Pontificia Universidad Católica de Chile
Facultad de Economía y Administración

EAE3709
Aplicaciones de Machine Learning en Economía

Identificación de Perfiles de Países de la OCDE a partir del Better Life Index: Un Enfoque de Clustering y Análisis Comparado del Bienestar Subjetivo

---

Nombres:
Valentina Flores, Ignacio Maluenda & Constanza Muñoz

Profesores:
Joaquín Pérez

Ayudantes:
Luis & Oscar

30 de junio, 2025

Abstract

El presente estudio explora la segmentación de países de la OCDE a partir de indicadores objetivos del **Better Life Index**[1], utilizando técnicas de **machine learning no supervisado**. En particular, se aplican los algoritmos K-Means y Gaussian Mixture Models (GMM) para identificar grupos de países con perfiles estructurales similares en dimensiones como educación, salud, ingresos, seguridad, entre otras, y comparar la agrupación obtenida con la variable auto-reportada *Life Satisfaction*. Se incorpora además un análisis de componentes principales (PCA) para reducir la dimensionalidad y se consideran aspectos geográficos para enriquecer la interpretación de resultados con posibles patrones regionales. Los resultados muestran que ambos modelos logran diferenciar grupos con patrones significativos de desarrollo social, aunque K-Means presenta una mejor cohesión interna según el Silhouette Score. Finalmente, al comparar las agrupaciones obtenidas, se revela que los clusters con mayor bienestar estructural tienden a reportar mayores niveles de satisfacción subjetiva. Este análisis contribuye a una comprensión empírica multivariada del bienestar, y ofrece herramientas exploratorias para apoyar el diseño de políticas públicas orientadas a mejorar la calidad de vida.

1. Introducción

En las últimas décadas, el concepto de bienestar subjetivo se ha impuesto como una dimensión clave para comprender el desarrollo social de los países. Más allá de indicadores económicos tradicionales como el PIB per cápita, la percepción que tienen las personas sobre su propia calidad de vida permite capturar aspectos fundamentales del progreso de los países, incluyendo satisfacción más allá de lo económico como salud, empleo, seguridad o el entorno comunitario. Esta visión ha motivado la creación de índices compuestos como el Better Life Index de la OCDE, que añaden un factor subjetivo al análisis de bienestar. Este enfoque multidimensional plantea nuevos desafíos analíticos, especialmente al momento de identificar patrones o agrupaciones de países según su perfil de bienestar.

En este proyecto se aplican técnicas de machine learning no supervisado para segmentar países de la OCDE según variables objetivas del Better Life Index (OCDE), con el fin de identificar perfiles de desarrollo social y analizar su relación con la satisfacción de vida autorreportada (bienestar subjetivo). Utilizando los algoritmos K-Means y Gaussian Mixture Models (GMM), se busca detectar patrones comunes entre países y evaluar si los grupos formados presentan diferencias sistemáticas en el bienestar subjetivo.

La pregunta de investigación que guía este estudio es: ¿Existen perfiles similares de bienestar subjetivo que comparten los países de la OCDE?, y de ser así, ¿qué modelo de clustering, entre KMeans y GMM, genera una segmentación más coherente y explicativa respecto a las percepciones de bienestar subjetivo?

[1] https://www.oecdregionalwellbeing.org/

2. Metodología

Para abordar el análisis de segmentación de países en función de su bienestar, se utilizaron dos algoritmos de aprendizaje no supervisado: K-Means y Gaussian Mixture Models (GMM). Ambos modelos tienen como objetivo descubrir estructuras subyacentes en los datos no etiquetados.

K-Means es un algoritmo de partición que divide los datos en k grupos, minimizando la varianza intra-cluster. Asume que los clusters son esféricos, de tamaño similar y equidistantes, lo que lo hace eficiente pero sensible a la forma y distribución de los datos. En contraste, GMM es un modelo probabilístico basado en una combinación de distribuciones gaussianas. Permite una mayor flexibilidad al capturar clusters elípticos o de diferente densidad, asignando a cada observación una probabilidad de pertenencia a cada grupo.

La base de datos fue recuperada de la página de la OCDE. La recopilación de los datos comenzó en el año 2011, y se basa en encuestas y percepciones de los ciudadanos de los países miembros de la organización. Desde la página de Better Life Index, se puede interactuar con el dataset para comparar los distintos features para los distintos países. Se recolecta y se limpia la información desde el dataset en formato excel separando por región (sheet 1) y por país (sheet 2). El proyecto en el cual se enmarca la contrucción de esta base busca analizar cómo la vida está mejorando para los 38 países que participan hoy en la OCDE. En total son 6248 observaciones.

```
Categorias = ['Vivienda', 'Ingreso', 'Empleo', 'Comunidad', 'Educacion', 'Ambiental', '
    Participacion Civica', 'Salud', 'Seguridad', 'Acceso a servicios', 'Life satisfaction
    index']
```

Para efectos del modelo, la variable subjetiva 'Life Satisfaction' fue excluida del proceso de clustering y posteriormente utilizada para validación externa. Durante el proceso de preparación de datos, se implementaron varias decisiones de feature engineering de acuerdo con el análisis exploratorio y la naturaleza del conjunto. Para los missing values faltantes en algunas regiones decidimos utilizar la información facilitada por la institución, la que corresponde a un promedio con la información que tenían en ese momento. Hay países como Japón e Islandia en donde no existían datos ni por región ni por país, por lo que quedaron como faltantes. Se creó la base de datos modificada manualmente (sheet 3) en la que se incorpora la información de los primeros dos dataset (ver **Tabla 1** y **Figura 3**).

Primero, se descartaron las observaciones con valores faltantes, dado que los datos disponibles de la OCDE eran relativamente completos. En cuanto a los outliers, si bien se detectaron valores extremos en algunas variables a través de boxplots, no se aplicaron transformaciones ni imputaciones para suavizarlos. Esta decisión se basa en que los valores observados representan realidades nacionales y, por tanto, eliminar o modificar estos datos podría suprimir información sustantiva (ver **Figura 4**).

Adicionalmente, se crea una variable PCA_material a partir de la matriz de correlaciones (ver **Figura 5**). Las variables como educación, salud, ingreso, vivienda, empleo presentan alta correlación entre ellas, probablemente porque sean dimensiones socioecómicas importantes que determinan el bienestar de una persona. El PCA nos va a ayudar a resumirlas en menos

dimensiones (una nueva variable) sin perder tanta información. Así evitamos que las variables más parecidas sean "doble contabilizadas" cuando hagamos clustering, y además mantenemos variables como Health o Civic engagement que no están tan correlacionadas y podrían aportar algo distinto (ver **Tabla 2**). Posteriormente, se aplicó escalamiento estándar (StandardScaler) a las variables numéricas para asegurar que todas las dimensiones tuvieran igual peso en la formación de los clusters. Esto disminuye la distorsión generada por los outliers. Luego, se utilizó análisis de componentes principales (PCA) para reducir la dimensionalidad y facilitar la visualización de los resultados en un espacio bidimensional, sin perder una proporción significativa de la varianza total explicada.

Una vez finalizado el proceso de preparación de datos, se aplicaron dos criterios complementarios para determinar el número óptimo de clusters: el método del codo y el coeficiente de silueta. El primero sugirió una estabilización en la varianza explicada a partir de k = 5, mientras que el segundo indicó un valor máximo en k = 2 y en k = 5. Dado lo anterior, el enfoque exploratorio multivariado del estudio, se optó por k = 5 como un punto de equilibrio entre calidad de segmentación, heterogeneidad estructural y capacidad interpretativa de los perfiles generados (Ver **Figura 6**).

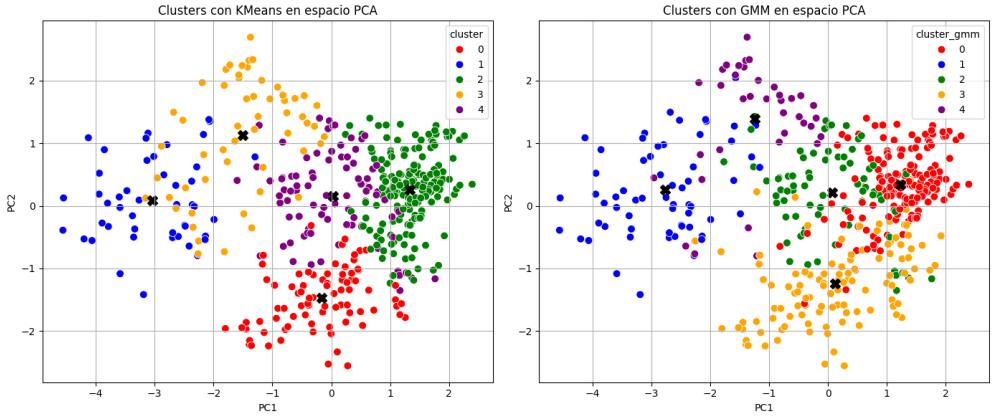
Con el valor óptimo de k definido, se entrenó el modelo K-Means utilizando las variables previamente estandarizadas. Paralelamente, se ajustó un modelo alternativo basado en Gaussian Mixture Models (GMM) con igual número de componentes, permitiendo comparar ambos algoritmos bajo condiciones similares. La visualización de los resultados se realizó mediante una proyección en dos dimensiones utilizando el Análisis de Componentes Principales (PCA), lo que facilitó la interpretación visual de la segmentación generada por ambos modelos.

3. Resultados

Se realiza una comparación usando el coeficiente de siluetas que mide qué tan similares son los elementos dentro de un mismo cluster (cohesión) en comparación con elementos de otros clusters (separación). Este índice toma valores entre -1 y 1, donde valores más altos indican mejor calidad de agrupamiento. Se obtienen los resultados: 0.3422 para KMeans y 0.2860 para GMM. Lo cual sugiere que el modelo KMeans proporciona una mejor segmentación global de los datos, en términos de separación y cohesión entre grupos. No obstante, al observar la estructura de los clusters proyectados en el espacio PCA y su comparación cruzada, esto es, comparar qué países captura cada grupo en cada modelo, se aprecia que ambos modelos capturan patrones similares en los datos, aunque con algunas diferencias notables en la asignación de subgrupos (Ver **Figura 1**).

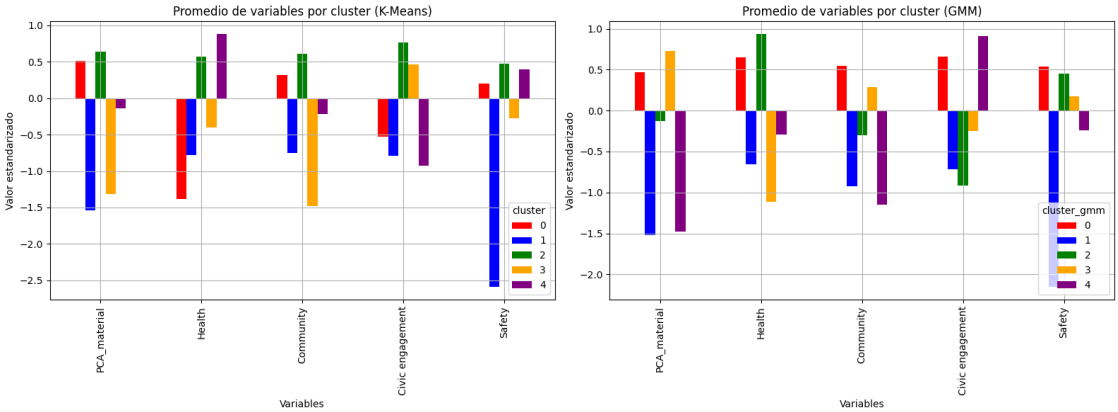
En particular, al comparar los clusters obtenidos por los dos algoritmos, se observa que el análisis de correspondencia entre los clusters generados por KMeans y GMM revela una relación aproximada entre los grupos formados por ambos algoritmos. Específicamente, el cluster 0 de KMeans se asemeja mayoritariamente al cluster 3 de GMM, el cluster 1 coincide casi exactamente con el cluster 1 de GMM, el cluster 2 se asocia principalmente con el cluster 0, el cluster 3 con el cluster 4, y el cluster 4 de KMeans con el cluster 2 de GMM. Esta correspondencia sugiere que ambos modelos capturan estructuras similares en los datos,

Figura 1: Comparación clusters - K Means y GMM



Nota: Elaboración propia.

Figura 2: Comparación clusters en ambos modelos



Nota: Elaboración propia.

aunque GMM tiende a subdividir algunos grupos identificados por KMeans, lo que podría reflejar una mayor sensibilidad a variaciones internas dentro de los clusters, a costa de una menor cohesión medida por el coeficiente de silueta.

Además, se realiza un análisis descriptivo de los perfiles de cada cluster en cada modelos con el objetivo de caracterizar las agrupaciones generadas. Para ello, se calculan los promedios de las variables consideradas dentro de cada grupo, lo que permite identificar patrones comunes en dimensiones clave del bienestar. Esta caracterización se complementa con información geográfica —como el país, la región y el continente— para enriquecer la interpretación de los resultados y vincular los perfiles emergentes con contextos territoriales específicos (Ver **Figura 2**).

La comparación de los promedios estandarizados de las dimensiones clave del bienestar (PCA_material, Health, Community, Civic engagement y Safety) para cada agrupación revela patrones consistentes entre los modelos K-Means y GMM, aunque con ciertas diferencias en la segmentación. En ambos casos, se identifican grupos de países con perfiles estructuralmente

similares que difieren en sus niveles relativos en las dimensiones evaluadas.

En el modelo K-Means, el cluster 2 exhibe valores positivos en todas las variables, lo que sugiere una agrupación de países con condiciones favorables tanto en dimensiones materiales como sociales. Este perfil de alto desempeño se replica parcialmente en el cluster 0 de GMM, aunque con diferencias más matizadas. De manera inversa, el cluster 1 de K-Means presenta valores consistentemente bajos, destacando como un grupo con deficiencias relativas en múltiples dimensiones. GMM, por su parte, descompone este patrón en subgrupos más específicos, siendo el cluster 4 el que concentra los niveles más bajos en Health, Civic engagement y Safety, lo cual indica una mayor sensibilidad del modelo a estructuras internas menos homogéneas.

En cuanto a los países agrupados, los resultados de ambos algoritmos reflejan patrones geográficos y estructurales consistentes con los niveles de desarrollo y características socioeconómicas de los países. Por ejemplo, tanto en K-Means (cluster 2) como en GMM (cluster 0) se agrupan países como Alemania, Suecia, Noruega, Australia, Canadá o Reino Unido, los cuales exhiben altos niveles en dimensiones como salud, condiciones materiales, participación cívica y seguridad. Este grupo representa economías desarrolladas con sistemas de bienestar consolidados y puntajes altos en el Better Life Index (ver **Tabla 3**)..

En el otro extremo, K-Means (cluster 1) y GMM (cluster 1 o 4) agrupan países como Colombia, Costa Rica y México, los cuales tienden a tener menores indicadores objetivos en comparación al resto de la OCDE. Estos países se caracterizan por desafíos estructurales persistentes, especialmente en seguridad, salud y condiciones materiales. Otros clusters reflejan perfiles mixtos. Por ejemplo, el cluster 4 de K-Means y el cluster 2 de GMM reúnen países como Chile, Grecia o Corea del Sur, que tienen niveles intermedios de bienestar: destacan en algunas dimensiones pero presentan rezagos en otras, como comunidad o participación cívica.

En síntesis, los agrupamientos permiten distinguir perfiles geográficos y estructurales claros: países nórdicos y de Europa Occidental en clusters de alto bienestar; países latinoamericanos en grupos de menor desempeño; y un conjunto intermedio de países con niveles de desarrollo más heterogéneos (ver **Tabla 4**).

Además de la evaluación interna, se incorporó una validación externa utilizando la variable auto-reportada "Life Satisfaction"del Better Life Index. Este análisis permitió contrastar los grupos formados por cada modelo con una medida subjetiva de bienestar no utilizada en el entrenamiento. Los resultados muestran que los clusters con mayores niveles en dimensiones objetivas tienden a tener, en promedio, mayores niveles de satisfacción con la vida. Por ejemplo, el cluster 2 de K-Means, caracterizado por altos niveles materiales, de salud y participación cívica, presentó un promedio de satisfacción de 7.90, seguido por el cluster 0 con 5.81. En contraste, el cluster 3, con los valores objetivos más bajos, tuvo una satisfacción promedio de apenas 2.27. De forma análoga, en el modelo GMM, el cluster 0, con mejor perfil objetivo, alcanzó un promedio de 7.65, mientras que el cluster 4, compuesto por países en condiciones más precarias, tuvo un promedio de 1.86. Estos resultados refuerzan la validez externa de los agrupamientos, al mostrar una correspondencia coherente entre las estructuras objetivas detectadas y las percepciones subjetivas de bienestar, lo que aporta robustez interpretativa y utilidad práctica a los perfiles identificados (ver **Tabla 5**).

4. Discusión

Aunque K-Means y GMM son técnicas comunes de clustering, responden a lógicas distintas: K-Means emplea asignación rígida, asignando cada observación a un único cluster mediante la minimización de la varianza intra-grupo, lo que lo hace eficiente. Mientras que GMM modela pertenencias probabilísticas de cada punto, lo que le permite capturar formas de cluster más complejas con formas elípticas o superposición entre grupos, aunque a costa de una mayor carga computacional y menor interpretabilidad directa. En este estudio, ambos modelos identificaron agrupaciones consistentes; sin embargo, K-Means mostró mayor cohesión interna (silhouette = 0.3422 vs. 0.2860). GMM, por su parte, logró detectar subestructuras más sutiles en ciertos grupos. Esta diferencia metodológica se reflejó también en la validación externa con la variable de satisfacción de vida, donde ambos modelos produjeron segmentaciones coherentes, aunque K-Means ofreció una jerarquización más clara de los perfiles.

En definitiva, los resultados respaldan el uso complementario de ambos enfoques. Las diferencias en los resultados entre K-Means y GMM provienen de sus fundamentos metodológicos: K-Means asigna puntos de forma rígida a un único cluster, mientras que GMM asigna probabilidades, permitiendo una mayor flexibilidad. En este estudio, K-Means mostró una mejor separación entre grupos según el coeficiente de silueta, mientras que GMM identificó subestructuras más complejas dentro de ciertos clusters. Esto sugiere que K-Means capta mejor agrupaciones bien definidas, mientras que GMM puede capturar gradientes o traslapes entre perfiles, lo que puede enriquecer la interpretación en contextos con alta heterogeneidad.

5. Conclusión

Este estudio exploró la capacidad de los modelos de clustering no supervisados, K-Means y Gaussian Mixture Models (GMM), para segmentar países de la OCDE en función de indicadores objetivos del bienestar provistos por el Better Life Index, y luego contrastar esas agrupaciones con la variable subjetiva de satisfacción con la vida (Life Satisfaction). La comparación entre ambos algoritmos reveló que K-Means obtuvo una mayor cohesión y separación entre grupos mientras que GMM permitió capturar patrones internos más complejos y sutiles dentro de los clusters. Ambos modelos produjeron agrupaciones similares en los niveles de bienestar subjetivo, lo cual valida externamente la coherencia de los perfiles generados.

En respuesta a la pregunta de investigación los resultados sugieren que sí existen perfiles similares de bienestar subjetivo en los países de la OCDE, y que estos perfiles pueden ser identificados mediante agrupamientos sobre variables objetivas del bienestar. Los países con altos niveles en dimensiones como salud, ingresos y seguridad tienden a compartir también niveles más altos de satisfacción de vida, aunque existen excepciones que podrían ser exploradas en mayor profundidad.

Para futuros análisis, se propone complementar el análisis con modelos supervisados, como Random Forest, para predecir directamente la satisfacción con la vida a partir de variables objetivas, evaluando su importancia relativa y posibles interacciones. Esto fue añadido como un acercamiento futuro a próximas extensiones en el código del proyecto.

Anexo

Estadísticas descriptivas del dataset

Tabla 1: Estadísticas descriptivas del dataset principal

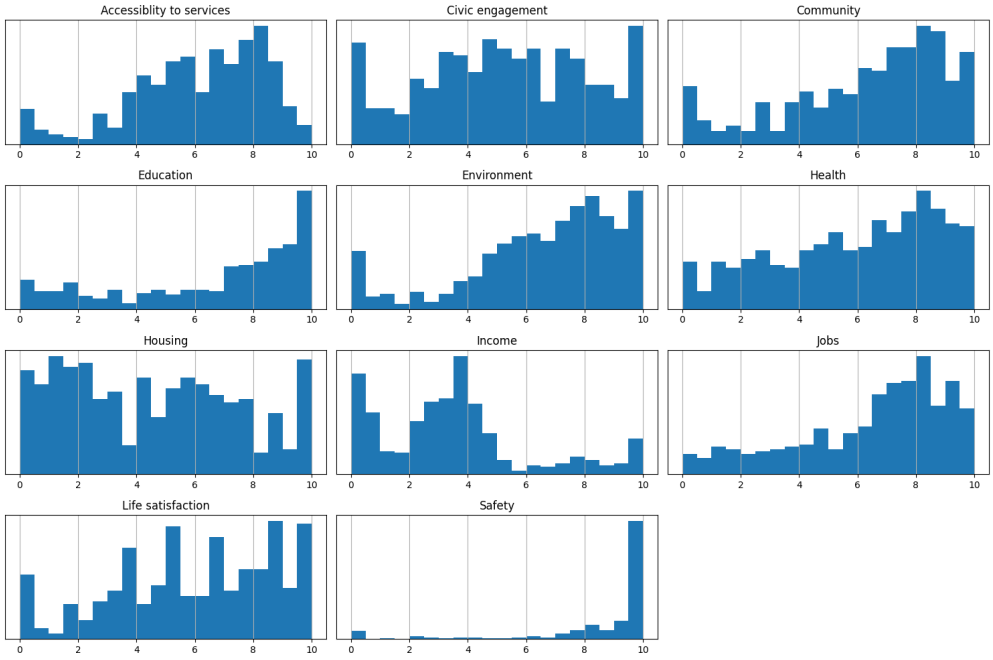
	Acceso a servicios	Participación Cívica	Comunidad	Educación	Ambiental
N	447	446	437	423	447
Media	6.09	5.22	6.29	6.59	6.59
SD	2.28	2.83	2.71	3.11	2.61
Min	0	0	0	0	0
25 %	4.75	3.2	4.6	4.45	5.2
50 %	6.5	5.2	7	7.7	7.1
75 %	7.9	7.48	8.4	9.2	8.5
Max	10	10	10	10	10

	Salud	Vivienda	Ingresos	Empleo	Seguridad	Life satisfaction index
N	447	444	435	432	444	437
Media	5.84	4.5	3.44	6.46	8.49	5.93
SD	2.76	2.95	2.57	2.58	2.47	2.76
Min	0	0	0	0	0	0
25 %	3.8	1.7	1.65	4.8	8.3	3.8
50 %	6.4	4.3	3.3	7.2	9.6	6.2
75 %	8.2	6.8	4.2	8.4	9.8	8.1
Max	10	10	10	10	10	10

Nota: Los datos provienen del Better Life Index 2024. Valores entre 0 y 10.

Distribución de los features

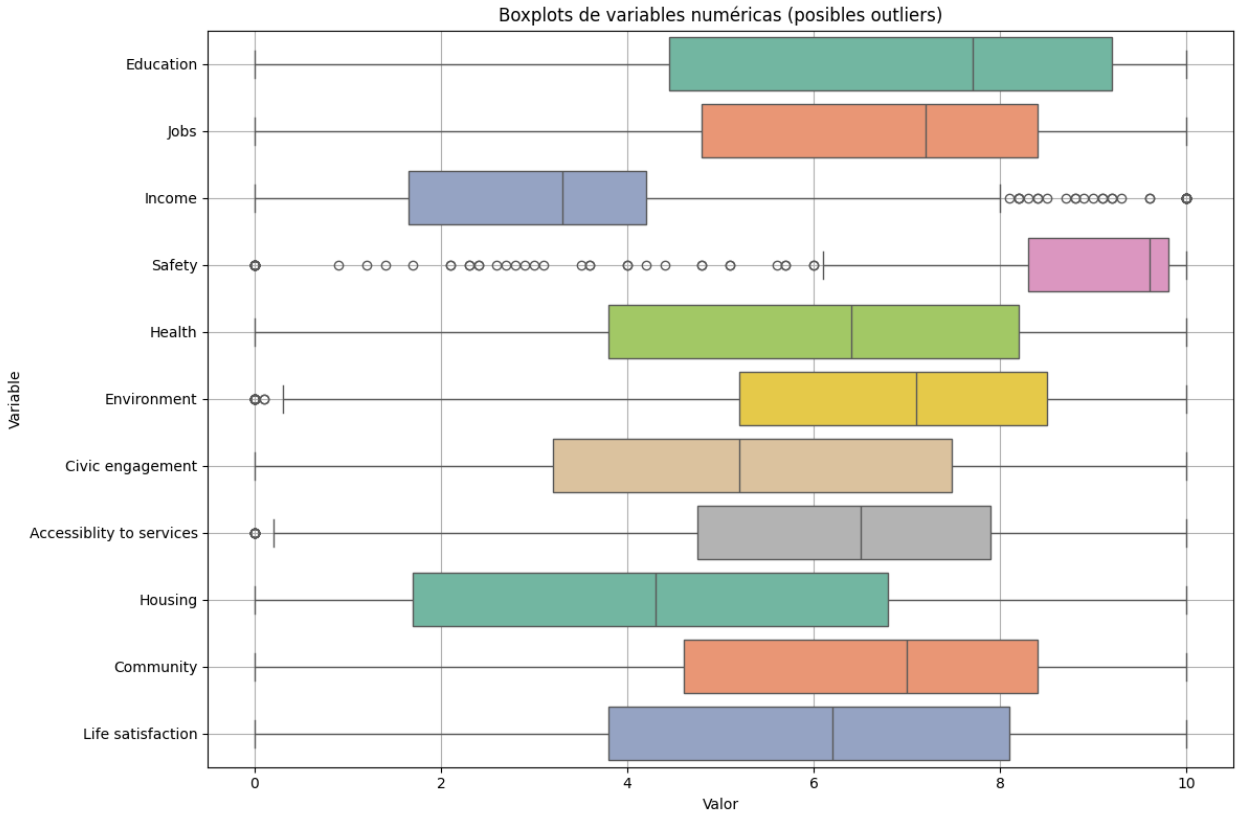
Figura 3: Distribución de puntajes por feature



Nota: Los datos provienen del Better Life Index 2024. Valores entre 0 y 10.

Visualización Outliers

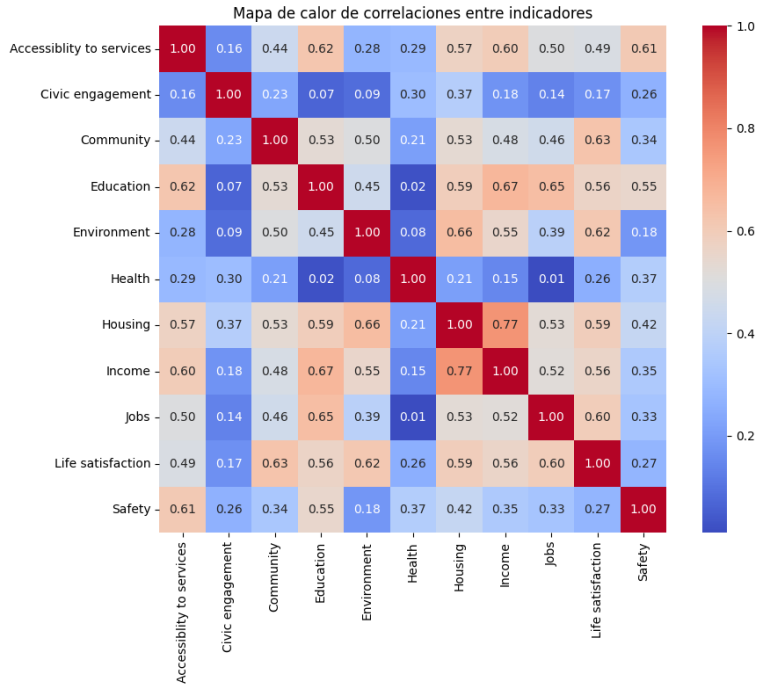
Figura 4: Boxplot - Valores extremos



Nota: Los datos provienen del Better Life Index 2024.

Matriz de correlaciones

Figura 5: Heatmap - Correlación entre features



Nota: Los datos provienen del Better Life Index 2024. Valores entre 0 y 10.

Estadísticas descriptivas del dataset al aplicar PCA

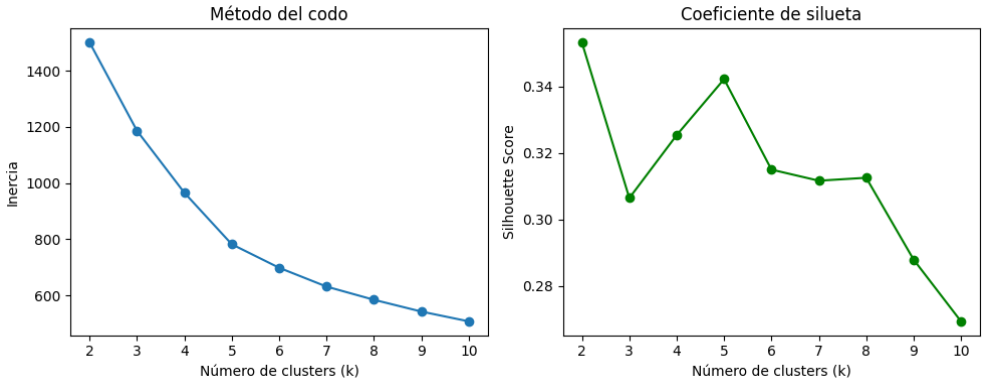
Tabla 2: Estadísticas descriptivas del dataset procesado

	PCA Material	Health	Community	Civic engagement	Safety
N	447	447	447	447	447
Media	0	5.84	6.31	5.21	8.48
SD	1.86	2.76	2.70	2.83	2.47
Min	-3.73	0	0	0	0
25 %	-1.51	3.8	4.6	3.2	8.3
50 %	0.51	6.4	7	5.2	9.6
75 %	1.22	8.2	8.4	7.45	9.8
Max	3.49	10	10	10	10

Nota: Elaboración propia.

Elección óptima de clusters

Figura 6: Elección óptima de clusters - Elbow Method y Sillhouette Score



Nota: Elaboración propia.

Tabla cluster según países

Tabla 3: Países por clusters para cada modelo

KMeans	Países (KMeans)	GMM similar	Países (GMM)
0	Canada, Czech Republic, Estonia, France, Hungary, Latvia, Lithuania, Poland, Portugal, Slovak Republic, Slovenia, United States	3	Canada, Czech Republic, Estonia, France, Hungary, Latvia, Lithuania, New Zealand, Poland, Slovak Republic, United States
1	Colombia, Costa Rica, Mexico	1	Colombia, Costa Rica, Mexico
2	Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Luxembourg, Netherlands, New Zealand, Norway, Spain, Sweden, United Kingdom, United States	0	Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Slovenia, Spain, Sweden, United Kingdom
3	Colombia, Costa Rica, Estonia, Greece, Italy, Mexico, Poland, Türkiye, United States	4	Colombia, Costa Rica, France, Mexico, Türkiye
4	Chile, Colombia, Costa Rica, Finland, France, Greece, Ireland, Israel, Italy, Korea, Portugal, Slovenia, Spain, Switzerland, United Kingdom, United States	2	Chile, Finland, France, Greece, Israel, Italy, Korea, Portugal, Spain, Switzerland, United States

Nota: Elaboración propia.

Zona geográfica (continentes) por clusters para cada modelo

Tabla 4: Zona geográfica (continentes) por clusters para cada modelo

Cluster KMeans	**Zonas Geográficas KMeans**	**Cluster GMM Similar**	**Zonas Geográficas GMM**
0	América del Norte, Europa	3	América del Norte, Europa, Oceanía
1	América del Sur, América Central, América del Norte	1	América del Sur, América Central, América del Norte
2	Oceanía, Europa, América del Norte, Asia Occidental	0	Oceanía, Europa, América del Norte, Asia Occidental
3	América del Sur, América Central, Europa, América del Norte, Asia Occidental	4	América del Sur, América Central, Europa, América del Norte, Asia Occidental
4	América del Sur, América Central, Europa, Asia Occidental, Asia Oriental, América del Norte	2	América del Sur, Europa, Asia Occidental, Asia Oriental, América del Norte

Nota: Elaboración propia.

Promedio de Life Satisfaction por clusters para cada modelo

Tabla 5: Promedio de Life Satisfaction por clusters para cada modelo

Cluster KMeans	Promedio LS KMeans	Cluster GMM Similar	Promedio LS GMM
0	5.812871	3	6.342191
1	4.386957	1	4.254098
2	7.901314	0	7.650617
3	2.279322	4	1.862001
4	5.169711	2	4.975921

Nota: Elaboración propia.