forms the basis of the symptoms being felt by the user.
2. Based on these symptoms, a list of possible diseases is determined, by the algorithm and in that list, the most probable disease is displayed.
3. A brief statement regarding the disease is given, including what tests to undergo and what precautions to take.

Limitations of the project are, project has been able to predict only few diseases and it can provide only the static map of the location of the recommended doctor. As with everything there is always room for improvement and enhancements. It is one of the unsaid laws in the field of software, that a major portion of the product life cycle is dedicated to updates and enhancement.

Here below are stated few of the possible updates, that could be expected:

1. Using Plugins from the Google Maps to ensure that the user has a map reference from his current position to the specific hospital.
2. Using Machine Learning to train the data set, and create a reference model
3. Using feedback from the users to train the model and make it more accurate.

ACKNOWLEDGEMENT

I consider this as a privilege to express a few words of gratitude to all those who guided and inspired me for the successful completion of this work especially Mrs. Ashwini BP, Assistant Profesor, Department of Computer Science, SIT.

REFERENCES

[1] Predictions in heart disease using techniques of data mining", Gandhi M, Singh S N, Computer Sc and Engg Dept, Amity University,Noida, Futuristic trends on Computer Analysis and Knowledge Management (ABLAZE),2015,IEEE,25-27February,2015.

[2] Intelligent heart disease prediction system using data mining techniques", Palaniappan Malaysia Univ. of Science and Technology, Jaya, Awang R, Computer System and Applications, 2008, AICCSA, IEEE/ACS Conference, March31, 2008-April4, 2008.

[3] "Automated Medical Diagnosis based on Decision Theory and Learning from Cases", Magnus Stensmo, Computer Science Division, UC Berkeley, CA (USA), Terrence J. Sejnowski, Computational Neurobiology Lab, The Salk Institute, CA, USA, World Congress on Neural Networks 1996 International Neural Network Society, 1227-1 231.

[4] "A Reliable Classifer Model Using Data Mining Approach for Heart Disease Prediction", I.S.Jenzi, P. Priyanka, Dr.P.Alli, Department of Computer Science and Engineering, Velammal College of Engineering and Technology , Madurai, Tamil Nadu, India, Volume 3, Issue 3, March 2013 ISSN: 2277 128X ,International Journal of Advanced Research in Computer Science and Software Engineering.

[5] "Data Mining Applications in Healthcare", Hian Chye Koh and Gerald Tan, Journal of Healthcare Information Management - Vol. 19, No. 2.