A Study of Different Knowledge Patterns in Diabetic Using Data Mining Techniques

Malpe Kalpana Devidas[1], Dr. Neeraj Sharma[2]

[1]Research Scholar, Department of Computer Science Engineering, of Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.
[2]Research Supervisor, Department of Computer Science Engineering, of Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India

Abstract

This research work was conducted on the design and implementation of a diabetes prediction system, a case study of Fudawa Health Centre. This research will help in automating prediction of diabetes even before clinicians arrived. The current process of carrying this activity is manually which tends not to analyzing data flexible for the doctors, and transmission of information is not transparent. The system was design using Java Programming Language, Weka Tool, and MySQL (Microsoft Structured Query Language) as the back end and a strategic approach to analyse the existing system was taking in order to meets the demands of this system and solve the problems of the existing system by implementing the naïve beyes classifier. The implementation of this new system will help to reduce the stressful process, doctors' face during prediction of diabetes, the result of the experiment shows that the proposed system has a better prediction in terms of accuracy.

***Keywords**: Diabetes, Data Mining, Weka Tool, Diagnosis, Prediction, Naïve Bayes Classifier, Technique.*

1. INTRODUCTION

The term "diabetes" is a disease that occurs when the blood glucose in the body, also called blood sugar, is too high. Blood glucose is the main source of energy and comes from the food we eat. According to doctors, diabetes occurs when a gland known as pancreas does not release a hormone called insulin in sufficient quantity. Insulin is a hormone that carries sugar from the bloodstream to various cells to be used as energy. Lack of insulin disrupts the body's natural ability to produce and use insulin accurately. As a result of this, high levels of glucose are released in urine. In the long -term, diabetes when not properly managed can lead to organ failure, cardiovascular diseases and disrupts other functions of the body. WHO (world Health organization) has listed diabetes as one of the four major NCDs (non communicable diseases) in the world today (World Health Day, 2016). Statistics released by WHO are alarming. Diabetes, as mentioned earlier, can lead to other major complicated cardiovascular diseases. According to WHO, 3.7 million fatalities occurs before the age of 70, and this high mortality rate is attributed to diabetes and cardiovascular diseases. Uncontrolled blood glucose level is the major factor behind diabetes. Diabetes is a health problem in Nigeria. It is the most common chronic diseases across all population and age groups. According to a report by WHO 2016, in Nigeria, 13% of deaths are related to diabetes. This report also shows that the harmful effects of this disease are increasing at a rapid speed. Diabetes is divided into two distinct types; type 1 diabetes enforces the need for artificially infusing insulin through medicines or by injections and type 2 diabetes, pancreas create insulin, but it is not effectively used by the body. The majority of people with diabetes are affected by type 2 diabetes. Diabetes was a common problem among adult's specifically middle-aged people but due to changing lifestyles diabetes affects children too. Type 1 diabetes is unpreventable because of the various external environmental stimulants which result in the destruction of body's insulin producing cells. However, changing lifestyle to achieve the required body weight and obtain the physical activities can help to prevent type 2 diabetes to enlarge. Diabetes is a chronic health problem with devastating, yet preventable consequences. It is characterized by high blood glucose levels resulting from defects in insulin production, insulin action, or both.1,2 Globally, rates of diabetes were 15.1 million in 2003 the number of people with diabetes worldwide is projected to increase to 36.6 million by 2030. Of these, 90-95% of these cases are adults with type 2 diabetes. Diabetes impacts men and women proportionately; there are over 12 million men with diabetes and 11.5 women with diabetes. Therefore, predicting diabetes manually sometimes seems not to be objective and it consumes a lot of time and cost. Diabetes treatment focuses on controlling blood sugar levels to prevent various symptoms and complications through diet and exercise.

Data mining is a relatively new concept used for retrieving information from a large set of data. Mining means using available data and processing it in such a way that it is useful for decision-making. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining thus has evolved based on human needs which can help humans in identifying relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Eapen, 2004). Data mining helps in pattern identification and categorizing data records by conducting cluster analysis, identification of odd records also called detecting anomalies and association rule mining or dependencies Frawley and Piatetsky (1996) describes data mining as the process of extracting implicit and previously undisclosed important information about data sets that can be used for effective decision-making. The process is termed as Knowledge Discovery in Database, Such discovered knowledge can be very useful in many areas of sciences, and health care is no different having a Knowledge Discovery in Database would help in predicting trends of many kinds of diseases and illness. So doctors, rather than depending on their own knowledge and experience, can use data mining and specifically Knowledge Discovery in Database to predict or to forecast and to predict trends that would lead to better diagnoses, reduce cost and save person-hours for the organization. Data mining is placed as a statistical interface, data mining lies in the interface of statistics, database technology, recognizing patterns, machine readable data, and intelligent expert systems (Obenshain, 2004). The prime objective of data mining is to extract information from data sources and alter it into a comprehensible assembly of information for more uses (Data Mining Curriculum, 2014). Data mining is a process that is used to locate correlations between data and form pattern of relationships among cluster fields in the enormous interactive database (Extract -Nature Biotechnology 18, 2000). With data mining techniques, doctors around the world will be able to predict illnesses effectively and be better equipped to manage potential high-risk candidates. Such analysis and predictions become critical if the objective is to provide relief to millions around the world. This research addressed the main challenging issue confronting the health care industry, which is lack of quality service at minimal cost implying from diagnosing to the predicting patients correctly(chaurasia and pal, 2013) or sometimes even understand the complications that may result from the diseases(srinavas et al 2010). This issue can lead to unfortunate clinical decision that can result in devastating consequences that are unacceptable (Apte and Dangare 2012). The availability of patients medical data has derived the need for clinicians and patients for alternative computer-based assessment tool that can assist in decision –making (soni et al 2011) for example, the physicians can compare analytical information of numerous patients with the matching condition and physicians can equally confirm their results with the conformity of other part of the country (srinavas et al 2010). This research applies naïve bayes classification technique on the dataset obtained from Fudawa health care centre, jos plateau state Nigeria. The dataset was preprocessed to remove noise and null fields using weka tool and it was further divided into training dataset and test dataset. The following parameters were used for detecting and classifying the diabetes into positive and negative class, the parameters are: age, insulin, smoke cigarette, agefirstsmoked, where survey was taking. Support vector machine is a method that uses the concept of computer science and statistics to analyze data and support pattern recognition, which are then used in classification which makes prediction based on the set of accepted input, in which every given input, there are two feasible classes that form the inputs (Madzov et al, 2009). Support vector machine is designed based on the principle of structural risk minimization principle with the basic idea of finding hypothesis with lowest error. However, the drawback of this learner is that its computation is highly expensive thereby running slow on high dataset and it does not offer probability estimate directly. It also does not perform very well on large dataset because higher training time is required. Naïve bayes classification is simple and particularly suited when the dimensionality of input is high. Despite its simplicity, it can outperform more sophisticated classification method. This classifier works on the assumptions that: the data must be categorical in nature, occurrences of attributes independent and predict accurately on high volume dataset.

2. RELATED CONCEPTS

The clinical presentation of diabetes in a patient is the symptomatic features presented by the patients. This feature is an indication of the disease cause and has direct impact in guiding clinicians about the decision to take. In case of classifying positive and negative diabetes, the following parameters were considered: age, insulin, smoke, age first smoked and where the survey was taking.

- Positive class label (P): patients can be confirmed to have positive diabetes, when the patient has one or more symptoms of diabetes and has also been confirm by laboratories. Since these features are the most occurring symptoms in a patient with diabetes.

- Negative class label (N): patients may have some of the parameters (symptoms) of positive diabetes, but after several trying of diagnostic test confirm, the diabetes is undetectable. This means that the existence of the signs may be as a result of the other concomitant disease.

Naïve Bayes Classifier

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. It works on the assumptions that: classifying categorical data, occurrences of an event independent and predict accurately on high dataset.

3. ALGORITHM

Let A be a training dataset. Suppose each tuple is represented by n-dimensional attribute vector X=(X_1, X_2…..Xn) which represents 'n' measurement on the tuple from 'n' B_2……Bn. Suppose that there are N classes: C_1, C_2……Cn. Given a tuple X, the classifier will predict that X belongs to the class having the highest probability (P) condition on Y that is, Y belongs to class C_1 if and only if

$$P(Ck|X) > P(Ci|X) \quad \text{for } 1 \le i \le n \quad i \neq k$$

The algorithm will maximize

$$P(C_1|X)=P(X|\ C_1)P(C_1)/\ P(X) \text{ and } P(C_2|X)=P(X|\ C_2)P(C_2)/\ P(X)$$

Hence,

$$P(C_1|X) > P(C_2|X) \quad \text{if and only if}$$

$$P(X|\ C_1)\ P(C_1) > P(X|\ C_2)\ P(C_2), \text{ since } P(X) \text{ is the same in both cases.}$$

Given the dataset with many attributes, it will be expensive to compute $P(X|\ C_1)$. Therefore, we assume that the values of the attributes are conditional independent. Thus

$$P(X|\ C_1) = \prod P(X_1|\ C_1). \text{ Therefore,}$$

$$P(X|\ C_1) = P(X_1|\ C_1)^*\ P(X_2|\ C_2)^*\ P(X_3|\ C_3)^*\text{……..}^*\ P(Xn|\ Cn)$$

The advantages of the Naïve bayes classifier are as follows:

- Ability to approximate probabilities for a class of any given instances and also it relative simplicity.

- It requires less model training time.

- It performs well in the present of irrelevant features

4. THE KNOWLEDGE DISCOVERY IN DATABASES

Knowledge Discovery in Databases (KDD) is the procedure used to attain important and useful knowledge from a large collection of previously collected data. The process involves selecting, preparing and cleansing the data from unnecessary information. Any previously available information is incorporated into the data sets. Data interpretations are conducted to achieve precise outcomes from available results as shown in Figure 1

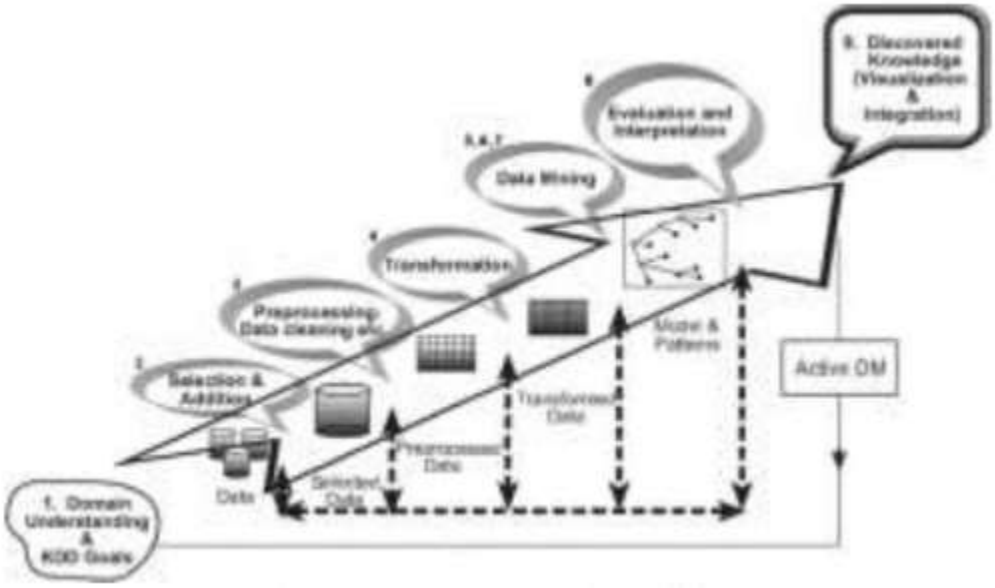

Figure 1: KDD Process (Courtesy Maimon, Rokhah, 2010)

Data Mining Techniques

There are two types of Data Mining task; the predictive model and descriptive model. Both models are explained as follows

Predictive Model

The predictive data-mining model predicts the future outcomes based on past records present in the database or with known answers. Data mining will help figure out the future credit risk of the applicant and predict future credit history of the applicant by using past data. Classification is known as the procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available. The resultant model is focused on analyzing a set of identified classes. Regression is a mathematical and statistical tool used widely in using numeric values for forecasting time series analysis. Prediction as the term implies means correctly envisioning the future using logical computation of available data.

Descriptive Model

This model is to discover patterns in the data and understand the relationships between the data attributes. Descriptive Model represents the main feature of the data, and summarizes. The collected knowledge can be used to develop marketing programs for targeting audience. Clustering examines data objects without referring to an identified class label. Summarization is to categorize the distinctive properties of data and point out if the data values are to be categorized as noise or outliers. This research classifies data mining as shown in the Figure 2.

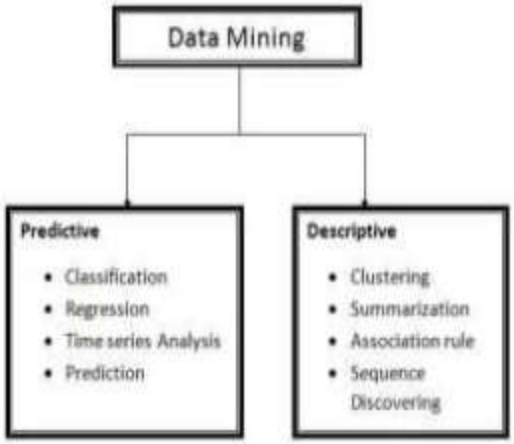

Figure 2: Data Mining Tasks

Using The Naïve Bayes In The Study

The naïve bayes method discussed in section 2 works as follows on our problem. Using the naïve bayes formula,

$$P(H|E) = P(E|H)\,P(H)/\,P(E)$$

Where H is the class

P (H|E) is a posterior probability of class given predictor

P (H) is the past (prior) probability of class

P (|H) is the probability of predictor given class

P (E) past (prior) probability of the predictor Class diabetes is calculated as:

- Positive class (p): patients may have diabetes if the probability of selected features point out that the probability of positive class is greater than negative class.

P (positive| patient) = P (patient| P) P (P)/ P (patient)

- Negative class (N): patients may not have diabetes if the probability of selected features point out that the probability of negative class is greater than positive class. P (Negative| patient) = P (patient| N) P (N)/ P (patient)

5. CONCLUSION

An Application using a data mining algorithm of classes' comparison has been developed to predict the occurrence of or recurrence of diabetes risks. In addition, the result of the application shows that the predictions system is capable of predicting diabetes effectively, efficiently and most importantly, timely. That means the application is capable of helping a physician in making decisions towards patient health risks. It generates results that make it closer to the real life situations. That makes the data mining more helpful in the health sector, which means that it is necessary for knowledge discovery in the healthcare's sector Much more than huge savings in costs in terms of medical expenses, loss of duty time and usage of critical medical facilities, The naïve bayes classifier based system is very useful for diagnosis of diabetes. The system can perform good prediction with less error and this technique could be an important tool for supplementing the medical doctors in performing expert diagnosis. In this method the efficiency of forecasting was found to be around 95% This application would be a tremendous asset for doctors who can have structured specific and invaluable information about their patients / others so that they can ensure that their diagnosis or inferences are correct and professional. Finally, the huge appreciations received from the doctors on having such software prove that in a place like, where diseases are on the rise, such applications should be

developed to cover the entire state. The common person stands to benefit from doctors having such a tool so that he/she can be better knowledgeable as far as personal health and wellbeing is concerned.

6. REFERENCES

[1]. Acharya, Rajendra, U and Yu, Wenwei, (2010).Data Mining Techniques in Medical Informatics. The Open Medical Informatics Journal, PMCID: PMC2916206.
[2]. Aflori C., and Craus, M., (May 2007) Grid Implementation of the Apriori algorithm Advances in Engineering Software, 38(5), pp. 295-300. A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.
[3]. Anbarasi M., (2010). 'Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,' International Journal of Engineering Science and Technology, 2(10), 5370- 5376.
[4]. Bronzino, D. Joseph, Medical Devices and Systems, 2006
[5]. Chauraisa V., and Pal, S.,(2013). 'Data Mining Approach to Detect Heart Diseases', International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2, (4), pp 56-66.
[6]. Clifton, Christopher (2010), Encyclopedia Britannica: Definition of Data Mining Retrieved 2016.
[7]. Data Mining Curriculum, ACM SIGKDD, 2006-04-30, retrieved 2016
[8]. Fayyed, Usama., (15 June1999), First Editorial By Editor-InChief, SIGKDD Explorations 1:1, doi: 10.1145/2207243.2207269
[9]. Fayyed, Usama., Piatetsky- Shapiro, Gregory., Smyth, Padhraic., (1996) From Data Mining to Knowledge Discovery in Databases.
[10]. Han J.,and Kamber, M. (2010). Data Mining: Concepts and Techniques, 2nd ed., the Morgan Kaufmann Series.
[11]. Han Jiawei., & Kamber, Micheline. (2001), Data Mining: Concepts and Techniques, pp. 5
[12]. Hastie, Trevor, Tibshirani Robert, Friedman, Jerome. (2009), 'The Elements of Statistical Learning', [13]. HninWintKhaing,(2011). "Data Mining based Fragmentation and Prediction of Medical Data", IEEE.