Aprendizaje Automático II: Tarea 2

Sergio Soler Rocha
Universidad Nacional de Educación a Distancia

1. Introducción

En este trabajo se aborda el problema de clustering no supervisado aplicado a un conjunto de datos astronómicos. Utilizando diversas técnicas de aprendizaje automático, como K-Means, Mezcla de Gaussianas y Clustering Jerárquico Aglomerativo, se pretende identificar la estructura subyacente de los datos. A través de un análisis exploratorio y el uso de métricas de validación interna, se determinará el número óptimo de clusters y se evaluará la efectividad de cada método.

2. Análisis exploratorio

El conjunto de datos contiene 43,351 entradas (filas), lo que indica una cantidad significativa de datos. Hay un total de 6 columnas en el conjunto de datos. Cada columna representa una variable numérica diferente y todas son del tipo float64, lo que significa que todas son variables numéricas de tipo continuo. Las columnas están etiquetadas como logP, logA11, R21, phi21, V-I, y WI, que son mediciones observaciones astronómicas. Cada una de las 6 columnas tiene 43,351 valores no nulos, lo que implica que no hay valores faltantes en el conjunto de datos. Esto es ideal porque significa que no necesitamos realizar imputación de datos antes de proceder con análisis más detallados o modelado. El conjunto de datos consume aproximadamente 2.0 MB de memoria. Este es un uso de memoria relativamente bajo, por tanto, manejar este conjunto de datos en memoria debería ser eficiente en la mayoría de los entornos modernos de computación.

A continuación presentamos información de la estadística descriptiva clave (media, desviación estándar, mínimo, máximo y cuartiles):

Cuadro 1: Estadísticas descriptivas de las variables

Statistic	**logP**	**logA11**	**R21**	**phi21**	**V-I**	**WI**
Mean	0.475023	0.467340	0.492901	0.533452	0.489598	0.531224
Std	0.289900	0.163169	0.061701	0.269675	0.116932	0.175788
Min	0.000000	0.000000	0.000000	0.000024	0.000000	0.000000
25 %	0.243317	0.334779	0.455108	0.340983	0.393517	0.381168
50 %	0.416001	0.443105	0.503527	0.524276	0.500953	0.577957
75 %	0.735259	0.605564	0.539280	0.770072	0.568497	0.662198
Max	1.000000	1.000000	1.000000	0.999990	1.000000	1.000000

En la media (Mean) Los valores cercanos a 0.5 en su mayoría podría indicar que, en promedio, las variables están distribuidas relativamente uniformemente sobre su rango. Los valores mínimos y máximos nos muestran el rango de datos. Las 6 variables tienen mínimos de 0 y máximos de 1, lo que podría indicar que han sido normalizadas o escaladas dentro de este rango.

Para entender mejor la naturaleza de los datos analizaremos los histogramas de las respectivas variables:

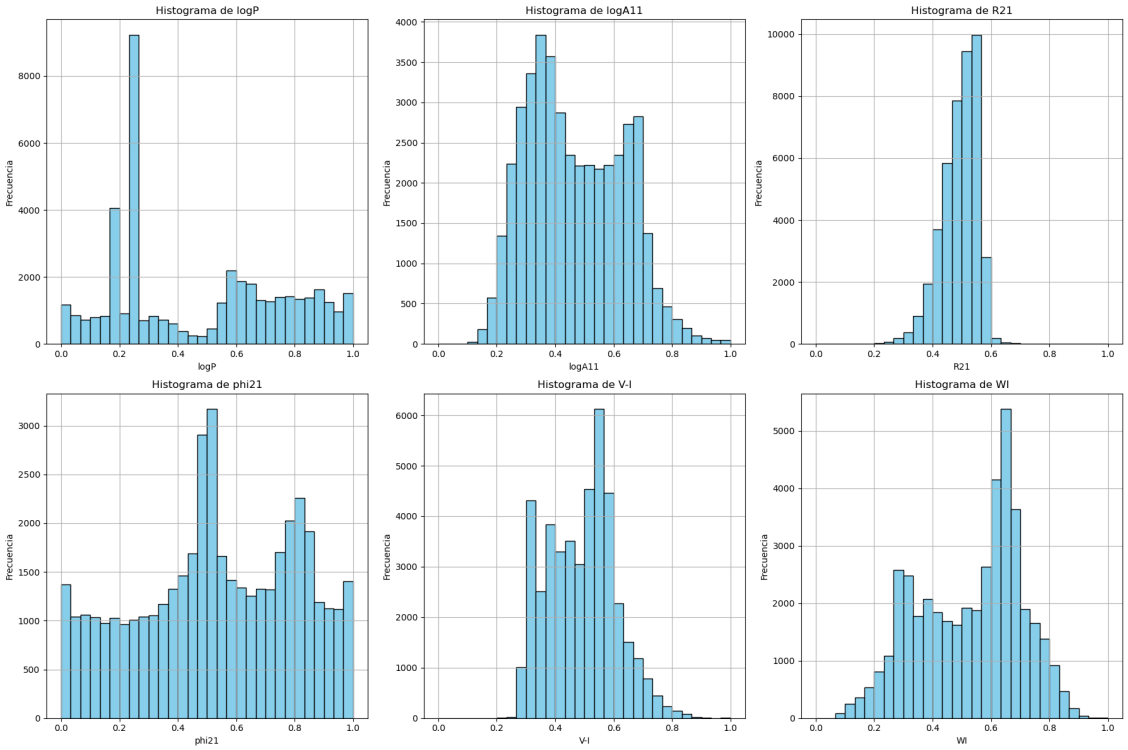

Figura 1: Histogramas de las variables del conjunto de datos

Podemos ver a simple vista que las variables LogP y phi21 paracen tener una distribución similar a una uniforme con algún pico en algunas zonas concretas. El resto de las gráficas tiene más semejanza a una distribución normal centrada en 0,5, Sin embargo, al realizar el test de normalidad D'Agostino no hay evidencias para asegurar de que sigan una distribución normal.

Otro punto importante a la hora de tratar con un problema de clustering es la presencia de outliers en los datos. Para ello, vamos a realizar un análisis de cajas, como se puede apreciar en la siguiente figura.

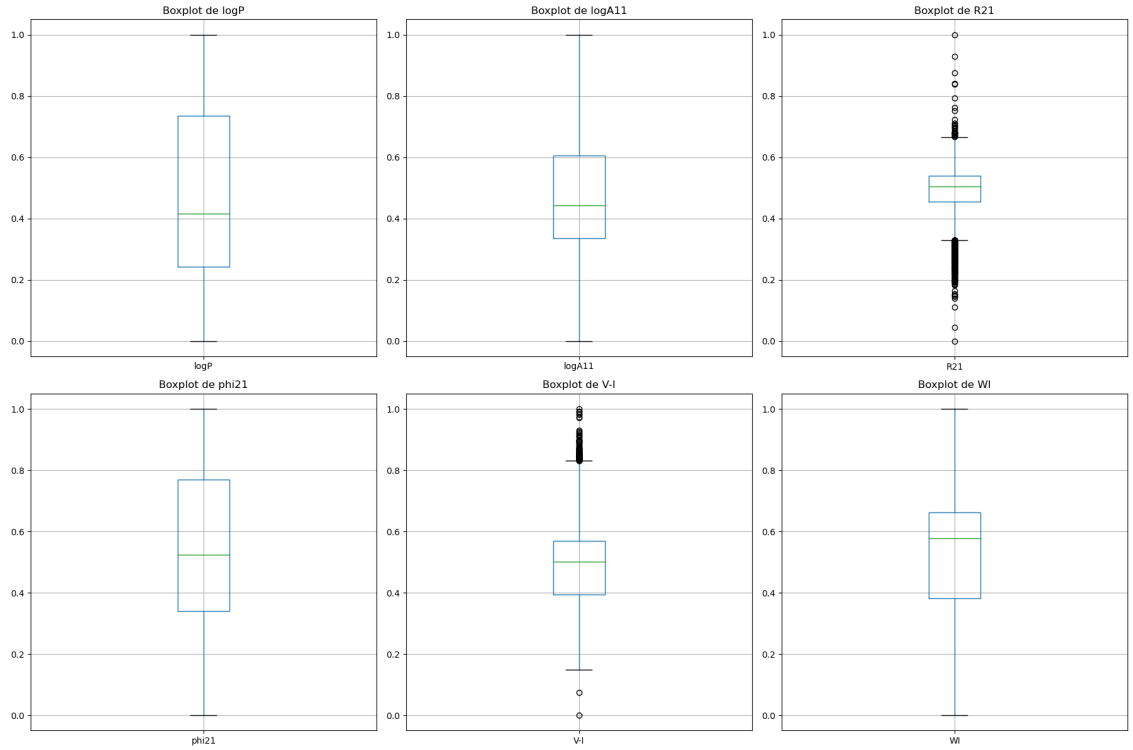

Figura 2: Análisis de cajas para identificar outliers en el conjunto de datos.

En general, observamos que los datos están centrados en torno a la mediana y presentan una uniformidad notable. Las dos únicas variables que contienen outliers son R21 y V-I; sin embargo, parece que estos están distribuidos de manera coherente, es decir, que aparecen debido a la naturaleza propia de las variables R21 y V-I, y no a ninguna anomalía o error al generar los datos. Por tanto,sabiendo esto, continuaremos el análisis teniendo en cuenta todos los datos sin descartar ningún valor atípico.

Por último en este apartado, analizaremos cual es la correlación lineal de las variables. En la siguiente figura se muestra la matriz de correlación de las 6 variables del conjunto de datos:

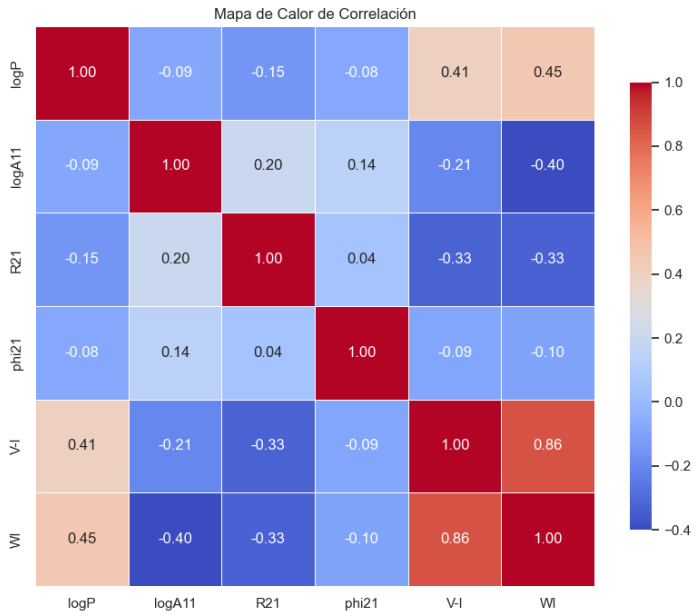

Figura 3: Mapa de calor de la matriz de correlación de las variables del conjunto de datos.

Por lo general, observamos que existe una correlación significativa entre algunas de las variables. La más destacada es la correlación entre WI y V-I, que es de 0.86, bastante cercana a 1. Este hecho podría tenerse en cuenta a la hora de realizar el clustering, ya que podríamos prescindir de una de estas dos variables en algún momento sin perder información relevante del análisis.

3. Clustering

Dado que estamos abordando un problema de clustering no supervisado con el conjunto de datos proporcionado, emplearemos tres algoritmos distintos de clustering para explorar la estructura subyacente de los datos. Los métodos seleccionados incluyen K-Means, Mezcla de Gaussianas (Gaussian Mixture Model, GMM) y Clustering Jerárquico Aglomerativo. Cada uno de estos algoritmos se aplica debido a sus características únicas y capacidades para identificar agrupaciones de datos basadas en diferentes principios:

- **K-Means**: Este método particiona los datos en K clusters en los que cada observación pertenece al cluster con la media más cercana. Es conocido por su eficiencia en grandes conjuntos de datos. La medida de distancia es la Euclidiana.
- **Mezcla de Gaussianas**: Este enfoque modela el conjunto de datos como una mezcla de varias distribuciones gaussianas. Se utiliza por su flexibilidad en la forma de los clusters y su capacidad para incorporar la covarianza entre las características en el modelo de clustering. La medida de distancia es la distancia Mahalanobis, que es una medida que toma en cuenta las correlaciones de los datos y la varianza en cada dirección, se utiliza para calcular la probabilidad de que un punto de datos pertenezca a una determinada distribución gaussiana dentro de la mezcla.
- **Clustering Jerárquico Aglomerativo**: Este algoritmo construye un árbol de clusters y es útil para proporcionar una visualización dendrográfica. Se emplea para comprender la jerarquía y particionar los datos a diferentes niveles de granularidad. La distancia utilizada es la Euclidiana para calcular la distancia entre puntos, combinada con métodos de enlace como el enlace simple (single linkage), enlace completo (complete linkage) o enlace promedio (average linkage) para determinar la distancia entre clusters.

Para evaluar y comparar la efectividad de cada algoritmo de clustering, se utilizarán tres métricas de validación interna: **Coeficiente de Silueta**, **Índice de Davies-Bouldin**, **Índice Calinski-Harabasz** que expondremos en el siguiente subapartado.

Utilizando estas métricas, inferiremos el número óptimo de clusters para cada método. Los resultados no solo ayudarán a identificar la configuración más adecuada para el agrupamiento sino también a comparar la eficacia de diferentes técnicas de clustering bajo diversas configuraciones

3.1. Coeficiente de Silueta

El Coeficiente de Silueta es una métrica ampliamente utilizada para evaluar la calidad de los clusters en análisis de datos no supervisados. Este coeficiente no solo toma en cuenta la cohesión interna de los clusters, sino también su separación respecto a otros clusters.

El Coeficiente de Silueta para una muestra individual se define como:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}} \tag{1}$$

donde:

- $a(i)$ representa la distancia media entre la muestra i y todos los otros puntos en el mismo cluster, reflejando la cohesión del cluster.
- $b(i)$ es la distancia media entre la muestra i y todos los puntos en el cluster más cercano del que i no es miembro, reflejando la separación del cluster.

El coeficiente varía de -1 a +1, donde:

- Valores cercanos a +1 indican que la muestra está bien emparejada a su propio cluster y lejos de otros clusters.
- Valores cercanos a 0 indican que la muestra podría estar en el borde entre dos clusters.
- Valores cercanos a -1 sugieren que la muestra podría estar asignada al cluster incorrecto.

El coeficiente de silueta se utiliza para:

- Determinar el número óptimo de clusters al variar el número de clusters y observar los cambios en el coeficiente medio de silueta.
- Se considera que el número de clusters que produce el valor más alto del coeficiente de silueta es el más óptimo para ese conjunto de datos, bajo el método de clustering que se esté utilizando

3.2. Índice de Davies-Bouldin

El Índice de Davies-Bouldin es una medida de evaluación interna utilizada para estimar la calidad de los clusters en un análisis de clustering. Este índice combina aspectos de cohesión interna y separación entre clusters para proporcionar una evaluación integral de la partición de un conjunto de datos.

El Índice de Davies-Bouldin para un conjunto de clusters se define como el promedio de las peores medidas de similitud entre cada cluster y su cluster más cercano. Matemáticamente, se expresa como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2}$$

donde R_{ij} es la medida de similitud entre los clusters i y j, definida por:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3}$$

Aquí, s_i y s_j representan las dispersiones dentro de los clusters i y j, respectivamente, y d_{ij} es la distancia entre los centroides de estos clusters.

- Un valor bajo de DB indica que los clusters son internamente densos y bien separados entre sí, lo cual es deseable en un buen clustering.
- Este índice se utiliza comúnmente para comparar la calidad de diferentes configuraciones de clustering o para determinar el número óptimo de clusters en un modelo.

3.2.1. Ventajas y Limitaciones

- **Ventajas:** El índice es fácil de calcular y no requiere conocimiento de etiquetas verdaderas, lo que es ideal para situaciones no supervisadas.
- **Limitaciones:** Puede ser sensible a outliers y no siempre efectivo para clusters de formas no esféricas o tamaños variados.

3.3. Índice Calinski-Harabasz

El índice Calinski-Harabasz, también conocido como la Varianza Ratio Criterion, es una métrica de evaluación interna utilizada para medir la calidad de los clusters en análisis de clustering. Este índice es especialmente útil para determinar el número óptimo de clusters dentro de un conjunto de datos.

El índice Calinski-Harabasz se define matemáticamente como:

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \tag{4}$$

donde:

- SSB es la suma de cuadrados entre los clusters.
- SSW es la suma de cuadrados dentro de los clusters.
- k es el número de clusters.
- n es el número total de puntos o muestras en el conjunto de datos.

La interpretación del índice Calinski-Harabasz es la siguiente:

- Un valor alto del índice Calinski-Harabasz sugiere que los clusters están bien separados y que cada cluster es densamente agrupado, lo que implica una buena calidad de clustering.
- Este índice evalúa la dispersión entre clusters en relación con la dispersión dentro de los clusters, proporcionando una medida clara y efectiva de la calidad del clustering.

El índice Calinski-Harabasz se utiliza comúnmente para:

- Comparar la calidad de diferentes configuraciones de clustering.
- Determinar el número óptimo de clusters, ayudando a los investigadores y analistas a tomar decisiones informadas sobre la segmentación de los datos.

3.3.1. Ventajas y Limitaciones

- **Ventajas:** El índice es relativamente fácil de calcular y proporciona una evaluación rápida de la calidad del clustering basada en criterios cuantitativos.
- **Limitaciones:** Puede ser sensible a la forma de los clusters y no siempre es adecuado para datos con estructuras de cluster complejas o atípicas.

3.4. Resultados de Evaluación de Clustering para K-Means

A continuación, se presentan los resultados de las métricas de evaluación interna para el algoritmo K-Means aplicado a un conjunto de datos.

Cuadro 2: Métricas de Evaluación para K-Means

Número de Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
2	0.335	1.272	24056.219
3	0.303	1.210	20410.944
4	0.290	1.266	20224.497
5	0.289	1.203	18961.879
6	0.289	1.133	17653.923
7	0.294	1.164	16929.124
8	0.290	1.131	16561.408
9	0.281	1.155	15826.040
10	0.282	1.173	15306.984

3.5. Resultados de Evaluación de Clustering para Mezcla de Gaussianas

A continuación, se presentan los resultados de las métricas de evaluación interna para el método de Mezcla de Gaussianas aplicado a un conjunto de datos.

Cuadro 3: Métricas de Evaluación para Mezcla de Gaussianas

Número de Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
2	0.331	1.283	23661.729
3	0.222	2.038	12453.748
4	0.126	1.985	11058.090
5	0.104	2.384	9059.644
6	0.112	2.466	8318.826
7	0.088	2.224	6954.512
8	0.108	2.010	8262.476
9	0.073	1.988	7375.955
10	0.078	1.879	7143.811

3.6. Resultados de Evaluación para Clustering Jerárquico Aglomerativo

A continuación, se presentan los resultados de las métricas de evaluación interna para el método de Clustering Jerárquico Aglomerativo aplicado a un conjunto de datos.

Cuadro 4: Métricas de Evaluación para Clustering Jerárquico Aglomerativo

Número de Clusters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
2	0.330	1.283	23641.697
3	0.277	1.380	17554.538
4	0.253	1.401	17056.641
5	0.252	1.323	15996.256
6	0.228	1.507	14742.615
7	0.231	1.391	13833.514
8	0.215	1.332	13338.470
9	0.222	1.284	12880.373
10	0.213	1.297	12418.592

4. Análisis de Métricas de Evaluación

4.1. K-Means

4.1.1. Coeficiente de Silueta

Valores: Los valores oscilan entre 0.281 y 0.335, con el valor más alto de 0.335 para 2 clusters.
Interpretación: El Coeficiente de Silueta más alto para 2 clusters indica que esta configuración ofrece una mejor cohesión y separación de los clusters comparado con configuraciones de mayor número de clusters. Los valores más altos sugieren que los puntos dentro de cada cluster están más cerca entre sí, y más lejos de los puntos de otros clusters, lo que implica que 2 clusters pueden ser una buena elección.

4.1.2. Índice de Davies-Bouldin

Valores: Los valores oscilan entre 1.132 y 1.271, con el valor más bajo de 1.132 para 6 clusters.
Interpretación: Un índice Davies-Bouldin más bajo indica mejor separación y compacidad de los clusters. El valor más bajo para 6 clusters sugiere que esta configuración podría tener clusters más densos y mejor separados que otras configuraciones.

4.1.3. Índice Calinski-Harabasz

Valores: Los valores disminuyen de 24056.219 para 2 clusters a 15306.984 para 10 clusters.
Interpretación: Este índice muestra una tendencia decreciente a medida que aumenta el número de clusters. El valor más alto para 2 clusters sugiere que los clusters están muy bien diferenciados y definidos en comparación con configuraciones de mayor número de clusters. Un valor más alto implica una mayor dispersión entre los clusters en comparación con la dispersión dentro de ellos.

4.2. Mezcla de Gaussianas

4.2.1. Coeficiente de Silueta

Valores: Oscilan entre 0.073 y 0.331, siendo el valor más alto de 0.331 para 2 clusters.
Interpretación: Este valor más alto para 2 clusters sugiere que la configuración con dos grupos ofrece una mejor cohesión y separación de los clusters en comparación con configuraciones de mayor número de clusters. Los valores altos indican que los puntos dentro de cada cluster están más cerca entre sí y más alejados de los puntos de otros clusters, lo que implica que 2 clusters pueden ser una buena elección desde la perspectiva del Coeficiente de Silueta.

4.2.2. Índice de Davies-Bouldin

Valores: Varían entre 1.282 y 2.466, con el valor más bajo de 1.282 para 2 clusters.
Interpretación: Un índice Davies-Bouldin más bajo indica una mejor separación y compacidad de los clusters. El valor más bajo para 2 clusters sugiere que esta configuración podría tener clusters más densos y mejor separados que otras configuraciones. Un valor bajo es deseable ya que implica que los clusters no solo están bien separados sino que también son compactos.

4.2.3. Índice Calinski-Harabasz

Valores: Los valores disminuyen desde 23661.729 para 2 clusters a 7143.811 para 10 clusters.
Interpretación: Este índice muestra una tendencia decreciente conforme aumenta el número de clusters. El valor más alto para 2 clusters sugiere que los clusters están muy bien diferenciados y definidos en comparación con configuraciones de mayor número de clusters. Un valor más alto implica una mayor dispersión entre los clusters en comparación con la dispersión dentro de ellos, indicando una estructura de clusters más clara y definida para configuraciones con menos clusters.

4.3. Clustering Jerárquico Aglomerativo

4.3.1. Coeficiente de Silueta

Valores: Oscilan entre 0.213 y 0.330, siendo el valor más alto de 0.330 para 2 clusters.
Interpretación: Este valor más alto para 2 clusters sugiere que esta configuración ofrece una

buena cohesión y separación de los clusters en comparación con configuraciones de mayor número de clusters. Los valores altos indican que los puntos dentro de cada cluster están más cerca entre sí y más lejos de los puntos de otros clusters. Esto implica que 2 clusters pueden ser una elección óptima para obtener una estructura clara y bien definida.

4.3.2. Índice de Davies-Bouldin

Valores: Varían entre 1.283 y 1.507, con el valor más bajo de 1.283 para 2 clusters.
Interpretación: Un índice Davies-Bouldin más bajo indica mejor separación y compacidad de los clusters. El valor más bajo para 2 clusters sugiere que esta configuración podría ofrecer clusters más densos y mejor separados que otras configuraciones. Un valor bajo es deseable ya que indica que los clusters no solo están bien separados sino que también son compactos, lo que es crucial para un buen clustering.

4.3.3. Índice Calinski-Harabasz

Valores: Los valores disminuyen desde 23641.697 para 2 clusters a 12418.592 para 10 clusters.
Interpretación: Este índice muestra una tendencia decreciente conforme aumenta el número de clusters. El valor más alto para 2 clusters sugiere que los clusters están muy bien diferenciados y definidos en comparación con configuraciones de mayor número de clusters. Un valor más alto implica una mayor dispersión entre los clusters en comparación con la dispersión dentro de ellos, indicando una estructura de clusters más clara y definida para configuraciones con menos clusters.

5. Analisis General

La consistencia entre las tres técnicas de clustering y las tres métricas de evaluación interna sugiere fuertemente que 2 clusters es el número óptimo para este conjunto de datos. Esta configuración ofrece la mejor cohesión y separación, lo cual es indicado por los altos valores de Calinski-Harabasz y Coeficiente de Silueta, así como los bajos valores en el Índice de Davies-Bouldin.

La siguiente tabla muestra el número óptimo de clusters determinado por diferentes métricas de evaluación interna para cada algoritmo de clustering utilizado en el análisis.

Cuadro 5: Número Óptimo de Clusters por Algoritmo y Métrica

Algoritmo	**Coeficiente de Silueta**	**Davies-Bouldin**	**Calinski-Harabasz**
K-Means	2	6	2
Mezcla de Gaussianas	2	2	2
Clustering Jerárquico	2	2	2

Para tener más certezas de que el número óptimo de agrupación es de 2 reduciremos las dimensiones utilizando el método de Análisis de Componentes Principales (PCA) que busca reducir la dimensionalidad de los datos identificando las direcciones (componentes principales) a lo largo de las cuales varían más los datos. Al reducir los datos a dos componentes principales, podemos graficar cada muestra en un plano bidimensional, donde el eje x puede ser el primer componente principal y el eje y el segundo. Esto facilita la observación de agrupaciones naturales o separaciones entre datos.

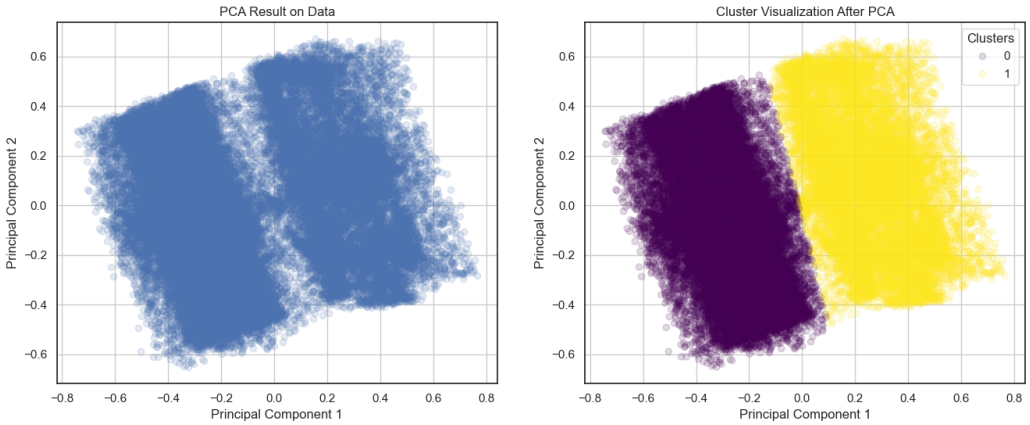

Figura 4: Visualización de los clusters después de aplicar PCA. Los datos están coloreados según los clusters identificados por K-means.

Como se puede apreciar en la Figura 4, hay dos grupos claramente diferenciados. Esto refuerza la decisión de elegir dos clusters como el número óptimo para aplicar el clustering.

6. Conclusiones

En este estudio, se han comparado tres algoritmos de clustering—K-Means, Mezcla de Gaussianas y Clustering Jerárquico Aglomerativo—para determinar el número óptimo de clusters para un conjunto de datos específico. A través de la aplicación de diversas métricas de evaluación interna, como el Coeficiente de Silueta, el Índice de Davies-Bouldin y el Índice Calinski-Harabasz, se ha llegado a la conclusión de que dos clusters son la configuración más adecuada para este conjunto de datos en particular.

La decisión unánime de dos clusters como el número óptimo por parte de dos de los tres métodos y la confirmación por parte del tercer método bajo ciertas métricas sugiere que esta configuración no solo es estadísticamente válida sino también significativa. La visualización de los datos post-reducción dimensional reafirma esta elección, mostrando una distinción clara y comprensible entre los dos grupos.

Es importante reconocer que, más allá de la validez estadística de los modelos de clustering, la interpretación de estos dos grupos debe estar anclada en una comprensión profunda del contexto de los datos. La separación en dos clusters debe tener una justificación o una interpretación que sea relevante para el campo de estudio o el propósito del análisis.

Referencias

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Springer, Enero, 13, 2017.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Second Edition. Springer, Junio 21, 2023.

[3] Julio-Omar Palacio-Niño, Fernando Berzal. *Evaluation Metrics for Unsupervised Learning Algorithms*. Dept. Computer Science and Artificial Intelligence, Universidad de Granada, Granada, Spain. jopalacion@correo.ugr.es, berzal@acm.org.