



CIMAT

SEGMENTACIÓN DE PAÍSES SEGÚN INDICADORES DE DESARROLLO

Introducción a Ciencia de Datos

Autor:
Iván García Mestiza

Profesor:
Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas, A. C.

1. Introducción

El análisis del desarrollo humano y económico de los países del mundo es una tarea relevante en muchas áreas de las ciencias sociales, en la economía, para la toma de decisiones de políticas públicas y para distintos organismos internacionales. El *World Bank Group* es una subdivisión del Banco Mundial (*World Bank*) que coordina el trabajo estadístico de datos y mantiene una gran cantidad de bases de datos macro, financieras y de distintos sectores [1]. La mayoría de sus datos provienen de los sistemas estadísticos de sus países miembros, y el Banco Mundial trabaja para ayudar a los países en desarrollo a mejorar la capacidad, eficiencia y efectividad de sus sistemas estadísticos nacionales, lo cual permite obtener datos confiables y de calidad.

Las bases de datos mantenidas por el Banco Mundial permiten tomar decisiones de gestión críticas y proveen información útil para las distintas actividades operacionales del organismo. Además, son útiles para el desarrollo de políticas públicas efectivas, monitorear la implementación de estrategias para la reducción de la pobreza, o monitorear el progreso hacia metas globales. Dicho organismo concentra información sobre 1600 indicadores sobre 217 economías, correspondientes a las siguientes áreas [2]:

- **Pobreza y desigualdad:** Son indicadores que miden la incidencia y profundidad de la pobreza de los países de acuerdo con definiciones internacionales, así como la desigualdad económica existente en ingresos y riqueza a lo largo de los países y regiones.
- **Sociedad:** Corresponde a los indicadores que describen a la sociedad en general, incluyendo salud, educación, nutrición, mortalidad, empleos, entre otros.
- **Medio ambientales:** Estos se relacionan con el uso de recursos naturales y permiten evaluar los efectos del cambio climático y el impacto ambiental del humano en el planeta.
- **Económicos:** Son indicadores relacionados con medidas económicas como el PIB per cápita, así como el balance de pagos, consumos y finanzas de los países en general.
- **Estados y mercados:** Analizan la inversión privada, el sector público, sistemas financieros, la infraestructura en comunicaciones y transporte, entre otros.
- **Enlaces globales:** Son indicadores sobre el tamaño y la dirección de los flujos y vínculos económicos, como el comercio, la equidad y la deuda, así como el turismo y la migración.

En el presente proyecto se hace una recolección sistemática de los datos correspondientes a diversos indicadores socioeconómicos, demográficos, ambientales y educativos para todos los países del mundo disponibles en la base del Banco Mundial, lo cual nos permite asegurar que se trabaja con información lo más completa, reciente y confiable posible, con el objetivo de hacer una segmentación de países en función de diversas métricas socioeconómicas y de desarrollo.

A través de la aplicación de diversas técnicas de ciencia de datos, se busca identificar patrones, similitudes y diferencias entre las naciones, lo que permitirá obtener una comprensión más profunda de su evolución y desafíos en el contexto global. Para la segmentación de los países se emplearán técnicas de aprendizaje automático no supervisado, clustering, reducción de dimensionalidad, redes neuronales y métodos estadísticos que faciliten la interpretación de los resultados.

Lo anterior no solo contribuye a una mejor visualización de las diferencias y similitudes entre países, sino que también puede servir como base para la formulación de políticas públicas y estrategias de desarrollo más efectivas, adaptadas a las características específicas de cada uno de los grupos de países identificados.

2. Recolección y preprocesamiento de datos

Los datos trabajados fueron extraídos a través de la API de *World Bank Data* en Python, usando el módulo *world_bank_data*. Por medio de esta biblioteca se pueden seleccionar los indicadores de los que se

desea extraer la información, sobre 217 países (territorios económicos). Sin embargo, a pesar de la existencia de una gran cantidad de indicadores, pocos de ellos tienen información suficiente (de hecho, muchos de ellos tienen datos para menos de la mitad de los países), por lo que el análisis se limitó a los indicadores que se muestran en la Tabla 1. Sin embargo, artículos como [3] hacen uso de indicadores similares, obteniendo resultados que explican de manera satisfactoria los datos, así que esta restricción no constituye un mayor problema. En el código adjunto se pueden observar la cantidad de datos faltantes por columna de distintos indicadores clave.

| Variable | Descripción | Unidades |
|--------------------------|---|--------------------------------|
| Económicos | | |
| gdp_per_capita | PIB per cápita | USD |
| gdp_growth | Tasa de variación anual del PIB | % anual |
| inflation | Inflación anual | % anual |
| unemployment | Porcentaje de la fuerza laboral sin empleo y buscando trabajo | % de la fuerza laboral |
| exports_gdp | Exportaciones como proporción del PIB | % del PIB |
| Sociales | | |
| life_expectancy | Esperanza de vida al nacer | Años |
| infant_mortality | Muertes de menores de 5 años por cada 1000 nacidos vivos | Muertes por 1000 nacidos vivos |
| fertility_rate | Tasa de fertilidad | Nacimientos por mujer |
| electricity_access | Porcentaje de la población con acceso a electricidad | % de la población |
| mobile_subscriptions | Suscripciones móviles | Suscripciones por 100 personas |
| Gubernamentales | | |
| control_corruption | Indicador WGI sobre control de la corrupción | Índice (-2.5 a +2.5) |
| government_effectiveness | Indicador WGI de efectividad del gobierno | Índice (-2.5 a +2.5) |
| population | Población total | Personas |
| urban_population | Proporción de población que vive en áreas urbanas | % de la población |

Tabla 1: Descripción de variables con unidades del World Bank Data.

Los índices de control de corrupción y efectividad del gobierno son provistos por el *World Bank Data*, y pertenecen a la clase llamada “Indicadores Mundiales de Gobernanza” (*Worldwide Governance Indicators*, WGI). Estos miden la distancia en variaciones estándar a la media mundial, calculada por esta misma organización.

Según la información de *World Bank Data*, muchos de los datos faltantes corresponden a países con infraestructura insuficiente para una recolección estadística adecuada. Por lo tanto, usar técnicas comunes de imputación, como reemplazar valores faltantes por la media, mediana o incluso *hot-deck*, podría sesgar de manera significativa la muestra. Para abordar este problema, se optó por usar *KNN Imputer*, que es un imputador más robusto que preserva las relaciones entre las características del conjunto de datos. Además, la imputación se realizó únicamente para los países con información para más de la mitad de los indicadores. Esto evitó la introducción de más información simulada que real, lo que podría sesgar aún más los resultados. Como resultado de esta reducción, la muestra trabajada consiste en 207 territorios, siendo los eliminados: Samoa Americana, Islas Vírgenes Británicas, Islas del Canal, Curazao, Islas Feroe, Gibraltar, Isla de Man, Nueva Caledonia, Islas Marianas del Norte y San Martín (parte francesa).

3. Análisis de datos

3.1. Consideraciones teóricas

Para realizar una segmentación de países se pueden usar varias técnicas de clasificación no supervisada, tales como k-means, clustering jerárquico y espectral. Existen diversas métricas que permiten evaluar cuál es el mejor algoritmo de clustering y el valor óptimo del número de clusters, tales como el Coeficiente de la Silueta, el Índice de Davies-Bouldin y el Índice de Calinski-Harabasz [4].

Por un lado, el Coeficiente de la Silueta permite comparar la distancia media entre los puntos de un cluster con la distancia al cluster más cercano, y mide qué tan buena es la separación entre clusters. Esta métrica es la más común y con mayor interpretabilidad, puesto que captura la cohesión intra-clusters y la separación entre ellos. Su rango es de -1 a 1 , en donde los valores cercanos a 1 indican que los clusters están bien separados del resto, valores cercanos a 0 indican que las fronteras de los clusters están muy encimadas, y valores negativos indican que la clasificación no es buena.

Por otra parte, el Índice de Davies-Bouldin calcula la dispersión intra-clusters y la distancia entre los centroides de una manera aproximada y más rápida que la Silueta, y mide la similitud entre clusters (los valores más chicos son considerados mejores). Entre sus principales ventajas destaca que es rápido de calcularse, y penaliza clusters que son muy cercanos o lejanos. Sin embargo, es sensible a los valores atípicos y puede ser inestable en altas dimensiones. Por último, el Índice de Calinski-Harabasz, también conocido como el Criterio de la Razón de Varianzas (*Variance Ratio Criterion*) mide la razón entre las varianzas intra-clusters y la varianza entre los diferentes clusters. Este índice favorece la separación de clusters, y proporciona fundamentos estadísticos que permiten escoger un número de clusters adecuado, pero es más sesgado hacia una cantidad mayor de clusters.

Para cada una de las técnicas de segmentación descritas en el resto de esta sección, se calcularán las tres métricas anteriores, lo cual permitirá fundamentar la elección de los clusters mostrados. Además, para visualizar las clasificaciones se hará una reducción de dimensionalidad por medio de un Análisis de Componentes Principales (PCA), mostrando los *loadings* de los tres que explican mejor la varianza. En el código adjunto se pueden ver las gráficas de los clusters por pareja de componentes principales, así como los grupos de los países identificados en mapas interactivos en formato html.

3.2. Segmentación por diferentes tipos de indicadores

Existen estudios, como [3], en donde se ha hecho un análisis de segmentación de países con los indicadores divididos por tipo: económicos, sociales, gubernamentales, entre otros. Sin embargo, este análisis puede resultar limitado cuando se quieren comparar las similitudes y diferencias entre las naciones de manera más transversal y completa. Así pues, primero realizaremos una segmentación con los distintos tipos de indicadores y luego incluyéndolos todos a la vez, para ver con cuáles de las técnicas se explican mejor los resultados.

3.2.1. Indicadores Económicos

Al hacer una reducción de dimensiones por medio de PCA, se obtiene que los primeros tres componentes principales explican el 78.43 % de la varianza, y los primeros cuatro logran explicar el 93.74 %. En la Figura 1 se muestran los loadings correspondientes para cada uno de los tres primeros componentes principales.

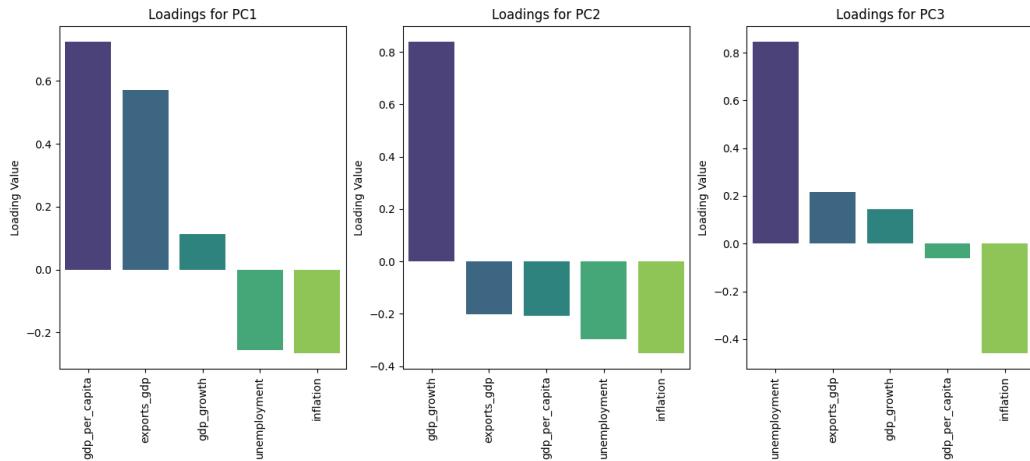


Figura 1: Loadings de PCA para indicadores económicos.

Como se puede observar, los indicadores que explican de mejor manera la varianza son el PIB per cápita y la cantidad de exportaciones, seguidos por el crecimiento anual del PIB, la tasa de desempleo y la inflación. Al realizar un análisis de clusters por medio de k-means, jerárquico y espectral, encontramos que el mejor método es el jerárquico, con un número de clusters óptimo igual a 5, y una distribución de países como se muestra en la Tabla 2, en donde se muestra la media de cada indicador.

| Cluster | Núm de países | PIB | Desempleo | Crecimiento del PIB | Inflación | Exportaciones |
|---------|---------------|----------|-----------|---------------------|-----------|---------------|
| A | 47 | 66771.26 | 4.422 | 1.786 | 2.366 | 70.387 |
| B | 134 | 16494.35 | 5.690 | 3.849 | 7.563 | 33.929 |
| C | 24 | 10276.91 | 18.079 | -0.263 | 7.294 | 44.550 |
| D | 1 | 26547.05 | 7.876 | -1.719 | 219.883 | 15.334 |
| E | 1 | 70297.40 | 10.165 | 43.372 | 2.903 | 34.625 |

Tabla 2: Distribución óptima de clusters económicos.

El cluster A se caracteriza por ser un grupo de países con un PIB medio bastante alto comparado al resto, con poca inflación, baja tasa de desempleo y cuya economía depende en gran parte de las exportaciones. En este se encuentran países como Estados Unidos, Canadá, Australia, Arabia Saudí, Japón y la gran mayoría de países de Europa Oriental, incluyendo Reino Unido, Suiza, Finlandia, Alemania, Italia, Francia, entre otros. Por otra parte, el cluster B tiene la mayoría de los países, una tasa de desempleo relativamente chica, pero buen crecimiento del PIB, en donde las exportaciones corresponden a la tercera parte de su PIB. A este corresponden la mayoría de los países de Latinoamérica, asiáticos y de África Oriental.

El cluster C comprende 24 países, siendo la mayoría pertenecientes a Sudáfrica, caracterizados por una tasa de desempleo muy alta, una economía decayente (pues el PIB tiene tasa de crecimiento negativa) y una inflación moderada. Por último, existen 2 clusters con un elemento cada uno. El D está conformado por Argentina, y su separación del resto se explica porque tiene niveles de inflación sumamente elevados, y un descenso del PIB más grande que el resto de los países. Por último, el E corresponde a Guayana Francesa, un territorio en América perteneciente a Francia, con una población de alrededor de 300 000 habitantes, siendo alrededor del 44 % menores de 20 años, y con diversos subsidios por parte de Francia.

A manera de evaluación de la robustez de la clasificación, podemos usar algoritmos de clasificación supervisada con las etiquetas siendo las de los clusters óptimos que encontramos anteriormente. Esto proporciona una forma de “verificar” los resultados anteriores: si los algoritmos como Random Forest o Redes Neuronales pueden clasificar de manera satisfactoria los clusters anteriores, entonces los grupos identificados corresponden a patrones reales en los datos, y esto indicaría que tienen una estructura que sí se puede separar e

identificar. Además, el entrenamiento de un algoritmo de este tipo permite predecir los cambios que puedan darse en los clusters con la llegada de nuevos datos.

Puesto que existen dos clusters de un elemento cada uno, no se puede esperar que los algoritmos de clasificación supervisada aprendan lo suficiente de ellos, por lo que se considerarán únicamente los clusters con una cantidad de países mayor a 2.

Al hacer una clasificación supervisada por el método de Random Forest se obtiene una exactitud de 0.919, la cual corresponde a la proporción de países que se clasificaron de manera correcta. Por otro lado, la sensibilidad media es de 0.901 y la especificidad media es de 0.945, por lo que el algoritmo aprendió lo suficientemente bien tanto a detectar a los pertenecientes a los clusters como a los que no, de manera balanceada. Esto permite confirmar que la clasificación de los clusters se realizó de manera adecuada, y por lo tanto concluir que los resultados obtenidos son significativos.

3.2.2. Indicadores Sociales

Haciendo una reducción de dimensiones por medio de PCA, se obtiene que los primeros tres componentes principales explican el 94.15 % de la varianza, y los primeros cuatro logran explicar el 97.92 %. En la Figura 2 se muestran los loadings correspondientes para cada uno de los tres primeros componentes principales.

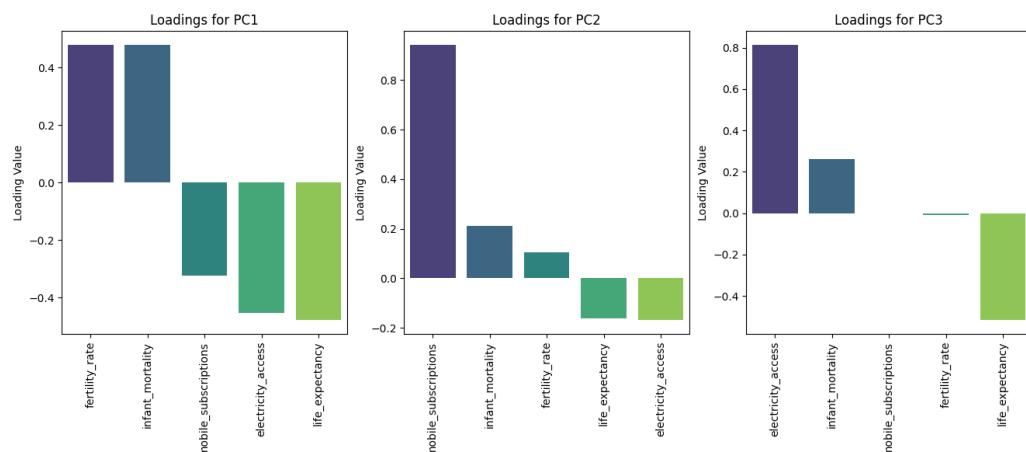


Figura 2: Loadings de PCA para indicadores sociales.

Como se puede observar, los indicadores que explican de mejor manera la varianza son la tasa de fertilidad y mortalidad infantil, seguidos por la esperanza de vida, acceso a electricidad y suscripciones móviles. Al realizar un análisis de clusters por medio de k-means, jerárquico y espectral, encontramos que el mejor método es nuevamente el jerárquico, con un número de clusters óptimo igual a 2, y una distribución de países como se muestra en la Tabla 3, en donde se muestra la media de cada indicador.

| Cluster | Núm países | Esperanza vida | Mortalidad inf | Fertilidad | Electricidad | Susc. móviles |
|---------|------------|----------------|----------------|------------|--------------|---------------|
| A | 163 | 76.158 | 11.148 | 1.908 | 97.695 | 127.890 |
| B | 44 | 63.801 | 43.822 | 4.234 | 50.788 | 83.559 |

Tabla 3: Distribución óptima de clusters sociales.

El cluster A contiene a la mayoría de los países, y se caracteriza por tener alta esperanza de vida, una tasa baja de mortalidad infantil, un acceso a la electricidad casi del total de su población (con media del 97.69 %) y que prácticamente todos los habitantes poseen dispositivos móviles. En este se encuentran prácticamente todos los países de América, Europa y Asia.

Por otra parte, el cluster B tiene 44 países, que abarcan la mayoría de África central, así como Haití, Afganistán, Pakistán y Papúa Nueva Guinea. Este bloque se caracteriza por tener una esperanza de vida menor, y una alta tasa de mortalidad infantil, del 4.3 %, casi 4 veces mayor a la del cluster 0. Además, podemos observar que apenas la mitad de su población tiene acceso a electricidad, y aquí se hacen claras las diferencias sociales con el resto del planeta.

En este caso el algoritmo de Random Forest también proporciona resultados satisfactorios, con una exactitud de 0.97, una sensibilidad media de 0.92 y una especificidad de 1. Lo anterior indica que el modelo aprendió muy bien a clasificar a los países, y por consiguiente, refuerza la validez de las conclusiones obtenidas.

3.2.3. Indicadores Gubernamentales

Haciendo una reducción de dimensiones por medio de PCA, se obtiene que los primeros dos componentes principales explican el 81.55 % de la varianza, y los primeros cuatro logran explicar el 97.88 %. En la Figura 3 se muestran los loadings correspondientes para cada uno de los tres primeros componentes principales.

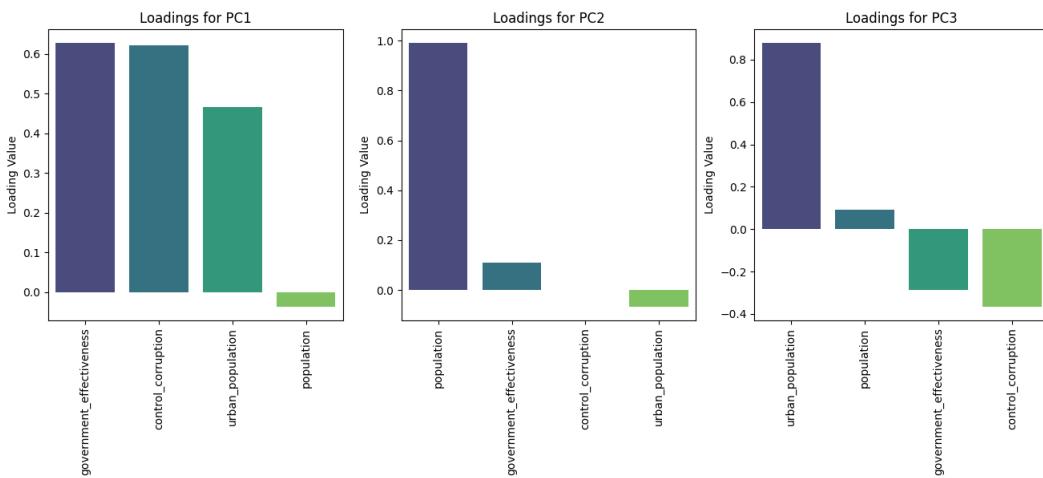


Figura 3: Loadings de PCA para indicadores gubernamentales.

Como se puede observar, los indicadores que explican de mejor manera la varianza son la efectividad del gobierno y el control de la corrupción, seguidos por el porcentaje de población en áreas urbanas y la población total. Al realizar un análisis de clusters por medio de k-means, jerárquico y espectral, encontramos que el mejor método es el jerárquico, con un número de clusters óptimo igual a 3, y una distribución de países como se muestra en la Tabla 4, en donde se muestra la media de cada indicador.

| Cluster | Núm países | Control corrupción | Efectividad gob | Población | Población urbana |
|---------|------------|--------------------|-----------------|---------------|------------------|
| A | 161 | -0.398 | -0.409 | 26 546 610 | 55.922 |
| B | 44 | 1.344 | 1.322 | 22 358 550 | 86.638 |
| C | 2 | -0.185 | 0.577 | 1 429 955 000 | 51.205 |

Tabla 4: Distribución óptima de clusters gubernamentales.

El cluster A contiene a la mayoría de los países, y se caracteriza por tener aproximadamente la mitad de su población en áreas urbanas, pero con poca efectividad gubernamental y alto nivel de corrupción. Estos últimos dos indicadores corresponden a desviaciones estándar con respecto a la media mundial, y a él pertenecen la mayoría de países en Latinoamérica, Asia, África y Europa Occidental.

Por otra parte, el cluster B tiene 44 países, muy similares a los del cluster A de los económicos, con una muy buena estructura gubernamental y la gran mayoría de su población perteneciente a áreas urbanas. Por

último, el cluster C está comprendido por China e India, siendo los países más poblados del mundo, que tienen buena efectividad gubernamental y aproximadamente la mitad de su población vive en zonas urbanas.

En este caso el algoritmo de Random Forest, entrenado únicamente con los clusters A y B, también proporciona resultados satisfactorios, con una exactitud de 0.98, una sensibilidad media de 0.99 y una especificidad de 1, de donde se concluye que esta clasificación es sumamente robusta.

3.3. Segmentación global de indicadores

La clasificación realizada en la sección anterior permite obtener una idea general de las características en común de los diferentes grupos de países. Sin embargo, como podemos apreciar, los clusters obtenidos no siempre se conservan a través de todos los tipos de indicadores, así que usar clusters económicos para dar sugerencias gubernamentales o sociales puede no ser completamente adecuado. Por consiguiente, exploraremos lo que sucede al hacer una segmentación usando todos los indicadores disponibles, permitiéndonos una comparación transversal entre los países de cada cluster, y por lo tanto, consiguiendo resultados con mayor validez estadística.

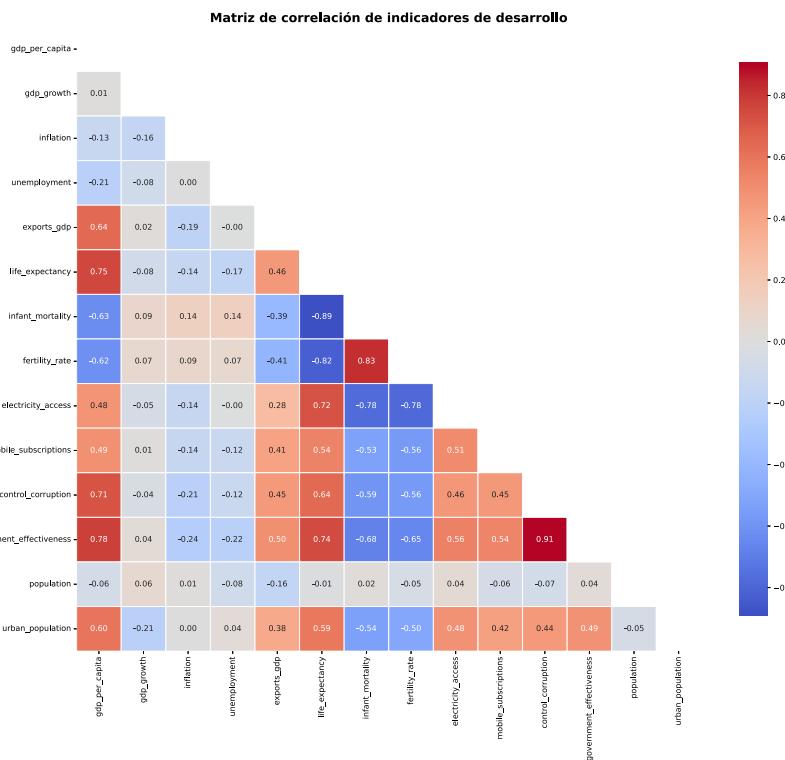


Figura 4: Matriz de correlación de todos los indicadores.

Como se puede observar en la Figura 4, existen indicadores fuertemente correlacionados, como el control de la corrupción con la efectividad del gobierno, la esperanza de vida con la mortalidad infantil y la tasa de fertilidad, y varios indicadores del mismo tipo entre sí. Sin embargo, también existen correlaciones entre indicadores de diversos ámbitos, como el PIB per cápita con la efectividad del gobierno y la esperanza de vida. Además, se puede observar que los indicadores con mayor correlación entre ellos y con el resto son los sociales y gubernamentales, mientras que los económicos tienen por lo general poca relación entre ellos. Esto sugiere que el agrupamiento económico es más robusto, puesto que consideró más indicadores que proporcionaban información más diversa que en los sociales o gubernamentales. Así pues, podemos hacer una clasificación no supervisada y luego métodos como Random Forest para obtener las características más importantes de cada grupo.

Al efectuar una reducción de dimensiones por medio de PCA, se obtiene que los primeros cinco componentes principales explican el 77.94 % de la varianza, mientras que los primeros seis explican el 83.17 %. En la Figura 5 se muestran los loadings correspondientes para cada uno de los tres primeros componentes principales.

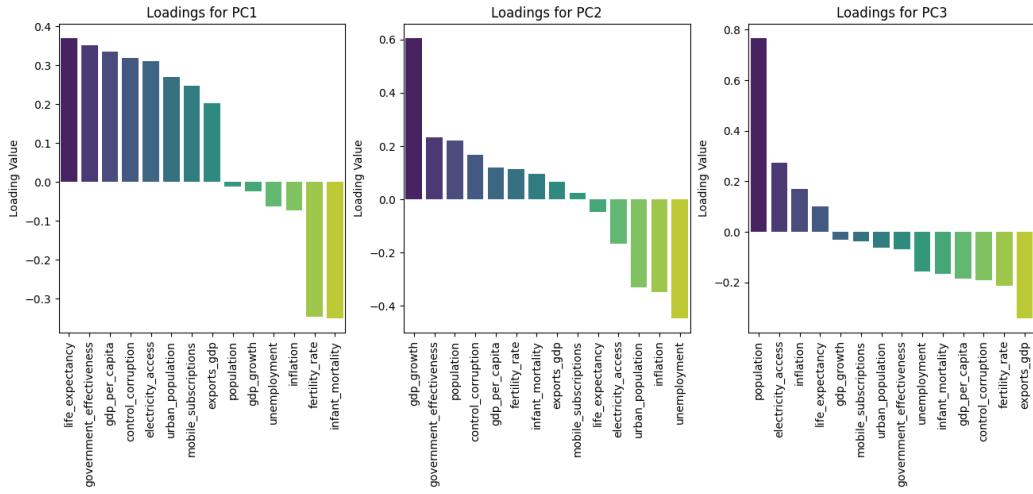


Figura 5: Loadings de PCA para todos los indicadores.

El primer componente principal logra explicar el 46.85 % de la varianza, y constituye una mezcla de los indicadores de diversos tipos, con énfasis en los gubernamentales y sociales, aunque la variable del PIB per cápita resulta significativa. Por otro lado, el segundo componente captura casi todo el resto de las variables económicas, dándole un peso muy alto al crecimiento del PIB, a la inflación y al desempleo, y el resto constituye una mezcla de indicadores sociales.

Al realizar un análisis de clusters por medio de k-means, jerárquico y espectral, encontramos que el mejor método es el k-means, con un número de clusters óptimo igual a 2, el primero de ellos constituido por 126 países y el segundo por 81. El primer cluster está compuesto por la mayoría de países de América, Europa y Asia, su PIB per cápita medio es de 40689.355 dólares, con una tasa de crecimiento media de 2.75 % anual, niveles de inflación alrededor del 5.79 %, y siendo las exportaciones el 52.85 % de su PIB total anual. Por otro lado, su población tiene una esperanza de vida de alrededor de 78 años, una tasa de fertilidad de 1.64 nacimientos por mujer, una mortalidad infantil del 0.81 %, una población en su mayoría urbana, con media del 73.3 %, y el 99.61 % de sus habitantes cuenta con electricidad. Por último, sus índices de control de corrupción y efectividad del gobierno son del 0.468 y 0.514, respectivamente.

Por otra parte, al segundo cluster pertenecen países como Guatemala, Honduras, Nicaragua, Haití y Bolivia en América, así como India, Irán, Afganistán, Pakistán, Papúa Nueva Guinea y la mayoría de los países de África. Su PIB per cápita medio es de 6976.845 dólares, con una tasa de crecimiento media de 3.56 % anual, niveles de inflación alrededor del 9.78 %, y siendo las exportaciones el 28.56 % de su PIB total anual. Por otro lado, su población tiene una esperanza de vida de alrededor de 66 años, una tasa de fertilidad de 3.58 nacimientos por mujer y una mortalidad infantil del 3.36 %, siendo sumamente mayor a la del primer cluster. Su población en su mayoría vive en áreas rurales, con el 45.45 % habitando en zonas urbanas, y el 69.23 % de sus habitantes cuenta con electricidad. Por último, sus índices de control de corrupción y efectividad del gobierno son del -0.795 y -0.882, respectivamente. Por lo tanto, este grupo de países tiene un menor desempeño en casi todos los indicadores con respecto a los del primer grupo.

Ahora bien, el Random Forest proporciona un método para realizar un análisis de las variables más importantes, siempre que este explique de manera satisfactoria los datos [5]. Al ejecutar dicho método de clasificación supervisada y evaluarlo por medio de Validación Cruzada, se obtuvo una media de exactitud de $0.961 (\pm 0.039)$, por lo que los clusters parecen ser bien diferenciados. La Figura 6 muestra las características más importantes que distinguen a los grupos, muchas de las cuales se describieron para los clusters anteriores.

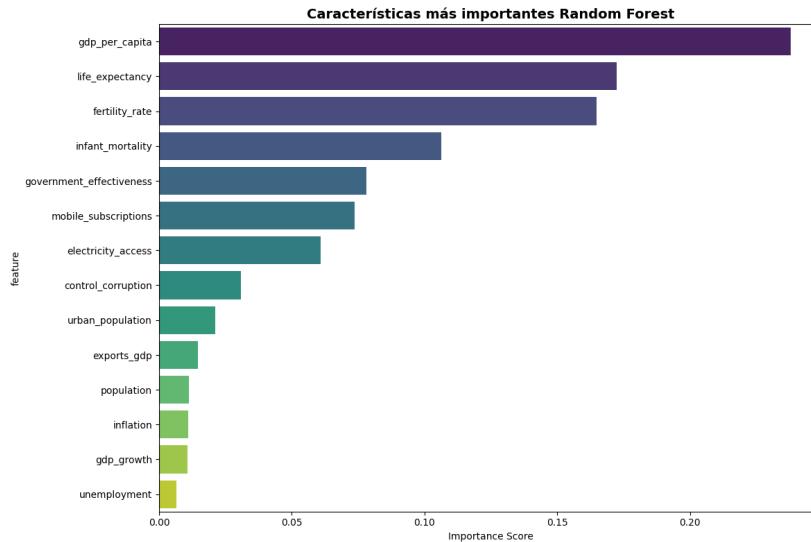


Figura 6: Características más importantes por Random Forest.

3.4. Aplicaciones de la Segmentación de Países

La segmentación de países permite realizar análisis más detallados para identificar las fortalezas y debilidades de cada grupo, considerando aspectos económicos, sociales, gubernamentales y otros. Esta información puede ser crucial para replantear o reforzar las políticas internas de cada país, con el fin de alcanzar las diversas metas globales. Además, organismos internacionales como el Banco Mundial, la ONU y la OMS pueden utilizar esta información para establecer programas de apoyo a las economías más débiles, a regiones con rezagos sociales o a países que enfrentan situaciones de salud complejas.

Una de las aplicaciones que podemos realizar es identificar en cada grupo los países con mejor desarrollo económico que de salud, y viceversa, comparado con los miembros de su mismo grupo. En la Figura 7 se muestra una gráfica del PIB per cápita vs Esperanza de Vida, con los dos clusters identificados en diferentes colores.

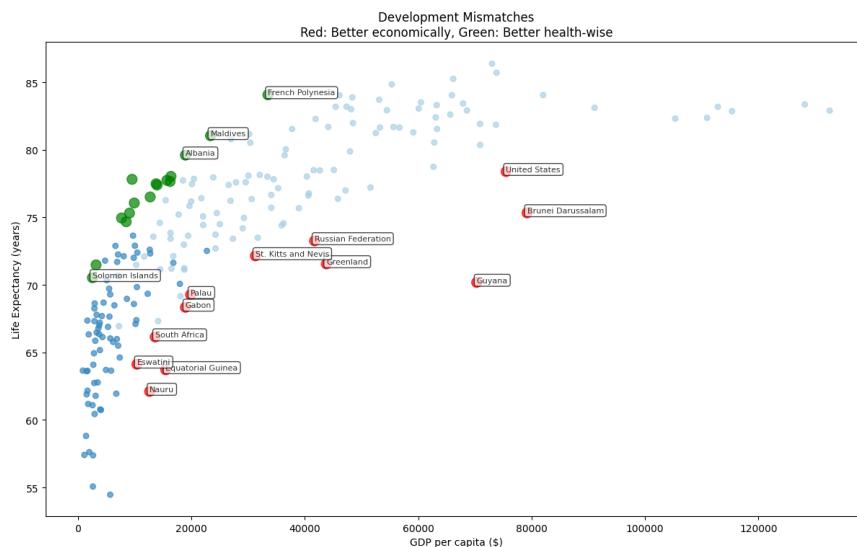


Figura 7: Países con mejor desempeño económico que de salud y viceversa.

Para cada país dentro de un cluster se hizo la diferencia entre la posición que ocupan en términos de PIB per cápita dentro de su grupo y la posición que tienen en términos de esperanza de vida. Los países con diferencias mayores con respecto a sus clusters fueron marcados con círculos de colores diferentes: rojo para aquellos que tienen mejores condiciones económicas que de salud, con respecto a su cluster, y verde para el caso contrario.

Además, en la figura anterior se puede ver claramente el comportamiento distinto de los países dentro de los clusters: los pertenecientes al de color azul intenso se caracterizan por tener bajo PIB per cápita y con poca varianza, con una esperanza de vida de menos de 75 años, y el resto de los países tiene una mayor varianza en términos del PIB y en general mayor esperanza de vida. Para los que toman decisiones de políticas públicas esta información puede ser bastante relevante, por ejemplo, países como Sudáfrica y Guinea Ecuatorial tiene un PIB más alto que el resto de los miembros de su cluster, así que podrían acceder a mejores préstamos, y regiones como la Polinesia Francesa o las Maldivas poseen excelentes condiciones de salud sin tener una gran economía, por lo que el resto de países de su grupo pueden adoptar medidas similares que les ayuden a mejorar su esperanza de vida.

4. Conclusiones

La segmentación de países con base en indicadores económicos, sociales y gubernamentales mediante técnicas de ciencia de datos constituye una herramienta poderosa para comprender patrones globales y clasificar territorios con características similares. Este enfoque permite analizar de manera transversal el desempeño de las naciones en diversas áreas, sin sacrificar la capacidad de clasificación ni interpretabilidad. Al comparar países dentro de un mismo grupo, es posible plantear estrategias para mejorar uno o varios aspectos con mayores probabilidades de éxito, evitando así proponer soluciones generales que no funcionen en todos los contextos.

Este tipo de análisis no solo es útil para gobiernos y organismos públicos que buscan diseñar y mejorar políticas públicas basadas en datos reales, sino que también proporciona información crucial para instituciones como bancos centrales y agencias de la ONU, que buscan implementar programas económicos y sociales orientados a disminuir las desigualdades en la población.

Como se observó, los grupos de países resultantes varían significativamente según el tipo de indicadores considerados. Aunque los países suelen clasificarse por su nivel económico, el análisis de indicadores sociales y gubernamentales permite identificar relaciones entre países que no siempre se investigan y que no necesariamente dependen de la cercanía geográfica. Al combinar distintos tipos de indicadores se reduce un poco la capacidad de segmentación, pues obtuvimos solamente 2 clusters en lugar de los 5 que resultan del agrupamiento económico. Sin embargo, al examinar las características de los grupos resultantes, podemos concluir que esto se debe a las enormes desigualdades sociales que existen entre muchos países africanos con el resto del mundo.

Aunque la segmentación realizada permite capturar patrones más complejos al considerar una mezcla de indicadores de distintos ámbitos, los resultados dependen fuertemente del conjunto de datos seleccionado, por lo que cambios en ellos pueden alterar de manera significativa la clasificación obtenida. Además, aunque existe una gran cantidad de indicadores del *World Bank Data*, muchos no están disponibles para todos los países, lo cual representa un desafío adicional, puesto que las naciones con menos recursos económicos suelen tener sistemas de recolección estadística limitados, dificultando un análisis profundo y preciso. Por otra parte, el comportamiento de los países cambia con el tiempo, por lo que estos análisis deben realizarse de manera periódica para tomar decisiones con información actualizada.

En conclusión, la segmentación de países basada en indicadores globales es una herramienta esencial para comprender las complejas relaciones entre naciones y analizar el desarrollo mundial. Esta permite identificar grupos homogéneos de naciones, mejorar la comparación internacional y guiar decisiones de política pública fundamentadas en datos y adaptadas a contextos específicos. Aunque presenta limitaciones metodológicas, como la dificultad de validar los clusters o la dependencia de información incompleta, podemos obtener una visión estructurada del panorama global, complementando y enriqueciendo las clasificaciones tradicionales.

Referencias

- [1] World Bank. *About the World Bank*. n.d. URL: <https://data.worldbank.org/about>.
- [2] World Bank. *World Development Indicators*. n.d. URL: <https://datatopics.worldbank.org/world-development-indicators/>.
- [3] Carolina Saraiva y Jorge Caiado. “Global Development Patterns: A Clustering Analysis of Economic, Social and Environmental Indicators”. En: *Sustainable Futures* 10 (2025). DOI: 10.1016/j.sftr.2025.100907.
- [4] Charu C. Aggarwal y Chandan K. Reddy, eds. *Data Clustering: Algorithms and Applications*. First. Boca Raton, FL: Chapman y Hall/CRC, 2014.
- [5] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2.^a ed. Springer, 2009.