COMPARACIÓN DE K-MEANS, DBSCAN y HDBSCAN PARA CLASIFICAR PERSONAS USANDO CUESTIONARIO DE ESTILOS DE VIDA

COMPARISON OF K-MEANS, DBSCAN, AND HDBSCAN FOR CLASSIFYING PEOPLE USING A LIFESTYLE QUESTIONNAIRE

Ing. David Alonso Díaz [1]davidad@uci.cu

Sheyla Torres Ricart[2] sheylatr@estudiantes.uci.cu

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

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

Resumen

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 comparar los resultados del agrupamiento de ciudadanos cubanos en función de sus hábitos y estilos de vida mediante laaplicación de los algoritmos de inteligencia artificial K-Means, DBSCAN y HDBSCAN.Para la realización de la investigación se

[1] ORCID: https://orcid.org/0000-0001-6109-7975Profesor del Centro de Estudios de Gestión de Proyectos y Toma de Decisiones, Facultad 3, Universidad de las Ciencias Informáticas. La Habana, Cuba. CP: 19370.

[2] 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.

[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. 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´´. Avenida Santa Catalina.

aplicó elmétodo: Análisis de Componentes 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 los resultados arrojaron que el algoritmo K-Means realiza para este conjunto de datos una mejor agrupación en cuanto a cohesión, separación y la similitud de los grupos.
Palabras Clave: algoritmos, agrupamiento, densidad, hábitos, vida.

Abstract

Currently, Information and Communication Technologies have emerged as one of the main engines that drive knowledge and research, forcing humans 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 compare the results of the grouping of Cuban citizens based on their habits and lifestyles by applying the artificial intelligence algorithms K-Means, DBSCAN, and HDBSCAN. For the realization of the research, the method: Principal Component Analysis was applied 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 yielded results that the K-Means algorithm performs for this data set a better grouping in terms of cohesion, separation, and similarity of the groups.
Keywords: algorithms, clustering, density, habits, life.

Introducción

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 comparar los resultados de los agrupamientos producidos por la aplicación de los algoritmos de inteligencia artificial K-Means, DBSCAN y HDBSCAN a un dataset que mide estilos de vida de ciudadanos cubanos.

Metodología

1. Algoritmos.

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 (Z. 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.).

El algoritmo DBSCANes el primer algoritmo basado en densidad, se definen los conceptos de punto central (puntos que tienen en su vecindad una cantidad de puntos mayor o igual que un umbral especificado), borde y ruido.

El algoritmo comienza seleccionando un punto p arbitrario, si p es un punto central, se comienza a construir un grupo y se ubican en su grupo todos los objetos denso-alcanzables desde p. Si p no es un punto central se visita otro objeto del conjunto de datos. El proceso continúa hasta que todos los objetos han sidoprocesados. Los puntos que quedan fuera de los grupos formados se llaman puntos ruido, los puntos que no son ni ruido ni centrales se llaman puntosborde.

```
for each o ∈ D do
    if o no está clasificado aún then
        if o es un objeto núcleo then
            Recoge todos los objetos alcanzables por densidad desde o
            y asígnalos a un nuevo grupo.
        else
            asigna o a RUIDO
```

Figure 2: Pseudo código de DBSCAN.

El algoritmo HDBSCANes una versión optimizada de DBSCAN, desarrollado por los mismos autores, que hereda los beneficios de los algoritmos jerárquicos y de densidad. Extiende DBSCANconvirtiéndolo en un algoritmo jerárquico extrayendo luego una estructuraplana de clusters(Ramirez Gomez, 2023).

El algoritmo comienza calculando la densidad de cada punto de datos basándose en la distancia a sus vecinos más cercanos.Esta densidad se usa para construir un árbol jerárquico de agrupamientos, el Árbol de Estabilidad. Luego Se aplica un algoritmo de corte en este árbol para obtener un agrupamiento plano, conservando

los agrupamientos más estables como clusters y considerando los menos estables como ruido(Ramirez Gomez, 2023).

```
  Input: Location data: LD, Parameter: Eps and Minpts,
S-Tree: Height
  Output: LD with cluster lable and Spatial_Tree was
built
1. DBSCAN_ OBJECT Root=Joint(LD,Eps,Minpts); // root
   node of Tree
2. ENQUEUE(Q, Root) ;         // push DBSCAN object into
   Queue
3. front:=0, last:=0, level=0;
4. while(Queue<>empty and front<=last) DO
5.   DBSCAN_ OBJECT node= DEQueue(Q); // Pull data from
   Queue
6.    front++;              //
7.    Data_OBJECT Childern    =DBSCAN.getCluster(node);
   //Call DBSCAN
8.    if(level > Height)
9.       break;
10.
11.    For i FROM 1 TO Childern.size  DO
12.      Data child=Childern.get(i);
13.      DBSCAN_ OBJECT Root=Joint(child,Eps,Minpts);
14.      ENQUEUE(Q,DBSCAN_ OBJECT) ;
15.    end For
16.
17.    if(front>last) // members in one level have been
   searched
18.       last= Q.size()+front-1;
19.       level ++;
20.    end if
21. end while
```

Figure 3: Pseudo código de HDBSCAN(Tanget al., 2009).

2. 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).

3. Métrica de validación interna.

Las Métricas de Validación Interna son indicadores que evalúan la calidad de un agrupamiento solo con la información de los datos. No necesitan información adicional al resultado del algoritmo de agrupamiento(Esteban et al., 2021).

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

Índice Silhouette: Este coeficiente evalúa la cohesión del grupo a(x) mediante el cálculo de la distancia promedio desde el centroide (x) a todos los otros puntos en el mismo clúster. También mide la separación de los grupos b(x) calculando la distancia promedio desde el centroide (x) a todos los puntos en el clúster 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 4: Fórmula del coeficiente de silhouette para el punto x.

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

Figura 5: Fórmula del coeficiente de silhouette para todo el agrupamiento, siendo N el número de grupos formados.

Índice Davies-Bouldin: es una métrica propuesta por David L. Davies y Donald W. Bouldin en 1979 para evaluar algoritmos de agrupamiento. Trata cada conglomerado de forma individual, mide su similitud con el conglomerado más cercano. Un valor más pequeño indica grupos más compactos y separados(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 6: 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 el centroide del clúster, σ_j es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d(c_i, c_j)$ es la distancia entre los centroides.

Índice Calinski-Harabasz:el Criterio de la Relación de Varianza, introducido por T. Calinski y J. Harabasz en 1974, evalúa algoritmos de agrupamiento. Mide la similitud de un objeto con su clúster (cohesión) y la separación entre clústeres. La cohesión se basa en las distancias al centroide del clúster, y la separación en la distancia de los centroides al centroide global(X. Wang & Xu, 2019). Para obtener el índice, se calcula la suma de las distancias al cuadrado entre clústeres (BSS)que se define como:

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

Figura 7: Fórmula de BSS, donde nk, es el número de observaciones en el clúster k, Ck es el centroide del clúster k, C es el centroide del conjunto de datos y K es el número de clústeres.

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

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

Figura 8: Fórmula de WSS, donde nk es el número de observaciones del 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 9: Fórmula del índice Calinski-Harabasz, donde N es el número total de observaciones.

4. 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

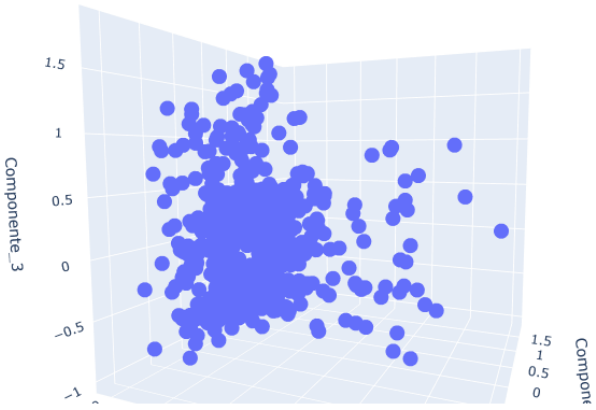

Figura 10: La figura muestra los datos espaciados en tres dimensiones.

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

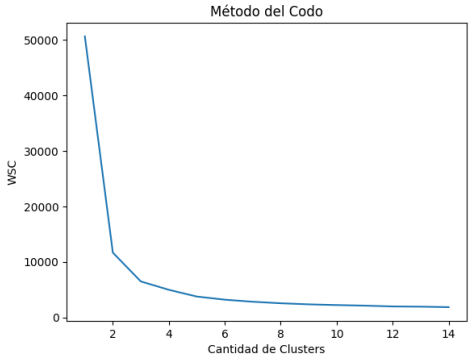

Figura 11: Gráfica resultante de la aplicación del método del codo al dataset.

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.

Además, se definieron los parámetros necesarios para los otros dos algoritmos en el caso de DBSCAN fueron definidos los parámetros épsilon (ε=3) y la cantidad de puntos iniciales para iniciar el agrupamiento (min_samples=3). Y en el caso del algoritmo HDBSCAN fueron definidos los parámetros ejemplares mínimos para armar un clúster (min_cluster_size=5) y la métrica que en este caso fue utilizada la distancia de cosenos (metric='cosein').

Posteriormente se aplicaronlos algoritmos propuestos y se agruparon las instancias quedando distribuidas como muestra la tabla2.

Tabla 2: Distribución de los elementos del dataset por grupos.

Grupos	K-Means	DBSCAN	HDBSCAN
-1	-	10	23
0	291	151	290
1	26	49	15
2	41	288	151
3	19	4	14
4	151	3	35
5	-	15	-
6	-	3	-
7	-	5	-

Como puede observarse los algoritmos DBSCAN y HDBSCAN no fueron capaces de agrupar todos los elementos, faltaron 10 y 23 respectivamente los cuales fueron catalogados como ruido. Por otra parte, el DBSCAN creo 8 grupos, tres más que los otros dos algoritmos, mostrando una menor capacidad de optimizar la culturización en el dataset.

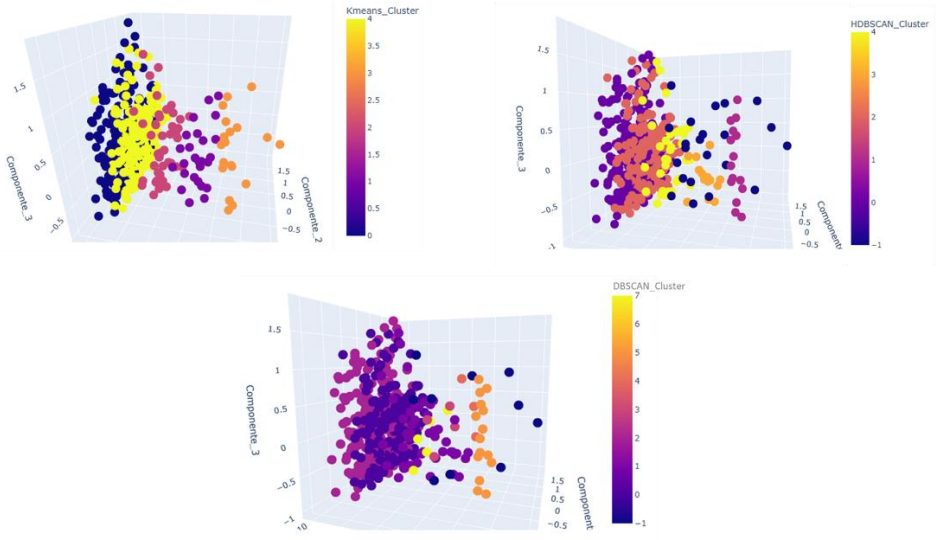

Figura 12: Datos espaciados agrupados según los algoritmos K-Mean, HDBSCAN (de izquierda a derecha) y DBSCAN (debajo).

Finalmente fueron aplicadas las métricas de validación interna para examinar los resultados delos agrupamientos como muestra la tabla 3.

Métricas	SC	DB	BBS
K-Means	0.50	0.68	1685.54
DBSCAN	0.47	1.34	436.27
HDBSCAN	0.51	1.65	670.5

Como puede apreciarse en relación al índice Silhouette los tres algoritmos cohesionan relativamente bien sus grupos, siendo el HDBSCAN el de mejor resultado, aunque el K-Means prácticamente tiene el mismo resultado. Con respeto al índice Davies-Bouldin, el de mejor desempeño es el algoritmo K-Means que mejora a los otros dos considerablemente,y finalmente en lo referente al índice Calinski-Harabasz, el algoritmo K-Means vuelve a mejorar al resto de algoritmos casi triplicando en resultado al HDBSCAN y cuadruplicando al DBSCAN, siendo este último el de peor desempeño.

Conclusiones

La investigación se en la comparación del rendimiento de los algoritmos K-Means, DBSCAN y HDBSCAN en cuanto a la conformación de grupos de ciudadanos cubanos en función de sus hábitos y estilos de vida, considerando para su estudio los datos recogidos por un diagnóstico realizado a 528 ciudadanos.

Luego de la aplicación delos algoritmos los resultados arrojados fueron sometidos a evaluación a través de métricas de validación interna las cuales arrojaron que el algoritmo K-Means realiza para este conjunto de datos una mejor agrupación en cuanto a cohesión, separación y la similitud de los grupos.

Como trabajo futuro se proyecta la aplicación de inteligencia artificial explicativa para ver su eficiencia en la extracción de los principales conceptos de cada grupo.

Referencias Bibliográficas

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.

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 to promote healthy lifestyles in children and adolescents: the case of overweight. *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

Ramirez Gomez, C. (2023). *Clasificación y detección de tópicos en Twitter: Caso de estudio elecciones presidenciales Colombia 2022.* http://repository.javeriana.edu.co/handle/10554/65484

Tang, M., Zhou, Y., Cui, P., Wang, W., Li, J., Zhang, H., Hou, Y., & Yan, B. (2009). *Discovery of Migration Habitats and Routes of Wild Bird Species by Clustering and Association Analysis* (Vol. 5678, p. 301). https://doi.org/10.1007/978-3-642-03348-3_29

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5), 052024. https://doi.org/10.1088/1757-899X/569/5/052024

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