Supplementary material

Uncovering Heterogeneous Cardiometabolic Risk Profiles in U.S. Adults: The Role of Social and Behavioral Determinants of Health

Ding Q, *et al. BMJ Open Diab Res Care* 2023; 11:e003558. doi: 10.1136/bmjdrc-2023-003558

Supplementary methods

Supplementary methods 1: Poverty-income ratio categories

We calculated poverty-income ratio (PIR) by using family income divided by the federal poverty level. PIR in this study was categorized as <1.30 (low family income; income below approximately $20,000 for a family of four); 1.31-1.85 (low middle income; income of $20,000 to $28,000 for a family of four); 1.86-3.50 (middle income; income $28,000 to $53,000 for a family of four); and >3.5 (high income; income of more than $53,000 for a family of four).[1]

Supplementary methods 2: Dietary pattern score

Dietary intake derived from the continuous NHANES dietary data[2] was evaluated on 5 diet elements, including 1) >4.5 cups/day of fruits and vegetables; 2) ≥2 servings/week of fish; 3) ≥3 servings/day of whole grains; 4) no more than 36oz/week of sugar-sweetened beverages; and 5) no more than 1500mg/day of sodium.[1] Participants received a score of 1 if they met one of the above five elements. Dietary pattern score was the sum of these five element scores and ranged from 0 to 5.[3]

Supplementary methods 3: Physical activity levels

We followed the Centers for Disease Control and Prevention's Physical Activity Guidelines for Americans[4] to create the physical activity variable in the study analysis. Aerobic physical activity was categorized into insufficiently active, active, and highly active.[3] Inactive was defined as not getting any moderate- or vigorous-intensity physical activity beyond basic movement from daily life activities; insufficiently active was defined as doing some moderate- or vigorous-intensity physical activity but less than 150 minutes of moderate-intensity physical activity a week or 75 minutes of vigorous-intensity physical activity or the equivalent combination; active was defined as doing the equivalent of 150 minutes to 300 minutes of moderate-intensity physical activity a week; and highly active was referred to as doing the equivalent of more than 300 minutes of moderate-intensity physical activity a week. This level exceeds the key guideline target range for adults. Such classification has been shown to be related to how much health benefit a person gets at a given level and how to become more active.

NHANES did not use consistent self-reported physical activity measures across the 20 years (1999-2018), and NHANES (1999-2006) and NHANES (2007-2018) used different coding systems or physical activity related variables. As a result, we organized the physical activity level variables separated in NHANES (1999-2006) and NHANES (2007-2018) dataset, and then used the estimation of the equivalent of moderate-intensity aerobic physical activity a week along with other supplementary variables to define our physical activity level variable across the 20 years. Due to limited number of individuals belonging to highly active and insufficiently active groups, we merged inactive and insufficiently active to create inactive, and merged active and highly active into active.

Ding Q, *et al. BMJ Open Diab Res Care* 2023; 11:e003558. doi: 10.1136/bmjdrc-2023-003558

Supplementary methods 4: Variable reduction analysis

Huang's k-prototype clustering procedure did not specify the method for variable reduction, and it has a risk of multicollinearity between attributes.[5] To strengthen our analytical approach, we performed variable reduction to select the least number of attributes (variables) that were judged to be representative of the three domains and were interpretable.

We followed SAS PROC VARCLUS procedure and performed principal component analysis on numeric variables in each domain.[6 7] Numeric variables were standardized with a mean of zero and one standard deviation. PROC VARCLUS uses a divisive hierarchical clustering method to create variable subgroups that are as dissimilar as possible. The iteration process includes three steps:

1. Choose a cluster with the greatest variability to split. Greatest variability in our study was measured by the largest eigenvalue associated with the second principal component (We chose Maxeign=0.7 in the analyses);
2. Split the chosen cluster into two clusters by finding the first two principal components, rotating the components, and then assigning each variable to the rotated component with which it has the highest squared correlation;
3. Iteratively assign the variables to the clusters to maximize the variance accounted for in the cluster components.

PROC VARCLUS stops either when each cluster has only one single eigenvalue greater than 1, or if the specified number of clusters, percent variation or second eigenvalue is reached.[7] The variable reduction for 20 numeric variables for all three domains reduced to 16 numeric variables that explain 87.1%-100% of the total data variance. For categorical variables, we removed variables if other included variables reflected similar physiological parameters or socio/behavioral characteristics. Categorical variables that were considered significant determinants of socioeconomic status (e.g., Medicaid) and lifestyle behaviors (e.g., alcohol use) were also included. Clinicians that were not directly involved in the study were consulted for selection of cardiometabolic risk factors.

Supplementary methods 5: K-prototypes cluster analysis

K-prototypes algorithm is an extension to the k-means algorithm for clustering large datasets with categorical and numeric attributes.[8] Most existing clustering algorithms can either handle large datasets efficiently but limited to numeric values or can handle mixed data by converting categorical variables (e.g., sex and race) into numeric values[9] or by discretizing numeric attributes and applying categorical clustering algorithms.[5] These strategies can lead to information loss and often are not efficient when clustering large dataset.[10] Huang proposed a cost function, an integration of Euclidean distance and simple matching dissimilarity, to handle mixed data.[6 8] The dissimilarity measure between two objects can be expressed as

$$s^r + \gamma s^c$$

s^r is the dissimilarity measures on numeric attributes defined by Euclidean distance, s^c is the dissimilarity measure on categorical attributes defined as the number of mismatches of categories between two objects, and γ is a weight to balance the two parts to avoid favoring either type of attribute.[8]

Huang later adopted the partition clustering techniques and introduced K-prototypes clustering algorithm to implement the cost function that considers both numeric and categorical attributes.[11] Like other clustering algorithms, the goal of K-prototype clustering is to put individuals into k groups or clusters, so that each person is put into only one group; individuals within a group share similar attributes; and individuals in different groups have different attributes. The k clusters form a partition of the dataset into k mutually exclusive subsets. Each person is described by their attribute vector x_i, i=1…,N, which is a combination of numeric variables (e.g., age) and categorical variables (e.g., sex). All attribute vectors contain the same suite of variables. The numeric attributes occupy the first p positions of the attribute vector, and the categorical attributes occupy the last q positions, giving:

$$x_i = (\underbrace{x_{i1}, x_{i2}, \ldots, x_{ip}}_{\text{numeric}}, \underbrace{x_{i(p+1)}, \ldots, x_{i(p+q)}}_{\text{categorical}})$$

The K-prototypes algorithm works in an iterative fashion as in the following steps[8 11].

1. Initialize the centroid locations of the clusters by selecting k individuals as "prototype" centroids.
2. Assign each person to the cluster with the nearest centroid.
3. Compute an overall cost of the allocation by calculating total distance of all individuals from their assigned centroids.
4. Update cluster centroids.
5. Re-allocate the person with the nearest centroid using the updated centroids.
6. Compute the overall cost again.
7. Iterate step 4 to 6 until no change in overall cost and cluster output.

Supplementary methods 5: K-prototype clustering steps in the study analyses

Step 1: In our analysis, we randomly select k individuals to serve as the initial centroids of the clusters. The initial centroids are given by the attribute vectors of the randomly chosen k individuals and are denoted by

$$c_l = (\underbrace{c_{l1}, c_{l2}, \ldots, c_{lp}}_{\text{numeric}}, \underbrace{c_{l(p+1)}, \ldots, c_{l(p+q)}}_{\text{categorical}}), l = 1, \ldots, k,$$

where c_{lj} represents the cluster-l, attribute-j centroid.

Step 2: After initializing the cluster centroids, we use the distance metric[8] to measure dissimilarity between individual I and cluster l:

$$d(x_i, c_l) = \sum_{n=1}^{p} \sqrt{(x_{in} - c_{ln})^2} + \sum_{n=p+1}^{p+q} \delta(x_{in}, c_{ln}),$$

where $\delta(a,b) = \begin{cases} 1 \ if \ a \neq b \\ 0 \ for \ a = b \end{cases}$

For individual I, the distances between its attribute vector and each of the l cluster centroids were computed using the distance metric $d(x_i, c_l)$, $l = 1, \dots, k$. The individual was then placed in the closest cluster with minimum distance. The process repeated for all individuals in our analytic dataset and each person was assigned to exactly one of the l clusters.

Step 3: After all individuals in the dataset were assigned to a cluster, we used the formular below to calculate the overall distance between individuals and their cluster centroid:
Cost function $j = \sum_{l=1}^{k} \sum_{i \in Ne} d(x_i, c_l)$

Step 4: The cluster centroids were updated by finding the middle for each cluster's attributes. For numeric variables, the centroids were updated to be the within cluster average value. The updated j-th attribute for cluster l is

$$c_{lj} = \frac{1}{n_l} \sum_{i \in Nl} x_{ij}, j = 1, \dots, p.$$

For categorical variables, attributes of each cluster will be updated using the mode function (denoted by $\mathbb{M}$) given by

$$c_{lj} = \mathbb{M}(X_{IJ} | i \ \in N_l)$$

Step 5: Based on the updated cluster centroids, we re-located each person to clusters using the minimum distance between the individual's attribute vector and the updated cluster centroids.
Step 6: Total cost was calculated again using cost function in Step 3.
Step 7: If the total cost was different from the previous iteration, we updated cluster centroids again, and re-allocated the individual. This was repeated until the total cost function was unchanged.
The best and final cluster was determined by having the minimum cost function over all randomly chosen initial cluster centroids.

Supplementary tables

Supplementary table 1. Variables (n=29) used to determine unique cardiometabolic risk clusters in U.S. adults without known diabetes, grouped by three commonly known risk factor categories.

Domain	Variables
Social determinants of health	Race/ethnicity
	Income evaluated by poverty ratio
	Education
	Marital status
	Health insurance status
	Government Insurance
	Private Insurance
	Medicaid
	Access to health care
	Age
	Sex
Lifestyle behavioral risk factors	Dietary intake measured by dietary pattern score
	Physical activity
	Smoking status
	Alcohol use
	Sleep duration
Cardiometabolic risk factors	Body mass index (BMI)
	Waist circumference
	Total cholesterol
	Triglycerides
	High-density lipoprotein cholesterol
	Low-density lipoprotein cholesterol
	Fasting glucose
	Hemoglobin A1c
	Average systolic blood pressure
	Average diastolic blood pressure
	Uric acid (mg/dL)
	Creatinine
	Blood urea nitrogen (mg/dL)

Supplementary table 2. Cluster summary for numeric variables from social determinant of health domain with 100% variance explained in its two member variables (n=38,476).

Cluster/Domain	Variable	R^2		
		Own cluster	Next closest	1-R**2 ratio
Social determinants of health				

Cluster 1	Age	1.00	0.012	0.000
Cluster 2	Poverty ratio	1.00	0.0121	0.000

Supplementary table 3. Cluster summary for numeric variables from lifestyle behavioral risk domain with 100% variance explained in its two member variables (n=38,476).

Cluster/Domain	Variable	R^2		
		Own cluster	Next closest	1-R**2 ratio
Lifestyle behavioral risk				
Cluster 1	Sleep duration	1.00	0.0006	0.000
Cluster 2	Diet pattern score	1.00	0.0006	0.000

Supplementary table 4. Cluster summary for numeric variables from cardiometabolic risk factor domain with 87.1% variance explained in its twelve member variables (n=38,476).

Cluster/Domain	Variable	R^2		
		Own cluster	Next closest	1-R**2 ratio
Cardiometabolic risk				
Cluster 1	BMI	0.949	0.073	0.055
	Waist circumference	0.949	0.143	0.060
Cluster 2	Total cholesterol	0.961	0.136	0.045
	LDL	0.961	0.046	0.041
Cluster 3	BUN	0.745	0.074	0.275
	Creatinine	0.745	0.098	0.282
Cluster 4	HbA1c	0.863	0.063	0.146
	Fasting glucose	0.863	0.063	0.145
Cluster 5	Mean systolic BP	0.705	0.066	0.316
	Mean diastolic BP	0.709	0.041	0.308
Cluster 6	Uric acid	1.00	0.115	0.000
Cluster 7	Triglyceride	1.00	0.089	0.000

Supplementary table 5. Twenty-one variables resultant from the variable reduction procedure of the 29 variables grouped by domain.

Domain	Selected Variables
Social determinants of health	Race/ethnicity
	Income evaluated by poverty ratio
	Education
	Marital status
	Private Insurance
	Medicaid
	Access to health care

	Age
	Sex
Lifestyle behavioral risk factors	Dietary intake measured by dietary pattern score
	Physical activity
	Smoking status
	Alcohol use
	Sleep duration
Cardiometabolic risk factors	Body mass index (BMI)
	Triglycerides
	Low-density lipoprotein cholesterol
	Fasting glucose
	Average diastolic blood pressure
	Uric acid (mg/dL)
	Blood urea nitrogen (mg/dL)

Supplementary table 6. Silhouette values for the K-prototype clustering analysis (K=2-10), random state =42

No. of clusters/Silhouette value	2	3	4	5	6	7	8	9	10
Weighted average silhouette	0.083	0.078	0.076	0.066	0.054	0.020	0.040	0.052	0.050

Supplementary table 7. Cost values for the K-prototype clustering analysis (K=2-10), random state =42.

No. of clusters/Cost value	2	3	4	5	6	7	8	9	10
Cost	420,000	390,000	370,000	360,000	345,000	330,000	325,000	320,000	310,000

Supplementary table 8. Regression analyses explore relationship between individual risk factor and cluster membership.

		Cluster 2	Cluster 3	
Variable		**OR (95% CI)**	**OR (95% CI)**	**P-value**
Physical Activity Score				<0.001
	Active	ref	ref	
	Inactive	0.67 (0.59,0.76)	0.43 (0.36,0.50)	
	NA	0.29 (0.04,2.08)	0.16 (0.02,1.54)	
Sex				<0.001
	Male	ref	ref	
	Female	0.02 (0.02,0.03)	0.02 (0.01,0.02)	
Poverty Ratio				<0.001
	<=1.3	ref	ref	
	<=1.85	3.23 (2.65,3.95)	3.00 (2.37,3.79)	
	>1.85	16.98 (12.92,22.30)	12.20 (9.16,16.23)	
	NA	3.85 (3.06,4.85)	3.38 (2.58,4.44)	
Race				<0.001
	White	ref	ref	
	Black	0.16 (0.13,0.19)	0.16 (0.13,0.20)	
	Hispanic	0.03 (0.02,0.04)	0.05 (0.04,0.07)	
	Other	0.17 (0.14,0.21)	0.18 (0.14,0.24)	
Education				<0.001
	HS	ref	ref	
	Some College	2.54 (2.16,2.98)	1.04 (0.84,1.29)	
	College+	0.68 (0.58,0.80)	0.46 (0.37,0.57)	
	NA	0.35 (0.30,0.42)	1.44 (1.19,1.74)	
Marital Status				<0.001
	Single	ref	ref	
	Married/Partner	42.88 (30.88,59.54)	22.80 (16.35,31.78)	

	DK/NA	5.50 (4.20,7.19)	5.96 (3.96,8.98)	
Medicaid				0.597
	No/NS	ref	ref	
	Yes	1.06 (0.88,1.29)	0.97 (0.75,1.27)	
Private Insurance				<0.001
	No/DK	ref	ref	
	Yes	0.84 (0.74,0.94)	1.06 (0.90,1.24)	
Care Access				0.974
	No place to go	ref	ref	
	Place to go	1.03 (0.91,1.16)	1.00 (0.85,1.18)	
	DK	0.17 (0.00,1.9e+12)	0.96 (0.00,5.4e+13)	
Smoking Status				<0.001
	Light	ref	ref	
	Significant	0.98 (0.84,1.13)	0.89 (0.72,1.09)	
	Heavy	0.91 (0.79,1.04)	0.76 (0.64,0.91)	
	DK	1.38 (1.03,1.84)	0.94 (0.65,1.37)	
Alcohol				<0.001
	No	ref	ref	
	Yes	1.26 (1.09,1.47)	1.11 (0.91,1.36)	
	DK	0.60 (0.51,0.72)	1.02 (0.81,1.29)	
Age		1.09 (1.09,1.10)	1.20 (1.19,1.21)	<0.001
Total Diet Score		0.55 (0.49,0.62)	0.57 (0.50,0.66)	<0.001
LDL		1.03 (1.02,1.04)	1.02 (1.02,1.03)	<0.001
Mean DBP		1.09 (1.08,1.11)	1.05 (1.04,1.06)	<0.001
Fasting blood glucose		1.02 (1.01,1.03)	1.05 (1.04,1.06)	<0.001
BMI		1.14 (1.12,1.15)	1.09 (1.08,1.11)	<0.001
Sleep hours		0.82 (0.77,0.87)	1.02 (0.97,1.08)	<0.001

Uric acid (mg/dL)	2.65 (2.35,2.99)	3.53 (3.07,4.05)	<0.001
Blood urea nitrogen (mg/dL)	1.12 (1.10,1.15)	1.38 (1.33,1.43)	<0.001
Triglycerides	1.00 (1.00,1.01)	1.00 (1.00,1.01)	<0.001

Abbreviations: BMI=body mass index; DBP=diastolic blood pressure; DK=don't know; HS=high school and less; LDL=low-density lipoprotein; NA=not available; OR=odds ratio; ref=reference group

Supplementary Table 9. Prevalence of undiagnosed diabetes and prediabetes among U.S. adults without known diabetes.

	Undiagnosed Diabetes			Prediabetes		
	Unweighted N	Weighted N	Weighted Percent	Unweighted N	Weighted N	Weighted Percent
All	1,310	4,269,098	2.7 (2.5, 2.9)	8,698	35,637,895	22.6 (21.9, 23.4)
Cluster 1	532	2,024,598	2.2 (1.9, 2.4)	4991	23,011,718	24.5 (23.5, 25.4)
Cluster 2	75	236,753	0.5 (0.4, 0.7)	1655	5,156,640	11.4 (10.7, 12.1)
Cluster 3	703	2,007,747	11.0 (9.9, 12.1)	2322	7,469,537	40.9 (38.9, 42.8)

Abbreviations**:** N=number of participants

Supplementary table 10. Prevalence of undiagnosed diabetes in adults without known diabetes, stratified by race/ethnicity.

	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic other	Overall
Unweighted n	486	311	384	129	1,310
Weighted n	2,534,436	603,073	761,749	369,840	4,269,098
All	2.4 (2.1,2.7)	3.4 (2.9,3.9)	3.2 (2.8,3.6)	3.3 (2.6,4.1)	2.7 (2.5,2.9)
Cluster 1	1.7 (1.4,2.1)	4.0 (3.3,4.7)	3.4 (2.7,4.0)	2.7 (1.7,3.6)	2.2 (1.9,2.4)
Cluster 2	0.2 (0.0,0.4)	0.8 (0.4,1.1)	0.7 (0.4,1.0)	0.8 (0.1,1.6)	0.5 (0.4,0.7)
Cluster 3	9.3 (7.9,10.6)	15.5 (13.1,17.9)	15.7 (13.2,18.2)	17.7 (12.3,23.1)	11.0 (9.9,12.1)

Supplementary table 11. Prevalence of pre-diabetes in adults without known diabetes, stratified by race/ethnicity.

	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic other	Overall
Unweighted n	3,686	1,950	2,424	908	8,968
Weighted n	23,068,110	4,118,900	5,604,052	2,846,833	35,637,895
All	22.0 (20.9, 23.1)	23.3 (22.2, 24.3)	23.6 (22.4, 24.3)	25.6 (23.7, 27.5)	22.6 (21.9, 23.4)
Cluster 1	22.3 (21.1, 23.5)	30.3 (28.8, 31.8)	31.2 (29.2, 33.1)	31.9 (29.1, 34.8)	24.5 (23.5, 25.4)
Cluster 2	7.8 (6.8, 8.8)	13.9 (12.6, 15.1)	15.3 (13.9, 16.7)	12.8 (10.5, 15.2)	11.4 (10.7, 12.1)
Cluster 3	42.0 (39.7, 44.4)	35.2 (31.9, 38.5)	38.4 (34.8, 42.1)	38.6 (32.2, 44.9)	40.9 (38.9, 42.8)

Abbreviations: n=number of participants

Supplementary table 12. Association between cluster membership and undiagnosed diabetes and prediabetes

Outcomes	Unadjusted			Adjusted for confounders*		
Cluster label	OR	95% CI	*p* value	OR	95% CI	*p* value
Undiagnosed diabetes						
Cluster 2: Healthy socioeconomically vulnerable young adults	Ref	-	<0.001	Ref	-	<0.001
Cluster 1: Middle-aged adults with multiple metabolic risk factors	5.17	4.05 to 6.59		3.66	2.78 to 4.81	
Cluster 3: Elderly adults with chronic conditions and low physical activity levels	22.51	17.70 to 28.62		14.87	10.94 to 20.22	
Prediabetes						
Cluster 2: Healthy socioeconomically vulnerable young adults	Ref	-	<0.001	Ref	-	<0.001
Cluster 1: Middle-aged adults with multiple metabolic risk factors	3.83	3.56 to 4.13		1.79	1.62 to 1.99	
Cluster 3: Elderly adults with chronic conditions and low physical activity levels	7.40	6.80 to 8.05		2.09	1.83 to 2.39	

*Logistic models adjusted for age, sex, race, alcohol use, smoking status, diet score, BMI, history of hypertension, physical activity, education, and PIR.

Abbreviations: BMI=body mass index; CI=confidence intervals; OR=odds ratio; PIR=poverty income ratio; ref=reference group

Supplementary table 13. The effect of interaction between cardiometabolic risk clusters and race in the relationships between clusters and undiagnosed diabetes and pre-diabetes.

	Cluster 2	Cluster 1	Cluster 3	P value for trend	P for interaction
Undiagnosed diabetes as dependent variable					
		OR (95% CI)	OR (95% CI)		
Adjusted model* + race x cluster	Ref	2.81 (1.54, 5.14)	11.50 (6.17,21.45)	<0.001	0.596
Prediabetes as dependent variable					
Adjusted model + race x cluster	Ref	1.42 (1.23,1.64)	1.50 (1.26,1.78)	<0.001	**0.003**

*Adjusted logistic model included: cluster membership, age, sex, race, fasting glucose, BMI, waist circumference, history of hypertension, physical activity, education, and PIR.

Abbreviations: BMI=body mass index; CI=confidence intervals; OR=odds ratio; PIR=poverty income ratio; ref=reference group

Supplementary table 14. Odds ratios (OR) for undiagnosed diabetes for cardiometabolic risk clusters; stratified by race/ethnicity group.

	Cluster 2	**Cluster 1**	**Cluster 3**	
Model	**ref**	**OR (95% CI)**	**OR (95% CI)**	**P-value**
Undiagnosed diabetes				
White				
Unadjusted		5.26 (2.93,9.44)	27.35 (15.32,48.82)	<0.001
Adjusted		2.13 (1.14,4.01)	8.37 (4.19,16.72)	<0.001
Black				
Unadjusted		6.61 (4.29,10.17)	22.20 (14.25,34.57)	<0.001
Adjusted		4.04 (2.45,6.68)	11.20 (6.31,19.91)	<0.001
Hispanic				
Unadjusted		6.55 (4.44,9.66)	25.56 (17.56,37.22)	<0.001
Adjusted		4.40 (2.76,7.04)	21.48 (12.98,35.55)	<0.001
Other				
Unadjusted		5.74 (2.61,12.65)	33.72 (15.35,74.06)	<0.001
Adjusted		5.95 (2.44,14.48)	37.21 (14.05,98.54)	<0.001

Abbreviations: CI=confidence intervals; OR=odds ratio; ref=reference group

References

1. Assistant Secretary for Planning and Evaluation. U.S. federal poverty guidelines used to determine financial eligibility for certain federal programs. Washington, D.C.: U.S. Department of Health and Human Services, 2021.
2. Ahluwalia N, Dwyer J, Terry A, et al. Update on NHANES Dietary Data: Focus on Collection, Release, Analytical Considerations, and Uses to Inform Public Policy. *Adv Nutr* 2016;7(1):121-34. doi: 10.3945/an.115.009258 [published Online First: 20160115]
3. Lloyd-Jones DM, Hong Y, Labarthe D, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. *Circulation* 2010;121(4):586-613. doi: 10.1161/circulationaha.109.192703 [published Online First: 20100120]
4. Centers for Disease Control and Prevention (CDC). *Physical activity guidelines for Americans*, 2nd edition Physical Activity 2022.
5. Nooraeni R, Arsa MI, Kusumo Projo NW. Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering. *Procedia Computer Science* 2021;179:677-84. doi: https://doi.org/10.1016/j.procs.2021.01.055
6. CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES; 1997.
7. Yeo D, Truxillo C, Huber M, et al. Applied Clustering Techniques Course Notes. Cary, NC: SAS Institute Inc. 2019:448.
8. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 1998;2(3):283-304. doi: 10.1023/A:1009769707641
9. Ralambondrainy H. A conceptual version of the K-means algorithm. *Pattern Recognition Letters* 1995;16(11):1147-57. doi: https://doi.org/10.1016/0167-8655(95)00075-R
10. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record* 1996;25(2):103-14.
11. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Workshop on Research Issues on Data Mining and Knowledge Discovery*; 1997.