Clasificación No Supervisada

Marco Ciccalè Baztán

1. Introducción

En este documento de detallará con exhaustividad todo el proceso que se ha seguido para abordar un problema de clasificación no supervisada. Para resolver este problema, se ha decidido experimentar con 3 algoritmos de aprendizaje no supervisado pertenecientes a diferentes familias de *clustering*: *KMeans* (*clustering* particional), *Agglomerative* (*clustering* jerárquico) y *Gaussian Mixture Models* (*clustering* probabilístico). El objetivo de este experimento es evaluar e interpretar cómo se comportan los diferentes algoritmos ante este problema.

2. Descripción del problema

El problema que se aborda en este experimento nace de una base de datos con información de un conjunto de 167 países alrededor del mundo en diferentes continentes. Dado este conjunto de datos, se busca agrupar los países en diferentes grupos para poder evaluar cuáles de ellos pueden necesitar ayuda humanitaria con más urgencia.

3. Metodología

3.1. Conjunto de Datos

El conjunto de datos que se ha usado para este experimento fue extraído de la plataforma *Kaggle* [1]. El tamaño del dataset son 167 filas y 10 columnas, cada fila del conjunto de datos corresponde a un país, y cada columna un valor de las siguientes medidas:

- `country`: El nombre del país (nominal).
- `child_mort`: Cantidad de muertes de niños menores de 5 años por cada 1000 nacimientos (continua).
- `exports`: Exportación de bienes per cápita (continua).
- `health`: Cantidad de recursos invertidos en salud per cápita (continua).
- `imports`: Importación de bienes per cápita (continua).
- `income`: Ingresos per cápita (continua).
- `inflation`: La inflación anual (continua).
- `life_expec`: La media de años de vida estimados (continua).
- `total_fer`: La media de niños que nacen por cada mujer (continua).
- `gdpp`: PIB per cápita (continua).

3.2. Preprocesado de los datos

Dada la naturaleza de los algoritmos de *clustering*, las variables no númericas en la mayoría de los casos no suponen una información relevante para el algoritmo. En este caso, el nombre del país es absolutamente irrelevante para los algoritmos. Por ello, se eliminó del conjunto de datos para los experimentos, aunque posteriormente fue útil para determinar los países que pertenecen a cada grupo.

Por otro lado, en el conjunto de datos, cada variable tiene un rango de valores completamente distinto. Esta característica afecta directamente a los resultados de los experimentos, ya que una variable con una magnitud en milares sería mucho más discriminante que una variable con una magnitud de unidades. Para solventar este problema, se utilizó la técnica `MaxAbsScaler` de **scikit-learn** [2], transformando los valores de cada variable a una misma escala (entre 0 y 1). Pudiendo así mantener la distribución original de cada variable a la vez que igualar el peso de cada variable en los experimentos.

3.3. Pruebas realizadas

3.3.1. Determinar el número de clusters óptimo de cada modelo

Para evaluar el comportamiento de *KMeans* para este problema, se siguieron los siguientes pasos:

1. Entrenar distintos modelos de *KMeans* con [2, 5] *clusters* utilizando la distancia euclídea.
2. Calcular las medias de los coeficientes de *silhouette* para cada uno de los 4 modelos.
3. Seleccionar el modelo con mayor media de coeficiente de *silhouette*.
4. Graficar los coeficientes de *silhouette* de cada cluster.

Para evaluar el comportamiento de *Agglomerative Clustering* para este problema, se siguieron los siguientes pasos:

1. Graficar el dendograma del conjunto de datos utilizando la distancia euclídea y el método *ward*.
2. Elegir el número de *clusters* óptimo dado el dendograma.
3. Seleccionar el modelo con mayor media de coeficiente de *silhouette*.

Para evaluar el comportamiento de *Gaussian Mixtures* para este problema, se siguieron los siguientes pasos:

1. Entrenar distintos modelos de *Gaussian Mixture* con [2, 8] *clusters* y obtener el *BIC* de cada modelo.
2. Graficar el *BIC* de cada modelo y seleccionar el número de *clusters* que minimiza el *BIC*
3. Seleccionar el modelo con mayor media de coeficiente de *silhouette*.

3.3.2. Mostrar los resultados de los modelos óptimos

Los pasos comunes después de haber obtenido el número de *clusters* óptimo para cada algoritmo son los siguientes:

1. Graficar la distribución de las 9 variables del conjunto de datos para cada *cluster*.
2. Graficar un mapa del mundo donde se muestra qué países pertenecen a cada *cluster*.

4. Resultados

4.1. KMeans (Clustering Particional)

Después de seguir los pasos detallados en la sección 3.3.1, el número de *clusters* óptimo (modelo que no tiene valores negativos en sus coeficientes de *silhouette* y que maximiza su media) es 2. La media de los coeficientes de *silhouette* del modelo con 2 *clusters* es `0.329`. La figura 1 refleja los coeficientes de *silhouette* de cada *cluster* y la media de los coeficientes.

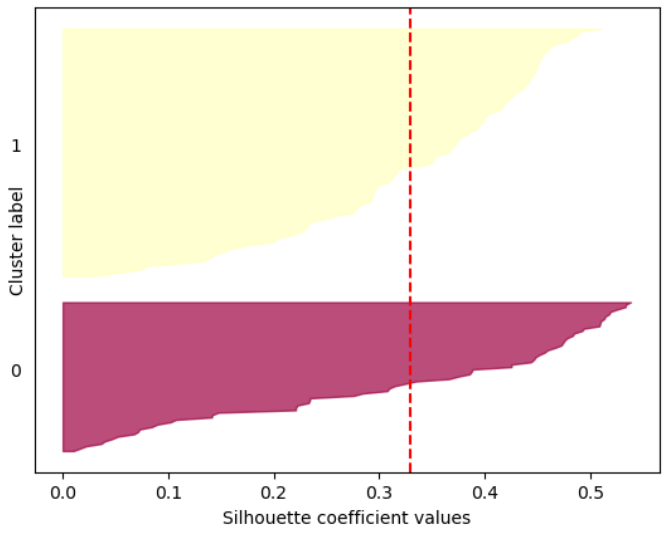

Figura 1: Coeficientes de *silhouette* para el número de *clusters* óptimo

Por otro lado, en la figura 2 se pueden observar las diferencias entre los *clusters* para cada variable. Lo más destacable es la diferencia entre las muertes infantiles, la cantidad de ingresos, la fertilidad y el PIB per cápita. Observando con detalle estas variables, se puede deducir que el *cluster* 0 seguramente englobe países en vías de desarrollo o subdesarrollados, ya que el PIB per cápita se concentra en un rango de valores muy pequeño y cerca de 0, al igual que los ingresos. Además, el número de muertes infantiles y la fertilidad parecen mucho mayores en el *cluster* 0, características que suelen ser atribuídas a países pertenecientes a estos grupos.

La figura 2 representa un mapa del mundo donde se muestra qué países pertenecen a cada cluster.

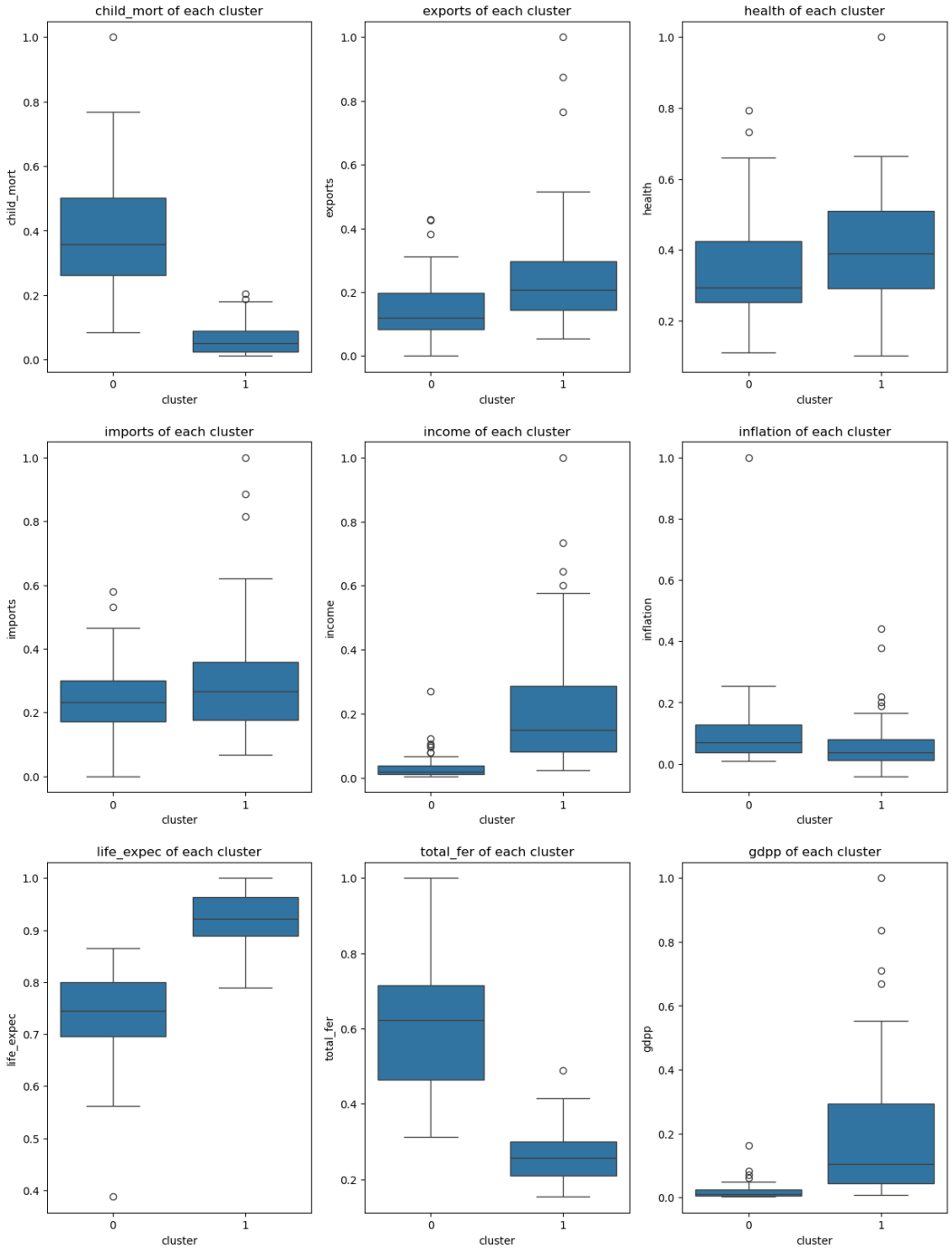

Figura 2: Distribución de las variables para cada *cluster* en *kmeans*

Figura 2: Distribución de las variables para cada *cluster*

4.2. Agglomerative Clustering (Clustering Jerárquico)

Después de seguir los pasos detallados en la sección 3.3.1, se puede identificar en el dendograma de la figura 3 que el número de *clusters* óptimo es 2, ya que es donde se puede aprecuar una distancia en el eje y más notoria entre las dos primeras ramas.

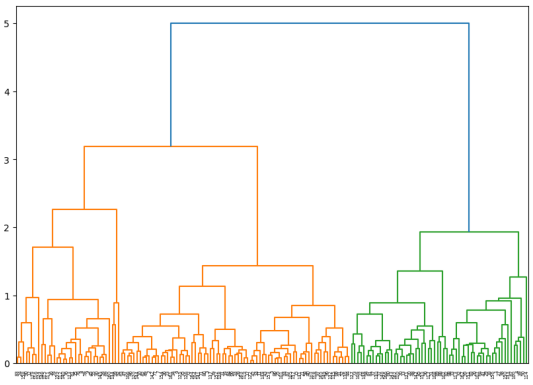

Figura 3: Dendograma del conjunto de datos

Por otro lado, en la figura 7 se puede observar que la diferencias entre los *clusters* para cada variable son prácticamente idénticas al caso anterior de *KMeans* en la sección 4.1, seguramente debido a que se está utilizando la distancia euclídea en ambos algoritmos como medida de distancia entre instancias, por lo que las conclusiones obtenidas con *KMeans* son perfectamente válidas para este caso también. Esta similitud entre los algoritmos también se puede observar en la figura 4, ya que también son prácticamente idénticos.

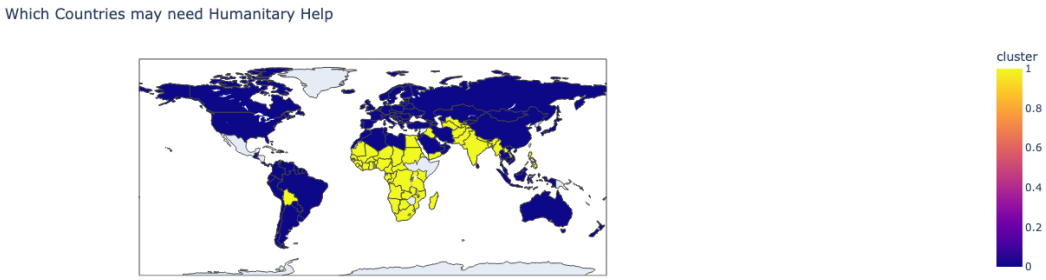

Figura 4: Países pertenecientes a cada *cluster*

4.3. Gaussian Mixture (Clustering Probabilístico)

Después de seguir los pasos detallados en la sección 3.3.1, se puede identificar en la figura 5 que el número de *clusters* que minimiza el *BIC* es 3, siendo así el número óptimo de *clusters*.

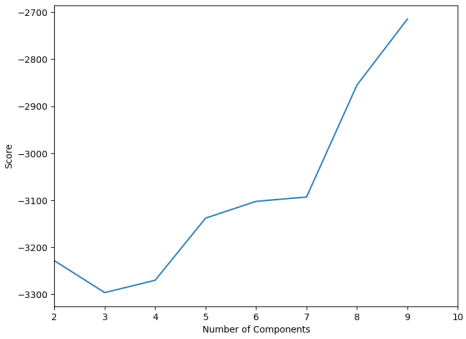

Figura 5: Gráfica que representa los valores *BIC* asociados a cada modelo con diferente número de *clusters*

Por otro lado, en la figura 6 se pueden observar las diferencias entre los *clusters* para cada variable. Lo más destacable de esta figura es que se pueden observar dos *clusters* (0 y 1) que comparten muchos rasgos sociales como la fertilidad más baja, la esperanza de vida más alta y la baja mortalidad infantil. Sin embargo, en rasgos más economicos, pareciera que el *cluster* 0 no está todavía en los niveles en los que está el *cluster* 1, como por ejemplo en el PIB per cápita, los ingresos y el gasto en salud. De estas conclusiones, se puede deducir que el *cluster* 2 representa a los países en desarrollo total, el *cluster* 0 representa a los países que están desarrollados al nivel del *cluster* 1 en temas sociales pero que todavía están atrás en temas económicos, y el *cluster* 1 representa a los países desarrollados.
La figura 6 representa un mapa del mundo donde se muestra qué países pertenecen a cada cluster.

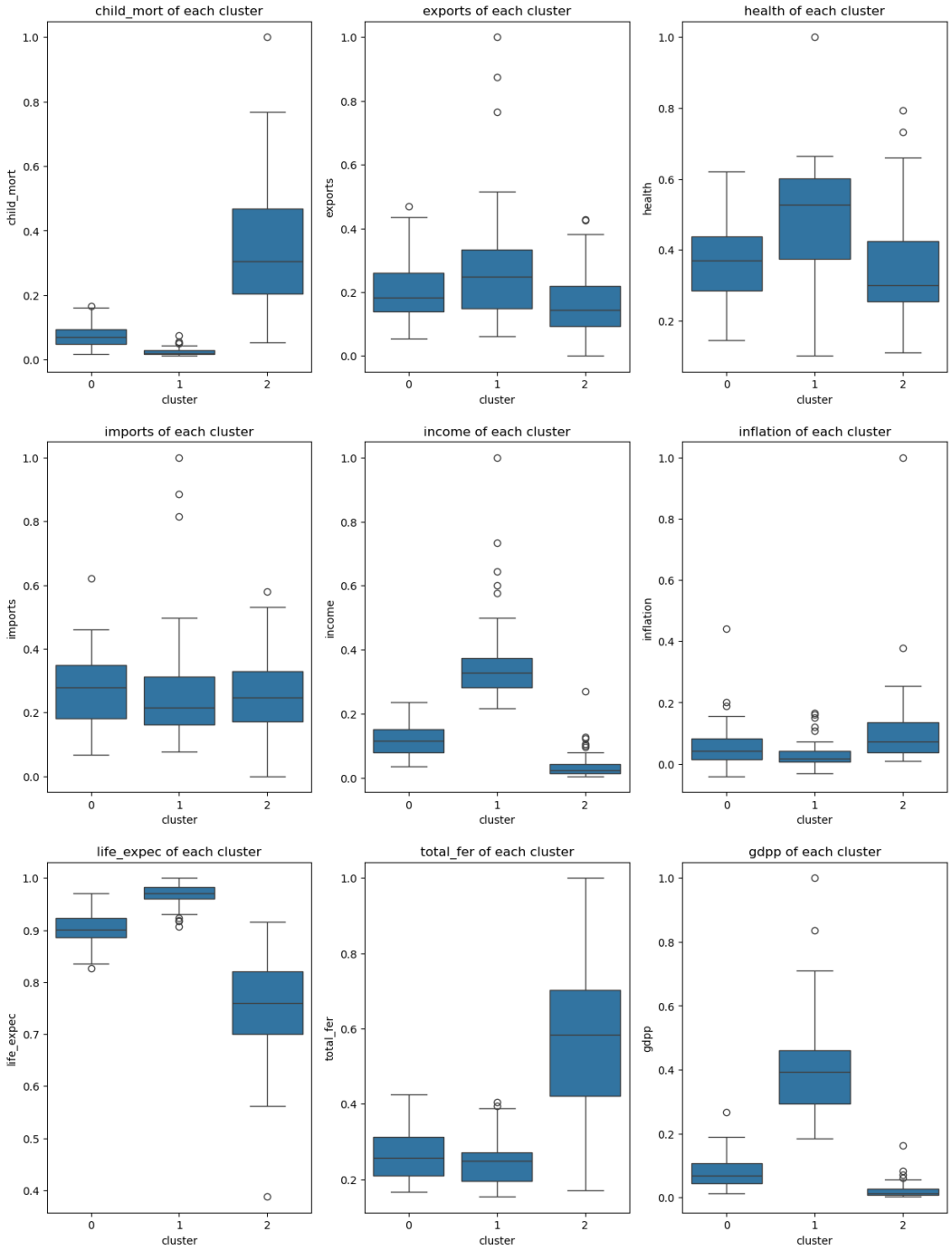

Figura 6: Distribución de las variables para cada *cluster* en *gaussian mixture*

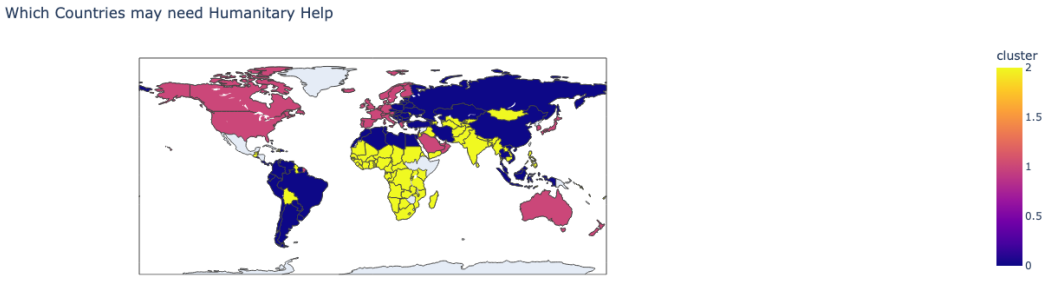

Figura 6: Países pertenecientes a cada *cluster*

5. Conclusiones

Después de haber realizado este experimento, se puede concluir que los algoritmos *KMeans* y *Agglomerative Clustering* a pesar de haber podido distinguir dos posibles grupos, la agrupación no parece muy acertada, ya que identifica países como China y Rusia como parte del *cluster* donde se encuentran países verdaderamente subdesarollados como Sudán. Esto también se ve reflejado formalmente en el bajo coeficiente de *silhouette* que se muestra en la figura 1 y la distancia no tan pronunciada entre las ramas del dendograma de la figura 3.

Por otro lado, el algoritmo *Gaussian Mixture* ha podido diferenciar con mayor precisión 3 *clusters* distintos, donde ha agrupado países desarrollados (españa, alemania...), países en vías de desarrollo (rusia, china...) y países subdesarrollados (nigeria, mali).

Una vez realizado este análisis sobre los resultados de los experimentos, se ha podido comprobar formalmente con diferentes técnicas qué países son los que más necesitan ayuda humanitaria, pudiendo así evaluar cómo distribuir las ayudas más eficientemente e intentando ver qué países pueden necesitarla aunque se piense que son países con un nivel de desarrollo más alto.

Referencias

[1] R. Kokkula, *Unsupervised Learning on Country Data*, 2024. dirección: `https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data` (visitado 12-05-2024).

[2] L. Buitinck, G. Louppe, M. Blondel et al., "API design for machine learning software: experiences from the scikit-learn project," en *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, págs. 108-122.

Anexo

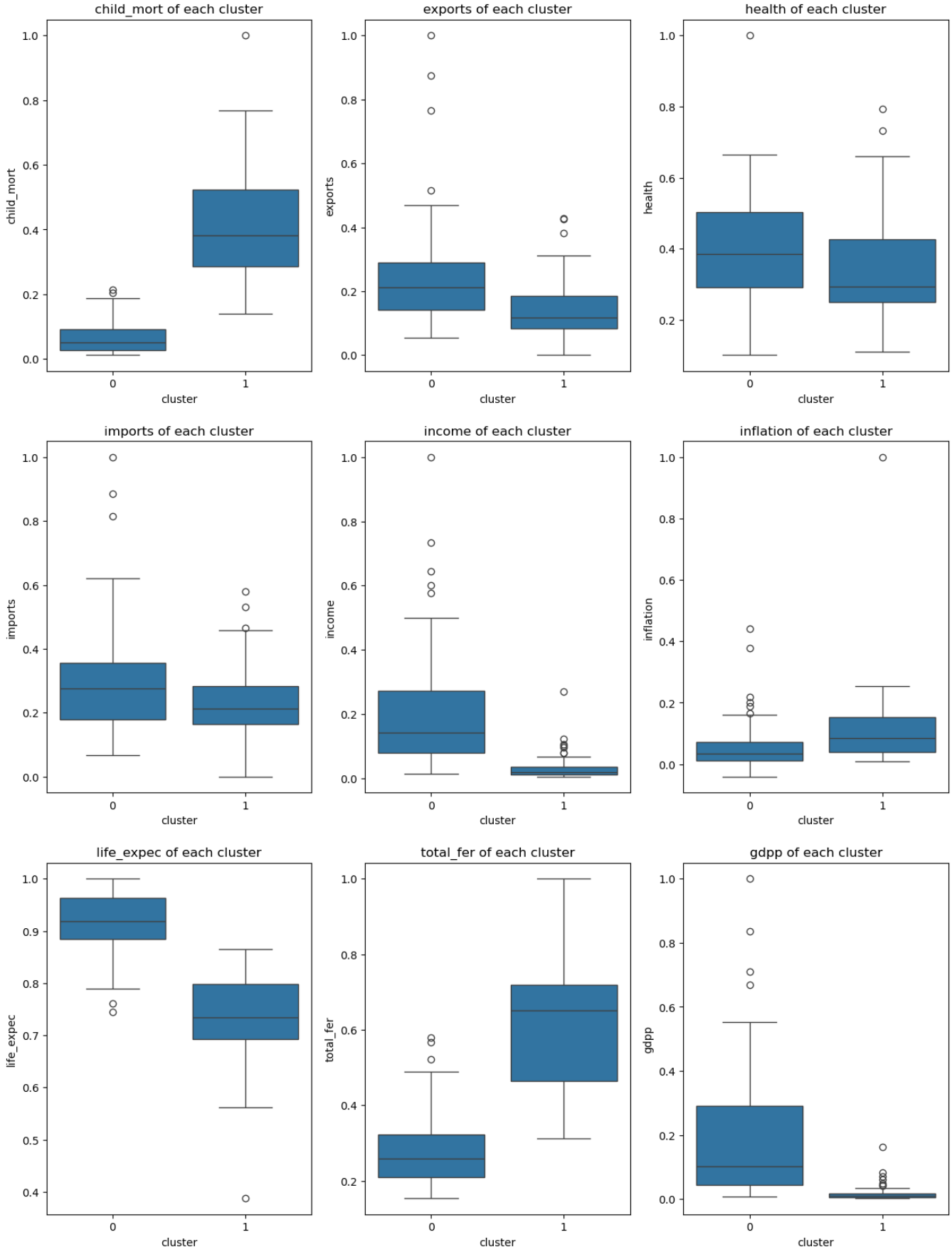

Figura 7: Distribución de las variables para cada *cluster* en *agglomerative clustering*