Amalgamatorial K-Means Clustering for Diabetes Mellitus Type 2: A Machine Learning Tool

Dr. P. SambasivaRao[1], Dr. Satyanarayana Gaddada[2], Uma Mahesh Kumar Gandham[3]

1. *Associate Professor. Departmentof CSE, Christu Jyothi Institute of Technology & Science, Jangaon, Jangaon District, Telangana, India.*
2. *Professor, Departmentof CSE, International School of Technology and Sciences for Women, Rajahmundry, East Godavari District, AP, India.*
3. *Associate Professor, Departmentof CSE, International School of Technology and Sciences for Women, Rajahmundry, East Godavari District, AP, India*

ABSTRACT: In order to effectively group patients with Diabetes Mellitus Type2 (DMT2) and underlying diseases (arterial hypertonia (AH), ischemic heart disease (CHD), diabetic polyneuropathy (DPNP), and diabetic micro angiopathy (DMA) into similar groups, a new, original procedure based on k-means clustering is being developed. A machine learning approach for finding structures in datasets is clustering. Based on clinical records, clustering has been shown to be effective for pattern recognition. With a predetermined number of descriptors and groups, theexplored combinatorial k-means approach investigates every potential k-means clustering. The predetermined criteria for the partitioning were as follows: optimal descriptors for each disease and group; each subgroup formed in this way was subject to partitioning into three patterns (good health status, medium health status, and degenerated health status). Each group of patients included patients with DMT2 and one of the underlying diseases. The parameter known as global variance, which is the total of all variance values for all clinical variables across all clusters, is used to determine which clustering is the best. In order to find the ideal clinical parameters, this global variance must be minimised. This methodology must determine a collection of factors that are presumptively capable of effectively separating each underlying condition in three different patient groupings. These four underlying disorders' hierarchical clustering could be utilised to create patient groups with linked clinical data. The suggested methodology provides inferred outcomes from complex data based on a connection to the group's health state and paints a picture of the accuracy of the continuing health status prediction rate.

Keywords:k-means, multivariate statistics, type 2 diabetes mellitus, underlying disorders, classification, and descriptors.

1. Introduction:

The clustering technique K-means (partitioning clustering) divides things into k groups (clusters). The supervised k-group beginning number must be specified for this method.

The iterative process is used to reduce the gap between each observation and its associated mean, with each group typically being centered around its mean. Discovering patterns in the diabetic data set and creating a model using k-means and hierarchical clustering analysis were the primary goals of the exploratory data investigations. The goal was to efficiently assess the survival ratio in a group of people who had been diagnosed with diabetes and were grouped by age.

In the recent years, machine learning (ML) approaches have been applied to identify the predictive patterns for the diabetes condition. Researchers have developed strong models using deep learning (DL) techniques as a result of the problem's increasing density.

The impact of this problem on society, however, was solved and lessened through the use of ML algorithms. The study by Marie-Sainte [1] graphed every ML and DL method based on initiatives for diabetes predictions released in the previous years. The recommendation is to use these accurate categorization and prediction models to diabetic disease and to strengthen their resilience by creating more intricate models that can be used.

The estimate of diabetes using various algorithms and approaches has undergone similar research. Anuja et alwork .'s [2] gave a support vector machine classification of the diabetic condition (SVM). Finding the most appropriate descriptors that can distinguish between a specified number of examples was done using a process based on k-means clustering. A categorization system for diabetes was put forth by Rajesh et al. [3] and is based on the C4.5 algorithms. By analysing the training data using data feature relevance analysis, the authors were able to attain a 91% classification rate. For classifying the diabetes diagnosis, Aiswarya et al. [4] employed decision trees and naive Bayes classifiers. A structure analysis based on a data mining technique for predicting diabetic disease was recommended by Harleen et al. [5].Pre-processing, feature extraction, and parameter evaluation are the three primary processes in the suggested framework. The cases of missing data or data that had irregularities in the sets were disregarded during the pre-processing stage. In order to improve the level of decision-making from the given results, hidden patterns and interactions within the data set are recorded in the process for feature extraction.

The achieved rates for the suggested system were 73.8% and 76.3%, respectively, when J48 and naive Bayes were used to evaluate it. In the context of machine learning techniques for the identification of diabetes, Haq et al. [6] have proposed a study. A clinical data set created from the patient's medical history called the diabetes data set has been used to test the proposed method. The hold out and K-fold model validation techniques omit one subject, and the suggested scheme's validity has been assessed using performance assessment measures such as accuracy, specificity, sensitivity, F1-score, receiver operating characteristic curve, and execution time. For the classification of the groups of healthy and diabetic participants, the decision tree classifier was utilised.For the feature selection, there was a good correlation with experimentally obtained results, and the predictive model's classification performance improved. The multilayer approach of the chosen feature set is what gives the executed procedure its capability. Six different techniques were put forth by Krishnaveni et al. [7] for the diabetic disease prediction model. Discriminant analysis, the KNN method, naive Bayes, SVM with a linear kernel function, and SVM with an RBF kernel function are the techniques that are used.According to the study's results utilising the aforementioned methodology, discriminant analysis was used by 76.3%, KNN by 71.1%, naive Bayes by 76.1%, SVM with a linear kernel function by 74.1%, and SVM with an RBF kernel function by 74.1%. However, in order to achieve the best prediction rate, multiple writers [8–10] have used a variety of techniques.

In a recent study, Ruy et al. [11] employed a novel screening algorithm for assessing undiagnosed diabetic mellitus (UDM) as critical for early medical intervention. The suggested screening strategy for calculating UDM in individuals with a family history of diabetes made it possible to validate it. The backward stepwise logistic regression technique was utilised to screen the UDM using data from distinct male, female, and combination groups. Both the establishment and confirmation of the screening models took place. A screening strategy for undiagnosed diabetes mellitus was used by Ryu et al. [12] to identify patients in a recent study.The screening model using machine learning techniques was used for undiagnosed DM. 11,456 participants' data were designated, and 4444 participants evaluated the model. A link between the machine learning model and earlier screening models was evident. According to the authors' research, a deep learning model for people with undiagnosed diabetes could help to lessen the disease's impact on the national economy and social conditions.

In this study, we were able to find an application of a novel technique that not only concentrated on finding the best descriptors but also had a strong ability to deal with missing data. There has been a lot of research done on the subject of missing data. It must be remembered that improper handling of missing data could result in biased findings, as in the case of multiple imputations.As an alternative, we created a unique methodology that does not make an effort to fill in the gaps in the data. To test if the strategy could find the optimal descriptors, computations using simulated data were made. The division of the patients into three clusters was further supported by the application of hierarchical cluster analysis. The major goal of the current study was to evaluate the efficacy of the suggested combinatorial k-means algorithm for accurately classifying type-2 diabetic patients with underlying conditions detected using the best suitable clinical descriptors.

2. Materials and Methods

2.1 Data from Clinical DMT2 with Four Underlying Conditions

Clinical DMT2 Data with Four Associated Conditions in this investigation, the clinical data of 52 subjects (diabetes mellitus type 2 (DMT2) patients) with at least one of four underlying diseases – arterial hypertonia (AH), ischemic heart disease (CHD), diabetic polyneuropathy (DPNP), and diabetic microangiopathy (DMA)—were used. 51 patients had AH, 48 had CHD, 51 had DPNP, and 47 had DMA out of the total of 52 patients.Age, height, weight, waist size, BMI (body mass index), systolic and diastolic blood pressure, Hgb (haemoglobin), Er (erythrocytes), Leu (leucocytes), Tro (thrombocytes), creatinine, ALAT, GGT (enzymes alanine aminotransferase, gamma glutamyl transferase), chol (cholesterol), HDL (high (concentrations of sodium, potassium, chloride, calcium, phosphorus).

All patients have access to some clinical factors, such as age, but generally speaking, some patients lack information regarding the values of certain variables. The patient's description depends on the data from 23 out of 26 clinical factors, each of which has an average of two missing values.

2.2 Tools for Combinatorial K-Means Cluster Analysis

A python tool called the combinatorial k-means clustering analysis tool was created to determine the separation of instances by k-means clustering using only a small number of descriptors (variables) chosen from a huge range of descriptors. The Python Notebook platform was used to create the combinatorial k-means clustering software, which makes use of NumPy, Pandas, Matplotlib, SciPy, and Itertools. The NumPy module was utilised to calculate data numerically in an effective manner. The data of the patients for each underlying disease was stored in spreadsheet files, which were read using the Pandas module. The Matplotlib module was used to represent the data graphically.K-means clustering was carried out using the submodule hierarchy of the SciPy submodule cluster. The module tools were used to perform all conceivable descriptor combinations. The user-selected Excel file was where the application pulled all the data from. The data was then normalised by dividing the data by each property's standard deviation. Different linking techniques, such as average, centroid, or Ward's, can be used to produce the k-means clustering. For all calculations in this study, the centroid approach was chosen. By default, the application used k-means clustering, using as input data all combinations of three descriptors picked from every property in the spreadsheet.Combinations with two descriptors are also available as an option. In this study, three variables were automatically clustered into all conceivable three groups. For each of the k-mean clustering, a property known as the global variance was calculated in order to rank all potential clustering:

$$\text{varGlobal} = \sum \sum \text{vari,k} \qquad [1]$$

where vari,k represents the descriptor i's variance within the cluster k. Equation (1) is calculated in this study and has nine terms that correspond to the three descriptors for each of the three clusters.

Some limitations were also used in order to choose a reliable clustering. Clusters with an unbalanced number of members could be produced by systems with random clustering. This would produce clusters made up of one or two elements that have a low variance for the group to which they belong. A legitimate clustering is defined as the situation when at least three elements for any of the clusters are involved in order to prevent this problem. Utilizing all three descriptors, the algorithm performed clustering. The proper number of cases with information for each combination of the three descriptors used in the study were chosen. In this manner, the most data was always used for each clustering.

The algorithm's workflow design diagram is shown below (Figure 1), and the source code may be found at https://github.com/smadurga/Combinatorial-K-means-clustering (Accessed on 5 February 2021).

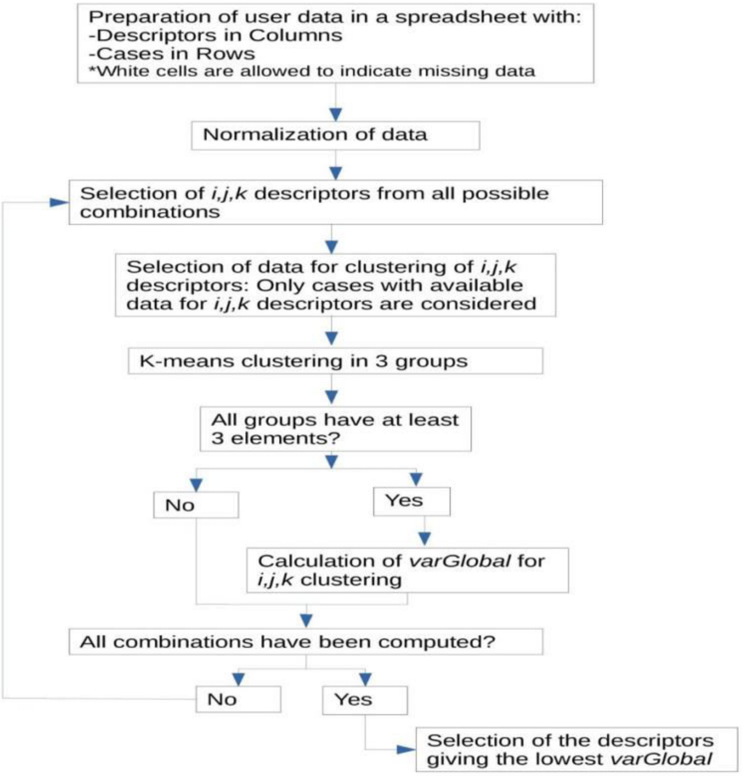
Preparation of user data in a spreadsheet with:
-Descriptors in Columns
-Cases in Rows
*White cells are allowed to indicate missing data
Normalization of data
Selection of i,j,k descriptors from all possible combinations
Selection of data for clustering of i,j,k descriptors: Only cases with available data for i,j,k descriptors are considered
K-means clustering in 3 groups
All groups have at least 3 elements?
No
Yes
Calculation of varGlobal for i,j,k clustering
All combinations have been computed?
No
Yes
Selection of the descriptors giving the lowest varGlobal

Figure 1: Amalgamatorial K-Means Clustering Work Flow Diagram

2.3 Analysis of Clusters in Hierarchy (HCA)

A non-supervised multivariate statistical technique called hierarchical clustering analysis (HCA) searches a data set for clusters—groups with similar patterns. By examining the patterns of similarities between objects in the parameter space and between parameters in the object space, this approach of data exploration permits the analysis of the structure of data. Different HCA techniques use different similarity measures and clustering techniques. The Euclidean distance is the similarity metric that is used the most frequently. The applied techniques for object and variable clustering are well known and frequently used as a useful algorithm.The following algorithms are frequently used: centroid linkage, average linkage method, complete linkage technique, and single linkage method. In order to accomplish the objectives of our study, the Ward's linkage approach was used. The procedure calls for normalising (standardising) the raw data (typically using z-normalization), determining the similarity measures (typically using Euclidean distances), choosing the best linkage option, plotting the clusters on a tree-like diagram (dendrogram), and counting the number of statistically significant clusters using a specific test (typically Sneath's criterion) [13,14]. The STATISTICA 8.0 software programme was utilised for clustering.

3. Result and Discussion

3.1 K-Means Clustering in Combination with Data Test Values

Test values with three descriptors that can define three clusters and other sets of descriptors with random values were developed in order to assess the program's capacity to identify the most suitable descriptors for categorization.

The programme creates three clusters initially, each with a unique center value for the designated X, Y, and Z descriptors. A set of values corresponding to non-correlated descriptors is then produced. The computer software should recognise the X, Y, and Z descriptors as those defining three clusters for each run of the combinatorial k-means clustering algorithm.The core values of the three clusters of X, Y, and Z descriptors were separated differently in each run. The combinatorial k-means clustering is anticipated to perform well in identifying the X, Y, and Z descriptors when there is a significant amount of separation between the three groups. These descriptors, however, are probably not recognised when the values of the groups are dispersed on the order of the separation of central values. A matrix with 30 columns (descriptors) and 45 cases was created to run this test. A normal distribution with a center point of 1 and a σ of 1w as used to create 37 columns of random numbers.Three groups were defined by the use of the additional three columns (X, Y, and Z). In particular, 15 examples with a normal distribution centered at point (p,0,0) and a dispersion of 1 for the X, Y, and Z descriptors were randomly created. X, Y, and Z descriptors were randomly generated in a further 15 cases using a normal distribution centered at point (0,p,0) using a factor of 1, and the final 15 cases with a factor of 1 were centered at (0,0,p). Using just one value of p, the three groups in the space of X, Y, and Z descriptors are divided in this set of values.It must be noticed that the variation in just one of the characteristics is what separates the pairs of groups. When the parameter p is 2 p, it is possible to determine the centroid's separation.

The combinatorial k-means clustering investigation was conducted using a variety of random test set values produced using various p values. Using the criterion of the smallest global variance, three sets of descriptors were determined to be the most suitable for categorization for each computation. Each time, it is assessed to see if the X, Y, and Z descriptors were chosen correctly.500 calculations using various random values were carried out to determine the likelihood that the relevant descriptors would be recognised. Table 1 shows the likelihood of getting the top-ranked X, Y, and Z descriptors out of all possible descriptor combinations (30 29 28 = 24,360) as a function of group separation. As can be seen in Table 1, there is a low likelihood that the X, Y, and Z descriptors will be accurately identified when the group separation is minimal relative to the value of. The randomly formed clusters are so closely packed together at low values of p that it is challenging to distinguish them from other randomly generated data.The X, Y, and Z descriptors are found to be the best ones for classification in 75% of situations when the parameter p is three times the value of. If the group separation is 3.5 times the value of, the success rate rises to 99%. As a result, the division of the groups according to the dispersion of the values in each of the clusters may serve as a sign of the suitability of the descriptors chosen.

Table 1 shows the likelihood of correctly identifying all 24,360 combinations of X, Y, and Z descriptors as a function of the p value used for cluster separation.

p	**Probability to Rank X, Y, and Z in First Position**
2.2	3
2.3	5
2.4	10
2.5	15
2.6	21
2.7	39
2.8	45
2.9	60
3	75
3.1	80
3.2	87
3.3	95
3.4	96
3.5	99

3.2 Combinatorial K-Means Clustering with DMT2 Patients' Underlying Diseases

For combinatorial k-means analysis, the 51 patients with arterial hypertonia as the underlying condition were chosen. BMI, HbA1c, and EAG were the clinical variables that received the highest ranking using the minimum global variance criteria (Table 2). Table 2 also displays the non-standardized values and accompanying standard deviations. It is clear that patients in cluster 1 had low BMI, HbA1c, and EAG levels.Cluster 2 has high BMI values and intermediate HbA1c and EAG values, while cluster 3 has intermediate BMI values and low HbA1c and EAG values (Table 2 and Figure 2). In order to better address the BMI factor, the three clusters are split. A new parameter is constructed taking into account the separation of the centroids relative to the dispersion of the data in order to quantify the descriptors that are able to better differentiate the clusters with:

[2]

where xi is the value of the descriptor's property in the cluster k, and I AB is determined by taking the standard deviation from each cluster i σ.

[3]

The parameter's ability to compare the results with the test case study and to select the descriptors that statistically differ from the others the most may be proven. It is important to keep in mind that the p' parameter will provide the parameter p access to the test case's data.

The findings for the BMI descriptor were shown in Table 2 as being appropriate for clustered parting. The HbA1c and EAG descriptors' lowest values of p' indicate that they are less important for cluster separation.

Table 2: The optimum k-means clustering variables for the underlying condition of arterial hypertonia

Descriptor	Cluster 1	Cluster 2	Cluster 3	p’
BMI	28 (1.2)	32 (1.6)	29 (1.3)	3.4
HbA1c	9.4 (0.8)	8.9 (0.7)	7.2 (0.6)	2.2
EAG	12.4 (1.3)	11.5 (1.1)	8.9 (0.9)	2.3

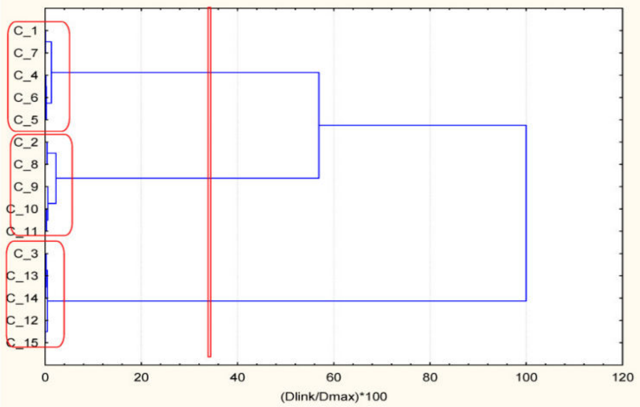

Figure 2: 15 sporadically chosen patients with diabetes mellitus type 2 (DMT2) and arterial hypertonia (AH) as an underlying ailment were clustered using a hierarchical dendrogram.

Therefore, k-means partitioning descriptors for AH underlying disease hierarchical clustering was carried out in order to demonstrate the effectiveness of the newly selected methodology that was picked best. In fact, three distinct clusters were discovered. With regard to the most ideal descriptor BMI, they matched to the predefined groups of similarity (good health status (cluster 1), intermediate health status (cluster 2), and worsened health status (cluster 3); the other two descriptors reflect the glucose level, and while they don't completely fit into the separation of the health position cited above, their discriminating probability (p') is lesser.It is possible to believe that the dominant metabolic syndrome and the blood glucose level are the two main factors influencing the health state in this scenario (less significant).

BMI, creatinine, and cholesterol are the most suitable descriptors in Table 3's analysis of CHD patients using combinatorial k-means clustering to determine the minimal global variance.

Table 3: variables of the best k-means clustering for the underlying illness of ischemic heart disease.

Descriptor	**Cluster 1**	**Cluster 2**	**Cluster 3**	**p'**
BMI	30 (1.4)	31 (1.3)	29 (1.2)	1.8
Creatinine	62 (7)	135 (9)	83 (16)	6.2
Cholesterol	6 (0.6)	4.2 (0.4)	4.1 (0.9)	2.5

Figure 3 displays a comparable control chart with the underlying issue of ischemic heart disease (CHD) and hierarchal clustering with the descriptors for patients with DMT2:

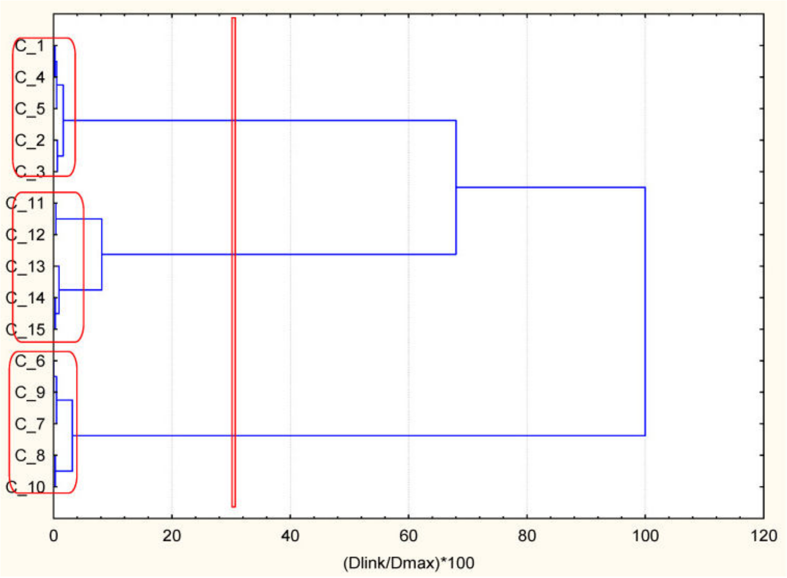

Figure 3: A hierarchical dendrogram was utilised to group together 15 patients who were randomly chosen and had underlying disorders of CHD and DMT2.

Figure 4 clearly demonstrates the division of 15 randomly chosen patients with DMT2 and CHD into three clusters. Because BMI and cholesterol are less discriminating parameters among the three best ones, the creatinine parameter (kidney function impact) is the best descriptor in this situation. Clusters 1 and 3 are similar to states of good health, severe health, and intermediate health, respectively. The optimal grouping of patients with underlying DPNP disease for BMI, HbA1c, and EAG parameters comes from combinatorial k-means analysis.

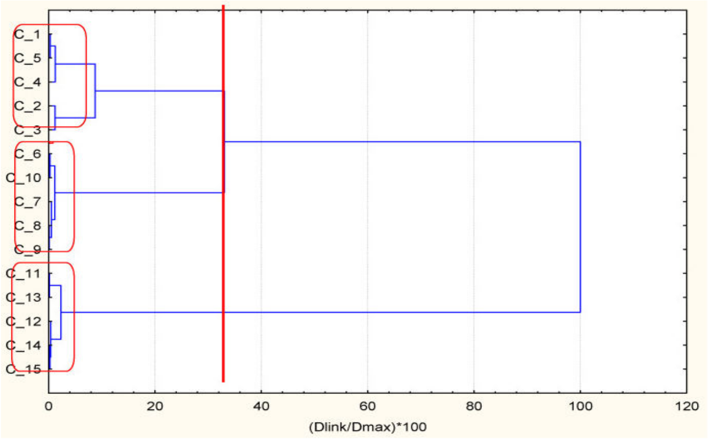

Figure 4: 15 randomly chosen individuals with DMT2 and diabetic polyneuropathy (DPNP) as underlying illnesses were clustered using a hierarchical dendrogram.

Hierarchical clustering does a good job of illustrating the division into the three groupings. The connection resembles the situation with the underlying disease in AH. Since the criteria relating to glucose levels are of secondary importance, BMI is the best indicator of the metabolic syndrome. The three clusters' health status, however, is extremely comparable (Table 4's descriptor values show this), and the separation into the three categories of health status is fairly conditional (especially cluster 1 and cluster 2). The patients with good health status who were chosen at random are included in Cluster 3. The optimal group separation for k-means clustering of BMI, creatinine, and cholesterol descriptors is provided by combinatorial k-means analysis for DMT2 patients with DMA as the underlying disease (Table 5).

Table 4: The optimum k-means clustering variables for the illness that underlies DPNP

Descriptor	Cluster 1	Cluster 2	Cluster 3	p’
BMI	32 (1.2)	34 (1.6)	30 (1.1)	3.4
HbA1c	9.4 (0.8)	9.0 (0.7)	7.3 (0.6)	2.1
EAG	12.4 (1.3)	11.7 (1.1)	8.9 (0.9)	2.2

Table 5: The optimal k-means clustering variables for the underlying condition of diabetic microangiopathy (DMA).

Descriptor	Cluster 1	Cluster 2	Cluster 3	p’
BMI	30 (1.4)	31 (1.3)	29 (1.2)	2
Creatinine	62 (7)	135 (9)	83 (15)	6.2
Cholesterol	6 (0.6)	4.2 (0.4)	4.1 (0.9)	2.5

The final case (patients with underlying DMT2 and DMA illnesses) completely replicates the pattern of case 2 (patients with DMT2 and CHD as underlying diseases). It is important to note that creatinine was chosen as the best descriptor to identify the three groups of the 15 randomly chosen objects.

4. Conclusion

In this study, we looked into a novel k-means procedure's potential for locating the descriptors that carry out clustering with the best possible separation between groups of object similarity. Using a combinatorial approach, this process chooses the descriptors that have the lowest global variance (i.e., the variance of all the descriptors in all the groups). Prediction separations are an appropriate technique that may be used to classify descriptors and to more effectively divide items into groups. It was shown that there is no prerequisite for this process that all variables for all objects must be known.

In conclusion, hierarchical clustering is proven to be effective and appropriate for screening investigations as well as in conjunction with the novel methodology to group the objects of interest into an initial set of clusters using the most advantageously chosen descriptors. This is the first study to examine the applicability of these partitioning-based models.

References:

1. Nedyalkova, M.; Madurga, S.; Simeonov, V. Combinatorial K-Means Clustering as a Machine Learning Tool Applied to Diabetes Mellitus Type 2. Int. J. Environ. Res. Public Health 2021
2. Larabi-Marie-Sainte, S.; Aburahmah, L.; Almohaini, R.; Saba, T. Current Techniques for Diabetes Prediction: Review and Case Study. Appl. Sci. 2019, 9, 4604. [CrossRef]
3. Anuja, V.; Chitra, R. Classification of Diabetes Disease Using Support Vector Machine. Int. J. Eng. Res. Appl. 2013, 3, 1797.
4. Rajesh, K.; Sangeetha, V. Application of Data Mining Methods and Techniques for Diabetes Diagnosis. Int. J. Eng. Innov. Technol. 2012, 2, 224–229.
5. Aiswarya, I.; Jeyalatha, S.; Ronak, S. Diagnosis of diabetes using classification mining techniques. Int. J. Data Mining Knowl. Manag. Process. 2015, 5. [CrossRef]
6. Harleen; Pankaj, B. A Prediction Technique in Data Mining for Diabetes Mellitus. J. Manag. Sci. Technol. 2016, 4, 1–12.
7. Haq, A.U.; Li, J.P.; Khan, J.; Memon, M.H.; Nazir, S.; Ahmad, S.; Khan, G.A.; Ali, A. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. Sensors 2020, 20, 2649. [CrossRef]
8. Krishnaveni, G.; Sudha, T. A novel technique to predict diabetic disease using data mining–classification techniques. Int. Conf. Innov. Appl. Eng. Inf. Technol. 2015, 3, 5.
9. Raj, A.; Vishnu, P.; Kavita, B. K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA. Int. J. Soft Comput. Eng. 2013, 2, 436.
10. Gadekallu, T.R.; Khare, N.; Bhattacharya, S.; Singh, S.; Maddikunta, P.K.R.; Ra, I.-H.; Alazab, M. Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model. Electronics 2020, 9, 274. [CrossRef]
11. Pranto, B.; Mehnaz, S.M.; Mahid, E.B.; Sadman, I.M.; Rahman, A.; Momen, S. Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. Information 2020, 11, 374. [CrossRef]
12. Ryu, K.S.; Kang, H.Y.J.; Lee, S.W.; Park, H.W.; You, N.Y.; Kim, J.H.; Hwangbo, Y.; Choi, K.S.; Cha, H.S. Screening Model for Estimating Undiagnosed Diabetes among People with a Family History of Diabetes Mellitus: A KNHANES-Based Study. Int. J. Environ. Res. Public Health 2020, 17, 8903. [CrossRef] [PubMed]

13. Ryu, K.S.; Lee, S.W.; Batbaatar, E.; Lee, J.W.; Choi, K.S.; Cha, H.S. A Deep Learning Model for Estimation of Patients with Undiagnosed Diabetes. Appl. Sci. 2020, 10, 421. [CrossRef]
14. Massart, D.L.; Kaufman, L. The Interpretation of Analytical Data by the Use of Cluster Analysis; John Wiley & Sons: New York, NY, USA, 1983.
15. Vogt, W.; Nagel, D.; Sator, H. Cluster Analysis in Clinical Chemistry: A Model; John Wiley & Sons: New York, NY, USA, 1987.