

CONSTRUCCIÓN DE UN INDICE DE PROGRESO SOCIAL EUROPEO: SPI-EU 2024

Pablo Montalvo Sanz

Master Data Science, Big Data & Business Analytics 2023-2024

Universidad Complutense de Madrid

<https://github.com/pmntlv/SPI-EU-2024.git>

Tabla de contenidos

Resumen	4
Introducción	4
Indicadores de Desarrollo Social	5
Fase I. Lectura y análisis descriptivo de dataset.....	6
I.1. PIB per cápita:	6
I.2. Datos perdidos:	6
I.3. Estadísticas descriptivas:	6
I.4. Outliers:	7
I.5. Distribución de los indicadores:	7
I.6. Correlación entre indicadores:	7
Fase II. Limpieza y normalización	8
II.1. Imputación de datos perdidos:	8
II.2. Consistencia del dataset:	8
II.3. Normalización:	9
Fase III. Agregación por componentes	10
III.1. Principales Consideraciones:	10
III.2. Cálculo de la Media Generalizada:	10
III.3. Análisis de Componentes Principales:	11
III.4. Análisis de sensibilidad:	11
III.5. Componentes por regiones:	12
III.6. Componentes y PIB per cápita:	12
Fase IV. Agregación por dimensiones	12
IV.1. Agregación:	12
IV.2. Análisis:	12
IV.3. Dimensiones y PIB per cápita:	13
Fase V. Agregación SPI-EU	13
V.1. Análisis:	13
V.2. SPI-EU por países:	13
V.3. SPI-EU por regiones:	14
V.4. GDP per cápita:	14
Fase VI. Clustering	15
VI.1. Ranking mejor y peor desempeño:	15
VI.2. Comparación SPI y PIB per Cápita	15
VI.3. Clasificación por SPI y PIB per cápita:	16
VI.4. K-Means Clustering:	17

Clasificación de 4 grupos:	17
Clasificación de 5 grupos:	17
VI.5. Clasificación por dimensiones:.....	18
Líneas de investigación futura	18
Conclusiones	18
Cuadro de Mando y Visualizaciones	19
Referencias	19

Resumen

El SPI-EU, o Índice de Progreso Social de la Unión Europea (Social Progress Index - EU¹), es una herramienta que mide el desarrollo social y bienestar en la Unión Europea a nivel de regiones (NUTS 2²). Este índice evalúa el progreso social en términos de calidad de vida, acceso a derechos y oportunidades, y sostenibilidad, proporcionando una visión más completa del desarrollo social en comparación con métricas económicas tradicionales, como el Producto Interior Bruto (PIB).

El objetivo de este trabajo es la construcción de un índice sintético compuesto, inspirado en el SPI-EU y desarrollado mediante código en Python. Se emplea un enfoque de análisis multivariante que incluye la normalización de los datos, el Análisis de Componentes Principales (PCA) y la agregación de resultados en distintos niveles. Los datos utilizados provienen de fuentes oficiales europeas³, lo que permite realizar comparaciones interregionales y ofrece una herramienta útil para la evaluación del progreso social en la Unión Europea.

Introducción

Este trabajo se divide en seis fases, para cada una de las cuales se ha creado un Jupyter Notebook que contiene el código Python necesario, generando un archivo CSV como resultado. En cada fase, el notebook correspondiente lee el CSV resultante de la fase anterior y guarda un nuevo CSV, que será leído en la fase siguiente, con excepción de la primera y la última fase.

Los cuadernos de Jupyter se presentan como anexos en formato IPYNB y HTML, junto con los datasets y archivos CSV correspondientes.

Debido al gran número de visualizaciones y cuadros generados durante el proceso de elaboración del índice, y considerando la limitación en la extensión de este trabajo, se ha decidido no incluir gráficos ni tablas en el cuerpo principal del informe. A lo largo del mismo, se hacen referencias a "figuras" que se encuentran en los anexos (gráficos, tablas, mapas, etc.).

Toda la documentación y anexos se encuentran en el repositorio de Github:

<https://github.com/pmntlv/SPI-EU-2024.git>

¹https://ec.europa.eu/regional_policy/information-sources/maps/social-progress_en

² <https://ec.europa.eu/eurostat/web/nuts>

³ <https://composite-indicators.jrc.ec.europa.eu/explorer/explorer/indices/eu-r-spi/eu-regional-social-progress-index>

Indicadores de Desarrollo Social

La Comisión Europea define 53 indicadores sociales de distinta naturaleza y medición, que se agrupan en 12 componentes según su naturaleza. A su vez, estos componentes se organizan en tres dimensiones:

Dimension	Component	Indicator
Basic needs	Nutrition and medical care	Infant Mortality
		Unmet dental needs
		Insufficient food
	Water and Sanitation	Satisfaction with water quality Lack of toilet in dwelling Uncollected sewage Sewage treatment
Foundations of wellbeing	Housing	Burdensome cost of housing Housing quality - dampness Overcrowding Lack of adequate heating
	Safety	Traffic deaths
		Safety at night
		Money stolen
	Basic education	Assaulted/Mugged
		Share of low-achieving 15-year-old in reading (level 1a or lower)
		Share of low-achieving 15-year-old in maths (level 2 or lower)
	Information and communications	Lower-secondary completion only
		Early school leavers
		High-speed broadband
	Health	Digital skills above the basic level
		Online interaction with public authorities
		Freedom of media
Opportunity	Environmental quality	Internet access
		Years of life lost caused by PM2.5, NO2 and ozone (SOMO35)
		Subjective health status
	Trust and governance	Standardised cancer death rate
		Standardised heart disease death rate
		Positive feelings
	Freedom and choice	NO2 air pollution
		Ozone (SOMO35) air pollution
		PM2.5 air pollution
	Inclusive society	Bathing water quality
		Trust in the national government
		Trust in the judicial system
	Advanced education	Trust in the police
		Voiced your opinion to a public official
		Female participation in regional assemblies
	Tolerance and respect	Institution quality index
		Freedom over life choices
		Job opportunities
	Lifelong learning	Teenage pregnancy
		Young people not in education, employment or training (NEET)
		Institutions corruption index

Fase I. Lectura y análisis descriptivo de dataset

Anexo:

- 00_Lectura y EDA.ipynb
- 00_Lectura y EDA.html

El dataset contiene 242 entradas, una por cada región, y 48 columnas, de las cuales 3 corresponden al nombre de la región, del país y al código NUTS respectivo. Por tanto, solo contamos con 45 indicadores de los 53 mencionados anteriormente.

I.1. PIB per cápita:

Hemos añadido al dataset un indicador adicional: PIB per cápita 2022⁴. Aunque trabajamos con índices de desarrollo en su edición de 2024, solo disponemos del PIB per cápita hasta 2022. Consideramos que el PIB debería influir positivamente en el desarrollo social de un país o región, ya que una mayor capacidad económica debería implicar una mayor inversión y mejora de las condiciones de vida en todas sus facetas. Sin embargo, no se puede hablar de un efecto inmediato, ya que se requiere un proceso de transformación que puede llevar tiempo, incluso años, y no se garantiza el mismo efecto positivo en todas las regiones. Consideraremos válido, para este trabajo, comparar el PIB per cápita de 2022 con los indicadores sociales de 2024.

I.2. Datos perdidos:

Los indicadores presentan algunos datos perdidos. Un análisis por indicador revela que no hay más de un 3.5% de valores faltantes en cada uno, como se observa en la **Figura 00.1** del Anexo.

Sin embargo, hay dos excepciones:

- Bathing water quality: 10%
- Share of low-achieving 15 years olds in reading (level 1a or lower): 7%

En la **Figura 00.2** se observa que algunos países presentan un 100% de datos faltantes en determinados indicadores:

- “Share of low-achieving 15 years old in reading...” en España.
- “Digital skills above basic level” (CY, EE, LU, LV, MT)

I.3. Estadísticas descriptivas:

En la **Figura 00.3** se presentan las principales estadísticas descriptivas de los 45 indicadores y del PIB per cápita. En este punto, no haremos una descripción detallada de todos los indicadores, pero es importante destacar la heterogeneidad de las métricas, lo que hace necesaria la normalización de los datos. Por ejemplo “Sewage

⁴https://ec.europa.eu/eurostat/databrowser/view/nama_10r_2gdp/default/table?lang=en

treatment, additional” tiene un formato de índice 0 - 100 mientras que “Infant mortality” tiene un rango de 0 - 9.70 y se refiere a número de eventos en un periodo de tiempo.

I.4. Outliers:

Los "outliers" son datos extremos que se alejan significativamente de la media en una distribución. Normalmente, estos valores son indicativos de posibles errores en la medición. Un análisis mediante Z-Score permite medir cuántas desviaciones estándar se encuentra un dato por encima o por debajo de la media de la distribución.

$Z = (X - \mu) / \sigma$; donde un Z-Score superior a 3 o inferior a -3 generalmente se considera un outlier.

Otro método para detectar valores atípicos es el análisis del rango intercuartílico (IQR). Véase **Figura 00.5** y **Figura 00.6** del Anexo.

La **Figura 00.7** muestra un gráfico box-plot de cada indicador por país y su referencia a la media de la UE. Los puntos más alejados, por encima o por debajo de los "bigotes", representan regiones de cada país en las que el indicador correspondiente muestra un valor inusual.

En este trabajo, se ha decidido no tratar estos valores atípicos como errores. Por un lado, confiamos en el trabajo de los estadísticos de Eurostat. Por otro lado, es razonable pensar que existen unas pocas regiones en cada país con datos significativamente diferentes a la media debido a razones geográficas o sociales.

En total, se observan outliers en 31 regiones de 11 países. En el mapa representado por la **Figura 00.8** se observa que la mayoría se concentran en regiones muy específicas:

- Europa del Este y Balcanes: Bulgaria, Rumanía, Croacia
- Irlanda
- Territorios de ultramar de Francia y Portugal

I.5. Distribución de los indicadores:

Los gráficos de distribución representados en la **Figura 00.9** del Anexo, nos permite recalcar la heterogeneidad de los rangos y la existencia de valores atípicos en algunos indicadores.

I.6. Correlación entre indicadores:

En el mapa de calor representado en la **Figura 00.10** del Anexo, se observan las correlaciones entre indicadores en una escala de -1 a 1. Los valores más cercanos a 1 (rojo más intenso) indican correlaciones positivas más fuertes; los valores más cercanos a -1 (azul más intenso) indican correlaciones negativas más fuertes; y los valores más cercanos a 0 (colores más tenues) indican poca o nula correlación. En general, se observan correlaciones positivas fuertes entre grupos de indicadores de la misma naturaleza. Algunos indicadores, como “Money Stolen” o “Subjective health status”, presentan muy poca correlación con el resto. Estas relaciones deberán ser observadas nuevamente después de la normalización de los datos.

Fase II. Limpieza y normalización

Anexo:

- 01_limpieza_normalizacion.ipynb
- 01_limpieza_normalizacion.html

Primero estandarizamos y renombramos los indicadores: eliminamos signos de puntuación, espacios, y abreviamos algunos nombres que eran demasiado largos.

II.1. Imputación de datos perdidos:

Descartamos la eliminación de datos, ya que cada registro representa una región y no podemos eliminar una región entera por un dato perdido. Además, ningún indicador presenta un porcentaje de datos perdidos lo suficientemente alto como para justificar su eliminación. Por lo tanto, se procede a imputar.

Existen muchas estrategias para la imputación de datos, como el uso de la media de la muestra, k-vecinos más cercanos, regresiones y técnicas de aprendizaje automático, entre otras. En este trabajo optamos por la imputación simple por la media, debido a que es un método sencillo y fácilmente interpretable. Sin embargo, no sería razonable imputar utilizando la media de toda la muestra, ya que pueden existir diferencias significativas entre países. Imputamos el dato perdido de un indicador, en una región, utilizando la media del país al cual pertenece esa región.

Algunos indicadores presentan un 100% de datos perdidos en ciertos países, especialmente en países pequeños con una o pocas regiones (exceptuando el caso de España). En estos casos, imputamos utilizando la media de la UE:

- digital_skills_above_basic:
 - CY: Chipre
 - EE: Estonia
 - LU: Luxemburgo
 - LV: Letonia
 - MT: Malta
- low_achieving_reading_level: España

La estrategia consiste en realizar la imputación en dos fases: primero, imputar los valores perdidos dejando aparte los países donde hay alguna variable completamente vacía, y en una segunda fase, imputar los valores perdidos de estos países considerando la media de la muestra restante (y ya sin datos perdidos).

II.2. Consistencia del dataset:

Antes de proceder a la construcción del índice, es fundamental evaluar la consistencia y adecuación de los datos mediante pruebas estadísticas específicas. Para ello, se utilizan dos herramientas: el **test de esfericidad de Bartlett** y el **Índice de Adecuación Muestral de Kaiser-Meyer-Olkin (KMO)**. Ambas pruebas permiten verificar si el conjunto de indicadores es adecuado para ser combinado en un único índice.

Utilizamos la prueba de esfericidad de Bartlett para determinar si los indicadores están suficientemente correlacionados entre sí. La prueba compara la matriz de correlaciones de Pearson con una matriz identidad, donde las correlaciones fuera de la diagonal principal son cero. El estadístico de prueba resultante sigue una distribución Chi-cuadrado. Si el valor p es menor que el nivel de significancia (por ejemplo, 0.05), se rechaza la hipótesis nula de que los indicadores no están correlacionados. En nuestro caso la prueba arroja un estadístico de 12691.98 con un valor p de 0.0. Este resultado indica que la hipótesis nula debe ser rechazada, lo que nos permite concluir que los indicadores están significativamente correlacionados entre sí.

El Índice de Adecuación Muestral de Kaiser-Meyer-Olkin (KMO) compara la correlación observada entre pares de indicadores con sus correlaciones parciales, es decir, después de eliminar el efecto de otros indicadores. Un valor KMO superior a 0.5 y cercano a 1 indica que las variables están bien correlacionadas y es adecuado construir un índice compuesto. En este caso, el índice global es aproximadamente 0.89, lo que sugiere que las correlaciones entre los indicadores son fuertes y apropiadas para la combinación en un índice compuesto.

II.3. Normalización:

Para normalizar los datos en una escala de 0 a 100, utilizamos el método de normalización “min-max”, que ajusta los valores a un rango específico basado en los valores mínimo y máximo de cada indicador. Esta normalización se realiza teniendo en cuenta la orientación del indicador: si está orientado “positivamente” o “negativamente”. Por ejemplo, para un indicador orientado negativamente, como el número de fallecimientos por accidentes de tráfico, los valores más altos se normalizarán hacia el extremo inferior de la escala (cercano a 0), mientras que los valores más bajos se normalizarán hacia el extremo superior (cercano a 100). En contraste, para un indicador orientado positivamente, como el índice de satisfacción con la calidad del agua, los valores más altos se normalizarán hacia el extremo superior (cercano a 100), y los valores más bajos hacia el extremo inferior (cercano a 0). Este enfoque asegura que todos los indicadores sean comparables en una escala uniforme. La fórmula es la siguiente:

$$x_{norm} = \begin{cases} \frac{100 (x - x_{min})}{(x_{max} - x_{min})} & \text{if } x \text{ is positively oriented} \\ \frac{-100 (x - x_{min})}{(x_{max} - x_{min})} + 100 & \text{if } x \text{ is negatively oriented} \end{cases}$$

Las **Figuras 01.1** y **01.2** muestran el mapa de correlaciones entre indicadores antes y después de la normalización. Se observa como muchas correlaciones negativas se vuelven positivas tras la normalización. Un detalle interesante es el indicador “insitutions corruption” que surge con una correlación muy negativa con casi todos los indicadores.

Una alternativa de normalización es la transformación **Z-Score**, aunque este trabajo se centra en el primer método, se ha de tener en cuenta en futuras líneas de investigación.

Fase III. Agregación por componentes

Anexo:

- 02_componentes_agregacion.ipynb
- 02_componentes_agregacion.html

El **Índice de Progreso Social EU-SPI** utiliza un método de agregación híbrido para combinar los datos de diferentes dimensiones y componentes. Este método incluye:

- **Media aritmética simple** para agregar los indicadores en componentes.
- **Media generalizada** para combinar los componentes en dimensiones y, finalmente, estas en el índice final.

En ambos casos, la Comisión Europea opta por no utilizar ponderaciones (edición 2020)⁵, lo que significa que cada indicador tiene el mismo peso en el cálculo del índice. Este enfoque busca reflejar de manera equitativa el desempeño en todas las dimensiones del progreso social, alineando el índice con el objetivo de medir un progreso social equilibrado y uniforme.

III.1. Principales Consideraciones:

- **Desempeño Equitativo:** Se considera que una región proporciona un nivel adecuado de progreso social solo si obtiene un desempeño satisfactorio en todas las dimensiones evaluadas. Esto asegura que el índice no favorezca a regiones que solo destacan en ciertas áreas mientras descuida otras.
- **Evitar Compensaciones:** Al no utilizar ponderaciones, se evita que un alto desempeño en una dimensión pueda compensar un bajo desempeño en otra. Sin ponderaciones, no hay riesgo de que una dimensión "menos importante" sea compensada por una "más importante", lo que podría ocurrir si se emplearan pesos. Este método garantiza que todas las dimensiones sean igualmente importantes en la evaluación del progreso social.

III.2. Cálculo de la Media Generalizada:

El índice correspondiente a una región, denotado por I_j se calcula como la media generalizada no ponderada de orden β de q indicadores, componentes o dimensiones

$$I_j^{(\beta)} = \begin{cases} \left(\frac{1}{q} \sum_{j=1}^q x_{ji}^{\beta} \right)^{1/\beta} & \beta \neq 0 \\ \left(\prod_{j=1}^q x_{ji} \right)^{1/q} & \text{for } \beta = 0 \text{ (geometric mean)} \end{cases}$$

⁵ https://ec.europa.eu/regional_policy/sources/work/202006_spi_en.pdf

En la ecuación, β representa una constante que se puede establecer en diferentes niveles para ajustar el grado de compensabilidad entre los elementos en el índice. Cuando $\beta = 1$, la media generalizada es igual a la media aritmética y la compensación entre categorías es perfecta. Para $\beta = 0$, la media generalizada es la media geométrica, por lo que una calificación muy baja en una categoría afecta severamente la calificación general. Para $\beta = 0.5$ la agregación permite alguna compensación, pero no totalmente.

- **Agregación de indicadores:** utiliza media aritmética simple ($\beta=1$)
- **Agregación de componentes y dimensiones:** utiliza una media generalizada con un valor de $\beta = 0.5$ (Media generalizada con compensación parcial)

La media generalizada entre componentes y dimensiones limita la capacidad de compensar un mal desempeño en un área con un buen desempeño en un área completamente diferente.

III.3. Análisis de Componentes Principales:

Antes de proceder con la agregación de los indicadores normalizados, es crucial evaluar la consistencia interna mediante un análisis de componentes principales (PCA). Este análisis permite verificar que los indicadores seleccionados sean robustos y significativos, asegurando que la agregación sea válida. Un alto grado de consistencia interna minimiza la compensabilidad entre los indicadores.

El PCA identifica el primer componente principal (PC1) como la combinación lineal de los indicadores que explica la mayor parte de la variabilidad en los datos. En este trabajo, los indicadores han sido seleccionados en función de sus cargas absolutas en el PC1, descartando aquellos que no superan un umbral específico. La elección de este umbral es subjetiva y, en este caso, se ha fijado en 0.4, lo que permite incluir indicadores con una contribución moderada a la variabilidad principal, asegurando una representación equilibrada y robusta en el índice compuesto.

Véase por ejemplo la **Figura 02.6**: tras analizar el componente "Safety" descartamos los indicadores "safety at night" y "traffic deaths" por tener cargas absolutas inferiores a nuestro umbral.

III.4. Análisis de sensibilidad:

Se realiza un análisis de sensibilidad para evaluar la robustez del método de agregación mediante tres enfoques:

1. **Análisis de Distintos β :** Se explora la sensibilidad del método de agregación al variar los valores de β en la media generalizada. Este análisis permite comparar cómo diferentes valores de β afectan la distribución de los resultados del índice compuesto. Por ejemplo, el componente "information and communications" es sensible a este parámetro, como se observa en la **Figura 02.18**. En general, valores menores de β otorgan mayor peso a los resultados bajos, lo que produce un índice promedio más bajo. Por el contrario, si se aumenta el valor de β , el índice promedio aumenta.
2. **Añadir o Eliminar Indicadores:** A partir del conjunto de datos completo, con todos los indicadores, se repite el proceso de agregación eliminando un indicador cada vez. Este enfoque permite evaluar el impacto de cada indicador en el resultado final del índice compuesto. Como se muestra en la **Figura 02.19**, eliminando "broadband at home" del componente "info comm", la distribución se desplaza, cambiando el valor promedio. Por ejemplo, al excluir "digital skills", el índice mejora, lo que indica que el componente es sensible a este indicador y que, en general, las regiones obtienen malos resultados en ese indicador.

3. **Bootstrapping:** Con este método se genera una gran cantidad de muestras (1000) mediante remuestreo con reemplazo de la muestra original. Para cada muestra bootstrap, se repite el proceso de agregación, lo que permite estimar la distribución de la media del índice compuesto. Este método proporciona una medida de la estabilidad y precisión del índice al evaluar cómo varía con diferentes muestras y estimar la incertidumbre asociada a la media. Continuando con el componente "info comm" (**Figura 02.20**), se observa que la media y la varianza obtenidas mediante bootstrap no se alejan significativamente de los datos originales. En el gráfico, se aprecia que la distribución obtenida mediante bootstrap es bastante irregular, mientras que la distribución original parece ajustarse bien a la media de los datos remuestreados.

III.5. Componentes por regiones:

Una vez obtenidos los índices compuestos a nivel de componentes, observamos en la **Figura 02.56** un mapa que muestra los resultados por regiones. En general, las regiones con los peores resultados se concentran en Europa del Este y los Balcanes, mientras que los mejores resultados se observan en los países escandinavos.

III.6. Componentes y PIB per cápita:

En **Figura 02.57** y **Figura 02.58** observamos la relación de cada componente con el PIB per cápita de 2022. En general, existe una relación positiva, excepto en el caso del componente "Safety", donde la relación es baja o incluso negativa. Las relaciones más fuertes se encuentran en los componentes de "advance education" y "trust in government". En resumen, podríamos concluir que el PIB tiene una menor influencia en los componentes relacionados con el desarrollo más básico.

Fase IV. Agregación por dimensiones

Anexo:

- 03_dimensiones_agregacion_minmax.ipynb
- 03_dimensiones_agregacion_minmax.html

IV.1. Agregación:

En esta fase, el proceso de agregación es similar a la anterior con dos excepciones:

1. Agregación mediante medias generalizadas con factor $\beta = 0.5$; Como se indicó en la fase anterior. Esta agregación está a medio camino de la media aritmética (compensación completa) y la media geométrica.
2. No seleccionamos componentes mediante el análisis de componentes principales (PCA): en la fase anterior ya eliminamos los indicadores menos relevantes y no aplicamos ponderaciones a componentes ni dimensiones.

IV.2. Análisis:

En las **Figura 03.10**, **Figura 03.11** y **Figura 03.13** del Anexo, se observa que los índices resultantes presentan valores altos en "Basic Needs" y "Foundations of Wellbeing". Estos índices tienen distribuciones con un leve sesgo hacia la izquierda, donde la media es menor que la mediana, lo que indica que algunos valores bajos (outliers) están afectando los promedios. La dimensión "Opportunity" muestra una distribución más equilibrada.

El índice de "Foundations of Wellbeing" y el de "Opportunity" presentan una correlación muy alta (0.86 sobre 1) según se observa en la **Figura 03.12**, mientras que no se observa una correlación tan fuerte con el índice de "Basic Needs".

La **Figura 03.13** muestra cómo se distribuye cada dimensión en los diferentes países, permitiendo identificar aquellos con peor desempeño. Como dato interesante, se destacan los outliers en Francia hacia la izquierda (valores más bajos) en el índice de "Basic Needs", probablemente relacionados con sus territorios de ultramar.

En las **Figuras 03.14** y **03.15** visualizamos la distribución de las tres dimensiones en el mapa de las regiones europeas. En general, para las tres dimensiones se observan valores más bajos en Europa del Este y los Balcanes, y valores más altos en el norte de Europa. Esta distribución se hace más evidente a medida que se avanza en desarrollo social: se encuentran puntuaciones altas en "Basic Needs" en zonas muy diversas como Europa Central, España, Italia y Grecia, mientras que en "Opportunity" casi todo el continente muestra puntuaciones medias, destacando claramente las zonas con mayor y menor desarrollo (las colas de la distribución).

IV.3. Dimensiones y PIB per cápita:

Como se indicó en la fase anterior, la correlación entre el PIB per cápita y los índices de desarrollo social aumenta a medida que se superan las necesidades más básicas y se avanza hacia objetivos más complejos. En las **Figuras 03.17** y **03.18** del Anexo se puede apreciar que la correlación de Pearson para las tres dimensiones es de 0.273 para "Basic Needs", 0.641 para "Foundations of Wellbeing" y 0.726 para "Opportunity".

Fase V. Agregación SPI-EU

Anexo:

- 04_SPI_agregacion_minmax.ipynb
- 04_SPI_agregacion_minmax.html

En esta fase, las dimensiones se agregan en un índice final por región, utilizando la media generalizada con un factor $\beta = 0.5$, de forma similar a la fase anterior.

V.1. Análisis:

La distribución del SPI presenta un ligero sesgo hacia la izquierda. Aunque muchas regiones alcanzan un nivel medio-alto de progreso social, una cuarta parte de las regiones exhibe niveles significativamente inferiores a la media, lo que refleja una disparidad notable en el desarrollo social entre las diferentes regiones de la UE. Esto se puede observar en las **Figuras 04.4** y **04.5** del Anexo.

V.2. SPI-EU por países:

La **Figura 04.6** muestra un gráfico de caja (box-plot) que representa la distribución del SPI y la media general por país. Los países presentan distribuciones dispares, aunque la mayoría se sitúan en un rango de puntuaciones entre 50 y 70, con una media general de 60.89. Se identifican fácilmente los países con peor desempeño: Bulgaria (BG), Croacia (HR), Rumanía (RO) y Eslovaquia (SK). Por el contrario, los países con mejor desempeño son Dinamarca (DK), Finlandia (FI), Países Bajos (NL) y Suecia (SE).

Un aspecto interesante de este gráfico es la distribución de Francia (FR), que está centrada en la media, pero con cuatro regiones que destacan por tener puntuaciones muy bajas (outliers), incluso por debajo de 40. Algo similar, aunque en menor medida, ocurre con Italia (IT). Por el contrario, Eslovaquia muestra puntuaciones medias-bajas en general, pero con una región que se destaca por encima de la media.

V.3. SPI-EU por regiones:

La **Figura 04.7** presenta varios mapas que muestran la distribución del SPI por regiones (excluyendo las regiones de ultramar para facilitar la representación), utilizando cuatro métodos distintos para clasificar los datos:

- **Equal Interval:** El rango de datos se divide en intervalos iguales, y cada intervalo representa una categoría en el mapa. Las regiones con las puntuaciones más altas y bajas se representan con colores más intensos (verde y marrón). Este método ofrece una clasificación simétrica y fácil de interpretar, con tres intervalos de puntuaciones bajas a medias, dos intervalos medios y tres intervalos de puntuaciones medias a altas.
- **Quantiles:** Este método divide los datos de manera que cada intervalo contenga aproximadamente la misma cantidad de datos, aunque los rangos de los intervalos pueden variar. Sin embargo, no es el mejor método para representar el índice, ya que concentra los datos más bajos en una sola categoría (24-48), debido al sesgo hacia la izquierda de la distribución.
- **Natural breaks:** Esta técnica agrupa los datos en clases que minimizan la variación dentro de cada clase y maximizan la diferencia entre clases. Aunque este método genera una clasificación algo asimétrica, es menos extrema que la de los cuantiles.
- **Std Mean:** Esta técnica utiliza las desviaciones estándar con respecto a la media para visualizar cómo se distribuyen los datos en relación con la media. Es un método excelente para identificar rápidamente los grupos más extremos en la distribución. En este caso, permite diferenciar claramente las regiones con los valores más bajos y altos del SPI.

V.4. GDP per cápita:

El coeficiente de correlación de Pearson entre el PIB per cápita y el SPI es de 0.677 (**Figura 04.9**), lo que indica una fuerte correlación positiva entre ambos índices. Aunque la relación no es perfectamente lineal, como se puede observar en la **Figura 04.8**, en términos generales, un mayor PIB tiende a estar asociado con un mejor desempeño en el SPI.

Fase VI. Clustering

Anexo:

- 05_Clustering_minmax.ipynb
- 05_Clustering_minmax.html

VI.1. Ranking mejor y peor desempeño:

Una utilidad del índice compuesto que hemos creado sería clasificar o segmentar las regiones y países de la UE, en función de su score. La primera y más inmediata clasificación es listar los países y regiones con mejor y peor SPI:

5 países con mejor score medio (Figura 05.2):

País	Score
Findania	81.59
Suecia	81.25
Dinamarca	78.45
Países Bajos	73.99
Irlanda	70.03

5 países con peor score medio (Figura 05.3):

País	Score
Hungría	49.47
Eslovaquia	48.73
Rumanía	36.94
Croacia	36.74
Bulgaria	31.74

Observamos que las notas medias por país son en general altas, con solo tres países por debajo de 40 sobre 100. Si bajamos al nivel región observamos notas más bajas:

10 regiones con mejor score (Figura 05.4):

País	Region	SPI	GDP
SE	Stockholm	84.63	56.62
FI	Helsinki-Uusimaa	83.90	47.32
FI	Pohjois- ja Itä-Suomi	83.57	29.52
SE	Östra Mellansverige	83.36	33.81
DK	Hovedstaden	82.28	73.26
SE	Övre Norrland	81.79	44.01
FI	Åland	81.42	36.85
SE	Mellersta Norrland	81.07	36.05
FI	Länsi-Suomi	81.07	30.86
DK	Midtjylland	81.05	42.84

10 regiones con peor score (Figura 05.5):

País	Region	SPI	GDP
FR	Guyane	34.95	6.35
HR	Grad Zagreb	34.60	20.93
BG	Severoiztochen	33.37	1.61
RO	Nord-Est	33.21	0.54
BG	Yuzhen tsentralen	32.67	0.72
RO	Sud-Vest Oltenia	31.98	2.50
BG	Severen tsentralen	31.95	0.36
RO	Sud-Muntenia	30.50	2.50
BG	Severozapaden	28.84	0.00
BG	Yugoiztochen	24.72	3.04

VI.2. Comparación SPI y PIB per Cápita

Al comparar el SPI con el PIB per cápita por región (GDP), observamos que el nivel económico de las regiones con mayor SPI no es necesariamente el más alto, presentándose una gran variedad de combinaciones. Sin embargo, en las regiones con un SPI más bajo, el PIB per cápita tiende a ser, en general, muy reducido.

En las **Figuras 05.7 y 05.8** se muestran las 10 mejores y 10 peores regiones en cuanto a SPI y dimensiones (Necesidades Básicas, Fundamentos de Bienestar y Oportunidades), destacando que no todas las regiones con buen SPI son las mejores en las tres dimensiones. Hay regiones que obtienen las calificaciones más altas en Necesidades Básicas, pero no aparecen en el top 10 del SPI (Creta). Sin embargo, en el caso de las regiones con un SPI bajo, el patrón es más evidente: si la región está a la cola en alguna de las tres dimensiones, probablemente estará a la cola del SPI. Ver en las **Figura 05.9 y Figura 05.10** el mapa de todas las regiones según SPI, dimensiones y PIB per cápita.

VI.3. Clasificación por SPI y PIB per cápita:

Clasificamos las regiones en función de sus puntuaciones alcanzadas para los dos índices que miden el desarrollo social.

Ya hemos observado en fases anteriores que existe una correlación positiva entre ambos indicadores, sin embargo, sus distribuciones son muy distintas (**Figura 05.11 y Figura 05.1**): La distribución del SPI es más simétrica y concentrada en torno a valores medios, indicando que la mayoría de las regiones están en niveles moderados, mientras que la distribución del PIB per cápita es más sesgada hacia la derecha, con una alta variabilidad y una concentración en la parte baja, lo que refleja una brecha significativa entre las regiones más ricas y más pobres.

En la **Figura 05.14** visualizamos la estructura jerárquica de agrupamiento de los datos en función de cuatro métodos de enlace. Estos métodos afectan la manera en que los clústeres o grupos se combinan en cada paso del proceso jerárquico. De un primer vistazo podemos deducir que el método “Ward”, método que minimiza la varianza dentro de los clústeres, y el método “Complete”, basado en la “distancia” máxima entre cualquier par de datos entre clústeres, ofrecen una separación más clara. Dependiendo de la “altura” en la que cortemos el dendograma estaríamos hablando de entre 4, 5 o 6 grupos diferenciados. Para evaluar qué número de grupos es el adecuado utilizamos el método Ward y extraemos dos métricas de evaluación:

- **Silueta:** El coeficiente de silueta mide cómo de bien cada punto está agrupado con respecto a otros puntos en su mismo clúster en comparación con puntos en clústeres diferentes. Cada punto recibe un valor de silueta entre -1 y +1: un valor cercano a +1 indica que el punto está bien agrupado, mientras que un valor cercano a -1 sugiere que el punto podría estar mal agrupado o en el clúster equivocado.
- **El índice de Calinski-Harabasz:** también conocido como el índice de varianza entre-clústeres, mide la compactibilidad de los clústeres y la separación entre ellos. Se basa en comparar la varianza dentro de los clústeres con la varianza entre clústeres. Un valor alto del índice indica que los clústeres están bien separados entre sí y que los puntos dentro de cada clúster son muy similares entre sí, sugiriendo una buena estructura de agrupamiento. En cambio, un valor bajo sugiere que los clústeres pueden estar solapados o que los puntos dentro de los clústeres no son muy homogéneos.

Según estas métricas, la segmentación en 4 o 5 grupos parecen las mejores opciones:

Nº de grupos	Silueta	Calinski-Harabasz
3	0.41	257.15
4	0.44	307.35
5	0.40	397.35
6	0.38	306.93

VI.4. K-Means Clustering:

Utilizamos el método K-means para clasificar los datos en 4 y 5 grupos, evaluando la opción más adecuada a través del método del “codo” y el índice de Davies-Bouldin. El método del codo nos ayuda a visualizar el punto en el que la reducción en la variabilidad interna de los clústeres comienza a estabilizarse, confirmando que 4 clústeres es la opción más viable (**Figura 05.15**). El índice de Davies-Bouldin nos proporciona una medida de la calidad del agrupamiento, evaluando la separación y cohesión de los clústeres (**Figura 05.16**). Los resultados indican que 4 grupos es la opción óptima, aunque explorar 5 grupos puede ser útil para capturar matices adicionales.

Clasificación de 4 grupos (Figura 05.17 y Figura 05.18):

1. Regiones con PIB muy bajo y SPI entre mínimo y medio.
2. Regiones PIB bajo y SPI medio-alto.
3. Regiones con PIB medio y SPI medio-alto.
4. Regiones con PIB alto y SPI medio-alto.

El último grupo, con PIB más alto, es muy pequeño, formado por las regiones que alargan la cola de la distribución del PIB muy a la derecha (**Figura 05.11**)

Grupo	GPD	SPI
1	00.00 - 20.93	24.72 - 58.23
2	07.07 - 31.93	46.79 - 71.32
3	24.40 - 56.62	62.48 - 84.63
4	61.54 - 100.00	59.02 - 82.28

Clasificación de 5 grupos (Figura 05.19 y Figura 05.20):

La clasificación en cinco grupos añade una partición a las regiones más desfavorecidas, diferenciando entre aquellas con PIB muy bajo y SPI bajos, frente a regiones con PIB bajo, pero SPI medio o alto.

En esta clasificación se observa algo de solapamiento y las métricas no son tan buenas como en la clasificación a cuatro, sin embargo, las diferencias son pequeñas y el hecho de añadir un grupo más específico en las notas más bajas de ambos índices hace la clasificación a cinco mucho más interesante.

Grupo	GPD	SPI
Vulnerable	00.00 - 20.93	24.72 - 45.68
Emerging	02.95 - 24.33	45.28 - 66.24
Stable	13.24 - 32.02	54.66 - 72.83
Prosperous	28.18 - 56.62	62.48 - 84.63
Affluent	61.54 - 100.00	59.02 - 82.28

En la **Figura 05.21** del Anexo observamos la distribución de los cinco grupos definidos en este punto, a través del mapa de regiones de la UE. El grupo de regiones mas vulnerable (Vulnerable) se concentra en las regiones de

Croacia, Bulgaria y Rumania, además de Sicilia, Lituania y la Guyana Francesa (Ver listado completo en **Figura 05.23**). Mientras que las regiones más prosperas (Prosperous y Affluent) se localizan en la península escandinava, Dinamarca, Irlanda y un “cinturón” de regiones centroeuropeas en el entorno de los Países Bajos, Bélgica, Luxemburgo, Alemania y Austria. El resto de la UE se divide en regiones con un desarrollo social y económico consolidado (Stable) y regiones en crecimiento con un SPI medio, pero bajo PIB (Emerging), estos dos últimos grupos se distribuyen en el clásico esquema de regiones “Centrales” (Francia, Alemania, Austria y regiones más industrializadas de España e Italia) y regiones de “Periferia” (Sur de España, Portugal, sur de Italia, Grecia, Chipre, Europa del Este). Comprobamos en la **Figura 05.22** que los grupos “Stable”, “Emerging” y “Prosperous” concentran la mayoría de las regiones.

VI.5. Clasificación por dimensiones:

Una forma interesante de clasificar las regiones sería a través de las tres dimensiones: Necesidades Básicas, Fundamentos de Bienestar Social y Oportunidades, sin embargo, la variabilidad en las calificaciones entre estas dimensiones da como resultado clasificaciones de hasta seis grupos diferentes, con cierto solapamiento que dificulta la interpretabilidad. En las **Figuras 05.28 y 05.29** podemos diferenciar con facilidad los grupos con altos y bajos valores en las tres dimensiones.

Líneas de investigación futura

Este trabajo sobre el Índice de Progreso Social Europeo (SPI-EU) 2024 proporciona valiosos conocimientos sobre el desarrollo social regional en la Unión Europea, pero también destaca áreas donde futuras investigaciones podrían mejorar la comprensión y aplicabilidad del índice. A continuación, se proponen algunas líneas de investigación futura: la incorporación de nuevos indicadores de carácter social, como la salud mental, podría ofrecer una visión más completa; la ponderación de indicadores permitiría un análisis más detallado; el análisis y seguimiento del índice a lo largo del tiempo podría revelar patrones en el progreso social; el estudio de los efectos de la migración interna y externa sobre el progreso social en distintas regiones proporcionaría más información; realizar comparaciones entre las regiones de la UE y países no pertenecientes a la UE que utilicen marcos similares de progreso social enriquecería el análisis; La aplicación de modelos de aprendizaje automático avanzados para la imputación de datos y clasificación mejoraría la precisión del SPI, haciéndolo una herramienta más robusta. Finalmente, desarrollar modelos predictivos basados en el SPI-EU podría ayudar a anticipar futuros cambios en el progreso social y guiar la formulación de políticas proactivas

Estas líneas de investigación no solo perfeccionarían el SPI, sino que también proporcionarían una comprensión más profunda de la dinámica del progreso social en Europa, contribuyendo a la formulación de políticas más específicas y eficaces.

Conclusiones

Los indicadores compuestos son una herramienta poderosa capaz de sintetizar información multivariante en un solo dato fácil de interpretar. Sin embargo, durante la construcción de estos índices, se toman decisiones subjetivas que pueden afectar el resultado final. En la creación de un índice de progreso social, como el que se aborda en este trabajo, el estadístico intenta tomar las decisiones más objetivas posibles, pero no siempre dispone de toda la información necesaria. Por lo tanto, es recomendable considerar las sugerencias, revisiones y análisis de sociólogos y otros expertos en ciencias sociales durante el proceso de elaboración del índice.

En este trabajo se ha desarrollado un Índice de Progreso Social Europeo (SPI-EU) 2024 para las regiones y países de la Unión Europea, proporcionando una visión detallada del desarrollo social en la región. Se ha comenzado con 45 indicadores sociales, que se han agrupado en 12 componentes y estos, a su vez, en 3 dimensiones, para finalmente calcular un único indicador por región. Se han utilizado métodos de normalización y agregación generalizada, siguiendo la metodología aplicada por la Comisión Europea en su edición de 2020. No se han aplicado ponderaciones. Se ha verificado la robustez del índice en cada fase de su construcción frente a diferentes indicadores. Además, se realizó un análisis de la relación entre el SPI-EU y un indicador económico tradicional, como el PIB per cápita, incluyendo la clasificación de las regiones en función de su desempeño.

El índice revela importantes disparidades entre las distintas zonas de la UE, destacando que las áreas más avanzadas, como los países escandinavos, presentan altos niveles de progreso social, mientras que las regiones de Europa del Este y los Balcanes muestran los índices más bajos. Aunque existe una correlación positiva entre el PIB per cápita y el SPI, el estudio demuestra que el desarrollo económico no siempre se traduce directamente en bienestar social, subrayando la importancia de considerar indicadores no económicos, como la salud, la educación y la inclusión. Las políticas de desarrollo deben adaptarse a las necesidades locales, priorizando no solo el crecimiento económico, sino también un progreso social integral que atienda tanto las necesidades básicas como la creación de oportunidades equitativas para todas las regiones.

Cuadro de Mando y Visualizaciones

Para facilitar la interpretación de los resultados del SPI, se ha elaborado un cuadro de mando con diversas visualizaciones y KPIs. Este dashboard, desarrollado en Tableau, proporciona una vista integral de los datos y permite un análisis interactivo del índice. Tanto el paquete de trabajo en Tableau como la base de datos en formato Excel se incluyen como anexos a este informe.

Referencias

Annoni, P & Bolsi, P. (2020), *The regional dimension of social progress in Europe: Presenting the new EU Social Progress Index*. European Commission Directorate-General for Regional and Urban policy. https://ec.europa.eu/regional_policy/sources/work/202006_spi_en.pdf

OECD/European Union/EC-JRC (2008), *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264043466-en>.

Nardo M, Saisana M, Saltelli A, Tarantola S. (2005), *Tools for Composite Indicators Building*. European Commission. JRC Publications Repository. <https://publications.jrc.ec.europa.eu/repository/handle/JRC31473>

“European Social Progress Index 2024”. https://ec.europa.eu/regional_policy/assets/social-progress/index.html#