Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 1, January-2018

DISEASE STATUS PREDICTION AND IDENTIFICATION

Pragati Shukla[1], Simran Lal[2], Gauri Kumbhar[3], Susmita Kulkarni[4], Mrs. V. V. Waykule[5]

[1]*Computer Engineering, AISSMS College Of Engineering*
[2]*Computer Engineering, AISSMS College Of Engineering*
[3]*Computer Engineering, AISSMS College Of Engineering*
[4]*Computer Engineering, AISSMS College Of Engineering*
[5]*Computer Engineering, AISSMS College Of Engineering*

Abstract- Electronic Health Records (EHRs) are the main source of information for assessment, diagnosis, and treatment of disease in clinical care. An EHR typically contains a patient's historical health data, collected over several years of patient care. This data includes both physician's clinical notes written in unstructured text recording their observations, assessments, and plans, as well as structured data such as ordered medications, vital signs measurements, laboratory test results, and procedures conducted. The system takes input and helps the user to predict the disease. The result for the same is provided. The proposed system assists doctor to predict disease correctly and the prediction makes patients and medical insurance providers benefited. The use of EHRs are very limited when the scenario in our country is taken into account. This can also benefit the physician since the patient history will be readily available and in a structured format. Through the visits the results will be stored and a record will be maintained. Thus, our system will enhance the usage of EHR to store data as well as to predict the disease accurately and efficiently.

Keywords-data mining; clinical decision support system; expert application; disease prediction; C4.5.

I. INTRODUCTION

Computational health informatics is an emerging research topic which involves various sciences such as biomedical, medical, nursing, information technology, computer science and statistics. Data mining techniques are applied to predict the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among huge clinical and diagnosis data. In medical science, doctor's facilities introduced different data frameworks with a lot of information to manage medical insurance and patient information but unfortunately, data are not mined to discover hidden information for effective decision. Clinical test outcomes are regularly made on the basis of doctor's perception and experience rather than on the knowledge enrich data masked in the database and sometimes this procedure prompts inadvertent predispositions, doctors expertise may not be capable to diagnose it accurately which affects the disease diagnosis system . In health care sector, the term information mining can mean to analyze the clinical information to predict patient's health status. So discovering interesting pattern from health care data, different data mining techniques are applied with statistical analysis, machine learning and database technology.

1.1 BACKGROUND

The current status of the healthcare sector in India is associated with low public spending (1% of GDP), high out-ofpocket payments (71%), a high level of anemia among young women (56%), high infant mortality (47/1,000 live births), and high maternal mortality (212/100,000 live births), etc. India has a mixed system of healthcare consisting of a large number of hospitals run by the Central Government and State Government as well as the private sector. In general, the level of use of ICT(Information and Communication Technology) in the healthcare sector in the country has been lower in comparison to other countries. At the same time, both union and State Governments are working on several fronts to make use of the opportunities covered by ICT. Private sector hospitals are also in the process of implementing ICT projects, including electronic patient records.

Some of the corporate hospitals in India, such as Max Health, Apollo, SankaraNethralaya, Fortis, etc., have implemented integrated ICT systems in place, covering all aspects, i.e., registration and billing as well as laboratory and clinical data. Max Healthcare hospitals started implantation of EHR in its hospitals in 2009 and achieved Stage 6 level of the EMR

Adoption Model, which is used by the HIMSS for assessment of the level of adoption of EMR systems in any hospital. Max Healthcare Group received the recognition for two of its hospitalsEast Wing, Saket and West Wing, Saket, New Delhi in 2012.

However, even in private hospitals, EMRs are rarely exchanged between hospitals. These remain in the same hospital and are referenced when the patient visits again. There is no authentic report on the number of patients whose EMRs/EHRs have been stored so far.

1.2 PROBLEM STATEMENT

With the increase in health care facilities, it is also necessary to store patient information for the ease of the physician and the government, so that better measures can be taken in future. Since EHRs are used in many private organizations for storing patient information, the same can be used to mine data and trace out patterns for better understanding. This can be done using various data mining techniques and the discovered patterns may help doctors for better decision making in the future.

1.3 PURPOSE

The proposed system assists doctor to predict disease correctly and the prediction makes patients and medical insurance providers benefited. The system focuses on diagnosis of diabetes as it is a great threat to human life worldwide. The system uses the Decision Tree,C4.5 Algorithm as supervised classification model. Finally, the proposed system calculates the accuracy of C4.5 and the experimental result demonstrates that the C4.5 provides better accuracy for diagnosis of diabetes. The proposed system will help in quick clinical decision making.

II . LITERATURE REVIEW

2.1 SUPERVISED LEARNING

Supervised machine learning is a machine learning algorithm that uses a labeled dataset for prediction. Labeled Data means an output is associated with every input. It means youhave input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$Y = f(X)$

The goal is to approximate the mapping function so well that when you have new input data (x) you can predictthe output variables (Y) for that data. In other words, supervised learning builds a model using this dataset that can make a prediction of the new or unseen dataset. This unseen or new dataset is called training dataset and helps in validating the model. Supervised algorithms can be used for classification and regression. Classification algorithm(s) like Support Vector Machines, Naive Bayes, Decision trees can be used to build the classifier for a system.

2.2DECISION TREE

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes (has two or more branches) and leaf nodes (represents classification/decision). The topmost decision node corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

2.3C4.5 DECISION TREE ALGORITHM

Decision trees generated by C4.5 are used for classification. C4.5 is often referred to as a statistical classifier. C4.5 algorithm is described as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date". Hence,also being used in the proposed system. C4.5 builds decision trees from a set of training data using the concept of information entropy. C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists.

III. SYSTEM DESIGN

3.1 SYSTEM ARCHITECTURE

The overall system design consists of following modules: (a) Data Collection. (b) Preprocessing (c) Data classification (d) Prediction of Output. Through the proposed application user (doctor, patient, physician etc.) can input the attribute values of disease and send it to the server with the help of internet. After applying the data mining approach the predicted result can be viewed on the user GUI. On the server, admin can load dataset of different diseases and apply different data mining algorithms to train dataset. Requested user inputs are collected and processed on server to predict the diagnosis result. For analyzing healthcare data, major steps of data mining approaches like preprocess data, replace missing values, feature selection, machine learning and make decision are applied on train dataset and ready to classify the test dataset. The system architecture is shown in Figure 1.

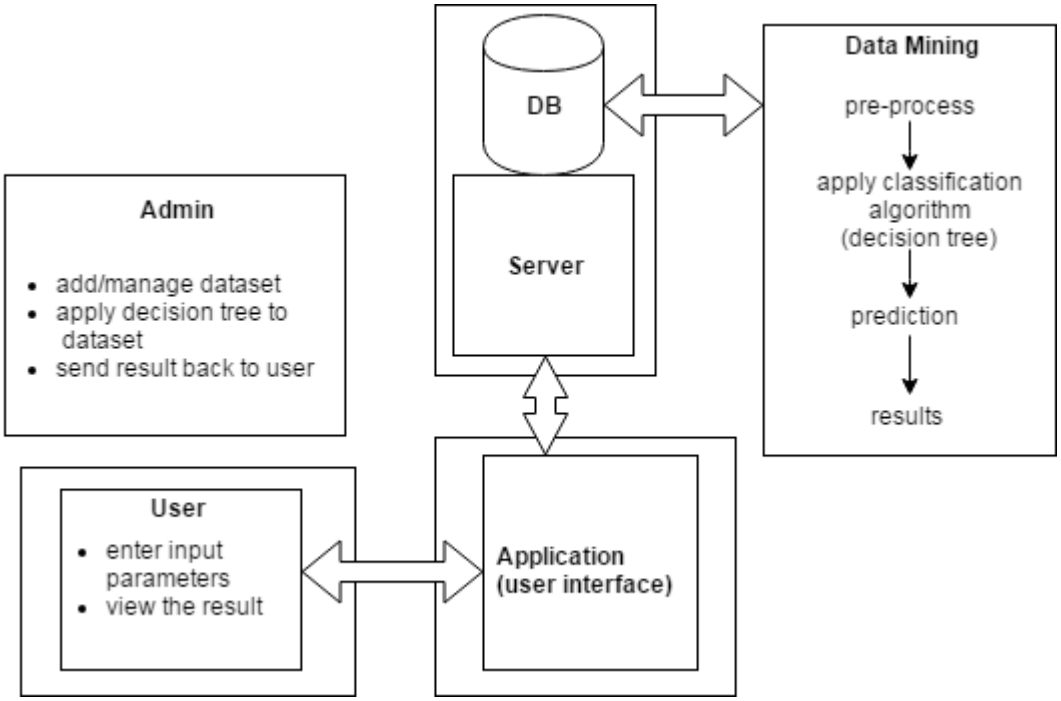

Figure 1: System Architecture

IV. CONCLUSION

An expert system is proposed for predicting the diseases like diabetes using data mining classification technique. The system gives benefit to the doctors, physicians, medical students and patients to make decision regarding the diagnosis of the diseases. The system uses C4.5 algorithm and uses the input features to give a decision as a result.Given the use of EHRs to form a decision tree the system thus encourages the uses of EHRs, so that the physicians have a structured representation of patient information,that can be easily available for them to use. Through the visits the results will be stored and a record will be maintained. Thus, our system will enhance the usage of EHR to store data as well as to predict the disease accurately and efficiently.

V. REFERENCES

[1] R. Fang, S. Pouyanfar, Y. Yang, S. Chen and S. Iyengar, "Computational health informatics in the big data age: a survey," ACM Comput. Surv., New York, Vol. 49, pp. 12-47, June 2016.

[2] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," J. of King Saud Uni. Comput. and Inform. Sci., ELSEVIER, Vol. 24, pp. 27-40, 2012.

[3] Purushottam, K.Saxena and R. Sharma, "Efficient Heart Disease Prediction System," Proced. Comput. Sci., ELSEVIER, Vol. 85, pp. 962 – 969, 2016.

[4] P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," Int. Res. J. of Eng. and Tech. IRJET, Vol. 02, pp. 1039-1043, June-2015.

[5] P. Bhandari, S. Yadav, S. Mote and D.Rankhambe, "Predictive system for medical diagnosis with expertise analysis," Int. J. of Eng. Sci. and Comput., IJESC, Vol. 6, pp. 4652-4656, April 2016.

[6] A. Iyer, S. Jeyalatha and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," Int. J. of Data M. & Know. Manag. Process, IJDKP, United Arab Emirates, vol. 5, pp. 1-14, January 2015.

[7] A. Naik and L. Samant. "Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer , Tanagra ,Orange and Knime," Int. Con. on Computa. Mod. and Sec., ELSEVIER, Vol. 85, pp. 662-668, 2016.

[8] N. Long, P. Meesad and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," Expert Syst. with App., ELSEVIER, Vol. 42, pp. 8221–8231, 2015.

[9] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana and A. Mendonca, "Data mining methods in the prediction of Dementia: A realdata comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," Maroco et al. BMC, Vol. 4, pp. 299-313, 2011.

[10] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.