RESEARCH ARTICLE OPEN ACCESS

Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm

Krati Saxena[1], Dr. Zubair Khan[2], Shefali Singh[3]
M-Tech Research Scholar[1&3], Professor[2],
Department of Computer Science Engineering,
Invertis University, Bareilly-243123,
Lucknow, UP-India

ABSTRACT
Diabetes is one of the major global health problems. According to WHO 2011 report, around 346 million people worldwide are suffering from diabetes mellitus. Diabetes Mellitus is a metabolic disease where the improper management of blood glucose levels lead to the risk of many diseases like heart attack, kidney disease and renal failure. In Diabetes Mellitus, body does not properly use the insulin hormone secreted by Pancreas gland. There are so many computerized methods for the diagnosis of Diabetes Mellitus but the main drawback of these methods is that the patient has to undergo several medical tests to provide the input values to the computerized diagnostic system which proves to be very costly and time consuming. With the rapid advancement in the field of Artificial Intelligence, there are so many techniques and algorithms in A.I. that can be effectively used for the prediction and diagnosis of various diseases. These algorithms in artificial intelligence prove to be cost-effective and time saving for diabetic patients and doctors. In this paper, we are diagnosing Diabetes Mellitus using K- Nearest neighbour algorithm which is one of the most important techniques of A.I. The dataset is taken from www.stanford.edu/~hastie/Papers /LARS/diabetes.data.

Keywords:- Diabetes Mellitus, K- Nearest Neighbor algorithm, Diabetes Dataset.

I. INTRODUCTION

Nowadays Diabetes Mellitus has become a common global health problem which may lead to several health complications such as cardio vascular diseases, renal failure, visual impairment etc. According to W.H.O report, a total of 300 million of the world population will be affected by diabetes by 2025. Insulin is a natural hormone which is secreted by pancreas in the human body. Insulin converts sugar into simpler molecules which is utilized by the body cells to generate energy. This conversion is affected due to the lack of insulin and the sugar starts getting accumulated in the blood stream. As a result, blood glucose level starts increasing and the person develops Diabetes Mellitus. It is divided into two types Type I and Type II. Type I diabetes is usually diagnosed in children and here the body does not produce insulin. Type II is generally diagnosed in adults. Here, either the body does not produce enough insulin or cells ignored the insulin. Diabetes is a chronic metabolic disease in which pancreas gland in the body does not produce enough insulin or the body cannot effectively use the produced insulin. Uncontrolled Diabetes may lead to Hyperglycemia or Raised blood sugar. It may lead to serious damage to the nerves and blood vessels.

Nowadays data mining tools and techniques are widely used in almost every field like healthcare systems, marketing, weather forecasting, E business, retails etc. Healthcare System is one of the new emerging research areas where data mining techniques and tools can be effectively applied.

Our health care systems are rich in information but they are poor in knowledge so there is a large need of having techniques and tools for extracting the information from the huge data set so that medical diagnosis can be done. Data Mining is a process of semi automatically analyzing large databases to find useful patterns. Data mining attempts to discover rules and patterns from data as it deals with large volumes of data, stored primarily on disk. Data mining mainly deals with knowledge discovery in databases.

Diabetes Mellitus is a clinical syndrome which is characterized by hyperglycemia due to absolute or relative deficiency of insulin. The lack of insulin affects the metabolism of the body. It causes increase in the blood sugar level. Here the body does not produce enough insulin. Diabetes is a disease that affects more than 100 million people in the world. If Diabetes is not recognized at the right time and treated properly at an early stage of patients, it would affect people and may lead to various complications like visual impairment, cardio vascular disease, leg amputation and renal failure.

The major common symptoms of diabetes include frequent urination with large volume of urine (polyurea), Excessive thirst (polydipsia), Extreme Hunger (polyphagia), Unusual, sudden, unexplained weight loss, increased fatigue, Feeling very tired, Feeling ill, sudden vision changes or Blurry vision, Nausea, Vomiting, Stomach pain. There are some other symptoms too that include Delayed wound healing, Bruises that won‘t go away Sores that won‘t heal, Tingling or numbness in hands or legs, A feeling of pins and needles in feet, Unexplained aches and pains, Burning paining legs, feet, arms ,Fruity smell of breath and sweat etc.

II. K - NEAREST NEIGHBOR ALGORITHM

KNN is a method which is used for classifying objects based on closest training examples in the feature space. KNN is the most basic type of instance-based learning or lazy learning. It assumes all instances are points in n-dimensional space. A distance measure is needed to determine the "closeness" of instances. KNN classifies an instance by finding its nearest neighbors and picking the most popular class among the neighbors.

Features of KNN

a) All instances of the data correspond to the points in an n-dimensional Euclidean space

b) Classification is delayed till a new instance arrives

c) In KNN, the Classification is done by comparing feature vectors of the different points in a space region.

d) The target function may be discrete or real-valued.

An arbitrary instance is represented by($a_1(x)$, $a_2(x)$, $a_3(x)$,.., $a_n(x)$), where $a_i(x)$ denotes features. Euclidean distance between two instances d(x_i, x_j)=sqrt (sum for r=1 to n ($a_r(x_i)$ - $a_r(x_j)$)^2). The k-nearest neighbor algorithm is simplest of all machine learning algorithms and it is analytically tractable.

In KNN, the training samples are mainly described by n-dimensional numeric attributes. The training samples are stored in an n-dimensional space. When a test sample (unknown class label) is given, k-nearest neighbor classifier starts searching the 'k' training samples which are closest to the unknown sample or test sample. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, …. Pn) and Q (q1, q2,..qn) is defined by the following equation:-

$$d(P,Q)=\sum_{i=1}^{n}(P_i-Q_i)^2$$

The Simple KNN algorithm is:

Algorithm:-

I. Take a sample dataset of n columns and m rows named as R .In which $n-1^{th}$ columns are the input vector and n^{th} column is the output vector.

II. Take a test dataset of $n-1$ attributes and y rows named as P.

III. Find the Euclidean distance between every S and T by the help of formula $EculideanDistance =$

$$\sqrt{\sum_{i=1}^{y}\sum_{j=1}^{m}\sum_{l=1}^{n-1}\left(R_{(j,l)}-P_{(i,l)}\right)^2}$$

IV. Then, Decide a random value of K. $K$is the no. of nearest neighbors.

V. Then with the help of these minimum distance and Euclidean distance find out the n^{th} column of each.

VI. Find out the same output values.

If the values are same, then the patient is diabetic, otherwise not. After this, the accuracy rate and the error rate of the data set is being calculated. The accuracy rate shows that how many outputs of the test dataset are same as the output of the data of different features of the training dataset. The error rate is showing that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset.

KNN is a highly effective inductive inference method for noisy training data and complex target functions. Target function for a whole space may be described as a combination of less complex local approximations. In Knn Learning is very simple and Classification is time consuming.

KNN has many disadvantages like it has high Computation cost since it needs to compute the distance of each test instance to all training samples. It requires large memory proportional to the size of training set. It has Low accuracy rate in multidimensional data sets with irrelevant features. It can be used for both prediction and classification. It is highly adaptive to local information. KNN algorithm uses the closest data points for estimation; therefore it is able to take full advantage of local information and form highly nonlinear, highly adaptive decision boundaries for each data point.

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function. Sensitivity (also called the *true positive rate*, or the recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). Specificity (sometimes called the *true negative rate*) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition).

TP, TN, FP and FN signify the following:

- True positive: Those Sick people who are correctly diagnosed as sick
- False positive: The Healthy people who are incorrectly identified as sick
- True negative: The Healthy people who are correctly identified as healthy
- False negative: The Sick people who are incorrectly identified as healthy

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

- **Sensitivity or True Positive rate (TPR)**
  - TPR = TP / (TP+FN)

- **Specificity or True Negative rate (TNR)**
  - TNR = TN / (TN+FP)

- **False Positive rate**
  - FPR = FP / (FP+TN)

- **False Negative rate**
  - FNR = FN / (FN+TP)

- **Accuracy**
  - Accuracy=TN+TP/(TN+TP+FN+FP)
- **Precision or Positive predicted value**
  - PPV = TP / (TP+FP)
- **Negative Predicted value**
  - NPV = TN / (TN+FN)

III. RELATED WORK

Asha Gowda Karegowda , M.A. Jayaram, A.S. Manjunath (2012) [1], used cascading K mean ad K nearest neighbor algorithm for categorization of diabetic patients in their paper. They classified diabetic patients by proposing results using KNN and K mean. Accuracy achieved by the proposed system is 82%.

HardikManiya, Mosin I. Hasan, Komal P. Patel (2011) [2], have done the Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis. W. Yu, and W. Zhengguo(2007) [3], have proposed A Fast KNN algorithm for text categorization.

Asha Gowda Karegowda, MA.Jayaram(2007) [4], have used integrated decision trees and artificial neural network for categorization of diabetes data. The accuracy achieved using DT is 78.21% and using ANN is 72.88%.

“Siti Farhanah, Bt Jaafar and Dannawaty Mohd Ali” [5] forecasted diabetes mellitus using artificial neural network. They have used back propagation and supervised training method is used for prediction of the disease. The inputs to the system are plasma glucose concentration, blood pressure, triceps skin fold, serum insulin etc. In this paper they described recurrent Neural Network and Time series convolution Neural Network in Diabetes and compared to linear model and nonlinear compartment model.

Rajeeb Dey and Bajpai and Gagan Gandhi and Barnal iDey[6] in their paper " Application of Artificial Neural Network techniques for Diagnosis of Diabetes Mellitus"Artificial Neural Network has been used with parameters Random Blood Sugar test results , fasting blood sugar test result , post plasma blood sugar test, sex , age and their occupation. Performance achieved by using this system is 92.5%.

Y. Angeline Christobel, P.Sivaprakasam[7], proposed a New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset.

IV. DATASET AND ATTRIBUTES

The dataset we have used in our work is taken from - **www.stanford.edu/~hastie/Papers/LARS/diabetes.data.**

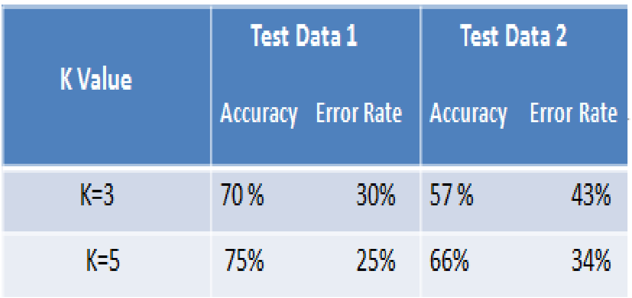
The dataset comprises of 11 attributes which are as follows:

1) Age (years)
2) Sex
3) Body mass index
4) Blood Pressure (mm Hg)
5) Plasma Glucose Concentration (Glucose tolerance test)
6) Triceps Skin fold
7) 2-Hour serum insulin
8) Diabetes Pedigree function
9) Cholesterol Level
10) Weight (kg)
11) Class variable (0 or 1)

V. PROPOSED WORK

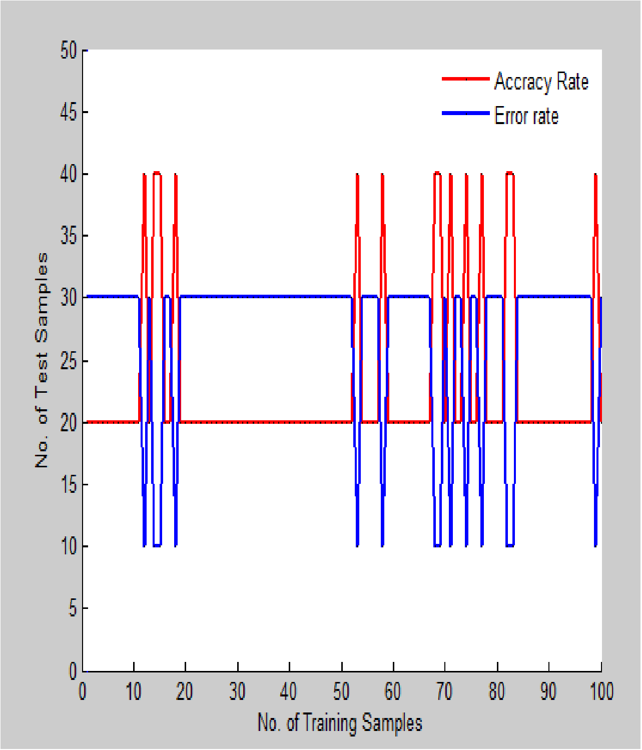
In our proposed work, we have taken one sample training dataset containing 100 rows and 11 columns of the above mentioned attributes. There are two test sample data containing 50 rows each. We have applied K- Nearest neighbor algorithm on the training and test sample data and obtained results for different values of K which is number of nearest neighbors. Accuracy and error rates have been calculated for K=3 and K= 5. The results have been evaluated using MATLAB.

VI. EXPERIMENTAL RESULTS

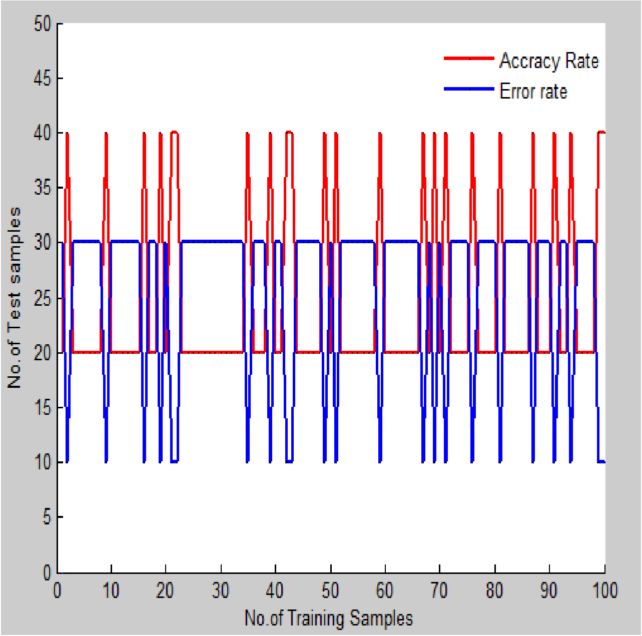


K Value	Test Data 1		Test Data 2	
	Accuracy	Error Rate	Accuracy	Error Rate
K=3	70 %	30%	57 %	43%
K=5	75%	25%	66%	34%

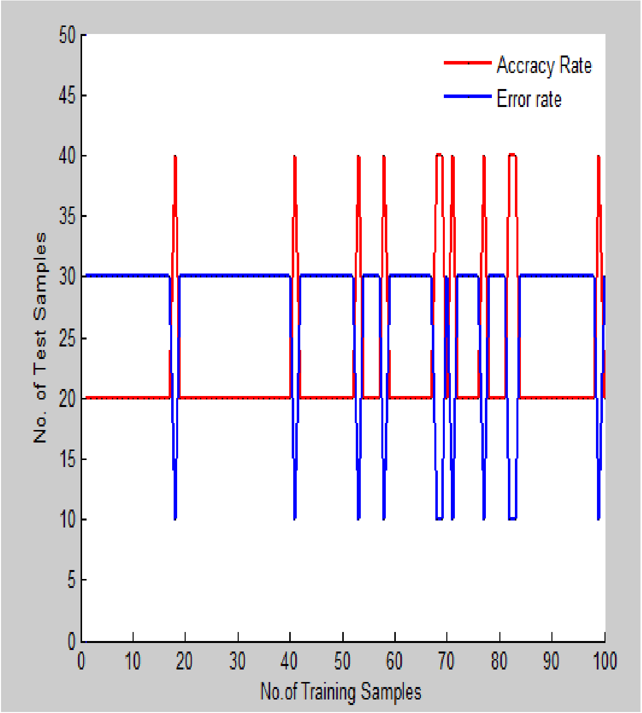
We have calculated accuracy and error rates for different values of k. The value accuracy and error rate increases as the value of K increases. Results are as follows:



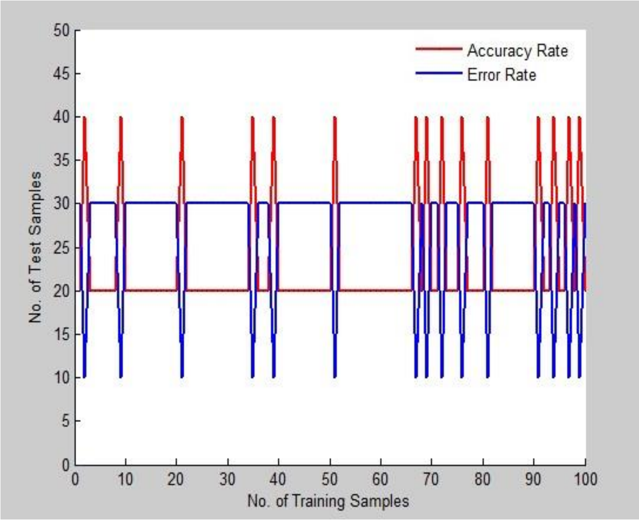
For K= 3 in Test data 1



For K=3 in Test data 2



For K=5 in Test data 1



For K=5 in Test data 2

In our work, we have developed a code for the diagnosis of Diabetes Mellitus in MATLAB using KNN. With this code we have evaluated True positive, true negative, false positive and false negative rates. Positive and Negative precision values and accuracies have also been calculated.

The results have been evaluated for different values of K which is the number of nearest neighbors. We have calculated the values for K=3 and K=5.

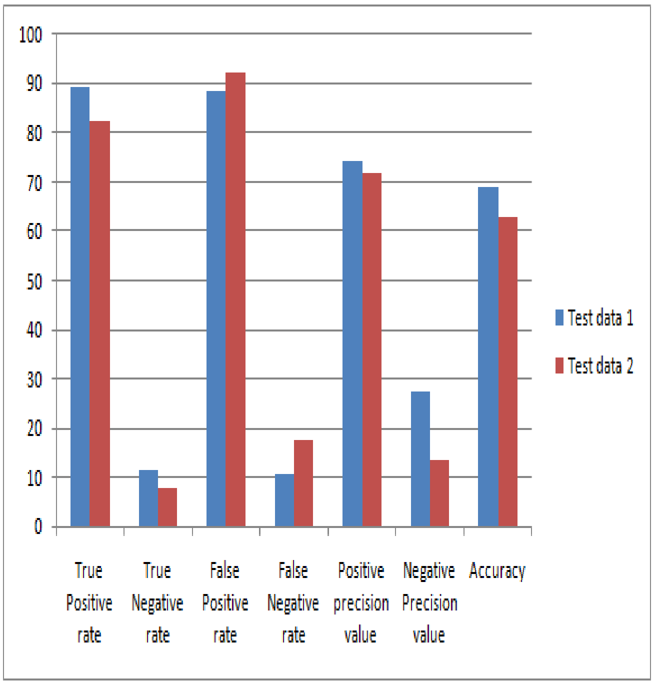
After the execution of the MATLAB Code developed for the diagnosis of diabetes mellitus, the following results have been generated.

For K=5,

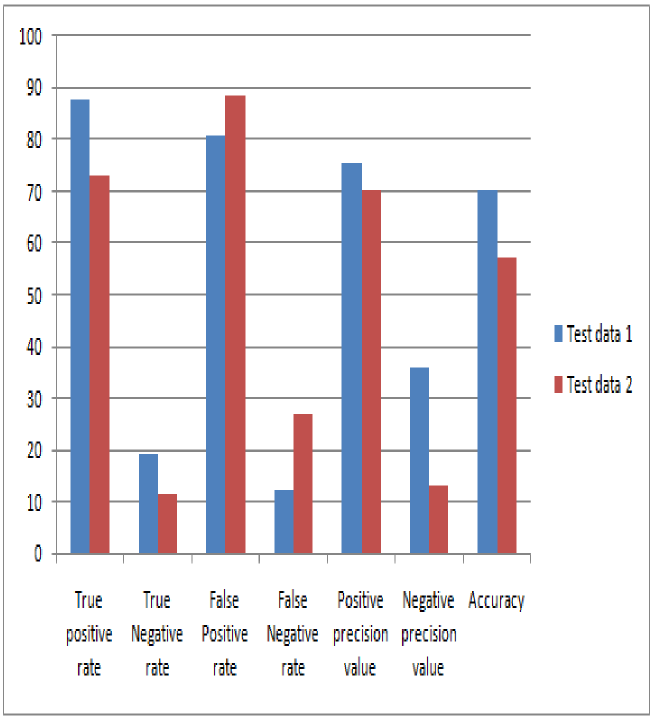
	Test Data 1	Test Data 2
True positive Rate	89.1892%	82.4324%
True negative Rate	11.5385%	7.6923%
False positive Rate	88.4615%	92.3077%
False negative Rate	10.8108%	17.5676%
Positive Precision Value	74.1573%	71.7647%
Negative Precision Value	27.2727%	13.3333%
Accuracy	69%	63%

For K=3,

	Test Data 1	Test Data 2
True positive Rate	87.8378%	72.9730%
True negative Rate	19.2308%	11.5385%
False positive Rate	80.7692%	88.4615%
False negative Rate	12.1622%	27.0270%
Positive Precision Value	75.5814%	70.1299%
Negative Precision Value	35.7143%	13.0435%
Accuracy	70%	57%



Graph showing results for test data 1 & test data 2 for K = 5



Graph showing results for test data 1 and test data 2 for K = 3

VII. CONCLUSION

Here we have used K- nearest neighbor algorithm for the diagnosis of diabetes mellitus. We have calculated accuracy and error rates for K=3, 5. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the training dataset. The error rate is showing that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset. The result is showing that as the value of k increases, accuracy rate and error rate will also increase. KNN is one of the most effective Artificial intelligence algorithms that is widely used for diagnostic purposes. More accurate and efficient results can be obtained through KNN.

VIII. FUTURE WORK

Different results can also be obtained using other AI Algorithms like K mean clustering, decision trees, ANN etc. Different datasets can be used with different techniques for calculating accuracies and error rates. Future works may address hybrid classification models using KNN with other techniques of AI. Simulations can be done using different tools other than MATLAB like WEKA etc. to get better and more accurate results.

REFERENCES

[1] Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath(2012) 'Cascading K-means Clustering and K-NearestNeighbor Classifier for Categorization of Diabetic Patients' IJEAT Vol.1 No.3 pp 147-151

[2] Hardik Maniya, Mosin I. Hasan, Komal P.Patel (2011) Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis, IJCA pp
22-26

[3] W. Yu, and W. Zhengguo (2007), A Fast KNN algorithm for text categorization, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp.3436-3441

[4] Asha Gowda Karegowda ,MA.Jayaram(2007) 'Integrating Decision Tree and ANN for Categorization of Diabetics Data' International Conference on Computer Aided Engineering, December 13-15, IIT Madras, Chennai, India.

[5] Siti Farhanah Bt Jaafar and DannawatyMohdAli, "Diabetes mellitus forecast using artificial neural networks", Asian conference of paramedical research proceedings, 5-7, September, 2005, Kuala Lumpur, MALAYSIA

[6] Rajeeb Dey and Vaibhav Bajpai and Gagan Gandhi and Barnali Dey, "Application of artificial neural network technique for diagnosing diabetes mellitus", 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10.

[7] Y. Angeline Christobel, P.Sivaprakasam, "A New Classwise k nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset".

[8] "Forecast of Diabetes using Modified Radial basis Functional Neural Networks" International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887) ,G. Magudeeswaran and D. Suganyadevi, Sreesaraswathi Thyagaraja College Pollachi-642 107, Tamil Nadu

[9] "Diagnosis of Diabetes Mellitus based on Risk Factors" International Journal of Computer Applications (0975 – 8887) Volume 10– No.4, November 2010

[10] Nahla H. Barakat, Andrew P. Bradley, and Mohamed Nabil H. Barakat,"Intelligible support vector machines for diagnosis of diabetes mellitus", IEEE transaction on information technology in Biomedicine, Vol. 14, No. 4, July 2010

[11] Mu Zhu, Wenhong Chen, John P. Hirdes, Paul Stolee (2007) 'The K-nearest neighbor algorithm predicted rehabilitationpotential better than current Clinical Assessment Protocol' Elsevier Vol.60 pp-1015-1021.