

Clasificación de Países Basada en Indicadores Socioeconómicos: Un Enfoque Basado en Datos



Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del
Conocimiento

Trabajo de Especialización

Autor: Eduardo Miguel Kiszkurno

Resumen

El presente trabajo tiene como objetivo clasificar países en función de indicadores socioeconómicos utilizando técnicas de machine learning. Se analizan datos del Banco Mundial, principalmente relacionados con el Ingreso Nacional Bruto (GNI) y otros indicadores de desarrollo económico y laboral. A través de técnicas de clustering como K-means y Análisis de Componentes Principales (PCA), se busca identificar patrones en los datos que permitan ofrecer una clasificación alternativa a la del Banco Mundial. El enfoque inicial se basa en el GNI per cápita, seguido de la incorporación de indicadores educativos y laborales para refinar los grupos formados.

Los resultados muestran que una clasificación basada únicamente en el GNI mejora la coherencia interna respecto a la categorización oficial, pero al incorporar indicadores educativos se logra una visión más compleja y matizada de las dinámicas socioeconómicas. Este estudio concluye que las técnicas de agrupamiento permiten una nueva perspectiva en la comprensión de las diferencias socioeconómicas entre países y destacan la importancia de considerar tanto variables económicas como educativas para obtener una clasificación más precisa.

Asimismo, se identifican patrones relevantes al integrar datos laborales, como la participación femenina en la fuerza laboral y las tasas de desempleo, lo que permite una visión más detallada de los factores que influyen en el desarrollo de los países. Las limitaciones relacionadas con la disponibilidad de datos también son abordadas, sugiriendo que futuros estudios deberían explorar la integración de más indicadores y nuevas técnicas de análisis para mejorar la robustez de las conclusiones.

Tabla de Contenidos

| | |
|--|-----------|
| INTRODUCCIÓN..... | 4 |
| METODOLOGÍA | 6 |
| FUENTES DE DATOS | 6 |
| HERRAMIENTAS Y LIBRERÍAS | 6 |
| PREPROCESAMIENTO Y LIMPIEZA DE LOS DATOS..... | 6 |
| SELECCIÓN DE FEATURES | 7 |
| ANÁLISIS DE INDICADORES..... | 7 |
| PREDICCIÓN DEL ÍNDICE DE INGRESO NACIONAL BRUTO (GNI) | 8 |
| VISUALIZACIÓN DE LOS CLUSTERS..... | 8 |
| ANÁLISIS EXPLORATORIO DE DATOS | 9 |
| DESCRIPCIÓN DE LOS DATOS | 9 |
| ANÁLISIS DE DATOS FALTANTES | 12 |
| EVALUACIÓN DE LA CATEGORIZACIÓN DEL BANCO MUNDIAL | 16 |
| CLUSTERING SOBRE EL GNI PER CÁPITA | 19 |
| CLUSTERING SOBRE INDICADORES DE EDUCACIÓN | 22 |
| RESULTADOS Y DISCUSIÓN..... | 26 |
| PRESENTACIÓN Y ANÁLISIS DE RESULTADOS OBTENIDOS | 26 |
| DISCUSIÓN DE LOS RESULTADOS Y SU RELEVANCIA | 27 |
| LIMITACIONES Y POSIBLES MEJoras | 27 |
| CONCLUSIÓN | 28 |
| RESUMEN DE LOS HALLAZGOS PRINCIPALES | 28 |
| CONCLUSIONES GENERALES Y SU RELACIÓN CON LOS OBJETIVOS DEL TRABAJO | 28 |
| RECOMENDACIONES PARA FUTUROS TRABAJOS | 28 |
| REFERENCIAS BIBLIOGRÁFICAS | 31 |
| ANEXOS..... | 32 |
| ANEXO 1: CÓDIGO FUENTE UTILIZADO EN EL ANÁLISIS..... | 32 |
| ANEXO 2: DESCRIPCIÓN DE ÁREAS CLAVE | 32 |
| ANEXO 3: MOTIVACIÓN PARA EL RECORTE DE DATOS | 33 |

Introducción

El desarrollo socioeconómico de los países está determinado por una interacción compleja de diversos factores. Este trabajo busca, mediante el uso de técnicas de aprendizaje automático, identificar y comparar estos indicadores y su influencia sobre el desarrollo económico; y de esta manera, entender las diferencias de Argentina con otros países.

El corpus de estudios en la materia es cuantioso y muy variado tanto en el enfoque, como en las técnicas de análisis empleadas.

El caso argentino es estudiado a lo largo de todo el mundo por sus particularidades. Naguib [21] 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 [2] hace un estudio histórico de nuestra performance económica caracterizando las siguientes etapas: pre-1913, 1913-1930s y 1930s-1950s.

El estudio de Landau [4] examina la relación entre la proporción del gasto de consumo del gobierno en el PIB y la tasa de crecimiento del PIB real per cápita en más de 100 países. Este trabajo es relevante para nuestra investigación ya que proporciona una base metodológica sólida para analizar cómo los diferentes tipos de gasto gubernamental pueden impactar en el crecimiento económico. En nuestro estudio, entender estos efectos puede ayudar a identificar factores promotores o inhibidores del crecimiento económico.

Weisskoff [6] investiga si el crecimiento económico en países en desarrollo ha conducido a inequidades en la distribución del ingreso, enfocándose en Puerto Rico, Argentina y México. Este análisis es fundamental para entender las dinámicas de distribución del ingreso en el contexto del crecimiento económico, un tema relevante en nuestro estudio. Weisskoff concluye que, aunque el crecimiento económico puede aumentar la riqueza nacional, a menudo exacerba las desigualdades de ingresos.

Timberlake y Kentor [5] exploran los determinantes estructurales de la urbanización periférica y sus efectos en el desarrollo nacional. Este trabajo aporta una perspectiva sobre cómo la urbanización puede influir en el desarrollo económico y social, un aspecto relevante para nuestro análisis. El estudio resalta que la urbanización no planificada puede llevar a problemas económicos y sociales significativos, lo que subraya la importancia de políticas urbanas bien diseñadas para promover un desarrollo equilibrado. Esto es particularmente pertinente para Argentina, donde la urbanización ha tenido impactos mixtos en el desarrollo regional.

El trabajo de Newfarmer y Mueller [22] analiza el poder económico y no económico de las corporaciones multinacionales en Brasil y México. Este estudio es relevante ya que muestra cómo la inversión extranjera directa puede afectar las economías locales, tanto positivamente como negativamente. Este análisis es útil para entender el impacto de la inversión extranjera en el crecimiento económico y cómo puede ser gestionada para maximizar sus beneficios y minimizar sus efectos adversos. En nuestra investigación, evaluar el papel de la inversión extranjera puede servir para comprender las dinámicas de crecimiento en un contexto globalizado.

Técnicas de Clustering (K-means)

El clustering es una técnica de aprendizaje no supervisado que se utiliza para agrupar un conjunto de objetos de manera que los elementos dentro de un mismo grupo (o clúster) sean más similares entre sí que con los de otros grupos. Las técnicas de clustering son fundamentales en el análisis exploratorio de datos, y a menudo permiten descubrir patrones y estructuras ocultas al ojo humano.

Una de las técnicas de clustering más utilizadas es K-means. Este algoritmo partitiona los datos en K clústeres, donde cada punto pertenece al clúster con el centroide más cercano. El objetivo de K-means es minimizar la variación dentro de cada clúster. Los centroides de los clústeres se actualizan iterativamente hasta que los cambios sean mínimos. Este método es especialmente útil por su simplicidad y eficiencia, aunque puede verse afectado por la elección inicial de los centroides y es sensible a los valores atípicos.

Análisis de series temporales (Arima)

ARIMA (AutoRegressive Integrated Moving Average) es una clase de modelos utilizada para analizar y predecir series temporales. Combina tres componentes principales: autorregresivo (AR), diferenciación (I) y promedio móvil (MA). El componente AR utiliza la relación entre una observación y un número de observaciones rezagadas anteriores. El componente I se aplica para hacer que la serie sea estacionaria, eliminando tendencias y estacionalidades. Finalmente, el componente MA modela el error de la predicción como una combinación lineal de errores pasados.

Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que se utiliza para transformar un conjunto de datos de alta dimensión en un espacio de menor dimensión. Este proceso se lleva a cabo identificando las direcciones (componentes principales) en las que varía más el conjunto de datos. PCA permite conservar la mayor cantidad posible de varianza en los datos originales mientras reduce el número de dimensiones, lo que facilita la visualización y el análisis. Los componentes principales son ortogonales entre sí y están ordenados de manera que el primer componente principal captura la mayor cantidad de varianza, el segundo componente captura la segunda mayor cantidad de varianza, y así sucesivamente.

Objetivos del trabajo

Este estudio tiene como objetivo principal desarrollar una clasificación alternativa de los países, integrando indicadores económicos y educativos, con el fin de identificar diferencias significativas respecto a la categorización proporcionada por el Banco Mundial. Para ello, se tomaron como referencia naciones con características socioeconómicas comparables a las de Argentina, tanto dentro de América Latina como de otras regiones. El propósito es analizar cómo la inclusión de indicadores educativos puede influir en el posicionamiento de Argentina en relación con países más desarrollados, generando una clasificación que ofrezca una perspectiva más matizada de la realidad socioeconómica global.

Desde un enfoque práctico, se busca aportar información útil para la formulación de políticas y estudios futuros sobre el crecimiento económico y social en Argentina. En este sentido, se analizarán variables clave como el Producto Interno Bruto (PIB), las tasas de empleo y desempleo, y el nivel de inversión en educación, evaluando su impacto en la clasificación obtenida y contrastándola con la categorización oficial del Banco Mundial.

Metodología

Fuentes de datos

Para este trabajo se utilizó un dataset publicado por el World Bank [7]. Los datos se encuentran disponibles para su descarga en el siguiente [enlace](#).

Con 189 países miembros y oficinas en más de 130 ubicaciones, el Banco Mundial es una asociación global que trabaja para encontrar soluciones sostenibles que reduzcan la pobreza y fomenten la prosperidad compartida en los países en desarrollo.

Hay un total de 1,463 indicadores numéricos de los países miembro sobre diversos aspectos, como Educación; Ciencia y Tecnología; y Crecimiento Económico entre otros. Estos indicadores se organizan en 20 áreas clave (ver Tabla 2 en los anexos de este documento) y cubren el período de 1960 a 2022.

Herramientas y librerías

Tanto para la descarga y preprocesamiento de los datos, como para el análisis exploratorio, y el modelado, utilizamos Python 3.9.7 [23] corriendo sobre la IDE Visual Studio Code Version 1.90.0 [24].

Dentro de Python [23] utilizamos las siguientes bibliotecas y librerías:

- Selenium webdriver [8]: Se utiliza para automatizar la descarga de datos desde la web.
- OpenPyXL [9]: Se utiliza para manipular los archivos Excel durante el preprocesamiento de datos.
- Pandas [10] y NumPy [11]: Se utilizan para manipulación y análisis de datos.
- Matplotlib [12], Seaborn [13], Holoviews [14] y bokeh [15]: Se utilizan para la visualización de datos.
- PDFPages [16] de Matplotlib: Para generar reportes en PDF con gráficos de los análisis realizados.
- Scikit-learn [17] : Se utiliza para aplicar las técnicas de clustering en la etapa de modelado.
 - StandardScaler
 - KMeans
 - silhouette_score
- Statsmodels [18]: Se utiliza para el análisis de series temporales, incluyendo modelos ARIMA.
- Jupyter Notebooks [19]: 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 [23], 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 y limpieza de los datos

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.

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 Webdriver [8], se realizaron las descargas automáticas de todos los archivos en formato xls.

- a. Estos archivos se convirtieron a CSV para facilitar su manipulación¹
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 se automatizó el proceso de cálculo de métricas relacionadas principalmente con los datos faltantes. Esto permitió buscar estrategias de recorte de los datos (seleccionando los features que menos datos nulos tuviesen).

Selección de features

En la sección Análisis exploratorio de datos se describen los detalles de las actividades de análisis desarrolladas sobre el dataset. Pero cabe destacar que, dada la complejidad para imputar los datos faltantes, el criterio principal para la selección de los features (indicadores) a utilizar durante el trabajo fue que la cantidad de nulos sea la menor posible.

La nulidad de los datos juega un rol fundamental, ya que los métodos utilizados no soportan el uso de valores nulos.

Análisis de Indicadores

En nuestro análisis, hemos utilizado K-means para clasificar y caracterizar los países en función de los indicadores socioeconómicos disponibles. Comenzamos con un experimento inicial utilizando el Ingreso per cápita para el año 2022, partiendo de la clasificación en cuatro categorías de ingreso según lo define el Banco Mundial. Este análisis inicial nos permitió establecer una línea de base para la clasificación de los países.

Posteriormente, realizamos varios experimentos adicionales en diferentes años y con distintos indicadores socioeconómicos para identificar diferencias significativas entre las nuevas categorizaciones y la realizada por el Banco Mundial. Esta exploración nos permitió observar cómo varían las características socioeconómicas de los países a lo largo del tiempo y cómo estas variaciones pueden influir en su clasificación.

Para evaluar la calidad de los agrupamientos obtenidos, tanto la coherencia interna de los grupos como su separación respecto a otros grupos, se utilizaron las métricas de Silhouette e Inercia.

El coeficiente de Silhouette mide qué tan similar es un elemento a los elementos de su propio grupo (cohesión) en comparación con los de otros grupos (separación). Este coeficiente se define para cada elemento como:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde:

- $a(i)$ es la distancia media entre el elemento i y los demás elementos de su mismo grupo.

¹ 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)

- $b(i)$ es la distancia media entre el elemento i y los elementos del grupo más cercano al que no pertenece.

El valor de $S(i)$ oscila entre -1 y 1, donde un valor cercano a 1 indica que el elemento está bien agrupado, mientras que valores cercanos a -1 sugieren que el elemento debería estar en otro grupo.

Por otro lado, la inercia mide la compacidad de los clústeres en términos de la distancia de los elementos al centroide del grupo al que pertenecen. La inercia total de un conjunto de datos se define como la suma de las distancias cuadradas de cada elemento a su centroide:

$$\text{Inercia} = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

Donde:

- C_k es el conjunto de elementos en el clúster k ,
- μ_k es el centroide del clúster k ,
- x_i es un elemento en el clúster k .

La inercia se utiliza para medir qué tan bien ajustados están los elementos a los centroides de sus respectivos clústeres; valores más bajos de inercia indican grupos más compactos.

Predicción del Índice de Ingreso Nacional Bruto (GNI)

En uno de los experimentos realizados, utilizamos el modelo ARIMA para predecir el Índice de Ingreso Nacional Bruto (GNI) de diferentes países. La serie temporal del GNI es crucial para entender las tendencias económicas y planificar políticas futuras. El supuesto subyacente es que predecir el GNI mediante ARIMA permitiría identificar patrones subyacentes, pero también facilitar la imputación de datos faltantes.

El proceso para aplicar ARIMA al GNI implica los siguientes pasos:

- **Visualización y análisis exploratorio de la serie temporal:** Comenzamos visualizando la serie temporal del GNI para identificar cualquier tendencia, estacionalidad o patrones.
- **Diferenciación para estacionarizar la serie:** Si la serie no es estacionaria, aplicamos diferenciación para eliminar la tendencia.
- **Selección de parámetros ARIMA (p, d, q):** Utilizamos métodos como el gráfico ACF (Autocorrelation Function) y PACF (Partial Autocorrelation Function) para identificar los valores adecuados de p (orden autorregresivo), d (diferenciación) y q (orden de promedio móvil).
- **Ajuste del modelo ARIMA:** Ajustamos el modelo ARIMA a la serie temporal del GNI utilizando los parámetros seleccionados.
- **Evaluación del modelo:** Evaluamos el modelo utilizando métricas de error y validación cruzada.
- **Predicción futura:** Utilizamos el modelo ajustado para hacer predicciones futuras del GNI.

Visualización de los clusters

Se utilizó PCA para visualizar la información de los clusters generados a partir de diversos subconjuntos de indicadores. Dado que los datos originales contienen múltiples indicadores como tasas de empleo, participación laboral femenina, Producto Nacional Bruto, entre otros, que pueden ser difíciles de analizar en un espacio multidimensional, PCA nos permite proyectar estos datos en un espacio bidimensional. Esta transformación facilita la visualización al condensar la información de múltiples variables en solo dos componentes principales que retienen la mayor parte de la variabilidad de los datos originales, lo cual ayuda a identificar patrones y relaciones entre los países dentro de los clusters.

Análisis exploratorio de datos

Descripción de los datos

El Banco Mundial categoriza sus indicadores en diversas áreas para proporcionar una visión integral del desarrollo socioeconómico de los países. Estas áreas abarcan desde la agricultura y el desarrollo rural, hasta la eficacia de la ayuda, el cambio climático y el crecimiento económico. Indicadores como la producción agrícola, las emisiones de CO₂ y el Producto Interno Bruto (PIB) son esenciales para evaluar el progreso y los desafíos en estos sectores. Además, incluye indicadores en sectores cruciales como la educación, la salud y la infraestructura, que miden aspectos como la tasa de matrícula en educación primaria, la esperanza de vida al nacer y el acceso a electricidad. Estos indicadores permiten un análisis detallado de las políticas y estrategias necesarias para fomentar un desarrollo sostenible y equitativo.

En términos de equidad y sostenibilidad, las áreas de género, medio ambiente y pobreza son particularmente relevantes. Los indicadores de participación laboral femenina, calidad del aire y tasa de pobreza internacional ayudan a comprender mejor las desigualdades y los impactos ambientales que afectan el desarrollo. Asimismo, las categorías de sector financiero, sector público y protección social y trabajo ofrecen datos sobre la estabilidad económica y la seguridad social, con indicadores como el crédito doméstico al sector privado, el gasto público como porcentaje del PIB y la tasa de desempleo. En conjunto, estos indicadores proporcionan un marco completo para evaluar el desarrollo económico y social de los países, permitiendo la formulación de políticas más efectivas y focalizadas.

Hay un total de 1,463 indicadores, organizados en 20 áreas clave (ver Tabla 2 en los anexos de este documento), 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](#).

Los indicadores se dividen en 20 categorías o áreas clave según se puede ver en la Figura 1.

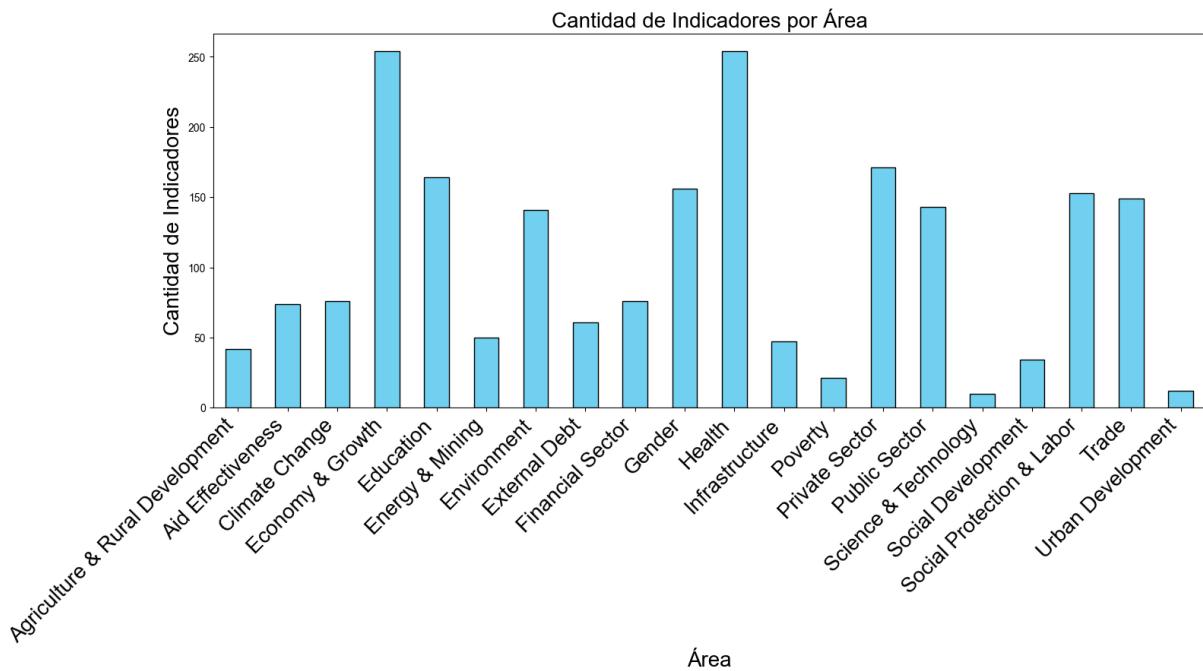


Figura 1: muestra las áreas clave y la cantidad de indicadores por área. Se observa que las áreas de 'Economy & Growth' y 'Health' tienen una cantidad significativamente mayor de indicadores en comparación con otras, lo que refleja la importancia de estos sectores en los estudios socioeconómicos globales.

Dentro del dataset, los países se categorizan por dos criterios: geográfico y económico. En el aspecto geográfico, el Banco Mundial distingue las siguientes regiones (Figura 2):

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

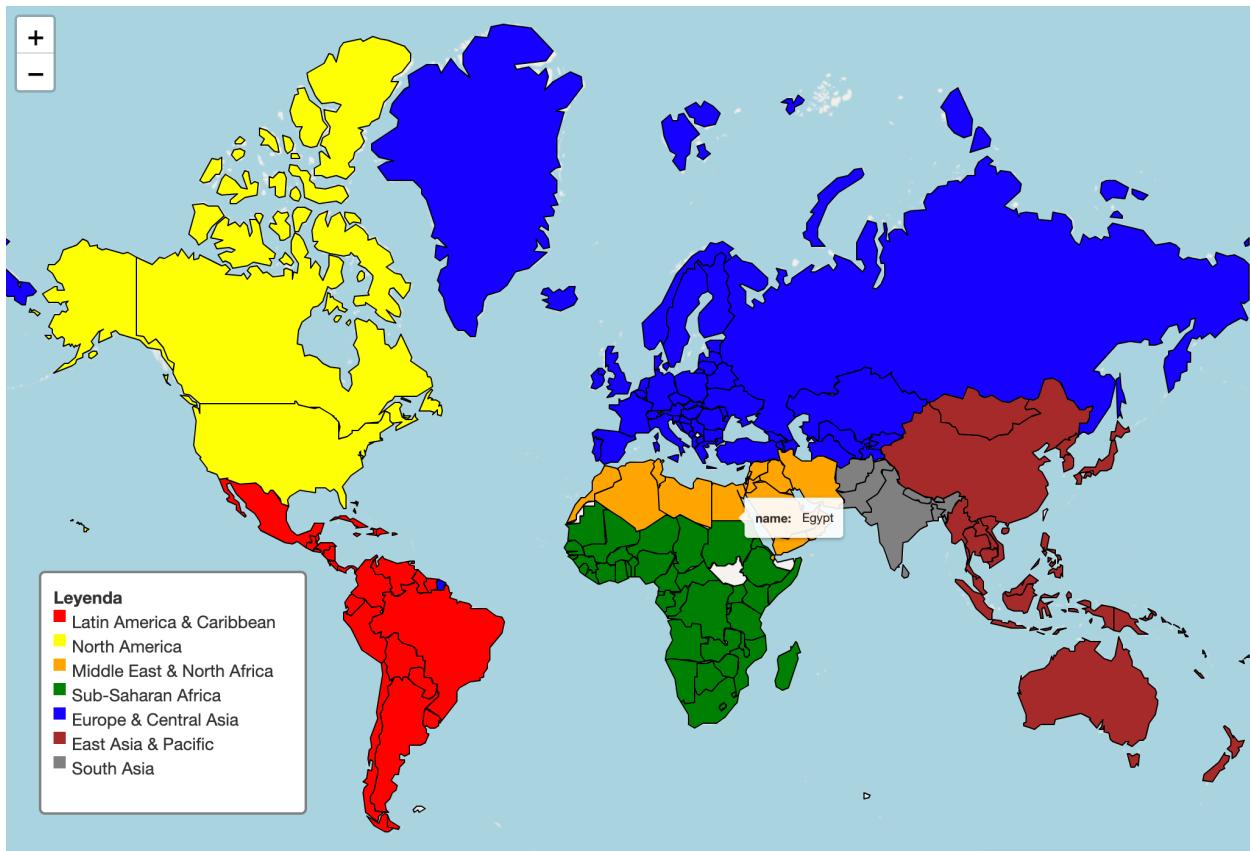


Figura 2: Mapa del mundo. Cada país se encuentra coloreado según su región definida en el dataset del Banco Mundial.

En cuanto al criterio económico, se 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”.

En la Figura 3 puede verse 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.

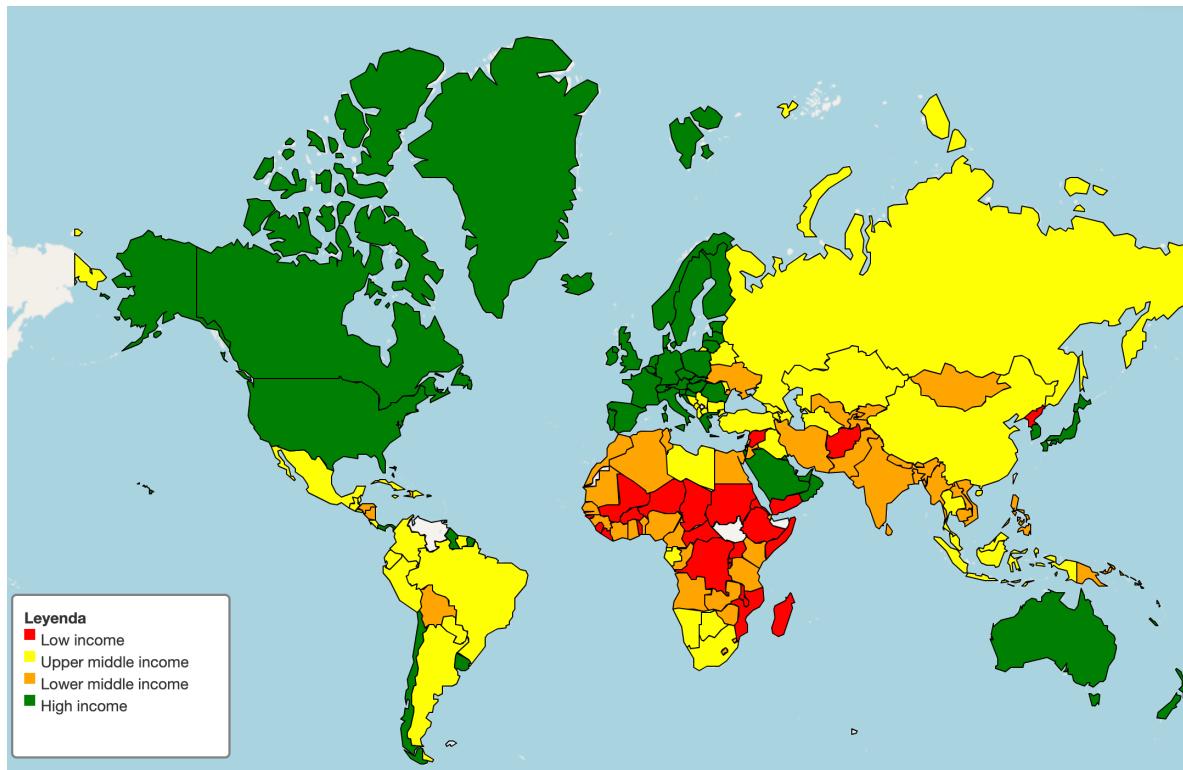


Figura 3: presenta el mapa mundial, coloreando a los países según su categoría de ingreso definida por el Banco Mundial. Este mapa es fundamental para visualizar cómo las naciones se agrupan según su Producto Nacional Bruto per cápita, lo que utilizamos para establecer una línea de base antes de aplicar nuestras propias categorizaciones a partir de K-means.

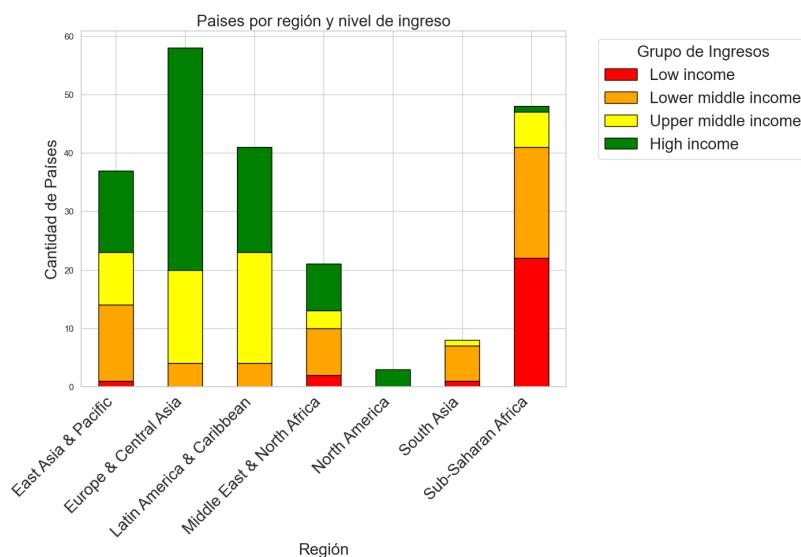


Figura 4: Muestra la cantidad de países en función de su categoría de ingreso agrupando por la región a la que pertenecen. Puede verse que la categorización utilizada por el Banco Mundial ubica a una gran cantidad de países dentro de la categoría “High Income”.

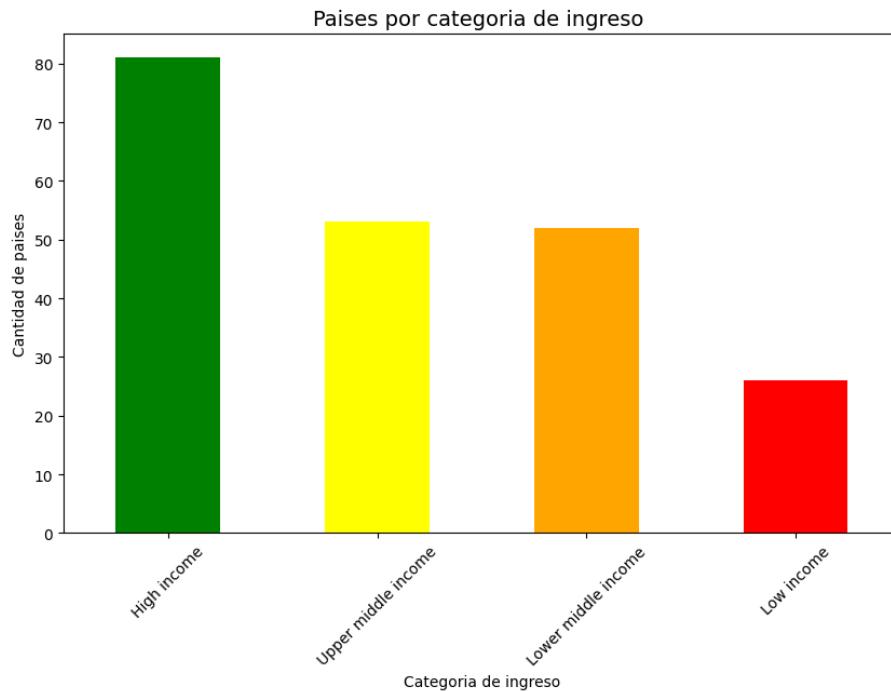


Figura 5: Cantidad países por categoría de ingreso. Se ve muy claramente, que la categoría "High Income" es la más poblada.

Análisis de 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.

En la Figura 6 puede verse un heatmap con los años en el eje de las X y los indicadores en el eje de las Y. Los nombres de los indicadores se ocultan para facilitar la visualización. Los valores del heatmap representan la tasa de valores nulos para un año y un indicador determinado. Un numero cercano a 0 indica mucha densidad de nulos, y un valor cercano a 1 indica que hay muchos datos no nulos.

Como puede apreciarse, la cantidad de nulos es muy alta en general.

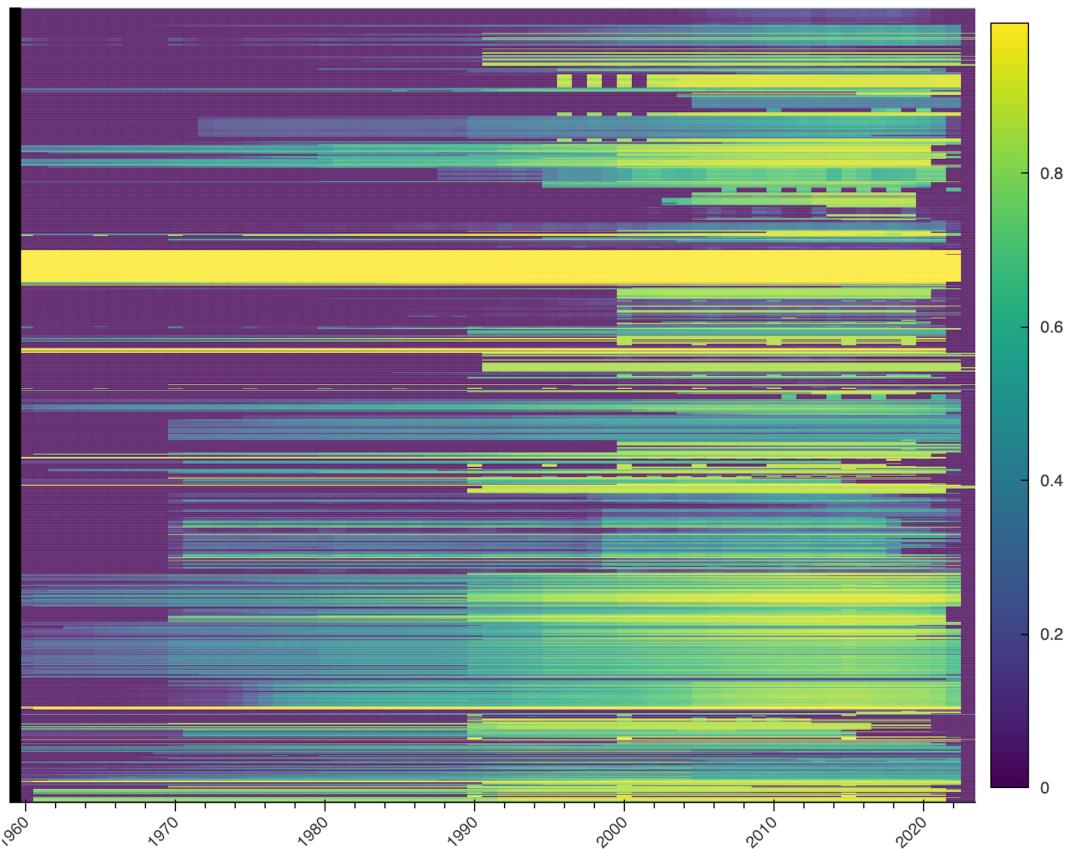


Figura 6: Heatmap de datos faltantes. En el eje de las x se visualizan los años, mientras que en el de las y se ven los indicadores. Las cajas del heatmap presentan la cantidad de datos que tienen valores distintos de nulo. Se removieron las etiquetas para facilitar la comprensión. Aun sin etiquetas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

Para este estudio, el indicador mas relevante es el GNI per capita, que es el utilizado por el Banco Mundial para agrupar a los distintos países en función de sus categorías de ingreso.

En particular, el NY.GNP.MKTP.CD, que segun la propia definición del Banco Mundial, representa el “GNI (current US\$)”, es el que utilizamos para generar los clusters que se describen mas adelante en este trabajo.

La Figura 7 muestra no solo el NY.GNP.MKTP.CD, sino todos los demás indicadores relacionados con el GNI y su tasa de nulos.

El NY.GNP.MKTP.CD es el que menos datos faltantes tiene junto con el NY.GNP.MKTP.CN, que el GNI en moneda local, lo cual invalida comparaciones entre países..

Como puede verse, a partir del año 1990 tiene tasas de nulos bastante razonables. En Particular, se observa que para el 2022 hay 29 datos faltantes². Esto no invalida el análisis posterior, pero de alguna manera limita los resultados.

² Esto es curioso, ya que estos 29 países faltantes se encuentran apropiadamente catalogados en la metadata provista por el mismo Banco Mundial.

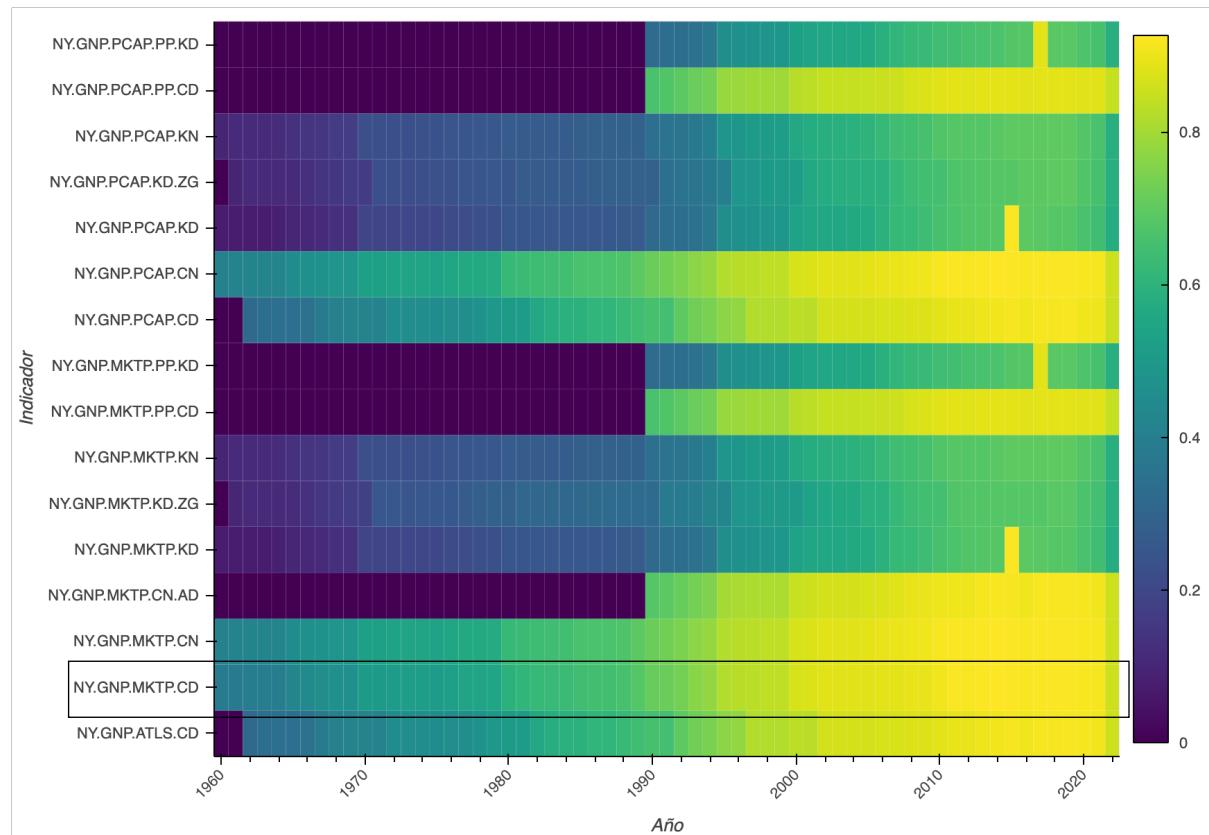


Figura 7: muestra un heatmap de los indicadores relacionados con el GNI y su tasa de datos faltantes distribuidos a lo largo de los años. Este gráfico destaca la dificultad de trabajar con datos incompletos, en particular a partir del año 1990, cuando la tasa de nulos disminuye pero sigue siendo significativa. A pesar de esta limitación, el análisis de los indicadores del GNI sigue siendo central en nuestro trabajo, ya que es la métrica principal para clasificar los países. El manejo de los datos faltantes mediante técnicas de imputación o recorte es clave para asegurar la calidad de los resultados.

Una de las componentes que nos propusimos incluir en el análisis es la educación de la población. Es interesante determinar cómo afecta la inversión en educación al desarrollo económico de una nación. No obstante, como puede verse en Figura 8, la cantidad de valores nulos que tienen estos indicadores hace imposible ejecutar cualquier algoritmo de clustering, ya que no se pueden eliminar ni siquiera imputar tantos valores.

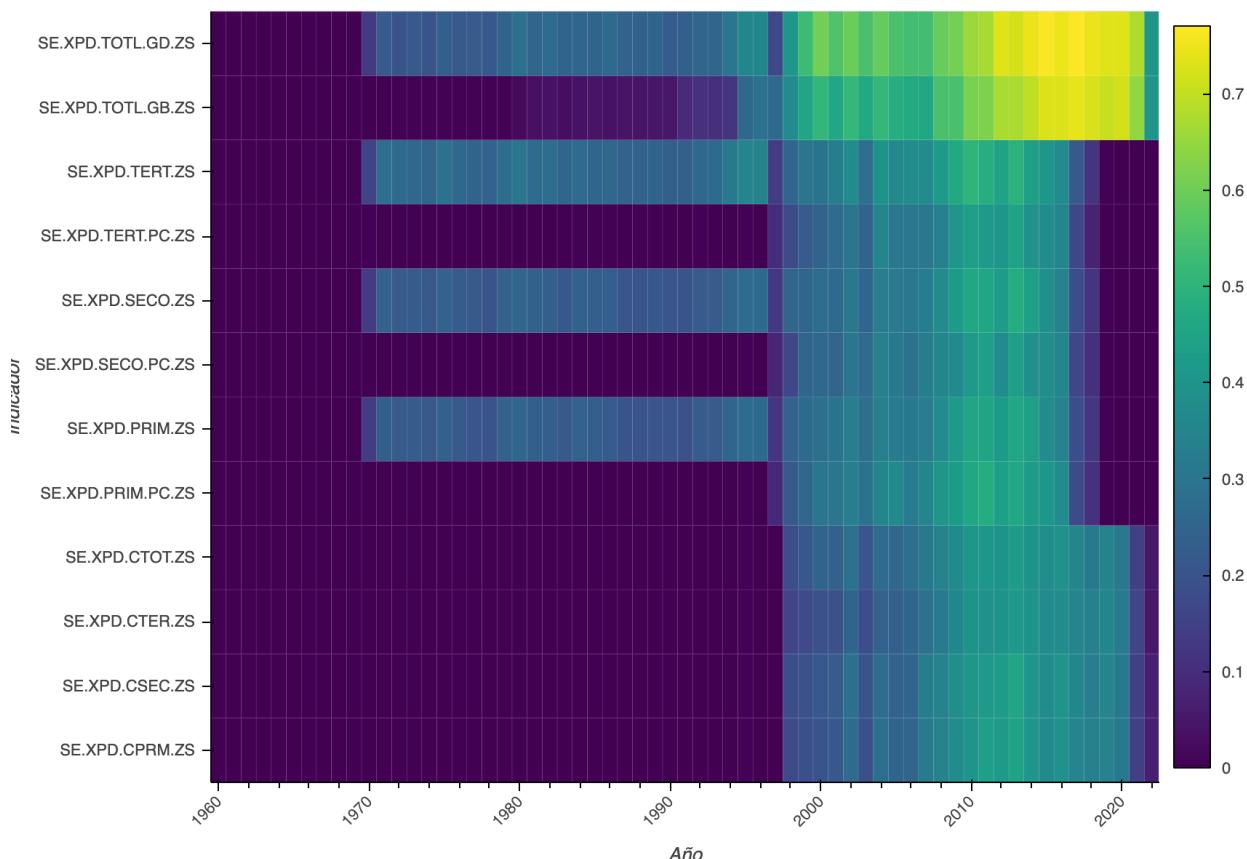


Figura 8: Heatmap que muestra la cantidad de datos faltantes por año para los distintos indicadores relacionados con gastos en Educación. El indicador SE.XPD.TOTL.GD.ZS' representa el porcentaje total del GDP en educación, el SE.XPD.TOTL.GB.ZS representa el gasto total en educación como porcentaje del gasto total del estado. La cantidad de datos faltantes hace imposible que podamos utilizar estos indicadores para nuestro estudio.

Descartados los gastos en educación, identificamos otro subconjunto de indicadores, también parte del área “Education”³ que son de utilidad y cuentan con una densidad de nulos aceptable:

SP.POP.TOTL (Población total): Representa el total de la población en un país o región.

SL.TLF.TOTL.IN (Fuerza laboral total): Representa el total de la fuerza laboral, incluyendo a todas las personas mayores de una edad específica que están disponibles para trabajar.

SL.UEM.TOTL.ZS (Desempleo total como porcentaje de la fuerza laboral total): Representa el porcentaje de la fuerza laboral total que está desempleada, según estimaciones modeladas por la Organización Internacional del Trabajo (OIT).

SL.TLF.TOTL.FE.ZS: (Fuerza laboral femenina, como porcentaje de la fuerza laboral total): Representa el porcentaje de mujeres dentro de la fuerza laboral total.

SL.UEM.TOTL.FE.ZS (Desempleo femenino, como % de la fuerza laboral femenina): Representa el porcentaje de mujeres dentro de la fuerza laboral femenina que está desempleada, según estimaciones modeladas por la Organización Internacional del Trabajo (OIT).

³ Estos datos son parte del área clave “Education”, aunque tal vez sería más apropiado catalogarlos dentro del área de “Economy & growth”.

SL.UEM.TOTL.MA.ZS (Desempleo masculino, como porcentaje de la fuerza laboral masculina): Representa el porcentaje de hombres dentro de la fuerza laboral masculina que está desempleada, según estimaciones modeladas por la Organización Internacional del Trabajo (OIT).

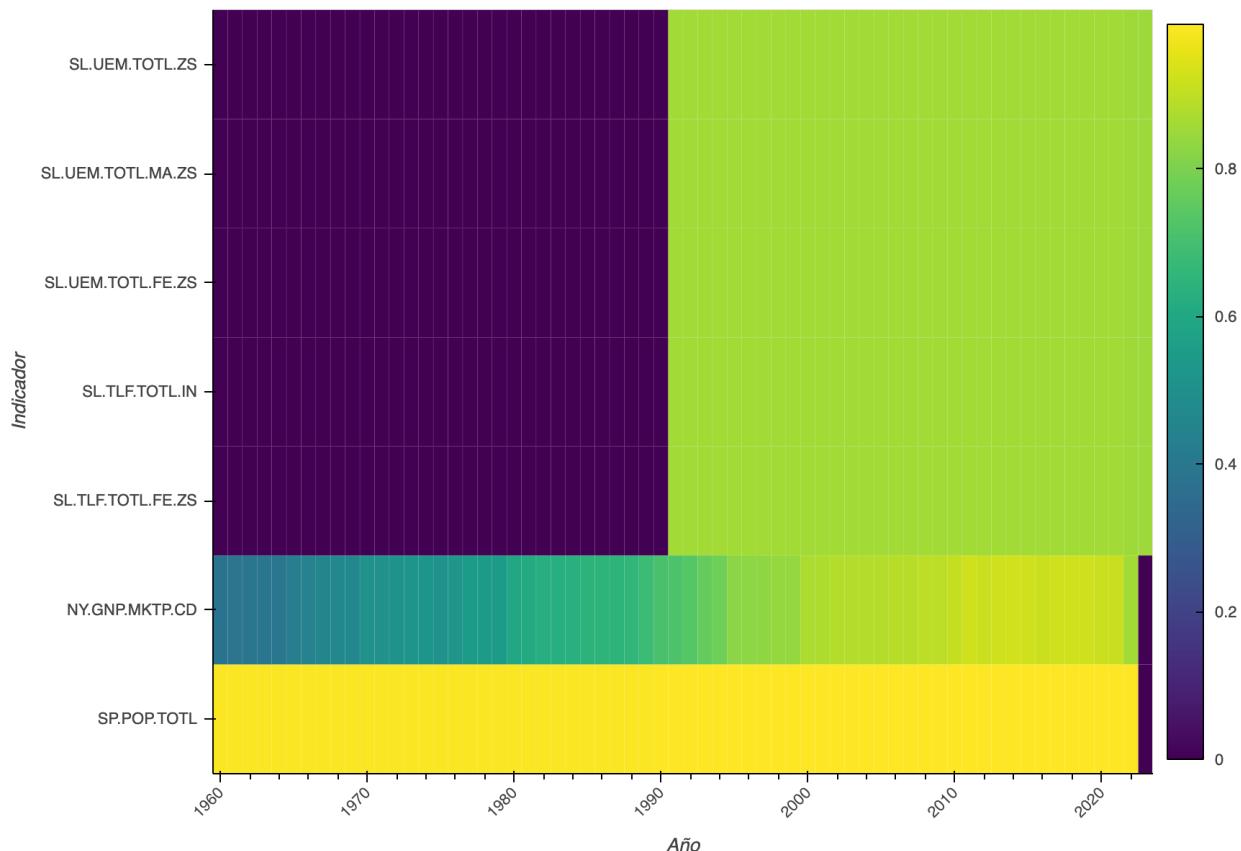


Figura 9: Heatmap con los indicadores relacionados con Educación y el índice GNI y su tasa de nulos distribuidos a lo largo de los años. Puede verse que los indicadores seleccionados para el análisis tienen mayores tasas de no nulos a partir de 1990. Este es el motivo principal por el cual se eligieron dichos indicadores.

Con esta lista de indicadores es factible utilizar un algoritmo de clustering para lograr una clasificación de los países.

Evaluación de la categorización del Banco Mundial

La categorización propuesta por World Bank descrita anteriormente muestra una inusual cantidad de países en la categoría más alta (~80 "High Income") y una muy baja cantidad de países en la categoría de menor ingreso (~30 "Low Income"). Mientras que en las categorías del medio se acumulan las restantes (~50 en cada una). Esto es de alguna manera anti-intuitivo.

El Boxplot de la Figura 10 se identifican los países de mayores ingresos, en primer lugar Bermuda, seguido por Noruega y Suiza.

A los efectos de este análisis, parece claro que utilizar sólo 4 categorías es poco para describir la variedad de contextos. Y tal vez, aquellos países que se ven en el cuartil superior deberían tener su propia categoría.

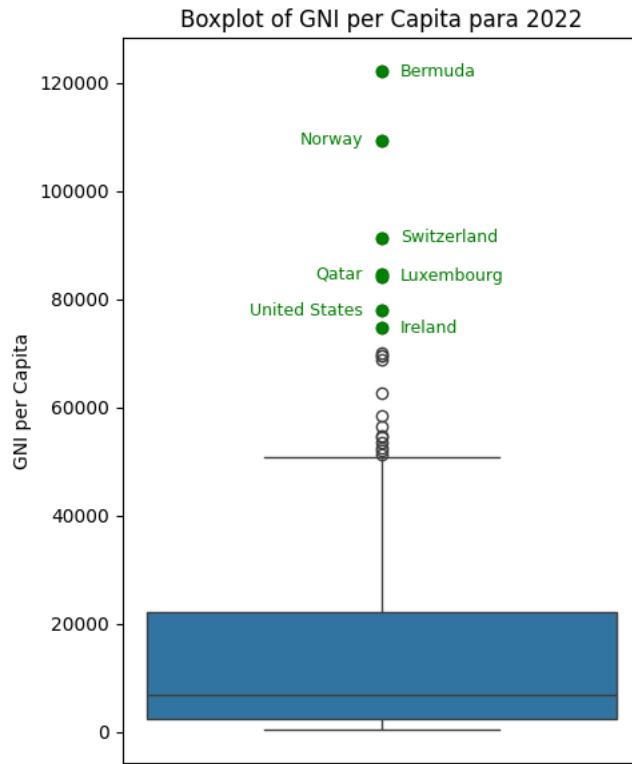


Figura 10: presenta boxplots de los indicadores asociados al GNI para el año 2022, detallando las distribuciones de (a) GNI total, (b) población total, y (c) GNI per cápita. Los países de mayores ingresos se destacan claramente del resto, con Bermude, Noruega y Suiza liderando el grupo de altos ingresos. Este gráfico pone de relieve las diferencias significativas entre países de ingresos bajos y altos. Argentina, situada en un punto intermedio entre 'Upper middle income' y 'High income', refuerza su posición dentro de los países de desarrollo socioeconómico medio-alto. Esta representación ayuda a visualizar las disparidades globales, que posteriormente serán analizadas más en detalle mediante técnicas de clustering.

El boxplot ofrece una vista limitada, y algo sesgada, de la distribución real de los ingresos por países. No obstante, ayuda a clarificar la diferencia de magnitudes entre los primeros países y los últimos. Como complemento, la Figura 11 muestra la distribución de los países en términos de su GNI index.

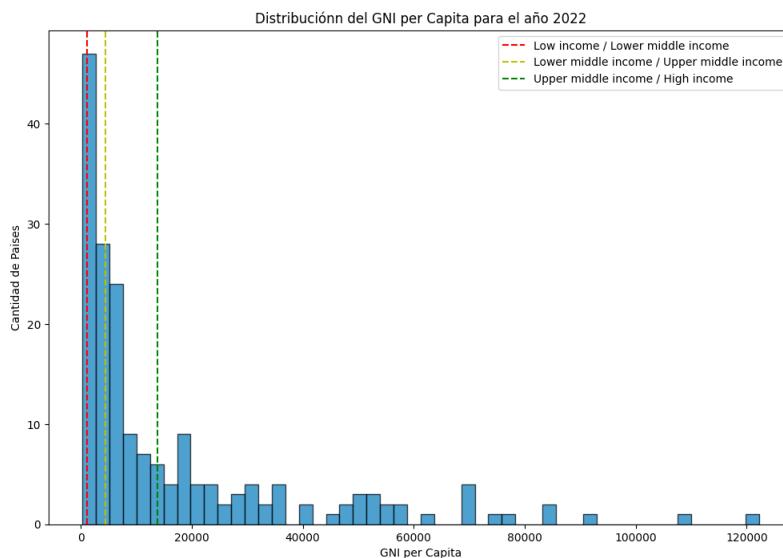


Figura 11: Distribución de los países en función de su grupo de ingreso. Las líneas amarilla, verde y roja marcan los límites a partir de los cuales los países se consideran Lower Middle Income, Upper middle income y High Income respectivamente.

La Figura 12 presenta la distribución del GNI per cápita para el año 2022, categorizada de acuerdo con los grupos de ingresos definidos por el Banco Mundial. La visualización se muestra en forma de scatter plot, donde los países están ordenados en el eje de las X por el índice GNI. Es importante destacar que, para simplificar la interpretación, no se muestran otros datos adicionales en este gráfico. Cada punto representa un país y se utiliza una codificación por colores para diferenciar los distintos grupos de ingresos: 'Low income', 'Lower middle income', 'Upper middle income' y 'High income'.

Es significativa la manera en que los países de la categoría más alta se destacan por encima del resto. Por otro lado, los países de ingresos bajos, indicados en rojo, parecen mezclarse más entre sí, mostrando menos dispersión y situándose en la parte inferior del gráfico. Los países de ingresos medios bajos y medios altos, en amarillo y naranja respectivamente, también muestran una distribución más compacta en comparación con los de ingresos altos.

El coeficiente de variación del dataset completo es de 132.14%. Lo cual indica que la variabilidad relativa del índice GNI es bastante alta en comparación con su media.

El grupo High income tiene una media de GNI de U\$S 41120.30 (con un CV de 59.58%), mientras que la categoría siguiente (Upper Middle income) es de U\$S 7973 (con un CV de 33.38%), lo cual representa 1.45 desviaciones estándar de diferencia. Los grupos Low income (CV de 30.61%) y Lower middle income (CV de 41.30%) están más cerca entre sí en términos de GNI per capita, con distancias cercanas a cero desviaciones estándar entre sí.

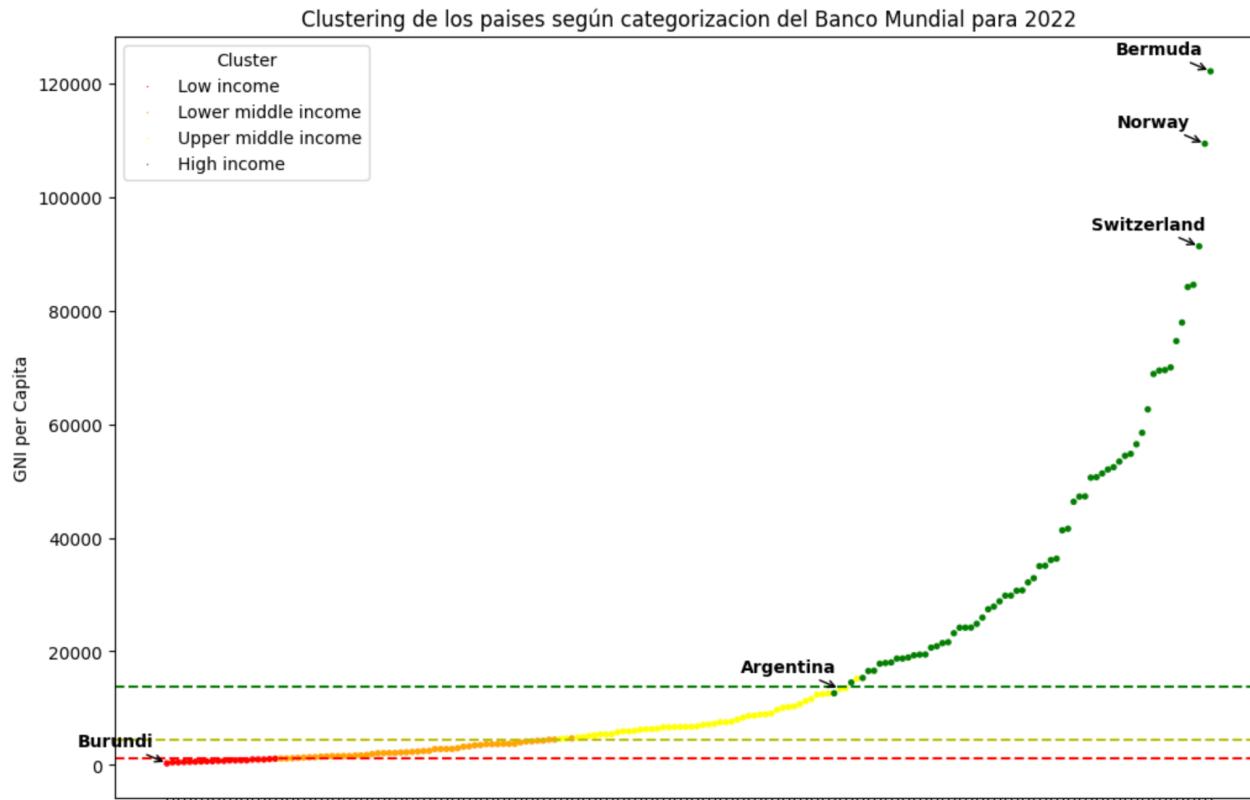


Figura 12: Muestra el GNI per cápita para los países según la categorización del Banco Mundial. En verde el grupo de más altos ingresos, con sus mayores valores en Bermuda, Noruega y Suiza, mientras que del lado de más bajos ingresos puede verse a Burundi. Argentina se sitúa en el límite entre Upper middle income y High income.

Haciendo un cálculo de la inercia y el coeficiente de Silhouette sobre los datos previamente escalados (restando la media y dividiendo por el desvío estándar) para estos Clústeres, se obtuvo: 796 para la inercia y 0.26 para el coeficiente de Silhouette. Esto sugiere que los clústeres están dispersos y muy cercanos entre sí y confirma lo que se advierte visualmente en la Figura 12.

Clustering sobre el GNI per cápita

Una primera clusterización utilizando K-means con 4 grupos sobre con el GNI per cápita como único feature, logra una clasificación de mejor calidad, con un coeficiente de silhouette de ~0.7 y una inercia de 12.72.

En esta categorización Argentina queda ubicada dentro de la categoría más 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).

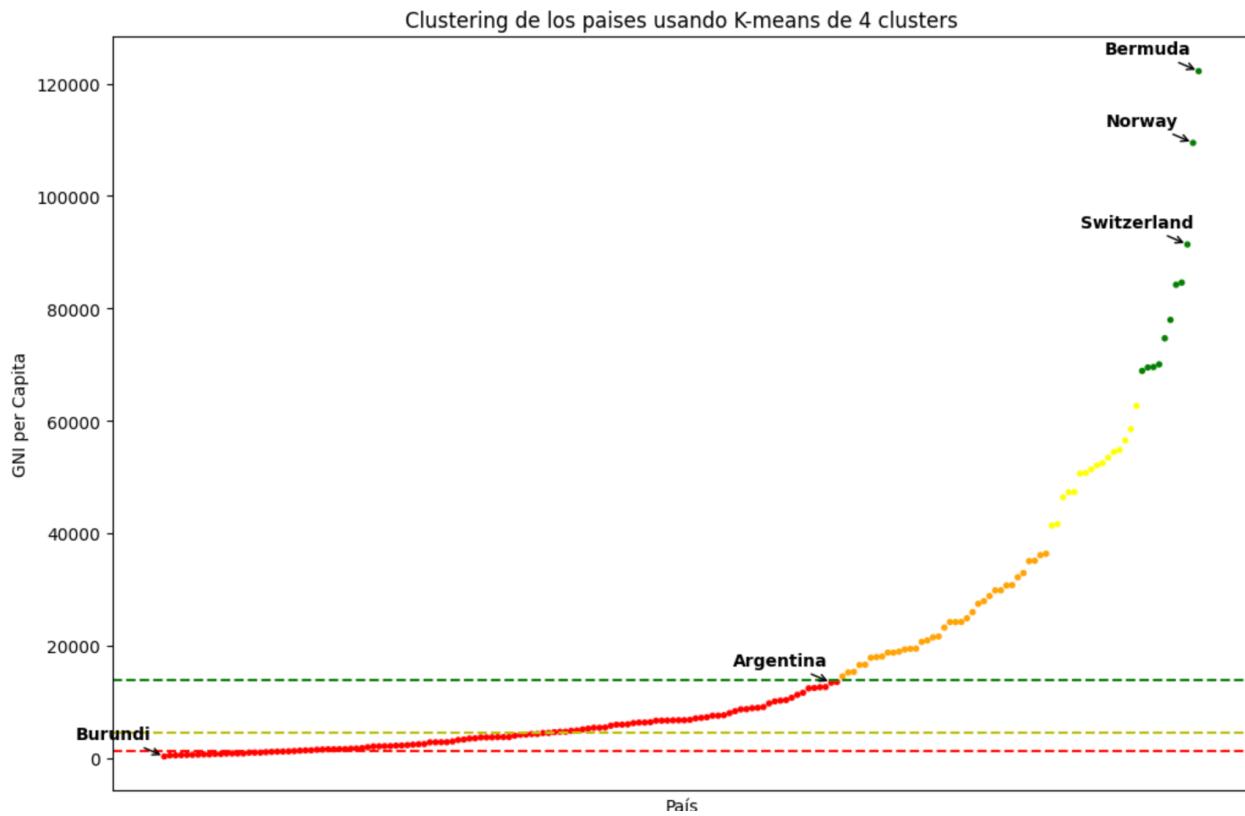


Figura 13: muestra el GNI per cápita para los países clasificados mediante K-means con 4 grupos, utilizando el GNI como único feature. Los países de mayores ingresos, como Bermuda y Suiza, se agrupan en un clúster de alta renta, mientras que Argentina aparece en el clúster de ingresos bajos. Este gráfico subraya cómo la clasificación mediante K-means puede diferir de la categorización tradicional del Banco Mundial, proporcionando una perspectiva alternativa basada exclusivamente en datos económicos.

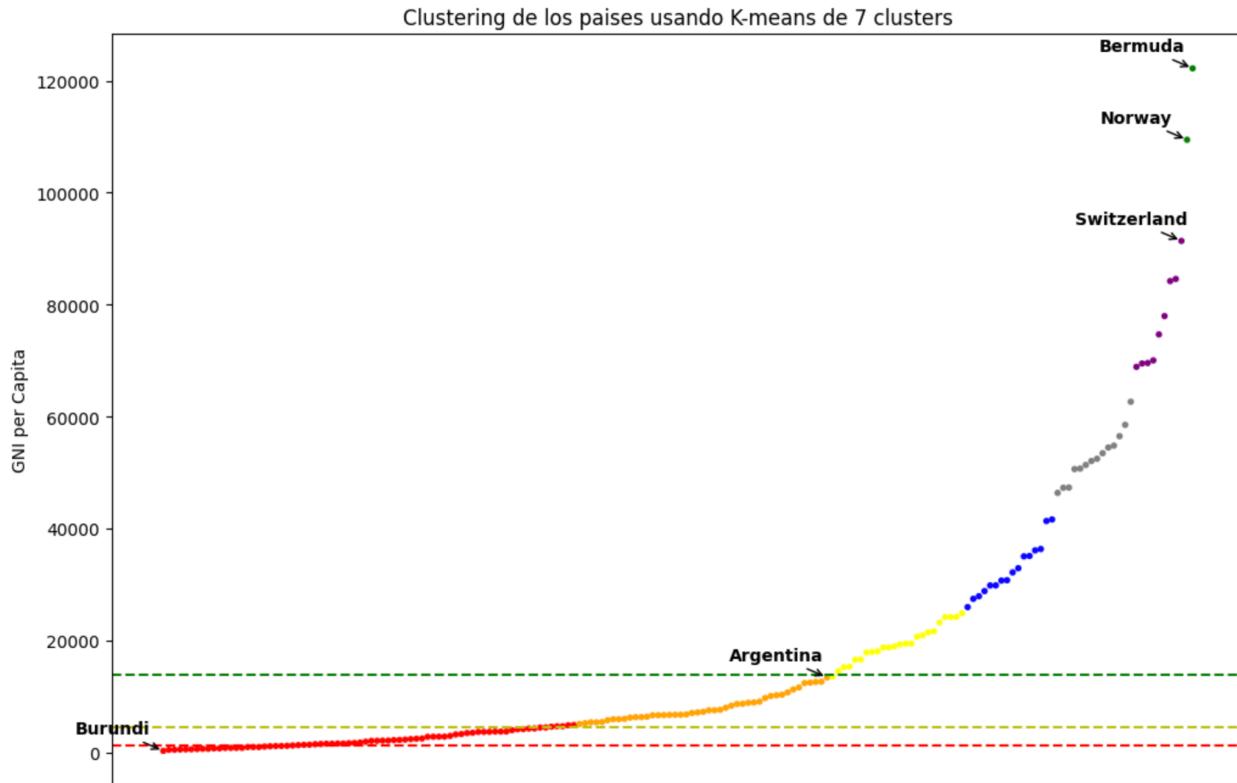


Figura 14: Muestra el GNI per cápita para los países categorizados utilizando K-means con 7 grupos sobre el GNI como único feature. En violeta el grupo de mas altos ingresos (que ahora cuenta únicamente con 2 elementos). Argentina se sitúa en el Cluster inmediatamente superior al de mas bajos ingresos.

La Figura 15 muestra las diferencias la clasificación del Banco Mundial y el algoritmo de Clustering. Muchos de los catalogados como High Income o Upper middle income por el Banco Mundial, en la nueva caracterización han pasado a las categorías de Low Income y Lower Middle Income.

Hay 29 países que tienen definida alguna categoría, pero no tienen datos relacionados con el GNI para 2022. Estos países, no tienen cluster asignado dentro de la nueva clasificación. En la Figura 15 se representan en la categoría “Indefinido”, que puede verse en la banda derecha del gráfico.

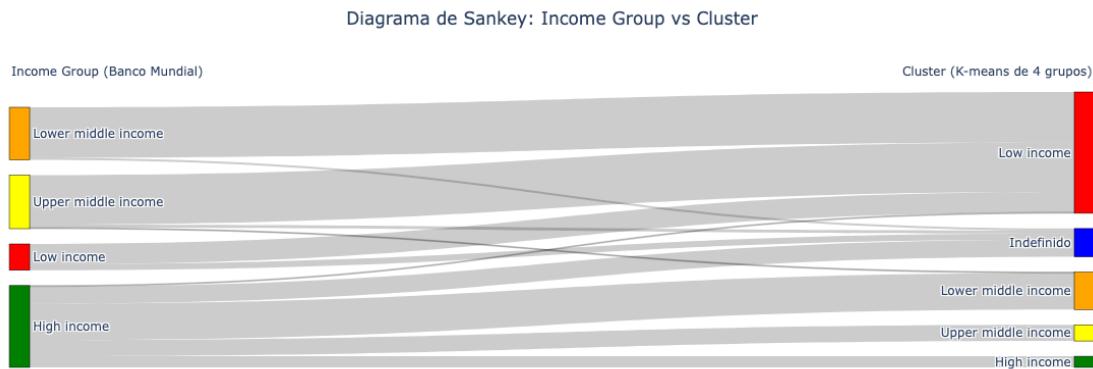


Figura 15: Diagrama de Sankey que muestra la forma en que se redistribuyen los países al aplicar K-means con 4 grupos sobre el GNI como único feature. En verde el grupo de mas altos ingresos. Argentina se sitúa en el Cluster de mas bajos ingresos.

Clustering sobre Indicadores de Educación

Utilizar únicamente el GNI ofrece una visión muy acotada. Es por esto que como parte de este trabajo decidimos incluir otros indicadores, en particular nos interesa analizar cómo afecta a esta clasificación el agregado de información referente a la Educación.

El área de Educación del Banco Mundial abarca una amplia gama de indicadores que proporcionan una visión detallada del estado y la evolución del sistema educativo en los países de todo el mundo. Estos indicadores se agrupan en varias categorías clave, tales como tasas de matrícula, deserción escolar, logro educativo y gasto en educación. Por ejemplo, las tasas ajustadas de matrícula neta en educación primaria (SE.PRM.TENR), desglosadas por género, permiten evaluar el acceso equitativo a la educación básica. Indicadores de deserción escolar en adolescentes (SE.SEC.UNER.LO.ZS) y en niños en edad primaria (SE.PRM.UNER.ZS) ofrecen datos cruciales para entender los desafíos en la retención de estudiantes. Además, los indicadores de duración de la educación obligatoria (SE.COM.DURS) y los gastos corrientes en educación primaria, secundaria y terciaria (SE.XPD.CPRM.ZS, SE.XPD.CSEC.ZS, SE.XPD.CTER.ZS) proporcionan información sobre la inversión en el sector educativo y su impacto potencial en la calidad de la educación.

Otra área importante dentro de la educación es el logro educativo, que se mide a través de indicadores como la tasa de finalización de la educación secundaria inferior (SE.SEC.CMPT.LO.ZS) y los niveles de logro educativo en la población adulta, desglosados por niveles de educación alcanzados, desde primaria hasta doctorado (SE.PRM.CUAT.ZS, SE.TER.CUAT.DO.ZS). Estos indicadores permiten evaluar no solo el acceso, sino también la permanencia y el éxito en el sistema educativo. Además, los indicadores de paridad de género en la educación (SE.PRM.GINT.FE.ZS, SE.ENR.TERT.FM.ZS) son cruciales para monitorear y promover la igualdad de oportunidades educativas entre hombres y mujeres. En conjunto, estos datos proporcionan una base sólida para analizar las políticas educativas, identificar áreas de mejora y desarrollar estrategias para aumentar la equidad y calidad en la educación a nivel global.

Se realizó un segundo Clustering utilizando K-means con 4 grupos sobre los indicadores de Educación descritos en la sección Análisis de datos faltantes (incluyendo también el GNI index y el total de población). Como es de esperarse, esta nueva clasificación proporciona una visión completamente distinta.

Para facilitar el análisis, luego del clustering se aplica PCA. La Tabla 1 muestra los loadings resultado de dicho proceso, mientras que en la Figura 16 se muestran visualmente para PCA1 y PCA2.

| Feature | PCA1 | PCA2 | PCA3 | PCA4 | PCA5 | PCA6 | PCA7 |
|---|--------|--------|--------|--------|--------|--------|--------|
| SP.POP.TOTL (población total) | 0,411 | 0,867 | 0,045 | -0,277 | -0,001 | -0,076 | 0,002 |
| NY.GNP.MKTP.CD (GNI) | 0,38 | 0,683 | -0,209 | 0,593 | 0,008 | -0,007 | 0 |
| SL.TLF.TOTL.IN (fuerza laboral total) | 0,428 | 0,877 | -0,008 | -0,216 | -0,015 | 0,08 | -0,001 |
| SL.UEM.TOTL.ZS (% desempleo) | -0,928 | 0,341 | -0,162 | -0,019 | 0,006 | -0,004 | -0,045 |
| SL.TLF.TOTL.FE.ZS (% fuerza laboral femenina) | 0,266 | -0,244 | -0,924 | -0,127 | -0,069 | -0,003 | 0,002 |
| SL.UEM.TOTL.FE.ZS (desempleo femenino) | -0,915 | 0,343 | 0,072 | 0,045 | -0,208 | -0,001 | 0,021 |
| SL.UEM.TOTL.MA.ZS (desempleo masculino) | -0,892 | 0,332 | -0,254 | -0,043 | 0,182 | 0,004 | 0,026 |

Tabla 1: Loadings del PCA aplicado al dataset de Educación

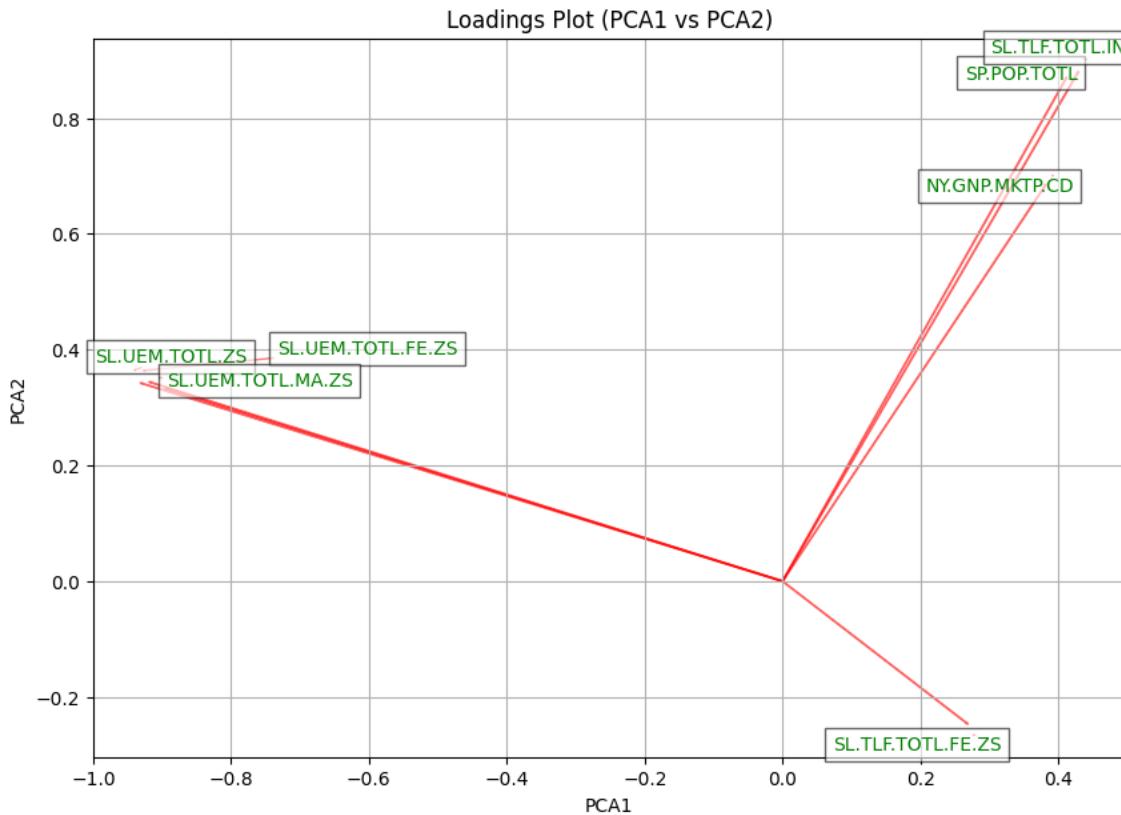


Figura 16: Biplot de los loadings de PCA (para los componentes 1 y 2) aplicado a los campos del área Education.

Las variables SP.POP.TOTL, NY.GNP.MKTP.CD, y SL.TLF.TOTL.IN están fuertemente alineadas y apuntan en la misma dirección en el cuadrante superior derecho, lo que sugiere que estas variables están positivamente correlacionadas entre sí.

Las variables SL.UEM.TOTL.FE.ZS, SL.UEM.TOTL.MA.ZS, y SL.UEM.TOTL.ZS están agrupadas en el cuadrante izquierdo, indicando que están correlacionadas y se encuentran fuertemente relacionadas con el desempleo.

PCA1 parece capturar la varianza relacionada con el tamaño de la población (SP.POP.TOTL), el producto nacional bruto (NY.GNP.MKTP.CD), y la fuerza laboral total (SL.TLF.TOTL.IN).

PCA2 captura la varianza relacionada con el desempleo, especialmente en mujeres (SL.UEM.TOTL.FE.ZS).

La Figura 17 presenta la visualización de los clusters usando PCA. Como puede verse, mientras los países que antes destacaban (tanto hacia arriba como hacia abajo) ahora forman parte del mismo cluster que Argentina, adicionalmente, ahora hay nuevos países que destacan sobre el resto. En este caso China, India y Estados Unidos.

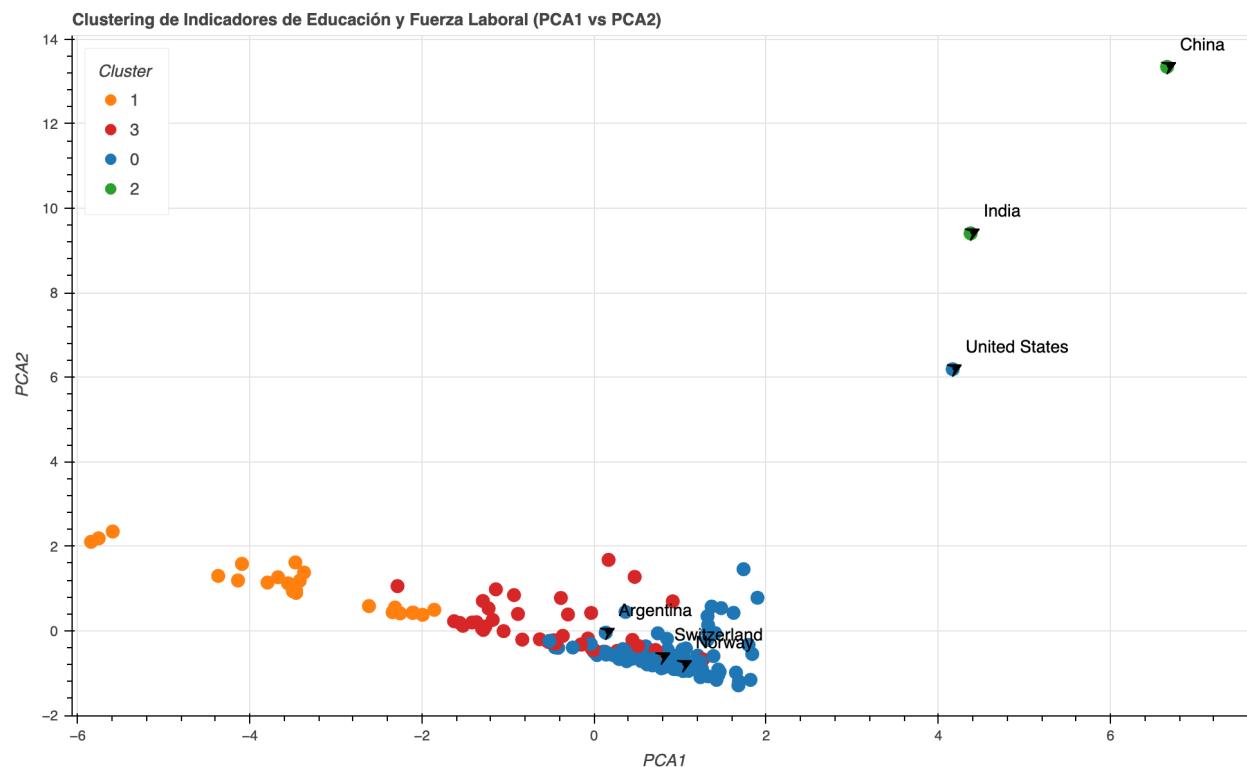


Figura 17: Muestra las componentes principales 1 y 2 resultado de aplicar PCA (con una varianza explicada de 77.43%) luego de hacer un Clustering de 4 clusters utilizando indicadores relacionados con la fuerza laboral y desempleo para el año 2022. Los clusters formados son completamente distintos. Se destacan China, India y Estados Unidos.

Clúster 0 (Azul): Economías Estables e Igualitarias

Este clúster incluye países como Estados Unidos, Suiza y Noruega. Las características promedio de estos países son una economía relativamente fuerte y estable, con bajas tasas de desempleo y una alta participación femenina en la fuerza laboral. Por lo tanto, el nombre "Economías Estables e Igualitarias" refleja adecuadamente sus características. Argentina forma parte de este cluster pues tiene un PBI relativamente alto, ~42% de participación femenina en la fuerza laboral y bajas tasas de desempleo (~8% femenino y ~6% masculino).

Clúster 1 (Naranja): Economías con Alta Desocupación

Este clúster representa países con menores ingresos y alta tasa de desempleo. Los países en este clúster enfrentan desafíos económicos significativos, con altas tasas de desempleo, especialmente entre las mujeres. El nombre "Economías con Alta Desocupación" captura bien la situación de estos países.

Clúster 2 (Verde): Economías Emergentes de Gran Escala

Este clúster incluye países con grandes poblaciones y economías emergentes. Estos países tienen grandes economías con baja tasa de desempleo, pero la participación femenina en la fuerza laboral es menor comparada con otros clústeres. El nombre "Economías Emergentes de Gran Escala" es apropiado para describir este grupo.

Clúster 3 (Rojo): Economías Vulnerables

Este clúster incluye países como Burundi, con economías más débiles. Estos países tienen tasas de desempleo más altas y menor participación femenina en la fuerza laboral, reflejando desafíos económicos y laborales significativos. El nombre "Economías Vulnerables" es adecuado para describir las características de este clúster.

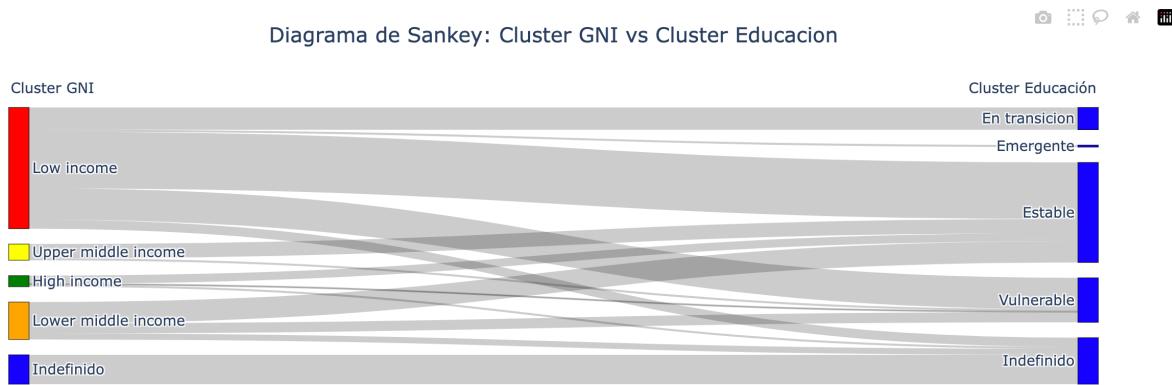


Figura 18: Diagrama de Sankey que muestra las diferencias entre los clustres conformados a partir del GNI en comparación con los clustres conformados a partir de los datos de Educación (Fuerza Laboral). Aquellos para los cuales no hay datos, se ven con la etiqueta de indefinido.

Los países considerados como Low Income se distribuyen, principalmente a las categorías de Estable y Vulnerable. Todos los países en la categoría High Income transicionan hacia la categoría Estable, con la única excepción de Qatar.

Resultados y discusión

Presentación y análisis de resultados obtenidos

En este trabajo se realizaron diversas técnicas de análisis y modelado para estudiar la relación entre indicadores socioeconómicos y la clasificación de países. Se aplicaron técnicas de clustering, específicamente K-means, para agrupar países según sus características económicas y laborales. Además, se utilizó PCA para reducir la dimensionalidad de los datos y facilitar la visualización de los clústeres.

Clustering y PCA

Utilizamos K-means para clasificar los países en cuatro clústeres basados en el GNI y en indicadores de la fuerza laboral y desempleo para el año 2022. Se aplicó PCA para reducir los múltiples indicadores a dos componentes principales, lo que permitió visualizar los clústeres en un gráfico bidimensional.

La primera clusterización utilizando únicamente el GNI arrojó mejores métricas que la clasificación propuesta por el Banco Mundial en cuanto a la calidad de los clusters (tanto Silhouette como la inercia mejoraron la clasificación propuesta por el organismo).

También se hicieron experimentos con distintos algoritmos y con distintas cantidades de clusters (todos disponibles en el código fuente anexo como parte de este trabajo) pero en pos de poder comparar con la clasificación mantuvimos la cantidad de clusters en 4.

Para la clusterización teniendo en cuenta los datos de fuerza laboral, los resultados del PCA mostraron que los dos primeros componentes principales explicaron el 77.43% de la varianza total, con un 43.47% explicado por el primer componente y un 33.96% por el segundo componente. Esto indica que los datos originales de alta dimensión se representaron adecuadamente en dos dimensiones, facilitando la interpretación visual de los clústeres.

Si bien los clusters logrados con los datos de fuerza laboral no son completamente distintos a los del GNI, permiten detectar otro tipo de comportamientos y permiten considerar aspectos de igualdad que el GNI enmascara.

Análisis de Series Temporales con ARIMA

Por cuestiones de tiempo, se alcanzó a hacer un único experimento (cuyo código se encuentra disponible como adjunto en el presente trabajo) para predecir la evolución del GNI para Argentina. Se aplicó un modelo ARIMA. Se utilizó la diferenciación para hacer la serie estacionaria, y se seleccionaron los parámetros p, d y q basados en los gráficos de ACF y PACF. El modelo ajustado mostró cierta capacidad predictiva, capturando las tendencias subyacentes en los datos históricos del GNI. Aunque es necesario mayor trabajo de entrenamiento, validación y pruebas para hacer afirmaciones al respecto.

Evolución de los factores socioeconómicos

La evolución de los factores socioeconómicos a lo largo de los años ofrece una visión general de cómo las economías y las sociedades han cambiado en respuesta a diversas políticas, eventos globales y desarrollos internos. Utilizando el clustering de 4 grupos para los años de 1980 a 2022, no se detectan grandes patrones observables ni muchos cambios significativos a lo largo del rango temporal. Esto puede deberse a que el uso de solo 4 clusters no logra captar una granularidad suficiente que facilite un análisis más detallado y revelador.

El clustering con un número reducido de grupos puede resultar en una agrupación demasiado amplia, donde países con características socioeconómicas muy diferentes terminan en el mismo clúster, diluyendo así las diferencias y similitudes significativas. Además, las transformaciones socioeconómicas suelen ser complejas y multifacéticas, requiriendo un número mayor de clústeres para capturar la diversidad de trayectorias de desarrollo de los países.

Otra consideración es la estabilidad de los clústeres a lo largo del tiempo. Los países pueden cambiar de clúster debido a cambios en sus indicadores económicos y sociales. Sin embargo, si los clústeres son demasiado generales, estos movimientos pueden no ser visibles o no proporcionar información útil. Esto sugiere que futuros estudios deberían considerar un mayor número de clústeres y quizás integrar análisis adicionales, como el uso de métodos de clustering jerárquico o análisis de trayectorias para entender mejor las dinámicas temporales.

Datos faltantes

La falta de datos fue un desafío significativo. Aunque dicha ausencia no invalida los resultados obtenidos, ciertamente limita su aplicabilidad y robustez. En cada año, los países que tienen (o no tienen) datos varían, lo que dificulta la tarea de comparar los clústeres entre sí, ya que las muestras subyacentes cambian.

Esta inconsistencia en los datos puede introducir sesgos y afectar la precisión de los clústeres formados. Por ejemplo, un país con datos faltantes en varios indicadores puede ser asignado a un clúster incorrecto debido a la información incompleta. Además, la variabilidad en los datos disponibles entre años impide una comparación directa y clara de la evolución de los clústeres, lo que es crucial para entender las dinámicas a largo plazo.

Para mitigar estos problemas, es fundamental explorar técnicas avanzadas de imputación de datos que permitan estimar los valores faltantes de manera confiable. Métodos como la imputación múltiple, el uso de modelos predictivos basados en machine learning, o técnicas de imputación basadas en series temporales pueden ofrecer soluciones efectivas. Además, la integración de datos de múltiples fuentes y la validación cruzada de estos datos pueden ayudar a mejorar la calidad y la cobertura de los datos disponibles, proporcionando una base más sólida para el análisis.

Discusión de los resultados y su relevancia

Los resultados obtenidos proporcionan una visión clara de cómo los países pueden ser agrupados según sus características socioeconómicas y laborales. La técnica de PCA demostró ser efectiva para reducir la complejidad de los datos y facilitar su visualización. Los clústeres identificados revelan patrones interesantes sobre la economía y el mercado laboral de diferentes países, destacando las similitudes y diferencias significativas entre ellos.

El uso del modelo ARIMA para predecir el GNI es particularmente relevante para la planificación económica y la formulación de políticas. La capacidad del modelo para capturar las tendencias históricas y hacer predicciones razonables puede ayudar a los responsables de políticas a tomar decisiones informadas sobre intervenciones económicas y sociales, pero realmente se necesita mas trabajo de modelado y validacion para poder conseguir resultados utilizables.

Aunque el análisis actual proporciona una visión útil de los factores socioeconómicos, no arroja luz sobre los cambios a lo largo del tiempo, las limitaciones impuestas por los datos faltantes y la granularidad del clustering deben ser abordadas en investigaciones futuras para obtener conclusiones más detalladas y robustas.

Limitaciones y posibles mejoras

Una limitación de este estudio es la dependencia de los datos disponibles, que en algunos casos presentan datos faltantes. Aunque se aplicaron estrategias para mitigar este problema, la precisión y la fiabilidad de los resultados pueden verse afectadas. Futuras investigaciones podrían beneficiarse de la integración de más fuentes de datos y la aplicación de técnicas avanzadas de imputación de datos faltantes.

Además, explorar otras técnicas de clustering, como el clustering jerárquico o el clustering basado en densidad, podría proporcionar perspectivas adicionales y complementar los hallazgos obtenidos con K-means. También se podría extender el análisis a un período más largo, utilizando datos de múltiples años para observar la evolución de los países en el tiempo y revelar tendencias a largo plazo.

Conclusión

Resumen de los hallazgos principales

Clustering Basado en GNI y Fuerza Laboral

- La técnica de K-means, aplicada tanto al GNI per cápita como a los indicadores de fuerza laboral, permitió identificar diferentes clústeres de países.
- La clasificación basada en GNI mostró una mejora en la calidad de los clústeres respecto a la clasificación del Banco Mundial.
- Los resultados del clustering basado en fuerza laboral revelaron patrones distintos, destacando la importancia de la participación femenina en la fuerza laboral y las tasas de desempleo.

Análisis de Componentes Principales (PCA)

- PCA facilitó la visualización de los datos reduciendo la dimensionalidad y manteniendo el 77.43% de la varianza total en dos componentes principales.
- Esta visualización permitió identificar outliers y patrones entre los clústeres, proporcionando una herramienta visual efectiva para el análisis de los datos.

Conclusiones generales y su relación con los objetivos del trabajo

En resumen, el uso de técnicas de clustering, y en particular de K-means, ha demostrado ser una herramienta poderosa para la clasificación y análisis de los indicadores socioeconómicos, permitiendo una mejor comprensión de las dinámicas y características de los países en estudio. Sin embargo, para mejorar la robustez y aplicabilidad de los resultados, es crucial considerar las estrategias mencionadas para manejar los datos faltantes y explorar enfoques más avanzados y diversos en futuras investigaciones.

El trabajo ha logrado clasificar y analizar los indicadores socioeconómicos de manera efectiva, utilizando técnicas de clustering y PCA para proporcionar una mejor comprensión de las dinámicas y características de los países en estudio. La utilización de K-means y PCA ha demostrado ser particularmente útil para identificar patrones y relaciones entre los indicadores seleccionados, ofreciendo una nueva perspectiva sobre la clasificación de países basada en sus características económicas y laborales.

Recomendaciones para futuros trabajos

Considerando los resultados y las limitaciones del análisis actual, se proponen varias direcciones para investigaciones futuras que podrían mejorar la comprensión de las dinámicas socioeconómicas y ofrecer una base más sólida para la formulación de políticas.

Mayor exploración de los datos Existentes

En este trabajo, se utilizó un subconjunto muy pequeño de los 1463 indicadores provistos por la fuente de datos original. Aunque es de esperar que muchos de estos indicadores sean irrelevantes para nuestro análisis específico, existe un potencial significativo en explorar más a fondo la vasta cantidad de datos disponibles. A continuación, se presentan algunas recomendaciones para futuras investigaciones sobre cómo aprovechar mejor este amplio conjunto de datos.

Identificación de Indicadores Relevantes

El primer paso hacia una mayor exploración de los datos es identificar qué indicadores adicionales pueden ser relevantes para el análisis. Esto puede implicar un proceso iterativo de selección y evaluación de indicadores basados en criterios específicos como su relación con el desarrollo económico, social y laboral. Utilizar técnicas de selección de características (feature selection) y análisis de correlación puede ayudar a identificar los indicadores más prometedores para estudios más detallados.

Análisis Multidimensional

El uso de un mayor número de indicadores permitirá realizar análisis multidimensionales más complejos y detallados. Esto puede incluir la aplicación de técnicas de análisis factorial, análisis de componentes principales (PCA) con más dimensiones, y modelos de redes neuronales que pueden manejar grandes volúmenes de datos para detectar patrones y relaciones no lineales. Estos enfoques pueden proporcionar una visión más rica y matizada de las dinámicas socioeconómicas.

Estudios Sectoriales

Explorar los datos existentes puede permitir estudios sectoriales específicos que aborden áreas como la educación, la salud, la infraestructura, y la tecnología, entre otros. Cada uno de estos sectores tiene indicadores específicos que, cuando se analizan en profundidad, pueden proporcionar información valiosa sobre las fortalezas y debilidades de diferentes países en esos ámbitos. Por ejemplo, indicadores relacionados con la salud pública pueden ofrecer perspectivas sobre cómo las políticas de salud afectan el desarrollo económico y social.

Uso de Datos Temporales Completos

Como posible trabajo futuro, se propone utilizar todos los años disponibles para todos los países y realizar clustering en función de la evolución de los países a lo largo del tiempo. Esta aproximación temporal podría revelar tendencias y patrones de desarrollo que no son evidentes cuando se analiza un solo año. Al considerar la evolución de los indicadores socioeconómicos a lo largo del tiempo, se pueden identificar patrones de crecimiento y desarrollo que se pierden en un análisis estático.

Exploración de Nuevas Técnicas de Clustering

Además de K-means, explorar otras técnicas de clustering, como el clustering jerárquico o el clustering basado en densidad, podría proporcionar perspectivas adicionales y complementar los hallazgos obtenidos con K-means. Estas técnicas alternativas pueden ofrecer una mayor flexibilidad y adaptabilidad a la estructura subyacente de los datos, permitiendo la identificación de subgrupos y patrones más sutiles en los datos.

Problema de los Datos Faltantes

La falta de datos es un desafío significativo que debe abordarse para mejorar la robustez y aplicabilidad de los resultados. A continuación, se describen varias estrategias que podrían ser implementadas para manejar los datos faltantes:

Búsqueda de Información Externa: Si bien algunos de estos indicadores podrían estar disponibles en datasets de otras organizaciones, este enfoque 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, como obtener datos conocidos en Argentina.

Imputar Datos Faltantes: Para los indicadores 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 que se imputan, y las conclusiones derivadas de esto deben ser tomadas con cautela. Técnicas como la imputación por la media, media móvil, último valor conocido, regresión y MICE (Multivariate Imputation by Chained Equations) pueden ser útiles.

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, como por ejemplo: cortes por Región, por indicadores o esquemas mixto

Referencias bibliográficas

- [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](#).
- [8] Selenium webdriver, [documentación oficial del framework](#).
- [9] OpenPyXL, [documentación oficial del framework](#).
- [10] Pandas, [documentación oficial del framework](#).
- [11] NumPy, [documentación oficial del framework](#).
- [12] Matplotlib, [documentación oficial del framework](#).
- [13] Seaborn, [documentación oficial del framework](#).
- [14] Holoviews, [documentación oficial del framework](#).
- [15] bokeh, [documentación oficial del framework](#).
- [16] PDFPages, [documentación oficial del framework](#).
- [17] Scikit-learn, [documentación oficial del framework](#).
- [18] Statsmodels, [documentación oficial del framework](#).
- [19] Jupyter Notebooks, [documentación oficial del framework](#).
- [20] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar (2019). Introduction to Data Mining, Second Edition. Pearson Education Limited.
- [21] Rania Ihab Naguib. [The effects of privatisation and Foreign Direct Investment on economic growth in Argentina](#). The Journal of International Trade & Economic Development.
- [22] Richard Newfarmer, W. Mueller. [Multinational corporations in Brazil and Mexico : structural sources of economic and noneconomic power](#). Report to the Subcommittee on Multinational Corporations of the Committee on Foreign Relations, United States Senate.
- [23] Python 3.9.7. [documentación oficial](#).
- [24] Visual Studio Code 1.9.0. [documentación oficial](#).

Anexos

Anexo 1: Código fuente utilizado en el análisis

Todo el código fuente desarrollado como parte de este trabajo se encuentra en [este](#) enlace.

Anexo 2: Descripción de áreas clave

| 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 2: 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.

Anexo 3: Motivación para el recorte de datos

Como se mencionó anteriormente, dependiendo del tipo de análisis que se deseé hacer, hay distintos recortes de los datos que pueden ser utilizados con relativa poca perdida.

Por ejemplo, en la Figura 19 puede verse un recorte de un conjunto de países Latinoamérica: Argentina, Bolivia, Chile, Colombia, Ecuador, México y Uruguay, pero esta vez agrupando los indicadores en sus respectivas áreas clave.

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.

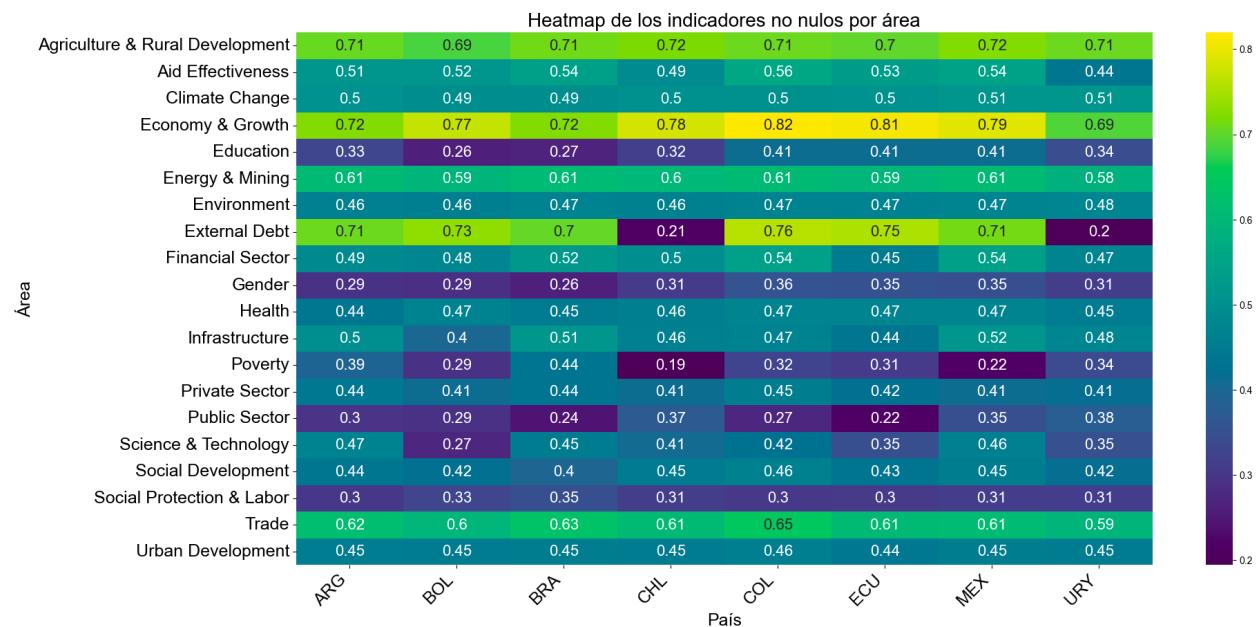


Figura 19: Heatmap con los datos faltantes para un conjunto de países seleccionados agrupando los indicadores por área.

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

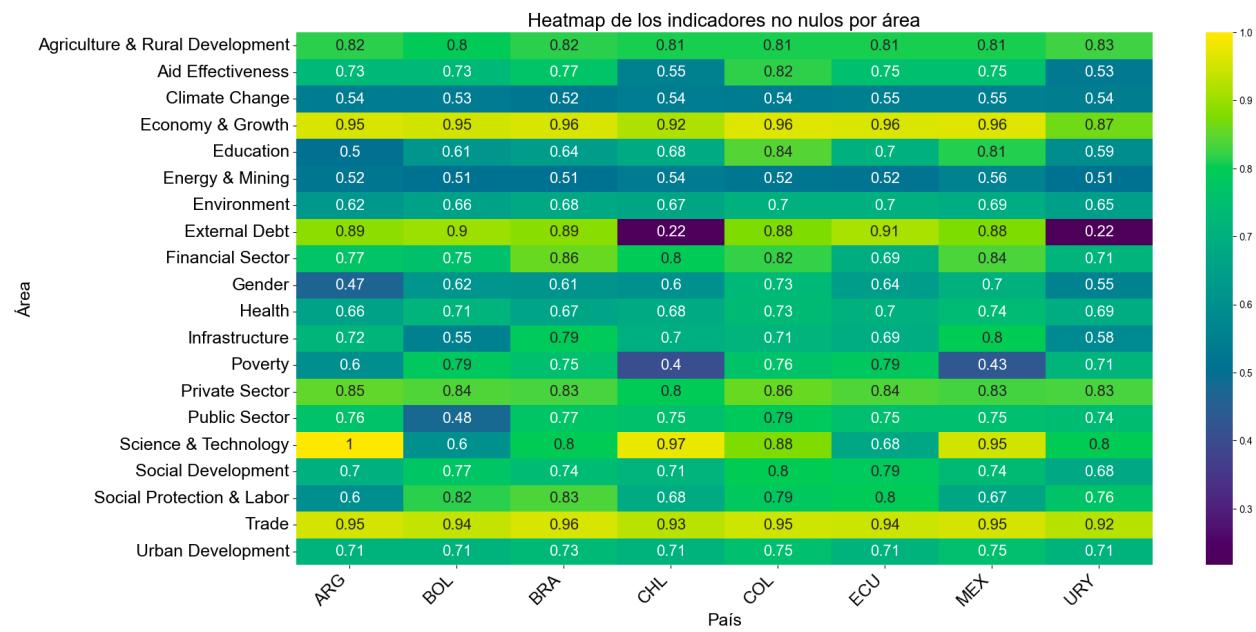


Figura 20: Muestra el mismo gráfico, en el periodo 2015-2018. La cantidad de datos faltantes disminuye notablemente, alcanzando picos de no nulos del 96% en algunos casos.

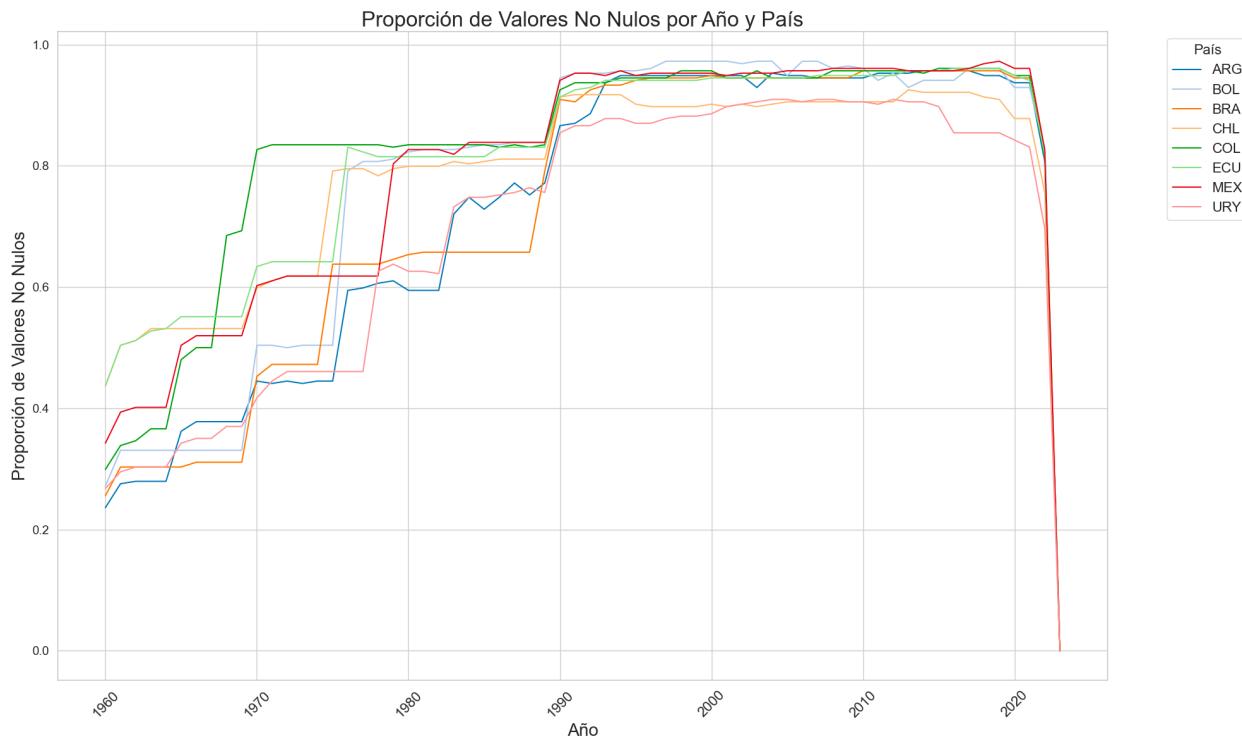


Figura 21: Cantidad de valores no nulos por año para el área de Economy & Growth

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