**NOTA:** Para el presente ejercicio se utiliza el dataset *Unsupervised Learning on Country Data (Dataset for Kmeans Clustering)* (2020) publicado por Rohan Kokkula en Kaggle. Este conjunto de datos incluye información relativa a factores socioeconómicos y de salud de diversos países, tales como la mortalidad infantil, los ingresos, el gasto en salud, la inflación, entre otros. Los datos se han utilizado como base para el análisis de agrupamiento y categorización de los países en función de su desarrollo.

Es importante tener en cuenta que las fuentes de este dataset no están verificadas y contiene limitaciones inherentes en cuanto a su actualidad, integralidad o precisión. Por lo tanto, se recomienda hacer una evaluación crítica de los resultados obtenidos y considerar que éste se trata de un ejercicio con datos ficticios – con un propósito únicamente académico – sin perjuicio de la validez metodológica y analítica. En una tarea para toma de decisiones “reales”, la calidad y ética relativa a los datos disponibles deben examinarse, contrastarse y validarse a profundidad y exponerse claramente – entre otras consideraciones.

El conjunto de datos (e información relacionada al caso) se ha utilizado en su forma original.

El dataset está disponible en:

<https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data>

**Contexto:**

“La organización no gubernamental HELP International es una entidad dedicada a contribuir al progreso de países en situaciones de vulnerabilidad socioeconómica y a su alivio en afectaciones causadas por desastres naturales.” (Kokkula, 2020)

“Recientemente, la ONG ha podido recaudar $10 millones [diremos que de USD] y el CEO quiere utilizarlos de manera estratégica y efectiva. Por lo tanto, se necesita identificar los países que requieran ayuda con más urgencia.” (ibid.).

**Objeto de consultoría:**

Se me ha encomendado la tarea de clasificar los países según factores socioeconómicos y de salud que influyen en su desarrollo para orientar las decisiones del CEO de HELP International sobre a qué países debe dirigirse la ayuda humanitaria con mayor urgencia.

Con base en este trabajo, la organización HELP International podrá tomar decisiones estratégicas sobre la distribución de ayuda humanitaria, maximizando el impacto de los recursos disponibles y contribuir a mejorar las condiciones de vida en los países más necesitados.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Descripción** | **Unidad de medida (original)\*** |
| **1.** | **country** | *Nombre del país.* | Caracteres |
| **2.** | **child\_mort** | *Muertes de niños menores de 5 años.* | Valores por cada 1,000 nacidos vivos |
| **3.** | **exports** | *Exportaciones de bienes y servicios per cápita.* | Dado como porcentaje del PIB per cápita (gdpp) |
| **4.** | **health** | *Gasto total en salud per cápita.* | Dado como porcentaje del PIB per cápita (gdpp) |
| **5.** | **imports** | *Importaciones de bienes y servicios per cápita.* | Dado como porcentaje del PIB per cápita (gdpp) |
| **6.** | **income** | *Ingreso neto por persona. Normalmente, calculado como:*  *Ingreso Bruto - (Impuestos + Otras deducciones obligatorias).*  *Es un reflejo del poder adquisitivo. Nota: No se especifica temporalidad (supongo anual).* | Valor real |
| **7.** | **inflation** | *Medición de la tasa de crecimiento anual del PIB total.* | Porcentaje |
| **8.** | **life\_expec** | *Promedio de años que viviría un recién nacido si los patrones actuales de mortalidad permanecieran constantes.* | Valor real |
| **9.** | **total\_fer** | *Número de hijos que nacerían por mujer si las tasas actuales de fertilidad por edad permanecieran constantes* | Valor real |
| **10.** | **gdpp** | *PIB per cápita. Mide el valor total de bienes y servicios producidos por un país en un año determinado, dividido por su población total. Da una idea de la producción económica general y el nivel de vida de un país.* | Calculado como el PIB total dividido por la población total |
| \*Nota: no se especifica moneda de cambio en variables económicas, supongo US$. | | | |

**Bibliografía**

EDA(Elgabry, 2019)

(Greenacre et al., 2022; StatisticalSongs, 2011)

(StatQuest with Josh Starmer, 2018)

(IBM Technology, 2024)

(StatQuest with Josh Starmer, 2022)

(Sefidian, 2022)

(stataiml, 2024)

(Kassambara, 2018)

(Gonzalez, 2020)

(Kashnitsky on Korolev, 2019)

Elgabry, O. (2019, March 2). *The Ultimate Guide to Data Cleaning*. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Gonzalez, L. (2020, September 8). Algoritmo Agrupamiento Jerárquico—Teoría. *Aprende IA*. https://aprendeia.com/algoritmo-agrupamiento-jerarquico-teoria/

Greenacre, M., Groenen, P. J. F., Hastie, T., D’Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 1–21. https://doi.org/10.1038/s43586-022-00184-w

IBM Technology (Director). (2024, June 10). *Principal Component Analysis (PCA) Explained: Simplify Complex Data for Machine Learning* [Video recording]. https://www.youtube.com/watch?v=ZgyY3JuGQY8

Kashnitsky, Y. (2019). *Topic 7. Unsupervised learning: PCA and clustering*. https://kaggle.com/code/kashnitsky/topic-7-unsupervised-learning-pca-and-clustering

Kassambara, A. (2018). Agglomerative Hierarchical Clustering. *Datanovia*. https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/

Kokkula, R. (2020). *Unsupervised Learning on Country Data*. https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

Sefidian, A. M. (2022, December 18). How to determine epsilon and MinPts parameters of DBSCAN clustering. *Amir Masoud Sefidian - Sefidian Academy*. https://www.sefidian.com/2022/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/

stataiml. (2024, April 4). *How to Choose Optimal Hyperparameters for DBSCAN*. Stataiml. https://stataiml.com/posts/how\_to\_set\_dbscan\_paramter/

StatisticalSongs (Director). (2011, June 26). *It had to be U - the SVD song* [Video recording]. https://www.youtube.com/watch?v=JEYLfIVvR9I

StatQuest with Josh Starmer (Director). (2018, April 2). *StatQuest: Principal Component Analysis (PCA), Step-by-Step* [Video recording]. https://www.youtube.com/watch?v=FgakZw6K1QQ

StatQuest with Josh Starmer (Director). (2022, January 10). *Clustering with DBSCAN, Clearly Explained!!!* [Video recording]. https://www.youtube.com/watch?v=RDZUdRSDOok