ORIGINAL ARTICLE

Using data mining methods for risk assessment and intervention planning in diabetic patients

Vanisree Ramanathan[1], Sharyu Mhamane[2], Jayesh Pawar[3], Nisha PK[4] Ujjwal Kumar[5], Shailesh Tripathi[6], Keerti B Pradhan[7], Sudip Bhattacharya[8]

[1]Department of Public Health, Dr Vishwanath Karad's MIT World Peace University, Pune
[2]Advanced Centre for Treatment Research and Education in Cancer, Centre for Cancer Epidemiology, Tata Memorial Centre, Navi Mumbai
[3] Jehangir Clinical Development Centre, Pune
[4]Department of Statistics, Savitribai Phule Pune University, Pune.
[5,8]Department of Community & Family Medicine, All India Institute of Medical Sciences, Deoghar, Jharkhand
[6]Department of Hospital Administration, Rajendra Institute of Medical Sciences, Ranchi, Jharkhand
[7]Department of Healthcare Management, Chitkara University, Punjab

CORRESPONDING AUTHOR
Dr Sudip Bhattacharya, Assistant Professor, Department of Community & Family Medicine, All India Institute of Medical Sciences Deoghar, Jharkhand 814152
Email: drsudip81@gmail.com

CITATION
Ramanathan V, Mhamane S, Pawar J, Nisha PK, Kumar U, Tripathi S, Pradhan K, Bhattacharya S. Using data mining methods for risk assessment and intervention planning in diabetic patients. Indian J Comm Health. 2024;36(2):278-284. https://doi.org/10.47203/IJCH.2024.v36i02.019

ARTICLE CYCLE
Received: 15/02/2024; Accepted: 05/04/2024; Published: 30/04/2024

Abstract

Introduction: Data mining in healthcare is a nascent arena of research in healthcare. Heterogeneity of Diabetes Mellitus in terms of clinical presentation calls for newer methods of research to study potential risk factors. **Aim:** The paper aims to use clustering techniques to identify the relationship between the four variables, namely the pre-prandial and postprandial sugar level, age and sex. **Methods:** The data was taken from a diagnostic laboratory in Wagholi, Pune. We conducted K-mean algorithm, EM algorithm, model-based clustering and t-mixture model. **Results:** It is evidenced that the data was best fitted to the t-mixture model. Our 50% samples were people with diabetes, 17% had prediabetes. Trivial correlation existed between age and sugar level. Males and females were equally at risk of having diabetes. Data presented concludes that age and sex have no effect on the risk of having diabetes. Data mining can be used to deduce meaningful clusters to drive plan-based interventions in the population. **Conclusion:** Methods of data mining can be used to deduce meaningful clusters in a heterogeneous dataset thus providing policymakers and healthcare researchers with novel information that will potentially contribute in formulating evidence-based policies.

Keywords

Cluster analysis, Data mining, Diabetes Mellitus, Medical Informatics

INTRODUCTION

In recent years, there has been a steady rise in both the occurrence and frequency of Diabetes Mellitus (DM). Approximately 422 million individuals, predominantly from low- and middle-income nations, are affected by diabetes. This number is projected to surge to 592 million cases by 2035. The global economic impact of DM is striking, amounting to US $1.31 trillion or 1.8% of the world's GDP. Remarkably, indirect costs constitute 34.7% of this total burden (1,2,3).

Diabetes Mellitus, marked by heightened levels of glucose in the bloodstream, stems from insufficient insulin production by the pancreas, leading to a persistent and enduring metabolic condition known as Diabetes Mellitus, (2). Diabetes Mellitus (DM) significantly elevates the risk of mortality, with approximately one out of every twelve deaths across all causes being attributable to this condition. (4) Diabetes Mellitus (DM) is categorized into two primary types: type 1 and type 2 diabetes. Type 1 diabetes is typically diagnosed at a younger age and is characterized by the presence of autoantibodies against pancreatic islet-cell antigens. Using this criterion, type 2 diabetes accounts for 75–85% of cases (5,6). The clinical presentation and progression of DM demonstrate variability, leading to challenges in generalizing and accurately identifying risk factors and parameters associated with the onset and advancement of the condition (5). As the public health burden of DM continues to grow, it is crucial to comprehend the shared characteristics that could potentially act as risk factors. Therefore, it is essential to embrace innovative research methodologies to explore the disease further. One emerging field of study in healthcare is data mining and the utilization of machine learning techniques (8). Data mining presents an enticing avenue for research, given the immense volume of healthcare data available and the need for sophisticated data analysis tools. Through data mining tools and techniques, it becomes feasible to uncover and understand concealed patterns within a dataset that may not be apparent through straightforward data presentation. Selecting the optimal clustering method and determining the appropriate number of clusters for healthcare data often proves intricate and demanding (9). Given the diversity within DM progression, cluster analysis offers a method to discern shared patterns among different individuals. Cluster analysis, a statistical method, categorizes objects or characteristics into clusters, ensuring that those within the same cluster exhibit greater statistical similarity compared to those in different clusters (10). The paper seeks to undertake a similar endeavor by employing clustering algorithms to identify significant clusters within the dataset of individuals with diabetes.

MATERIAL & METHODS

The dataset comprises of information about two variables, viz., fasting sugar level and postprandial sugar level (PP) along with the associated age, and sex of the individuals. This information is collected from 132 individuals who approached a particular diagnostic laboratory in Wagholi, Pune in the month of November 2022. All the cases (no sampling done) related to testing of fasting sugar level and postprandial sugar level during the month of November 2022 is considered. The objective of the study is to explore the data to identify potential meaningful clusters using various clustering and data mining techniques. The methods employed include the K-means clustering algorithm, Expectation-Maximization (EM) algorithm, model-based clustering, and t-mixture model methods for data analysis.

RESULTS

A K-means clustering algorithm has been carried out first (K-means clustering is an algorithm which groups the unlabeled data set into different clusters in such a way that each dataset belongs to only one group that has similar properties. Here K is the number of clusters which pre-defined) and it gives an optimal number of clusters as K=2, K=8. Multiple outlier points exist in the data and K-mean considers them as either separate clusters or includes them in the nearby cluster. Thus, if the number of clusters is less, the cluster variability increases and if the number of clusters is more, each outlier is considered as a single cluster.

The data has been checked for cluster tendency using Hopkin's statistic and H = 0.8147 > 0.5 indicates there is a strong clustering tendency. According to the visual assessment of cluster tendency, there are at least 3 clusters present in this data.

An EM algorithm, (Expectation- Maximization (EM) algorithm is an iterative procedure to find maximum likelihood estimates (MLE) or maximum a posteriori (MAP) estimates of the parameters in a statistical model, where the model is based on the unobserved latent variables. The algorithm performs an expectation step (E-step) and a maximization step (M-step) alternatively. At each E-step, the function corresponding to the expectation of the log- likelihood function is calculated using the current estimates of the parameters and each M-step computes the parameters which maximize the log-likelihood function found in the E-step. The E-step and the M-step are repeated until converges) (11) has been carried out for the data and it gave three clusters with a number of observations 126, 4 and 2 and with mixing proportions 0.0455, 0.1724, 0. 7803. An EM algorithm with different initialization has been carried to check the convergence rate of each the initialization methods. Different initializations like K- means, small em, emEM, RndEM etc. has been adopted and EM algorithm has been carried out. Both RndEM and emEM converges fast compared to other initialization and divides the data into 3 clusters with 43, 38 and 51 points with corresponding mixing proportions 0.3543, 0.2786, 0.3672 and with a log likelihood = -300.5575, Akaike Information Criterion (AIC) = 659. 1150 and Bayesian Information Criterion (BIC) = 742.7164.

Model-based clustering (MBC) is a statistical technique used to group data, assuming that the observed multivariate data arises from a finite mixture of component models. Each component model represents a probability distribution, typically a parametric multivariate distribution. For instance, in a multivariate Gaussian Mixture Model (GMM), all components follow a multivariate Gaussian distribution, while in a t-mixture model, each component adheres to a multivariate t distribution. (12) has been carried out for the data. Different information criteria are used to select the best model which is suitable for the data. AIC, BIC and ICL are commonly used information criterias. According to AIC, BIC and Integrated Complete-data Likelihood (ICL), the best model is the Gaussian mixture model with K=3 components. However, there are a lot of uncertainty points that exist in this fit. The data has been checked for multivariate normality. According to Mardia's skewness and kurtosis test, Doornik-Hansen test and Royston's test, the data does not follow a multivariate normal distribution. Individual variables are tested for univariate normality and none of them was found to be following a normal distribution.

Figure 1- Uncertainty plot using a t-mixture model with an Integrated Complete data Likelihood Criterion (bigger blue dots and red dots indicates doubtful points (or we are not sure in which cluster they belongs to).

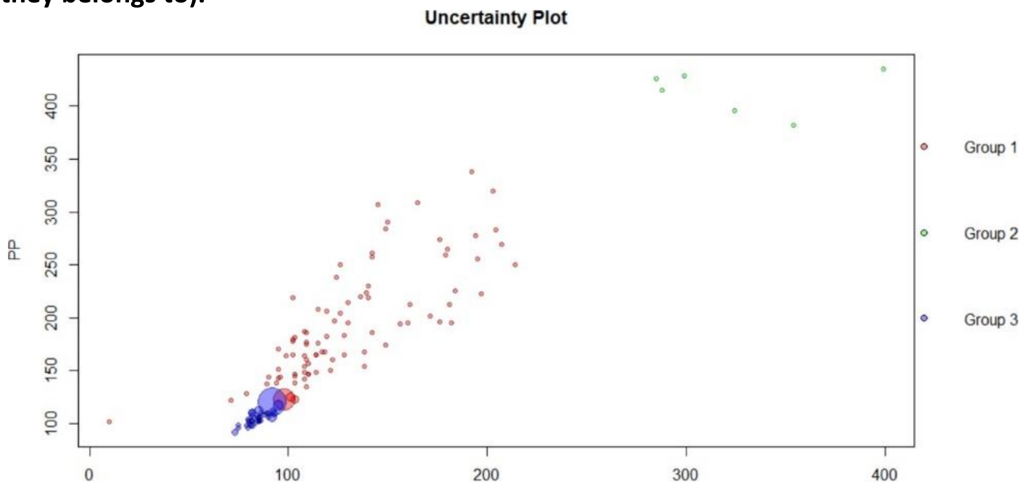
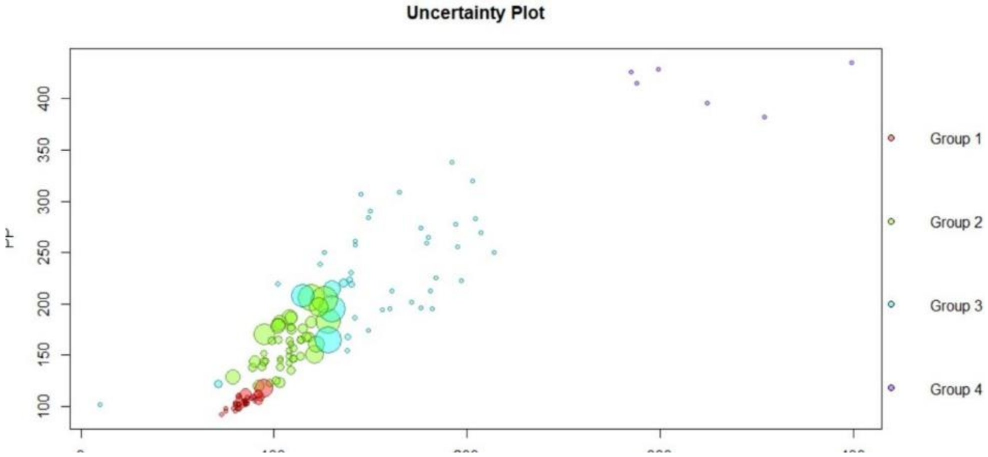


Figure 2-Uncertainty plot using t-mixture model with Bayesian Information Criterion.



A t-mixture model has been fitted to the data and both ICL and BIC choose the t-mixture model with different K. ICL chooses the best model as the model with K=3 with an ICL value -3665 and BIC chooses the best model with K=4 with a BIC value -3660.28. An uncertainty plot gives an idea about the points which are not certain that in which clusters these points are to be included.

The uncertainty plot shows less number doubtful points compared to the Gaussian mixture model (GMM) fit. The uncertainty points are less in ICL selected model (Figure 1) compared to that of the BIC-selected model and the likelihood function is also maximum in the ICL-selected model (Figures 2).

Thus, we can conclude that the best model is a t-mixture model with three clusters (K=3), that is, the data possess three clusters.

Some outlier points were present in the data and one individual with a fasting sugar level of 10 is doubtful as it must be a measurement recording error.

These points have been removed from the data and the analysis has been carried out again. The optimum number of clusters for a K-means clustering as K=3 and (Figure 3) shows K-means clustering with K=3.

(Figure 4) shows the GMM fit with K=3, after removing the outlier points as the best model, according to ICL. Here also we can see a lot of uncertainty points. A multivariate normality test has been carried out and it was found that the data does not follow a multivariate normal distribution. Therefore, fitting a Gaussian mixture model is not very appropriate in this context.

A t-mixture has been fitted to the data after removing the outlier points. ICL chooses the model with K=2 as the best model with an ICL value -3234 and BIC chooses the best model as the model with K=3 (Figure5).

Figure 3- K-means clusters with 3 distinct clusters (after removing the outliers).

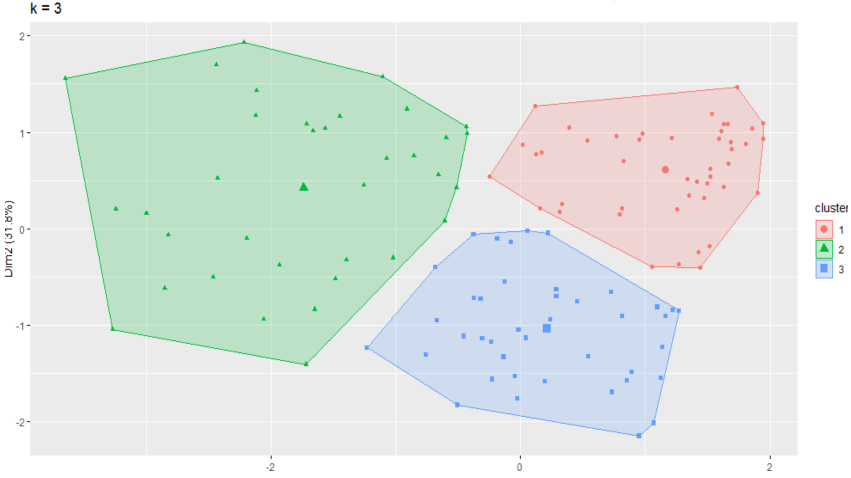


Figure 4- Uncertainty plot using Gaussian Mixture Model (after removing the outliers)

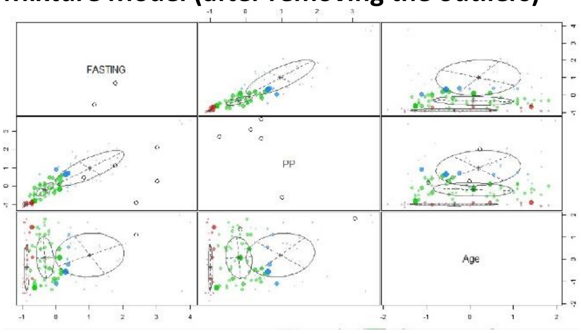
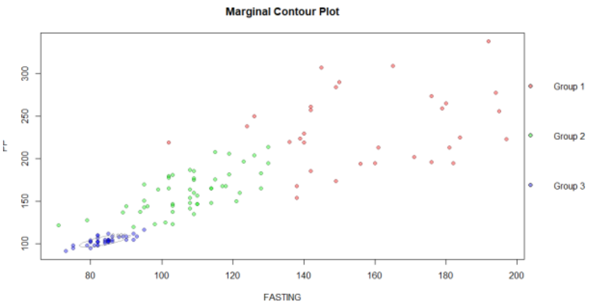


Figure 5 Uncertainty plot using t-mixture model with Bayesian Information Criterion (after removing the outliers).



Number of uncertainty points are less in both the cases compared to the GMM fit. Note that the number of uncertainty points are lesser in ICL selected model, as compared to the BIC selected model (Figures 6 & 7). Thus, we conclude that the best model is a t- mixture model with K=2 and the data possess only two clusters after eliminating the outliers.

Figure 6- Uncertainty plot using t-mixture model with Integrated Complete data Likelihood criterion (after removing the outliers).

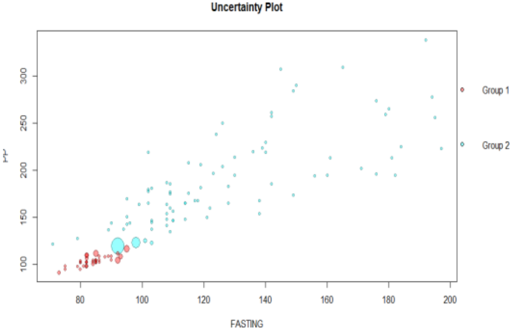
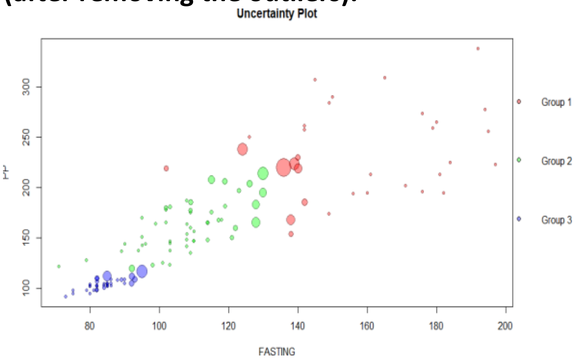
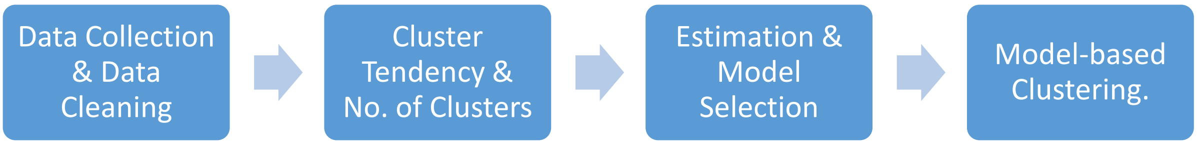


Figure 7- Uncertainty plot using t-mixture model with Bayesian Information Criterion (after removing the outliers).



The flow chart of the entire process can be escribed as follows:



Data Collection & Data Cleaning

Cluster Tendency & No. of Clusters

Estimation & Model Selection

Model-based Clustering.

DISCUSSION

A fasting sugar level more than 125 is considered as diabetes and a fasting sugar level between 100 and 125 is considered as pre-diabetes and a fasting sugar level less than100 is treated as normal category. Most of people shows symptom of diabetes after the age of 30 according to this data. There are people with high blood sugar level in all age categories starting from age 30 to age 80. Almost half of the people (50%) from this data were suffering from diabetes and the remaining 33% falls under normal category and the remaining 17% comes under the pre- diabetes category. Also, males and females are almost equally at a risk of suffering from diabetes and pre-diabetes. The correlation between age and sugar level is very small (0. 2464) and the correlation between sex of the person with diabetes is also very small (-0.0961), according to this data. Therefore, we can conclude that age and sex have no effect on being a person suffering from diabetes according to this data.

However, a few studies show an increased risk and positive correlations between young age and DM and gender and DM with women being at a higher risk of DM. (13,14,15,16) A study conducted in China to identify the unhealthy lifestyle among elderly population showed that females were at lower risk of suffering from DM (13) while Min Kyung Lee reported no difference between both sexes. (17)

From Figure 1, we can see that most of the observations lie in the category of having diabetes. The first cluster indicates the normal category (blue spots), second cluster includes both the pre-diabetes and diabetes categories (red spots) and the third cluster consists of patients with higher sugar levels (green spots). Figure 1 indicates 4 clusters with normal, pre-diabetes, diabetes and high diabetes and we can consider people with higher diabetes as outliers.

After removing the outliers, Figure 6 indicates that there exist only two clusters, the first one normal group and the second group consisting of both having pre-diabetes and diabetes. In Figure 5, we can see three clusters which represent normal, people having pre-diabetes and diabetes. Even though Figure 5 has a larger number of uncertainty points, it gives a matching result with the sugar level limits according to the benchmark sugar testing levels.

This study has its own limitations, especially with respect to the sample size. It is based only on the total number of test cases that occurred in a particular laboratory in Pune during the month of November 2022. There is a possibility of extending this to more laboratories operating in Pune and thereby increasing the sample size. Also, a temporal extension is possible, by repeating the same exercise during another periods of time. Another limitation is that in terms of the non-availability of information related to other possible related variables. In most cases, the laboratories are not willing to share the test details, even for carrying out academic research.

Conclusion

The primary causes of ill health and early mortality worldwide are non-communicable diseases (NCDs), which include cardiovascular diseases, type 2 diabetes, malignancies, and chronic respiratory disorders. (18) This method in different populations and healthcare systems would merit replication. (19) Methods of data mining can be used to deduce meaningful clusters in a heterogeneous dataset.

Recommendation

These types of research when implemented on a wider spectrum of population for larger data can help in identifying newer peculiarities of existing health adversities thus providing policymakers and healthcare researchers with novel information that will potentially contribute in formulating evidence-based policies.

Authors Contribution

VR, SM, JP and NPK-Conceptualization, data analysis and first draft writing, UK, ST and SB-reviewing and finalizing the draft.

Financial Support and Sponsorship

Nil.

Conflict of Interest

Nil.

Acknowledgement

The authors are thankful to T.V Ramanathan for a critical review and valuable input in the discussion.

Declaration of generative AI and AI assisted technologies in the writing process

The authors haven’t used any generative AI/ AI assisted technologies in the writing process.

References

1. World Health Organization. Diabetes. Accessed December 4, 2022. https://www.who.int/health-topics/diabetes
2. Guariguata L, Whiting DR, Hambleton I, et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract. 2014;103(2):137-149.
3. Bommer C, Heesemann E, Sagalova V, et al. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. Lancet Diabetes Endocrinol. 2017;5(6):423-430.
4. Sarría-Santamera A, Orazumbekova B, Maulenkul T, et al. The Identification of Diabetes Mellitus Subtypes Applying Cluster Analysis Techniques: A Systematic Review. Int J Environ Res Public Health. 2020;17(24):9523.
5. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their

association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361-369.

6. Eby EL, Edwards A, Meadows E, et al. Evaluating the relationship between clinical and demographic characteristics of insulin-using people with diabetes and their health outcomes: a cluster analysis application. BMC Health Serv Res. 2021;21(1):669.
7. Wang X, Gao H, Xu H. Cluster Analysis of Unhealthy Lifestyles among Elderly Adults with Prediabetes: A Cross-Sectional Study in Rural China. Diabetes Ther Res Treat Educ Diabetes Relat Disord. 2019;10(5):1935-1948.
8. Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? BMC Public Health. 2020;20(1):798.
9. Ogbuabor G, F. N U. Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. Int J Comput Sci Inf Technol. 2018;10(2):27-37.
10. Robertson L, Vieira R, Butler J, et al. Identifying multimorbidity clusters in an unselected population of hospitalised patients. Sci Rep. 2022;12(1):5134.
11. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol. 2008;26(8):897-899.
12. Meila M, Heckerman D. An Experimental Comparison of Model-Based Clustering Methods. Mach. Learn. 2001;42: 9-29.
13. Wang T, Zhao Z, Wang G, et al. Age-related disparities in diabetes risk attributable to modifiable risk factor profiles in Chinese adults: a nationwide, population-based, cohort study. Lancet Healthy Longev. 2021;2(10):e618-e628.
14. Bahour N, Cortez B, Pan H, et al. Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers. GeroScience. 2022;44(1):415-427.
15. Nguyen QM, Xu JH, et al. Correlates of Age Onset of Type 2 Diabetes Among Relatively Young Black and White Adults in a Community. Diabetes Care. 2012;35(6):1341- 1346.
16. Zhang L, Yang H, Yang P. The Correlation between Type 2 DiabetesMellitus and Cardiovascular Disease Risk Factors in the Elderly. Appl Bionics Biomech. 2022;2022: e4154426.
17. Lee MK, Han K, Kwon HS. Age-specific diabetes risk by the number of metabolic syndrome components: a Korean nationwide cohort study. Diabetol Metab Syndr. 2019;11(1):112.
18. Uddin R, Lee EY, Khan SR, et al. Clustering of lifestyle risk factors for non- communicable diseases in 304,779 adolescents from 89 countries: A global perspective. Prev Med. 2020; 131:105955.
19. Lefèvre T, Rondet C, Parizot I, et al. Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area. Divaris K, ed. PLoS ONE. 2014;9(12):e115064.