Introducción

El desarrollo socioeconómico de los países está determinado por una interacción compleja de factores. Este estudio se enfoca en identificar y comparar los factores más influyentes sobre las tendencias de crecimiento de Argentina, utilizando como referencia países con desarrollos similares tanto en Latinoamérica como en otras regiones seleccionadas.
Con una perspectiva práctica, buscamos proporcionar datos relevantes que puedan servir como referencia en la elaboración de estudios futuros. Nos centraremos en comprender: ¿Qué indicadores que dan cuenta de factores socioeconómicos han influido significativamente en el desarrollo de Argentina durante las últimas seis décadas y cuál es su correlato en países con desarrollos comparables?

Metodos y Materiales

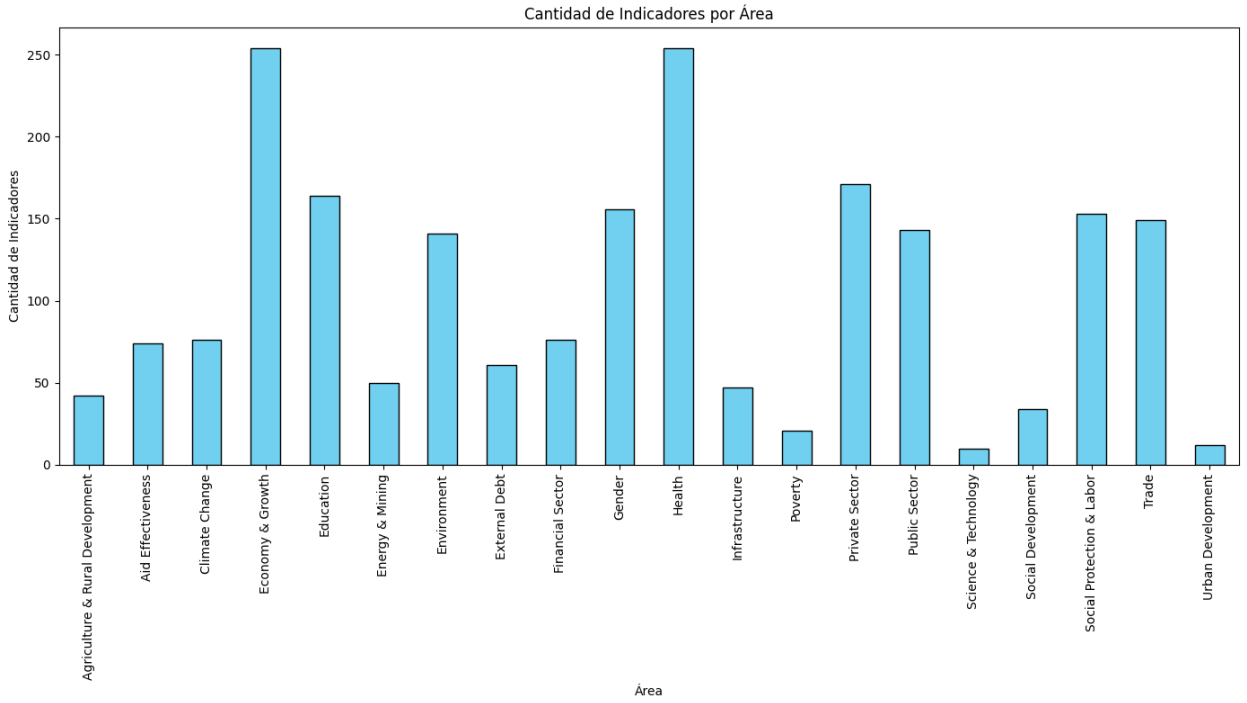
Datos

Trabajaremos con un dataset del Banco Mundial que incluye 1463 indicadores de países agrupados en 20 en áreas clave, tales como Educación, Ciencia y Tecnología, y Crecimiento económico, abarcando el período de 1960 a 2021. Los datasets se encuentran disponibles en este link de descarga.

Dentro de la pagina del Banco Mundial, los datos se almacenan y acceden de manera individual. Esto es, cada indicador se encuentra en un archivo excel con la información de ese indicador para todos los paises y todos los años.

Tal como se menciona anteriormente, los indicadores se encuentran agrupados por distintas areas, cada indicador puede pertenecer a mas de un área.

En la siguiente figura puede verse la distribución de indicadores por area.

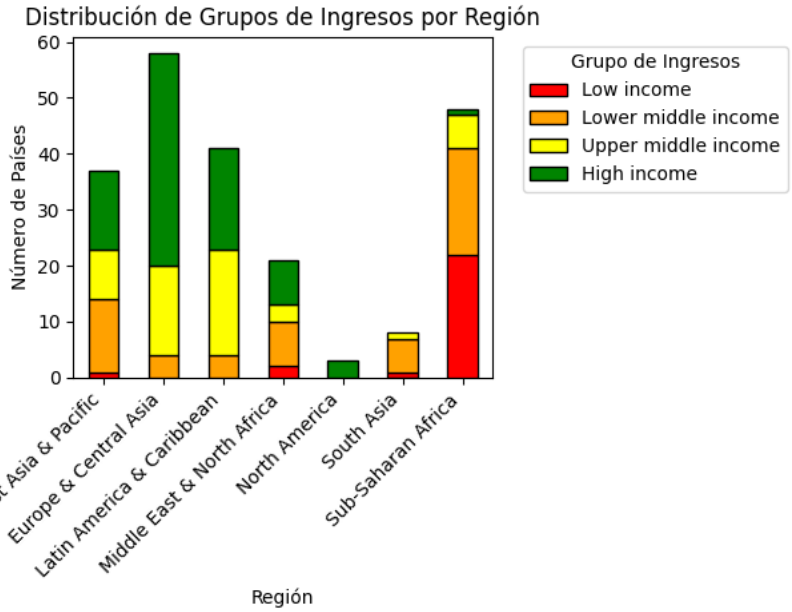


Dentro del dataset, los paises se categorizan por dos criterios: Geografico y economico. Las regiones en las que distingue el banco mundial a los paises son las siguientes:

- 'Latin America & Caribbean'
- 'North America'
- 'Middle East & North Africa',
- 'Sub-Saharan Africa',
- 'Europe & Central Asia',
- 'East Asia & Pacific',
- 'South Asia'

Mientras que las distintas categorias de ingresos en las que clasifica a los paises son las siguientes:

- **Low income**: "*Low-income economies are those in which 2022 GNI per capita was $1,135 or less*".
- **High income**: "*High-income economies are those in which 2022 GNI per capita was more than $13,845*".
- **Lower middle income**: "*Lower-middle-income economies are those in which 2022 GNI per capita was between $1,136 and $4,465*".
- **Upper middle income:** "*Upper-middle-income economies are those in which 2022 GNI per capita was between $4,466 and $13,845*".



Preprocesamiento

Los datos se encuentran agrupados en distintas planillas por cada indicador. El trabajo de preprocesamiento consiste en los siguientes pasos:

1. Descarga automática: Utilizando Selenium, se realizaron las descargas automaticas de todos los archivos en formato xls.
2. Generacion de metadata: Luego de bajar los archivos, estos se procesan y se genera archivo maestro con la metadata necesaria para el estudio. Esta metainformacion incluye, las areas en las que los indicadores se encuentran agrupados, los nombres y descripciones de los significados, los codigos y nombres de los paises, etc.
3. Agrupamiento: Para facilitar el uso y manipulacion de los datos, se procesaron los archivos individuales para generar un consolidado por area, que agrupa a todos los indicadores de esa area para todos los paises.

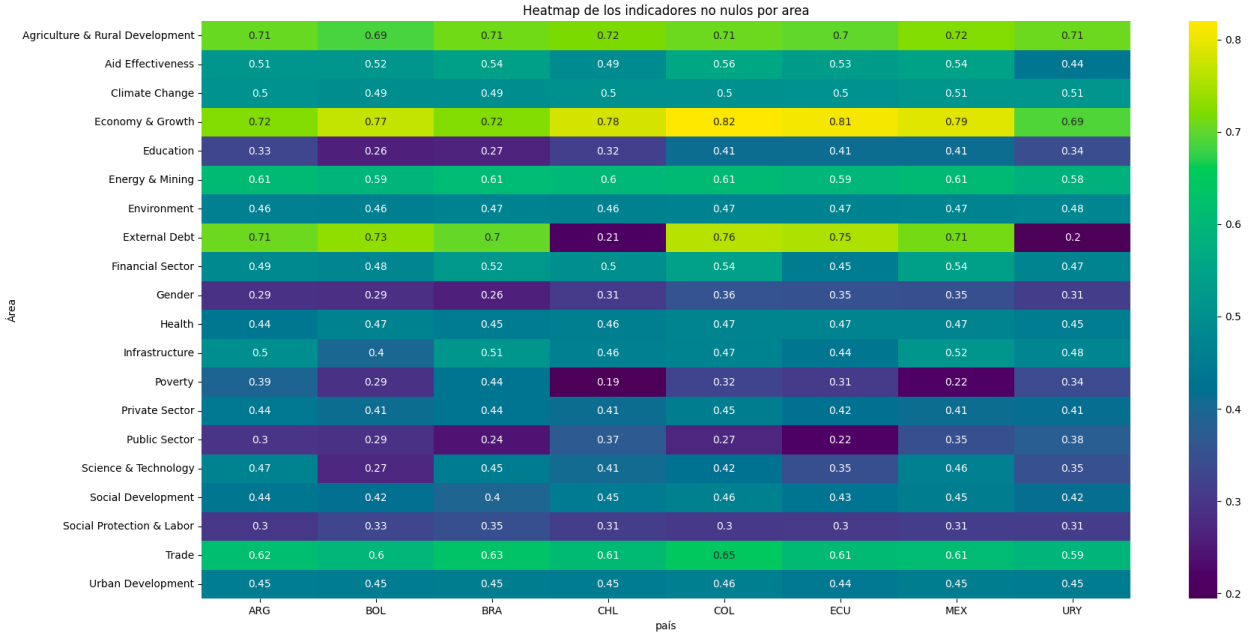
Analisis y Modelado

Nuestro análisis constará de: (1) Métodos descriptivos que incluyen técnicas estadisticas y visualización de datos para caracterizar y comparar países; (2) Técnicas de clustering como K-means para identificar agrupaciones de países con patrones de desarrollo similares; y (3) Análisis de series temporales utilizando modelos como ARIMA para evaluar la evolución de los indicadores a lo largo del tiempo.

Métodos descriptivos

El foco estara dado en entender si los datos existentes permiten la utilización de las tecnicas de clustering y de analisis de series temporales que pensamos utilizar.

En un primer analisis de un subconjunto de paises de Latinoamerica, puede verse que hay una gran cantidad de datos faltantes (utilizando todos los años del dataset).



Esto puede dificultar la tarea si dichos faltantes no se tratan adecuadamente. No obstante, como puede verse en la siguiente figura, filtrando por los años de 2015 a 2018 los ratios de datos faltantes mejoran.

Técnicas de clustering

El principal desafio es determinar que hacer con los datos faltantes. Creo que dada la cantidad de datos incompletos, imputar los valores faltantes no es una buena idea y puede llevar a conclusiones erroneas.
Creo que lo mas efectivo sera realizar cortes especificos en los datos y ejecutar las tecnicas sobre espacios de datos mas pequeños y completos.
Algunas estrategias de corte que tengo pensado explorar son las siguientes:

- Por region: Elegir solo algunos subconjuntos de paises de determinadas regiones que tengan mayor completitud
- Por rangos de fechas: Hay años especificos donde se encuentran muchos menos datos faltantes.
- Por indicadores: Seleccionare subconjuntos de indicadores que tengan la mayor cantidad de datos posibles

Cabe destacar, que si para un corte dado, hay solo unos pocos datos faltantes, eventualmente podria intentar imputar los valores con distintas tecnicas (Media, regresión o MICE podrian ser efectivos en estos casos) y evaluar los resultados obtenidos.

Análisis de series temporales

Los mismos problemas de datos faltantes afectan a este tipo de tecnicas. No obstante, en este caso creo que el impacto puede llegar a ser aun mayor.
Utilizare este tipo de tecnicas de manera aun mas acotada, comparando Argentina con otros paises seleccionados en funcion de los clusters a los que pertenezca. Esto es, comparando la evolución de Argentina en determinados indicadores clave y su relacion con paises especificos dentro de sus mismos clusters.
En principio utilizaré ARIMA para el analisis de series temporales.

Herramientas

Para el análisis de datos y modelado utilizaremos las siguientes herramientas:

- Python: Utilizaremos diversas bibliotecas para distintas tareas:
  - Selenium: Para la automatización de la descarga de datos desde la web.
  - OpenPyXL: Para manipulación de archivos Excel durante el preprocesamiento de datos.
  - Pandas y NumPy para la manipulación y análisis de datos.
  - Matplotlib y Seaborn para la visualización de datos.
  - Scikit-learn para las técnicas de clustering
  - Statsmodels para el análisis de series temporales, incluyendo modelos ARIMA.
  - PDFPages de Matplotlib: Para generar reportes en PDF con gráficos de los análisis realizados.

- Jupyter Notebooks: Para documentar y presentar los análisis de manera interactiva.
- Microsoft Excel: Para la gestión inicial de los datos descargados y la generación de la metadata necesaria.

Resultado esperado

Esperamos que este enfoque analítico revele patrones subyacentes en la trayectoria de desarrollo de los países estudiados, ofreciendo así una base de conocimiento que pueda servir para futuras investigaciones y para la formulación de políticas más efectivas hacia el desarrollo sostenible.