AN EFFICIENT CLASSIFICATION ALGORITHM FOR THE PREDICTION OF DISEASES USING DATA-MINING

Ekta Rani[#1], Gaurav Sharma[*2]

[#1]Student& CSE & Indus International University Una ,(India)

*[*2] Assistance Professor & CSE & Indus International University Una, (India)*

ABSTRACT

Different disease prediction defines with the task of that disease prediction which has to be cured at an utmost importance. We need to have proper knowledge for the diseases we tend to see in our daily life. The insulin dependent diabetic disease is more dangerous than that of the non-dependent insulin diabetic patient as it causes early deaths of the patients suffering from the disease. Harm to the small blood yacht of the eye is also frequent that could escort to sightlessness which is due to retinopathy. Kidney malfunction of patients, exaggerated by diseases, is moderately common. The generally risk of bereavement among people pretentious with diseases like diabetes, heart patients is twofold the threat of their peers without other diseases. Categorized by missing insulin in body and entail insulin not avoidable with current knowledge, Symptoms contain thirst, starvation, heaviness, eye predicament and weakness. Different datasets were taken by using various classifiers including J48 and Naïve Bayes in the WEKA tool. Overall Naïve Bayes gives the best accuracy in terms of time taken and no. of correct instances being taken. Full-featured analysis gave Naïve Bayes as the best classifier while reduced featured analysis gave another result. In this paper, future work comprises of a proposed model where clustering algorithms are to be applied on the same dataset that gives better accuracy in less time as it will provide better results than previous work done.

Keywords: Classification, Disease, Data mining, Health care.

I. INTRODUCTION

Data mining techniques has been used intensively and expansively by many organizations and can deeply promote all parties concerned in the healthcare production. Several factors have aggravated the use of data mining applications in healthcare. There are plenty of inhabited tackle to predict aging infection in progress. Much illness correlated to one disease that arises all through the heal of one infection. The alleviate can have an effect on the tolerant body continuing the side possessions of the one infection problem for the most part right through parenthood that crosses over 40 years of period are the fatalities of the illness. The current work is focused to grasp the feature reliable for the Disease imitation.

Seeing all the disease prediction early than the disease in order to become more adverse, it can be treated with ease. The intention of this research exertion is:

a) To employ patient's record, health data, and file for determining and identifying illnesses, and afford resolution sustain to therapeutic proficient

b) This can help in detecting early onset of the disease, identification of disease stages and treatment plans.

c) To work with large number of features and attributes in the dataset, and identify the significance of some features over others.

The quality of accent recognition is also one of the resolutions that can be used to recognize the disease. The Data Mining algorithms also award with leading solutions to be familiar with the factors responsible for the derivation of illness discovery. There are several features that are accountable for the grounds for such infection incidence. Depression in addition is solitary of the input factors of the Disease. The changes take position in the body is proficient to be envisaging as peripheral factors like modify in the voice and changes in shade of skin, shade of hair etc. All the citizens of similar infection don't require having indistinguishable symptoms. The modish indicator may put together up or transform with illustration as the Disease improvement. Groups will familiarity about both cruise and non cruise indicator.

II. DIFFERENT ALGORITHMS CLASSIFIED

1. ZeroR Algorithm:

It is the simplest classification method which relies on the target and ignores all predictors. ZeroR simply predicts the majority category simply predicts the majority class. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods

2. J48 Algorithm:

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

3. Naïve bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes classifier assumes that the effect of the value of a predictor on a given class is independent of the values of other predictors.

4. OneR ALGORITHM

The idea of the OneR (one-attribute-rule) algorithm is to find the one attribute to use that makes fewest prediction errors. It turns out that very simple association rules, involving just one attribute in the condition part, often work disgustingly well in practice with real-world data.

III. METHODOLOGY

1. Clustering

Clustering is an unsubstantiated machine learning move toward, but can it be used to get better the precision of administer machine learning algorithms as well by huddling the data summit into similar groups and using these cluster labels as self-determining variables in the supervised machine learning algorithm.

2. K-Nearest Neighbor Classifiers (KNN)

The term data mining is used mostly by statisticians, data analysts and the management information systems (MIS). The difference between data mining and knowledge discovery is that the latter is the application of different intelligent algorithms to extract patterns from the data, whereas the knowledge discovery is the overall process that is involved in discovering knowledge from data. Methods used before computers were introduced into health care use manual analysis to find patterns or extract knowledge from the database. Let us take any field like banking, mechanic, healthcare, and marketing; there will always be a data analyst to work with the data and analyzing the final results. The analyst acts like an interface between the data and knowledge. We can, using machine intelligence to assist the analyst to produce similar results or knowledge from the data.

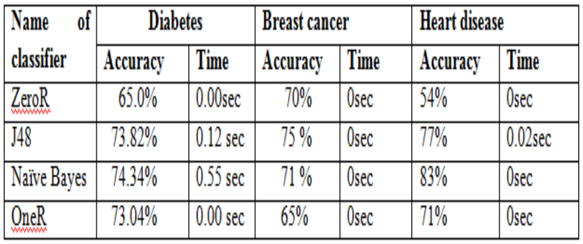
IV. REVIEW SURVEY

Due to the vast amount of data that was being created humans invented algorithms that produce results once a query is supplied. Although these tools perform very well, they can be used to perform only routine tasks. This has led to the creation of machine intelligence algorithms that can perform tasks supplied by humans and make decisions without human supervision. From the evolution of machine intelligence came data mining. In data mining, algorithms seek out patterns and rules within the data from which sets of rules are derived.

According to the World Health Organization (WHO), India has ranked 154 out of 195 countries in the field of health care systems. A study shows that India had fewer practicing physicians and limited care beds per one 150 people than the median of some countries in the Organization for Economic Cooperation and Development (OECD). Different diseases such as diabetes, heart disease and breast cancer have become most common disease.

Data mining technology provides customer oriented approach towards new and hidden patterns in data, from which the knowledge is being generated, the knowledge that can help in providing of medical and other services to the patients. With the future development of information communication technologies, data mining will achieve its full potential in the discovery of knowledge veiled in the medical data. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires, and conditions of the patients and to make adequate and optimal decisions about their treatments.

V. COMPARISION OF DIIFERENT TOOLS WITH VARIOUS DISEASES



Name of classifier	Diabetes		Breast cancer		Heart disease	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
ZeroR	65.0%	0.00sec	70%	0sec	54%	0sec
J48	73.82%	0.12 sec	75 %	0sec	77%	0.02sec
Naïve Bayes	74.34%	0.55 sec	71 %	0sec	83%	0sec
OneR	73.04%	0.00 sec	65%	0sec	71%	0sec

Table 1. Comparison in Accuracy and Time taken for different classifiers on various diseases.

VI. EXPERIMENTAL RESULTS

Classification is one of the most popular techniques in data mining. Comparisons of algorithms based on their accuracy, learning time and error rate according to disease. From the result obtained, the frequently used classification techniques ZeroR,J48,Naïve Bayes and OneR are analyzed, on the medical dataset to find the best classifier according to diseases. The performance indicating accuracy, specificity, precision, error rate are calculated for the various dataset given. With proper data processing techniques, the accuracy of the classifier is calculated. The result shows that the performance of Naïve Bayes classifier/technique is significantly superior to the other three techniques for the classification of diabetes data. For Breast cancer disease, J48 classifier has given better accuracy as compared to the other three techniques. For Heart disease, Naive Bayes classifier gives better accuracy as compared to other three techniques.

VII. FUTURE SCOPE

Medical related information's are volumetric in nature and it could be derived from different birthplaces which are not entirely applicable in a feature. In this work, we have performed a literature survey on various papers. In future, we are planning to propose an effective disease prediction system to predict the various diseases with better accuracy using different data mining classification techniques such as Naïve Bayes, J48, etc. To enhance the model’s accuracy for the future work, there are additional measurements need to be added in the dataset such as the nutrition system and the exercise for the patients.

REFERENCES

1. M.Inbavalli and G. Tholkappia Arasu(2016): Multi-Attribute Density Estimation Based Location Selection Approach in Multi-Agent Disease Prediction Model for Decision Support System Using Diagnosis Pattern and Data Mining.
2. TARIG MOHAMED AHMED(2016): Using data mining to develop model for classifying diabetic patient control level based on historical medical records.
3. Tarigoppula.V.S.Sriram, M. Venkateswara Rao(2016): A Comparison And Prediction Analysis For The Diagnosis Of Parkinson Disease Using Data Mining Techniques On Voice Datasets.
4. G.V. Satya Narayana, DSVGK Kaladhar (2016): Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms.
5. Mrs. R. Vidhu, Mrs. S. Kiruthika (2016): A New Feature Selection Method for Oral Cancer Using Data Mining Techniques.
6. Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen, “ Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method”, IJSETR Volume 2, Issue 10, October 2013
7. M.Akhil jabbar, B.L Deekshatulu and Priti Chandra, “ Classification of Heart Disease using K-Nearest Neighbor and Genetic Algorithm”, International Conference on (CIMTA) 2013.
8. Daljeet Kaur A and Aman Paul , “ Performance Analysis of Different Data mining Techniques over Heart Disease dataset”, International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 - 5161
9. K.Vembandasamy R.Sasipriya and E.Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm”, IJISET Vol. 2 Issue 9,
10. Ruchika Rana, Jyoti Pruthi, “Heart Disease Prediction using Naive Bayes Classification in Data Mining”, IJSRD - International Journal for Scientific Research & Development|
11. Prerana T H M, Shivaprakash N C and Swetha N, “ Prediction of Heart Disease Using Machine Learning Algorithms- Naive Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS”, International Journal of Science and Engineering Vol-3, Number 2-2015
12. Ms Manaswini Pradhan(2014): Data Mining and Health Care: Techniques of Application.
13. Geetha R. R., Sivagami, G.: Parkinson Disease Classification using Data Mining Algorithms.International Journal of Computer Applications.
14. A Proficient Heart Disease Prediction Method Using Different Data Mining Tools(K.Manimekalai(2016)

15. Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery(Heling Jiang, An Yang, Fengyun Yan and Hong Miao)

16. Basagic R., Krupic D., Suzic B., ―Automatic Text Summarization, Information Search and Retrieval‖, WS 2009, Institute for Information Systems and Computer Media, Graz University of Technology, Graz, 2009

17. Arora R., and Ravindran B., ―Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization‖, Proc. Eighth IEEE International Conference on Data Mining (ICDM 2008), IEEE Press, pp. 713-718, 2008.

18. USING DATA MINING TO DEVELOP MODEL FOR CLASSIFYING DIABETIC PATIENT CONTROL LEVEL BASED ON HISTORICAL MEDICAL RECORDS (Tarig Mohamed)