International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 07, July-2015

Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms

P.Swathi Baby[1]
[1]CSE, Raghu Engineering College,
Visakhapatnam, India.

T. Panduranga Vital[1]
[1]CSE, Raghu Engineering College,
Visakhapatnam, India.

***Abstract*: Data mining techniques has been used as a recent trend for achieving diagnostics results, in medical fields such as kidney dialysis. Data mining concepts are used to examine a rich collection of data from different perspectives and deriving useful information. This project intends to diagnosis and prediction system based on predictive mining. This paper covers the essential problems associated with the data mining method and its usage within the medical field. The data has been collected from Visakhapatnam district during 2014 and 2015 with 690 instances and 49 attributes. This has been done in cases of kidney diseases. It was proved that the data mining is applicable within the medical sector and can improve the various medical applications .The K-Means(KM) algorithm is a major role in determine the number of clusters k for large datasets The data mining classification techniques are done based on namely ADT trees, Naïve Bayes , J48 are analyzed on kidney disease data set.**

Keywords: Kidney, Statistical Analysis, Classifications, Data Mining, Machine Learning.

I.INDRODUCTION

The extraction of data from huge unknown, potential information is known as data mining. [1] In the present contemporary world, data mining became popular in the health care fields due to its efficient requirements for effective detecting of obscure and important information from the data. The data being collected and stored automatically and finally processed to better analytical methods. This huge amount of data being processed effectively using various techniques. Medically, data mining is used to analyses the information obtained from research reports, the data which obtained are processed using the data mining tool techniques.[4] Here, in this paper we obtain a detail explanation of the present research that is being carried out on kidney datasets using various data mining techniques.

II. METHODOLOGY

The data has been collected from Visakhapatnam district during 2014 and 2015 with 690 instances and 49 attributes(Gender, Age, bath, blood group, body temp, breakfast, coffee, dinner items, diseases, drinking, fast foods , food habits, gender, height, job position, kidney stone, fruits intake, leafy-veg, lunch items, meals , milk, no drink, non-veg, coffee intake, milk, tea, smoke, place, fruits preferred, leaf, pregnant, relation members, salt consumption, scatter plot, sleeping, smoke, soft drinks, surgeries, sweat, tea, type soil, type water, tablets used , water consumption, weight, yoga)[2,15].

The dataset has been analyzed using Weka and Orange software's. The frequency of patients has been analyzed from the collected data from Visakhapatnam district. The questionnaire has been framed based on the present personal profile, living and food habits, travelling mode and usage of previous history, and internal and external factors [2,3]. Here the data has been analyzed with the help of Weka Software (Waikato Environment for knowledge Analysis, "it is a popular suite of machine learning software written in Java.[4] Here the Weka software consists huge collection of algorithms and visualization tools for predictive modeling and analysis of data. It also supports standard data mining tasks, more specifically like Visualization, Regression, Data processing, Features Selection. [3]

A. Model:

Data modeling is a progression in which numerous sets of data are collected and combined and further analysis is done to uncover the patterns or relationships. [11] The main aim of data modelling is to provide a future result using the past data. Here in this paper it shows the Processing model.

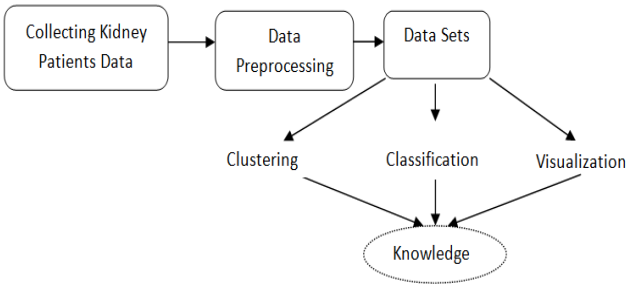

Fig. 1 Processing model

B. Clustering Analysis:

In data mining clustering is an essential process. The cluster or clustering analysis is the process of combining a group of items into a same group and into their relationships. The K-Means (KM) algorithm is a major role in determines the number of clusters k for large datasets. It needs to predefine the k value which itself difficult and tough to calculate the occurrence of number of clusters in the data.[7,8] There are no competent and universal methods to select the best number of clusters.

C. Classifiers Used:

In this project we used the Classification algorithm. The Weka software has been categorized into different groups those are Bayes, Tree based classifiers etc. Here few more selected group of algorithms are chosen from the group of algorithms those are Bayes Net & Naive Bayes (from Bayes), K Star (from Lazy), AD Tree, J48, Random Forest.[6]

1. *ADTress*

Here the classification is performed with the help of machine learning in an Alternating Decision Tree (ADTree). An ADTress resides on two different nodes those are decision nodes and prediction nodes. AD Trees constantly have prediction nodes for root and leaves. An instance is categorized by an AD Tree using succeeding paths in which every decision nodes is true and converts the prediction nodes that are to be traversed.

2. *J48*

Here a decision tree is generated with the help of J48. The classification in J48 is the base obtained from C4.5. It is an open source Java implementation of C4.5 algorithm. In J48 it consists of binary trees**[15]**.

2. *Naïve Bayes*

Naïve Bayes depends on probabilistic knowledge. Here it is only effective to multiply possibilities when the events are independent.[5] Here the Bayes theorem depends on the Naïve Bayesian classifier for an assumption among predictors. To find a useful large datasets Naive Bayesian is best an easy to obtain.

4. *Random Forest*

In Random forest it obtain class as output from many decision trees it is a joint classifier which contain many decision tree. Here it develops lots of decision tree based on Random selection of data and Random selection of variables. Without pruning the classification tress are obtained from the Random Forests.

III.RESULTS AND DISCUSSION:

Fig. 2 data has been provided the statistical relationship of the collected datasets. It shows the personnel profile of the collected data from kidney patients. There are more number of instances observed in Males (55.5%) compared with females (44.5%). The disease in urban area are more (61.3%) compared to rural (38.7%). The average age groups observed at 44.32±11.26, Height as 163.30±8.24cms and weight as 64.66±10.49Kgs.

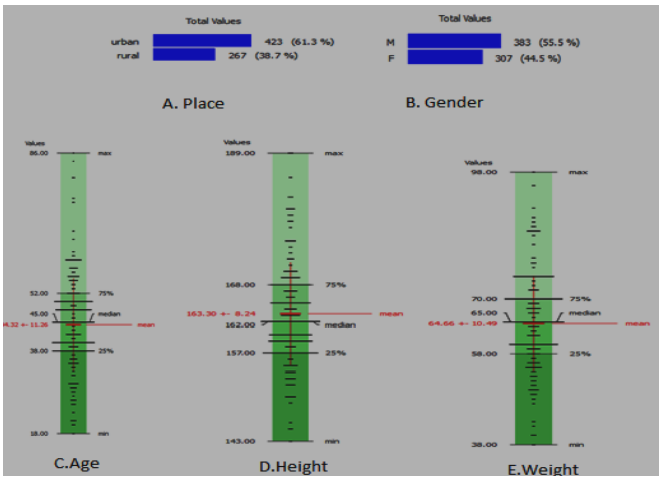

Fig. 2 Personal Profile

Fig.3 shows the food habits of kidney patients. Out of 690 patients 379(54.9%) prefer drinking Coffee, 382(55.4%) prefer drinking milk, 534(77.4%) prefer drinking Tea, 464(67.2%) prefer Non-Veg and 226(32.8%) prefer Veg out of total 690,276(40.0%) prefer fast foods, 221(32.0%) prefer to drink.

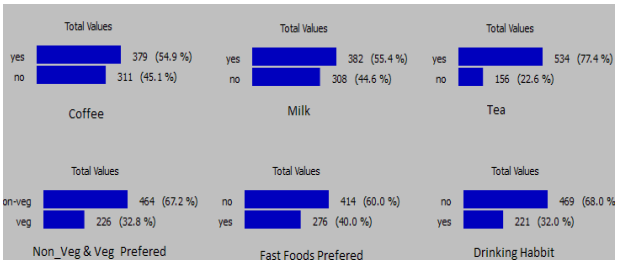

Fig. 3 Food and Drinking Habits

Fig. 4 shows the various diseases occurred to the patients. More number of people are suffered with the diabetes 88(12.8%), BP 78 (11.3%) compared to the rest of the diseases analysis. Kidney stones occurred before and diagnosed in the early stages 208(30.1%) is found through analysis. The kidney stone history indicates the person affected with the kidney stone diseases and diagnosed and again reformation of the stone occurred in 208 cases (30%) during the analysis.

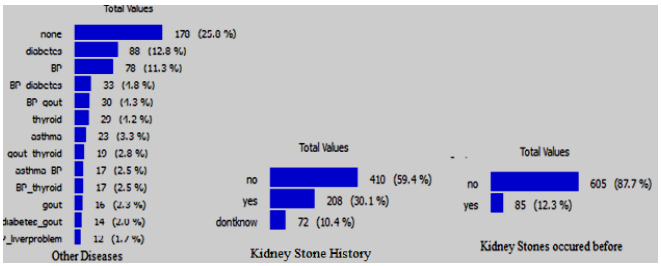

Fig. 4 Other Diseases, Kidney stone history, Kidney stones occurred before

Fig. 5 provides the sensations which are occurred during the kidney diseases starting stages and during the different stages of kidney analysis those are vomiting sensations (33.6%) and sweating sensations (42.6%) which is high as per the analysis.

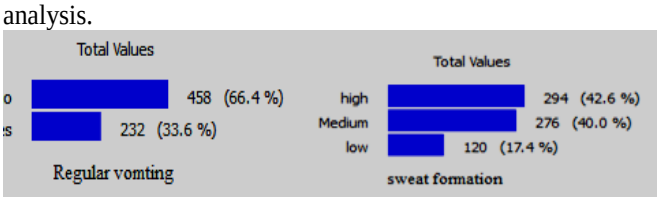

Fig. 5 Sensations occurred during the different stages

Fig. 6 shows the intake of unhealthy foods and liquids at regular intervals of time that leads to the kidney disease. High rate of leafy-veg consumed is sorrel spinach (22.5%), High rate of smoking in a week (65.2%), number of times the intake of Non-veg a week is 226(32.8%).

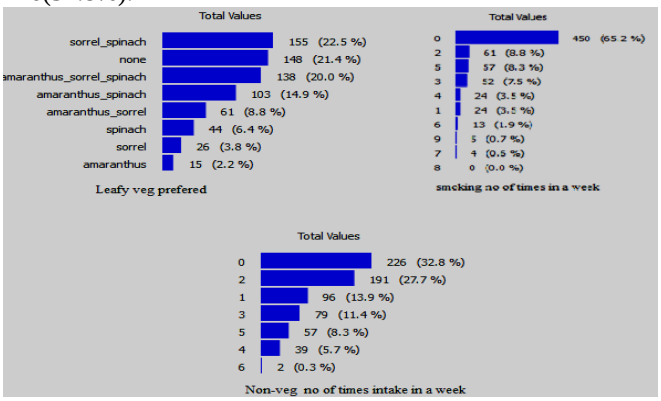

Fig. 6 Unhealthy habits that leads to kidney diseases

K-MEANS ALGORITHM:

Fig 7 Shows the K-Means (KM) algorithm is a major role in determine the number of clusters k for large datasets.[10,13] Hence the value's selected as random. The key challenge in the clustering process is sensitivity in the selection of the initial partition, in order to overcome this issue implement the hybrid algorithms to select best number of clusters.[8,9] Here the output obtained by using the Simple K-Means with two (2) Clusters namely 0 and1, the number of iteration performed are 5. The cluster instances obtained for Cluster 0:303(44%), Cluster 1: 387(56%).The cluster 0 is classified as gender: F (female), age(40.9637),weight(62.429),Blood Group(O+ve) and so on. As well the Cluster 1 is classified as gender: M (male), age (46.9432) Weight (66.4083), Blood Group (A+ve) and so on.

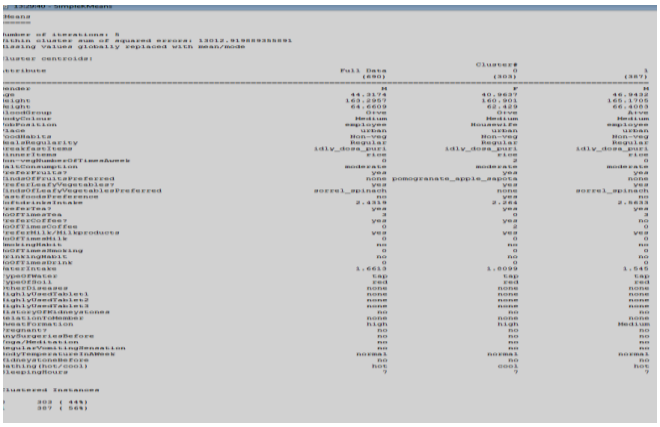

Fig. 7 K-Means Clustering

CLASSIFICATION ALGORITHMS:

Fig. 8 provides Cluster analysis using AD Trees. Here 0.14secs is the total amount of time taken to build the model. Correctly classified instances 648 (93.913%). Incorrectly classified instances 42(6.087%).The kappa statistics is 0.8771.The kappa (first syntax) calculates the kappa-statistic measure of interpreter agreement when there are two unique raters and two or more ratings.[12]

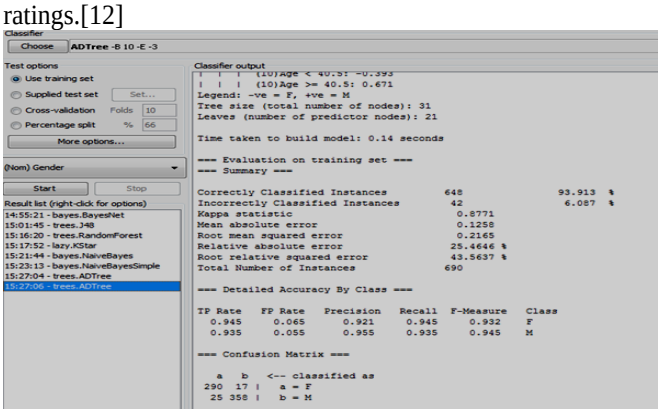

Fig. 8 Provides cluster analysis using AD Tress results

Fig. 9 Shows the visualization of the data sets using the AD Trees. [14] Here the analysis is based on the gender in such a way that the attributes are considered those are the job positions, kinds of food intake, smoking habit. Based on the gender the visualization of all these are obtained in the AD Trees. The fig. shows AD Tree (Alternative Decision Tree). It gives Very interesting predicting values. The female Kidney patients predicted value is Negative (-ve) and Male is Positive (+ve).most of the Females are Housewives (value is -2.09).most of male kidney patients habituated by drinking alcohol it is cause of kidney problem. Most of Non-vegetarian Male patients are eating bellow 3.5 days a week (0.113). Most of Non-vegetarian female patients are eating above 3.5 days a week (-0.747).

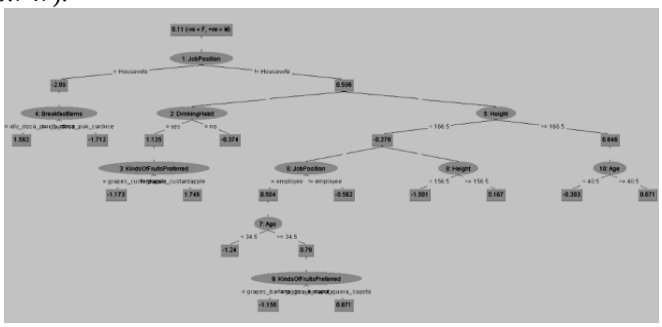

Fig. 9 Visualization in AD Trees using gender

Fig. 10 shows the classification of the data sets using the J48 algorithm. Here the time taken to obtain the accurate data is 0.13secs. The correctly classified instances (677) it obtained 98.1159% accuracy. The incorrectly classified instances are (13) it obtained 1.8841% incorrect data out of the total instances (690).

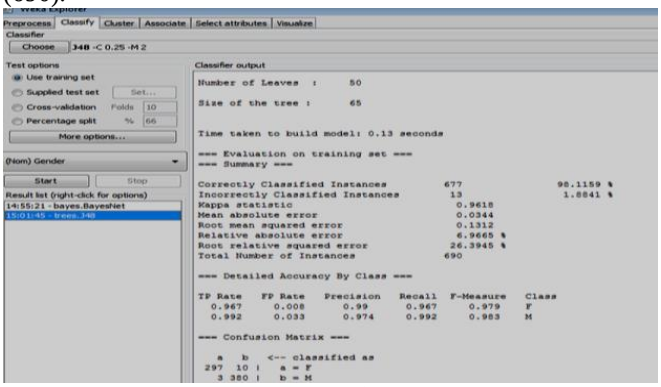

Fig. 10 J48 Classification results

Fig. 11 shows the visualization of the data sets using the J48 Trees. Here the analysis is based on the gender in such a way that the attributes are considered those are drinking habit, here according to the Fig. 11 male suffering with diseases are167 due to alcohol consumption. Here the different types of soils are also considered the analysis obtained in such a way that the black soil which is the best out of all types of soil. The people who are staying in black soil will never leads to any kidney diseases.

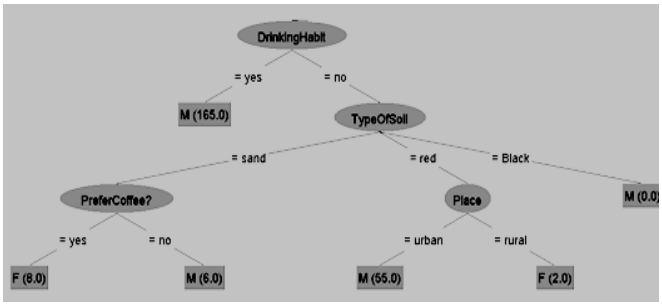

Fig. 11 Visualization in J48 tree

Fig. 12 shows the performance study of Algorithm. The **Naïve Bayes** is a statistical classifier which is independent among attributes. The advantage of using naive Bayes is that one can work with the naïve Bayes model without using any Bayesian methods.

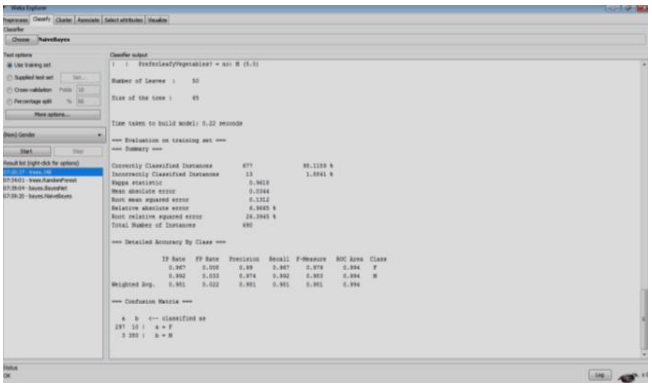

Fig. 12 Naïve Bayes

Fig 13 shows Threshold Curve with class female curve which ROC is 0.9877. ROC graphs are used to visualize the performance of the data, but there are some complications that arise when they are used in research. Roc curve is a graphical plot that explains the presentation of binary classifiers.

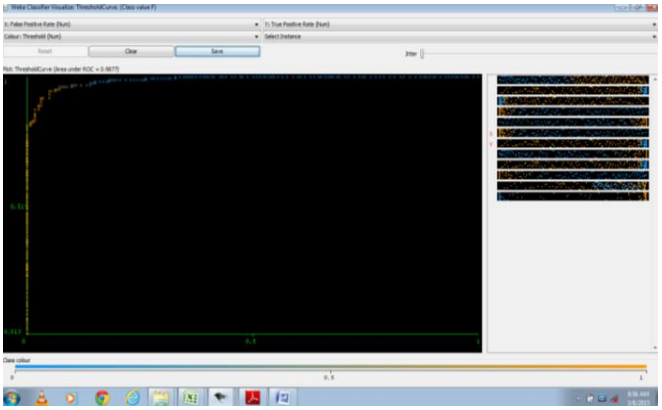

Fig.13 ROC curve

Fig. 14 shows the classification of the data sets using the K-Star algorithm. Here the time taken to obtain the accurate data is 0 sec. The correctly classified instances (699) it obtained 100%

accuracy. The incorrectly classified instances obtained 0% out of the total instances (690).

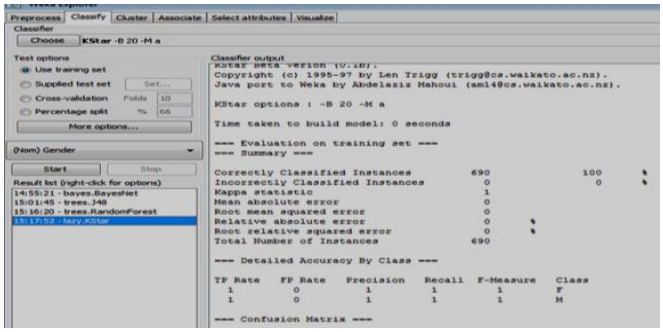

Fig. 14 K-Star Classification results

Table 1 shows the exactness that obtained while running the various algorithms in Weka. The observations that the best algorithms K-Star and Random Forest for this Dataset where Build the models are less time(0 sec and 0.6 sec) and the ROC values are 1(very accuracy).

Algorithm used	Time Taken	Correctly classified	ROC
AD tress	0.14sec	93.913	0.989
J48	0.12sec	98.5507	0.993
Kstar	0sec	100	1
Naivebaye	0.23sec	92046	0.987
NaiveBayesSimple	0.01sec	92.7536	0.988
NaiveBayesNet	0.14sec	94.2029	0.996
RandomForest	0.6sec	100	1

Table 1: Performance study of each algorithm

IV. CONCLUSION

Here in this research paper the analysis is obtained with the attributes like Age, Gender, Height, weight, Tea, coffee, food intake, alcohol consumption, and smoking are major factors/attributes in kidney diseases occurrences and survey has performed. The data sets which are collected from the various hospitals are processed through the data mining techniques tool such as Weka and Orange. Here the Machine learning algorithms such as AD Trees, J48, K star , Naïve Bayes, Random forest are used for the performance study of each algorithm which gives the Statistical analysis and predicting kidney diseases using the algorithms.

ACKNOWLEDGEMENTS

Authors would like to thank management and staff of Raghu Engineering College, Visakhapatnam, India for their kind support and providing lab facilities.

V. REFERENCES

[1] Velide Phani Kumar and Lakshmi Velide "A Data Mining approach for prediction and treatment of diabetes disease" IJSIT, 2014, 3(1),073-079.

[2] Bala, Suman, and Krishan Kumar. "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique." (2014).

[3] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of healthcare information management 19.2 (2011): 65.

[4] DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi "Statistical and data mining aspects on kidney stones: A systematic review and meta-analysis" Volume 1 , Issue 12 , 2012

[5] D. Pedro and M. Pazzani "On the optimality of the simple Bayesian classifier under zero-one loss". Machine Learning, 29:103–137, 1997.

[6] Richardson, Matthew, and Pedro Domingos. "Markov logic networks." Machine learning 62.1-2 (2006): 107-136.

[7] Ng, Raymond T., and Jiawei Han. "E cient and E ective Clustering Methods for Spatial Data Mining." Proc. of. 1994.

[8] Krishna, K., and M. Narasimha Murty. "Genetic K-means algorithm." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 29.3 (1999): 433-439.

[9] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[10] AhamedShaffeq, BM & Hareesha ..et all"Dynamic Clustering of Data with Modified Data with Modified k-means Algorithm" International Conference on Information and Computer Networks, ICICN-2012, pp. 221-225.

[11] Lior Rokach and Oded Maimon "The Data Mining and Knowledge Discovery Handbook", pages 321-352. Springer, 2005.

[12] Kaladhar, D. S. V. G. K., Krishna Apparao Rayavarapu, and Varahalarao Vadlapudi. "Open Access Scientific Reports." (2012).

[13] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7 (2002): 881-892.

[14] Kuo, Wen-Jia, et al. "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images." Breast cancer research and treatment 66.1 (2001): 51-57.

[15] Gaganjot Kaur, Amit Chhabra "Improved J48 Classification Algorithm for the Prediction of Diabetes" International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014