

Actividad 3 - Análisis del proceso de ML no supervisado de agrupación, K-MEANS & DBSCAN



Online
Universidad
Espíritu Santo

**Inteligencia
Artificial**

Unidad 2

Nombre de la materia

Aprendizaje Automático - MIAR0525

Nombre del alumno

David Alejandro Narváez Mejía

Nombre del Profesor

Gladys Villegas R. PhD(C)

Fecha

16/05/2024

Introducción

Temas

Tema 1

Generalidades del
Proyecto

Tema 2

Estadística Descriptiva y
EDA

Tema 3

Metodología y técnicas
aplicadas

Tema 4

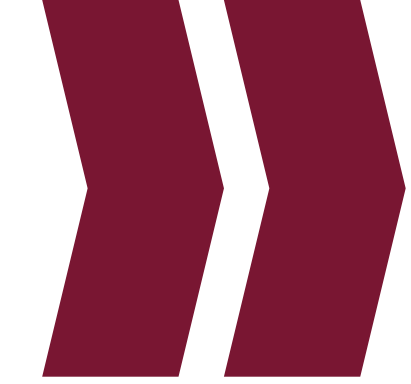
Conclusiones y Caso de
Estudio

Referencias:

[1] <https://www.kaggle.com/datasets/hrhuynguyen/2d-spatial-dataset/data>

[2] Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (3rd ed.). O'Reilly Media.

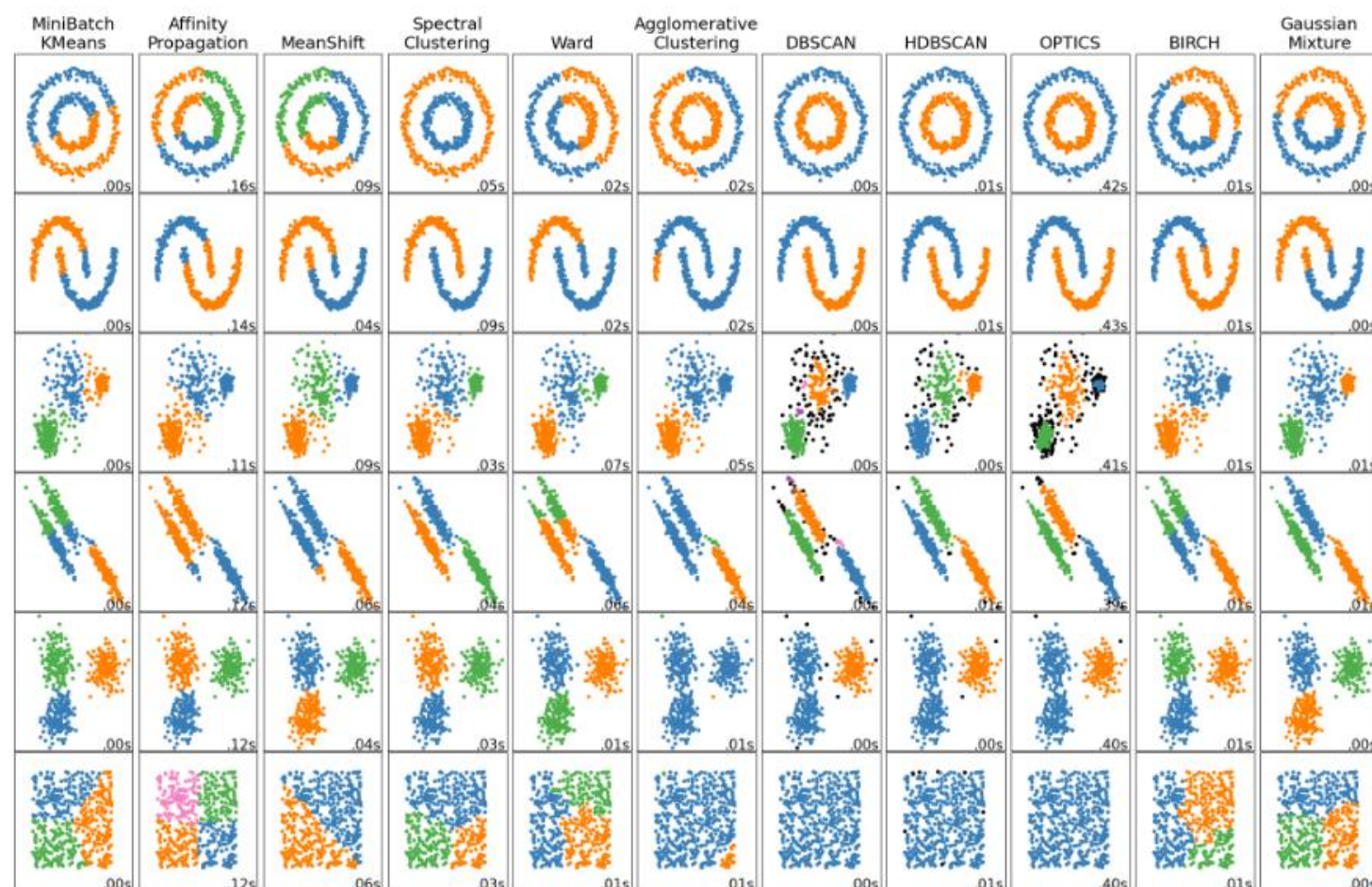
[3] <https://github.com/DAVOALEJO1987/CLUSTERING-K-MEANS-DBSCAN.git>



Objetivo

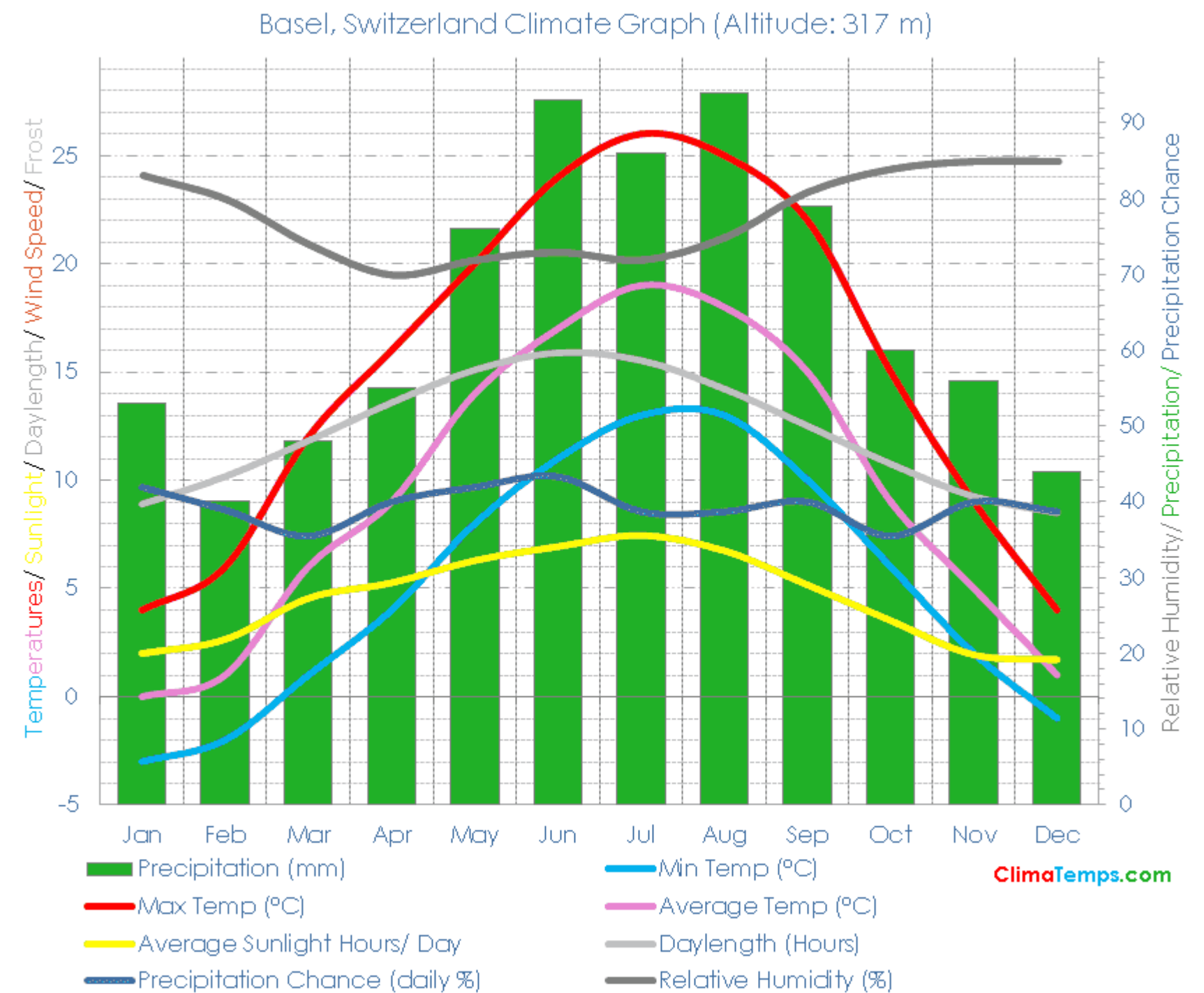
01

Analizar el proceso de ML no supervisado de clasificación, K-MEANS & DBSCAN del dataset meteorológicos de la ciudad de Basilea.

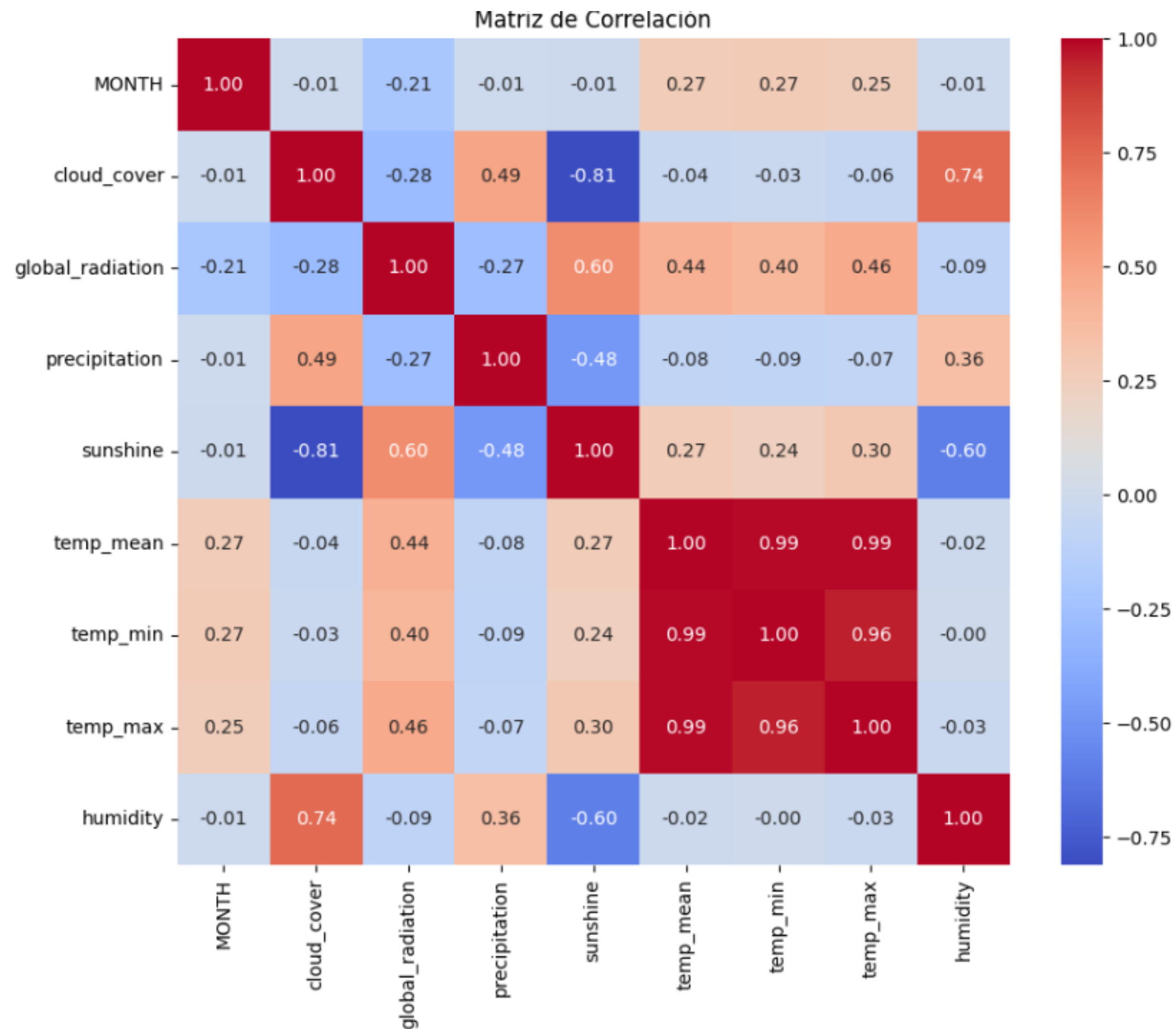


A comparison of the clustering algorithms in scikit-learn #

Tema 1: Dataset

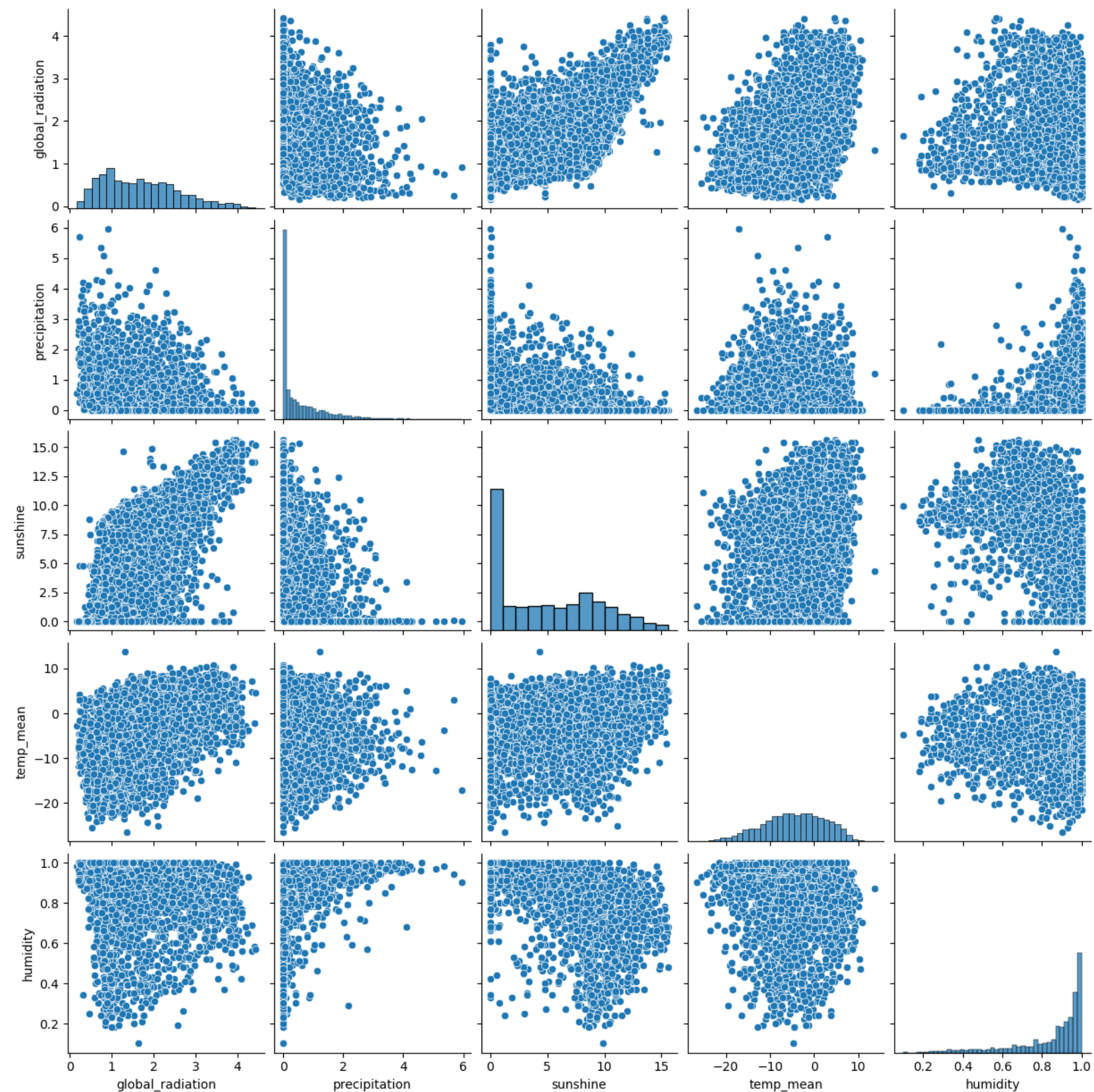


Tema 2: Estadística Descriptiva y EDA



Tema 2: Estadística Descriptiva y EDA

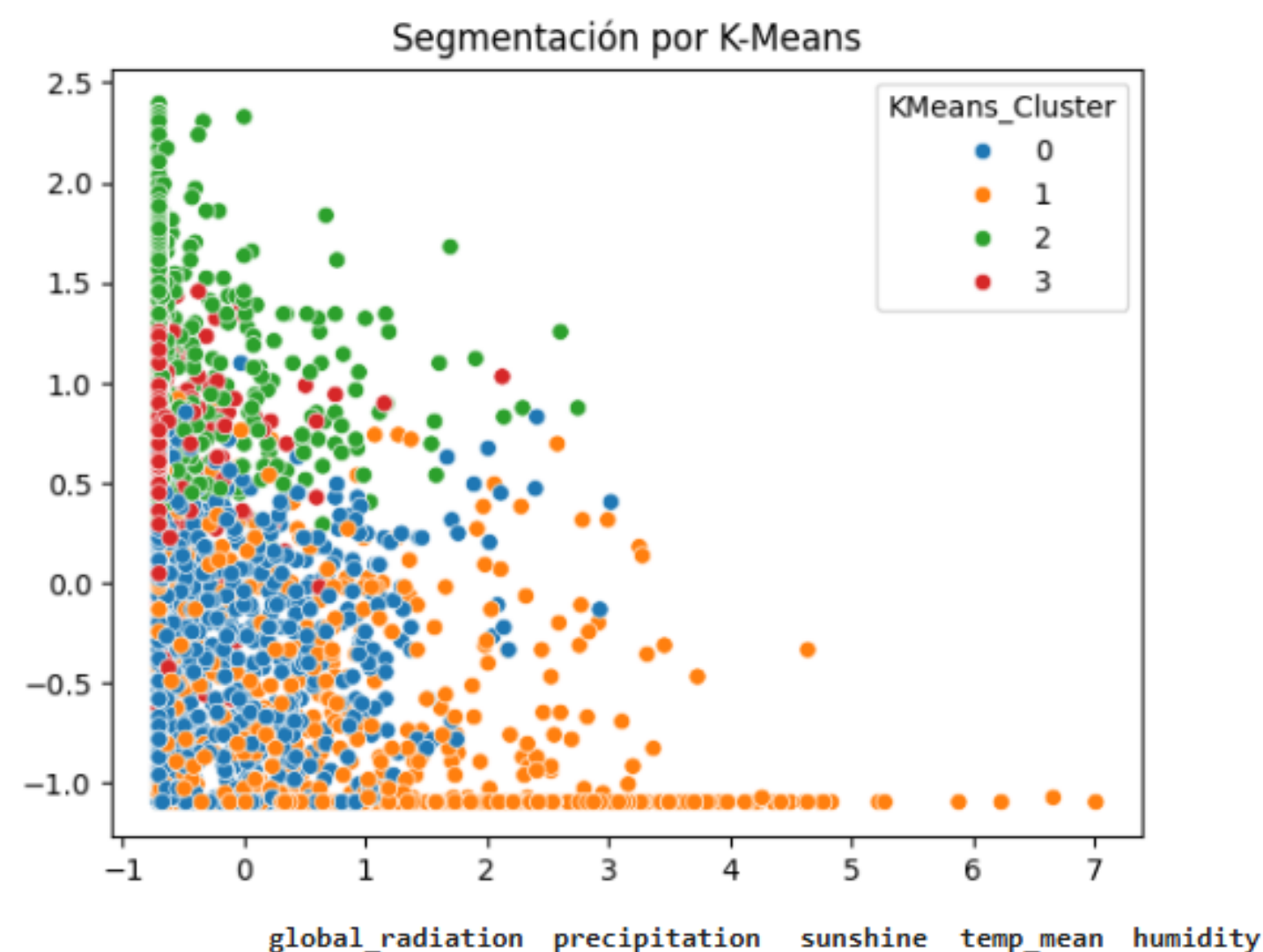
	global_radiation	precipitation	sunshine	temp_mean	humidity
count	3654.000000	3654.000000	3654.000000	3654.000000	3654.000000
mean	1.693919	0.541475	4.891078	-4.626327	0.853952
std	0.898277	0.771348	4.470904	6.987080	0.174900
min	0.170000	0.000000	0.000000	-26.600000	0.100000
25%	0.930000	0.000000	0.000000	-9.400000	0.800000
50%	1.600000	0.180000	4.300000	-4.400000	0.930000
75%	2.300000	0.840000	8.700000	0.700000	0.970000
max	4.420000	5.950000	15.600000	13.800000	1.000000



Tema 3: Metodología y técnicas aplicadas

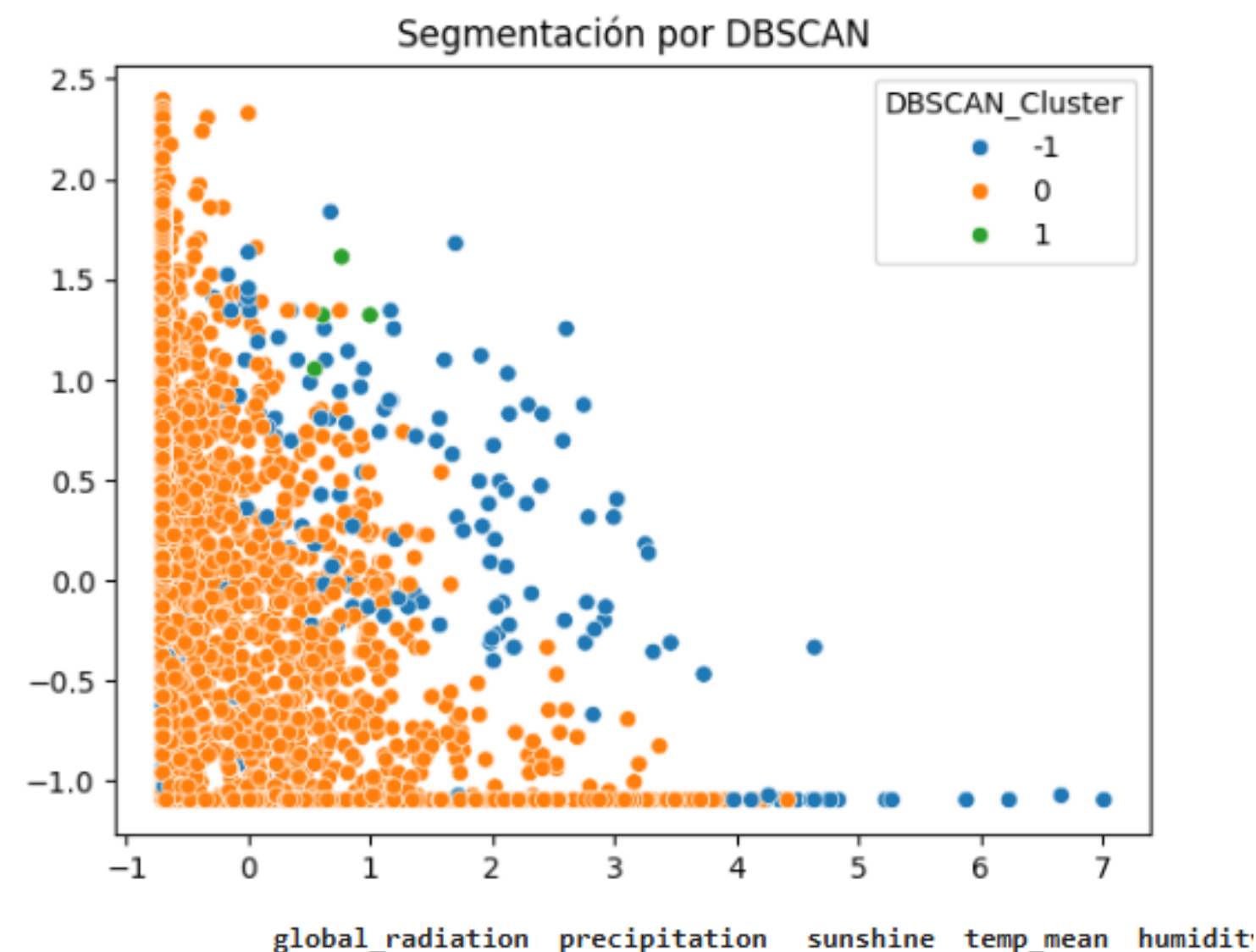
Método	Tipo	Ventajas	Desventajas	Aplicación en el código
K-Means	Clustering	<ul style="list-style-type: none">- Rápido y eficiente en grandes volúmenes de datos- Fácil de interpretar- Genera clusters bien definidos si son esféricos	<ul style="list-style-type: none">- Sensible a valores atípicos- Necesita definir el número de clusters (<code>k</code>)- No detecta formas no lineales	Segmentación de condiciones climáticas
DBSCAN	Clustering basado en densidad	<ul style="list-style-type: none">- Detecta clusters de cualquier forma- No requiere definir <code>k</code>- Detecta outliers automáticamente	<ul style="list-style-type: none">- Sensible a los parámetros <code>eps</code> y <code>min_samples</code>- Difícil con densidades variadas	Identificación de eventos extremos y ruido
PCA	Reducción dimensional	<ul style="list-style-type: none">- Reduce dimensiones preservando varianza- Mejora visualización- Facilita modelado posterior	<ul style="list-style-type: none">- Supone relaciones lineales- Puede perder interpretabilidad	Proyección de datos a 2D para graficar clusters
t-SNE	Reducción dimensional (no lineal)	<ul style="list-style-type: none">- Excelente para visualizar estructuras complejas- Captura relaciones no lineales locales	<ul style="list-style-type: none">- Computacionalmente costoso- No útil para modelado predictivo- No conserva escalas	Visualización clara de separación entre clusters

Tema 3: Metodología y técnicas aplicadas



KMeans_Cluster

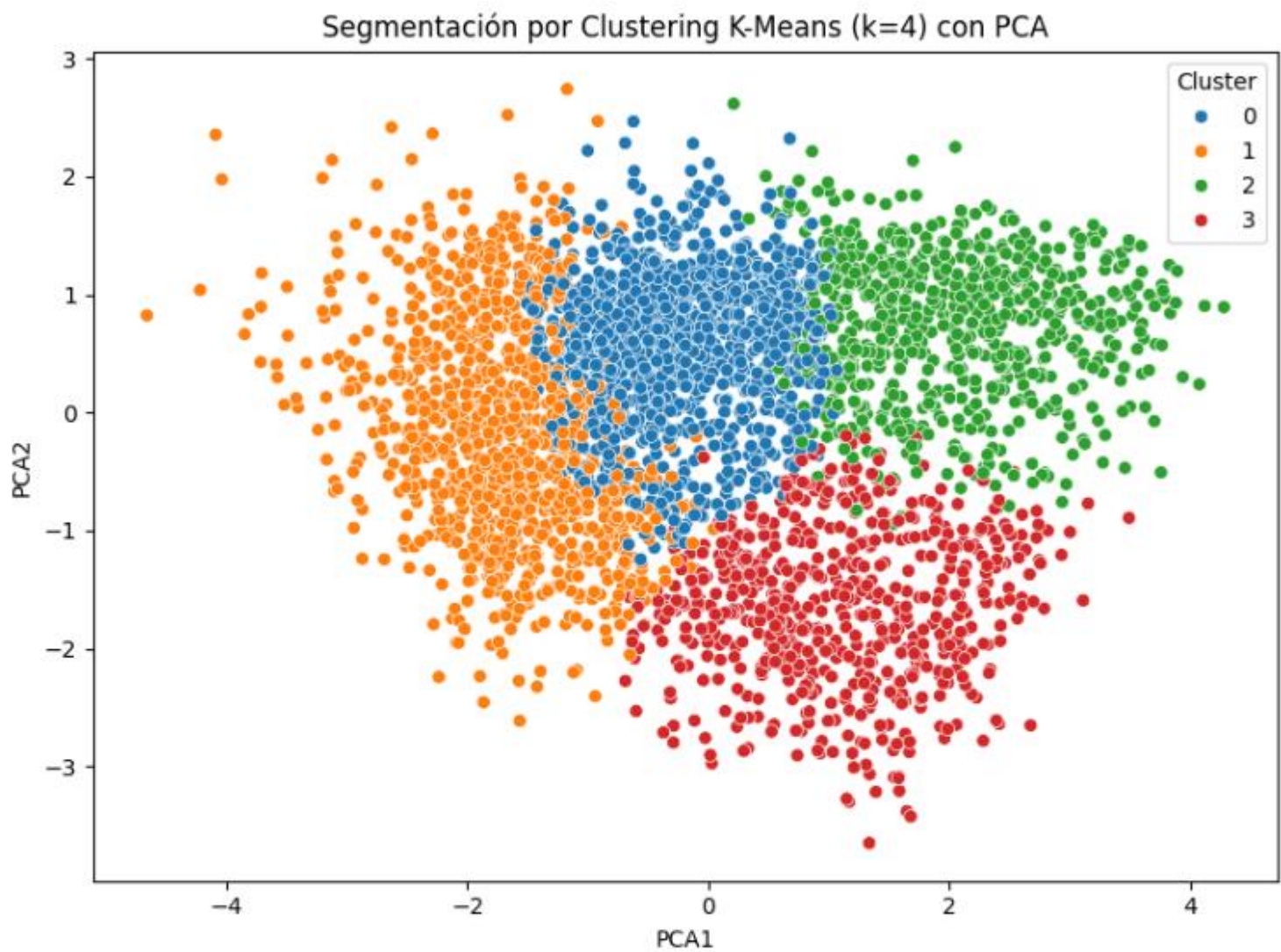
0	1.736848	0.400428	2.938765	-1.616955	0.946428
1	1.004649	1.237731	1.074446	-10.013469	0.948625
2	2.924340	0.152463	10.688435	1.329388	0.804585
3	1.356258	0.061726	8.517258	-8.165323	0.565726



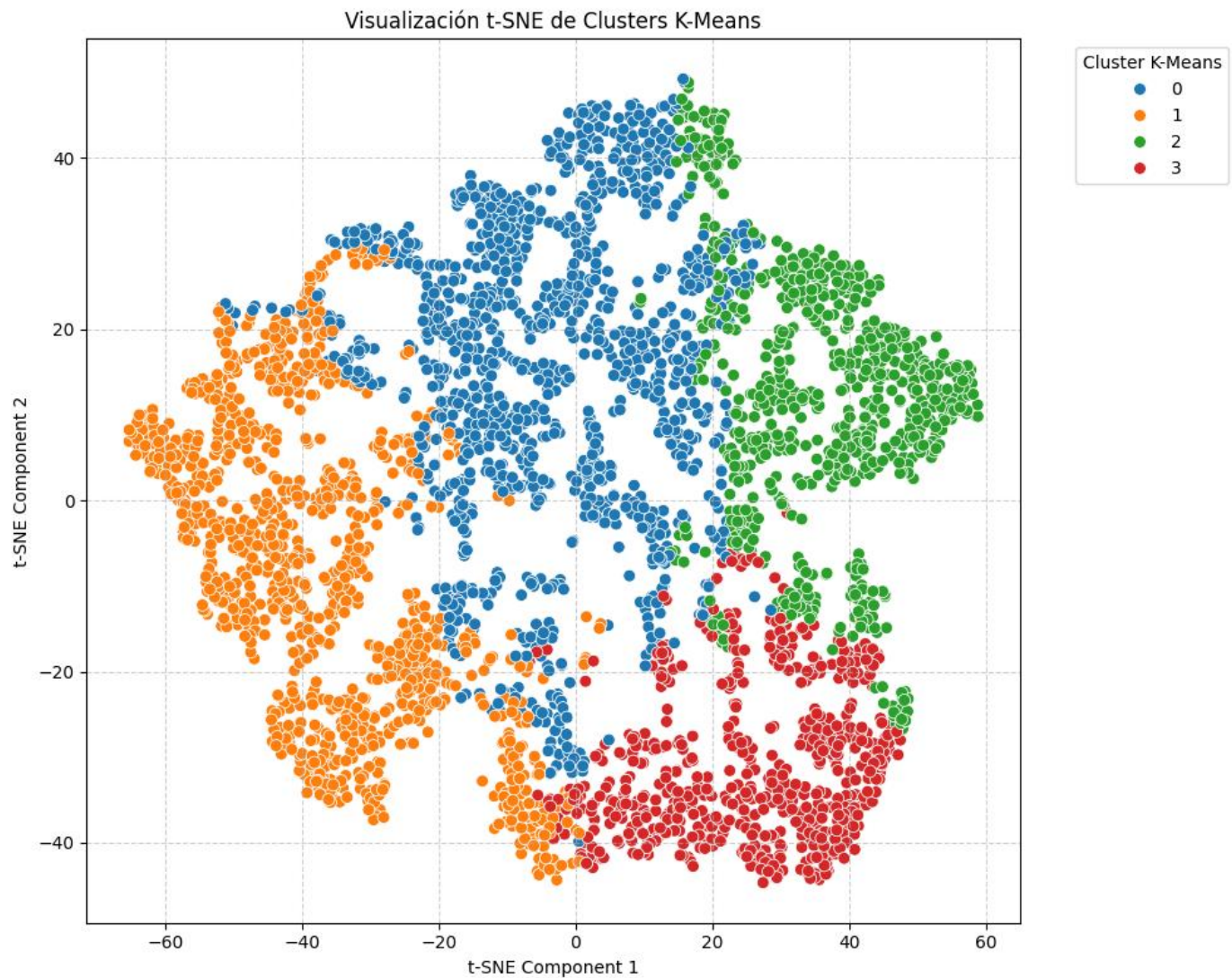
DBSCAN_Cluster

-1	1.875185	1.380556	5.629167	-5.735185	0.781891
0	1.680804	0.488043	4.837740	-4.569365	0.858571
1	3.165000	1.102500	10.825000	6.350000	0.772500

Tema 3: Metodología y técnicas aplicadas



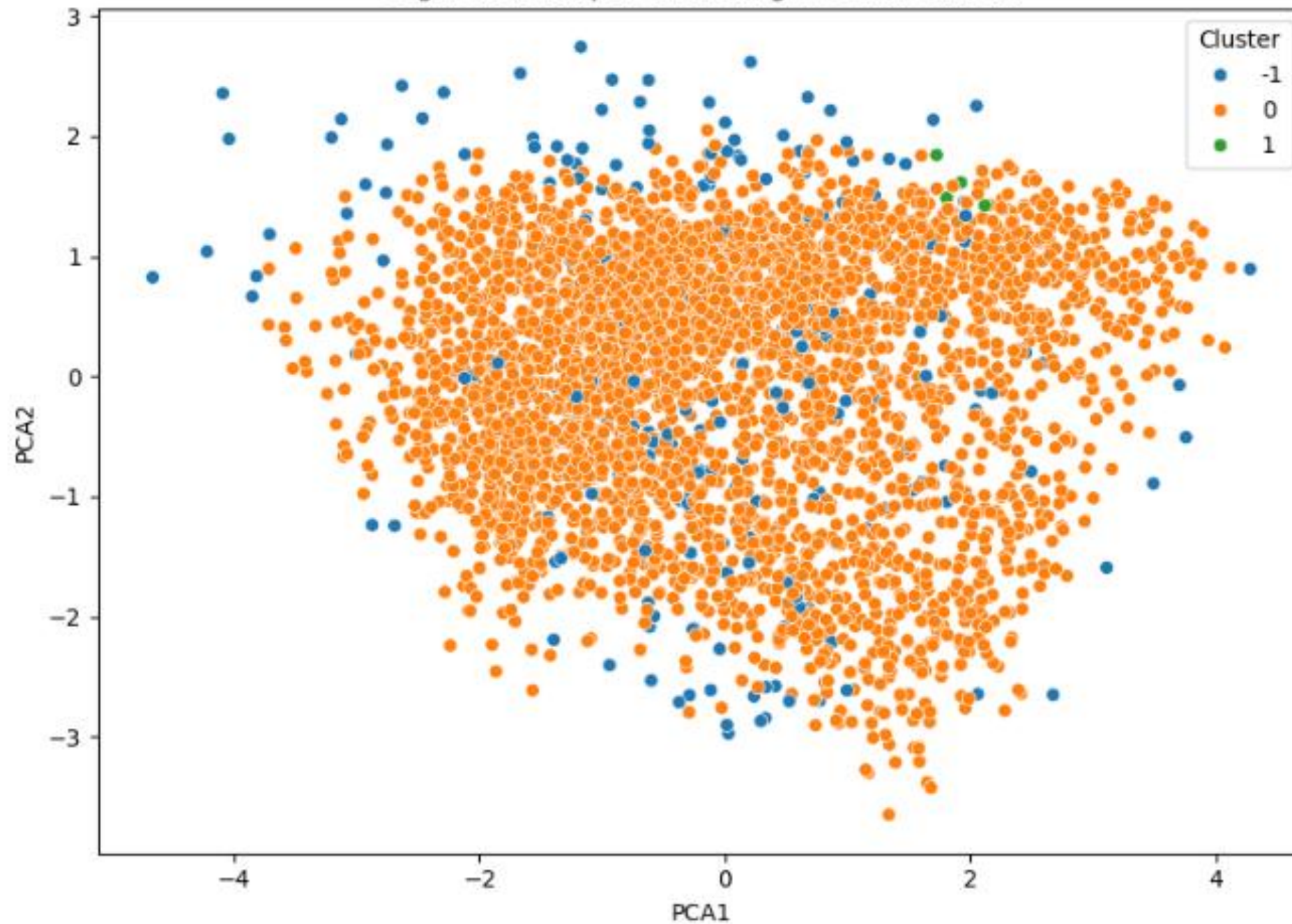
	global_radiation	precipitation	sunshine	temp_mean	humidity
KMeans_Cluster					
0	1.736848	0.400428	2.938765	-1.616955	0.946428
1	1.004649	1.237731	1.074446	-10.013469	0.948625
2	2.924340	0.152463	10.688435	1.329388	0.804585
3	1.356258	0.061726	8.517258	-8.165323	0.565726



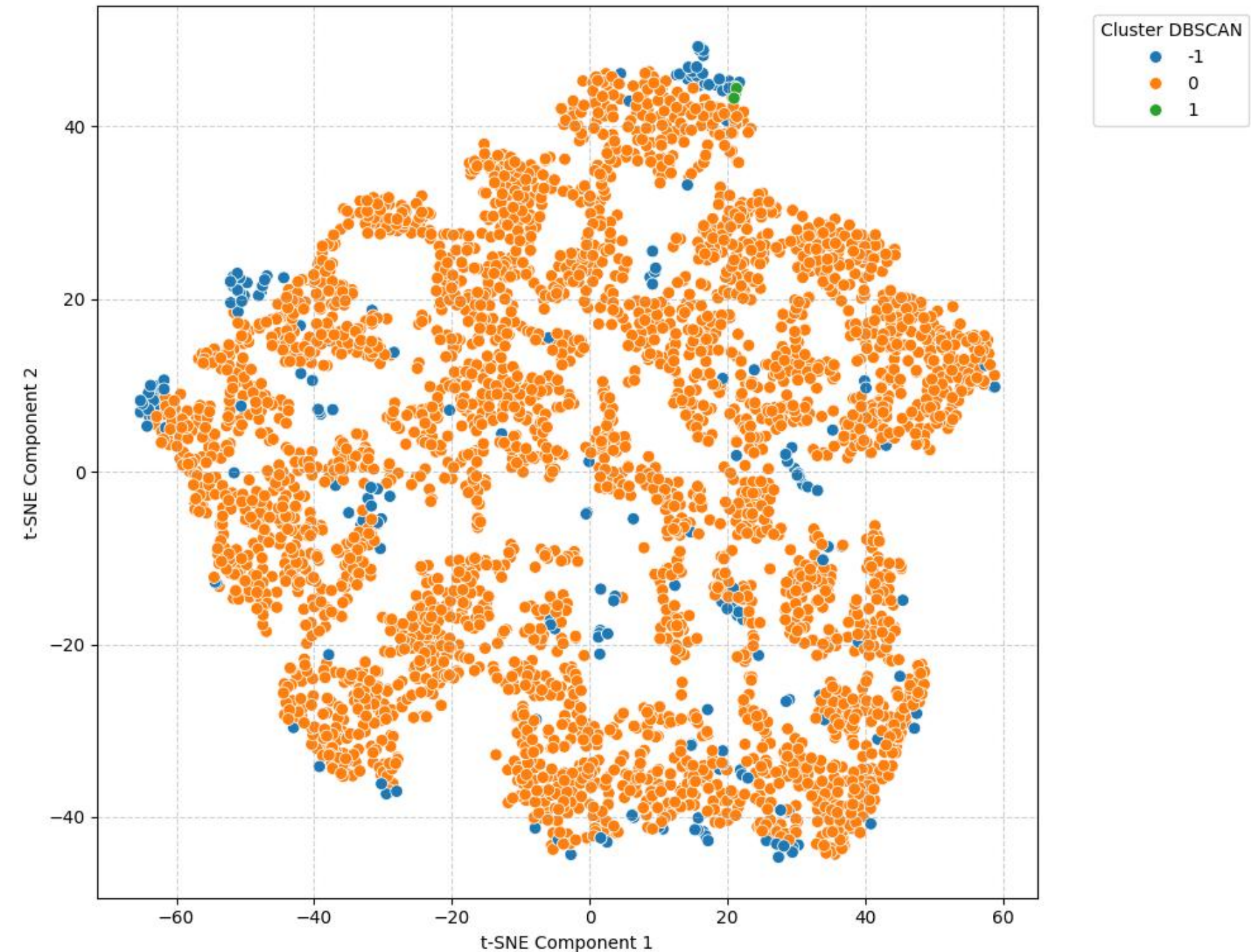
	global_radiation	precipitation	sunshine	temp_mean	humidity
KMeans_Cluster					
0	1.736848	0.400428	2.938765	-1.616955	0.946428
1	1.004649	1.237731	1.074446	-10.013469	0.948625
2	2.924340	0.152463	10.688435	1.329388	0.804585
3	1.356258	0.061726	8.517258	-8.165323	0.565726

Tema 3: Metodología y técnicas aplicadas

Segmentación por Clustering DBSCAN en PCA

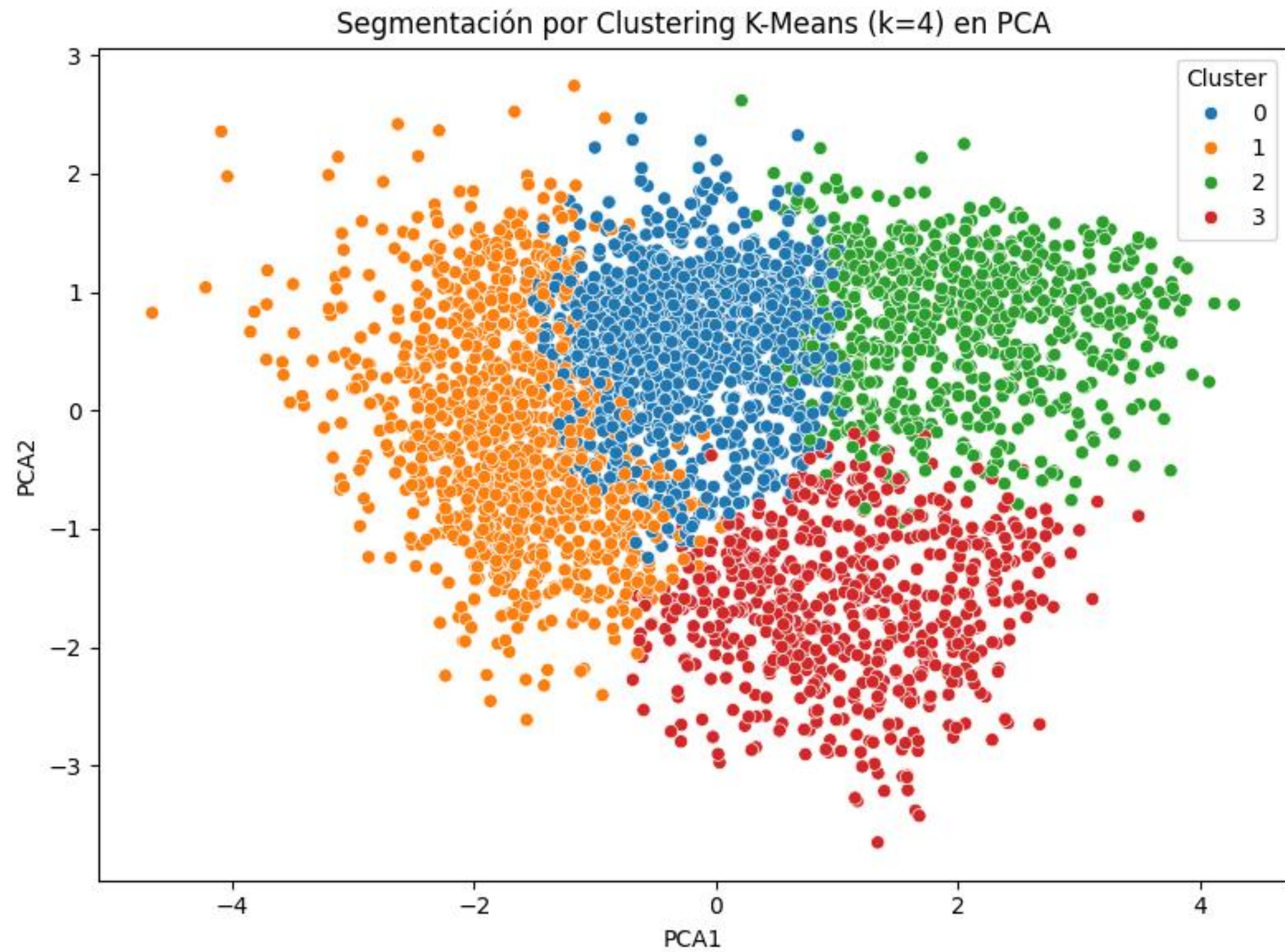


Visualización t-SNE de Clusters DBSCAN



Probando el ML

K-Means (0.25) ofrece mejor definición de grupos y es más confiable para interpretar y clasificar condiciones climáticas.



Caso de estudio WATERGEN

¿Cómo puede WATERGEN aprovechar la IA?

1. Optimización de generación de agua.
2. Geolocalización inteligente.
3. Modelos de predicción climática.
4. Gestión eficiente de energía.

