**INFORME MODELO MACHINE LEARNING**

**Empresa Farmacéutica**



**PROYECTO**: Análisis de la Esperanza de Vida al Nacer para Factibilidad de lanzamiento de multivitamínico a nivel Global

**1. Introducción**

En el epicentro de la toma de decisiones informadas se encuentra el análisis exploratorio de datos (EDA), una práctica esencial que arroja luz sobre los matices y patrones ocultos dentro de conjuntos de datos vastos y complejos. Este informe se sumerge en un EDA exhaustivo, basado en ocho conjuntos de datos extraídos del Banco Mundial, cada uno de los cuales abarca factores cruciales que potencialmente inciden en la esperanza de vida.

Los datos, procedentes de una fuente confiable como el Banco Mundial, se han seleccionado meticulosamente para ofrecer una visión holística de los 44 países en estudio, distribuidos a lo largo de los 5 continentes. Estos conjuntos abarcan aspectos que van desde indicadores económicos y sociales hasta datos demográficos, permitiéndonos desentrañar las complejidades que subyacen a la variabilidad en la esperanza de vida.

A través de este EDA, no solo buscamos describir y visualizar la distribución de variables clave, sino también identificar relaciones y tendencias significativas entre los diversos factores. Al comprender la interconexión entre variables como el Producto Interno Bruto, la educación, la nutrición y otros indicadores relevantes, se espera revelar insights valiosos que orientarán la toma de decisiones estratégicas.

**Data sets**

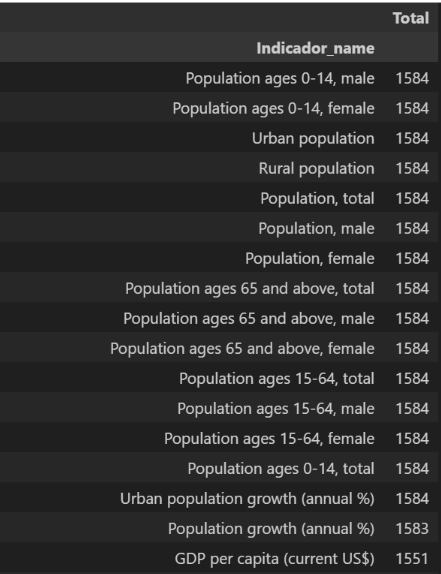
Los conjuntos de datos fueron descargados directamente desde la página oficial del Banco Mundial, accediendo específicamente al siguiente enlace:

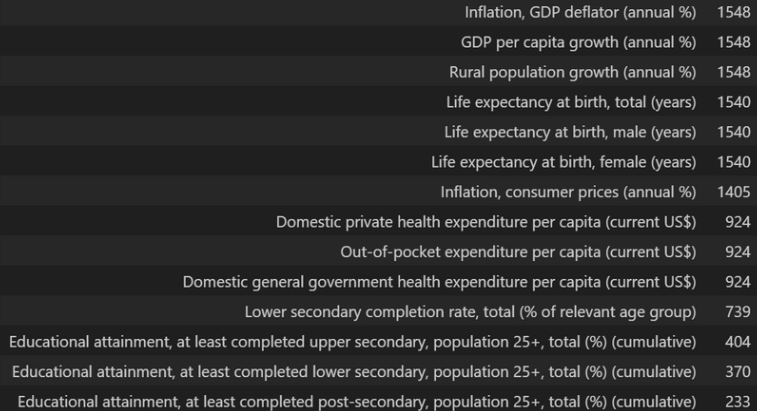
<https://databank.worldbank.org/databases>

Este enlace proporciona acceso a una amplia variedad de series que se pueden filtrar según las necesidades específicas. Para el presente proyecto, se ha optado por descargar los factores más alineados con los objetivos establecidos, los cuales incluyen las siguientes categorías:

* Longevidad
* Población
* Economía
* Educación
* Inversión en Salud

**Selección de indicadores para el EDA**

Después de hacer una exploración de los datos nos damos cuenta que algunos de los indicadores seleccionados presentan una cantidad considerable de valores nulos, por lo que decidimos quedarnos con aquellos que presentan por lo menos 1000 registros de un total de 1584 no nulos.

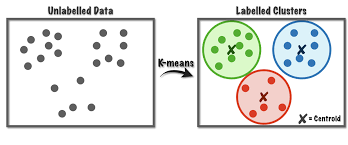


**Selección de variables para el modelo**

* País: Variable categórica que identifica la ubicación geográfica del registro.
* Año: Variable temporal que indica el año al que corresponde cada registro.
* GDP per capita (current US$): El Producto Interno Bruto (PIB) per cápita en dólares estadounidenses corrientes. Es un indicador económico que muestra la producción promedio por individuo en el país.
* Inflation, GDP deflator (annual %): Indica el cambio porcentual en el nivel general de precios en la economía.
* Inflation, consumer prices (annual %): Representa el cambio porcentual en el índice de precios al consumidor.
* Life expectancy at birth, total (years): La esperanza de vida al nacer en años.
* Population growth (annual %): La tasa de crecimiento poblacional anual, que indica la variación porcentual en la población en un año dado.
* Ratio\_population ages 65 and above: La proporción de la población total que tiene 65 años o más.
* Ratio\_urban population: La proporción de la población que vive en áreas urbanas en comparación con la población total.
* Urban population growth (annual %): La tasa de crecimiento anual de la población urbana.

**Elección del modelo**

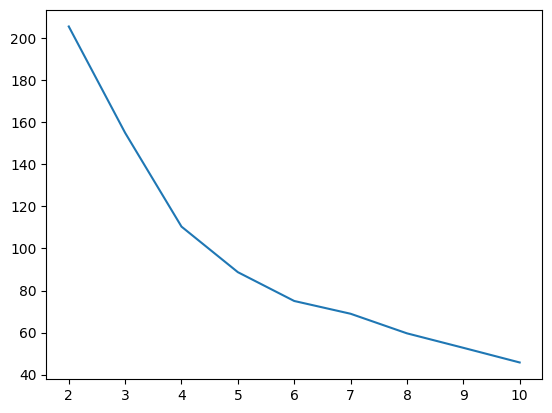
Se ha decidido el uso de un modelo no supervisado con el objetivo de segmentar los países en base a sus indicadores y poder recomendar los que se encuentren en una mejor situación.

El algoritmo elegido es K-means.  


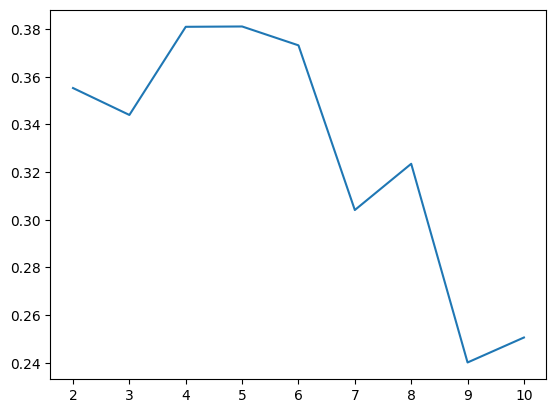
Se ha decidido tomar como referencia el promedio de los valores registrados en los últimos 5 años (2017-2021).



**Selección del número óptimo de clusters**

**Método del codo**

En el eje x tenemos el número de clusters y en el eje y la suma de las distancias al cuadrado. El "codo" en el gráfico es el punto donde la disminución en la suma de las distancias se hace más gradual después de haber descendido abruptamente. Este punto sugiere el número óptimo de clusters, ya que agregar más clusters a partir de ese punto no proporcionará una mejora significativa en la reducción de la suma de distancias.

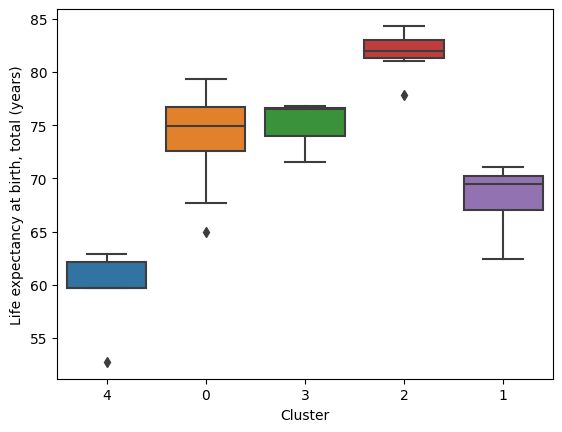
**Coeficiente de silhoutte**

El coeficiente de silueta oscila entre -1 y 1. Un valor alto indica que la muestra está bien ubicada en su cluster y que está separada de los clusters vecinos, mientras que un valor bajo indica que la muestra podría estar asignada al cluster incorrecto o que está muy cerca de los límites del cluster.

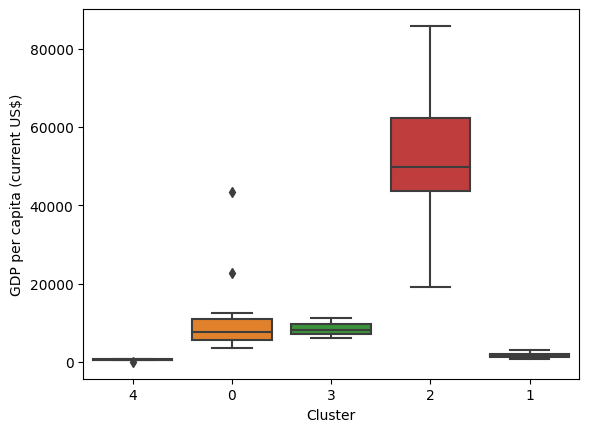
En base a los gráficos anteriores el número de clusters elegido es 5

**Visualizaciones de los diferentes indicadores**

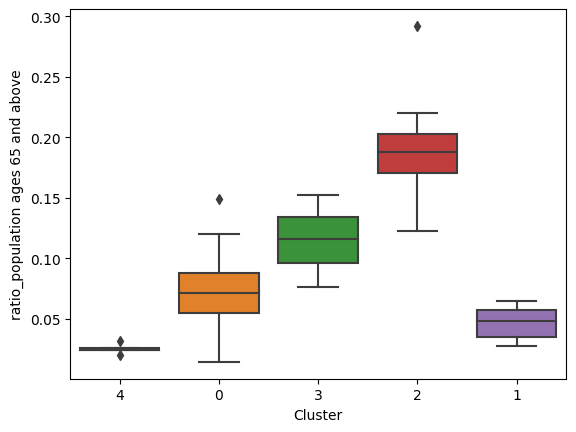
Observamos que el cluster 2 agrupa los valores mayores para el indicador de “Life expectancy at birth, total (years)”



Observamos que el cluster 2 agrupa los valores mayores para el indicador de “GDP per capita (US$)”



Observamos que el cluster 2 agrupa los valores mayores para el indicador de “Ratio\_population ages 65 and above”



**Elección del cluster para la recomendación de países a la empresa**

En base a los gráficos anteriores observamos que el cluster número 2 resalta sobre los demás en cada uno de los indicadores por lo que es el elegido.

**Países recomendados**