Data mining Applications in Healthcare Domain

1Dr. Satish Chandra Pandey
1Assistant Professor
1Pratap Universty Jaipur

Abstract - **Data mining is the progression of dig out valuable information from huge data sets over the use of statistics, algorithms, machine learning & Database Management System. The huge amounts of data produced by healthcare organizations are too complex and capacious to be processed and analyzed by outmoded methods. It provides the approach and knowledge to transform these mounds of data into valuable statistics for decision making. In the present research paper decision tree data mining technique has been applied in healthcare domain to identify the risk factors associated with the onset of diabetes. It has been found that out of the five attributes (Age ,Gender, Body Masss Index,Excersie per week, Waist Hip Ratio)Age is the most effective and gender is the learnt effective attribute on the onset of diabetics. It is also concluded that as the age of the person increases the probability of getting affected by diabetics also increases. The paper also highlights the limitations of data mining and discusses some future directions.**

keywords **- Attributes, Body Mass Index, Waist Hip Ratio, Classification, Decision Trees**

I. INTRODUCTION:

Data mining is comparatively newly advanced approach and technology which appeared into reputation in 1997[1].It aims to identify valid novel, potentially useful and understandable correlations and patterns in data [2] by combining through copious datasets to sniff out patterns that are too subtle or complex for humans to detect [3] .

Data mining has been used extensively in healthcare domain because of four reasons.

Firstly, the tremendous amounts of data generated by healthcare transactions are extremely complex and voluminous and can not be analyzed by traditional methods alone. Data mining provides a powerful methodology and technology to transform these mounds of data into useful informations for decision making [4].

Secondly, the existence of medical insurance fraud and abuse which have lead many healthcare insurers to attempt to reduce their loses by using Data mining tools to help them find and track offenders [5]. Recently there has been reports of successful Data mining applications in healthcare “fraud and abuse” detection [6].Thirdly, Data mining applications can benefit healthcare providers such as hospitals, clinics, and physicians and patients by identifying effective treatments and best practices [7, 8].Fourthly, the use of Data mining applications in healthcare is the realization that Data mining can generate information which are very useful to all parties involved in the healthcare industry.

In this paper we have applied decision tree [9] Data mining technique in healthcare domain to identify certain variables which are associated with the onset of diabetes. The purpose of this data mining application is to identify high risk individuals so that appropriate information can be communicated to patients.

II. RELATED WORK:

A literature survey reveals many results on diabetes in different countries. The diabetic data warehouse was molded by a huge cohesive healthcare system in the New Orleans area with 30383 diabetic patients. The classification and regression tree approach was applied to analyze these data sets [10].

Indrani Bose [11] in her research paper has explained the different steps involved in mining diabetic’s data and to show using case studies how Data mining has been carried out for recognition & finding of diabetics in Hongkong, USA, Poland & Singapore.

Abdullah et al [12] have shown in their article the predictive analysis of diabetic treatment using a regression based Data mining technique. The oracle data miner (ODM) was hired as a software mining instrument for expecting modes of tracking diabetes. The support vector machine algorithm was used for investigational analysis.

Monali et al [13] have mentioned in their article about analysis of the uniqueness of medical Data mining, overview of healthcare decision support systems currently used in medicine, identification and selection of the most common Data mining algorithms implemented in the modern HDHSS and Companion between different algorithms in Data mining.

Ravi Sankall et al[14] ,this study attempts to use Data mining method to analyze the data bank of diabetic’s diseases and diagnose the diabetes diseases. This homework contains the execution of FCM and SVMM and testing it on a set of medical data responded to diabetic’s diagnosis problem.

III. METHODS & MTEHODOLOGY:

The data mining process for identifying the most effective attribute is divided into seven steps as under [15].which is shown in following **Figure-1.**

(i) Objective Selection:

The business objectives are decided prior to the Data mining process. These business objectives may depend on the inferences or results which will be drawn by Data mining processes. This can be achieved by joint effort of the data analyst who can translate the objectives identified by the analyst into a well defined Data mining problem.

(ii) Data preparations:

After significant the intentions the data is organized for the mining progression. This process contains of the following sub steps:

- **Data Selection:**

  There are two parts to select data for Data mining. The leading fragment, finding data tends to be more automated in nature than the subsequent part. The second part identifying data requires significant input by a domain expert for the data. A domain expert is one who is intimately familiar with the business purposes and aspects or domain of the data to be examined.

- **Data Cleaning:**

  Data cleaning is the process of ensuring that for Data mining purposes the data is uniform in terms of key and attributes usage. Data cleaning is separate data from data enrichment and data because data cleaning attempts to correct misused or in correct attributes in existing data. Data enrichment by contrast adds new attributes to existing data while data transformation changes to form or structure of attributes in existing data to meet specific Data mining requirements.

- **Data Enrichment:**

  Data enhancement is the progression of adding innovative characteristics such as designed fields or data from exterior sources to standing data. It can include merging internal data with exterior data, gained from either one different departments or companies or vendors that sell standardized industry relevant data.

- **Data Transformation:**

  Data transformation is the process of changing the form or structure of existing attributes.

(iii) Searching of a Database:

This segment is to select & inspect significant data sets of a Data mining database in order to determine their feasibility to resolve the problem. Penetrating the database is a time overwhelming process and needs a decent user interface and computer system with good processing speed.

(iv) Creation of Data mining Model:

This stage is to select variables to act as analysts. New-fangled variables are also built dependent upon the prevailing variables along with defining the range of variables in order to support inexact information.

(v) Building of Data mining Model:

This phase is to generate several Data mining models & to select the best of these models. Construction a Data mining model is an iterative process. The Data mining prototypical which we select can be an artificial neural network, a decision tree, or an association rule model.

(vi) Evaluation of Data mining Model:

This stage is to estimate the correctness of the designated Data mining model. In Data mining the assessing restriction is data accuracy in order to test the operational of the model. This is for the reason that the material generated in the simulated environment varies from the external environment.

(vii) Deployment of Data mining Model:

This phase is to organize the constructed & the assessed Data mining model in the external occupied environment. A observing system should monitor the operational of the model and produce reports about its presentation. The fact in the report helps to increase the performance of selected Data mining model.

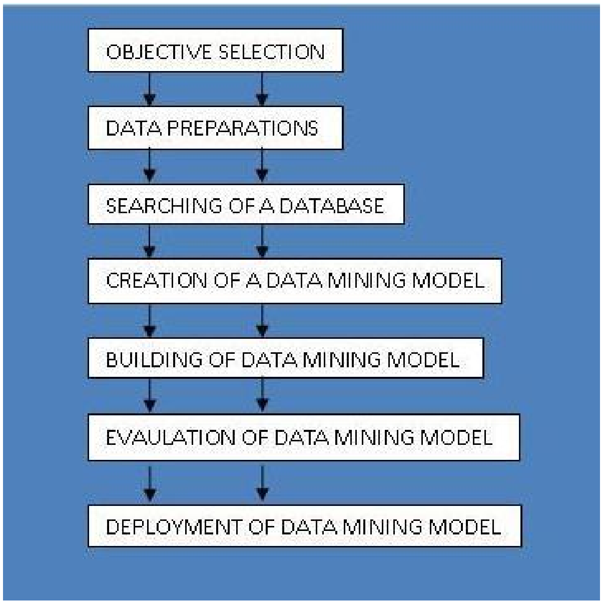
OBJECTIVE SELECTION
DATA PREPARATIONS
SEARCHING OF A DATABASE
CREATION OF A DATA MINING MODEL
BUILDING OF DATA MINING MODEL
EVAULATION OF DATA MINING MODEL
DEPLOYMENT OF DATA MINING MODEL

Fig-1 Steps of Data mining Process

IV. EXPERIMENTAL STUDY:

To discuss a data mining application in healthcare we take Healthcare data from various health care institutions (Hospitals, Clinics, and dispensaries etc.) situated in Tarai region (Basti, Siddharthnagar, Sant Kabir Nagar, Distts.)With the help of questionnaire. The No of patients examined was 1010 (including diabetic patients).

Here we are interested to find out how certain variables are associated with the onset of diabetes. Thus the aim of this Data mining application is to identify the high risk factors associated with the onset of diabetes so that appropriate information can be communicated to patients. The data set contains nine variables of interest: Gender, Age, Body Mass Index (BMI), Waist Hip Ratio (WHR), Height, Smoking status Religion, Cast, & the number of times a patient exercises per weak out of which only four attributes (Age, BMI, WHR & Exercise per weak) are most effective and have been taken into consideration.

The data set comprises of 150 or 14.86% of positive diabetic cases and 860 or 85.15% of negative non diabetic cases. The total no of patients examined was 1010.

After reviewing the work of Breault et al (2002) on the Data mining of a diabetic data warehouse. We have decided that decision trees is an appropriate data mining technique to use to find out how certain variables are associated with the onset of diabetes.

These all steps are showing in **Figure-2**.

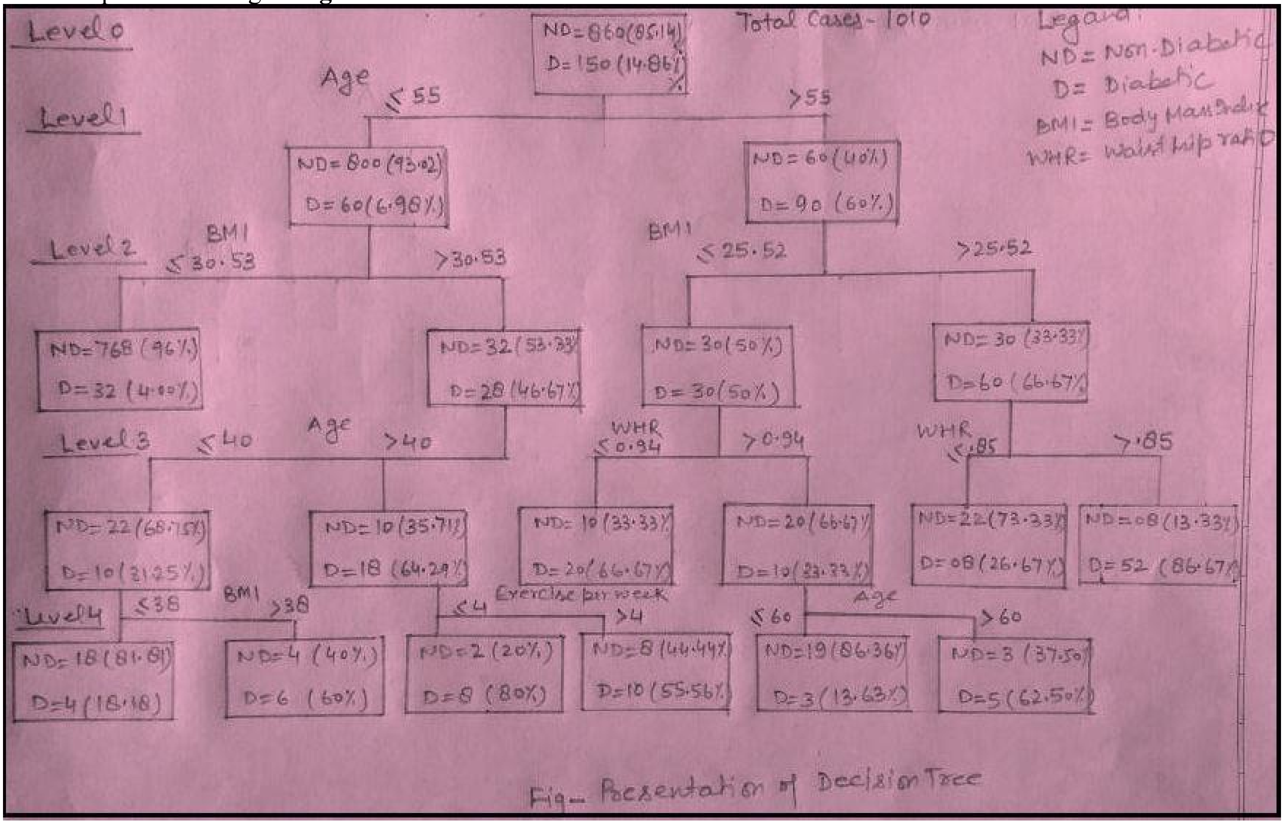

Fig-2 Presentation of the decision tree

V. RESULTS AND DISCUSSIONS:

Results are summarized as shown in the Fig-2 Age, Body Mass Index (BMI), Waist Hip Ratio (WHR) and the No of exercise per week are significantly associated with the onset of diabetics. Fig-2 gives a visualization of decision tree and facilitates interpretation of the results. The generated decision tree can be interpreted as follows .The result shows that the age is the most important factor associated with the onset of diabetes (level 1 Fig-2) with individuals older than age 55 showing significant higher risk of diabetes, compared with their counterparts.

At the next level 2 (Fig-2) the Body Mass Index (BMI) is the next most important factor associated with the onset of diabetes. In particular individuals younger than 55 with a body mass index (BMI) of less than 30.53 have a very low risk of diabetes (probability of only 4.00% in the cohort).Moreover, increasing levels of BMI is associated with increasing risk of diabetes .For individuals older than 55 with a BMI greater than 25.52, the risk is 66.67% .The Waist Hip Ratio (WHR) is the next most important factor (level 3 Fig-2) with increasing WHR associated with an increased risk of diabetes. For example, the highest risk individuals in the data base are those above 55 years old with BMI of more than 25.52 and WHR above 0.85 – their risk probability is 86.67 in this cohort. The exercise per week (level 4 Fig-2) is the next most important factor associated with the onset of diabetes .The greater the no of exercise per week the lower is the no of diabetic cases.

In a similar manner the remaining nodes in the decision tree can be interpreted. Nodes further down in the decision tree are less important in view of their smaller sample sizes and also because of their more restrictive sub-setting at lower levels. It is also concluded that gender & religion is not effective attributes associated with the onset of diabetes. Thus decision tree can help health organizations to identify the high risk individuals and appropriate messages can be communicated to them. For example, healthcare organizations can launch a health promotions campaign to educate people that large BMI and WHR are risk factors associated with the onset of diabetes. It can also scan through its patient data bases to identify individuals for further counseling or medical checkups.

VI. CONCLUSIONS:

It is also concluded that larger is the no of exercise per week ,the smaller is the chance for being diabetic. Moreover the decision rules also illustrate that diabetes is independent of gender, religion, cast. The decision rules show that age is the prominent risk factor with the onset of diabetes. The larger the age of a person, the larger is the probability of being diabetic.

REFERENCES:

[1] Trybula,W.J.,Data mining and knowledge discovery .Annual Review of information Science and technology 32,197-229.
[2] Chung,H.M & Gray.P Data mining.Journal of Management Information Systems 16(1),pp-11-16,(1999).
[3] Kreuze D,"Debugging hospitals technology review", 104(2), 32. . (2001).
[4] Biafore S., Predictive solutions bring more power to decision makes, health management technology, 20(10), pp.12-14,(1999).
[5] Chirsty T,"Analtytical tools help health firms, Fig-ht, fraud ,insurance & Technology", 22(3), pp.22-26, (1997).
[6] Millay A. "healthcare Data mining, Health management technology"21 (8), pp. 44-47, (2000).Ingram M.,Internet privacy threatened following terrorist attacks on VS,URL :http://www.wsws.org/articles/2001/Sep 2001/isps24, shtml.(2001).
[7] Gillespie G."There's gold in them that' databases", Health Data Management, 8(11), 40-52. (2000).
[8] Kolar,H,R,"Caring for healthcare, health management technology", (22-14),46-47. (2001).
[9] Data mining and Warehousing by S. Prabhu & N. Venkatesan,New Age International(P) Limited,Publication,New Delhi,pp.23-24
[10] Breault ,Joseph L.,Goodall Colin R.,Fose Peter J., "Data mining a diabetic data warehouse" .Artifical Intelligence in Medicine 26,pp-37-54.(2002).
[11] Bose Indrani, "Data mining in diabetic's diagnosis and detection" (2006).
[12] Abdullah A. Aljumah., Ahamad Mohammad Gulam. Siddiqui Mohammad Khubeb,"Application of Data mining:Diabets health care in young and old patients", Journal of King Saud University Computer & Information Science 25, pp.127-136, (2013).
[13] Dey Monali, Ratanargy Siddarth Swarup.,"Study and Analysis of Data mining algorithms for Healthcare decision support system, IJCSIT, pp. 470-477, Vol-5 No-1, (2014).
[14] Sanakal Ravi,Smt. Jayakumari T.,"Prognosis of diabetes using data mining approach ,Fuzzy C means clustering and support vector machine" ,IJCIT Journal ,Vol-11,No-2,(2014).
[15] Data mining, (BPB publications, B-14, Connaught place, New Delhi-1) pp-5-6.