A black and white logo with a person sitting on a chair

Description automatically generated

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 3](#_Toc168760259)

[Marco teórico 3](#_Toc168760260)

[Relevamiento de trabajos previos y relevantes 3](#_Toc168760261)

[Conceptos y técnicas de ciencia de datos utilizados en el trabajo 3](#_Toc168760262)

[Técnicas de clustering 3](#_Toc168760263)

[Análisis de series temporales 3](#_Toc168760264)

[El problema de los datos faltantes 4](#_Toc168760265)

[Metodología 5](#_Toc168760266)

[Presentación y descripción de los datos utilizados 5](#_Toc168760267)

[Preprocesamiento y limpieza de los datos 9](#_Toc168760268)

[Análisis exploratorio de datos (AED) 10](#_Toc168760269)

[Datos faltantes 10](#_Toc168760270)

[Categorización de los países 13](#_Toc168760271)

[Métodos descriptivos 17](#_Toc168760272)

[Descripción de las técnicas de análisis y, si corresponde, de modelado 17](#_Toc168760273)

[Descripción de la selección de características 17](#_Toc168760274)

[Descripción de las métricas de evaluación de los modelos 17](#_Toc168760275)

[Descripción de los métodos estadísticos utilizados 17](#_Toc168760276)

[Herramientas 18](#_Toc168760277)

[Resultados y discusión: 18](#_Toc168760278)

[Conclusión: 18](#_Toc168760279)

[Bibliografía y Referencias 18](#_Toc168760280)

# 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 periodo 1971–2000; mientras que Taylor (1994) hace un estudio histórico de nuestra performance económica caracterizando las siguientes etapas: pre-1913, 1913-1930s y 1930s-1950s. Asimismo, Weisskoff (1970) intenta responder la pregunta de si el crecimiento económico en países en desarrollo llevó 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?

# Marco teórico

## Relevamiento de trabajos previos y relevantes

Comentar papers que encontré

## Conceptos y técnicas de ciencia de datos utilizados en el trabajo

### Técnicas de clustering

Se utilizaron distintas KNN y K-means para clasificar y caracterizar a los distintos países en función de distintos determinados indicadores socioeconómicos disponibles en el dataset del Banco Mundial.

Partiendo desde la clasificación en 4 categorías de ingreso según define el Banco Mundial, se realizó un primer experimento que utiliza únicamente el Ingreso per cápita para el 2022.

Luego, se realizaron distintos experimentos sobre distintos años evaluando distintos indicadores para buscar diferencias sustanciales entre las nuevas categorizaciones y la realizada por el banco Mundial.

Los clústeres resultantes se evaluaron utilizando el coeficiente de silhouette, que mide la similaridad entre los puntos dentro del mismo clúster en comparación con los puntos fuera del clúster; y La inercia, para k-means, que mide la suma de las distancias al cuadrado entre cada punto y el centroide de su clúster.

Otro camino de exploración que dejamos como posible trabajo futuro es utilizar para los distintos clústeres, todos los años disponibles para todos los países, creando así clústeres en función de la evolución de los países en el transcurso del tiempo.

### 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.

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.

### El problema de los datos faltantes

A continuación, se describen las opciones evaluadas junto con los resultados obtenidos.

#### Búsqueda de información externa

Si bien algunos de estos indicadores podrían estar disponibles en datasets de otras organizaciones. Éste no puede considerarse como el método por defecto. El tiempo de búsqueda y validación es una limitante, y no hay garantía de éxito. Esta técnica puede utilizarse en casos muy específicos (algún indicador conocido en Argentina, por ejemplo).

#### Imputar datos faltantes

Para los indicadore con hasta el 5% de datos faltantes se intenta imputarlos utilizando diversas técnicas, luego evaluar las conclusiones a las que esto conduce. Es necesario ser muy cuidadosos con la forma en se imputan, y las conclusiones derivadas de esto deben ser tomadas con cautela.

#### Segmentar los datos

Dado que hay periodos de tiempo que parecen tener mayor volumen de datos completos. Lo mejor es explorar estrategias de corte que eviten la necesidad de imputar los datos faltantes de manera masiva.

* Cortes por región: Elegir solo algunos subconjuntos de países de determinadas regiones que tengan mayor completitud de datos[[1]](#footnote-1).
* 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: Se seleccionan los subconjuntos de indicadores que tienen la mayor cantidad de datos posibles.

#### Esquema mixto

Cabe destacar; qué si para un corte dado, hay solo unos pocos datos faltantes, se intenta imputarlos con distintas técnicas:

* 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

# Metodología

## Presentación y descripción de los datos utilizados

El trabajo se realizó sobre un conjunto de datos del Banco Mundial que comprende 1,463 indicadores, organizados en 20 áreas clave (ver Tabla 1), como Educación; Ciencia y Tecnología; y Crecimiento Económico; cubriendo el período de 1960 a 2022. Los conjuntos de datos están disponibles para su descarga en el siguiente [enlace](https://data.worldbank.org/indicator).

|  |  |  |
| --- | --- | --- |
| **Área** | **Descripción** | **Ejemplos de indicadores** |
| Agricultura y Desarrollo Rural | Incluye indicadores sobre producción agrícola, uso de la tierra, insumos agrícolas y desarrollo rural. | Producción agrícola, Uso de fertilizantes, Tierra arable |
| Eficacia de la Ayuda | Mide la efectividad y el impacto de la ayuda internacional. | Flujos netos de ayuda oficial al desarrollo, Asistencia bilateral |
| Cambio Climático | Se centra en los efectos del cambio climático y las medidas de mitigación. | Emisiones de CO2, Consumo de energía renovable |
| Economía y Crecimiento | Comprende datos sobre el crecimiento económico, la estructura económica y la productividad. | Producto Interno Bruto (PIB), Crecimiento del PIB per cápita |
| Educación | Incluye estadísticas sobre el acceso a la educación, la calidad educativa y los resultados de aprendizaje. | Tasa de matrícula en educación primaria, Gasto público en educación |
| Energía y Minería | Se enfoca en el suministro y el consumo de energía, así como en la explotación de recursos minerales. | Producción de energía, Consumo de electricidad per cápita |
| Medio Ambiente | Cubre temas como la biodiversidad, la calidad del aire y el agua, y la gestión de residuos. | Áreas protegidas, Índice de calidad del aire |
| Deuda Externa | Proporciona datos sobre la deuda externa de los países y su sostenibilidad. | Deuda externa total, Pagos de servicio de la deuda |
| Sector Financiero | Incluye estadísticas sobre la banca, los mercados financieros y el acceso al financiamiento. | Crédito doméstico al sector privado, Capitalización bursátil |
| Género | Examina las desigualdades de género en diversas áreas como la educación, la salud y el empleo. | Tasa de participación laboral femenina, Diferencia salarial entre géneros |
| Salud | Cubre aspectos de la salud pública, el acceso a servicios de salud y los resultados de salud. | Esperanza de vida al nacer, Mortalidad infantil |
| Infraestructura | Se centra en el desarrollo y la calidad de la infraestructura básica y avanzada. | Acceso a electricidad, Infraestructura de transporte |
| Pobreza | Mide la incidencia y la severidad de la pobreza. | Tasa de pobreza internacional (menos de $1.90 al día), Índice de Gi |
| Sector Privado | Incluye datos sobre el desarrollo del sector privado y la actividad empresarial. | Número de empresas nuevas registradas, Facilidad para hacer negocios |
| Sector Público | Examina la eficiencia y la calidad del gobierno y el sector público. | Gasto público como % del PIB, Índice de gobernabilidad |
| Ciencia y Tecnología | Proporciona datos sobre innovación, investigación y desarrollo tecnológico. | Gasto en I+D como % del PIB, Número de patentes registradas |
| Desarrollo Social | Incluye estadísticas sobre bienestar social, cohesión social y servicios sociales. | Acceso a servicios de saneamiento, Índice de desarrollo humano |
| Protección Social y Trabajo | Se enfoca en la seguridad social, las condiciones laborales y la protección del empleo. | Cobertura de seguridad social, Tasa de desempleo |
| Comercio | Examina el comercio internacional y la integración económica. | Exportaciones de bienes y servicios, Balanza comercial |

Tabla 1: Áreas clave en las que el World Bank agrupa los distintos indicadores socioeconómicos de las naciones y su descripción junto a algunos indicadores relevantes por área.

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.

A graph of blue rectangular bars with black text

Description automatically generated with medium confidence

Figura 1: Cantidad de indicadores por Área

Dentro del dataset, los países se categorizan por dos criterios: geográfico y económico. Las regiones en las que distingue el Banco Mundial 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

A map of the world with different colors

Description automatically generated

Figura 2: Mapa de los países según su región en el mapa.

El Banco Mundial categoriza a los países en función de su GNI, que es la suma del valor agregado por todos los productores residentes más cualquier impuesto sobre productos (menos subsidios) no incluido en la valoración de la producción, más los ingresos netos primarios (compensación de empleados e ingresos por propiedad) recibidos del extranjero. Los datos se expresan en dólares estadounidenses corrientes. Las categorías son las siguientes:

* **Low income**: “*Low-income economies are those in which 2022 GNI per capita was $1,135 or less*”.
* **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*”.
* **High income**: “*High-income economies are those in which 2022 GNI per capita was more than $13,845*”.

La Figura 3 muestra el mapa mundial con los países coloreados según su categoría de ingresos. La Figura 4 muestra los países agrupados por geografía y categoría de ingresos. La Figura 5 muestra las cantidades totales de países por cada categoría de ingreso.

A map of the world with different colors

Description automatically generated

Figura 3: Los países según su nivel de ingreso dentro del mapa.

A graph with different colored bars

Description automatically generated

Figura 4: Muestra la cantidad de países en función de su categoría de ingreso agrupando por la región a la que pertenecen

A graph of different colored rectangular shapes

Description automatically generated with medium confidence

Figura 5: Cantidad total de países por categoría de ingreso

## Preprocesamiento y limpieza 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.
   1. Estos archivos se convirtieron a CSV para facilitar su manipulación[[2]](#footnote-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:
   1. Catálogo de indicadores: Con su código, nombre, descripción y ruta de acceso al archivo de origen dentro del repositorio de código.
   2. Catálogo de indicadores por área clave: Cada una de las áreas clave junto con sus indicadores y su dirección url.
   3. 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.

Cálculo de Métricas: Dada la cantidad de archivos e indicadores, el análisis exploratorio se hace extremadamente complejo y costoso. Es por esto se automatizó el proceso de cálculo de métricas. Esto permitió analizar las cantidades de datos faltantes para los distintos experimentos.

## Análisis exploratorio de datos (AED)

### Datos faltantes

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

Como puede verse en la Figura 6, el volumen de datos hace dificultosa la tarea de visualizarlos todos juntos.

A close-up of a screen

Description automatically generated

Figura 6: Heatmap de datos faltantes. En el eje de las x se visualizan los años, mientras que en el de las y se ven los indicadores. Las cajas del heatmap presentan la cantidad de datos que tienen valores distintos de nulo. Se removieron las etiquetas para facilitar la comprensión. Aun sin etiquetas puede apreciarse la alta taza de datos faltantes

En la Figura 5 puede verse un recorte de los datos a un conjunto de países Latinoamérica: Argentina, Bolivia, Chile, Colombia, Ecuador, México y Uruguay, pero esta vez agrupando los indicadores en las áreas antes descritas.

Aquí 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.

A chart with numbers and symbols

Description automatically generated with medium confidence

Figura 7: Heatmap con los datos faltantes para un conjunto de países seleccionados agrupando los indicadores por área.

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

A chart with numbers and symbols

Description automatically generated with medium confidence Figura 8: Datos faltantes para 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.

A graph with colored lines

Description automatically generated

Figura 9: Cantidad de valores no nulos por año para el área de Economy & Growth

Si bien hay una amplia cantidad de datos faltante, es posible plantear estrategias de mitigación que permitan llevar adelante el análisis propuesto en este trabajo.

### Categorización de los países (GNI per cápita)

En la siguiente figura pueden verse todos los indicadores relacionados con el GNI y su taza de nulos. En particular, nos enfocamos en el NY.GNP.MKTP.CD, que refleja el GNI total de los países. En este caso, se observa que para el 2022 hay 29 datos faltantes. Esto no invalida el análisis posterior, pero de alguna manera limita los resultados.

A graph showing different colored squares

Description automatically generated

Como se vio anteriormente, la categorización propuesta por World Bank muestra una gran cantidad de países en las categorías más altas y muy baja cantidad de países en la categoría de menor ingreso.

A graph with a line and a line

Description automatically generated with medium confidence

Figura 10: Boxplots de los indicadores asociados al GNI para el año 2022: (a) GNI total, (b) Población total, (c) GNI per cápita. Los potenciales outliers que se ven en el grafico son: (1) Bermuda, (2) Noruega, y (3) Suiza. Mientras que el mínimo es Burundi.

A graph of a number of blue and green bars

Description automatically generatedFigura 11: Distribución de los paises en función de su grupo de ingreso. Las líneas amarilla, verde y roja marcan los límites a partir de los cual los países se consideran Lower Middle Income, Upper middle income y High Income respectivamente.

En la siguiente grafica puede verse la distribución del GNI per cápita para el año 2022 según lo categoriza el Banco Mundial. Cabe destacar que es notoria la cantidad de países de ingresos altos, mientras que los de ingresos bajos parecen mezclarse entre sí.

A chart of different colored dots

Description automatically generated

Figura 12: Scatterplot que muestra el GNI per cápita para los países según la categorización del Banco Mundial

Haciendo un cálculo de la inercia y el coeficiente de silhouette para estos Clústeres, el resultado obtenido es: 796 para la inercia y 0.26 para el coeficiente de silhouette. Esto sugiere que los clústeres están dispersos y muy cercanos entre sí.

Por otro lado, el grupo High income tiene una media de GNI de 41120.30, mientras que la categoría siguiente (Upper Middle income) es de 7973, lo cual representa 1.453505 desviaciones estándar de diferencia. Los grupos Low income, Lower middle income y Upper middle income están más cerca entre sí en términos de GNI per capita, con distancias cercanas a cero desviaciones estándar entre sí.

Utilizando kmeans con 4 grupos sobre el GNI per cáputa, se logra una clasificacion de mejor calidad, con un coeficiente de silhouette de ~0.7 y una inercia de 12.72. En este caso, Argentina queda ubicada dentro de la categoria mas baja de ingresos ( Figura 13). Asimismo, aplicando el mismo algoritmo con 7 clusters, se logra una clasificación que tiene ~0.6 de coeficiente de silhouette y ~3.5 de inercia (Figura 14).

A chart of different colored dots

Description automatically generated

Figura 13: Países categorizados según su ingreso de acuerdo con el criterio del Banco Mundial

A chart of different colored dots

Description automatically generated

Figura 14: Scatterplot que muestra la distribución de los países en función en 7 clústeres

### 

### Métodos descriptivos

Nuestro análisis consta 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.

## Descripción de las técnicas de análisis y, si corresponde, de modelado

## Descripción de la selección de características

## Descripción de las métricas de evaluación de los modelos

## Descripción de los métodos estadísticos utilizados

## 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

# Resultados y discusión:

* 1. Presentación y análisis de resultados obtenidos
  2. Discusión de los resultados y su relevancia
  3. Limitaciones y posibles mejoras

# Conclusión:

* 1. Resumen de los hallazgos principales
  2. Conclusiones generales y su relación con los objetivos del trabajo
  3. Recomendaciones para futuros trabajos

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](https://www.tandfonline.com/doi/abs/10.1080/17565529.2021.1873724). [The Journal of International Trade & Economic Development](https://www.tandfonline.com/doi/abs/10.1080/17565529.2021.1873724).

[2] Alan M. Taylor (1994).[Three Phases of Argentine Economic Growth](https://www.nber.org/papers/h0060). National Bureau of economic research.

[3] Robert J. Barro (1996). [Determinants of Economic Growth: A Cross-Country Empirical Study](https://www.nber.org/papers/w5698). National Bureau of economic research.

[4] Daniel Landau (1986). [Government and Economic Growth in the Less Developed Countries: An Empirical Study for 1960-1980](https://www.journals.uchicago.edu/doi/abs/10.1086/451572). Universidad de Connecticut

[5] Michael Timberlake, Jeffrey Kentor (1986). [Economic Dependence, Overurbanization, and Economic Growth: A Study of Less Developed Countries](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1533-8525.1983.tb00715.x). 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](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4991.1970.tb00759.x). The review of income and wealth.

[7] World Bank Open Data, [Free and open access to global development data](https://data.worldbank.org/).

1. En todo caso, tomara prioridad para el análisis la completitud de datos de Argentina, que es el país donde se enfoca este trabajo. [↑](#footnote-ref-1)
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) [↑](#footnote-ref-2)