Aprendizaje no supervisado

Claudia Lissette Gutiérrez Díaz[1]*

Palabras Clave
Número de Clusters, Elbow Method, Silhouette Coefficient, Davies-Bouldin Index, Nearest Neighbords K-Means, DBSCAN

Fecha: 17 de octubre de 2023

[1] *Licenciada en Ciencias Computacionales, FCFM, UANL*

Índice

1. Introducción

En este artículo, exploraremos dos metodologías de agrupamiento (clustering) ampliamente utilizadas: K-Means y DBSCAN. Nuestra investigación se enfocará en determinar los parámetros óptimos para estas técnicas mediante métodos de selección como el Método del Codo (Elbow Method) y el enfoque de Vecinos Más Cercanos (Nearest Neighbors).

Además, evaluaremos la efectividad de ambas metodologías utilizando métricas clave, como la Silueta (Silhouette) y el Índice Davies-Bouldin. Estas métricas nos proporcionarán una comprensión más profunda de la calidad de los grupos generados por K-Means y DBSCAN.

A medida que avanzamos en este artículo, exploraremos cómo estas metodologías de agrupamiento pueden aplicarse en diversas situaciones y cómo seleccionar los mejores parámetros y métricas para lograr resultados precisos y significativos.

2. Descripción del Conjunto de Datos de la Planta de Energía Solar

El conjunto de datos utilizado en este estudio se recopiló de una planta de energía solar ubicada en Villa de Arista, en el estado de San Luis Potosí, México. El objetivo de este conjunto de datos es analizar y predecir la generación de energía en kilovatios-hora (kWh) a partir de diversas condiciones ambientales. Se realizaron procesos de filtrado, diferenciación y normalización para obtener un conjunto de datos preparado de la siguiente manera:

- **Generación de Energía (kWh):** Esta es la variable dependiente en nuestro conjunto de datos y representa la cantidad de energía generada por la planta de energía solar en kilovatios-hora en un intervalos de una hora.
- **Condiciones Ambientales:** Las siguientes variables independientes representan las condiciones ambientales que podrían influir en la generación de energía solar:

  1. **Temperatura (°C):** La temperatura ambiente en grados Celsius.
  2. **Probabilidad de Lluvia (%):** La probabilidad de precipitación en forma de lluvia en porcentaje.

3. **Dirección del Viento (grados):** La dirección del viento en grados.
4. **Índice UV:** El índice de radiación ultravioleta que mide la intensidad de la radiación solar.
5. **Condición del Cielo:** Una variable que describe las condiciones del cielo, que podría ser una categoría como "despejado", "poco nuboso", "nublado", "medio nublado", "cielo nublado", "cielo cubierto".
6. **Dirección Categórica del Viento:** La dirección del viento categorizada en términos de puntos cardinales (por ejemplo, "norte", "sur", "este", "oeste").
7. **Velocidad de la Ráfaga (km/h):** La velocidad del viento en kilómetros por hora.
8. **Temperatura del Punto de Rocío (°C):** La temperatura a la cual el aire se satura y la humedad relativa es del 100 %.

- **Preparación del Conjunto de Datos:** Las variables independientes fueron sometidas a un proceso de filtrado mediante el método de selección de características exhaustivo para identificar las más relevantes para la predicción de la generación de energía solar. Además, se eliminó la estacionalidad mediante el método de diferenciación. Los datos independientes se normalizaron para garantizar la consistencia en las escalas.

Temp.	Prob.Lluvia	Dir.Viento	IndiceUV	CondCielo	CatDir.Viento	Vel.Ráfaga	DPT
0.447368	0.527950	0.494949	0.444444	0.750000	0.375000	0.494208	0.507246
0.473684	0.552795	0.494949	0.444444	0.750000	0.375000	0.494208	0.507246
0.473684	0.534161	0.494949	0.444444	0.250000	0.375000	0.494208	0.507246
0.447368	0.503106	0.494949	0.444444	0.250000	0.375000	0.494208	0.492754

Cuadro 1. Muestra de las condiciones ambientales de Villa de Arista ya normalizadas

Este conjunto de datos preparado se utiliza para realizar análisis estadísticos y modelado predictivo, con el propósito de entender cómo las condiciones ambientales afectan la generación en la planta de energía solar.

3. Metodología

3.1 Métricas para determinar el número óptimo de Clusters

En el ámbito del análisis de clusters, la elección del número óptimo es una etapa crucial que influye en la efectividad y la interpretación de los resultados. Para abordar este desafío, se han desarrollado diversas métricas que permiten evaluar y cuantificar la calidad de las agrupaciones en función de la cantidad de clusters utilizados. Exploraremos los siguientes dos engoques para determinar el número óptimo de clusters o el parámetro óptimo para contabilizarlos: el Método del Codo (Elbow Method) y el enfoque de Vecinos Más Cercanos (Nearest Neighbors). Estas técnicas ofrecen perspectivas valiosas y proporcionan una guía fundamental para los profesionales y científicos de datos que buscan segmentar sus datos de manera efectiva y con fundamentos sólidos.

3.1.1 Elbow Method

Como se menciona en «*Review on determining number of Cluster in K-Means Clustering*»**(Kodinariya, Makwana et al., 2013)**, el método del codo es uno de los más viejos utilizado para determinar el número de clusters en un conjunto de datos, es una técniva visual, como se puede determinar en la Fig. 1. La idea es que se inicie con una cantidad de 2 clusters y luego se vayan incrementando en uno. Conforme se avance, habrá un cluster cuyo costo caiga dramáticamente y después de eso se estabiliza, éste sería el número K del Cluster que se necesita.

El proceso implica los siguientes pasos:

1. **Realizar el clustering:** Comienza por aplicar el algoritmo de clustering, como K-Means (*3.3.1*), a tus datos con un rango de valores para el número de clusters, generalmente desde 1 hasta un número máximo que se considera razonable para tu conjunto de datos.

2. **Calcular la métrica:** Después de realizar el clustering para diferentes valores de k, calcula una métrica que evalúe la calidad de las agrupaciones. La métrica más comúnmente utilizada es la suma de las distancias cuadradas intra-cluster 1. En el caso de K-Means, esta métrica se conoce como la inercia o la suma de los cuadrados intra-cluster (SSW) dada por la siguiente fórmula:

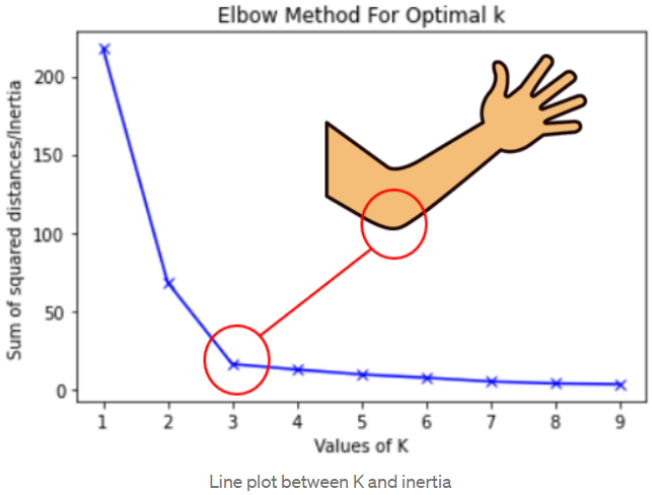

Figura 1. Ejemplo de visualización del método del codo

$$\text{Inercia} = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

Donde:

- K es el número de clusters.
- C_i es el conjunto de datos del i-ésimo cluster.
- x es un punto de datos dentro del cluster C_i.
- μ_i es el centroide del cluster C_i.
- $||x - \mu_i||^2$ representa la distancia euclidiana al cuadrado entre el punto de datos x y el centroide μ_i del cluster C_i.

3. **Crear un gráfico:** Representa la métrica calculada en el paso anterior en un gráfico, con el número de clusters en el eje horizontal (x) y la métrica en el eje vertical (y).

4. **Identificar el "codo":** Examina el gráfico y busca el punto en el que la métrica deja de disminuir significativamente y comienza a aplanarse. Este punto se asemeja a un "codo" en el gráfico, y su ubicación indica el número óptimo de clusters. En otras palabras, el número de clusters en el punto del codo se considera el número ideal para la estructura de tus datos.

El Método del Codo es una técnica útil para determinar el número óptimo de clusters, pero no siempre proporciona una solución definitiva. En algunos casos, el "codo" puede no ser claramente visible, y es posible que debas combinarlo con otras métricas y enfoques para tomar una decisión informada. Además, su efectividad depende de la forma y la estructura de tus datos, por lo que es importante utilizarlo en conjunto con otras técnicas de selección de clusters.

3.1.2 Nearest Neighbors

El Método de los Vecinos Más Cercanos es una técnica utilizada en el contexto del análisis de clusters para determinar el número óptimo. Aunque es menos común que el Método del Codo, puede ser útil en situaciones donde la identificación del "codo" en la curva de inercia no es clara o cuando los datos tienen características especiales.

El Nearest Neighbors Method se realiza siguiendo estos pasos:

1. **Aplicación de Clustering:** Al igual que en el Método del Codo, comienzas por aplicar un algoritmo de clustering, como K-Means (3.3.1), a tus datos para una serie de valores de k (el número de clusters). Puedes realizar múltiples ejecuciones del algoritmo para diferentes valores de k.

2. **Calcular Distancias:** Luego, para cada punto de datos en tu conjunto, calculas la distancia a su vecino más cercano. Puedes utilizar diversas métricas de distancia, como la distancia euclidiana, la distancia de Manhattan o cualquier otra métrica adecuada a tus datos y necesidades. En este estudio, usaremos la distancia euclidiana.

$$\text{Distancia Euclidiana} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

3. **Obtener Distancias Mínimas:** Para cada valor de k, calculas la distancia mínima entre todos los puntos de datos y sus vecinos más cercanos. Idealmente, a medida que incrementas k, esperas que estas distancias mínimas disminuyan.

4. **Identificar el Punto de Inflexión:** Finalmente, buscas el valor de k en el que la disminución de las distancias mínimas se estabiliza o presenta un punto de inflexión. Este valor de k se considera el número óptimo de clusters según el método de los Vecinos Más Cercanos.

El Método de los Vecinos Más Cercanos es especialmente útil cuando los datos tienen una estructura interna que se manifiesta en la distancia a sus vecinos más cercanos. Sin embargo, es importante recordar que no es una técnica infalible y puede no funcionar bien en todos los casos. Por lo tanto, se recomienda utilizarlo junto con otras técnicas de selección de clusters y métricas para determinar el número óptimo de clusters de manera más confiable. Para este estudio, utilizaremos este método para determinar el valor óptimo de ε, utilizado en la técnica DBSCAN (3.3.2). Recomiendo abiertamente mirar el siguiente video explicativo de este método (*Cómo elegir el epsilon en DBSCAN*).

3.2 Métricas para evaluar Clusters

Las métricas para evaluar clusters son herramientas que se utilizan para medir la calidad y la coherencia de los grupos o clusters generados por algoritmos de clustering en análisis de datos. Estas métricas son fundamentales para determinar la idoneidad de un conjunto de clusters. A continuación, se describirán 2 de las métricas que se utilizarán en el estudio.

3.2.1 Silhouette Index

El Índice de Silueta (Gutiérrez-García, 2023), es una métrica utilizada para evaluar la calidad de los grupos formados por algoritmos de clustering, como K-Means (3.3.1). Mide cuán similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación). El Índice de Silueta proporciona una medida de cuán bien separados están los grupos.

A continuación, una descripción del Índice de Silueta:

- Para cada punto de datos en el conjunto de datos, se calculan dos valores:
  - $a(i)$: La distancia promedio del punto de datos a todos los demás puntos de datos en el mismo grupo. Mide qué tan cerca está el punto de datos de los otros puntos en su grupo.
  - $b(i)$: La distancia promedio más pequeña del punto de datos a todos los puntos de datos en cualquier otro grupo al que no pertenezca. Mide cuán bien separado está el punto de datos de otros grupos.
- Se calcula la puntuación de la Silueta para cada punto de datos utilizando la fórmula:

$$\text{Silueta}(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}}$$

  Esta fórmula devuelve un valor en el rango [-1, 1], donde un valor alto indica que el objeto se ajusta bien a su propio grupo y está poco ajustado a grupos vecinos.
- Se calcula el promedio de las puntuaciones de Silueta de todos los puntos de datos para obtener el Índice de Silueta general para el clustering.

Interpretación del Índice de Silueta:

- Si el Índice de Silueta se acerca a 1, sugiere que el clustering es apropiado, con grupos bien definidos y separados.
- Si el Índice de Silueta está cerca de 0, indica grupos superpuestos o grupos mal definidos.

3.2.2 Davies-Bouldin Index

El Índice Davies-Bouldin (Gutiérrez-García, 2023) es una métrica utilizada para evaluar la calidad de los clusters en un problema de clustering. Se basa en la idea de que los clusters de alta calidad deben ser cohesivos (los puntos dentro del mismo cluster deben estar cerca) y bien separados (los clusters deben estar distantes entre sí). A continuación, se muestra la metodología para calcular el índice:

1. **Cohesión intra-cluster (a_i):** Para cada cluster "i", se calcula la cohesión, que mide cuán cerca están los puntos dentro del mismo cluster. La cohesión se calcula como la distancia promedio entre todos los pares de puntos en el mismo cluster. La fórmula para calcular la cohesión es:

$$a_i = \frac{1}{n_i} \sum_{j, j \neq i} d(x_j, x_i)$$

   Donde:

   - a_i es la cohesión del cluster "i".
   - n_i es el número de puntos en el cluster "i".
   - x_j y x_i son puntos de datos en el cluster "i".
   - $d(x_j, x_i)$ es la distancia entre los puntos x_j y x_i.

2. **Separación inter-cluster (s_{ij}):** Para cada par de clusters "i" y "j", se calcula la separación, que mide cuán separados están los clusters. La separación se calcula como la distancia entre los centroides de los clusters "i" y "j". La fórmula para calcular la separación es:

$$s_{ij} = d(c_i, c_j)$$

   Donde:

   - s_{ij} es la separación entre los clusters "i" y "j".
   - c_i es el centroide del cluster "i".
   - c_j es el centroide del cluster "j".
   - $d(c_i, c_j)$ es la distancia entre los centroides de los clusters "i" y "j".

3. **Índice Davies-Bouldin (DB_i):** Para cada cluster "i", el índice Davies-Bouldin se calcula como la relación entre la cohesión intra-cluster más baja y la separación inter-cluster más alta. La fórmula para calcular el índice Davies-Bouldin es:

$$DB_i = \max_{j, j \neq i} \left(\frac{a_i + a_j}{s_{ij}} \right)$$

   Donde:

   - DB_i es el índice Davies-Bouldin del cluster "i".
   - a_i es la cohesión del cluster "i".
   - a_j es la cohesión del cluster "j".
   - s_{ij} es la separación entre los clusters "i" y "j".

Interpretación del Índice de Davies-Bouldin (DB):

- Un valor bajo del Índice Davies-Bouldin (DB) indica que los clusters son de alta calidad, con buena cohesión intra-cluster y separación inter-cluster. En otras palabras, los clusters están bien definidos y distantes entre sí.

- Un valor alto de *DB* indica que los clusters pueden estar superpuestos, mal definidos o que la cohesión intra-cluster es baja. Esto sugiere una calidad de clustering deficiente.

- En la práctica, se busca minimizar el valor de *DB*. Además, el *DB* se utiliza para comparar diferentes resultados de clustering y seleccionar el número óptimo de clusters, ya que permite evaluar qué número de clusters proporciona la mejor separación y cohesión de los datos.

3.3 Técnicas de Clustering

El clustering es una técnica fundamental en el campo de la minería de datos y el análisis de patrones. Permite agrupar datos en conjuntos o clusters con características similares, lo que facilita la identificación de patrones y la segmentación de datos. Dos de las técnicas de clustering más utilizadas son K-Means y DBSCAN.

3.3.1 Método K-Means

El método K-Means (Gutiérrez-García, 2021) es un algoritmo de clustering utilizado para agrupar datos en clusters o grupos con características similares. Funciona a través de los siguientes pasos:

1. **Inicialización:** Se seleccionan aleatoriamente K centroides iniciales, donde K es el número de clusters deseado.

2. **Asignación de puntos:** Cada punto de datos se asigna al centroide más cercano, creando así K grupos iniciales.

3. **Actualización de centroides:** Se calcula el nuevo centroide para cada grupo, tomando el promedio de todos los puntos asignados a ese grupo.

4. **Reasignación de puntos:** Se vuelven a asignar los puntos a los centroides más cercanos, considerando los centroides actualizados.

5. **Repeticiones:** Los pasos 3 y 4 se repiten hasta que no haya cambios significativos en la asignación de puntos o se alcance un número máximo de iteraciones.

El objetivo del algoritmo K-Means es minimizar la distancia entre los puntos de datos y los centroides dentro de cada cluster. Al final, se obtienen K clusters donde los puntos dentro de cada cluster son similares entre sí, y los clusters están separados entre sí.

K-Means es un método de clustering ampliamente utilizado en aprendizaje automático y análisis de datos debido a su eficiencia computacional. Es importante tener en cuenta que el resultado del K-Means puede depender de la inicialización de los centroides, por lo que a veces es útil ejecutar el algoritmo varias veces con diferentes inicializaciones para obtener el mejor resultado.

Para el estudio, realizaremos mediante el Método del codo (3.1.1) la elección del número de clusters más óptimo. Para esto, se realizó un ciclo desde 2 clusters hasta 9 (el número total de variables independientes), luego se mide la inercia para cada ejercicio con el número de clusters (Cuadro 2).

Número Clusters	Inercia
2	1599.126291
3	1178.265700
4	966.261048
5	860.249912
6	817.326355
7	762.381069
8	719.458877
9	690.290272

Cuadro 2. Inercia para cada uno de los cálculos con k-clusters mediante el método K-Means

Una vez que se ha calculado la media de las distancias cuadradas entre los grupos, se procede a su representación gráfica, como se muestra en la Figura 2. La interpretación visual de esta gráfica se lleva a cabo con la ayuda de las distancias calculadas al trazar una línea recta entre el punto de inicio y el punto final de la gráfica, y luego midiendo la distancia desde el punto *x*, correspondiente al número de clusters del método hacia dicha línea recta. Según este análisis, se determina que, con una distancia máxima de *373.2*, el número óptimo de clusters para el conjunto de datos de este estudio es de **4**.

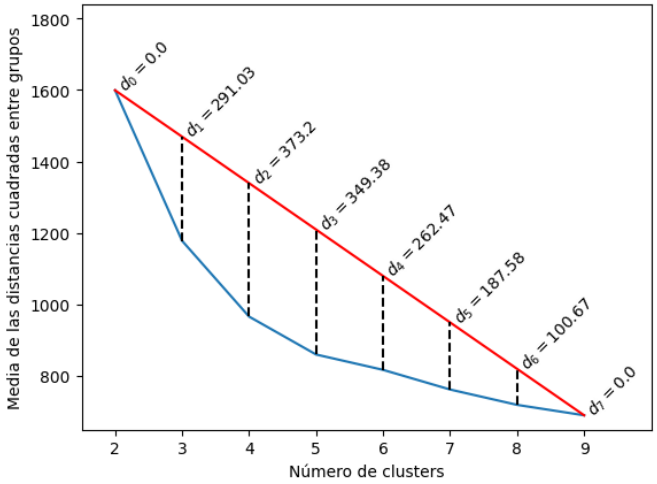

Figura 2. Inercia para cada uno de los cálculos con k-clusters mediante el método K-Means

3.3.2 DBSCAN

DBSCAN (Gutiérrez-García, 2022) es un algoritmo de clustering basado en la densidad de los puntos de datos. A diferencia del método K-Means (3.3.1), DBSCAN puede identificar clusters de diferentes formas y tamaños automáticamente, sin necesidad de especificar previamente el número de clusters. Su funcionamiento se basa en la noción de que un cluster es una región densa de puntos, separada por regiones de baja densidad. Este algoritmo es especialmente útil para la detección de outliers y la identificación de clusters no esféricos.

El algoritmo DBSCAN asigna puntos a tres categorías diferentes:

1. **Núcleo (*Core Points*):** Estos son puntos que poseen al menos un número mínimo de puntos vecinos que se encuentran a una distancia específica. Estos puntos conforman el núcleo de un cluster.

2. **Puntos Limítrofes (*Border Points*):** Son puntos que están dentro de la distancia de vecindad de un núcleo pero no tienen suficientes vecinos para ser considerados núcleos. Estos puntos se asignan a un cluster existente.

3. **Outliers (*Puntos Atípicos*):** Son puntos que no son núcleos ni puntos límite y quedan fuera de todos los clusters.

DBSCAN es capaz de identificar clusters de diferentes densidades y formas, lo que lo convierte en una opción robusta para muchos problemas de clustering. Además, no requiere la especificación previa del número de clusters, lo que lo hace especialmente útil en situaciones donde esta información no es conocida de antemano.

Sin embargo, es factible optimizar el proceso de cálculo mediante el análisis del parámetro ε. Para llevar a cabo este análisis, utilizaremos la metodología de *Nearest Neighbors* (ver Sección 3.1.2). En este estudio, empleamos un conjunto de 10 puntos, que incluyen el punto núcleo para calcular las distancias entre el punto núcleo y sus vecinos más cercanos. A continuación, determinamos el vecino más lejano para cada punto y representamos estos resultados en una gráfica, como se muestra en la Figura 3. De manera similar al método del codo (ver Sección 3.1.1), seleccionamos visualmente el valor que corresponde a la primera caída drástica en la gráfica. En este caso, identificamos un valor óptimo de ε igual a **0.21**.

4. Resultados y Discusión

Una vez determinado el número óptimo de clusters para el método K-Means y el ε óptimo para el método DBSCAN, podemos calcular las métricas y comparar resultados. Recordemos que para las métricas, utilizaremos el coeficiente de silueta (ver Sección 3.2.1) y el índice de Davis-Bouldin (ver Sección 3.2.2).

En el caso de K-Means, los 4 clusters se repartieron de acuerdo al Cuadro 3 . Calculando las métricas de Silueta y de Davis-Bouldin, obtuvimos para el primero un coeficiente de ***0.3915***, mientras que el segundo arrojó ***1.0396***.

Para el caso de DBSCAN, de acuerdo al ε de 0.21 calculado previamente, tuvimos una cantidad de 5 clusters, los cuales se repartieron de acuerdo al Cuadro 4 . Calculando las métricas de Silueta y de Davis-Bouldin, obtuvimos para el primero un coeficiente de ***0.2316***, mientras que el segundo arrojó ***4.7528***.

Si comparamos las métricas de Silueta y de Davis-Bouldin entre los métodos de K-Means y DBSCAN, y recordando que para el coeficiente de Silueta, entre más cercano a 1, es mejor y que para el índice de Davis-Bouldin, entre más cercano a 0,

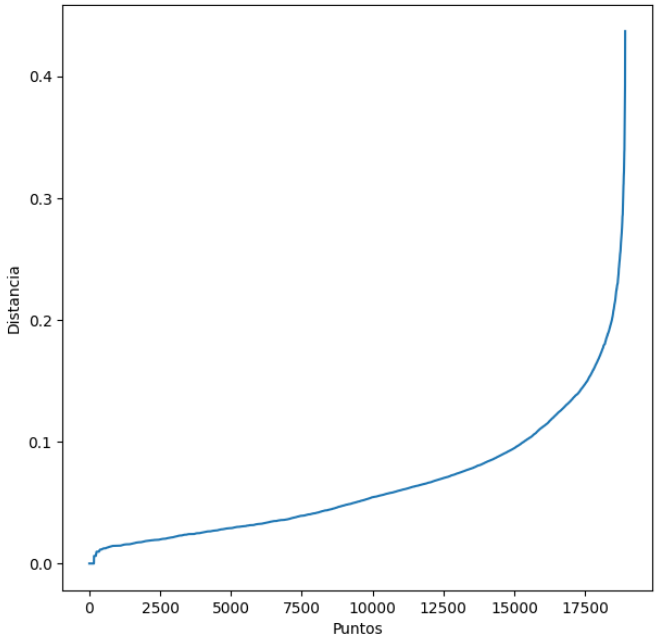

Figura 3. Distancia de cada punto con respecto a su vecino más lejano de entre los 10 más cercanos al núcleo

Cluster	Conteo
0	6868
1	5474
2	3310
3	3270

Cuadro 3. Dispersión de los puntos de acuerdo al método K-Means

es mejor. Podemos concluir que para este conjunto de datos, **el método K-Means es mejor**. Sin embargo, los números que arrojaron las métricas no son buenos, lo cual indica que los datos se encuentran muy dispersos y juntos.

5. Conclusión

En este estudio, evaluamos dos métodos de clustering ampliamente utilizados, K-Means y DBSCAN, aplicados a nuestro conjunto de datos de condiciones ambientales. Nuestra investigación revela que, en términos de la segmentación de datos, el método K-Means arrojó resultados más efectivos en comparación con DBSCAN para nuestro conjunto de datos en particular.

K-Means proporcionó una partición clara de los datos en clusters definidos, lo que resultó beneficioso en la identificación de grupos con características ambientales similares. Sin embargo, es importante destacar que, aunque K-Means demostró ser más efectivo en la partición de datos, las métricas de evaluación de los clusters resultaron ser modestas. Esto podría deberse a la naturaleza de nuestros datos o a la elección de los parámetros del algoritmo.

En contraste, DBSCAN se caracteriza por su capacidad para identificar clusters de diferentes formas y tamaños, y no

Cluster	Conteo
-1	124
0	2259
1	993
2	959
3	9664
4	4923

Cuadro 4. Dispersión de los puntos de acuerdo al método DBSCAN

requiere la especificación previa del número de clusters. A pesar de estas ventajas, no obtuvo resultados tan definidos en nuestro conjunto de datos específico.

En resumen, K-Means demostró ser la opción preferida para segmentar nuestro conjunto de datos de condiciones ambientales, pero es esencial continuar evaluando y ajustando los parámetros del algoritmo para mejorar las métricas de evaluación de los clusters. Los resultados aquí presentados ofrecen un punto de partida para futuras investigaciones y análisis en el ámbito de las condiciones ambientales y el clustering.

Referencias

Gutiérrez-García, J. (2021, Diciembre 6). *K-means (o K-medias) para detección de Clusters: Algoritmo e implementación con Python*. Código Máquina. https://youtu.be/mICySHB0fh4?si=UjvxJsebYPDjMSBg

Gutiérrez-García, J. (2022, Enero 31). *Identifica Clusters con DBSCAN: Algoritmo paso a paso e implementación con Python*. Código Máquina. https://youtu.be/HMis89lGdkA?si=1HeitI7VdRIXzHXN

Gutiérrez-García, J. (2023, 24 de Julio). *¿Qué tan buenos son tus Clusters? Métricas para Clustering con Python: Silueta y Davies Bouldin*. YouTube. https://youtu.be/b920s9nXGao?si=NhF8QGsh8AJ-QBu3

Kodinariya, T. M., Makwana, P. R., et al. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.