USO DEL ALGORITMO K-MEANS PARA CLASIFICAR CIUDADANOS CUBANOS MEDIANTE UN CUESTIONARIO DE ESTILOS DE VIDA

USE OF THE K-MEANS ALGORITHM TO CLASSIFY CUBAN CITIZENS THROUGH A LIFESTYLE QUESTIONNAIRE

Sheyla Torres Ricart[1] sheylatr@estudiantes.uci.cu,

Ing. David Alonso Díaz[2] davidad@uci.cu,

Dra.C. Natalia Martínez Sánchez[3] natalia@uci.cu,

Dr.C. Silvano Merced Len[4] silvano@uccfd.cu

Resumen

El término 'estilos de vida saludable' se ha transformado en un referente que caracteriza tanto a la sociedad en su conjunto como al individuo en su singularidad, ya que promueve un desarrollo equilibrado del ser humano, mejorando su calidad de vida. En la actualidad, las Tecnologías de la Información y la Comunicación se han erigido como uno de los principales motores que propulsan el conocimiento y la investigación, obligando al ser humano a avanzar en términos tecnológicos. En esta dirección, este trabajo pretende desarrollar grupos poblacionales de cubanos en cuanto a sus estilos de vida mediante el análisis de datos recopilados en un diagnóstico. Para esto, se presenta un estudio que utiliza una metodología basada en el análisis clúster para identificar similitudes entre los ciudadanos cubanos en cuanto a sus hábitos y estilos de vida. El objetivo específico de la investigación es realizar el agrupamiento de ciudadanos cubanos en función de sus hábitos y estilos de vida mediante el algoritmo de inteligencia artificial K-Means. Para la realización de la investigación se aplicaron los métodos: del codo para obtener un buen valor de k pudiendo así determinar la cantidad de grupos y Análisis de Componentes

[1] ORCID: https://orcid.org/0009-0008-2698-0453Estudiante de 4to año de Ingeniería en Ciencias Informáticas, Facultad de Tecnologías Educativas (FTE), Universidad de las Ciencias Informáticas. La Habana. Cuba.

[2] ORCID: https://orcid.org/0000-0001-6109-7975 Profesor del Centro de Estudios de Gestión de Proyectos y Toma de Decisiones, Facultad 3, Universidad de las Ciencias Informáticas. La Habana, Cuba.

[3] ORCID:https://orcid.org/0000-0002-2065-1746Profesora de la Universidad de las Ciencias Informáticas. Directora de formación del profesional. Ministerio de Educación Superior. Calle F, La Habana.

[4] ORCID: https://orcid.org/0000-0001-5131-0429Rector de la Universidad de las Ciencias de la Cultura Física y el Deporte ``Manuel Fajardo´´.

Principales para permitir la visualización del dataset. Además, se definieron las métricas de validación interna a aplicar. Como resultado de la investigación se obtiene que la aplicación de las métricas al resultado del algoritmo arrojó que se efectuó un buen proceso de agrupamiento, con *clusters* cohesionados y relativamente bien definidos.
Palabras Clave: algoritmo, agrupamiento, hábitos, estilos, vida.

Abstrac

The term ‘healthy lifestyles’ has become a reference that characterizes both society as a whole and the individual in their uniqueness, as it promotes a balanced development of the human being, improving their quality of life. Currently, Information and Communication Technologies have emerged as one of the main engines that drive knowledge and research, forcing the human being to advance in technological terms. In this direction, this work aims to develop population groups of Cubans in terms of their lifestyles through the analysis of data collected in a diagnosis. For this, a study is presented that uses a methodology based on cluster analysis to identify similarities among Cuban citizens in terms of their habits and lifestyles. The specific objective of the research is to group Cuban citizens based on their habits and lifestyles using the K-Means artificial intelligence algorithm. For the realization of the research, the elbow methods were applied to obtain a good value of k thus being able to determine the number of groups and Principal Component Analysis to allow the visualization of the dataset. In addition, the internal validation metrics to be applied were defined. As a result of the research, it is obtained that the application of the metrics to the result of the algorithm showed that a good clustering process was carried out, with cohesive clusters and relatively well defined.
Keywords: algorithm, clustering, habits, styles, life.

Introducción

El término ‘estilos de vida saludable’ se ha transformado en un referente que caracteriza tanto a la sociedad en su conjunto como al individuo en su singularidad, ya que promueve un desarrollo equilibrado del ser humano, mejorando su calidad de vida(Petrides et al., 2019). A nivel global, se ha vinculado tradicionalmente con aspectos como la nutrición, el ejercicio físico, el sueño y el consumo de sustancias, y en menor medida, con factores relacionados con la salud mental como la conexión social(Zaman et al., 2019).

Parece ser que la mayoría de las iniciativas para alcanzar este referente se basan en la elaboración de listas de comportamientos individuales y colectivos, con actividades concretas que se deben realizar o evitar, con el objetivo de aumentar la longevidad. El desarrollo equilibrado del ser humano implica no solo la satisfacción de necesidades básicas, sino también un crecimiento personal, lo cual repercute en la calidad global de nuestras vidas (Petrides et al., 2019).

En la actualidad, las Tecnologías de la Información y la Comunicación (TIC) se han erigido como uno de los principales motores que propulsan el conocimiento y la investigación, obligando al ser humano a avanzar en términos tecnológicos. Por esta razón, se destaca la relevancia de su uso, así como los pros y contras que conllevan para fomentar un estilo de vida saludable. Las TIC han revolucionado nuestra forma de vida, incluyendo cómo accedemos al conocimiento y llevamos a cabo nuestras actividades cotidianas. Han desplegado un abanico de oportunidades para la educación y el aprendizaje, la investigación y el desarrollo de nuevas tecnologías (Martínez et al., 2023).

En relación con la promoción de un estilo de vida saludable, las Tecnologías de la Información y la Comunicación (TIC) pueden ser un recurso valioso para informar y educar acerca de la relevancia de la actividad física y una alimentación balanceada(Diego-Cordero et al., 2017). Existen múltiples aplicaciones y herramientas digitales que pueden asistir a las personas en el seguimiento de su actividad física, como el conteo de pasos, la distancia recorrida, las calorías quemadas, el ritmo cardíaco, la dieta, entre otros aspectos vinculados con la salud, proporcionando retroalimentación en tiempo real. En este contexto, hay una gran variedad de aplicaciones y dispositivos electrónicos disponibles que pueden facilitar a las personas el registro de su actividad física y compartirle recomendaciones al respecto.

Para que las recomendaciones entregadas a las personas sean acordes al estilo de vida que estas llevan se hace que las mismas sean clasificadas en grupos que reflejen sus estilos de vida. En esta dirección, este trabajo pretende desarrollar grupos poblacionales de cubanos en cuanto a sus estilos de vida mediante el análisis de datos recopilados en un diagnóstico. Para esto, se presenta un estudio que utiliza una metodología basada en el análisis clúster para identificar similitudes entre los ciudadanos cubanos en cuanto a sus hábitos y estilos de vida. El estudio

consideró datos recogidos por un diagnóstico realizado a 528 ciudadanos cubanos de todo el país y grupos etarios, encuestados en el año 2023.

El objetivo específico de la investigación es realizar el agrupamiento de ciudadanos cubanos en función de sus hábitos y estilos de vida mediante el algoritmo de inteligencia artificial K-Means.

Metodología

1. Algoritmo.

El algoritmo K-Means es un algoritmo particional, es decir, divide los objetos en un número de clústeres pre especificado, sin atender a una estructura jerárquica (Wang et al., 2020), puede aplicarse para problemas de "agrupación por similitud" y puede ayudar al investigador a una comprensión cualitativa y cuantitativa de grandes cantidades de datos N-dimensionales (MacQueen, 1967).

El algoritmo K-Means inicia con una solución preliminar, la cual se obtiene de forma aleatoria a partir del conjunto de datos. El objetivo es mejorar esta solución de manera iterativa hasta alcanzar un mínimo local. En primer lugar, se dividen los elementos del conjunto en un grupo de M clústeres, asociando cada objeto con el centroide del clúster más cercano de la iteración previa. Luego, se recalculan los centroides de cada clúster, considerando la nueva partición. La calidad de la nueva solución debe ser igual o superior a la anterior. El algoritmo prosigue mientras exista mejora (MacQueen, 1967). Este algoritmo puede codificarse siguiendo el esquema a continuación:

```
procedure K-Means (X, P, C): Solución
    begin
        repeat
            for xi in X
                Pi := EncontrarCentroideMásCercano(xi, C)
            for Ci in C
                Ci := CalcularCentroide(X,P,i)
        until NoHayMejora
    end
```

Figura1: Pseudo código de K-Means(Lunar, s. f.).

La principal ventaja del algoritmo K-Means es que es un método sencillo y rápido. El algoritmo K-Means se adapta fácilmente con heurísticas ya que es fácil de implementar incluso para grandes conjuntos de datos. Por lo que ha sido ampliamente usado en muchas áreas como segmentación de mercados, visión por computadoras, geoestadística, astronomía y minería de datos en agricultura.

También se usa como pre-procesamiento para otros algoritmos, por ejemplo, para buscar una configuración inicial.

2. Medidas de Similitud.

Las medidas de similitud establecen la forma en que se determina la proximidad que hay entre los datos. Como medida se similitud en la aplicación del algoritmo, en este trabajo se empleará la distancia euclídea. La misma se entiende la longitud de la línea recta que conecta los dos puntos en el espacio euclidiano, esto se deduce por el teorema de Pitágoras(Chambers, 1999). Es una medida directa de la distancia más corta entre los puntos y cumple con las propiedades de una métrica, como la no negatividad, la simetría y la desigualdad triangular.

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

Figura 2: Formula N dimensional de la distancia euclídea.

3. Selección de la K.

Un aspecto importante en el funcionamiento del algoritmo K-Means es la selección del parámetro K, el cual tiene como función definir la cantidad de clusters a crear. Entre las principales técnicas para una correcta selección de la K se encuentran: V-measure / Normalised Mutual Information score(Ahmadinejad et al., 2023), validación cruzada de k-fold(Ziggah et al., 2019)y el método del codo, este último será el utilizado en la presente investigación.

El método del codo se emplea para determinar el número óptimo de agrupaciones. Este proceso implica calcular la suma de las distancias al cuadrado (SSE) entre los puntos y el centroide de cada clúster. Este cálculo se realiza de manera iterativa para diferentes valores de K (Alvarado-Ruiz et al., 2023).

Posteriormente, se grafican los SSE obtenidos en función del número de agrupaciones. Se busca el punto de inflexión donde la curva converge, lo que indica que agregar más clústeres no mejoraría significativamente la calidad de la segmentación (Alvarado-Ruiz et al., 2023).

Este punto de inflexión se conoce como “codo”, de ahí el nombre del método. Finalmente, se ejecuta nuevamente el algoritmo K-means con el número de clústeres y se verifica la calidad de la segmentación. Es importante mencionar que el método del codo puede no funcionar en todos los casos, ya que es un método

heurístico. Por lo tanto, es necesario revisar la calidad de la segmentación (Alvarado-Ruiz et al., 2023).

4. Reducción de la dimensionalidad.

La reducción de la dimensionalidad es una técnica que transforma datos de alta dimensionalidad a un espacio de menor dimensionalidad, disminuyendo así el número de variables o características en un conjunto de datos, pero conservando la información esencial. Esta técnica se utiliza principalmente para comprimir datos, reducir el ruido y como paso previo a la clasificación. Además, permite visualizar conjuntos de datos de alta dimensionalidad que, debido a su gran cantidad de atributos, serían imposibles de representar gráficamente sin esta técnica, como lo es en este caso (Dorado Valín, 2023).

Una de los algoritmos más utilizados en la reducción de la dimensionalidad y que utilizaremos en esta investigación es el método:Análisis de Componentes Principales (PCA). Esta es una técnica lineal y no supervisada de reducción de la dimensionalidad. Su objetivo es identificar las direcciones principales de variabilidad en los datos y representarlos en un espacio de menor dimensionalidad. De esta manera, logra condensar casi toda la información en unos pocos componentes. Transforma un conjunto de variables correlacionadas en un conjunto de variables ortogonales, conocidas como componentes principales. El primer componente principal es el que explica la mayor variabilidad en el conjunto de datos, y así sucesivamente, hasta llegar al número de componentes deseados (Dorado Valín, 2023).

5. Métrica de validación interna.

Las Métricas de Validación Interna son medidas utilizadas para evaluar la calidad de un agrupamiento (clustering) basándose únicamente en la información de los datos. Estas métricas no requieren información externa o adicional al resultado del algoritmo de agrupamiento(Esteban et al., 2021).

En este artículo utilizaremos tres de las métricas más utilizadas en el clustering:

ÍndiceSilhouette: Este es un coeficiente que mide cohesión del grupo a(x), calculando la distancia promedio del centroide (x)a todos los demás puntos en elmismo clúster y la separación de los grupos b(x), calculando la distancia promedio del centroide (x)a todos los demás puntos en elclúster más cercano(Esteban et al., 2021). Este coeficiente está acotado entre los valores -1 y 1, siendo -1 una mala agrupación, 0 una agrupación indiferente y 1 una buena agrupación.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Figura 3: Fórmula del coeficiente de silhouette para el punto x.

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

Figura 4: Fórmula del coeficiente de silhouette para todo el agrupamiento, siendo N el número de grupos formados.

Índice Davies-Bouldin: es una métrica introducida por David L. Davies y Donald W. Bouldin en 1979 para evaluar algoritmos de agrupamiento. Este es un esquema de evaluación interna, donde la validación de qué tan bien se ha realizado la agrupación se realiza utilizando cantidades y características inherentes al conjunto de datos.El índice DB trata a cada conglomerado individualmente y busca medir qué tan similar es al conglomerado más cercano a él, en este mientras más pequeño es el valor del mismo más compacto y separados están los grupos (Esteban et al., 2021).

$$DB = \frac{1}{k}\sum_{i=1, i\neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Figura 5: Fórmula para calcular Índice Davies-Bouldin, Donde k es el número de clústeres, σ_i es la distancia promedio entre cada punto en el clúster i y elcentroide del clúster, σ_j es la distancia promedio entre cada punto del clúster j y el centroide delclúster, y $d(c_i, c_j)$ es la distancia entre los centroides.

Índice Calinski-Harabasz: también conocido como el Criterio de la Relación de Varianza, es una métrica introducida por T. Calinski y J. Harabasz en 1974 para evaluar algoritmos de agrupamiento. Este es un esquema de evaluación interna, donde la validación de qué tan bien se ha realizado la agrupación se realiza utilizando cantidades y características inherentes al conjunto de datos. El índice CH es una medida de cuán similar es un objeto a su propio clúster (cohesión) en comparación con otros grupos. Aquí, la cohesión se estima en función de las distancias desde los puntos de datos en un clúster a su centroide del clúster y la separación se basa en la distancia de los centroides del clúster desde el centroide global.Para obtener el índice en primer lugar se debe obtener la suma de las

distancias al cuadrado entre clústeres (BSS, Between-ClusterSumofSquares) que se define como:

$$BSS = \sum_{k=1} K n_k |C_k - C|^2,$$

Figura 6: Fórmula de BSS, donde donde nk, es el número de observaciones en el clústerk,Ckes el centroide del clúster k, C es el centroide del conjunto dedatos y K es el número de clústeres.

Por otro lado, la suma de las distancias al cuadrado dentro de cada clúster (WSS, Within-ClusterSumofSquares) se puede obtener mediante la expresión:

$$WSS = \sum_{i=1}^{n_k} |X_{ik} - C_k|^2$$

Figura 7: Fórmula de WSS, donde nkes el número de observacionesdel clúster k y X_{ik}es la observación i del clúster k.

De este modo el índice Calinski-Harabasz se puede definir como:

$$CH = \frac{\frac{BSS}{K-1}}{\frac{WSS}{N-K}}$$

Figura 8: Fórmula del índice Calinski-Harabasz, donde N es el número total de observaciones.

6. Datos.

Los datos forman parte de un diagnóstico realizado por los autores, validado con software estadísticos y criterio de expertos de la Universidad de las Ciencias de la Cultura Física y el Deporte ``Manuel Fajardo´´ (UCCFD), en el año 2023. El diagnóstico recoge datos de ciudadanos cubanos de todas las provincias del país y todos los grupos etarios. Se confeccionó partiendo del cuestionario FANTASTICO desarrollado por la universidad canadiense de McMaster(Betancurth Loaiza et al., 2015). El dataset que con el que se trabajó está conformado por 528 filas y 30 columnas.

Resultados y Discusión

Primeramente, se procedió a realizar una reducción de la dimensionalidad al dataset para que pudiese ser graficado, al cual se le aplico el método PCA para reducir el dataset a 3 dimensiones.

Tabla 1: La tabla muestra una sección del dataset luego de la aplicación del método PCA.

	Componente_1	Componente_2	Componente_3
0	-2.693950	-0.307269	-0.402018
1	26.321687	-0.196449	-0.696198
2	0.308041	-0.502812	1.004303
3	5.308881	-0.874072	0.903206
4	1.304210	1.479802	0.045528
...	...	...	...
523	-4.694813	-0.110118	1.097810
524	-4.698110	0.367654	1.122991
525	-1.689702	0.526301	0.076379
526	-6.686951	-0.037436	-0.299987
527	-4.691671	-0.443299	0.325121

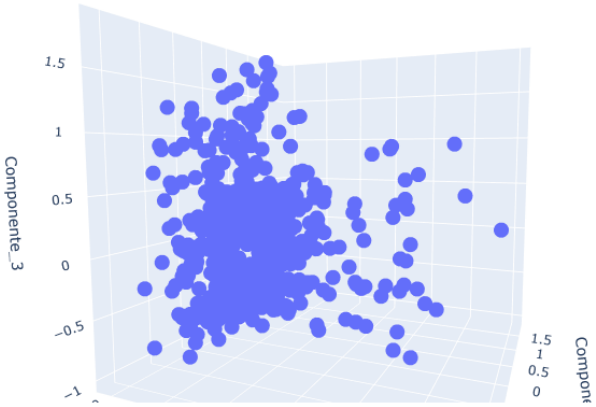

Figura 9: La figura muestra los datos espaciados en tres dimensiones.

Seguidamente se procede al cálculo del método del codo para definir una k eficiente para la ejecución del algoritmo.

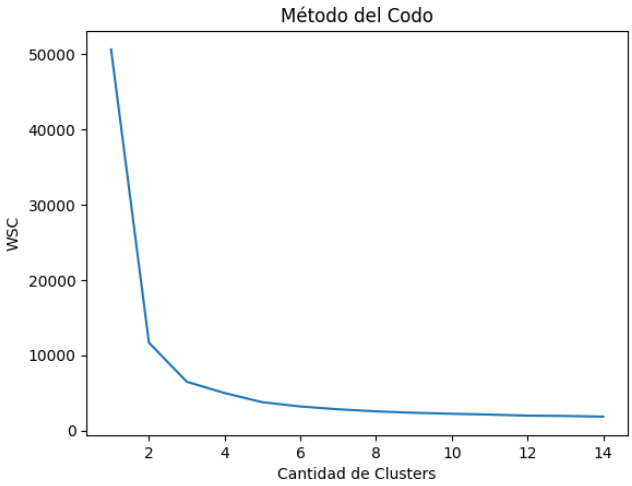

Figura 10: Gráfica resultante de la aplicación del método del codo al dataset.

Luego de obtenidos los resultados de la aplicación del método del codo se decidió aplicar el algoritmo con una k= 5, con lo cual se forman 5 grupos. Esta decisión se toma atendiendo que el quinto clúster la función realiza el punto de inflexión, luego de esto comienza de converger definitivamente.

Posteriormente se aplicó el algoritmo propuesto y se agruparon las instancias quedando distribuidas como muestra la figura 11.

col_0	Count
row_0	
0	291
1	26
2	41
3	19
4	151

Figura 11: Cantidad deinstancias por grupo conformado.

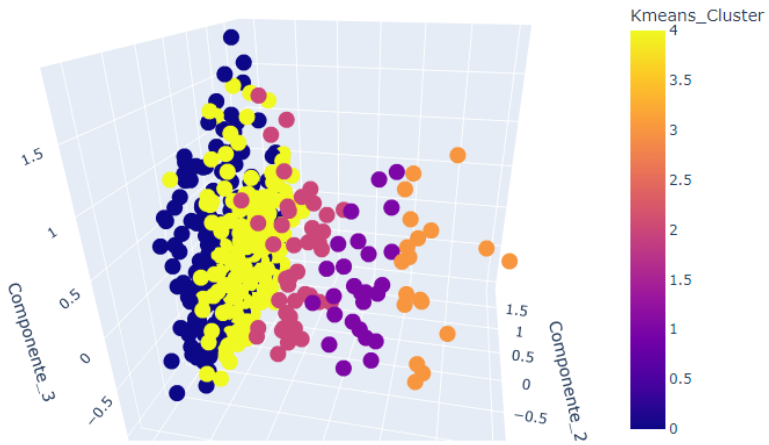

Figura 12: Datos espaciados agrupados.

Finalmente fueron aplicadas las métricas de validación interna para examinar los resultados del agrupamiento obteniéndose un índice Silhouette (SC=0.50), considerado de efectivo. Un índice Davies-Bouldin (DB=0.68), denotando un buen desempeño y finalmente un índice Calinski-Harabasz (CH=1685.54), considerado de una buena efectividad.

Conclusiones

La investigación se centra la conformación de grupos de ciudadanos cubanos en función de sus hábitos y estilos de vida a través de la aplicación del algoritmo K-Means, considerando para su estudio los datos recogidos por un diagnóstico realizado a 528 ciudadanos.

Luego de la aplicación del algoritmo los resultados arrojados fueron sometidos a evaluación a través de métricas de validación interna las cuales arrojaron que se efectuó un buen proceso de agrupamiento, con *clusters* cohesionados y relativamente bien definidos.

Como trabajo futuro se proyecta la aplicación del algoritmo Beta 0, perteneciente al Reconocimiento Lógico Combinatorio de Patrones (RLCP), el cual permitirá extraer los principales conceptos de cada grupo, lo cual disminuirá la complejidad temporal de las comparaciones entre los nuevos registros y los grupos.

Referencias Bibliográficas

Ahmadinejad, N., Chung, Y., & Liu, L. (2023). J-score: A robust measure of clustering accuracy. PeerJ Computer Science, 9, e1545. https://doi.org/10.7717/peerj-cs.1545

Alvarado-Ruiz, D. A., Ordaz-Hernández, K., Lara-Cadena, G. L., Díaz-Jiménez, M. de L. V., & Castelán, M. (2023). Caracterizacion del crecimiento de colonias bacterianas utilizando segmentación de imágenes con K-means. Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI, 11(Especial2), Article Especial2. https://doi.org/10.29057/icbi.v11iEspecial2.10711

Betancurth Loaiza, D. P., Vélez Álvarez, C., & Jurado Vargas, L. (2015). Validación de contenido y adaptación del cuestionario Fantastico por técnica Delphi. Revista Salud Uninorte, 31(2), 214-227.

Chambers, P. (1999). Teaching Pythagoras' Theorem. Mathematics in School, 28(4), 22-24.

Diego-Cordero, R. de, Fernández-García, E., & Romero, B. B. (2017). Uso de las TIC para fomentar estilos de vida saludables en niños/as y adolescentes: El caso del sobrepeso = Use of ICT topromotehealthylifestyles in children and adolescents: the case ofoverweight. REVISTA ESPAÑOLA DE COMUNICACIÓN EN SALUD, 79-91. https://doi.org/10.20318/recs.2017.3607

Dorado Valín, A. (2023). Análisis del impacto de las medidas de distancia en técnicas de reducción de la dimensionalidad. https://ruc.udc.es/dspace/handle/2183/33866

Esteban, A., Zafra, A., & Ventura, S. (2021). Estudio comparativo de medidas de disimilitud para Clustering Multi-Instancia.

Lunar, R. (s. f.). Comparación de Algoritmos Metaheurísticos para el Problema de Clustering. Recuperado 31 de octubre de 2023, de

https://www.academia.edu/5000545/Comparaci%C3%B3n_de_Algoritmos_Metaheur%C3%ADsticos_para_el_Problema_de_Clustering

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics: Vol. 5.1 (pp. 281-298). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Martínez, H. A. M., González, J. P. R., & García, M. I. B. (2023). Uso de las TIC y su influencia en estilos de vidas saludables en los estudiantes. Polo del Conocimiento, 8(5), Article 5. https://doi.org/10.23857/pc.v8i5.5551

Petrides, J., Collins, P., Kowalski, A., Sepede, J., & Vermeulen, M. (2019). Lifestyle Changes for Disease Prevention. Primary Care, 46(1), 1-12. https://doi.org/10.1016/j.pop.2018.10.003

Wang, Z., Zhou, Y., & Li, G. (2020). Anomaly Detection by Using Streaming K-Means and Batch K-Means. 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), 11-17. https://doi.org/10.1109/ICBDA49040.2020.9101212

Zaman, R., Hankir, A., &Jemni, M. (2019). Lifestyle Factors and Mental Health. PsychiatriaDanubina, 31(Suppl 3), 217-220.

Ziggah, Y. Y., Youjian, H., Tierra, A. R., Laari, P. B., Ziggah, Y. Y., Youjian, H., Tierra, A. R., & Laari, P. B. (2019). Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana. EarthSciencesResearchJournal, 23(1), 67-77. https://doi.org/10.15446/esrj.v23n1.63860