



TÉCNICA DE ANÁLISIS EXPLORATORIO Y NO SUPERVISADO "CLUSTERING"

*Desarrollado por Adrián Armando Araneda Toro
Y Alex Sebastián Meléndez Suazo*

Julio 2022

Versión 0.2



Medidas de Distancia

- Todos los métodos de clustering tienen un elemento en común, para poder llevar a cabo las agrupaciones necesitan definir “una Medida de Distancia”, y luego cuantificarla, con el objetivo de cuantificar la similitud o diferencia entre las observaciones (puntos en el plano cartesiano).
- El término distancia se emplea entonces dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión.

b. Mundo No Supervisado
¿Qué Buscan?



Medidas de Distancia

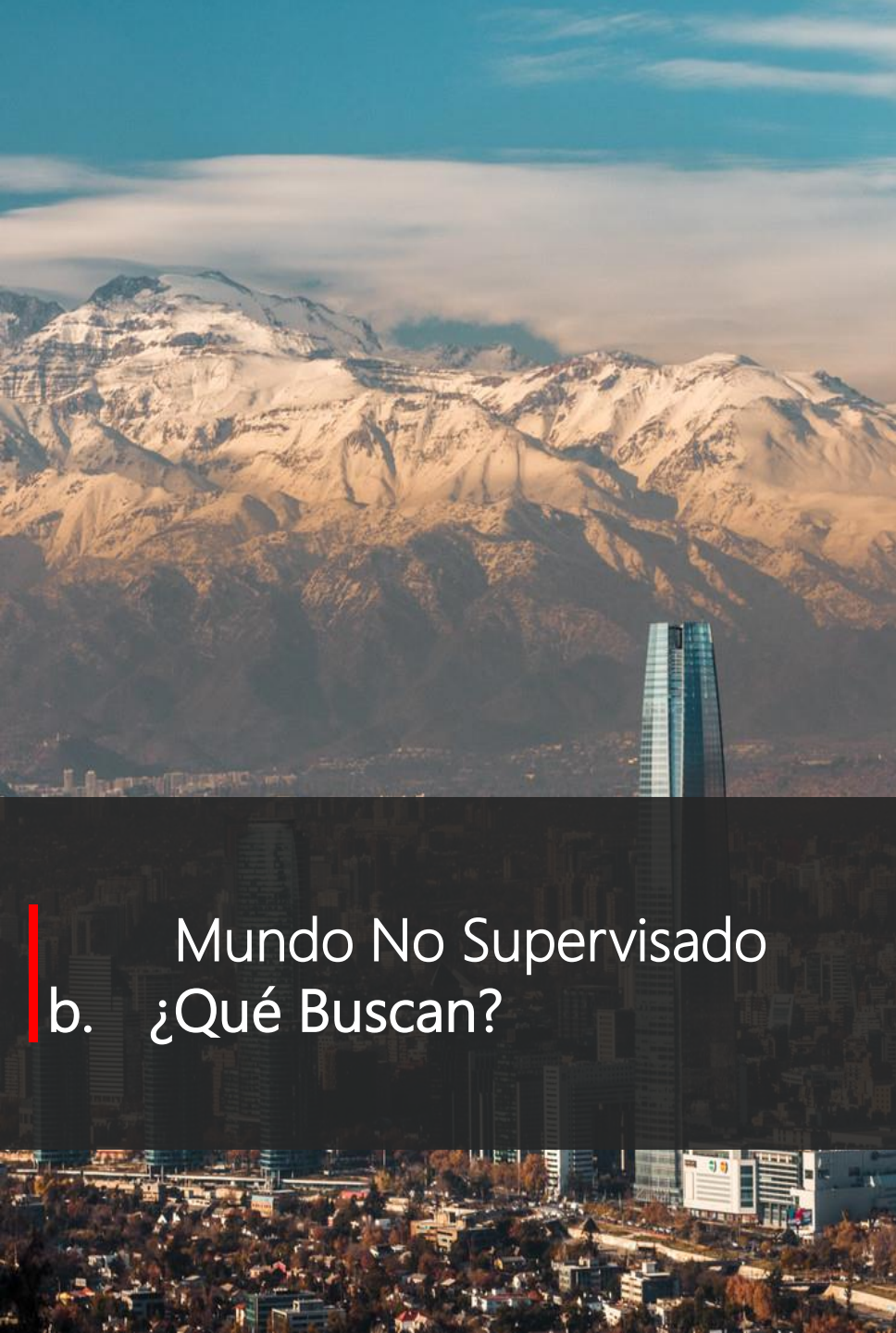
- A. **Distancia Euclidiana:** La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y) , la distancia euclídea entre p y q viene dada por la ecuación:

$$d_{euc}(p, q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

- B. **Distancia de Manhattan:** La distancia de Manhattan, también conocida como taxicab metric, rectilinear distance o L1 distance, define la distancia entre dos puntos p y q como la sumatoria de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por outliers (es más robusta) que la distancia euclídea debido a que no eleva al cuadrado las diferencias.

$$d_{man}(p, q) = \sum_{i=1}^n |p_i - q_i|$$

Mundo No Supervisado
b. ¿Qué Buscan?



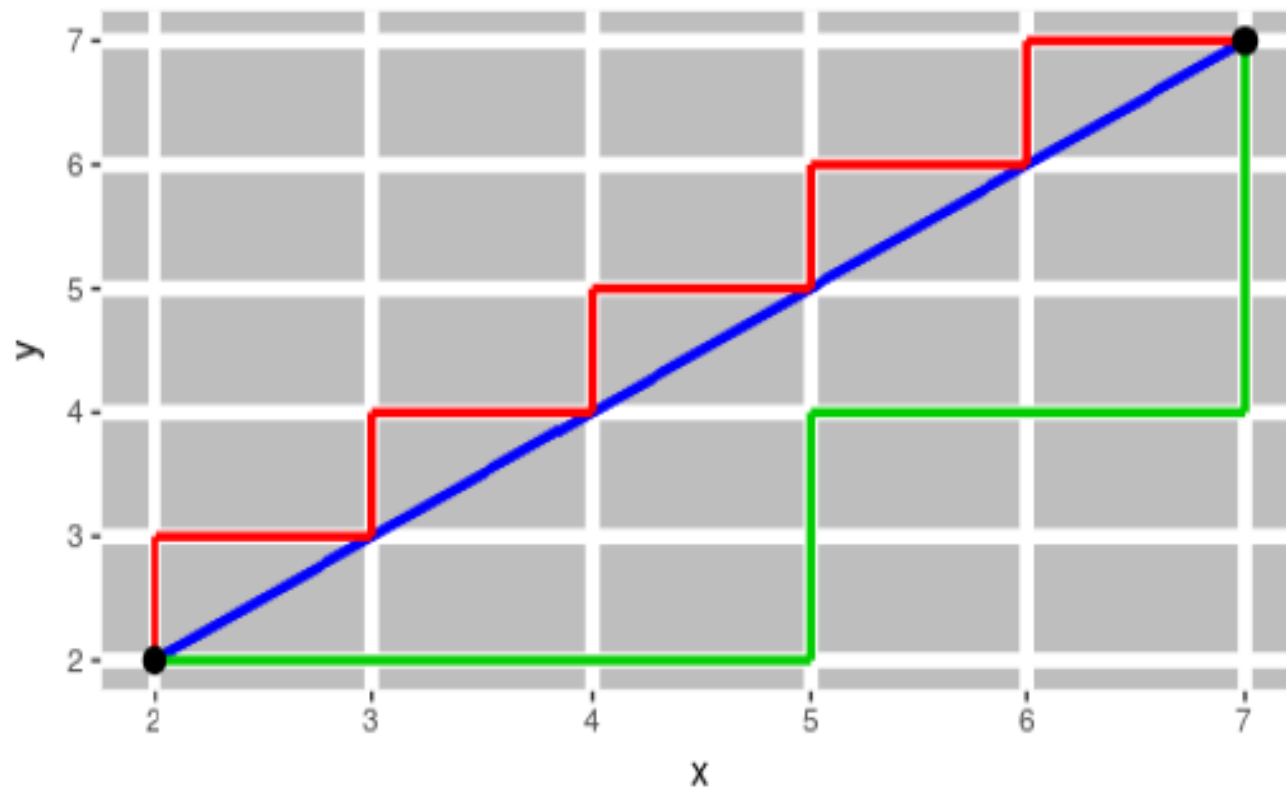
b. Mundo No Supervisado

¿Qué Buscan?



Medidas de Distancia

La siguiente imagen muestra una comparación entre la distancia euclídea (segmento azul) y la distancia de manhattan (segmento rojo y verde) en un espacio bidimensional. Existen múltiples caminos para unir dos puntos con el mismo valor de distancia de manhattan, ya que su valor es igual al desplazamiento total en cada una de las dimensiones.





Calidad de los clusters

Una vez seleccionado el número adecuado de clusters y aplicado el algoritmo de clustering pertinente se tiene que evaluar la calidad de los de los mismos, de lo contrario, podrían derivarse conclusiones de agrupación que no se corresponden con la realidad. Pueden diferenciarse tres tipos de estadísticos empleados con este fin:

- Validación interna de los clusters: Emplean únicamente información interna del proceso de clustering para evaluar la bondad de las agrupaciones generadas. Se trata de un proceso totalmente unsupervised ya que no se incluye ningún tipo de información que no estuviese ya incluida en el clustering.
- Validación externa de los clusters (ground truth): Combinan los resultados del clustering (unsupervised) con información externa (supervised), como puede ser un set de validación en el que se conoce el verdadero grupo al que pertenece cada observación. Permiten evaluar hasta qué punto el clustering es capaz de agrupar correctamente las observaciones. Se emplea principalmente para seleccionar el algoritmo de clustering más adecuado, aunque su uso está limitado a escenarios en los que se dispone de un set de datos de validación.
- Significancia de los clusters: Calculan la probabilidad (p-value) de que los clusters generados se deban únicamente al azar.

b. Mundo No Supervisado
¿Qué Buscan?



Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

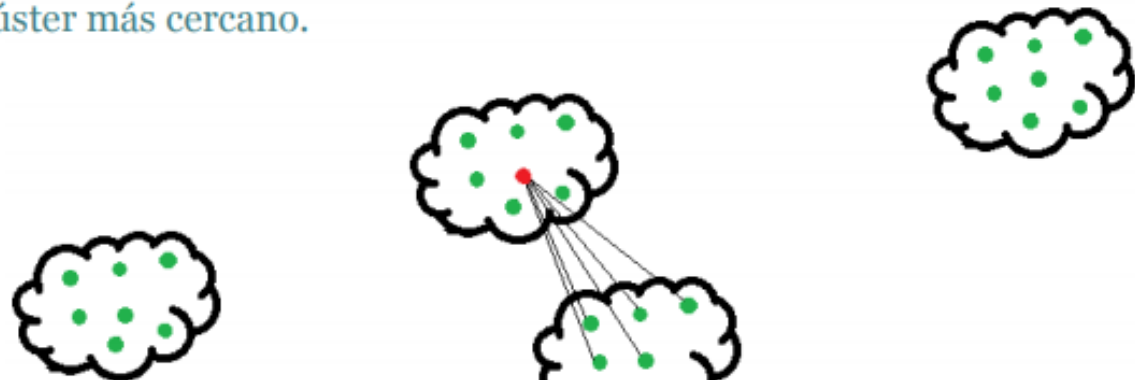
La idea principal detrás del clustering es agrupar las observaciones de forma que sean similares a aquellas que están dentro de un mismo cluster y distintas a las de otros clusters, es decir, que la homogeneidad (también llamada compactness o cohesión) sea lo mayor posible a la vez que lo es la separación entre clusters.

- **Cohesión:** El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.
- **Separación:** Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

- **Cohesión $a(x)$:** distancia promedio de x a todos los demás puntos en el mismo clúster.



- **Separación $b(x)$:** distancia promedio de x a todos los demás puntos en el clúster más cercano.



Mundo No Supervisado
b. ¿Qué Buscan?



Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

Sum of Squared Within (SSW)

Suma de cuadrados para la distancia intra Clusters.

Medida interna especialmente usada para evaluar la **Cohesión** de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster C_i y m_i el centroide del clúster C_i .

Sum of Squared Between (SSB)

Suma de cuadrados para la distancia entre Clusters.

Es una medida de separación utilizada para evaluar la distancia inter-clúster (**Separación**)

$$SSB = \sum_{j=1}^k n_j dist^2(c_j - \bar{x})$$

Siendo k el número de clústeres, n_j el número de elementos en el clúster j, c_j el centroide del clúster j y \bar{x} es la media del data set.

Mundo No Supervisado
b. ¿Qué Buscan?



Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

- Cohesión se mide como **within cluster sum of squares (SSE)**

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

- Separación se mide como **between cluster sum of squares (BSS)**

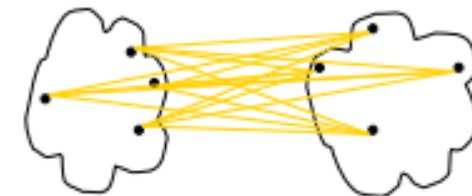
$$BSS = \sum_i |C_i| (m - m_i)^2$$

Medidas internas: Cohesión y separación

- Enfoque basado en grafos de proximidad
- Cohesión: suma de los pesos de todos los arcos en un cluster
- Separación: suma de los pesos entre nodos del cluster y de otros clusters



cohesion



separation

Mundo No Supervisado
b. ¿Qué Buscan?



Calidad de los clusters

Silhouette width

El coeficiente de silueta contrasta la distancia media a elementos en el mismo grupo con la distancia media a elementos en otros grupos. Los objetos con un valor de silueta alto están considerados bien agrupados, los objetos con un valor bajo pueden ser ruido o anomalías. Estos índices trabajan bien con k-means, y es también utilizado para determinar el número óptimo de grupos:

El valor de la silueta es una medida de cuán similar es un objeto a su propia agrupación (cohesión) en comparación con otras agrupaciones (separación). La silueta va de -1 a +1, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con las agrupaciones vecinas. Si la mayoría de los objetos tienen un valor alto, entonces la composición de la agrupación es apropiada.

- Calcula la media de las distancias (llámese a_i) entre la observación i y el resto de observaciones que pertenecen al mismo cluster. Cuanto menor sea a_i mayor la similitud que tiene con el resto de observaciones de su cluster.
- Calcula la distancia promedio entre la observación i y el resto de clusters. Entendiendo por distancia promedio entre i y un determinado cluster C como la media de las distancias entre i y las observaciones del cluster C .
- Identifica como b_i a la menor de las distancias promedio entre i y el resto de clusters, es decir, la distancia al cluster más próximo (neighbouring cluster).
- Calcula el valor de silhouette como:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Mundo No Supervisado
b. ¿Qué Buscan?



Mundo No Supervisado

b. ¿Qué Buscan?

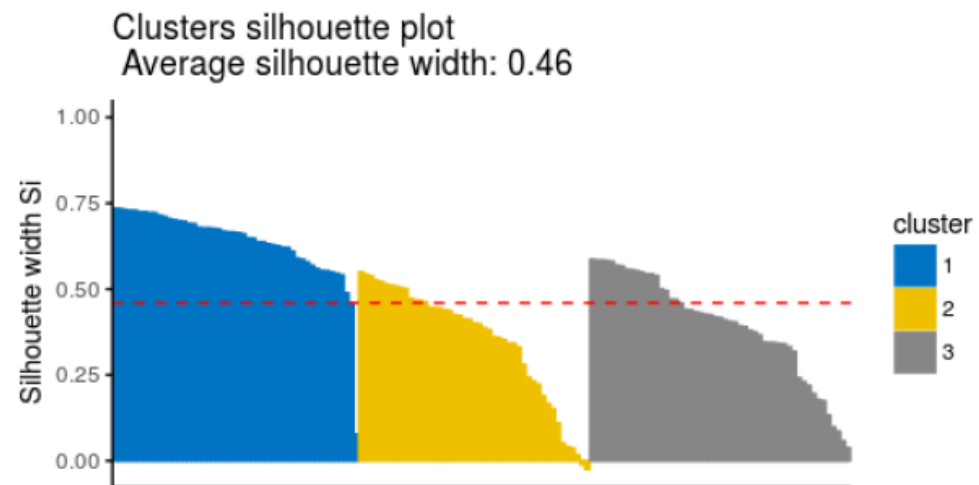


Calidad de los clusters

Su valor puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al cluster correcto. Cuando su valor es próximo a cero significa que la observación se encuentra en un punto intermedio entre dos clusters. Valores negativos apuntan a una posible asignación incorrecta de la observación. Se trata por lo tanto de un método que permite evaluar el resultado del clustering a múltiples niveles:

- La calidad de asignación de cada observación por separado. Permitiendo identificar potenciales asignaciones erróneas (valores negativos de silhouette).
- La calidad de cada cluster a partir del promedio de los índices silhouette de todas las observaciones que lo forman. Si por ejemplo se han introducido demasiados clusters, es muy probable que algunos de ellos tengan un valor promedio mucho menor que el resto.
- La calidad de la estructura de clusters en su conjunto a partir del promedio de todos los índices silhouette.

##	cluster	size	ave.sil.width
## 1	1	50	0.64
## 2	2	47	0.35
## 3	3	53	0.39

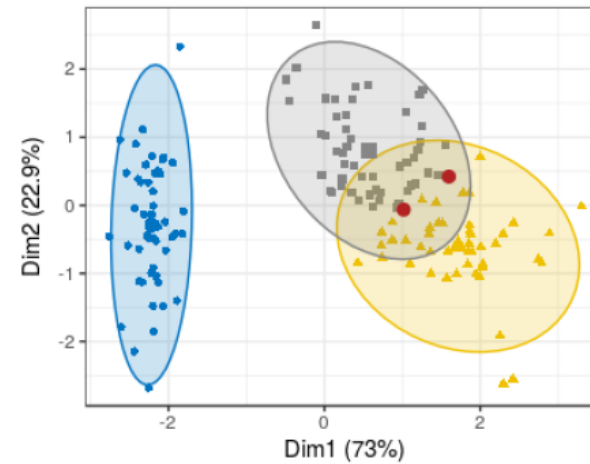




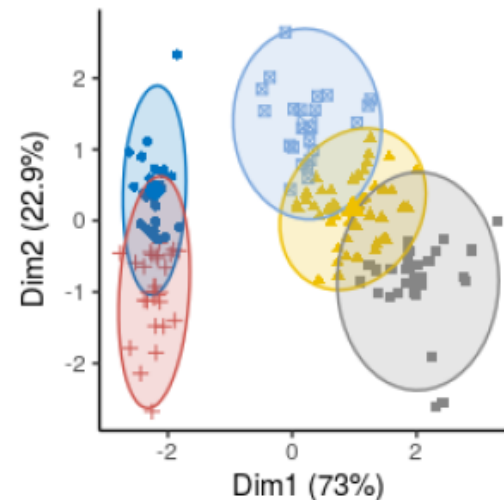
Calidad de los clusters

El cluster número 2 (amarillo) tiene observaciones con valores de silhouette próximos a 0 e incluso negativos, lo que indica que esas observaciones podrían estar mal clasificadas. Viendo la representación gráfica del clustering, cabe esperar que sean observaciones que están situadas en la frontera entre los clusters 2 y 3 ya que solapan.

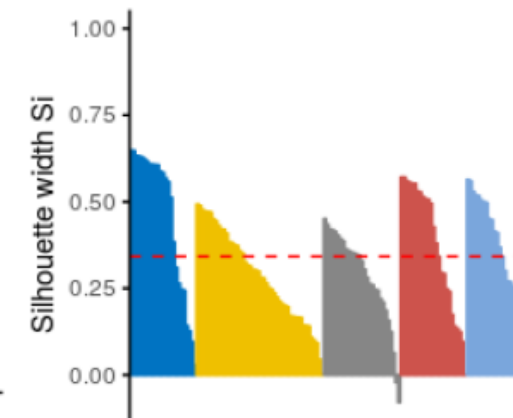
Cluster plot



Cluster plot



Clusters silhouette plot
Average silhouette width:



b. Mundo No Supervisado
¿Qué Buscan?





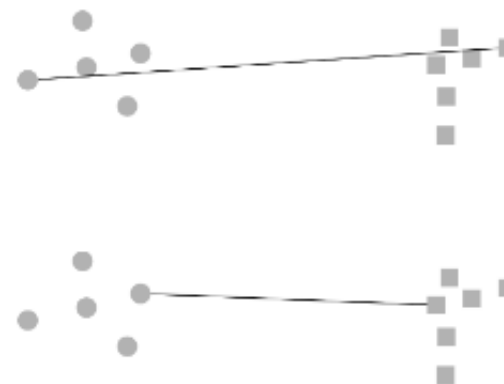
Calidad de los clusters

Índice Dunn

El objetivo de este índice es identificar un conjunto de clústeres que sean compactos, con una varianza pequeña entre los miembros del clúster, y que éstos estén bien separados de los miembros de otros clústeres. Un valor más alto del índice de Dunn indica un mejor rendimiento del algoritmo de clustering.

El índice de Dunn tiene un valor entre cero e infinito, y debe ser lo más alto posible. Por lo tanto, la distancia entre los miembros de un clúster debe ser lo más baja posible, y la distancia entre los clústeres lo más alta posible.

Por ejemplo, en el caso de la distancia entre clústeres puede utilizarse la distancia más corta entre dos puntos de diferentes clústeres, o la distancia más larga, o la distancia entre los centroides. También pueden utilizarse diferentes indicadores de la distancia dentro de un clúster.



$$D = \frac{\text{separacion minima interclusters}}{\text{separacion maxima intracluster}}$$

Mundo No Supervisado
b. ¿Qué Buscan?



Calidad de los clusters

Medidas de estabilidad

Las medidas de estabilidad son un tipo particular de validación interna que cuantifican el grado en que varían los resultados de un clustering como consecuencia de eliminar, de forma iterativa, una columna del set de datos. Todas ellas son relativamente costosas desde el punto de vista computacional ya que requieren repetir el clustering tantas veces como columnas tenga el set de datos. Dentro de esta familia de medidas se encuentran:

- Average proportion of non-overlap (APN): mide la proporción media de observaciones que no se asignan al mismo cluster cuando se elimina una columna del set de datos en comparación a cuando se incluyen todas.
- Average distance (AD): mide la media de las distancias promedio intra-cluster empleando todos los datos y eliminando una columna a la vez.
- Average distance between means (ADM): mide la media de las distancias entre centroides empleando todos los datos y eliminando una columna a la vez.
- Figure of merit (FOM): mide la media de la varianza intra-cluster de la columna eliminada, empleando la estructura del clustering, calcula con las columnas no eliminadas.

Los valores de APN, ADM, y FOM pueden ir desde 0 a 1, siendo valores pequeños un indicativo de alta estabilidad. En el caso de AD ocurre lo mismo pero sus valores pueden ir de 0 hasta infinito.

b. Mundo No Supervisado
¿Qué Buscan?