Prediction Of Diabetes And Clustering Based On Its Levels Using Fuzzy C Means Algorithm

S. Jamuna, Dr. K. Mohan Kumar

Abstract: Diabetes forecast framework is exceptionally valuable framework within the healthcare field. Clustering is making set of things having a place to the same sets of comparable and diverse sets of divergent. Applications of clustering includes in numerous areas like pharmaceutical, showcasing, fund, www etc. In this work diabetic patients' information are taken from authenticated web resources. Their resulting blood test reports are given as input to the proposed frame work to find the three levels of diabetes like no diabetes, pre diabetes and diabetes using various computations. Also, the diabetes dataset are grouped using Fuzzy C-Means algorithm. This grouping task is very much useful for the medical practitioner to give effective treatment

Keywords: Cluster analysis, clustering, Pima Indian Diabetes, Fuzzy C-Means, Diabetes and Prediction

1 INTRODUCTION

Diabetes is one of the illnesses that are spreading like plagues within the whole world. It is seen that in each era diabetes affects people from children, young people and old people from ancient age. Pro-long diabetes can cause failure of organs such as liver, kidneys, heart, and stomach. Also it is related with the disorders of retinopathy and neuropathy. Diabetes mellitus could be a disorder characterized by metabolic clutter and unusual rise within the concentration of blood sugar [1]. Diabetes occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels [2]. The number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014. Diabetes prevalence has been rising more rapidly in middle- and low-income countries. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths [3]. Diabetic persons should suffer many health related problems if they do not maintain their glucose level. The doctors prescribe tablet with more care based on their type of diabetes. At present they write the medicines using their experience. If they use assisting tool with machine learning algorithms for their analysis definitely that will increase the accuracy in diagnosis. Machine learning algorithms are used for this purpose. Machine learning is a field of computer science which gives computers an ability to learn without being explicitly programmed. Machine learning is used in a variety of computational tasks where designing and programming explicit algorithms with good performance is not easy.

- *Mrs. S. Jamuna is currently pursuing Ph.D as a Part Time Research Scholar in PG & Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.PH+91- 9976350354. E-mail: sjamunacs@gmail.com*
- *Dr. K. Mohan Kumar is currently working as Head Of the Department , PG & Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India. PH-+91 9443805042. E-mail: tnjmohankumar@gmail.com*

Fuzzy C Means Clustering algorithm is one of the better algorithms for clustering [4]. So in this paper the level diabetes are analyzed and clustered using machine algorithm FCM. So the main objectives of this work are finding the level of diabetes in a diabetic person and provide a to software tool to assist doctors to diagnose their patients through clustering method.

2 RELATED WORKS

Various works have been done for diabetes determination by utilizing data mining procedures. R.Nithya et al.,(2015) focused on three clustering algorithms namely Hierarchical clustering, Density based clustering and K-means clustering algorithm for grouping the data instances into subsets on their similarity. Diabetes dataset was taken for the analysis of comparison of the algorithms on the basis of their class attributes the number of clustered instances and the execution time taken [5]. Dr. K. Mohan Kumar, S. Jamuna (2018), compared the advantages and Disadvantages of various prediction algorithms for knowledge discovery [6]. Zeynel Cebeci and Figen Yildiz (2015) compared K-means(KM) and Fuzzy C-means(FCM) algorithms on the basis of their computing performance and clustering accuracy for the given data set and concluded that no algorithm is the best for all cases only the dataset selection determined the suitable algorithm and 2D and/or 3D scatter plots of datasets gave better result for understanding the structure of clusters in datasets[7]. Amatul et al.(2013) reviewed the data mining applications used specially on an open source diabetes dataset. They emphasized the mining patterns of independent and dependant variables in the dataset and compared the non-processed and pre-processed data for classification accuracy and found that the pre-processed data gave the better accuracy result. So that pre-processing was the vital point in the data mining technique [8]. Jayaram et al. (2012) used K-means clustering and K- Nearest Neighbor classifier(KNN) with Genetic Algorithm (GA) and Correlation based Feature Selection (CFS) for classifying Pima Indian Diabetic Database in three stages where K-means clustering was utilized to recognize and omit the incorrectly classified data in the primary stage. In the secondary stage, Genetic Algorithm and correlation based feature selection were used to extract the data as input for the next level. In the final stage, an enhanced classification of PIDD was done using K-Nearest Neighbor classifier with the accuracy of 96.68%[9]. Vijayalakshmi et al.(2012) developed a clustering algorithm

based on b-Coloring technique to classify PIDD. They implemented and performed experiments by comparing their approach with KNN classification and K-means clustering. The results showed that the clustering based on coloring technique is much way better than other clustering approaches in terms of exactness and purity. It is used to evaluate the quality of cluster [10]. Padmaja.P et al.,(2008) used four kinds of algorithms k-means, Partitioning Around Medoids (PAM),Minimum Spanning Tree (MST) and Nearest Neighbor to generate five types clusters which help to estimate the range of people especially women with particular characteristics affected by diabetes and the stage in which they were that is whether in early stage or advanced stage. The results were analysed using a graph to find out the best algorithm among the four and they found that the Partitioning Around Medodis(PAM) algorithm produced quality clusters[11]. Han et al. (2008) utilized information mining strategies through Fast Mineworker for the classification of diabetes information investigation and diabetes prediction model. A Choice tree and ID3 calculation were utilized for expectation with 72% and 80% of exactness individually [12].

3. METHODOLOGY

The clinical diabetic dataset is collected from https://data.world/data-society/pima-indians-diabetes-database. This dataset deals with 1050 patients and their therapeutic conditions. The following Table 1 shows the various parameters which are related to diabetic person. The Table 2 shows the values of these parameters of those patients.

TABLE 1
DIABETES PARAMETERS

No.	Name
1.	Patient's ID
2.	Gender (G)
3.	Age
4.	Blood pressure
5.	Glucose (GLS)
6.	Skin thickness
7.	Serum insulin
8.	Body mass index (weight/(height)2
9.	Pregnancies

3.1 PROPOSED MODEL

The proposed model has been implemented using Python code and run in Microsoft Windows environment. The parameter Glucose (GLS) plays a vital role to determine the diabetes. So, in this tool it is used to predict the diabetic persons. The diabetes forecast framework model shown in the following Figure 1, finds the three stages of diabetes. The main objective of the proposed framework is finding the stage of diabetes using the Glucose level in blood and clustering the patients based on all the parameters in the diabetes dataset.

TABLE 2
DIABETES DATASET

Patients ID	Gender	Age	Blood Pressure	Glucose	Skin Thickness	Serum insulin	BMI	Pregnancies
1	M	50	72	148	35	0	33.6	6
2	F	31	66	85	29	0	26.6	1
3	M	32	64	183	0	0	23.3	8
4	M	21	66	89	23	94	28.1	1
5	F	33	40	137	35	168	43.1	0
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
.		.	.	.	.	.	.	.
1046	F	30	74	116	0	0	25.6	5
1047	F	26	50	78	32	88	31	3
1048	M	29	0	115	0	0	35.3	10
1049	M	53	70	197	45	543	30.5	2
1050	F	54	96	125	0	0	0	8

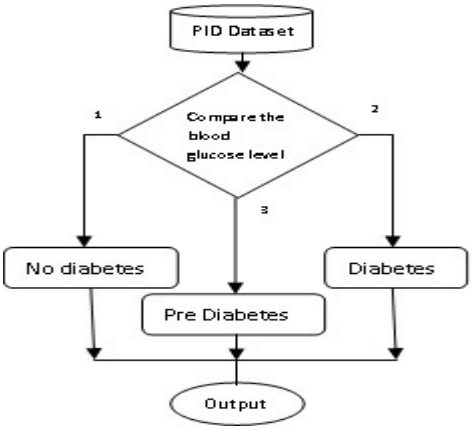
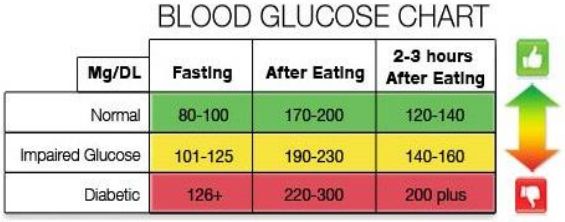


Figure 1: *Diabetic forecast framework Model*

In this model comparison of blood glucose level is done by using the conditions mentioned in the following Table 3.

TABLE 3
Blood Glucose chart [13]



3.1 FUZZY C-MEANS CLUSTERING ALGORITHM

FCM could be a information clustering calculation in which each information point belongs to a cluster to a degree indicated by a participation grade. FCM utilizes fluffy partitioning such that a given information point can have a place to a few bunches with the degree of belongingness indicated by enrollment grades between and 1. However, FCM still employments a fetched function that's to be minimized whereas attempting to partition the information set [14]. The participation lattice U is permitted to have components with values between 0 and 1. In any case, the summation of degrees of belongingness of a information point to all clusters is continuously break even with to unity:

$$\sum_{i=1}^{c} Uij = 1 \text{for all} j = 1, \dots, n. \quad \text{---------- (1)}$$

The cost function for FCM is

$$J(U,c1\dots..Cc)=\sum_{i=1}^{c} Ji \sum_{i=1}^{c} 0 \sum_{j=1}^{n} U \frac{m}{ij} d \frac{2}{ij} \quad \text{----------(2)}$$

Where uij is between 0 and 1;
ci is the cluster center of fuzzy group i; and d ij =||Ci-Xj|| is the Euclidean distance between the ith cluster center and the j^{th} data point; and m ∈[1,∞] is a weighting exponent. The necessary conditions for Equation to reach its minimum are:

$$c=\frac{\sum_{j=1}^{n} u \frac{m}{ij} x j}{\sum_{j=1}^{n} u \frac{m}{ij}} \quad \text{---------- (3)}$$

And

$$Uij=\frac{1}{\sum_{k=1}^{c} \left(\frac{dij}{dkj}\right)} \frac{2}{(m-1)} \quad \text{------- (4)}$$

The calculation works iteratively through the going before two conditions until no more enhancements are taken note. In a bunch mode operation, FCM decides the cluster centers ci and the enrollment framework U using the taking after steps:

Step 1: Initialize the relationship matrix U with arbitrary values between 0 and 1 such that the constraints in Equation (1) are satisfied.
Step 2: Compute c fuzzy cluster centers; ci, i = 1, …, c, using Equation (3).
Step 3: Compute the cost function according to Equation (2). Stop if either it is below a certain acceptance value or its improvement over previous iterationis below a certain threshold.
Step 4: Compute a new U using Equation (4).
Go to step 2[15, 16].

4 RESULT AND DISCUSSION

The following Figures 2 to 6 shows the screen shots of various windows of the developed software tool.

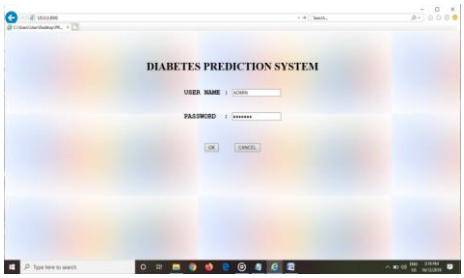


Figure 2: *Login window*

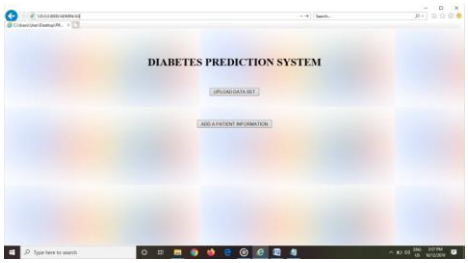


Figure 3: *Data Upload option window*

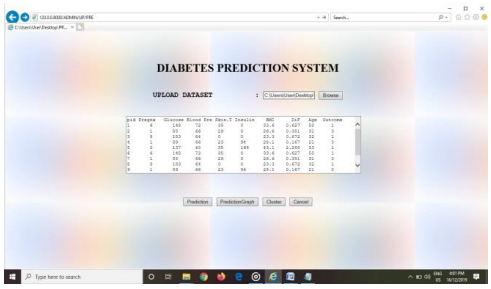


Figure 4: *Upload a Dataset window*

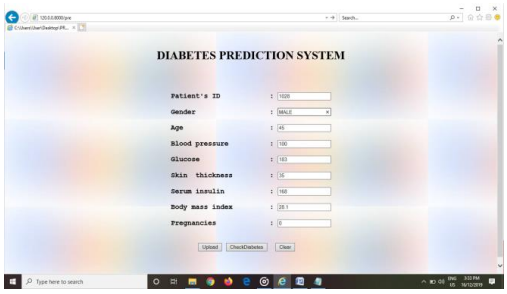


Figure 5: *Upload a Patient's information Window*

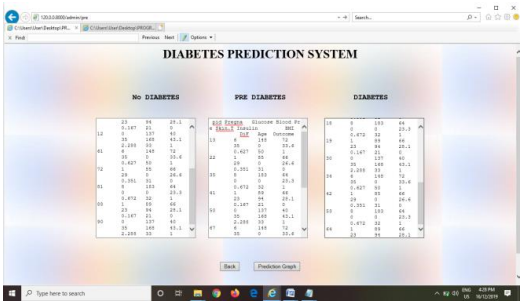


Figure 6: *Window for the Classification of Patients*

The developed software tool produced the following graphical

output for the given diabetes dataset shown in Figure 7.

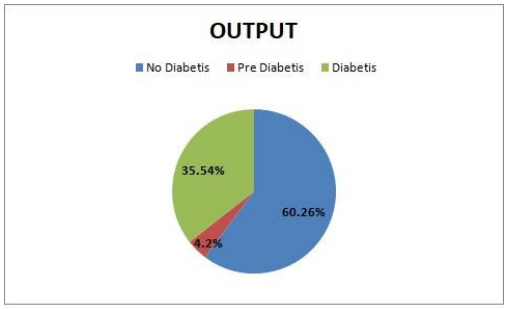


Figure 7: *Overall diabetes patients' Percentage*

The above Figure 7 clearly shows 60.26% of people are in no diabetes level, 4.2% of people are in pre diabetes level and 35.54% of people in diabetes level. The software tool also produce the following graphical output after applying FCM algorithm for the given diabetes dataset shown in Figure 8.

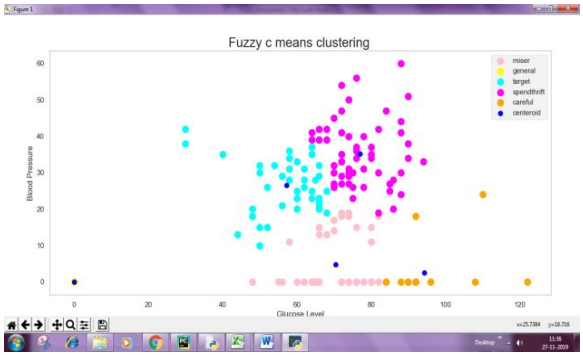


Figure 8: *Fuzzy clustering points*

In the above Figure 8, the purple dots represent no diabetes, the orange dots represent pre diabetes, and the light blue dots represent diabetes. This clustering analysis can be used by the medical practitioner to give effective treatment.

5. CONCLUSION AND FUTURE WORK

This work clearly explains how to predict the levels of diabetes in an effective methodology for the given input dataset using a software tool. Also it clusters the input dataset using FCM algorithm to foresee the patients with diabetic malady. It helps to the restorative specialists, medical students and medical attendants to analyze diabetes infection. In near future this tool will be improved by showing the previous treatment history of patients.

6 REFERENCES

[1] Yihong Donga, Yueting Zhuanga, Ken Chenc, Xiaoying Taib, "A hierarchical clustering algorithm based on fuzzy graph connectedness", Fuzzy Sets and Systems, Vol. 157 (13), 2006, pp. 1760–1774.

[2] https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio et al. "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies". Emerging Risk Factors Collaboration. Lancet. 2010; 26; 375:2215-2222.

[4] Dr. K. Mohan Kumar1, S. Jamuna, "Comparative study on Machine learning Clustering Algorithms using Medical related Datasets", International Journal of Management, Technology And Engineering, Volume IX, Issue III, MARCH/2019.

[5] R.Nithya, P.Manikandan, and D.Ramyachitra, "Analysis of clustering technique for the diabetes dataset using the training set parameter", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4 (9), 2015, pp. 166–169.

[6] Dr. K. Mohan Kumar and S. Jamuna, "Comparitive study on Datamining techniques for Healthcare Information System", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-8, Aug 2018.

[7] Zeynel Cebeci and Figen Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures", Journal of Agricultural Informatics, Vol. 6 (3), 2015, pp. 13–23.

[8] Amatul,Zehra and tuty Asmawaty, Abdul Kadir and M.A.M., Aznan(2013), "A Comparative Study on the Pre-Processing and Mining of Pima Indian Diabetes Dataset", 3rd International Conference on software Engineering & Computer Systems(ICSECS-2013),20-22 August 2013,Universiti Malaysia Pahang. PP 1-10.

[9] Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath(2012), "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering Advanced Technology, Vol.1 No.3 pp 147-151.

[10] Vijayalakshmi, d., Thilagavathi, k.,"An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis", International Journal of applied mathematical research, Vol.1 (4) pp. 520-530,2012.

[11] P. Padmaja, V. Srikanth, N. Siddiqui, D. Praveen, B. Ambica, V. B. V. E. Venkata Rao, and V.J.P. Raju Rudraraju, "Characteristic evaluation of diabetes data using clustering techniques", International Journal of Computer Science and Network Security, Vol. 8 (11), 2008, pp. 244–251.

[12] Han, J., Rodriguze, J.C ., Beheshti, m., "Diabetes data analysis and prediction model discovery using rapid miner", Second International Conference on Future Generation Communication and Networking. 978-0-7695-3431-2 ,2008.

[13] https://www.diabetes.co.uk/ diabetes care/blood-sugar-level-ranges

[14] Juraj Horvath, "Image Segmentation using Fuzzy C-means", Slovak grants KEGA 3/120603 and VEGA 1/2185/05, 2006.

[15] Rajshekhar Ghogge, "Brain Tumour Detection Using K-means and Fuzzy C-means Clustering Algorithm", International Journal of Science, Engineering and Technology Research(IJSETR), Volume 3, Issue 7, July 2014.

[16] Ravi Sanakal, Smt. T Jayakumari, "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine", International Journal of Computer Trends and Technology– Vol.11 No.2 – May 2014.