**Acta de Constitución**

**del Proyecto**

***HELP***

***13 de septiembre de 2024***

DOCUMENTO EN DESARROLLO

***Luis Felipe Villota Macías***

**Tabla de contenido**

*\*La presente acta de constitución ha sido hecha en base a* [*www.pmoinformatica.com*](http://www.pmoinformatica.com)

**Información general del proyecto**

| Proyecto | HELP. Un proyecto de clasificación de países para la asignación de ayuda humanitaria |
| --- | --- |
| Fecha de preparación | 13 de septiembre 2024 |
| Cliente | HELP International |
| Gerente de Proyecto | Luis Felipe Villota Macías |

**Contexto breve**

“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.” 1

“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.

**Objetivos específicos**

* Realizar un análisis exploratorio de los datos (EDA) para comprender la distribución de los indicadores clave.
* Aplicar un análisis de componentes principales (PCA).
* Implementar algoritmos de clustering para agrupar países según patrones similares de desarrollo.
* Comparar los modelos
* Identificar los países que pertenecen a los grupos más necesitados y presentar estos hallazgos a HELP International.
* Proporcionar recomendaciones específicas basadas en el análisis para guiar la inversión estratégica de los $10 millones recaudados.

**Datos**

**Consideraciones sobre el material empírico.** 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>

| Variable | Descripción | Unidad o expresión de medida (original) |
| --- | --- | --- |
| **country** | Nombre del país. | Caracteres |
| **child\_mort** | Muertes de niños menores de 5 años. | Valores por cada 1000 nacidos vivos |
| **exports** | Exportaciones de bienes y servicios per cápita. | Dado como porcentaje del PIB per cápita (gdpp) |
| **health** | Gasto total en salud per cápita. | Dado como porcentaje del PIB per cápita (gdpp) |
| **imports** | Importaciones de bienes y servicios per cápita. | Dado como porcentaje del PIB per cápita (gdpp) |
| **Income** | Ingreso neto por persona. | Valor real |
| **Inflation** | Medición de la tasa de crecimiento anual del PIB total. | Valor real |
| **life\_expec** | Promedio de años que viviría un recién nacido si los patrones actuales de mortalidad permanecieran constantes. | Valor real |
| **total\_fer** | Número de hijos que nacerían por mujer si las tasas actuales de fertilidad por edad permanecieran constantes | Valor real |
| **gdpp** | PIB per cápita. | Calculado como el PIB total dividido por la población total |

**Metodología**

### 1. Preparación y Exploración de Datos

* Limpieza de Datos: Identificación y tratamiento de valores atípicos, datos faltantes y inconsistencias.
* Análisis Exploratorio de Datos (EDA): Visualización de distribuciones, correlaciones y tendencias en los datos.
* Ingeniería de Características: Creación de nuevas variables relevantes basadas en el conocimiento del dominio y los insights del EDA.

### 2. Técnicas de Reducción de Dimensionalidad

* Análisis de Componentes Principales (PCA): Para reducir la dimensionalidad de los datos mientras se preserva la máxima varianza.
* t-SNE (t-Distributed Stochastic Neighbor Embedding): Para visualizar estructuras de alta dimensionalidad en un espacio bidimensional o tridimensional.

### 3. Algoritmos de Clustering

* K-Means: Para agrupar observaciones en K clusters basados en la similaridad de sus características.
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Para identificar clusters de forma arbitraria y detectar anomalías.
* Hierarchical Clustering: Para crear una jerarquía de clusters, útil para explorar diferentes niveles de granularidad en los datos.

### 4. Técnicas de Asociación

* Reglas de Asociación: Para descubrir relaciones interesantes entre variables, especialmente útiles en el análisis de factores socioeconómicos interrelacionados.

### 5. Validación y Evaluación

* Métricas de Cohesión y Separación: Como el coeficiente de silueta y la inercia para evaluar la calidad de los clusters.
* Validación Cruzada: Para asegurar la robustez de los modelos y evitar el sobreajuste.
* Visualización de Resultados: Uso de técnicas de visualización avanzadas para interpretar y comunicar los hallazgos.

### 6. Interpretación y Extracción de Insights

* Caracterización de Clusters: Análisis detallado de las características distintivas de cada cluster identificado.
* Análisis de Importancia de Variables: Identificación de las variables más influyentes en la formación de patrones y grupos.
* Narración de Datos: Construcción de narrativas coherentes basadas en los patrones descubiertos, relacionándolos con el contexto socioeconómico de Pakistán.

## Herramientas y Tecnologías

* Lenguaje de Programación: Python
* Bibliotecas Principales: scikit-learn, pandas, numpy, matplotlib, seaborn
* IDE: Jupyter Notebooks en Google Colab para facilitar la colaboración y reproducibilidad

## Consideraciones Éticas

* Protección de la privacidad y confidencialidad de los datos sensibles.
* Evaluación y mitigación de posibles sesgos en los datos y algoritmos.
* Interpretación responsable de los resultados, considerando el contexto sociocultural.

## Resultados Esperados

* Identificación de patrones socioeconómicos no evidentes en los datos crudos.
* Insights accionables para informar políticas y estrategias.

**Limitaciones**

**Bibliografía**

EDA2

1. Kokkula R. Unsupervised Learning on Country Data. 2020. Accessed September 5, 2024. https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

2. Elgabry O. The Ultimate Guide to Data Cleaning. Medium. March 2, 2019. Accessed September 6, 2024. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4