ISSN:2394-3386

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 4
April 2017

Predictive Risk Modeling of Diabetes Using Data Mining

Inderjeet Kaur,
*Department of Computer Science,
Lokmanya Tilak College of Engineering,
University of Mumbai,
Mumbai, India.*

Aishwarya Deshpande,
*Department of Computer Science,
Lokmanya Tilak College of Engineering,
University of Mumbai,
Mumbai, India.*

Varun Sriram,
*Department of Computer Science,
Lokmanya Tilak College of Engineering,
University of Mumbai,
Mumbai, India.*

Chaitrali Chaudhari
*Department of Computer Science,
Lokmanya Tilak College of Engineering,
University of Mumbai,
Mumbai,India*

ABSTRACT

Data mining is a new technology, which helps organization to process data through various algorithms. Following paper uses the data mining technology to predict if a person is diabetic or not.
Data collected is used for analysis and prediction of the diabetes.

KEY ELEMENTS:

1. ***DATA MINING***
2. ***CLASSIFICATION MODEL***
3. ***DECISION TREE***
4. ***WEKA TOOL***

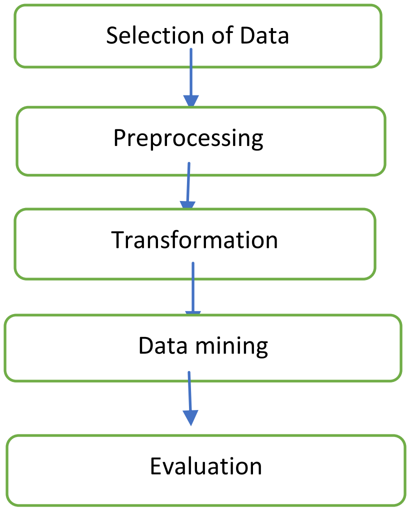
1.INTRODUCTION

With rapid refinement taking place over the globe, advancement in technology is significant. The area of medicine is not far behind. However, the desired expectations have not been met as far as diagnosis of some diseases is concerned. One such disease namely diabetes has been taken into consideration. Diabetes affects millions of people and is a significant long-lasting health problem. However, safeguarding diabetes in control is a tough task as self controlling measures have to be adopted for a significantly large percentage of people across the globe. Diabetes is a condition that causes high blood glucose. It cannot be entirely cured but can thoroughly be managed. There are primarily two types of diabetes, the first one is Type-1(insulin-dependent), and it is treated with systematic insulin injections and a nutritious diet. The second one is Type 2 (insulin resistance). The same is treated by adopting diet changes, adequate amount of exercise and prevention of smoking apart from other minor measures. Regular medication, medicines and injections play a critical role in the management of diabetics.
Early Prediction of diseases can reduce the fatal rate of human. There are very large and enormous data available in hospitals and medical related institutions.

2. Role of Data Mining Techniques in Diabetes prediction

The process of discovering knowledge in data and application of data mining methods refers to the term knowledge discovery in databases(KDD).

OUTLINE STEPS OF KDD PROCESS



Data mining involves a process that transforms the data into information through classifying, merging, sorting, transmitting, retrieving etc data processing can be manual or computer based.

To understand data mining functionality we need to understand basic terms in data mining which includes **1) Association**: The objects in relational databases , or any other information repositories are considered for finding frequent patterns.

2)Classification: Classification is a data mining technique used for systematic placement of grop data.

3)Regression : Used to predict for individuals on the basis of information gained from previous samlple of similar individuals.

4)Cluster Analysis: Clustering is a data mining technique a kind of machine learning technique used to place data elements into related groups without advance knowledge of group defination .

5)Forecasting: Discovering patterns in data that can lead to predictions

6)Outlier Analysis : Based on the above functionality data mining tasks can be classified into two categories
a)Descriptive Mining: To derive patterns like correlation, trends etc. which summarizes the underlying relationship in data.
b)Predictive Mining: Predict the value of a specific attribute based on the value of other attribute.

3.LITERATURE SURVEY

HISTORY:

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history.Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning.

Machine learning is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previouslyhidden trends or patterns within.

4. Data Mining Techniques

Classification analysis

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified. A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

Clustering Analysis

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. A result of this analysis can be used to create customer profiling.

Regression Analysis

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting.

5.PROPOSED SYSTEM

Input:

Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and
Kidney Diseases dataset pre-processed .

Output:

either tested-positive or tested-negative .

Procedure:

1.The dataset is pre-processed using libraries of WEKA tools.
Following operations are performed on the dataset
2.Replace Missing Values with the mean of the attribute split the data into train and test
3.from the data splitted into train and test the algorithm c4.5 predicts if the person id diabetic or not
4.finally the resullts of the test are displayed as tested positive or tested negative on the basis of the inputs given

6.TOOLS AND ALGORITHMS USED

Various data mining tools are available each has its pros and cons. For the analysis of diabetic data classification algorithms are used to find which treatment is effective for patients with different age groups and to find the efficient classification algorithm for the analysis. Weka version 3.6.12 is used to find the efficiency of algorithm. Weka is a graphical user interface which is used for data analysis and predictive modeling written in java. In weka, main interface called Explorer is a component based knowledge flow interface and Experimenter which provides comparison of the performance of different machine learning algorithms .

Decision tree learning:

Type of learning: supervised, concept learning, divide-and-conquer strategy. Strategies for concept learning:

1)Covering: generate a rule, exclude the data covered by it and continue with the rest.
2)Divide-and-conquer: split the data in subsets and apply the algorithm recursively to the subsets.
A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

C4.5:

The core algorithm for building decision trees is called C4.5 which employs a top-down, greedy search through the space of possible branches with no backtracking.A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous).

Features of c4.5

- C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan to address the following issues not dealt with by ID3:
- Avoiding overfitting the data
- Determining how deeply to grow a decision tree.
- Reduced error pruning.
- Rule post-pruning.
- Handling continuous attributes.
- e.g., temperature
- Choosing an appropriate attribute selection measure.
- Handling training data with missing attribute values.
- Handling attributes with differing costs.
- Improving computational efficiency.

C4.5 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

Entropy= $-p\log_2 p - q\log_2 q$,

where p is the no of positive outcomes and q the no of negative outcomes.

Gain: The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gainIt is calculated as:(X splits into T subsets)

Gain(T,X)=Entropy(T)-Entropy(T,X)

WEKA TOOL:

Weka is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing tilities in c, and a makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

Free availability Portability, since it is fully implemented in the Java -based programming language and thus runs on almost any modern computing platform. A comprehensive collection of data preprocessing and modeling techniques. Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, regression, visualization, and featureselection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using JavaDatabaseConnectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.

7.METHODOLOGY

Algorithm steps:

```
function ID3 (R: a set of non-categorical attributes,
                  C: the categorical attribute,
             S: a training set) returns a decision tree;
  begin
Step1: If S is empty, return a single node with value Failure;
Step 2:If S consists of records all with the same value for the categorical attribute, return a single node with that value;
Step 3: If R is empty, then return a single node with as valuethe most frequent of the values of the categorical attributethat are found in records of S
Step 4: Let D be the attribute with largest Gain(D,S)
              among attributes in R;
             Let {dj| j=1,2, .., m} be the values of attribute D;
             Let {Sj| j=1,2, .., m} be the subsets of S consisting
               respectively of records with value dj for attribute D;
             Return a tree with root labeled D and arcs labeled
               d1, d2, .., dm going respectively to the trees

                 ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
  end ID3;
```

Let us understand the working of the above algorithm using an example:

We consider a golfing example The categorical attribute specifies whether or not to Play. The non-categorical attributes are:

ATTRIBUTE	POSSIBLE VALUES
outlook	sunny, overcast, rain
Temperature	hot, sweet, cold
humidity	high , normal
windy	true, false

We have a training set as below:

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	false	Don'tPlay
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

We need to find the attribute that will be the root node in our decision tree. The gain is calculated for the four attributes. The entropy of the set S:

Entropy (S)=-9/14*log2(9/14)-5/14*log2(5/14)=0.94

Calculation for the first attribute Gain(S, Outlook) =

Entropy (S)-5/14*Entropy (Ssun)
-4/14*Entropy (Srain)
-5/14* Entropy (Sovercast)
=0.94 –5/14*0.9710-4/14*0 –5/14*0.9710

Gain(S, Outlook) = 0 .246
Calculation of entropies:
Entropy (SSunl) = -2/5*log2(2/5)-3/5* log2(3/5) = 0.9710
Entropy (Srain) = -4/4*log2(4/4)-0* log2(0) =0
Entropy (Sovercast) = -3/5* log2(3/5) -/5* log2(2/5) =0.9710
As well we find for the other variables:
Gain(S, Wind) = 0.048
Gain(S, Temperature) = 0.0289
Gain(S, Humidity) = 0.1515
Outlook attribute has the highest gain, so it is used as a decision attribute in theroot node of the tree.
Final tree looks like below figure:

```
Outlook
                         / |    \
                        /  |     \
     Overcast     /  |sunny  \rain
/   |       \
            Play   Humidity   Windy
                         /  |        | \
/   |        |  \
<=75 />75|   true|    \false
                  /    |        |    \
          Play   Don'tPlayDon'tPlay Play
```

8.CONCLUSIONS

Diabetes mellitus occurs throughout the world, but is more common in the more developed countries. The greatest increase in prevalence is, however, occurring in low- and middle-income countries including in Asia and Africa, where most patients will probably be found by 2030. The increase in incidence in developing countries follows the trend of urbanization and lifestyle changes, including increasingly sedentary lifestyles, less physically demanding work and the global nutrition transition, marked by increased intake of foods that are high energy-dense but nutrient-poor (often high in sugar and saturated fats). The risk of getting type 2 diabetes has been widely found to be associated with lower socio-economic position across countries.

hence with the proposed system model we can provide easy and quick medical help to the patients on the early detection of diabetes

ACKNOWLEDGEMENT

I remain immensely obliged to **Prof.Chaitrali Chaudhari** for providing me invaluable support in gathering resources for me either by way of information or computer also her guidance and supervision which made this work successful.

I would like to thank **Prof. Shobha Lolge**, Project Coordinator, **Dr. P. J. Nikumbh**, Head of Computer Engineering Department, **Dr. Subhash Shinde**, Vice Principal and **Dr. VivekYakkundi**, Principal and faculty and staff of Computer Engineering Department and **Lokmanya Tilak College of Engineering,** Navi Mumbai for their invaluable support.

I would like to say that it has indeed been a fulfilling experience for working out this project topic

REFERENCES

[1] Oxford dictionaries, the oec: Facts about the language. http://oxforddictionaries.com/page/ oecfactslanguage/ the-oec-facts-about-the-language, June 2011.

[2] American Diabetes Association., Living with Diabetes: Insulin Basics, June7, 2013:http://www.diabetes.org/living-with-diabetes/treatment-andcar e/medication/insulin/insulin-basics.html

[3] SwastiSinghal, Monika Jena, A Study on Weka Tool for Data Preprocessing, Classification and Clustering, IJITEE, 2 (2013), Issue-6. [

[4]C4.5 implementationhttp://www.otnira.com/2013/03/25/c4-5/