Propuesta Inicial

Repositorio: https://github.com/Caanar/Proyecto-Final-Clustering-del-Desarrollo-Paises

Resumen: El desarrollo socio-económico de los países del mundo es un factor fundamental que incide en la calidad de vida de las personas, afectándolas de manera positiva o negativa dependiendo del nivel de ingresos que tengan, el porcentaje de gasto en salud que disponga el gobierno, el porcentaje de importaciones y exportaciones per cápita, la expectativa de vida, la inflación y el nivel de precios del país, el producto interno bruto (PIB), entre otras. Estas variables que son únicas en cada país hacen que la distribución de los recursos no sea equitativa y por eso las organizaciones sin ánimo de lucro buscan ayudar a aquellos países que no cuentan con un nivel de desarrollo alto para mejorar la calidad de vida de su población, sin embargo, con una simple base de datos no se logra identificar a plenitud cuáles de estos países son candidato de ayuda.
En este sentido, a pesar de que cada país enfrenta retos distintos y presenta diferentes externalidades, el principal objetivo de este proyecto es agrupar a los países que presenten similitudes de acuerdo con sus variables, con el fin de identificar las características de cada uno de los grupos, entender sus necesidades en común y así tener herramientas que ayuden a analizar la información de una manera clara, permitiendo tomar acciones concretas y óptimas que beneficien al conjunto de países que presenten el nivel de desarrollo socio-económico más bajo.

Introducción

Cada país enfrenta una serie de retos y situaciones que afectan su nivel de desarrollo socio-económico y que a su vez tiene implicaciones en sus habitantes, sin embargo, a pesar de que cada uno de los países cuenta con variables 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? ya que como se mencionó anteriormente, 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 en cuenta que en este momento contamos únicamente con una base de datos cruda. Asimismo, a través del aprendizaje no supervisado buscamos disminuir

las dimensiones de la base ya que se cuentan con bastantes variables y con esta cantidad se dificulta realizar la clasificación pertinente.

Revisión preliminar de la literatura

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. Si bien el objetivo varía bastante, podemos encontrar usos muy similares. 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.

Descripción de los datos

Para abordar el problema que presenta el cliente El dataset que utilizaremos (datos originales disponibles en https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data?resource=download ; los datos también pueden ser encontrados en nuestro repositorio de GitHub) cuenta con 9 variables para cada país. Estas, siendo todas numéricas y continuas son: country (nombre del país), child_mort (mortalidad de niños menores a 5 años por cada 1000 nacimientos), exports (exportaciones de bienes y servicios per cápita, recogido como % de GDP per capita), health (inversión en salud, recogido como % de GDP per capita), imports (importación de bienes y servicios, recogido como % de GDP per capita), Income (ingresos netos por persona), Inflation (medida de la tasa de crecimiento anual del GDP total), life_expec (número promedio de años que un recién nacido viviría), total_fer (número de niños que nacerían por mujer si las tasas de edad-fertilidad se mantienen), gdpp (GDP per capita). Para limpiar estos datos se realizarán varias verificaciones. En primer lugar se revisará que no haya datos duplicados, es decir, que cada país aparezca una única vez. Asimismo, se verificará que todos los países cuenten con información en todas las variables, es decir, que no haya datos perdidos. En caso de encontrar datos perdidos procederemos a reemplazarlos con el promedio de la variable respectiva con el objetivo de no perder información existente ni tampoco sesgar el modelo hacia ninguna dirección. Por otro lado, a partir de los descriptivos se revisará que los valores que tome cada variable entren en el rango de lo esperado, y que los outliers, en caso de que existan, sean

realistas y no datos incorrectos. A continuación se presentan los descriptivos para cada una de las 9 variables.

child_mort	
Mean	38.27
Median	19.3
Mode	4.5
Standard Deviation	40.33
Range	205.4
Minimum	2.6
Maximum	208

exports	
Mean	41.11
Median	35
Mode	35
Standard Deviation	27.41
Range	199.89
Minimum	0.11
Maximum	200

health	
Mean	6.82
Median	6.32
Mode	5.2
Standard Deviation	2.75
Range	16.09
Minimum	1.81
Maximum	17.9

imports	
Mean	46.89
Median	43.3
Mode	58.9
Standard Deviation	24.21
Range	173.93
Minimum	0.07
Maximum	174

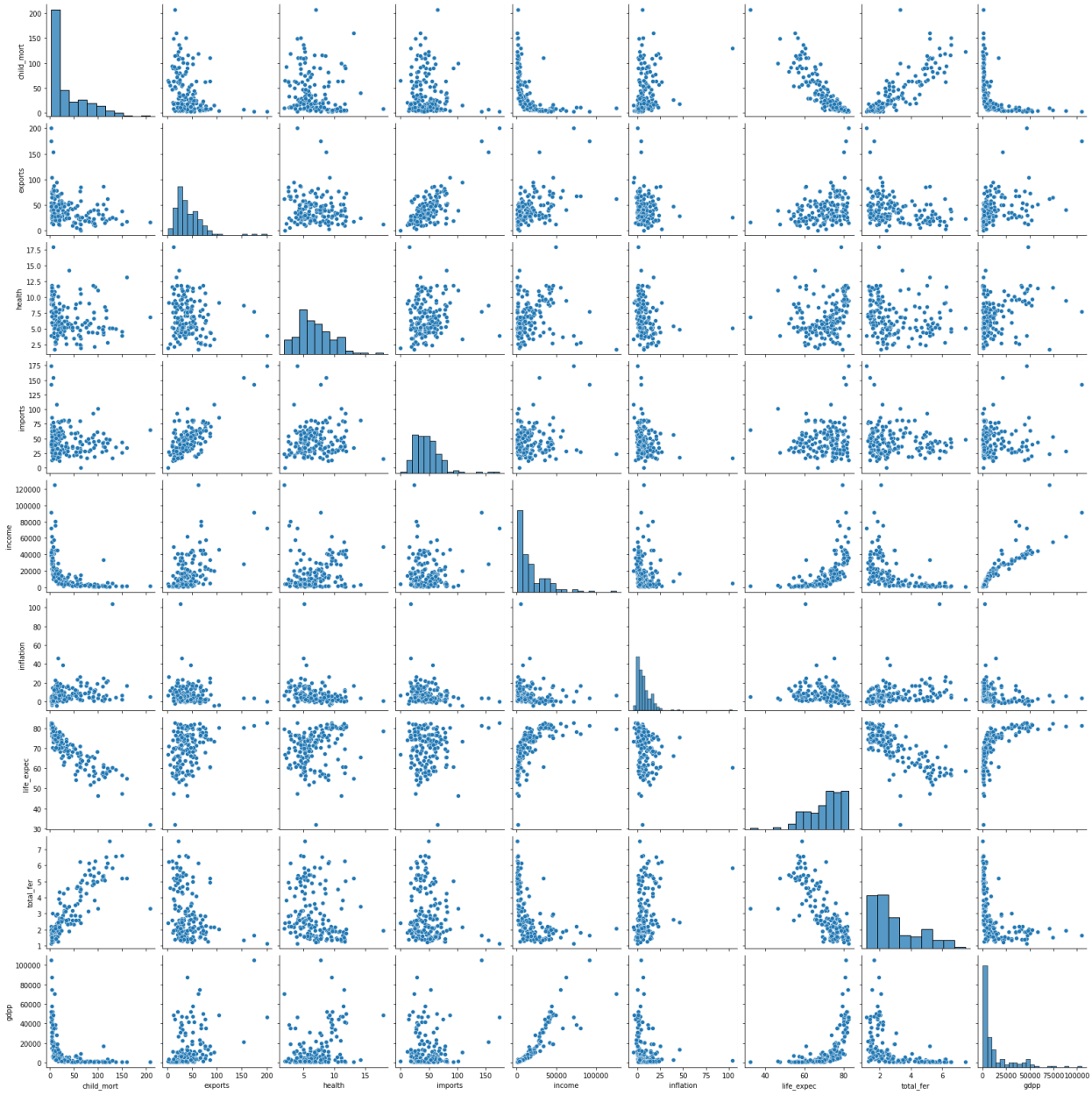
inflation	
Mean	7.78
Median	5.39
Mode	16.6
Standard Deviation	10.57
Range	108.21
Minimum	-4.21
Maximum	104

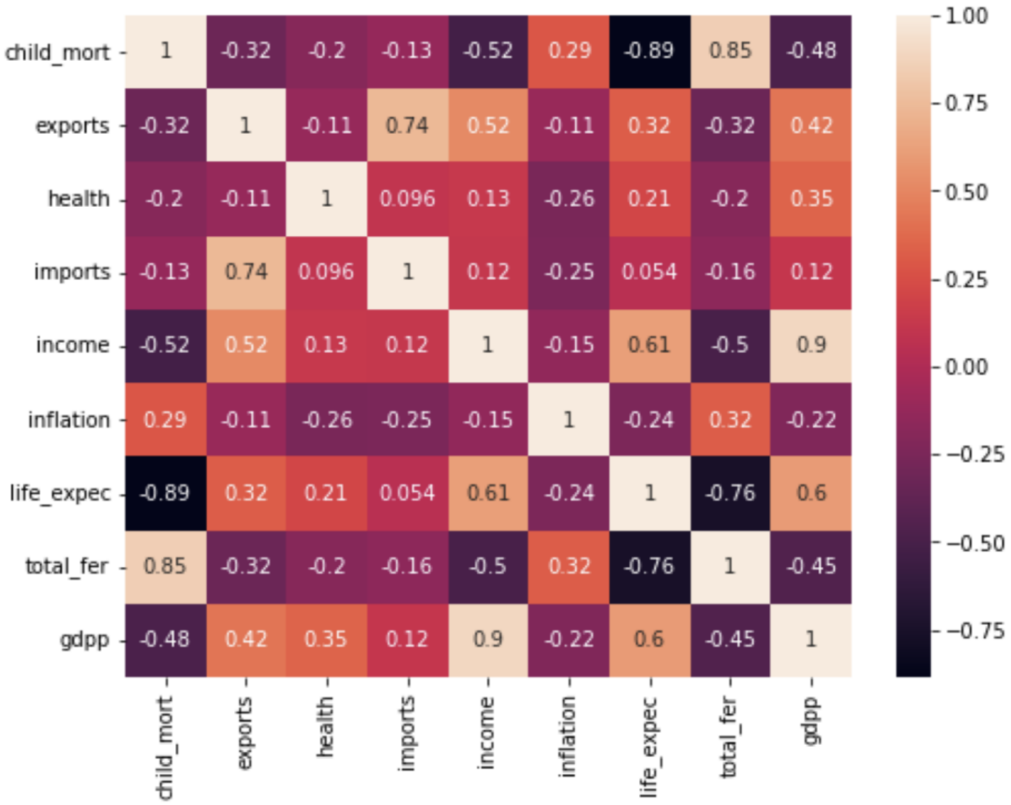
life_expec	
Mean	70.56
Median	73.1
Mode	76.3
Standard Deviation	8.89
Range	50.7
Minimum	32.1
Maximum	82.8

total_fer	
Mean	2.95
Median	2.41
Mode	1.92
Standard Deviation	1.51
Range	6.34
Minimum	1.15
Maximum	7.49

gdpp	
Mean	12964.16
Median	4660
Mode	1310
Standard Deviation	18328.7
Range	104769
Minimum	231
Maximum	105000

Income	
Mean	17144.68862
Median	9960
Mode	41100
Standard Deviation	19278.0677
Range	124391
Minimum	609
Maximum	125000


child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp



Propuesta metodológica

Es importante al menos considerar una reducción de dimensiones debido a que en principio no sabemos cómo cada una de las variables está correlacionada con las demás. Pero a simple vista resalta que variables como health, life expectancy, child mortality y total fertility podrían teóricamente agruparse dentro de una categoría general de "salud". Sin embargo, consideramos prudente primero visualizar las correlaciones y realizar una reducción de dimensión en caso de ser oportuno. Ahora, independientemente de esta decisión, luego se procederá a utilizar distintos algoritmos de clustering para explorar posibles agrupaciones de países. Entre los candidatos pertinentes para abordar este problema consideramos que el uso de distintos algoritmos como K-Medias, K-Medoides y Clustering Jerárquico y DBSCAN podrían brindarnos diferentes perspectivas complementarias a la hora de tomar una decisión sobre cuáles son los países que deberían tener prioridad en la inversión de recursos. Lo ideal en este caso sería comparar las diferentes agrupaciones para explorar si llegan a un consenso. Ahora, en cuanto a la reducción de dimensión utilizaremos el análisis de componentes principales (PCA).

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