Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 7, July 2017, pg.166 – 172

AN ANALYSIS OF RISK FACTORS FOR DIABETES USING DATA MINING APPROACH

Miss. N. Vijayalakshmi[1], Miss. T. Jenifer[2]

[1]Asst. Professor, Dept. of M.C.A., Shrimati Indira Gandhi College, Trichy- 2

[2]Research Scholar in Computer Science, Shrimati Indira Gandhi College, Trichy-2

[1]Email: nvijimca@gmail.com, [2]Email: jenifermca2007@gmail.com

Abstract- One of the most significant health issue faced by men and women these days is diabetes. Although several factors are considered to lead to diabetes, it would be worth enough to find the most predominant factors causing this problem to gain a better understanding of the issue. Data mining and statistical analysis go hand in hand in identifying these factors from a clinical database containing primary data pertaining to significant factors relating to diabetics/non-diabetics in men and women. The sample population encompasses both diabetics and non-diabetics men and women relating to a good age spread. Data mining techniques like association rule mining, classification using decision tree induction, clustering, prediction using a decision tree approach and building an application based on the knowledge gained for predicting the probability of diabetics in a men and women have been used to thoroughly attain our objectives.

Keywords- Data mining, classification, prediction, association rules, statistical analysis, clustering

I. INTRODUCTION

Diabetes is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Diabetes mellitus is a complex group of diseases caused by a number of reasons. Individuals suffering from diabetes have hyperglycemia (high blood sugar) either because there is low production of insulin or body cells do not use the produced insulin. There are three main type of diabetes. These are typ1, type2, gestational diabetes. Common causes of diabetics includes increased frequency of urination, especially at night, frequently feeling thirsty, weakness and fatigue, unexplained loss of weight, genital itching or thrush, blurred vision, increase in healing time of cuts and wounds.

This paper uses data mining and statistical analysis techniques to identify the dominant factors causing diabetes in men and women. Already factors that are thought to be significant like age, BMI, High Cholesterol, Hyperthyroid, Hypertension, Sleeplessness, Arthritis, Vision problems, Skin problems, Kidney problems, Amputation due to unhealed wounds, Numbness/Tingling/Irritation are only considered. Among these the most significant ones leading to diabetes are identified. Characteristics of each significant factor are studied in diabetic and non-diabetic men and women leading to knowledge discovery of highly significant causes of diabetes in general. The entire data set is also subject to classification using two different decision tree induction methods and a comparative study of the methods is also undertaken. An attempt is also made to predict diabetics in men and women using the knowledge gained through decision tree induction and this is used to build a software model for the same. Association rules that govern diabetes are also generated using Association rule mining. Clustering is used to perform descriptive data mining.

II. LITERATURE SURVEV

The diabetes risk score model considering Age, BMI, waist circumference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables [1]. The authors indicate risk factors, in the final model, including blood pressure, cholesterol, back pain, fatty food, weight index or alcohol index [2]. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model [3].There are various data mining techniques and algorithm used for finding the diabetes. Neural Network, Artificial neural fuzzy interference system, K-Nearest-Neighbor (KNN), Genetic Algorithm, Back Propagation algorithm etc[4]. These techniques and the algorithms provide the better result to the people and the doctors regarding the diagnosis of the diabetes. From these results the people can predict he is affected with the diabetes or not.

Predict the human use the inputs from complex tests conducted in labs and also predict the disease based on risk factors such as tobacco smoking, alcohol intake, age, family history, diabetes, hypertension, high cholesterol, physical inactivity, obesity. In this study, we are using three different kinds of clustering techniques named as Hierarchical clustering; Density based clustering, and Simple K-Means clustering. [5]Weka is used as a tool. They computed a new variable age new as nominal variable, dividing in to three group's young age, middle age and old age and the target variable diabetes_diag_binary is a binary variable. They found 34% of the population whose age was below 20 years was not affected by diabetes. [6]33.9% of the population whose age was above 20 and below 45 years was not affected by diabetes. 26.8% of the population whose age was above 45 years was not diabetic.

To avoid the dangerous complications of the diabetes, patients should control a blood glucose level as the HbA1c (accumulative blood glucose level for 3 months) should be less than 7%. [7] In this paper a new predicted model has been developed by using data mining techniques [8]. The model aims to classify the diabetic patients into two classes which are: under control (HbA1c < 7%) and out of control (HbA1c > 7%). The treatments plans for 10061 diabetic patients were used to build the model. After comprehensive survey for classification techniques, three algorithms have been selected which were Naive Bayes, Logistic and J48 [9]. By using WEKA application, the model has been implemented.

Taking into account the prevalence of diabetes among men and women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool [10]. Classification is a method used to extract models describing important data classes or to predict the future data.

III. METHODOLOGY

A. Data Sources

There are numerous factors causing diabetics in men and women. Sample population consisting of 337 patients who are getting treated in an I.S. nursing home research center in Trichy are taken. Physical and environmental history of parents and siblings are taken into account. Data pertaining to 202 diabetics and 135 non-diabetics were collected for the same attributes. A questionnaire was created consisting of various parameters regarding the factors influencing diabetics. Those questionnaires were distributed to the patients who are visiting the centre for weekly/monthly check ups. The response was satisfying. Out of several independent attributes collected from outpatients, it is clear that only some of the factors really play a vital role in causing diabetics in men and women.

B. Statistical Analysis

Weka is a work bench 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. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. WEKA is a popular data mining tool. It is used to analyze the most significant factors causing diabetics. It is also used to perform statistical analysis of each individual attribute.

C. Data Mining

Data Mining may be defined as the composite of techniques employed to detect patterns in large datasets to extract hidden pieces of information. It is fairly new technique used to discover concealed patterns in the behavior of data. While statisticians have for some time been performing Data Mining manually, recent advances in statistical software, computer power and storage capabilities have enabled us to easily and accurately extract hidden patterns from databases.

1) Use of classification techniques

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In learning the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. Decision tree induction is a popular technique used for classification and prediction.

2) Use of k-means clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as pre-processing approach for attribute subset selection and classification.

3) Use of associations rule mining

Association Rule Mining is a popular and well researched method for discovering interesting relations between variables in large databases. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used.

IV. RESULTS AND DISCUSSION

A. Use of statistical analysis on the sample data using WEKA revealed the following facts:

1. Patients with age greater than 35 are more likely to become Diabetics than the others.
2. Patients with BMI >25 are more prone to Diabetes.
3. Parent with a history of diabetes mellitus may make the patient more prone to diabetes.
4. Siblings with a history of diabetes mellitus may make the patient more prone to diabetes.
5. Patients with arthritis and vision problems may be more prone to diabetes.
6. Patients who consume non-vegetarian food may be at a high risk of acquiring diabetics when they have the above indications.

B. Use of CFS Subset Evaluator to identify the most deterministic factors causing diabetes produced the following results. BMI, Age, Arthritis, Vision Problems, Sleeplessness, Parent with diabetes, Siblings with diabetes are the significant factors leading to diabetics. This also agrees with our observations gained through statistical analysis.

C. Two different classification techniques used produced the following results:

TABLE I

STRATIFIED CROSS VALIDATION SUMMARY OF TWO DIFFERENT CLASSIFICATION TREES

===Stratified Cross –Validation=== ====Summary===				
	J48 pruned tree technique		**Random tree**	
No. of leaves	14		-	
Size of tree	27		193	
Time taken to build model	0.08 seconds		0.05 seconds	
No. of records/attributes	337/10		337/10	
Correctly Classified Instances	272	80.7122%	253	75.0742%
Incorrectly Classified Instances	65	19.2878	84	24.9258%
Kappa statistic	0.5898		0.4835	
Mean absolute error	0.2557		0.2493	
Root mean squared error	0.3967		0.4993	
Relative absolute error	53.2368%		51.8859%	
Root relative squared error	80.9559%		101.8782%	
Total Number of Instances	337		337	

TABLE II

COMPARATIVE STUDY OF TWO DIFFERENT CLASSIFICATION TREES – DETAILED ACCURACY BY CLASS

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
J48 Pruned Tree	0.881	0.304	0.813	0.881	0.846	0.593	0.807	0.810	DIABETES
	0.696	0.119	0.797	0.696	0.743	0.593	0.807	0.740	NON-DIABETES
Random Tree	0.782	0.296	0.798	0.782	0.790	0.484	0.743	0.755	DIABETES
	0.704	0.218	0.683	0.704	0.693	0.484	0.743	0.600	NON-DIABETES

Hence we find that J48 pruned tree is a relatively better technique in terms of accuracy in classifying the given record sets with an accuracy of 81%

D. Use of k-means clustering on the given data set produced the following results:

TABLE III

RESULTS ANALYSIS OF K-MEANS CLUSTERING ON THE GIVEN DATA SET

Initial Cluster Centroids			**Final Cluster Centroids (after 3 iterations)**		
	Cluster 0	**Cluster 1**	**Full Data (337)**	**Cluster 0 (80)**	**Cluster 1 (257)**
Age	70	48	51.7151	63.025	48.1946
BMI	22.662709	22.230987	26.3815	25.5264	26.6477
Sleeplessness	YES	NO	NO	YES	NO
Arthritis	YES	NO	NO	YES	NO
Problem in Vision	YES	NO	NO	YES	NO
Numbness/Tingling/Irritation	NO	NO	NO	NO	NO
Kidney Problems	YES	NO	NO	NO	NO
Parent Sugar	YES	NO	NO	YES	NO
Sibling Sugar	NO	NO	NO	YES	NO
Result	Non-diabetes	Diabetes		Non-diabetes	Diabetes

Within cluster sum of squared errors 492.74

Clustered Instances

0 80 (24%)
1 257 (76%)

E. Use of Association rule mining on WEKA for the given data yielded the following results:

Minimum support: 0.75 (253 instances) Minimum metric <confidence>: 0.9
Number of cycles performed: 5
Generated sets of large itemsets:
Size of set of large itemsets L (1): 4
Size of set of large itemsets L (2): 5
Size of set of large itemsets L (3): 2

TABLE IV

ASSOCIATION RULE MINING WITH WEKA ON GIVEN DATASET

Association rules found	Support	Confidence	Lift	Level	No. of records	Convergence
NUMBNESS/TINGLING/ IRRITATION=NO 327 ==> BMI='All' 327	327	1	1	0	0	0
KIDNEY PROBLEMS=NO 300 ==> BMI='All' 300	300	1	1	0	0	0
. NUMBNESS/TINGLING/ IRRITATION=NO KIDNEY PROBLEMS=NO 293 ==> BMI='All' 293	293	1	1	0	0	0
SLEEPLESSNESS=NO 264 ==> BMI='All' 264	264	1	1	0	0	0
SLEEPLESSNESS=NO NUMBNESS/TINGLING/ IRRITATION=NO 258 ==> BMI='All' 258	258	1	1	0	0	0
SLEEPLESSNESS=NO 264 ==> NUMBNESS/ TINGLING/ IRRITATION =NO 258	258/264	0.98	1.01	0.01	1	1.12
BMI='All' SLEEPLESSNESS=NO 264 ==> NUMBNESS/ TINGLING/ IRRITATION =NO 258	258/264	0.98	1.01	0.01	1	1.12
SLEEPLESSNESS=NO 264 ==> BMI='All' NUMBNESS/TINGLING/ IRRITATION=NO 258	288/264	0.98	1.01	0.01	1	1.12
KIDNEY PROBLEMS=NO 300 ==> NUMBNESS/ TINGLING/ IRRITATION =NO 293	293/300	0.98	1.01	0.01	1	1.11
BMI='All' KIDNEY PROBLEMS=NO 300 ==> NUMBNESS/TINGLING/ IRRITATION=NO 293	293/300	0.98	1.01	0.01	1	1.11

F.A C program to predict diabetics based on the J48 decision tree was rested with the given training set. An accuracy of **81%** was obtained.

V. CONCLUSION

In this paper, we made an attempt to use **data mining as a tool** for analyzing clinical data records of both diabetic and non-diabetic patients. Data mining is a powerful tool which is currently used for extracting significant information from historical data. This information can be used for further decision making and prediction. WEKA was used for applying various data mining techniques like Statistical analysis, Associative rule mining, and Clustering, Classification and subset evaluation. This has been very helpful in extracting key information regarding diabetics. Two classification methods were used on the same record set to produce almost similar results at varying levels of accuracy. Among them, J48 pruned tree has been found to be more accurate. Clustering is also carried out to verify the output of previous methods. From the information gained an attempt is also made to build a decision tree model for prediction of diabetics. The accuracy of prediction is 81%.

ACKNOWLEDGEMENT

We acknowledge with thanks the support given by I.S. Nursing Home and Research Centre, Trichy for having provided us with primary clinical data relating to diabetics on 337 diabetics and non-diabetics men and women patients only for the purpose of research on diabetics.

REFERENCES

[1]. K. Meena, N. Vijayalakshmi, "An Analysis of Risk Factor for Diabetes using Data Mining Approach", *Indian Journal of Public Health Research and Development*, Vol. 6, Issue No. 2, pp 112-117, April-June 2015.
[2]. Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", *International Journal of Computer Applications*, Vol.10, Issue No.4, November.2010.
[3]. P.Thangaraju, B.Deepa, T.Karthikeyan, "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, Issue No. 8, August 2014.
[4]. J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," Diabetes Care, 26:3 (2003), 725-731.
[5]. P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", *International Journal of Innovative Science, Engineering & Technology*, vol. 1 Issue 6, August 2014, pp. 334-339.
[6]. Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.
[7]. Varsha Kavi and Divyesh Joshi , "A Survey on Enhancing Data Processing of Positive and Negative Association Rule Mining", *International Journal of Computer Sciences and Engineering*, Volume-02,Issue-03, Page No (139-143), Mar -2014.
[8]. Sharma, Trilok Chand, and Manoj Jain. "WEKA Approach for Comparative Study of Classification Algorithm."
[9]. P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool". *International Journal of Scientific & Engineering Research*, Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool.
[10]. K. R. Lakshmi and S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability", *International Journal of Scientific & Engineering Research*, Volume 4, Issue 6, June-2013.

