Documento del proyecto Clustering del Desarrollo de Países

Título del proyecto: Clustering del Desarrollo de Países
Integrantes: Sergio Pinzón, Camilo Arrieta y Juana Sánchez

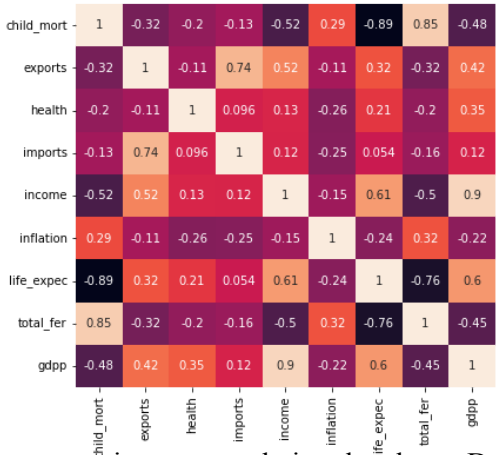
Resumen. En los últimos años se han realizado varios intentos por agrupar países según métricas de desarrollo económico y humano, sin embargo partimos del hecho de que existe una gran diversidad entre los perfiles de desarrollo socio-económico para cada país en el mundo. Esta diversidad es precisamente lo que crea la necesidad de agruparlos de acuerdo con sus características compartidas según el objetivo para así lograr entender las necesidades que cada uno de los grupos tienen en común. En este contexto actuamos desde la intención de una ONG que dispone de cierta cantidad de recursos limitados a distribuir entre los países que más atrás se encuentren en su desarrollo. Siendo así, contamos con 9 variables de desarrollo para cada país, con las cuales utilizamos distintas estrategias de clustering (K-Medias, K-Medoides, Clustering Jerárquico y DBSCAN) para identificar el grupo de países menos privilegiado y hacer una recomendación de inversión para los recursos de la ONG de la forma más eficiente y equitativa. En el ejercicio realizado encontramos que los países africanos son los que presentan menores niveles de desarrollo ya que en las diferentes mediciones presenta los niveles más bajos.

Introducción.
Cada país enfrenta retos y situaciones que afectan su nivel de desarrollo socio-económico y que a su vez tienen implicaciones en la calidad de vida de sus habitantes, sin embargo, a pesar de que cada países cuenta con características distintas, entre ellos hay similitudes y por esta razón en este proyecto la pregunta a responder es ¿Cómo se pueden agrupan los países en cuanto a su nivel general de desarrollo?
Nuestra motivación para responder a esta pregunta reside en el hecho de que existe una gran diversidad entre los perfiles de desarrollo socio-económico para cada país en el mundo. Teniendo en cuenta lo indicado, consideramos importante agruparlos de acuerdo con sus características compartidas para así lograr identificarlos de una mejor manera y llegar a entender las necesidades que cada uno de los grupos tienen en común. En este contexto actuamos desde la intención de una ONG que dispone de cierta cantidad de recursos limitados a distribuir entre los países que más atrás se encuentren en su desarrollo. Siendo así, una agrupación de países se vuelve relevante en la medida en que nos ayude a hacer una recomendación basada en datos sobre cuáles países se encuentran en el grupo menos privilegiado, y así invertir los recursos de la ONG de la forma más eficiente y equitativa. Por otro lado, consideramos que es una pregunta que se puede abordar desde el aprendizaje no supervisado dado que buscamos realizar una clasificación y clusterización de los países, teniendo sólo la disponibilidad de datos crudos para cada país.
El uso de técnicas de clustering es muy común a lo largo de la literatura para abordar diferentes problemas que involucran variables demográficas y socioeconómicas. Dado que el propósito de los algoritmos de clustering es agrupar datos según sus similitudes, es entendible que sea una técnica apropiada para el abordaje de problemas sociales y económicos que requieren de perfilar conjuntos de la población, ya sea para una intervención específica o con propósitos investigativos. En la literatura encontramos ejemplos como el trabajo de Krylovas, Kosareva y Dadelo (2020), quienes buscan, a nivel europeo, agrupar países según distintos índices de desarrollo humano y de actividad física infantil.
Del mismo modo, es importante agrupar países por sus características ya que como lo demostraron Buendía Rice y Edgar Alejandro (2013) existen características de dotación de factores heredados, políticas y de desarrollo de infraestructura que van a generar una ventaja competitiva en algunos países sobre otros y esto se va a ver reflejado directamente en el desarrollo económico.

Materiales y Métodos.
En este caso se utilizaron los datos de países disponibles en https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data?resource=download para obtener información de cada país. Todo el procesamiento, verificaciones y análisis se realizaron en Python

(disponibles en nuestro repositorio de GitHub: https://github.com/Caanar/Proyecto-Final-Clustering-del-Desarrollo-Paises). Se hicieron las verificaciones correspondientes para no tener datos duplicados ni perdidos a la hora de correr los análisis. Todas las variables allí son numéricas y continuas (a excepción de la que representa el nombre de cada país). A continuación se muestra en qué consiste cada variable y se presentan los descriptivos y algunas correlaciones (ir a la sección de resultados para ver el correlograma completo). La variable ***country*** contiene el nombre de los 167 países en la base. ***child_mort*** representa la mortalidad de niños menores a 5 años por cada 1000 nacimientos. Esta variable está fuertemente correlacionada de manera negativa (-0.89) con la variable life_expec y se correlaciona fuertemente (0.85) con total_fer. Esto posiblemente se deba a que los paises con fertilidad más alta son los más pobres. ***exports*** muestra las exportaciones de bienes y servicios per cápita (como % de PIB per cápita). Esta variable tiene una correlación media-alta (0.74) con la variable imports. ***health*** refleja la inversión en salud (como % de PIB per capita). ***imports*** muestra la importación de bienes y servicios (como % de PIB per capita). En promedio las importaciones son mayores que las exportaciones. Si bien no es por mucho, esto puede estar reflejando las tendencias de mercado global actuales. ***income*** contiene los ingresos netos por persona. ***inflation*** es la medida de la tasa de crecimiento de los precios. Es notable que hay países en los que la tasa de inflación tiende a disminuir, si bien en general tiende a aumentar. ***life_expec*** refleja el número promedio de años que un recién nacido vivirá. La expectativa de vida parece estar negativamente correlacionada (-0.76) con la fertilidad total. Esto, al igual que con la mortalidad, probablemente se debe a que los países más pobres tienen las tasas de fertilidad más altas. ***total_fer*** contiene el número de niños que nacerían por mujer si las tasas de edad-fertilidad se mantienen. ***gdpp*** muestra el PIB per cápita.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00
mean	38.27	41.11	6.82	46.89	17144.69	7.78	70.56	2.95	12964.16
std	40.33	27.41	2.75	24.21	19278.07	10.57	8.89	1.51	18328.70
min	2.60	0.11	1.81	0.07	609.00	-4.21	32.10	1.15	231.00
25%	8.25	23.80	4.92	30.20	3355.00	1.81	65.30	1.80	1330.00
50%	19.30	35.00	6.32	43.30	9960.00	5.39	73.10	2.41	4660.00
75%	62.10	51.35	8.60	58.75	22800.00	10.75	76.80	3.88	14050.00
max	208.00	200.00	17.90	174.00	125000.00	104.00	82.80	7.49	105000.00



Como las variables se encontraban en diferentes escalas, consideramos pertinente estandarizar los datos. Dado que algunas de las variables se encuentran altamente correlacionadas, se realizó una reducción de dimensiones por medio de Análisis de Componentes Principales (PCA), buscando conservar al menos el 90% de la varianza. Con los datos resultado del PCA se utilizaron distintos algoritmos de clustering para explorar posibles agrupaciones de países (Clustering K-Medias, K-Medoides y Clustering Jerárquico). En todos los casos, dado que las variables son cuantitativas, utilizamos la distancia euclidiana. Los hiperparámetros en cada uno de estos algoritmos fueron identificados y utilizados según el caso (Ej. en el caso del número de clusters en K-Medias y K-Medoides se utilizó el coeficiente de Silhouette y el método del codo para elegir el número adecuado). Como se puede ver en la carpeta 'Códigos' del repositorio, se corre el PCA con 5 componentes (varianza explicada del 94.53%). Encontramos que para K-Medias el número óptimo de clusters se encuentra alrededor de 6, al igual que para K-Medoides. Con esto se formulan los clusters según K-Medias y se calcularon los descriptivos para cada variable, agrupando por cada clúster. Para realizar el modelo de K-Medoides se utilizaron 6 clusters y se calcularon los descriptivos agrupados. Luego, para correr el algoritmo de Clustering Jerárquico, se realizó un dendrograma que mostró una división bastante profunda, por lo que para tener mayor claridad sobre los clusters a conformar, se probó con la elección de cuatro 4 clusters, si bien en este caso la elección correcta tampoco fue evidente. Por todo lo anterior, al final se decidió conservar los resultados del algoritmo de clustering K-Medias, por razones que se explicarán más adelante.

Resultados y Discusión.

Al implementar el algoritmo de K-Medias, como se dijo anteriormente, se utilizaron los datos resultado del PCA, es decir, que se utilizaron 5 dimensiones para correr el algoritmo. Se realizaron gráficas con los coeficientes de Silhouette y las varianzas intra-cluster y se encontró de manera muy definida que el número óptimo de clusters se encuentra alrededor de 6. Además, se utilizaron diferentes estados aleatorios y se utilizó el que obtuvo la menor varianza intra-cluster para mitigar la asignación aleatoria propia del algoritmo. Siendo así, el algoritmo nos regresa 4 clusters bastante grandes, y 2 clusters que cuentan cada uno con 3 países únicamente (ver tabla más adelante). Por razones de dimensionalidad y de complejidad del problema no fue posible obtener gráficos sencillos, por lo que se utilizaron medidas de correlación entre variables, descriptivos agrupados por clusters, y scatter plots bivariados señalando los clusters para complementar el análisis. Ahora, es claro que los clusters 5 y 6 (ver más abajo) están constituidos por datos relativamente extremos. El cluster 5 abarca Luxemburgo, Malta y Singapur (mundialmente reconocidos por sus altos ingresos y su pequeño tamaño). Por otro lado, el cluster 6 está constituido por Mongolia, Nigeria y Venezuela (representan el extremo del subdesarrollo). Los otros 4 clusters parecen ser más homogéneos en cuanto a los países que contienen como se observa en los gráficos bivariados.

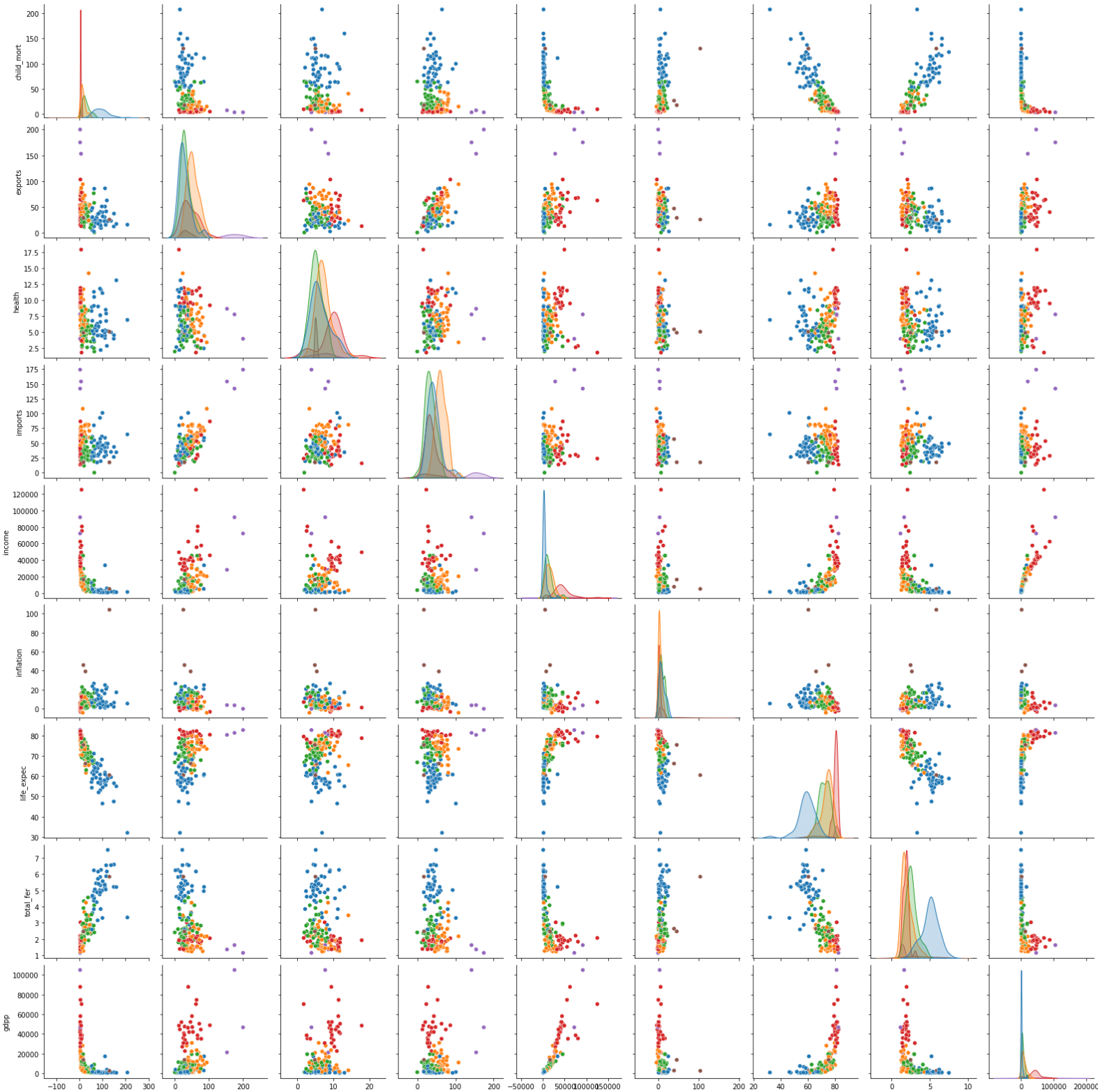
El primer resultado importante a analizar proviene de la correlación entre las medidas de desarrollo que encontramos en la base y que se encuentra relacionada en el siguiente gráfico:

Algunas son bastantes predecibles, este es el caso de exportaciones e importaciones donde por la teoría económica y simple definición de balanza comercial de los países, van a tener una gran correlación. Ahora bien, hay algunas correlaciones interesantes tales como las de ingreso y tasa de mortalidad infantil, una correlación negativa de la cual podemos inferir que a mayor ingreso se tiene una menor tasa de mortalidad infantil. Si bien este análisis es interesante, no podemos fundamentar nuestras recomendaciones en correlaciones ya que sería muy difícil generar políticas para cada país con base en solo esta correlación.

Dicho lo anterior y realizando el algoritmo de K-Medias explicado previamente, a continuación se relacionan los seis (6) cluster con los países contenidos en ellos:

Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	Cluster 6
Afghanistan	Lesotho	Albania	Lebanon	Algeria	Iran	Australia	Qatar	Luxembourg	Mongolia
Angola	Liberia	Antigua and Barbuda	Lithuania	Argentina	Iraq	Austria	Spain	Malta	Nigeria
Benin	Madagascar	Bahamas	Macedonia	Armenia	Jamaica	Belgium	Sweden	Singapore	Venezuela
Botswana	Malawi	Bahrain	Malaysia	Azerbaijan	Kazakhstan	Brunei	Switzerland		
Burkina Faso	Mali	Barbados	Maldives	Bangladesh	Libya	Canada	United Arab Emirates		
Burundi	Mauritania	Belarus	Mauritius	Bolivia	Morocco	Denmark	United Kingdom		
Cameroon	Mozambique	Belize	Micronesia	Brazil	Myanmar	Finland	United States		
Central African Republic	Namibia	Bhutan	Moldova	Cape Verde	Nepal	France			
Chad	Niger	Bosnia and Herzegovina	Montenegro	Chile	Oman	Germany			
Comoros	Pakistan	Bulgaria	Panama	China	Peru	Greece			
Congo, Dem Rep	Rwanda	Cambodia	Paraguay	Colombia	Philippines	Iceland			
Congo, Rep.	Senegal	Costa Rica	Poland	Dominican Republic	Romania	Ireland			
Cote d'Ivoire	Sierra Leone	Croatia	Serbia	Ecuador	Russia	Israel			
Equatorial Guinea	South Africa	Cyprus	Seychelles	Egypt	Samoa	Italy			
Eritrea	Sudan	Czech Republic	Slovak Republic	El Salvador	Saudi Arabia	Japan			
Gambia	Tanzania	Estonia	Slovenia	Gabon	Sri Lanka	Kuwait			
Ghana	Timor-Leste	Fiji	Solomon Islands	Grenada	St. Vincent and the Grenadines	Netherlands			
Guinea	Togo	Georgia	South Korea	Guatemala	Suriname	New Zealand			
Guinea-Bissau	Uganda	Guyana	Thailand	India	Tajikistan	Norway			
Haiti	Yemen	Hungary	Tunisia	Indonesia	Tonga	Portugal			
Kenya	Zambia	Jordan	Ukraine	Turkey					
Kiribati		Kyrgyz Republic	Vietnam	Turkmenistan					
Lao		Latvia		Uruguay					
				Uzbekistan					
				Vanuatu					

Ahora bien, una vez tenemos los países agrupados por sus características, podemos ver comportamientos o tendencias por cada medición del desarrollo económico para de esta manera realizar políticas focalizadas a países con características similares (en términos de desarrollo económico). En las siguientes gráficas precisamente encontramos dicha relación donde cada cluster se va a representar por su respectivo color:


child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp

Para este problema elegimos el algoritmo K-Medias debido a que los coeficientes de silhouette y la varianza intra cluster fueron mucho más estables para este algoritmo. Para K-Medoides fueron mucho más variables estas pruebas, por lo que la elección del número de clusters resultaba mucho menos clara, y asimismo esto reflejaba una menor capacidad por parte de este último algoritmo para clasificar este tipo de datos (consideramos que en este caso los datos extremos deben representarse en todo su peso, y K-Medias se comporta mejor en esto). Por otro lado, el algoritmo de Clustering Jerárquico, mostraba desde el comienzo un dendrograma algo sobrepoblado, por lo que igualmente no se encontró un punto claro para la elección de clusters. Esto último probablemente se debe a que la estructura subyacente de los datos no es jerárquica, por lo que este algoritmo no se desempeña bien en el caso.

En cuanto a las potenciales limitaciones del algoritmo, encontramos una alta sensibilidad a datos extremos. En este caso, dicha característica jugó a nuestro favor, pero en caso de incluir más variables es posible que esto cambie. Como limitación puede decirse que la agrupación de los clusters 5 y 6, si bien es informativa, dificulta de cierta manera las posibles recomendaciones, puesto que si se tienen solo 3 países en el cluster menos desarrollado, y una cantidad mucho más alta en el segundo cluster menos desarrollado, se vuelve complicada la elección.

Para futuros estudios consideramos que la inclusión de otras variables de desarrollo humano y socio-economico podrían complementar e incluso modificar las recomendaciones presentadas anteriormente. Asimismo, se podrían aplicar otras técnicas de clustering como DBSCAN para explorar nuevas posibilidades.

Conclusión.

Se partió con un set de datos de 9 variables sobre el desarrollo socio-económico de cada país. Se normalizaron los datos para hacer comparables las variables y se utilizó el análisis de componentes principales (PCA) para reducir las dimensiones. Con los datos resultado, se utilizaron distintas estrategias de clustering (K-Medias, K-Medoides y Clustering Jerárquico) para explorar posibles agrupaciones y se eligió el clustering K-Medias como el mejor algoritmo para identificar el grupo de países menos privilegiado en este caso y para generar una recomendación de inversión a la ONG. En el ejercicio realizado encontramos que los países africanos son los que presentan menores niveles de desarrollo ya que en las diferentes mediciones presenta los niveles más bajos.

Como podemos evidenciar en las gráficas de indicadores de desarrollo por cluster, hay dos cluster que saltan a la vista, el cluster #1 (color azul) y el cluster #4 (Color rojo). En estos no encontramos outliers sino contrastes ya que el cluster #4 presenta los mejores niveles de desarrollo en los diferentes indicadores o categorías. Por ejemplo, en la tasa de mortalidad infantil vemos que el cluster#1 se encuentra ubicado en los valores más altos mientras que el cluster#4 en los valores más bajos, Del mismo modo vemos que, en términos de ingresos, el cluster#4 cuenta con unos ingresos significativamente más elevados respecto de el cluster#1. Ahora bien, de la tabla de países que conforman cada cluster podemos inferir que, en su mayoría los países menos desarrollados (los del cluster#1) son los países africanos y de esto se pueden desprender dos (2) tipos de sugerencias a aplicar en dichos países, monetarias y sociales.

En primera instancia y respecto de las políticas monetarias evidenciamos que se encuentra una situación bastante retadora en dichos países ya que en términos relativos son los que perciben un ingreso menor y adicional a esto tienen una inflación mayor por lo que el poder adquisitivo de estos países se encuentra en niveles muy bajos. Ahora bien, este no es un problema fácil de solucionar pero si podemos ver las características de los países incluidos en el cluster #4 (El cluster con mejores niveles de desarrollo) y es que países como EEUU cuentan con una fortaleza en las instituciones notable que les permite combatir la inflación mediante la principal herramienta: la tasa de interés. Por lo que se debería buscar fortalecer dichas instituciones centrales en los países africanos y así controlar el problema de la inflación para que se cuente con un poder adquisitivo mayor.

De acuerdo a los datos y modelos realizados, les recomendamos a las ONGs enfocarse o dar prioridad en la distribución de los recursos a los países del cluster#1 que en su mayoría son Africanos para así aumentar la calidad de vida en dichos países e incentivar el desarrollo económico. Del mismo modo y en aras a combatir la desigualdad, se podrían realizar campañas en los países del cluster #4, los países más desarrollados para apoyar a los países africanos.

Bibliografía

Grajales, J. D. S., Guerrero, V. A. B., & Arroyave, D. E. C. (2022). Análisis de clustering para entender la vulnerabilidad poblacional relacionada con el COVID-19. Investigación e Innovación en Ingenierías, 10(2), 7-24.

Krylovas, A., Kosareva, N., & Dadelo, S. (2020). European countries ranking and clustering solution by children's physical activity and human development index using entropy-based methods. Mathematics, 8(10), 1705.

Polyakov, M. V., Shevchenko, G. Y., & Bilozubenko, V. S. (2018). Clustering of countries in global landscape of knowledge economy development. Науковий вісник Полісся, 1(1 (13)), 176-183.

Trushkina, N., & Shiposha, V. (2021). Clustering of the national economy as an effective tool for the sustainable development of european countries. Three Seas Economic Journal, 2(4), 63-70.

Buendía Rice, Edgar Alejandro (2013). El papel de la Ventaja Competitiva en el desarrollo económico de los países. Análisis Económico, XXVIII(69),55-78. ISSN: 0185-3937. Disponible en: https://www.redalyc.org/articulo.oa?id=41331033004