PDD: Predictive Diabetes Diagnosis using Datamining Algorithms

M. S. Geetha Devasena
Department of Computer Science and Engineering
Sri Ramakrishna Engineering College
Coimbatore, India
msgeetha@srec.ac.in

R. Kingsy Grace
Department of Computer Science and Engineering
Sri Ramakrishna Engineering College
Coimbatore, India
kingsygrace.r@srec.ac.in

G. Gopu
Department of Biomedical Engineering
Sri Ramakrishna Engineering College
Coimbatore, India
hod-bme@srec.ac.in

Abstract— **Data analytics is used to obtain the useful insights from small or large data set to conclude some useful information and also used for future recommendation and decision making. Predictive Analytics uses the data mining, machine learning techniques to make predictions about future. It involves analysis of available data. The predictive analytics in health care is primarily used to determine patients having initial stages of diabetes, asthma, heart disease and other critical lifetime disease. The proposed method PDD uses data mining algorithms to predict the type2 diabetes. The data mining algorithms used in the proposed system are K-Means Clustering and Random Forest. The predictive model, PDD provides better results in terms of accuracy when compared to hierarchical clustering and Bayesian network clustering with random forest prediction.**

Keywords— Type 2 diabetes, Prediction, K-Means Clustering, Random Forest

I. Introduction

For long years data mining is used for searching knowledge from the data [1, 2, 3, 4]. Diabetes is not like other diseases, it prolong to the patients and also causes other diseases. Insulin secretion problem in pancreas is called diabetes. Two forms of diabetes also exist. These are groups 1 and 2[5]. Type 1 diabetes' other name is called insulin-dependent, or childhood-onset diabetes. This is because of less insulin secretions in the body.

The ineffective use of insulin in the body triggers Type 2 diabetes which is also other way known as adult-onset diabetes and also as non-insulin-dependent diabetes. Also produced by less physical activity and increased body weight. Diabetes is one of the biggest increasing challenges in India. The diabetic population is increased by 8.7% in the age group of 20 to 70. The non-communicable disease, diabetes causes other diseases and the growing incidence of diabetes is influenced by a combination of factors such as rapid urbanization, sedentary lifestyles, unhealthy diets, tobacco use and increased life expectancy.

According to World Health Organization (WHO) the number of people in diabetes will double in next decade. In India, The diabetics in India is estimated as 31,705,000 at present and it will reach 79,441,000 in the year 2030. The WHO assessment in the year 2013 showed that 63 million diabetic. According to International Diabetes Federation Atlas in the year 2015, 69.2 million Indians are diabetic. Diabetes is rapidly gaining the status of a potential disease in India and diabetes prevalence is expected to double globally. The increase in diabetic is from 171 million in 2000 to 366 million in 2030 and the maximum increase rate is evident in India [6]. Worldwide currently 143 million people suffer from this major disease diabetes mellitus. This number alarmingly shows rapid growth. Five percent of India's population suffers from diabetes mellitus. This disease is controlled and managed by timely detection and periodic health check-ups at regular intervals.

Diabetes is most prevailing disease and both men and women are getting affected. The upper middle class people are affected more than the lower middle class people. The Figure 1 depicts the diabetes prevalence in India [7] and Figure. 2 Samples the prevalence of diabetes and prediabetes in 15 Indian states[8].More Research is required to predict the diabetic in developing countries.

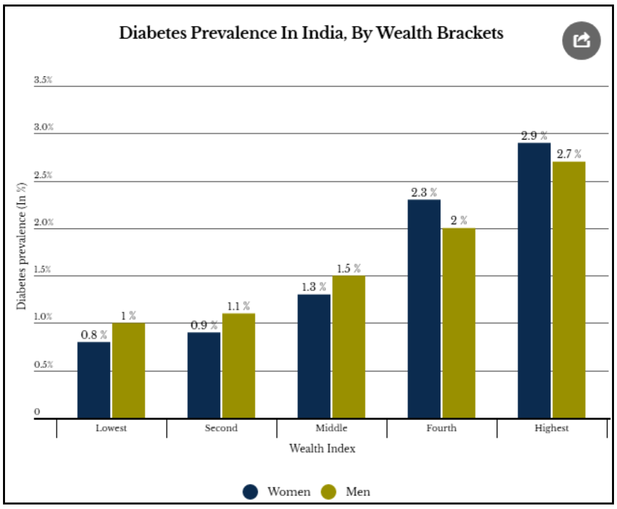


Fig. 1. Diabetes Prevalence in India [7]

Sadri Sa'di et al. implemented various algorithms for diagnosis of type 2 diabetes such as Naive Bayes, J48 and RBF Network using Weka tool [9]. The performance comparison showed that Naive Bayes gave accuracy rate of

978-1-7281-4514-3/20/$31.00 ©2020 IEEE

76.95% and performed better than RBF and J48. Daniah Almadn and Abdolreza Abhari have compared type 2 diabetes diagnosis using classification models.

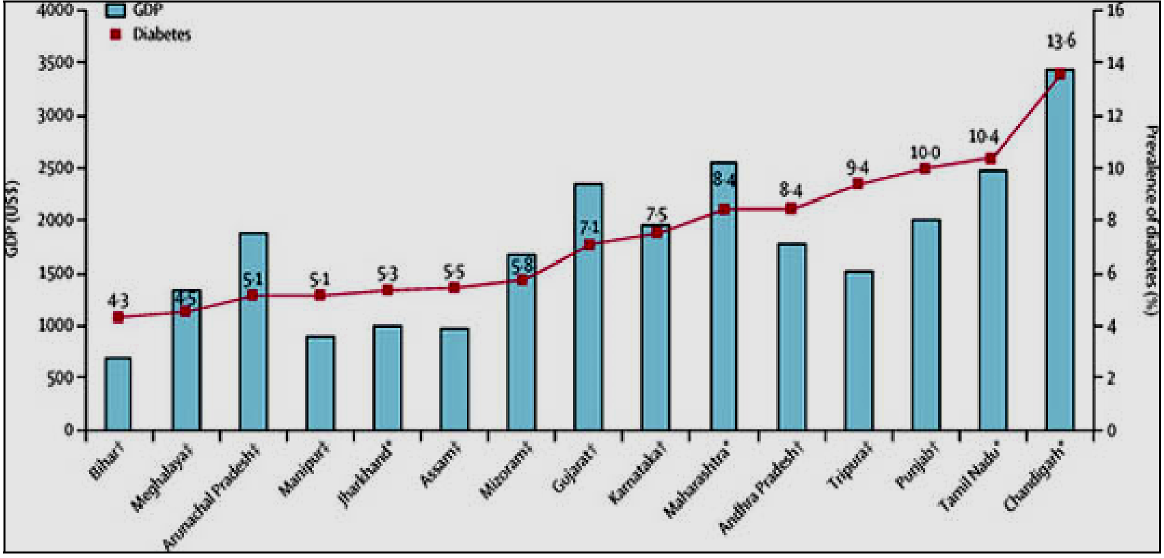


Fig. 2 Prevalence of diabetes and prediabetes in 15 states of India [8]

This paper implemented classification algorithms for diagnosis of type 2 diabetes and also compared the performance of the classification algorithms such as Logistic Regression and Support Vector Machine. The authors have proposed Fuzzy Expert System (FES) along with Fuzzy Inference System (FIS) to predict the type 2 diabetes. The comparisons of all the algorithms are done using Weka and MATLAB [10].

Mahmoud Heydari et al. have compared support vector machine, decision tree, 5-nearest neighbor, and Bayesian network classification algorithms for the diagnosis of type 2 diabetes in Iran. 95.03 % accuracy in 5-nearest neighbor algorithm [11] proves that the data mining algorithms provides better results in type 2 diabetes diagnosis. Angus G. Jones et al. proposed a detailed study [12] on type 2 diabetes and the effect prediction on b-Cell Failure. The study is carried out with 620 type 2 diabetes patients.

The HbA1c ‡58 mmol/mol (7.5%) is there in patients and they are assessed the therapy for 6 months. The effect of b-cell failure and the corresponding glycemic response is evaluated. Linear regression is used for prediction and the ANOVA tool is used for implementation. The research concluded that GLP-1RA therapy is suited for diabetes patient with b-Cell Failure.

II. PDD: Predictive Diabetes Diagnosis

The proposed method uses K-means clustering algorithm for clustering the input dataset and random forest algorithm for predicting the type 2 diabetes from the input parameters. K-Means is a simplest unsupervised learning algorithms which solve the well-known problem of clustering.

Using the K-Means algorithm the clustering and other data mining problems are easily resolved. It is the most commonly used algorithm for different applications including vector quantization, density estimation and characterization of workload behavior. This clustering algorithm is widely used algorithm and iterative in nature, hence called as Lloyd’s algorithm. K-Means is usually performed better with other learning algorithms.

Considering the initial set of cluster centers, criterion to converge this algorithm proceeds to find similar data points and underlying patterns.

The computational speed of k-means algorithm is based on the amount of data, cluster center computation and the requirement of number of iterations to converge. K-means is simple in nature. The Figure 3 depicts the fitness function with the various parameters such as number of clusters, number of cases and distance function. The Process Flow of K-Means Algorithm is shown in Figure 4.

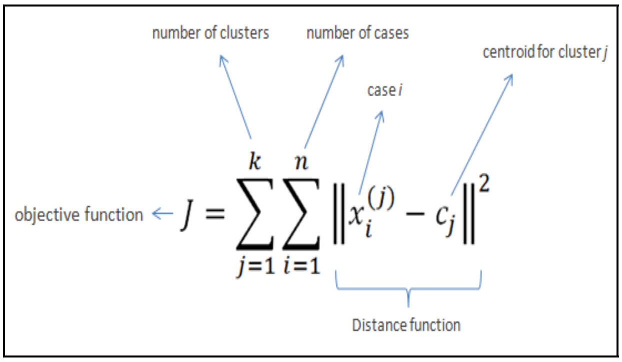

number of clusters
number of cases
centroid for cluster j
case i
objective function ← $J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$
Distance function

Fig. 3 Fitness Function of K-Means

K-Means Algorithm

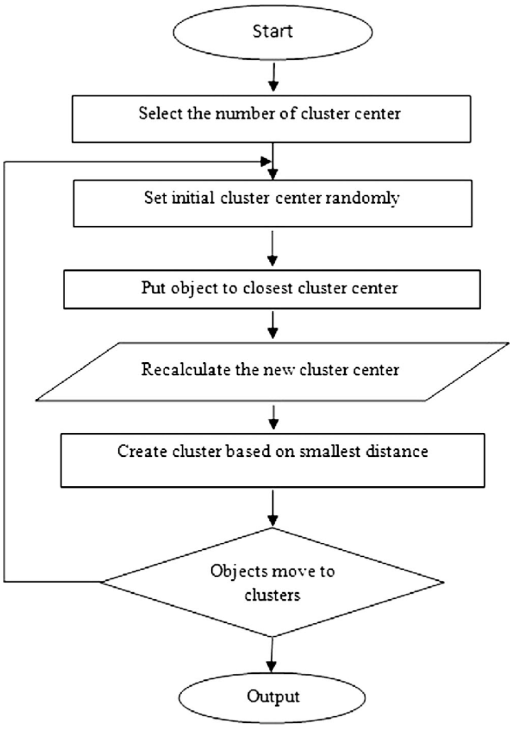
Step 1: Selects K centroids randomly.

Step 2: Identify the closest centroid and assign respective data points.

Step 3: Calculate average of cluster data points to act as centroid.

Step 4: Assign data points to centroids closest to them.

Step 5: Continue phases 3 and 4 until the results are reassigned or the number of iterations reached is full.


Start
Select the number of cluster center
Set initial cluster center randomly
Put object to closest cluster center
Recalculate the new cluster center
Create cluster based on smallest distance
Objects move to clusters
Output

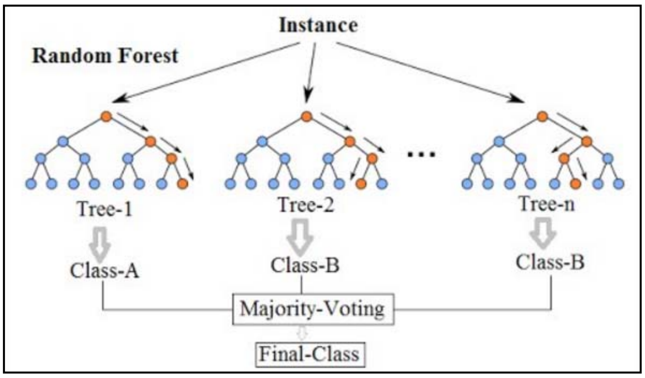


Fig. 4 Process Flow of K-Means Clustering Algorithm

Fig. 5 Process Flow of Random Forest.

R uses an effective algorithm by Hartigan and Wong (1976) to divide the observations into k groups, so that the number of squares of the observations to their cluster centers is a minimum. In step 2 and step 4, at each measurement, the smallest value of s is assigned to the cluster. Clustering of K-means can accommodate data sets larger than approaches to hierarchical clusters. In fact, findings aren't assigned to a cluster indefinitely. Using means thus means that all variables must be constant and outliers affect the approach.

Random forest algorithm was first suggested by Bell Labs ' Tin Kam Ho in 1995, as a classification and regression learning system for ensemble. Leo and Adele extend this algorithm further during 2001. The key idea is to create a higher percentage of decision trees (base learners) and the purpose is to reduce the association of errors between classifiers using a random selection of features to be split into each node.. At the training stage, the lot of decision trees are created and for each individual tree the mean prediction is found.

The common problem in Random forest decision trees is the training set tends to over fit. From Figure 5, it is clear that random forest algorithm combines individual tree and predicted the final class.

When the trees are built for each pair of cases proximities are need to be computed. The proximity is incremented by a factor of one if the same terminal node is occupied by two cases. Then normalization of proximities is done by dividing with the tree count. The missing data replacement, outliers identification and low-dimensional views illumination are done with proximities.

III. Results and Discussion

The Diabetes Dataset [13] is used from UCI repository. The dataset consists of features are shown in Table 1. These nine features are used for diagnosis of type 2 diabetes. Table 1 contains nine features, out of these nine features the input features are first eight and the ninth feature is the output feature. The presence of diabetes is indicated by Sick value equal to 1 (one) and the absence of diabetes is indicated by Sick value 0 (zero).

The K-means algorithm is executer to cluster and the prediction is done using random forest algorithm. The prediction accuracy of random forest is depicted in Figure 6 with other clustering algorithms such as hierarchical clustering and bayesian network.

TABLE 1. Features used for Type II Diabetes Diagnosis

S. No.	Feature	Description and Feature Values
1.	Number of times Sick	Numerical Values
2.	Concentration of plasma-glucose	Numerical Values
3.	Pressure on Diastolic Blood	Numerical Values in (mm Hg)
4.	Thickness of Triceps Skin Fold	Numerical Values in mm
5.	2-Hour Insulin Serum	Numerical Values in (mu U/ml)
6.	The Index of Body Mass (BMI)	Numerical Values in (weight in kg/ height in m2
7.	Pedigree Diabetes Feature (DPF)	Numerical Values
8.	Age	Numerical Values
9.	Stage two diabetes disorder	Sick=1, Normal=0

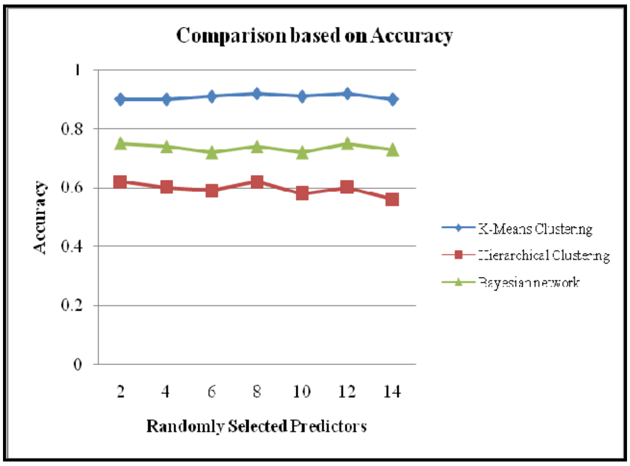


Fig. 6 Comparison of Prediction Accuracy

IV. Conclusion

Diabetes is the most common dangerous disease that can lead to additional problems such as heart attack, stroke, blindness, nerve damage, kidney failure, disease of the blood vessels and sexual disempowerment. The proposed method using K-Means and random forest provides better accuracy for predicting type 2 diabetes. To enhance the proposed method the fuzzy system and deep learning method could also be used.

Acknowledgment

The authors gratefully acknowledge the use of facilities at Sri Ramakrishna Engineering College, Coimbatore, India. Also like to convey their sincere thanks to the Principal and the Management for their support and guidance.

References

[1] F.S. GHAREHCHOPOGH,"Approach and Review of User Oriented Interactive Data Mining", 4th International Conference on Application of Information and Communication Technologies (AICT2010), Digital Object Identifier: 10.1109/ICAICT.2010.5611792, IEEE, Tashkent, Uzbekistan, pp.1-4, 12-14 October 2010.

[2] C. Olaru, L. Wehenkel, "Data Mining" , in IEEE Computer Applications in Power, Vol. 12, no. 3, pp. 19-25, July 1999.

[3] M. Chen, J. Han, P. S. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and data Engineering, pp. 866-883, 1996.

[4] Q. Luo, "Advancing knowledge discovery and data mining", 1st international Workshop on Knowledge discovery and data mining (WKDD'08), Adelaide, South Australia, pp. 3-5, 2008.

[5] Diagnosis and Classification of Diabetes Mellitus - NCBI – NIH, by American Diabetes Association - 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797383/, DIABETES CARE, VOLUME 33, SUPPLEMENT 1, JANUARY 2010

[6] Seema Abhijeet Kaveeshwar, Jon Cornwall , The current state of diabetes mellitus in India, Australas Med J. 2014; 7(1): 45–48, doi: 10.4066/AMJ.2013.1979

[7] Prevalence of diabetes and prediabetes in 15 states of India: http://www.thelancet.com/journals/landia/article/PIIS2213-8587(17)30174-2/fulltext?elsca1=tlpr

[8] Diabetes Prevalence in India: https://www.indiaspend.com/

[9] Sadri Sa'di, Amanj Maleki, Ramin Hashemi, Zahra Panbechi, Kamal Chalabi, " COMPARISON OF DATA MINING ALGORITHMS IN THE DIAGNOSIS OF TYPE II DIABETES", International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015

[10] Daniah Almadn, Abdolreza Abhari, "Comparative analysis of classification models in diagnosis of type 2 diabetes", 2016 Society for Modeling & Simulation International (SCS) SpringSim-MSM, 2016 April 3-6, Pasadena, CA, USA

[11] Mahmoud Heydari, Mehdi Teimouri, Zainabolhoda Heshmati, Seyed Mohammad Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran", International Journal of Diabetes in Developing Countries, June 2016, Volume 36, Issue 2, pp 167-173

[12] Angus G. Jones, Timothy J. McDonald, Beverley M. Shields, Anita V. Hill, Christopher J. Hyde, Bridget A. Knight, and Andrew T. Hattersley, "Markers of b-Cell Failure Predict Poor Glycemic Response to GLP-1 Receptor Agonist Therapy in Type 2 Diabetes" Diabetes Care Publish Ahead of Print, published online August 4, 2015, DOI: 10.2337/dc15-0258

[13] UCI Machine Learning Repository: Data Sets, http://mlr.cs.umass.edu/ml/datasets/