Análisis de Clustering para la Asignación Estratégica de Ayuda Humanitaria

Un Análisis de Segmentación Basado en Indicadores Socioeconómicos y de Salud

---

Autor:
Pablo García

Fecha:
28 de octubre de 2025

Índice

Resumen Ejecutivo

Este informe detalla el análisis de datos realizado para la ONG -Ayuda Internacional- con el fin de segmentar 167 países en grupos homogéneos y así optimizar la asignación de un fondo de 10 millones de dólares. A través de un análisis exploratorio y la aplicación del algoritmo de clustering K-Means, se identificaron tres clústeres principales que representan distintos niveles de vulnerabilidad socioeconómica y de salud.

El análisis concluye que los países del **Clúster 2** (principalmente en África Subsahariana) son los que presentan una necesidad de ayuda más urgente, caracterizados por una alta mortalidad infantil y un bajo ingreso per cápita. Se recomienda priorizar estos países para la primera fase de asignación de recursos.

1. Introducción

La ONG -Ayuda Internacional- se enfrenta al desafío de distribuir de manera eficaz y estratégica un fondo de 10 millones de dólares para combatir la pobreza y asistir en desastres naturales. Para ello, es fundamental tomar decisiones basadas en datos que permitan identificar a las naciones más necesitadas.

El objetivo de este proyecto es aplicar técnicas de machine learning no supervisado, específicamente clustering, para categorizar países utilizando factores socioeconómicos y de salud. El resultado es una segmentación clara que sirve como guía para la distribución de la ayuda humanitaria.

2. Análisis Exploratorio de Datos (EDA)

El análisis inicial se centró en comprender la estructura del dataset, la distribución de las variables y las relaciones entre ellas.

2.1. Descripción de las Variables

El conjunto de datos contiene 10 características para 167 países. La Tabla 1 detalla cada una de estas variables.

Cuadro 1: Descripción de las Variables del Dataset.

Variable	Descripción
`country`	Nombre del país.
`child_mort`	Muertes de niños menores de 5 años por cada 1000 nacidos.
`exports`	Exportaciones per cápita (como % del PIB per cápita).
`health`	Gasto total en salud per cápita (como % del PIB per cápita).
`imports`	Importaciones per cápita (como % del PIB per cápita).
`income`	Ingreso neto por persona.
`inflation`	Tasa de crecimiento anual del PIB total.
`life_expec`	Esperanza de vida al nacer (años).
`total_fer`	Tasa de fertilidad total (hijos por mujer).
`gdpp`	PIB per cápita.

2.2. Análisis de Correlación

Para identificar relaciones lineales entre las variables, se generó una matriz de correlación (Figura 1). Las observaciones más destacadas son:

- Una **fuerte correlación positiva (0.90)** entre `income` y `gdpp`, lo cual es esperado, ya que un mayor PIB per cápita suele traducirse en un mayor ingreso neto.
- Una **fuerte correlación negativa (-0.89)** entre `child_mort` y `life_expec`. A medida que la mortalidad infantil disminuye, la esperanza de vida aumenta.
- Una **fuerte correlación positiva (0.85)** entre `child_mort` y `total_fer`. Países con mayores tasas de fertilidad tienden a presentar una mayor mortalidad infantil, un patrón común en naciones en desarrollo.

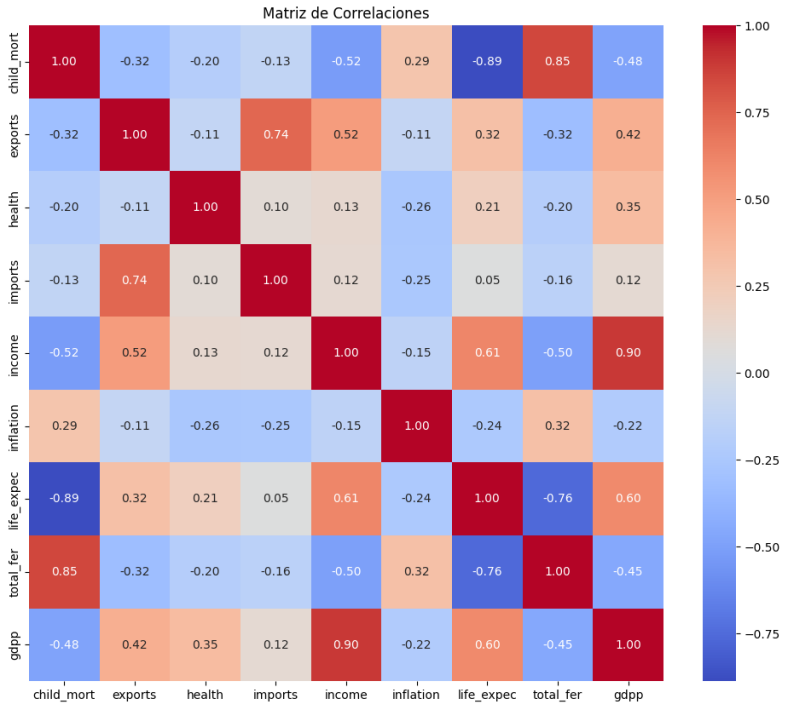
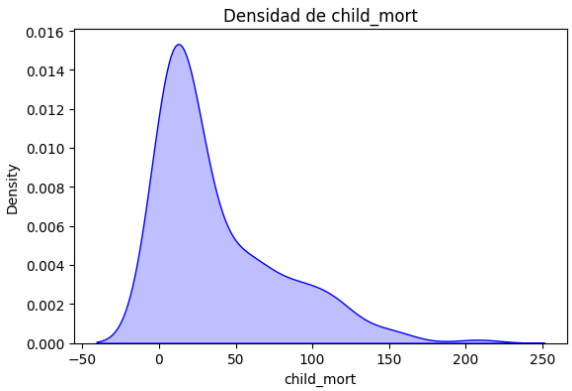


Figura 1: Matriz de correlación de las variables numéricas.

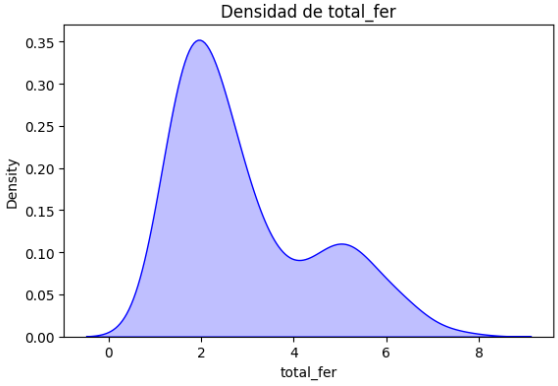
2.3. Distribución de Variables Clave

El análisis de las distribuciones de las variables más relevantes revela patrones importantes:

- **Mortalidad Infantil y Fertilidad**: Ambas distribuciones están sesgadas a la derecha, indicando que la mayoría de los países tienen valores bajos, pero existe una çola larga"de países con tasas muy altas. La fertilidad muestra un comportamiento bimodal, sugiriendo dos grupos de países con patrones demográficos distintos.

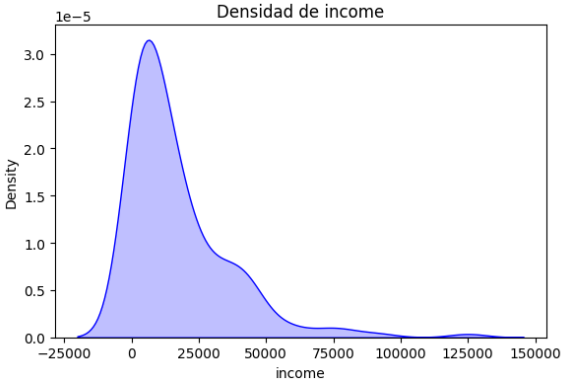


(a) Distribución de Fertilidad

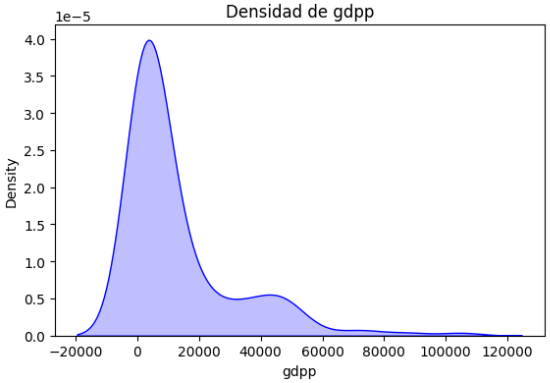


(b) Distribución de Mortalidad Infantil

- **Ingreso y PIB**: Estas variables económicas presentan un sesgo a la derecha extremadamente pronunciado. Esto refleja la gran desigualdad económica a nivel global, con una gran cantidad de países con ingresos bajos y un número reducido de países con ingresos muy elevados.



(a) Distribución de Ingreso



(b) Distribución de PIB

2.4. Ingeniería de Características: Creación de Indicadores

Trabajar directamente con las nueve variables numéricas presenta dos desafíos: la **multicolinealidad** (variables altamente correlacionadas como ‘income‘ y ‘gdpp‘) y la **dificultad de interpretación** de los clústeres resultantes. Un clúster definido en un espacio de 9 dimensiones es poco intuitivo.

Para solucionar esto, se aplicó una estrategia de ingeniería de características, consolidando las variables en tres indicadores temáticos, como se detalla en el Cuadro 2.

Cuadro 2: Agrupación de Variables en Indicadores Agregados.

Indicador Agregado	Variables Constituyentes
Salud	`child_mort`, `health`, `life_expec`, `total_fer`
Comercio	`exports`, `imports`
Finanzas	`income`, `inflation`, `gdpp`

Cada indicador se calculó sumando sus variables componentes, previamente normalizadas por su media. Esta normalización inicial asegura que, dentro de cada grupo, variables con escalas muy diferentes (ej. 'income' en miles y 'inflation' en unidades) contribuyan de forma equitativa al indicador final. Por ejemplo:

$$I_{\text{Finanzas}} = \frac{\text{Income}}{\text{Media(Income)}} + \frac{\text{Inflation}}{\text{Media(Inflation)}} + \frac{\text{GDPP}}{\text{Media(GDPP)}} \tag{1}$$

Esta transformación reduce el problema de 9 a 3 dimensiones, haciendo los clústeres finales directamente interpretables en términos de "Salud", Çomercioz "Finanzas".

2.5. Preprocesamiento Final: Escalado de Indicadores

Aunque los indicadores ya combinan variables normalizadas, sus propias distribuciones y rangos resultantes no son homogéneos. El indicador de *Salud*, por ejemplo, muestra un rango y varianza mayores que el de *Comercio*.

Dado que el algoritmo K-Means se basa en la distancia euclidiana, es sensible a las diferentes escalas de las variables. Una variable con un rango mayor podría dominar el proceso de clustering. Para evitar esto y asegurar que los tres indicadores tengan la misma importancia, se aplicó una **Normalización Min-Max**.

Este método fue elegido sobre la estandarización porque las distribuciones de los indicadores no son perfectamente gaussianas. La normalización reescala cada indicador a un rango común de [0, 1], preservando la forma de la distribución original y garantizando que todos los ejes del espacio de características tengan la misma escala.

2.6. Selección del Modelo y Número Óptimo de Clústeres

Se eligió el algoritmo **K-Means** por su robustez y eficacia en la creación de grupos bien diferenciados, lo cual es ideal para este problema de segmentación.

Para determinar el número óptimo de clústeres (K), se utilizaron tres métodos de validación:

1. **Método del Codo**: Mostró una inflexión clara en K=3 (ver Figura 4).
2. **Puntuación de Silueta**: El coeficiente de silueta alcanzó su valor máximo para K=3, indicando una buena calidad de agrupación (ver Figura 5).
3. **Gap Statistic**: Este método también señaló K=3 como el número óptimo de clústeres al maximizar la "brecha.^entre la dispersión intra-clúster observada y la esperada bajo una distribución nula (ver Figura 6).

La convergencia de los tres métodos proporcionó una fuerte evidencia para seleccionar **K=3**.

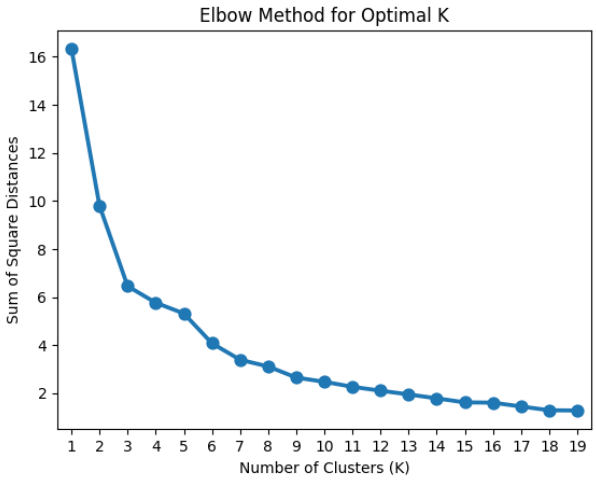


Figura 4: Método del Codo para la selección de K, mostrando la inercia en función del número de clústeres.

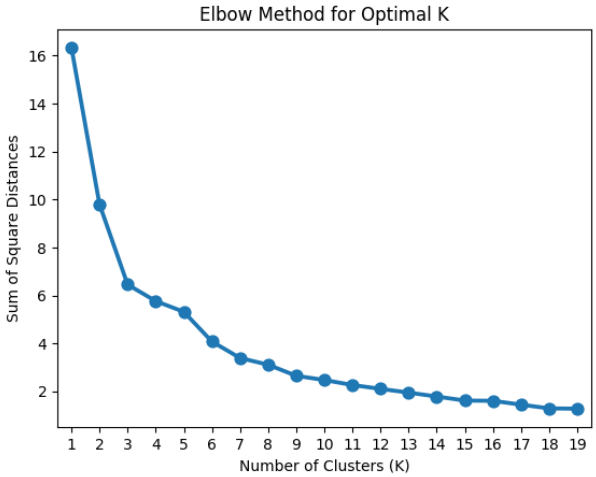


Figura 5: Puntuación de las siluetas para cada K.

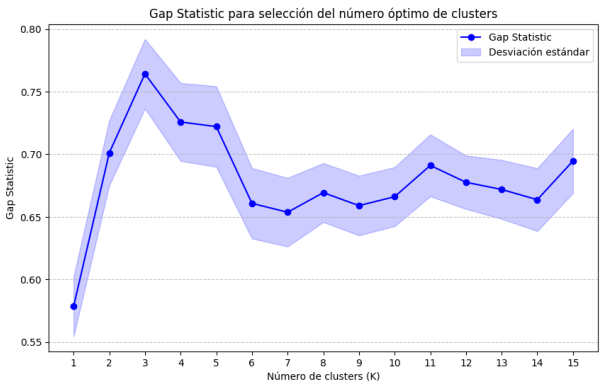


Figura 6: Gap Statisctic para cada K.

3. Análisis de Resultados

El modelo K-Means, entrenado con K=3, ha generado tres grupos de países con perfiles socioeconómicos y de salud claramente diferenciados. Esta sección se dedica a la caracterización detallada de dichos clústeres para poder asignarles un nivel de necesidad de ayuda.

3.1. Caracterización de los Clústeres por Variables Clave

Para interpretar el significado de cada clúster, se analizó la distribución de las dos variables consideradas más representativas por la ONG: el Ingreso Neto (`income`) y la Mortalidad Infantil (`child_mort`). Los diagramas de caja (boxplots) de la Figura 7 comparan estas distribuciones para cada uno de los tres clústeres identificados.

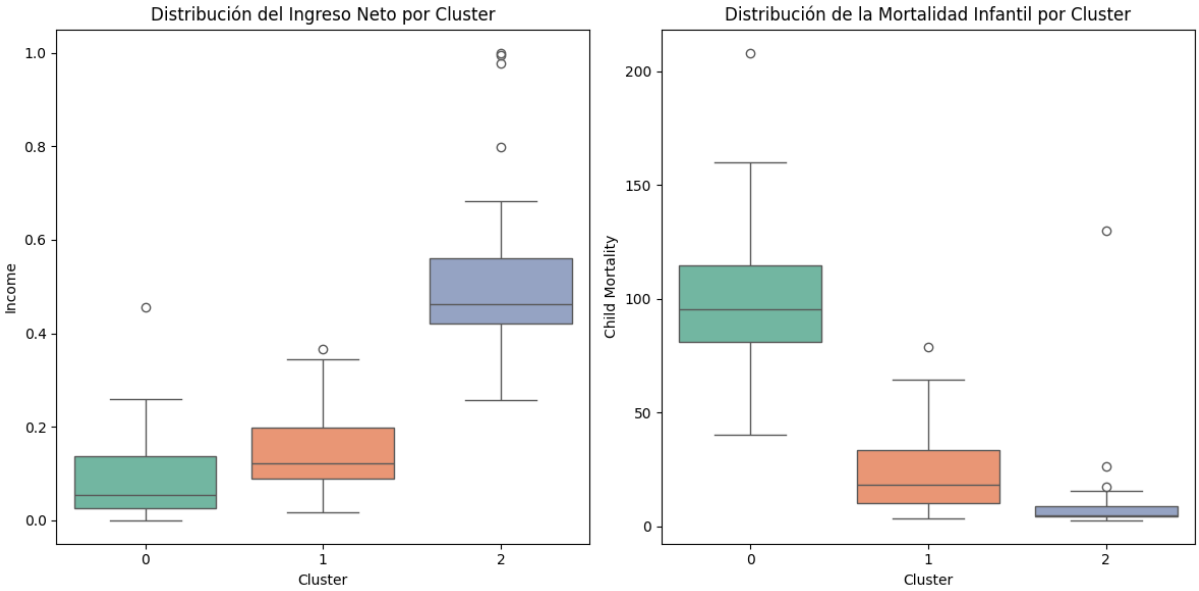


Figura 7: Distribución del Ingreso Neto y la Mortalidad Infantil por Clúster.

3.1.1. Análisis del Ingreso Neto por Clúster

El gráfico de la izquierda revela una clara jerarquía económica entre los grupos:

- **Clúster 0 (Verde):** Este grupo presenta la **mediana de ingresos más baja**. La caja (rango intercuartílico) es muy compacta y se sitúa en la parte inferior del gráfico, indicando que la mayoría de los países en este clúster tienen ingresos consistentemente bajos.
- **Clúster 1 (Naranja):** Muestra un nivel de ingresos **intermedio**. Su mediana es superior a la del Clúster 0, pero significativamente inferior a la del Clúster 2, lo que sugiere una mayor variabilidad económica entre estos países.
- **Clúster 2 (Azul):** Este clúster agrupa a los países con los **ingresos netos más altos**. La mediana es la más elevada y la distribución es muy dispersa, con valores atípicos (*outliers*) que representan a las economías más ricas del mundo.

3.1.2. Análisis de la Mortalidad Infantil por Clúster

El gráfico de la derecha muestra un patrón inverso y aún más pronunciado, reflejando la situación sanitaria de cada grupo:

- **Clúster 0 (Verde):** Este grupo sufre de una **mortalidad infantil drásticamente alta**, con una mediana que se sitúa alrededor de 95 muertes por cada 1000 nacimientos. La amplitud de la caja y la presencia de valores atípicos extremos (superiores a 200) señalan una crisis de salud pública severa.
- **Clúster 1 (Naranja):** La mortalidad infantil es **considerablemente más baja** que en el Clúster 0, con una mediana en torno a 20-30. Representa un nivel de desarrollo sanitario intermedio.
- **Clúster 2 (Azul):** Este clúster se caracteriza por una **mortalidad infantil extremadamente baja**, con una mediana cercana a cero. Esto es indicativo de sistemas de salud muy efectivos y un alto nivel de bienestar general.

3.2. Definición de los Perfiles y Niveles de Ayuda

Al combinar ambos análisis, se pueden definir perfiles claros y accionables para cada clúster, lo que permite asignarles un nivel de ayuda coherente, como se resume en el Cuadro 3.

Cuadro 3: Perfiles de los Clústeres y Nivel de Ayuda Asignado.

Métrica	**Clúster 2**	**Clúster 0**	**Clúster 1**
Ingreso Neto	Muy Alto	Muy Bajo	Intermedio
Mortalidad Infantil	Muy Baja	Muy Alta	Intermedia
Perfil del País	Desarrollado	Subdesarrollado	En vías de desarrollo
Nivel de Ayuda	**0 (No necesita)**	**2 (Urgente)**	**1 (Moderada)**

La visualización geográfica de estos resultados (Figura 8) es una herramienta poderosa para comunicar las conclusiones. El mapa coroplético colorea cada país según el nivel de ayuda asignado, revelando patrones muy marcados.

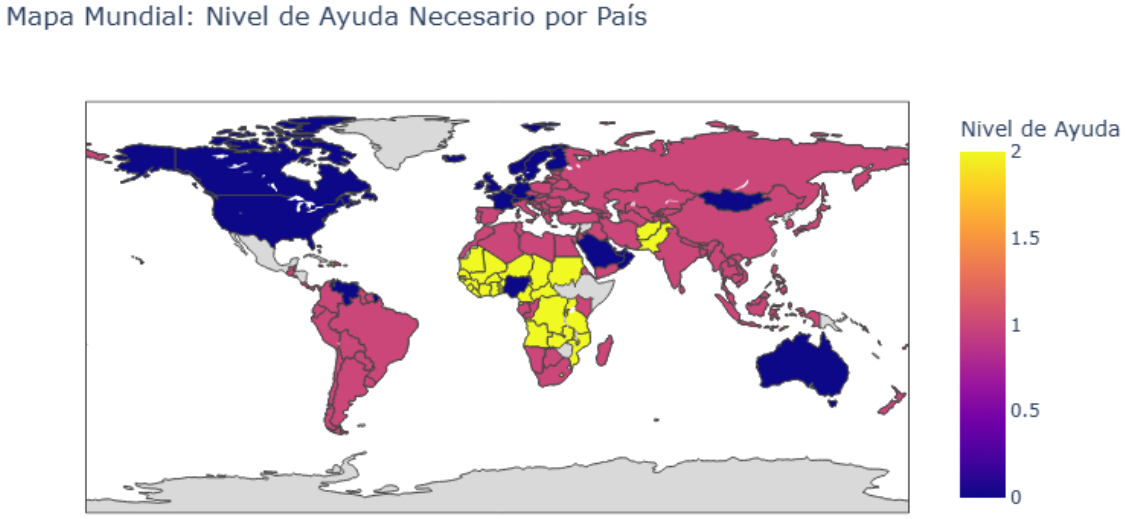


Figura 8: Mapa Mundial del Nivel de Ayuda Necesario por País.

El mapa revela patrones geográficos muy marcados que se alinean con las divisiones socioeconómicas globales conocidas:

3.2.1. Nivel 2 (Amarillo): Países con Necesidad de Ayuda Urgente

Este es el grupo de máxima prioridad, resaltado en amarillo brillante en el mapa. Estos países se concentran de manera abrumadora en:

- **África Subsahariana**: Naciones como Nigeria, Angola, Congo, Chad, Sierra Leona, y Mozambique forman un bloque casi continuo, destacando una crisis regional en términos de salud y economía.
- **Sur de Asia**: Países como Afganistán y Pakistán también caen en esta categoría, indicando condiciones de desarrollo muy precarias.
- **Otras regiones**: Se incluye también Haití en el Caribe, confirmando su conocida situación de vulnerabilidad.

Interpretación: La concentración geográfica de este grupo subraya que la pobreza extrema y las crisis de salud son problemas estructurales que afectan a regiones enteras, las cuales deberían ser el foco principal de las intervenciones de .ªyuda Internacional".

3.2.2. Nivel 1 (Rosa/Magenta): Países con Necesidad de Ayuda Moderada

Este grupo, coloreado en tonos rosados, representa a las naciones en vías de desarrollo. Geográficamente, son un grupo muy diverso y se encuentran en:

- **América del Sur**: La mayoría de los países del continente, como Brasil, Argentina, Colombia y Perú.
- **Norte de África y Sudáfrica**: Países como Egipto, Libia y Sudáfrica.

- **Asia y Medio Oriente**: Incluye a potencias emergentes como China e India, y a naciones como Indonesia, Irán y Arabia Saudita.
- **Europa del Este**: La mayoría de los países de esta región, incluyendo Rusia y Ucrania.

Interpretación: Estos países, aunque no están en una situación tan crítica como los del Nivel 2, todavía tienen poblaciones vulnerables y se beneficiarían de ayuda para consolidar su desarrollo.

3.2.3. Nivel 0 (Azul Oscuro/Morado): Países que No Necesitan Ayuda

Este clúster, representado en los colores más oscuros y fríos de la escala, agrupa a las naciones más desarrolladas del mundo. Se ubican principalmente en:

- **América del Norte**: Estados Unidos y Canadá.
- **Europa Occidental y del Norte**: La mayoría de los países de la Unión Europea y Escandinavia.
- **Oceanía**: Australia y Nueva Zelanda.
- **Asia Oriental**: Economías avanzadas como Japón y Corea del Sur.

Interpretación: Estos países tienen economías robustas y sistemas de salud eficientes. No son el objetivo de los programas de ayuda, pero pueden ser considerados como potenciales donantes o socios estratégicos.

4. Conclusiones y Recomendaciones

4.1. Recomendación Principal

Basado en los resultados, se recomienda a -Ayuda Internacional- **priorizar la asignación de fondos a los países del Clúster 2 (Nivel de Ayuda 2)**. Este grupo, que incluye naciones como Afganistán, Angola, Burundi, Chad, Congo, Haití, Nigeria y Sierra Leona, entre otros, muestra los indicadores más críticos y, por lo tanto, se beneficiaría de manera más significativa de la ayuda humanitaria.

4.2. Pasos Siguientes y Mejoras

Aunque el modelo actual proporciona una guía clara, se podrían considerar las siguientes mejoras en un futuro:

- **Análisis de Componentes Principales (PCA)**: Utilizar PCA sobre las 9 variables originales antes del clustering podría capturar la varianza de los datos de una manera diferente y potencialmente revelar agrupaciones más sutiles.
- **Exploración de otros algoritmos**: Probar modelos como el clustering jerárquico podría ofrecer una perspectiva taxonómica de las relaciones entre países.

El código fuente completo de este análisis está disponible en el siguiente repositorio de GitHub: https://github.com/tu_usuario/tu_repositorio