

BUILDING PREDICTIVE MODEL FOR DIABETICS **DATA USING K MEANS ALGORITHM**

D. Christy Sujatha *

D.Maghesh Kumar *

M.ChandrakumarPeter*

Abstract

In this digital era data is growing exponentially every day and the digitized data of patients health record is increasing due to the development of innovative digital devices used in the hospitals. Health care analytics plays a major role in big data because it is much more useful to prevent a disease from happening *than* to cure after it has occurred. In this paper the healthcare dataset of pregnant women from Indian healthcare is taken to analyze and the predictive model is built using k means algorithm. The k means algorithm is used to categorize the patients into Healthy and Diabetic clusters. The developed model is tested with the sample data and the accuracy of the predictive model using k means algorithm is found to be 78%.

Keywords:

Health care,
Predictive Model,
Cluster

*** Assistant Professor ,Department of Software Engineering,
Periyar Maniammai Institute of Science and Technology, Thanjavur, India**

1. Introduction

Big data and health care analytics are becoming a part of our every day life and many companies and research groups are working to treat diabetics. Nowadays most of the hospitals are equipped with well-established latest devices and the digitization of health care records are growing exponentially [1]. Diabetes leads to serious complications or even premature death. Hence it is very important to give an awareness to the patients to take proper treatment by analyzing the indications of diabetics. Using the existing data of the patients, we can predict the diabetic signs of the patients by building predictive model. Machine learning algorithms are used to classify and diagnosis the diseases to build the predictive model[2]. In this paper the healthcare dataset of pregnant women from Pima Indian Data set is taken to analyze and to build the predictive model. The k means algorithm is used to categorize and cluster the patients into Healthy and Diabetic. The developed model is tested with the sample data and the accuracy of the predictive model using k means algorithm is found to be 78%.

The organization of the paper is as follows: Session 2 includes the existing related work related to the diabetics patient analytics using various algorithms. Session 3 briefs about the steps involved in k means algorithm. Session 4 describes the implementation of k means algorithm to build the predictive model and the model is tested with testing data. Session 5 concludes with further enhancement of this paper.

2. RELATED WORK

Several researchers developed and design a system for diabetes prediction based on various algorithms and methods. The author M. Kothainayaki et al., defines the Classification of diabetic's data set and the k-means algorithm to categorical domains. They used the missing value algorithm to replace the null values in the data set which also used to improve the classification rate and cluster the data set using two attributes namely plasma and pregnancy attribute [3]. Mustafa S. Kadhm et al., proposed a system used K-nearest neighbor algorithm for eliminating the undesired data and the proposed classification approach based on Decision Tree (DT) to assign each data sample to its appropriate class[4]. Gagandeep Singh et al. [5] deals with diabetes and proposed a data mining system using simple K-Means and nearest neighbor hierarchical clustering. The main objective is to find the effects of diabetes on the people grouped by age and evaluating the survival ratio in efficient manner. Accuracy, Sensitivity and Specificity are different

metrics were evaluated. V. Anuja et al. [6] proposed a system for diabetes disease classification using Support Vector Machine (SVM). Aiswarya et al. [7] used J48 Decision Tree and Naïve Bayes as classifiers for classify the diagnosis of diabetes. Rajesh et al. [8] proposed a system for diabetes classification based on using C4.5 algorithm for classification.

3. BUILDING PREDICTIVE MODEL

Machine Learning algorithm is commonly used to build the predictive model. Clustering is a method of unsupervised machine learning technique and is a common technique for statistical data analysis used in many fields. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. Clustering algorithms can be classified into two main categories Linear clustering algorithms and Non-linear clustering algorithms. [9]

- **Linear clustering algorithm**
 - k-means clustering algorithm
 - Fuzzy c-means clustering algorithm
 - Hierarchical clustering algorithm
 - Gaussian(EM) clustering algorithm
 - Quality threshold clustering algorithm
- **Non-linear clustering algorithm**
 - MST based clustering algorithm
 - kernel k-means clustering algorithm
 - Density-based clustering algorithm

K-means [10] is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The algorithm is composed of the following steps:

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centres.
3. Assign objects to their closest cluster centre according to the *Euclidean distance* function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2,$$

where $\|x_i^{(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j , is an indicator of the distance of the n data points from their respective cluster centres.

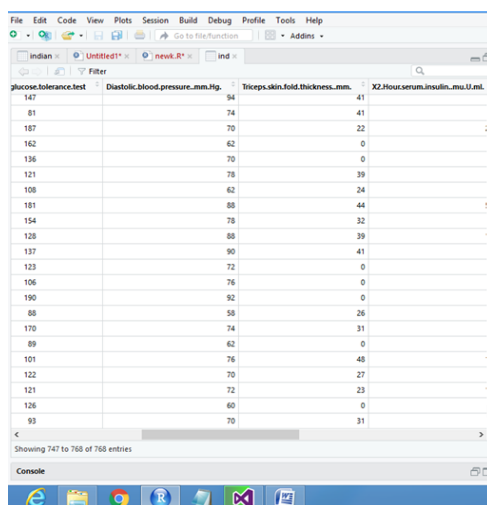
4. EXPERIMENTAL RESULTS

4.1 SYSTEM CONFIGURATION

Core i5 processor, Memory with 4GB RAM and 1TB Hard Disc is used to implement the analytics model . The open source software R tool programming language is used to build the predictive model, to find the statistical computation and to produce the graphical output.

4.2 DATA SET

The online data in csv format from Pima Indian Dataset is loaded which has the following attributes: Number of times pregnant, Plasma glucose concentration a two hours, Diastolic blood pressure , Triceps skin fold thickness, 2-Hours Serum insulin , Body mass index , Diabetes pedigree function and Age. Then the statistical data like min, max ,mean and median values are found.



glucose.tolerance.test	Diastolic.blood.pressure.mm.Hg.	Triceps.skin.fold.thickness.mm.	X2.Hourserum.insulin.mu.U.ml.
147	94	41	
85	74	41	
187	70	22	
162	62	0	
136	70	0	
121	78	39	
108	62	24	
181	88	44	
154	78	32	
128	68	39	
137	90	41	
123	72	0	
106	76	0	
190	92	0	
88	58	26	
170	74	31	
89	62	0	
101	76	48	
122	70	27	
121	72	23	
126	60	0	
93	70	31	

Figure 1. Loaded data

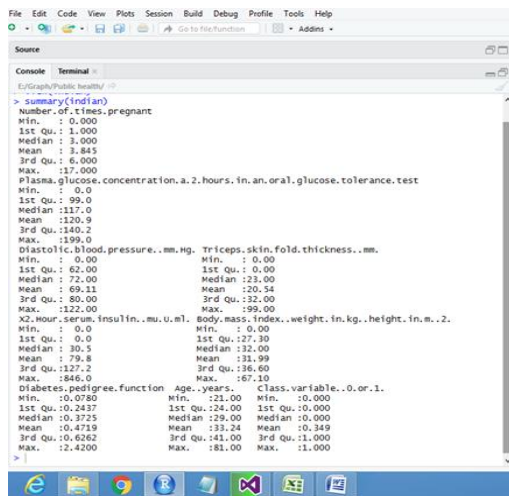


Figure 2. Statistical data of the given data set

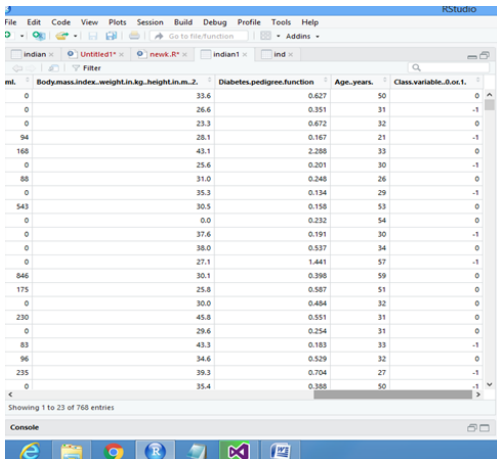
4.3 DATA PREPROCESSING

Data preprocessing describes the type of processing performed on raw data to prepare it for another processing procedure. It is possible to get quality output only if we process with quality input data. Hence it is necessary to preprocess the input in order to remove the null values and duplication of data. It plays an important role and more challenging process due to the high volume, velocity and variety of data. Hence the collected data is cleaned and the null values are removed.

4.4 DATA TRANSFORMATION AND REDUCTION

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.

In this paper the normalization process is implemented where the data is scaled to fall within the specified range such as -1.0 to 1 (or) 0 to 1.0. This process of minimizing the size of data increases the storage efficiency and reduces the cost.



	Bodymass.index	weight.in.kg	height.in.m.2	Diabetes.pedgree.function	Age.years	Class.variable.0,oc.1
0	33.6			0.627	50	0
0	26.6			0.351	31	-1
0	23.3			0.672	32	0
94	28.1			0.167	21	-1
168	43.1			2.288	33	0
0	25.6			0.201	30	-1
88	31.0			0.248	26	0
0	35.3			0.134	29	-1
543	30.5			0.158	53	0
0	0.0			0.232	54	0
0	37.6			0.191	30	-1
0	38.0			0.537	34	0
0	27.1			1.441	57	-1
846	30.1			0.398	59	0
175	25.8			0.587	51	0
0	30.0			0.404	32	0
230	45.8			0.551	31	0
0	29.6			0.254	31	0
83	43.3			0.183	33	-1
96	34.6			0.529	32	0
235	39.3			0.704	27	-1
0	35.4			0.388	50	-1

Figure 3. Transformed data

4.5 DATA ANALYTICS MODEL

The predictive analytics using k means algorithm is developed using R tool. To test the predictive analytics model, the data set is split into 2 sets namely training data set and testing data set. 300 records are selected randomly to train the data set and 92 records are used to test the building model. These datasets are selected at random the representation of the actual population. The training data set is significantly larger than the testing data set.

4.6 RESULT AND DISCUSSION

K means algorithm groups the data set into Healthy and Diabetic clusters.

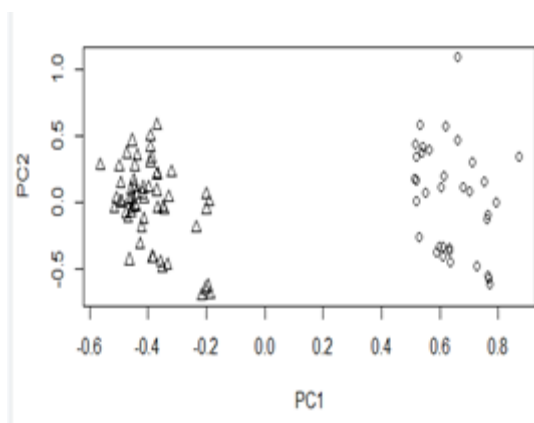


Figure 4. Data clusters of Healthy and Diabetic

Accuracy of the predictive model = $(TP + TN) / (TP + TN + FP + FN)$.
[11]

True Positive (TP) - Measures the proportion of positives that are correctly identified.

True Negative (TN) - Measures the proportion of negatives that are correctly identified.

False Positive (FP) - Result that indicates a given condition has been fulfilled, when it actually has not been fulfilled.

False Negative (FN) - It is where a test result indicates that a condition failed, while it actually was successful.

The accuracy of the implemented predicted model using k means algorithm is evaluated as 78 %

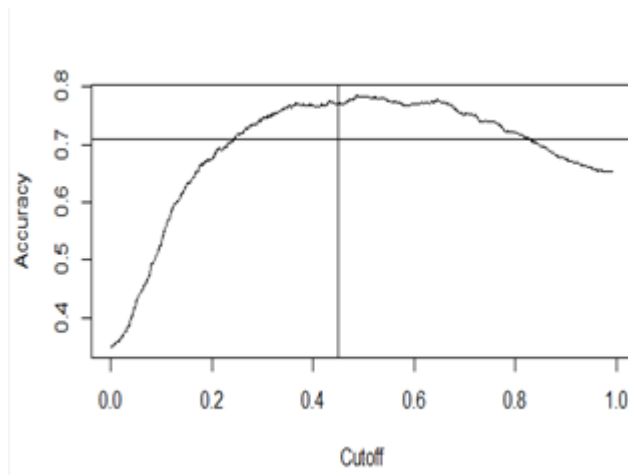


Figure 5. Accuracy of the predictive model

5. CONCLUSION

In this paper the healthcare dataset of pregnant women from Indianhealthcare is taken to build the predictive model using k means algorithm. The developed model is tested with the sample data and the accuracy of the predictive model using k means algorithm is found to be 78%. It is also planned to implement other clustering algorithms in R tool and accuracy and sensitivity of all algorithms may be compared.

REFERENCES

- [1] The Digitization of the Healthcare Industry: Using Technology to Transform Care CISCO
- [2] [Anju Jain.,2015],” Machine Learning Techniques For Medical Diagnosis: A Review” In: 2nd International Conference In Science , Technology And Management
- [3] [M. Kothainayaki et al., 2013], “Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm” In : Journal of Applied Information Science Volume 1, issue 1, June 2013.

- [4] [Mustafa S. Kadhm et al.,2018], “An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach” In : International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4038-4041
- [5] [Gagandeep Singh et al.],“Diabetes classification using k-means” In : <http://acfa.apeejay.edu> Apeejay On line Journal Of Computer Science And Applications.
- [6] [Subhagatachattopadhyay et al.] , “A comparative study of k meansc-means Algorithm and entropy-based k means Clustering algorithms” .
- [7] [Bottou et al., 1995], “Convergence properties of the K-means algorithms” In :Tesauro, G. and Touretzky, D. (Eds.) Advances in Neural Information Processing Systems 7, 585-592, The MIT Press, Cambridge.
- [8] [David Arthur et al.,] “k-means++: The Advantages of Careful Seeding”,
- [9] <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>.
- [10] [Tapas Kanungo et al., 2002], “An Efficient k-Means Clustering Algorithm: Analysis and Implementation” In: Ieee Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002
- [11] [Wen Zhu et. Al., 2010] , “Sensitivity Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations” In : Health care and Life Sciences