Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Taller de Tesis I – Entrega II

Grupo 2

Miguel Kiszkurno

Tabla de Contenidos

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.

El caso argentino es estudiado a lo largo de todo el mundo por sus particularidades. Gonzalez (2012) utiliza modelos de series temporales para estimar los efectos de las privatizaciones y la inversión directa extranjera sobre el crecimiento del país en el periodo1971–2000; mientras que Taylor (1994) hace un estudio histórico de nuestra performance económica caracterizando 3 etapas: pre-1913, 1913-1930s y 1930s-1950s. Asimismo, Weisskoff (1970) intenta responder la pregunta de si el crecimiento económico en países en desarrollo lleve a inequidades en la distribución del ingreso, analizando en detalle los casos de Puerto Rico, Argentina y México.

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?

Métodos y Materiales

Datos

Utilizaremos un conjunto de datos del Banco Mundial que comprende 1,463 indicadores, organizados en 20 áreas clave, como Educación; Ciencia y Tecnología; y Crecimiento Económico; cubriendo el período de 1960 a 2023[1]. Los conjuntos de datos están disponibles para su descarga en el siguiente enlace.

En el sitio web del Banco Mundial, los datos se almacenan y acceden de manera individual; es decir, cada indicador se encuentra en su propio archivo Excel con información correspondiente a todos los países y todos los años. Un mismo indicador puede pertenecer a varias áreas. En la Figura 1 se muestra la distribución de dichos indicadores.

[1] Los datos de 2023 se encuentran aún incompletes. Es por esto que serán descartados para el análisis propuesto.

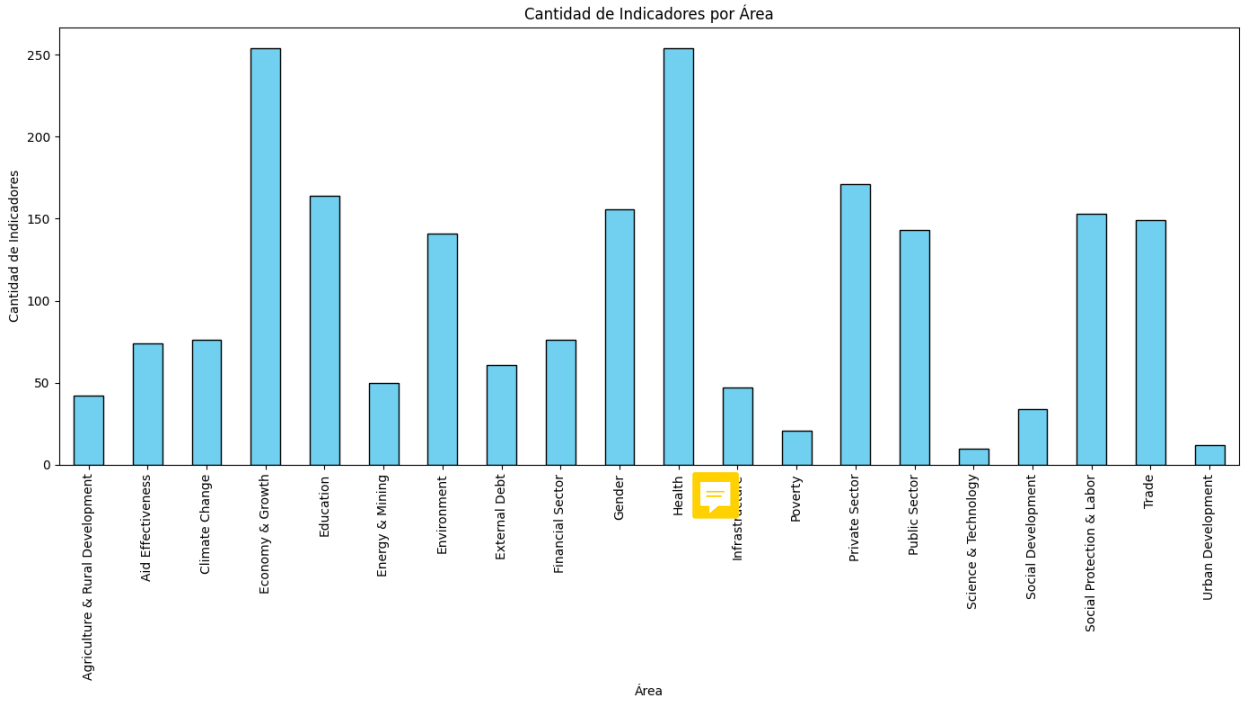

Figura 1: Cantidad de indicadores por Área

Dentro del dataset, los países se categorizan por dos criterios: ~~G~~eográfico y económico. Las regiones en las que distingue el ~~b~~anco ~~m~~undial a los países son las siguientes (Figura 2):

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

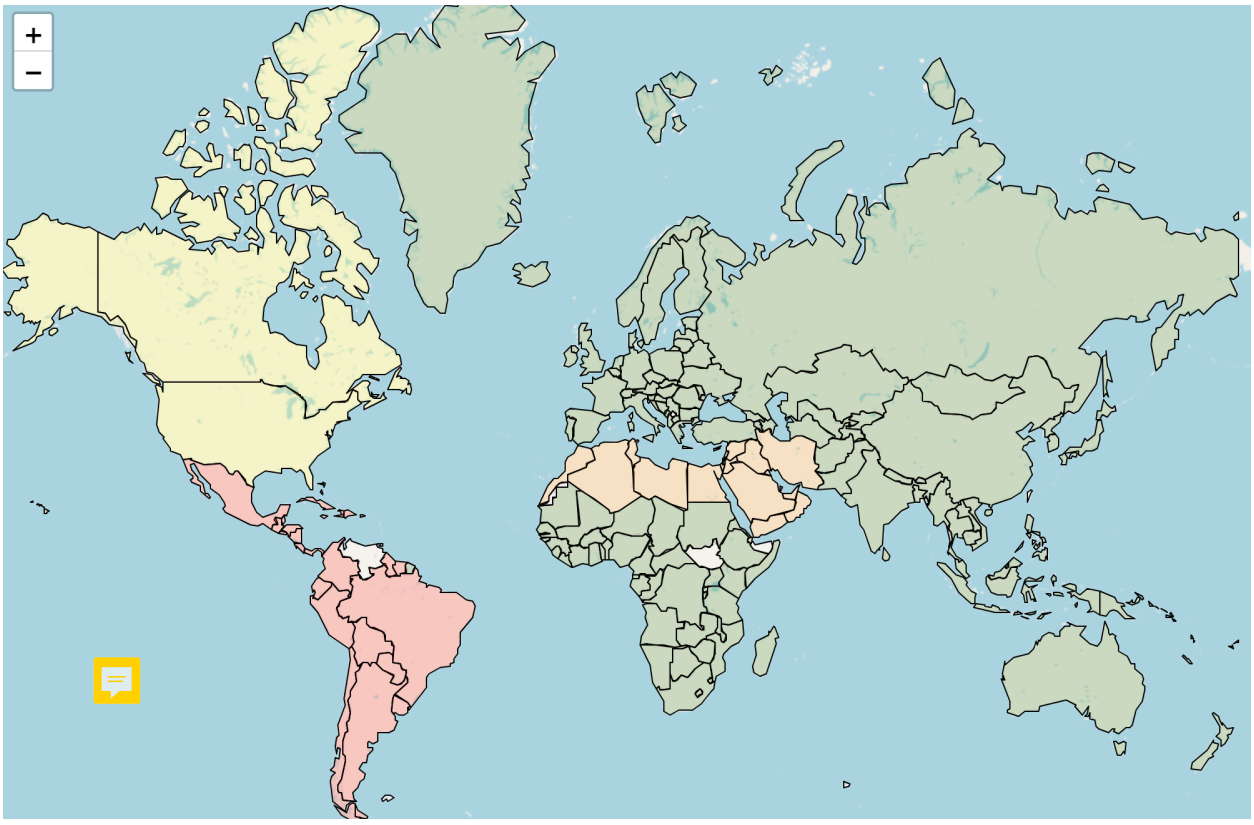

Figura 2: países según su región

Por otro lado, las distintas categorías de ingresos en las que clasifica a los países son las siguientes (Figura 3 y Figura 4):

- **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*".

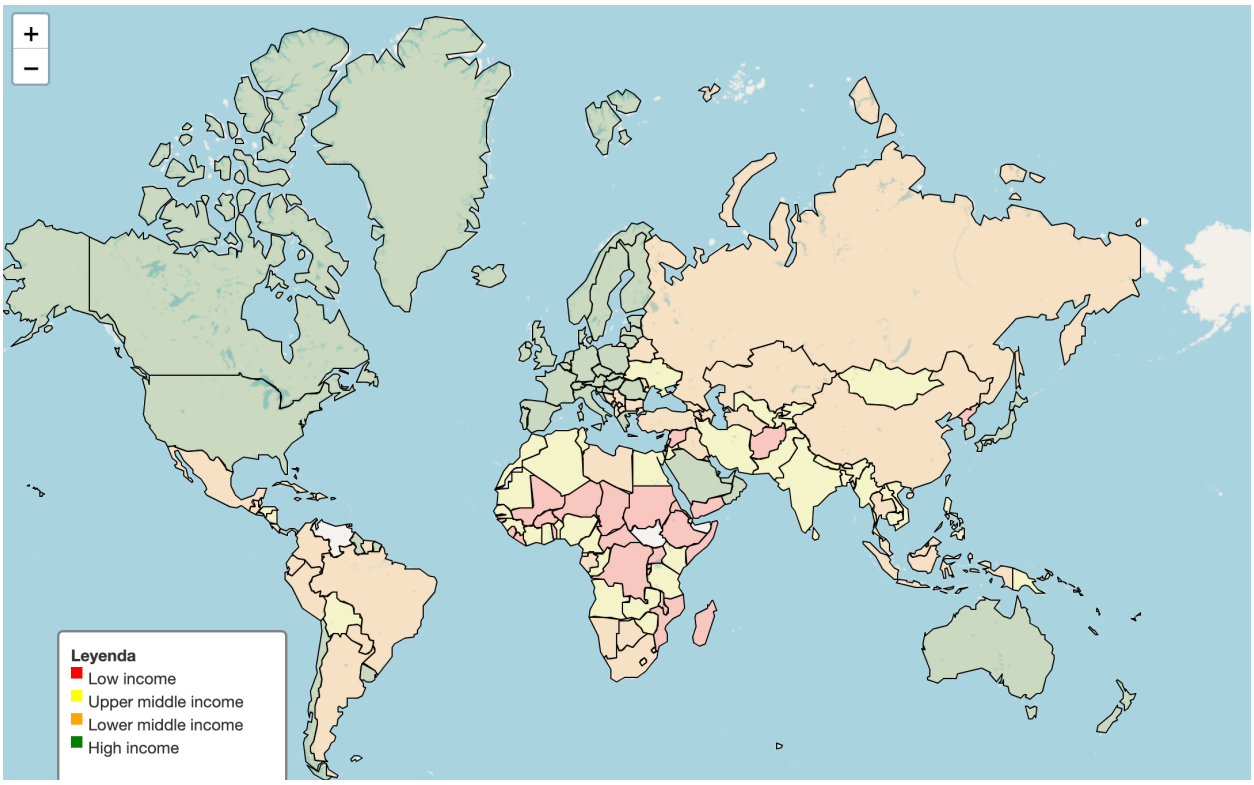

Figura 3: Países según su nivel de ingreso

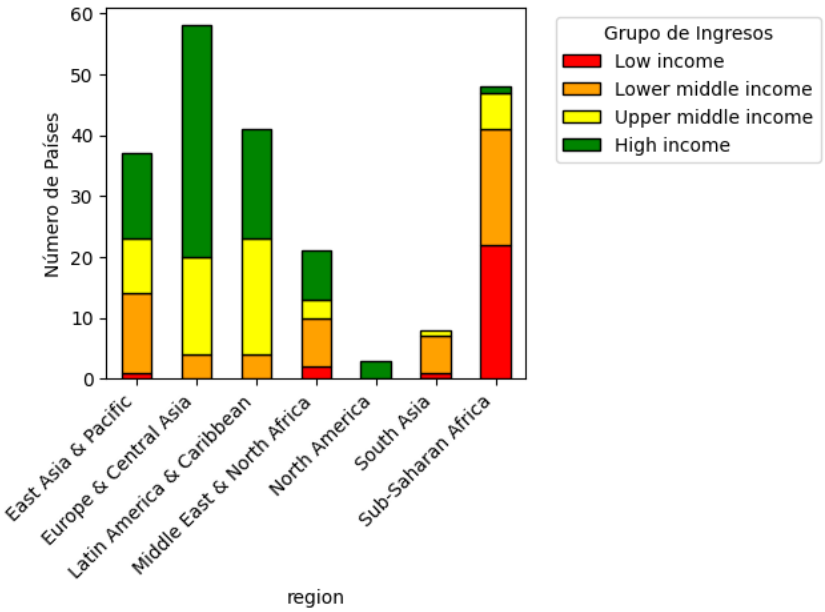

Figura 4: Países por su categoría de ingreso

Herramientas

Tanto para la descarga y preprocesamiento de los datos, como para el análisis exploratorio, y el modelado, utilizaremos python corriendo sobre la ide visual studio code.

Dentro de python usaremos las siguientes bibliotecas y librerías:

- Selenium: Se utiliza para automatizar la descarga de datos desde la web.
- OpenPyXL: Se utiliza para manipular los archivos Excel durante el preprocesamiento de datos.
- Pandas y NumPy: Se utilizan para manipulación y análisis de datos.
- Matplotlib y Seaborn: Se utilizan para la visualización de datos.
- PDFPages de Matplotlib: Para generar reportes en PDF con gráficos de los análisis realizados.
- Holoviews y bokeh: Se utilizan para la visualización dinámica de datos
- Scikit-learn: Se utiliza para aplicar las técnicas de clustering en la etapa de modelado
    - StandardScaler
    - KMeans
    - silhouette_score
- Statsmodels: Se utiliza para el análisis de series temporales, incluyendo modelos ARIMA.
- Jupyter Notebooks: Se utiliza para la etapa de análisis exploratorio. Permite no solo ejecutar código sino también documentar y presentar los análisis de manera interactiva.

Adicionalmente a las librerías de Python, se utilizaron los siguientes programas de oficina:

- Microsoft Excel: Para la gestión inicial de los datos descargados y la generación de la metadata necesaria.
- Microsoft Word: Para la elaboración de este documento

Preprocesamiento de los datos

Los datos se encuentran agrupados en distintas planillas. Cada indicador tiene su propio archivo, donde se encuentran los países como cabecera de las filas y los años como cabecera de la columna.

Es por esto por lo que para generar datasets fácilmente utilizables, hubo que realizar las siguientes tareas de preprocesamiento:

1. Descarga automática: Utilizando Selenium, se realizaron las descargas automáticas de todos los archivos en formato xls.

    a. Estos archivos se convirtieron a CSV para facilitar su manipulación[2]
2. Generación de metadata: Luego de bajar los archivos, estos se procesan y se genera archivo maestro con la metadata necesaria para el estudio. Esta metainformación incluye:
    a. Catálogo de indicadores: Con su código, nombre, descripción y ruta de acceso al archivo de origen dentro del repositorio de código.
    b. Catálogo de indicadores por área clave: Cada una de las áreas clave junto con sus indicadores y su dirección url.
    c. Catálogo de países: Los códigos, nombres, nivel de ingreso, agrupamiento geográfico y comentarios de creadores de los datos
3. Agrupamiento: Para facilitar el uso y manipulación de los datos, se procesaron los archivos individuales para generar un consolidado por área, que agrupa a todos los indicadores de esa área para todos los países.
4. Cálculo de Métricas: Dada la cantidad de archivos e indicadores, el análisis exploratorio se hace extremadamente complejo y costoso. Es por esto que genere un proceso que calcula métricas principalmente sobre los datos faltantes (que son críticos para las técnicas de modelado que pienso utilizar).

Análisis y Modelado

Nuestro análisis constará de: (1) Métodos descriptivos que incluyen técnicas estadísticas 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

Dado el origen de los datos, tome la decisión de aceptarlos como válidos. Entonces, el análisis exploratorio se basa principalmente en entender con la mayor precisión posible cuales son los datos que faltan, y buscar estrategias que ~~me~~ permitan imputar o recortarlos para poder aplicar las técnicas de clustering y de análisis de series temporales propuestas.

El volumen de datos hace dificultosa la tarea de explorarlos todos juntos, así que en una primera instancia me enfoque en hacer un primer análisis acotando los países de estudio a un pequeño conjunto de Latinoamérica: Argentina, Bolivia, Chile, Colombia, Ecuador, Méjico y Uruguay.

En la Figura 5 se visualiza el porcentaje de datos existentes sobre datos totales para cada país para cada área de interés (agrupando todos los indicadores de dicha área para todos los años. Hay escasas áreas donde este porcentaje supere el 80%. En particular, Argentina tiene un máximo de 72% en Economy & Growth.

[2] El World Bank también provee los datos en formato CSV, no obstante, ~~hacia~~ más compleja la tarea de descarga ya que por cada indicador requería descargar más de un csv (uno para el indicador y otros dos con metadata)

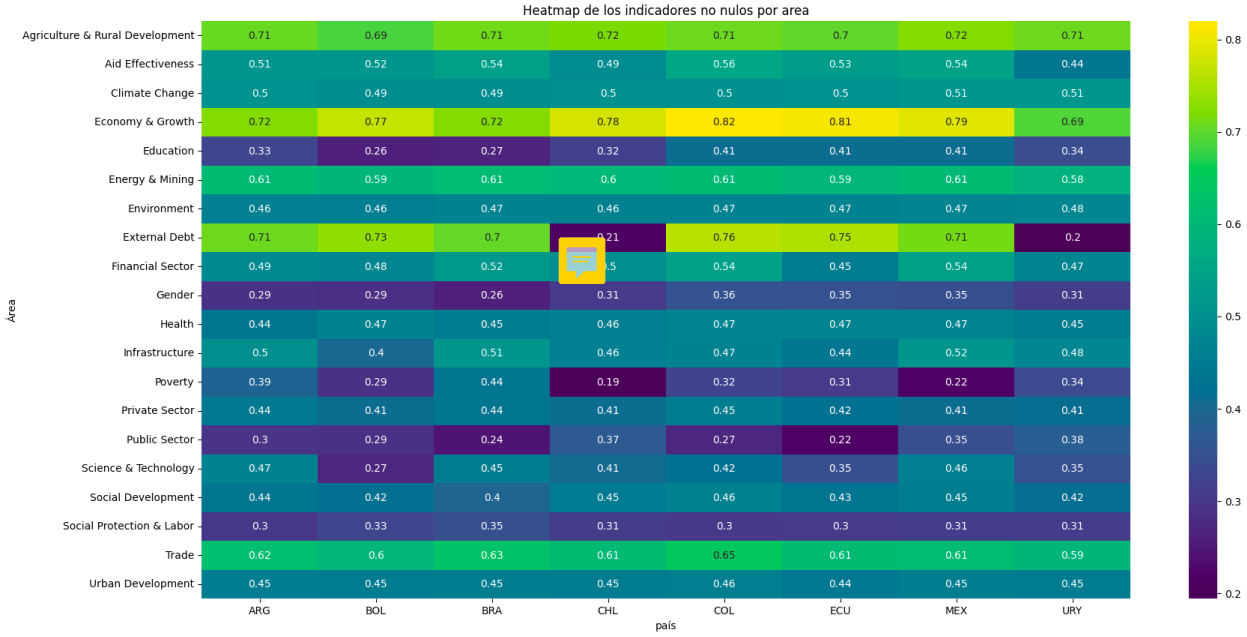

Figura 5: Datos faltantes para ~~un~~ conjunto de países seleccionados

La Figura 6~~, que~~ muestra el mismo gráfico, ~~al~~ periodo 2015-2018 la cantidad de datos faltantes disminuye notablemente, alcanzando picos de no nulos del 96% en algunos casos.

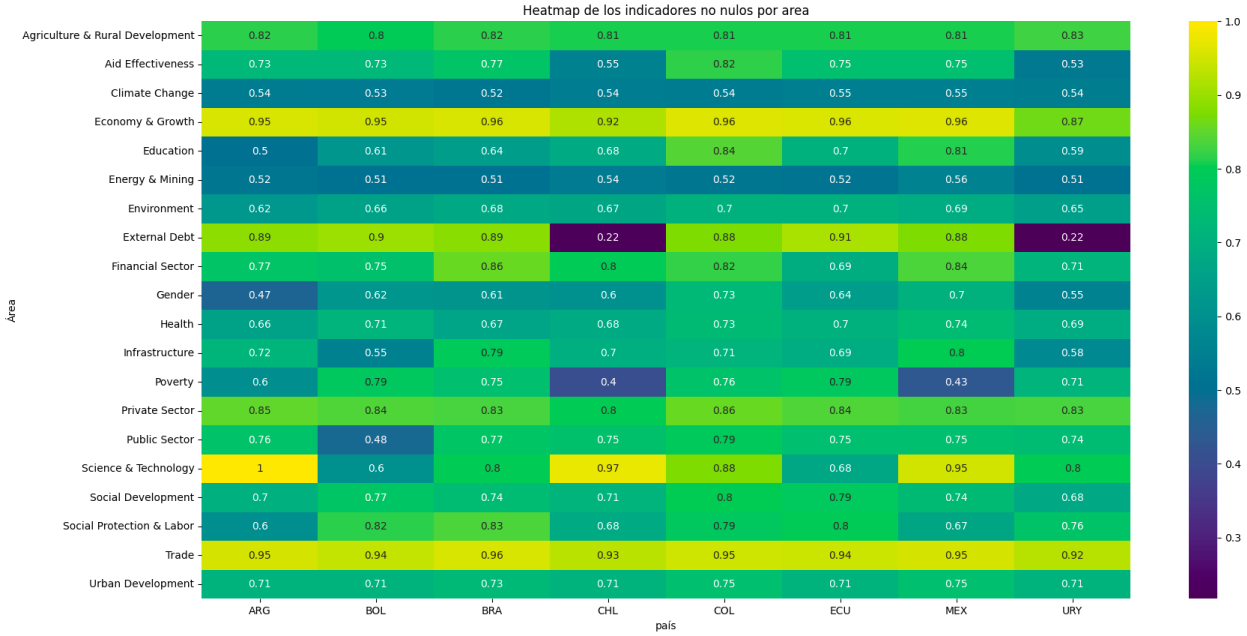

Figura 6: Datos faltantes para ~~un~~ conjunto de países seleccionados (años 2015 a 2018)

Analizando más específicamente el área de Economy & Growth, puede verse que en el periodo 1994-2018 es cuando más datos se tienen.

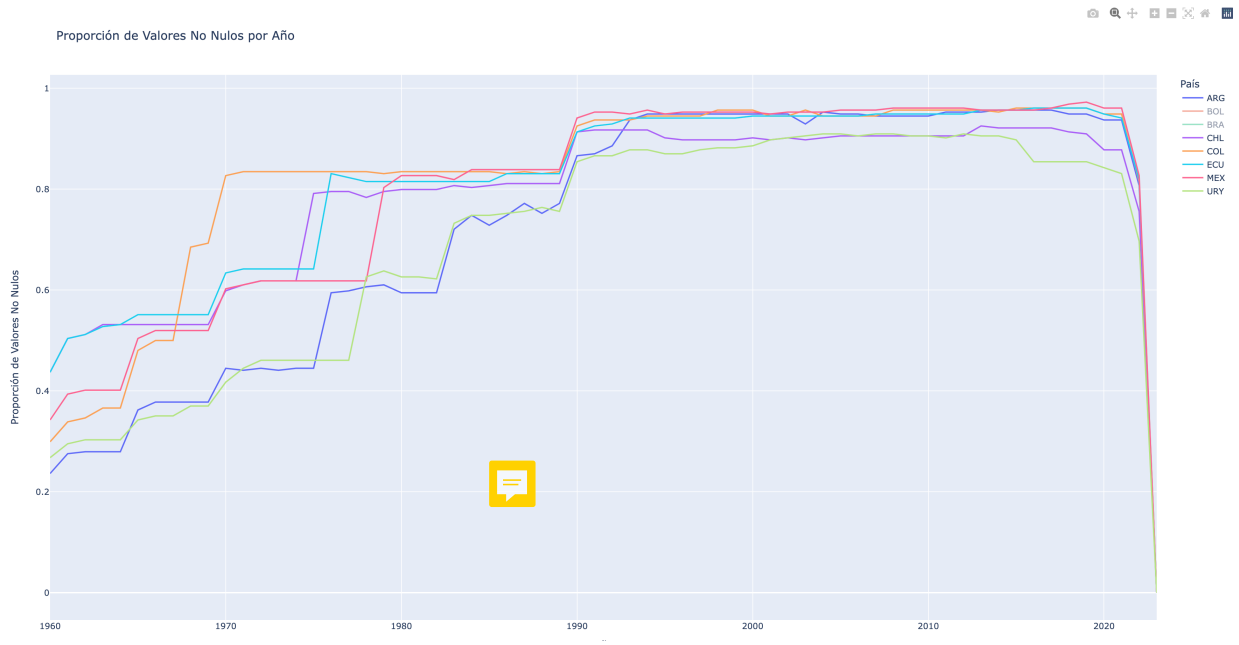

Figura 7: Cantidad de valores no nulos por año para el área de Economy & Growth

Creo que si bien se requiere mayor análisis permite plantear estrategias de mitigación ante la falta de datos.

Búsqueda de información externa

Si bien ~~hay~~ algunos de estos indicadores ~~que~~ podrían estar disponibles en otros datasets~~. Creo que no puede~~ considerarse como el método por defecto~~. Llevaría demasiado~~ tiempo y no hay garantía de que puedan conseguirse los datos necesarios. ~~Usaría esta técnica~~ en casos muy específicos (algún indicador conocido en Argentina por ejemplo).

Imputar datos faltantes

Para los casos en que consiga datos con unos pocos faltantes (tal vez menos del 5%) podemos intentar imputarlos usando diversas técnicas y ver a que conclusiones nos lleva. En principio sería muy cuidadoso con la forma en se imputan los datos y tomaría con cautela las conclusiones derivadas de esto.

Segmentar los datos

Dado que hay periodos de tiempo que parecen tener mayor volumen de datos completos~~. Me inclinaría a~~ explorar estrategias de corte que eviten la necesidad de imputar los datos faltantes de manera masiva.

Tengo pensado explorar las siguientes estrategias:

- Cortes por región: Elegir solo algunos subconjuntos de países de determinadas regiones que tengan mayor completitud de datos[3].
- Por rangos de fechas: Hay años específicos donde se encuentran muchos menos datos faltantes. Un ejemplo es el periodo 1990-2020 (para la muestra antes descripta), pero otro periodo realmente interesante para Argentina es el 1994-2018, que tiene más de 700 indicadores con ratio = 1 (esto es, sin datos faltantes).
- Por indicadores: Seleccionare subconjuntos de indicadores que tengan la mayor cantidad de datos posibles.

Esquema mixto

Cabe destacar, que si para un corte dado, hay solo unos pocos datos faltantes, eventualmente podría intentar imputar los valores con distintas técnicas. Exploraré las siguientes:

- Media o media móvil: Tomar la media (o media móvil) del país en el indicador dado
- Ultimo valor conocido: En general (salvo crisis y catástrofes) los indicadores económicos tienden a moverse lentamente, entonces es de esperar que el valor de un año sea muy parecido al del año anterior sumado a una tendencia.
- Regresión
- MICE

Técnicas de clustering

Una vez resuelto (o mitigado) el problema de los datos faltantes intentare aplicar al menos 2 técnicas de clustering. En principio KNN y K-means

Un último paso de preprocesamiento antes de ejecutar los algoritmos será escalar los datos, ya que KNN calcula los clusters usando el concepto de distancia, no escalar los datos podría generar clusters defectuosos.

Para evaluar los clusters resultantes utilizare el coeficiente de silhouette, que mide la similaridad entre los puntos dentro del mismo cluster en comparación con los puntos fuera del cluster; y La inercia, para k-means, que mide la suma de las distancias al cuadrado entre cada punto y el centroide de su cluster.

En función de los grupos que los algoritmos generen, hare un análisis de los grupos conformados usare PCA para graficarlos en 2d.

Análisis de series temporales

Los mismos problemas de datos faltantes afectan a este tipo de técnicas. No obstante, en este caso creo que el impacto puede llegar a ser aún mayor.

[3] En todo caso, tomara prioridad para el análisis la completitud de datos de Argentina, que es el país donde se enfoca este trabajo.

Utilizare este tipo de técnicas de manera aún más acotada, comparando Argentina con otros países seleccionados en función de los clusters a los que pertenezca. Esto es, comparando la evolución de Argentina en determinados indicadores clave y su relación con países específicos dentro de sus mismos clusters.

En principio utilizaré ARIMA para el análisis de series temporales.

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.

Bibliografía y Referencias

[1] Fernando Antonio Ignacio González, Silvia London, María Emma Santos (2012). The Journal of International Trade & Economic Development. The Journal of International Trade & Economic Development.

[2] Alan M. Taylor (1994).Three Phases of Argentine Economic Growth. National Bureau of economic research.

[3] Robert J. Barro (1996). Determinants of Economic Growth: A Cross-Country Empirical Study. National Bureau of economic research.

[4] Daniel Landau (1986). Government and Economic Growth in the Less Developed Countries: An Empirical Study for 1960-1980. Universidad de Connecticut

[5] Michael Timberlake, Jeffrey Kentor (1986). Economic Dependence, Overurbanization, and Economic Growth: A Study of Less Developed Countries. The sociological quarterly, official Journal of the Midwest Sociological Society.

[6] Richard Weisskoff (1970). INCOME DISTRIBUTION AND ECONOMIC GROWTH IN PUERTO RICO, ARGENTINA, AND MEXICO. The review of income and wealth.

[7] World Bank Open Data, Free and open access to global development data.