ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

A STUDY ON DATA MINING AND STATISTICAL METHODS USED IN DIABETES MELLITUS DIAGNOSIS.

DR. M. Mayilvaganan [1], R. Deepa[2] and P. Nandakumar[3].

1. Associate Professor, Dept. of Computer Science, PSG College of Arts & Science, CBE Tamil Nadu.
2. Research Scholar, Dept. of Computer Science, PSG College of Arts & Science, CBE Tamil Nadu.
3. Research Scholar, Dept. of Electronics, PSG College of Arts & Science, CBE Tamil Nadu.

Manuscript Info

Manuscript History:

Received: 18 May 2016
Final Accepted: 19 June 2016
Published Online: July 2016

Key words:
Diabetes Mellitus [DM], Insulin, Hyperglycaemia, Type -I DM, Type-II DM, Data Mining Techniques, Statistical methods, classification methods, cluster methods, Metrics.

Corresponding Author

R.Deepa.

Abstract

Diabetes is one of the most prevalent diseases in the world today with high mortality and morbidity rate, thus one of the biggest health problems in the world. Diagnosis of diseases is a vital role in medical field. The use of data mining on medical data brings important, valuable and effective achievement, which can enhance the medical knowledge to make necessary decision. The paper is organized as follows; it first gives a study done on diabetes and its types. Second it explains the Data Mining techniques and Statistical method used to predict Diabetes. Then the paper ends by concluding with summary of investigated methods.

Introduction:-

Diabetes mellitus[DM], or simply diabetes, is a chronic disease that occurs when the pancreas is no longer able to make insulin, or when the body cannot make good use of the insulin it produces. Insulin is a hormone made by the pancreas, that acts like a key to let glucose from the food we eat pass from the blood stream into the cells in the body to produce energy. All carbohydrate foods are broken down into glucose in the blood. Insulin helps glucose get into the cells. Not being able to produce insulin or use it effectively leads to raised glucose levels in the blood (known as hyperglycaemia). Over the long-term high glucose levels are associated with damage to the body and failure of various organs and tissues. Figure -1 shows the Glucose and insulin flow in body.

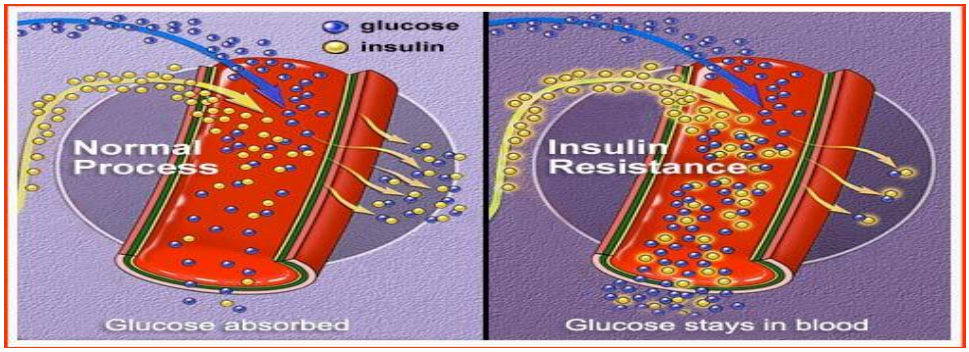

Figure 1:- The Glucose and insulin flow in body.

Diabetes mellitus is a clinically and genetically heterogeneous group of disorders that have one common Feature - abnormally high levels of glucose in the blood due either to insulin deficiency or to resistance of the body's cells to the action of insulin.

Types of diabetes:-

There are three main types of diabetes:

Type-I diabetes used to be called juvenile-onset diabetes or insulin dependent diabetes. It is usually caused by an auto-immune reaction where the body's defence system attacks the cells that produce insulin. People with Type-I diabetes produce very little or no insulin. The disease may affect people of any age, but usually develops in children or young adults. People with this form of diabetes need injections of insulin every day in order to control the levels of glucose in their blood. If people with Type-I diabetes do not have access to insulin, they will die.

Type-II diabetes used to be called non-insulin dependent diabetes or adult-onset diabetes, and accounts for at least 90% of all cases of diabetes. It is characterized by insulin resistance and relative insulin deficiency, either or both of which may be present at the time diabetes is diagnosed. The diagnosis of type-II diabetes can occur at any age. Type-II diabetes may remain undetected for many years and the diagnosis is often made when a complication appears or a routine blood or urine glucose test is done. It is often, but not always, associated with overweight or obesity, which itself can cause insulin resistance and lead to high blood glucose levels. People with Type-II diabetes can often initially manage their condition through exercise and diet. However, over time most people will require oral drugs and or insulin.

Gestational diabetes (GDM) is a form of diabetes consisting of high blood glucose levels during pregnancy. It develops in one in 25 pregnancies worldwide and is associated with complications to both mother and baby. GDM usually disappears after pregnancy but women with GDM and their children are at an increased risk of developing Type-II diabetes later in life. Approximately half of women with a history of GDM go on to develop Type-II diabetes within five to ten years after delivery.Table-1 shows the Normal Glucose Level chart.

Table-1: Normal Glucose Level

Blood sugar classification	**Fasting Blood Sugar Levels**	**Post Meal Blood Sugar Level**
Normal	70-100 mg/dl	70-140 mg/dl
Prediabetes	101-125 mg/dl	141-200 mg/dl
Diabetes	125 mg/dl and above	200 mg/dl and above

Symptoms, Diagnosis and Treatment:-

The common symptoms of a person suffering from diabetes are:

- Polyuria (frequent urination)
- Polyphagia (excessive hunger)
- Polydipsia (excessive thirst)
- Weight gain or strange weight loss
- Healing of wounds is not quick, blurred vision, fatigue, itchy skin, etc.

Urine test and blood tests are conducted to detect diabetes by checking for excess body glucose. The commonly conducted tests for determining whether a person has diabetes or not are

- A1C Test
- Fasting Plasma Glucose (FPG) Test
- Oral Glucose Tolerance Test (OGTT).

Though both Type-I and Type II diabetes cannot be cured they can be controlled and treated by
Special diets, regular exercise and insulin injections. The complications of the disease include the earlier diagnosis of diabetes; risk of the complications can be dodged.

Complications of Diabetes:-

Long-term complications of diabetes develop gradually. The longer you have diabetes — and the less controlled your blood sugar — the higher the risk of complications. Eventually, diabetes complications may be disabling or even life-threatening. Possible complications include:

- Cardiovascular disease.
- Nerve damage (neuropathy)
- Kidney damage (nephropathy)
- Eye damage (retinopathy).
- Foot damage.
- Skin conditions.
- Hearing impairment.
- Alzheimer's disease

Data Mining:-

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides [13]. Data mining is process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [14]. KDD process may Consists several steps: like data selection, data cleaning, data transformation, pattern searching i.e. data mining, finding presentation, finding interpretation and finding evaluation [15].Data mining technique are applied to analyse medical data for decision-making to guide the physicians. Figure-2 and Figure-3 Shows the KDD Process and Data mining methods.

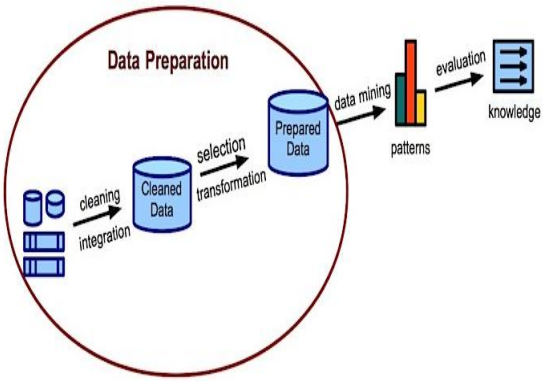

Figure 2:- KDD Process

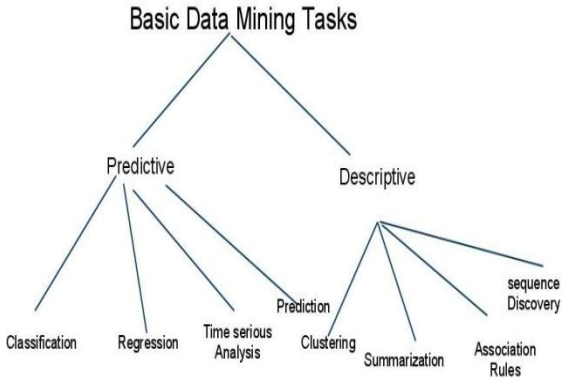

Figure 3:- Data mining methods.

Data Mining Techniques used in predicting diabetes Mellitus:-

The Table-2 present the Study made on reviewed papers with regard to diabetes analyses. It present the various techniques and tool used in predicting the diabetic mellitus and their finding.

Table 2:- Data Mining Techniques and Tools used in predicting diabetes Mellitus.

Sno	Authour Year	Methods and Tool used	Result
1	Najmeh hosseinpour, Saeed Setayeshi, Karim Ansari asi, Mohammad Mosleh 2012[1]	Bayesian, Functional, Rule based, Decision Tree and Ensemble Weka Tool	Ensemble classifier with logistic core has better performance in comparison with other classifier.
2	Sukhjinder Singh, Kamaljit Kaur 2013[2]	Neural Network, AN Fuzzy interference system, KNN, ML, principal component analysis(PCA) Weka tool	The outcome study is PCa combined with neural Network for classification and achived the best accuracy 71%
3	K.Jothi, M.Ramya, S.Kauser, L.Thomas Robinson, S.Raj Anand 2014[3]	ARIMA Model And Metabo System	The result predict future blood glucose so that awaiting dangerous hyper glycemia can be incidental in advance and preventive measures can be taken.

4	Ravi sankal, T.Jayakumari 2014[4]	FCM, SVM, SMO Weka tool	The result of the research shows FCM as best among other with accuracy of 94.3%.
5	P.Radha, Dr.b.Srinivasan 2014[5]	C4.5, SVM, KNN, PNN, BLR Tanagra tool	BLR algorithm plays a vital role in Data mining. BLR has low computing time and 75% accuracy compared with others.
6	Sadri sa'di, Amanj maleki, Rami hashemi, Zahra panbechi, kamal chalabi 2015 [6]	Navi bayes, RBF Network, J48 Weka tool	The result revealed thet naïve Bayes having accuracy rate of 76.95%.
7	Aiswarya Iyer, S.Jeyalatha, Ronak Sumbaly 2015 [7]	Decision Tree, (ID3, C4.5,C5, J4.8, CART and CHAID),Naïve Bayes Weka tool	The study proves that both model are efficient in diagnosis of diabetes using the percentage split of 70:30 of the dataset.
8	M.Durairaj, G.kalaiselvi 2015 [8]	Artificial Neural Network (ANN) C4.5, classifier, Suppourt vector Machine (SVM), K-Nearest Neighbour (KNN).	Comparison results of different data mining techniques shows that the ANN gives highest accuracy above 89% result than other similar techniques.
9	SrideivanaiNagarajan R.M.Chandrasekaran 2015 [9]	K-means algorithm, Naive Bayes, Random Tree, Simple Cart and Simple Logistic	This paper demonstrates creation of expert clinical system for the diagnosis of the diabetic mellitus using clustering and classification techniques of data mining.
10	Sukhjinder Singh, Kamaljit Kaur 2015 [10]	Neural Network, Artificial neural fuzzy interference system, K-Nearest-Neighbor (KNN), Genetic Algorithm, Back Propagation algorithm, Principal Component Analysis (PCA)	Using the data mining technique the health care management predicts the disease and diagnosis of the diabetes and then the health care management can alert the human being regarding diabetes based upon this prediction.
11	Dr.Omprakashjadhav, Nita Jivraj 2016 [11]	Statistical method, classification method, cluster method	Discussed various statistical method used to detect risk factor on diabetes and prevention measures. Result of the research shows that ANN gives more accurate prediction 89% than other similar techniques.
12	Dr.Renuka Devi , J.Maria Shyla 2016 [12]	Navie Bayer, MLP, Bayesian Network, C4.5, KNN, ANFIS, PLS-LDA, ANN, J4.8 Weka tool	J4.8 classifier gives 99.87% of higher accuracy.

Statistical Techniques and methods used in predicting diabetes Mellitus:-

The Table-3 present Statistical methods used to analyses diabetes mellitus. It present the various techniques and tool used in predicting the diabetic mellitus and their finding. Statistical analysis presents the statistical methods with 4 different categories and considered 3 parameters in which for Type-I DM and Type-II DM mostly parametric & non-parametric methods are used. And for comparison between Type-I DM and Type-II DM, distributions and parametric and non-parametric methods are equally used [11].

Table 3:- Statistical methods used in predicting diabetes Mellitus.

S.No	Diabetes Type	DISCRIPTIVE STATISTICS	DISTRIBUTIONS	REGRESSION	PARAMETRIC & NON-PARAMETRIC METHODS
1	**Type-1**	Mean, Standard Deviation	Poisson Distribution With Log Link	Meta-Analysis, Linear Regression Model, Meta Regression Analysis,Linear Mixed Model With Random Effect, Pearson Correlation Coefficient	Cochran's Q Test, I2 Statistics, Students t-Test, Two Sample t-Test, One Sample t-Test, One Way ANOVA With Post-Hoc, Tuckey's Test Random Effect Meta-Analysis, Random Effect, Fixed Effect Model, Bayesian Approaches, Funnel Plot, ANOVA
2	**Type-2**	Standard Deviation, Mean, Median, Interquartile Range, Frequencies, Percentage, Graphical & Diagrammatic Representation	Log Link With A Binomial Distribution,Identity Link With A Normal Distribution	RRR-Reduced Rank Regression, PCAPrincipal Component Analysis, PLS-Partial Least Square, Robust Generalized Estimating Equations, Least Square Analysis, Logistic Regression, Restricted Cubic Spline regression Model,Multiple logistic Regression,Multivariable Logistic Regression, Log Linear Poisson Regression Model, Exploratory Linear Regression Analysis	Two Sample t-Test,Factor Analysis, Cluster Analysis, Hazard Ratio, Proportional Hazard Ratio Test, Trim & Fill Method, Sensitivity Analysis, Forest Plot, Begg Funnel Plot, Egger's Test, Likelihood Ratio Test, Mantel (Log-Rank) Test, Stratified Analysis, Chi-Square Test, Pearson Chi-Square Test Students t-Test, Cochran's Q Test, I2 Statistics, Random Effect Model, Wilcoxon Rank Sum Test,Wilcoxon Matched Pair
3	**Comparison between Type-I and Type -II**	Median(range or Interquartile Range)	2-sided skew corrected inverted score test, binomial distribution	Global Logarithmic linear method	Mann-Whitney U Test maximum likelihood analysis

Metric used in Performance Evaluation

A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results Table-4 shows the confusion matrix.

Table 4:- Confusion Matrix.

	Classified as Healthy	**Classified as Not Healthy**
Actual Healthy	**TP(True positive)**	**FN(False Negative)**
Actual Not Healthy	**FP(False positive)**	**TN(True Negative)**

From the confusion matrix to analyse the performance criterion for the classifiers in disease detection accuracy, precision, recall have been computed for all datasets. Accuracy is the percentage of predictions that are correct. The

precision is the measure of accuracy provided that a specific class has been predicted. Recall is the percentage of positive labeled instances that were predicted as positive [5]. The fitness criteria are calculated as follows:

Sensitivity = TP / (TP + FN)
Specificity = TN / (FP+TN)
Accuracy = (TP + TN) / (TP + FP + TN + FN)
where

- TP is True Positive: Diabetic patients correctly diagnosed as Diabetic
- FP is False Positive: Healthy people incorrectly identified as Diabetic
- TN is True Negative: Healthy people correctly identified as healthy
- FN is False Negative: Diabetic

Conclusion:-

The main goal of medical data mining algorithm is to get best algorithms that describe given data from multiple aspects. The study made gives a various data mining and statistical method used in diagnosis diabetes Mellitus. Diet plays a main role in preventions and treatment of diabetes. Various factor are responsible for Type-I and Type-II diabetes. Awareness is needed to the people about self management and have a methodology which provide valuable information regarding improvement of healthcare using the application like smartphone So with the help of newer statistical application there is need to more study the causes of increasing diabetes in people mostly in youth because diabetes have long term complications such as retinopathy, neuropathy & nephropathy in diabetic patients. And there is also need to use proper methods, because poor methods affect the reliability of prediction model & ultimately compromise the accuracy of result. It is recommended to diagnosis diabetes with other methods such as Neuro Fuzzy Networks and compare with the algorithm used in this study, to determine the better method to diagnose the diabetes.

References:-

1. Najmeh hosseinpour, Saeed Setayeshi, Karim Ansari asi, Mohammad Mosleh, "Diabetes diagnosis by using computation intelligence algorithms",-International journal of advance research in computing science and software engineering, vol 2, issues 12, December 2012.
2. Sukhjinder Singh, Kamaljit Kaur,"A review on diagnosis of diabetes in data mining"- Internation Journal of Science and Research (IJSR), vol 4, issues 6, June 2015.
3. K.Jothi, M.Ramya, S.Kauser, L.Thomas Robinson, S.Raj Anand, "Prediction of hyperglycemia in diabetic patients using data mining techniques"-International journal of scientific Engineering and Technology vol 3, no 5 pp: 668-670.
4. Ravi sankal, T.Jayakumari, "prognosis of diabetes using data mining approach fuzzy c mean clustering and support vector machine"-International journal of computer Trends and Technology (IJCIT) vol 11 number 2 may 2014.
5. P.Radha, Dr.B.Srinivasan "International Journal of Innovative science science Engineering & Technology vol 1 issue 6 august 2014.
6. Sadri sa'di, Amanj maleki, Rami hashemi, Zahra panbechi, kamal chalabi, "Comparison of Data Mining algorithms in the diagnosis of type- II diabetes" – International journal on computational science & Application(IJCSA) vol 5 no 5 october 2015.
7. Aiswarya Iyer, S.Jeyalatha, Ronak Sumbaly, International journal of Data Mining & Knowledge Management process (IJDKP)) vol 5 no 1January 2015.
8. Durairaj M., Kalaiselvi G.,"Prediction of Diabetes Using Soft Computing Techniques- a Survey" International journal Of scientific & technology research, volume 4, ISSUE 03, PP.190-192, 2015.
9. NagarajanSrideivanai and Chandrasekaran R. M.,"Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques", Indian Journal of Science and Technology, Vol 8(8), PP. 771– 776, 2015.
10. SinghSukhjinder, Kaur Kamaljit, "A Review on Diagnosis of Diabetes in Data Mining", International Journal of Science and Research (IJSR); Volume 4, Issue 6, PP.2406-2408, 2015.
11. Dr.Omprakashjadhav, Nita Jivraj, "Study of risk factors and preventions for diabetes using statistical methods : A Systematic review", International journal of innovative Research in science,Engineering and Technology vol 5, issues 4, April 2016.
12. Dr.M.Renuka Devi, J.Maria Shyla, "Analysis of various data mining Techniques to predict diabetes Mellitus",- International journal of Applied engineering Research", ISSn 0973-4562 vol 11 number 1 2016.
13. HianChyeKoh and Gerald Tan: Data Mining Applications in Healthcare. Journal of Healthcare Information Management, Vol 19, No 2.
14. P. Giudici: Applied Data Mining Statistical Methods for Business and Industry. Wiley & sons, 2003.
15. G.Piatetsky-shapiro, U.Fayyed and P.Smith: From data mining to Knowledge discovery: An overview.
16. Advances in knowledge Discovery and Data Mining.pages 1-35, MIT Press, 1996.