Gyancity Journal of Engineering and Technology,
Vol.4, No.2, pp.21-28, July 2018
ISSN:2456-0065 DOI: 10.21058/gjet.2018.42003

A Machine Learning Based Approach for Diabetes Detection and Care in Bangladesh

Syed Sifat Rahman
Department of Computer Science and Engineering
University of Asia Pacific
Email:syedsifat02@gmail.com

Hasan Mahmud
Department of Computer Science and Engineering
University of Asia Pacific
Email:hasan404cse@gmail.com

Md.Rafeed Talukder
Department of Computer Science and Engineering
University of Asia Pacific
Email: rafeed717@gmail.com

Apubra Daria
Department of Computer Science and Engineering
University of Asia Pacific
Email:apurbadaria8@gmail.com

Shammi Akhtar
Assistant professor
Department of Computer Science and Engineering
University of Asia Pacific
Email: shammi@uap-bd.edu

Abstract

Now-a-days Diabetes is an alarming issue all over the world. In Bangladesh, many individuals are affected by it. Due to overpopulation and lack of proper education, it is very difficult to provide sufficient care for diabetes patients. This paper presents a system that can detect if anyone has diabetes. A machine learning algorithm, KNN is used on a supervised dataset to detect diabetes. It also shows nearby doctor chambers through location tracking. The dataset was collected from different hospitals of Bangladesh. Due to security reason, we cannot disclose the names of the institutions from where we collected all the data. However, supervised training in such situation shows a great accuracy although no such work has found for Bangladesh region. This paper studies two different algorithms on the dataset. These are, KNN and K-means. Between them, the proposed approach achieves 99.78% accuracy, which is so far the best for detecting diabetes. We have used total number of 6219 data of different diabetes affected patients. Through this system, one can easily know if he has diabetes by giving test reports and consult with nearest certified doctors with location tracking. Resulting in saving time and money used to detect diabetes and to find preferable doctors as one do not have to go to doctors the very first time and see his health condition sitting right at their home.

Keywords**:** *K-means, KNN, unsupervised training, supervised training, scaling, confusion matrix, clustering.*

1. Introduction

This paper primarily aims at detecting diabetes. Diabetes is one of the most talked topics in the world. It was first mention by Hesy-Ra, an Egyptian physician in 1552 B.C. Among 7 billion souls, 8.8% people of all over the world are now a victim of diabetes. Approximately 80% of people with diabetes live in low and middle-income countries. A great number of people

in Bangladesh is also affected by diabetes. The number is around 8.4% or 10 million according to research published in WHO bulletin in 2013. It will increase to a whopping 13% within 2030. It is due to population growth, aging, urbanization, and increasing prevalence of obesity and physical inactivity. It is greatly seen in rural areas, as people tend to be much lazy and physically inactive in developed areas. In Bangladesh, people spends large amount of money and time just to visit doctors and to do medical tests. Each year thousands of dollars is spend just to check if they have diabetes. In this paper, we have proposed a system that will be very much beneficial to all the diabetes-affected patients. Through this system, one can easily check whether he/she has diabetes or not. One can also look for the nearest doctor chamber and get necessary suggestions and others. Anyone who has access of internet can use this. As mentioned earlier, the proposed system gives an accuracy of 99.78%, which is very good. It is also as similar as how human would recognize. It splits the data into two sets and trains them. After that prediction is performed. We also used another method namely, K-means that only showed 96.38% accuracy. It separates all the data between two clusters. One who has diabetes and other one has not. Then it performs prediction. Both of the methods are completely different from each other. Among them KNN shows much better accuracy than unsupervised K-means. As a result, we moved on with the proposed method. This will save some money and time as one does not have to go to the doctors at the very first time. They can just do the medical test by themselves and know the result in seconds. After assuring, one will consult with doctors. The rest of the paper is organized as follows: Section 2 contains the background on KNN and K-means. Section 3 introduces our proposed approach, details of our used dataset, how we managed it and necessary processing. Section 4 describes different experiments we conducted and the methods followed for them. Section 5 discusses our output and the impact. Finally, conclusions are made in Section 6 along with our future plan.

2. Background

2.1. KNN (K-Nearest Neighbor)

K nearest neighbor algorithm is a learning algorithm, which is broadly used although it is very simple and gives good result in classification of linear data. It works better in a large dataset. KNN can be used for regression and classification problems. As our target value of the dataset is not continuous, we performed KNN classification. It works in such a way that a sample data is compared with all previously inputted samples in terms of Euclidean distance, the most popular distance measurement. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y) across all input attributes.

$$ED(x,y) = \sqrt{\sum_{i=1}^{n}(xj - yj)^2}$$

Other popular distance measures are Hamming Distance, Manhattan Distance and Minkowski Distance.

From the previous dataset, K samples of less distance to new sample x are selected as neighbors of sample x. Finally, x belongs to that class which has the most votes. Most difficult part is to find the value of K. For which value of K we get the best output is considered as the value of K. So we have to try different odd numbers so there can never be any conflict between two classes of same votes. For the value of K = 5, we got the best output. Therefore, it is the value of K for our dataset the procedures to compute prediction and accuracy is describes later on this paper.

Gyancity Journal of Engineering and Technology,
Vol.4, No.2, pp.21-28, July 2018
ISSN:2456-0065 DOI: 10.21058/gjet.2018.42003

2.2. K-means

K-means is a type of unsupervised clustering, which is used on unlabeled data. The objective of this algorithm is to find groups in the data, where the numbers is predefined by the value of K. The entire dataset will be clustered into K number of groups. Here we know that our whole dataset can be divided between two classes. Those who has diabetes and those who do not. So the value of K is 2 in this case, resulting the number of centroids being 2. Firstly, among all the data any two is chosen as the centroids. Then, assigns each data to the group that has the closest centroid. This is done by computing Euclidean distance of each data from each centroids. After assigning all the data to a group, both the centroids are repositioned by computing average of all the data of the same group. This step is repeated until the centroids no longer move. Finally, we get two groups of data; one has diabetes and one do not. After reshaping the target value, we can get the accuracy of our dataset.

3. Proposed approach and dataset

3.1. Proposed Approach

As mentioned earlier this system predicts whether an individual has diabetes or not. Easy access and understanding makes it user-friendly. After a user gives input of proper data, our system checks whether he/she has diabetes or not by using KNN method. It gives better accuracy than KMeans. Firstly, we calculate Euclidean distance of the input value from each attribute. Then, the class of least distance is assumed the class it belongs to. Here a user also can use the facility to see the doctor chamber/hospitals near him.

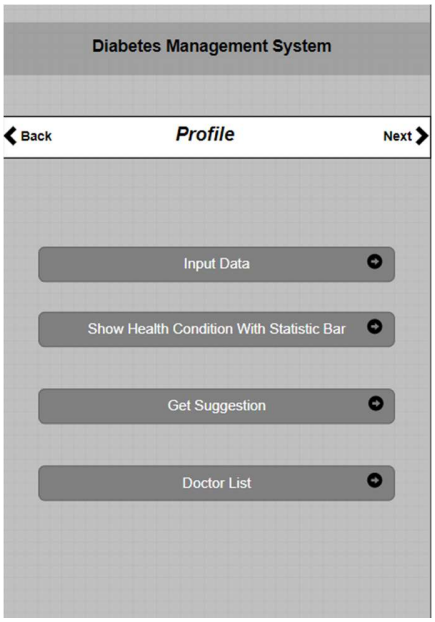
Figure 1.shows the profile of the user, from where he/she can choose what to do. One can entry data,observe his/her physical condition via statistic bar,get suggestions and nearby doctor list.

Figure 2.describes the entities one has to entry.

Figure 3.shows the statistic bar of the user depending on the values he inserted. Through it, he can easily observe his physical condition by comparing recent values with previous.

Figure 4.shows some suggestions based on the input value.

Figure 5.tracks the user's location and suggests nearest doctors.



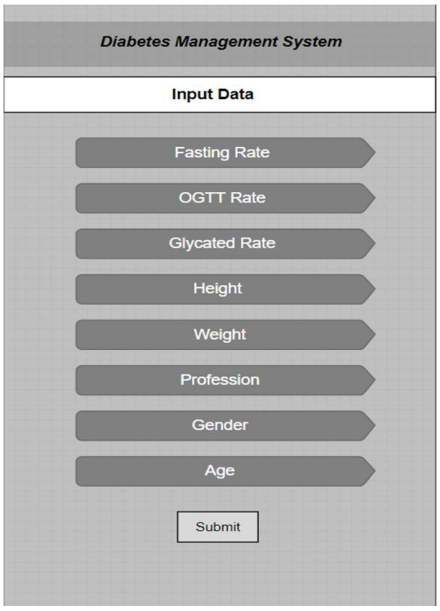


Figure 1.User profile

Figure 3.Statistic bar

Figure 2. Data to be inserted

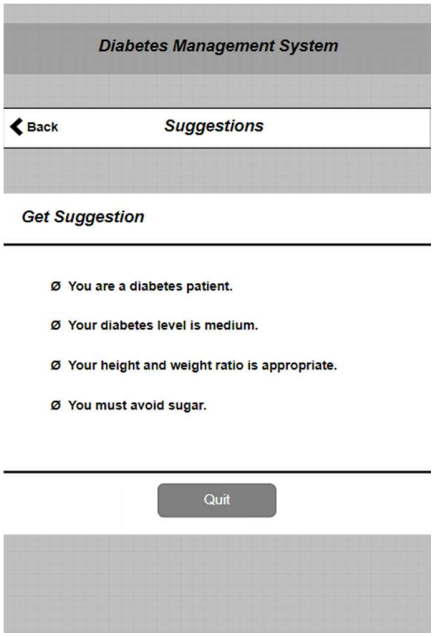


Figure 4. Suggestions

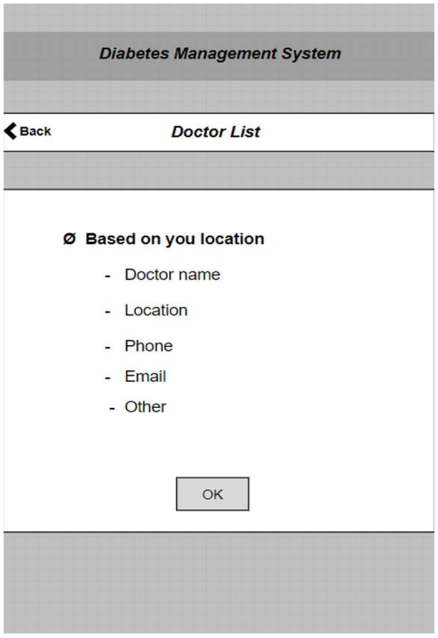


Figure 5.Nearest doctor list

Gyancity Journal of Engineering and Technology,
Vol.4, No.2, pp.21-28, July 2018
ISSN:2456-0065 DOI: 10.21058/gjet.2018.42003

3.2. Dataset and Preprocessing

As supervised dataset is used in this system, we will need many data. The more data we will work with the more accurate this system will be. That is why we have tried to maximize the number of data. We have collected our dataset from different hospitals. We managed to collect a total number of 6219 data (showed in Figure 6). There exist three types of data in our dataset. These are glycated, fasting and OGTT. In our dataset, we have an extra column named' SeqNum' which we do not need. Therefore, we dropped down that column. Doing so, it increased our accuracy a big time.

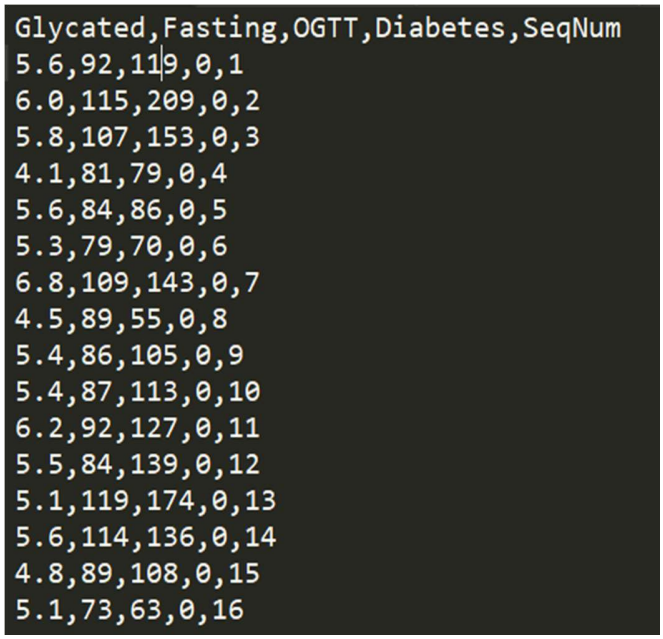


Figure 6.Dataset

```
[[ 0.6121182  -0.11787819  0.1364961 ]
 [ 1.1650852   1.33652905  2.11634037]
 [ 0.8886017   0.83064827  0.88443727]
 ...,
 [ 1.1650852   0.45123768  2.33632306]
 [ 0.0591512  -0.87669936 -0.83142777]
 [ 0.6121182  -0.43405368 -0.32546757]]
```

Figure 7.Scaled dataset

It is to be noticed that the data we used in our system is little bit skewed and largely sparse. The range of glycated is much smaller than other two. Working with this data will not give us proper prediction. Scaling the data brings all values onto one scale eliminating the sparsity. Therefore, it shifts the distribution of each attribute to have a mean of zero and a standard deviation of one. Therefore, we had to process them so that we can apply the algorithm correctly. Doing so, machine can easily relate all of them without much hustle. After scaling, the features look something like Figure 7.

4. Experiment

As mentioned earlier, we split entire dataset into two sets training and testing by 20%. So, for training set we get 4975 and for test set 1244 data.Then we fit the training data in the machine for training. We use the test set against the training set for accuracy. Machine is learned from the random train set of 4975 data and from that, test set of 1244 data is analyzed with respect to

train set. Accuracy is computed by analyzing how accurately test set scores by learning from train set. We take an input of an unknown data which we will predict if it is in the range of diabetes. For that, each value of each features is subtracted from the new value and adds them. So we get Euclidean distance of the new data from each attribute. Then the minimum difference is considered to be the class, where the new value belongs to. Here the value of k is 5. So the mean of closest value of 5 will be the predicted class. We also computed confusion matrix. Confusion matrix is often used to describe the performance of a classification model on a set of test data for which the true values are known.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4885
1	0.99	0.92	0.95	90
avg / total	1.00	1.00	1.00	4975

Figure 8.Confision matrix

From there we got precision, recall and f1 score. These are used to understand the prediction and accuracy.
Precision is the ratio of correctly predicted positive observations to the total predicted positive observation.
Recall is the ratio of correctly predicted positive observations to the all observations in actual class –yes
F1 Score is the weighted average of Precision and Recall.
In Figure 8, we can see that the avg/total is of precision, recall and f1-score is 1.00 that is the best value we can get. In average, all the classes of "0" and "1" were classified correctly.

5. Results and analysis

Among two algorithms KNN and KMeans, KNN gives the best output of 99.78% accuracy, where KMeans gives 96.38% of accuracy. We computed confusion matrix for both the algorithms. In KNN, precision, recall and f1-score are all 100%. Which is quite fascinating. Meaning, false positive rate is null, we successfully labeled all the features, and for f1-score, it is similar to accuracy. It is to be noticed that, in KMeans the values of precision, recall and f1-score are slightly different. The values are 99%, 96% and 97% for precision, recall and f1-score respectively. From this discussion we can say that KNN gives better accuracy over KMeans.

6. Conclusion

This paper presents the implementation of supervised training, using KNN for diabetes detection. To demonstrate the effectiveness of this approach, we tested the dataset with two different algorithm in the dataset. For the supervised algorithm we got an accuracy of 99.78%, which is the best reported result on this dataset so far. Apart from demonstrating the utility of unsupervised training in the context of Diabetes detection, our results also indicate that such supervised training can be useful even when the data sets are independently (and blindly) collected. Until now, supervised training has shown much better prediction and accuracy than unsupervised training. If we can manage to manage bigger and better dataset, both accuracy and prediction can be improved.

ISSN:2456-0065 DOI: 10.21058/gjet.2018.42003

In future, we are targeting to merge telecommunication with this system. If succeed, then it will be a revolutionary step towards handling diabetes in Bangladesh. One can easily access any sort of diabetes related treatment and consult with specialists sitting right at their home.

7. References

[1] Mbusa C Takenga, Rolf-Dietrich Berndt,Olivier Musongya, Henri Takenga,"An ICT-Based Diabetes Management System Tested for Health Care Delivery in the African Context",International Journal of Telemedicine and Applications, 2014.
[2] Si D., Bailie R, Wang Z, Weeramanthri T,"Comparison of diabetes management in five countries for general and indigenous populations: an internet-based review",BMC Health Serv Res.,2010.
[3] Lanzola G., Capozzi D, D'Annunzio G, Ferrari P, Bellazzi R, Larizza C,"Going mobile with a multiaccess service for the management of diabetic patients",J Diabetes Sci Technol.,2007.
[4] Liang X., Wang Q, Yang X, Cao J, Chen J, Mo X, Huang J, Wang L, Gu D,"Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis",Diabet Med.,2011.
[5] Davide Capozzi,Giordano Lanzola,"Utilizing Information Technologies for Lifelong Monitoring in Diabetes Patients",J Diabetes Sci Technol.,2011.\
[6] Giordano Lanzola,Davide Capozzi,Nadia Serina,Lalo Magni,Riccardo Bellazzi,"Bringing the Artificial Pancreas Home: Telemedicine Aspects",J Diabetes Sci Technol.,2011.
[7] Bellazzi R.,"Telemedicine and diabetes management: current challenges and future research directions",J Diabetes Sci Technol.,2008.
[8] Rohlfing CL, Wiedmeyer HM, Little RR, England JD, Tennill A, Goldstein DE,"Defining the relationship between plasma glucose and HbA(1c): analysis of glucose profiles and HbA(1c) in the Diabetes Control and Complications Trial",Diabetes Care., 2002.
[9] Takenga MC, Berndt RD, Kuehn S, Preik P, Stoll N, Thurow K, Kumar M, Behrendt S, Weippert M, Rieger A, Stoll R,"Stress and fitness monitoring embedded on a modern telematics platform",Telemed J E Health,2012.
[10] SiDiary-Diabetes Management Sofware,http://www.sinovo.de/.
[11] Diabetes Managemment Software, "Bayer's GLUCKOFACTS DELUXE Software",2013,http://www.bayerdiabetes.com/sections/outproducts/software.aspx.

Syed Sifat Rahman
Department of Computer Science and Engineering
University of Asia Pacific

Hasan Mahmud
Department of Computer Science and Engineering
University of Asia Pacific

Md.Rafeed Talukder
Department of Computer Science and Engineering
University of Asia Pacific

Apubra Daria
Department of Computer Science and Engineering
University of Asia Pacific

Shammi Akhtar
Assistant professor
Department of Computer Science and Engineering
University of Asia Pacific