Tecnologico de Monterrey,
Campus Guadalajara

PCA, K-Means y Regresión Lineal

Análisis Multifacético de Datos Socioeconómicos Mundiales

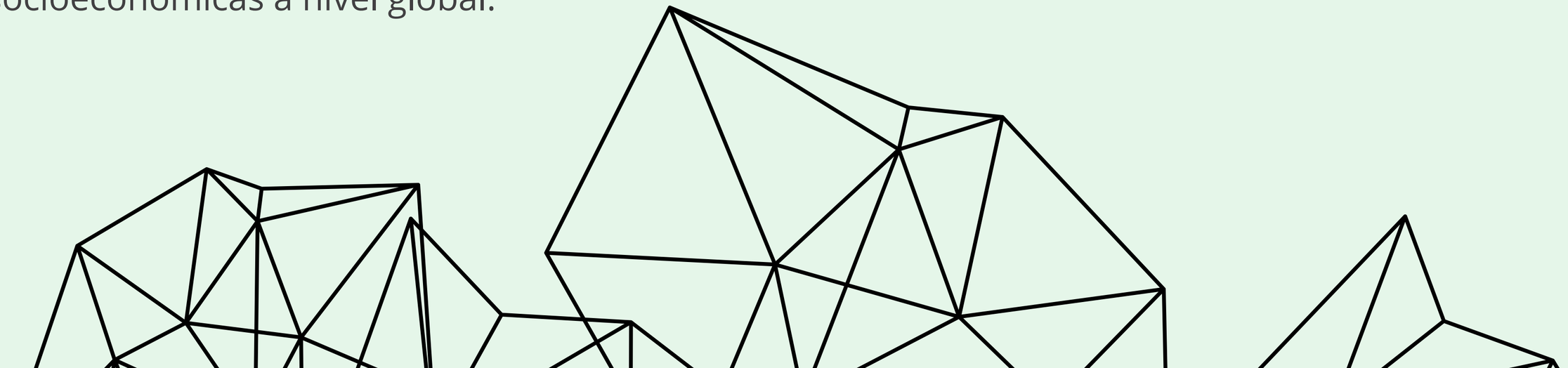
Presentada por
Claudio J. Gonzalez-Arriaga
Gabriel E. Melendez-Zavala

Ingenieria e Ciencia Datos y Matematicas

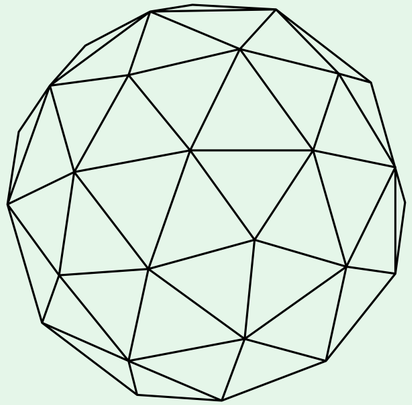
Noviembre 2023

Abstract

Este trabajo aborda la exploración profunda de datos socioeconómicos de países a nivel mundial mediante técnicas avanzadas de análisis matemático. Se empleó el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de la información, seguido de un análisis detallado de las correlaciones entre los componentes principales. Posteriormente, se imple-mentó el algoritmo deagrupamiento K-Means para clasificar los países en clusters significativos. Finalmente, se llevó a cabo una regresión lineal para prever el comportamiento de las exportaciones basándose en los componentes principales obtenidos. Este enfoque integral permite una comprensión profunda de las interrelaciones entre diversas variables socioeconómicas a nivel global.

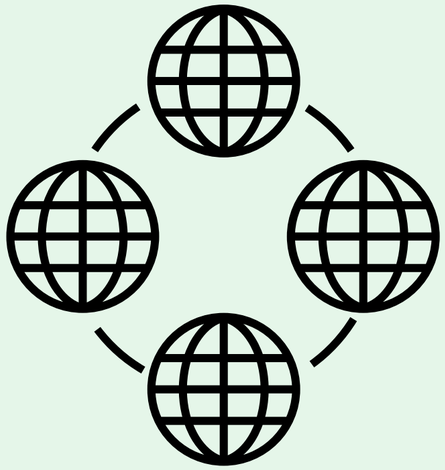


El Problema



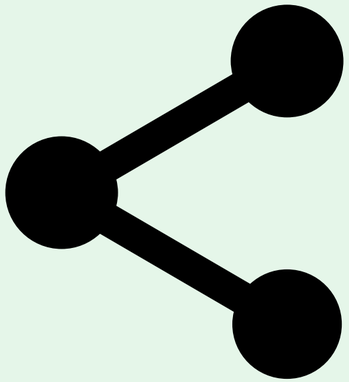
Problema A

Con la creciente disponibilidad de datos, la capacidad para analizar y comprender la compleja red de factores que influyen en el desarrollo de los países se ha vuelto una prioridad



Problema B

En un mundo donde la inter-conexión entre naciones es cada vez más intrincada, la comprensión de las dinámicas socioeconómicas globales se vuelve esencial para abordar desafíos y oportunidades emergentes.

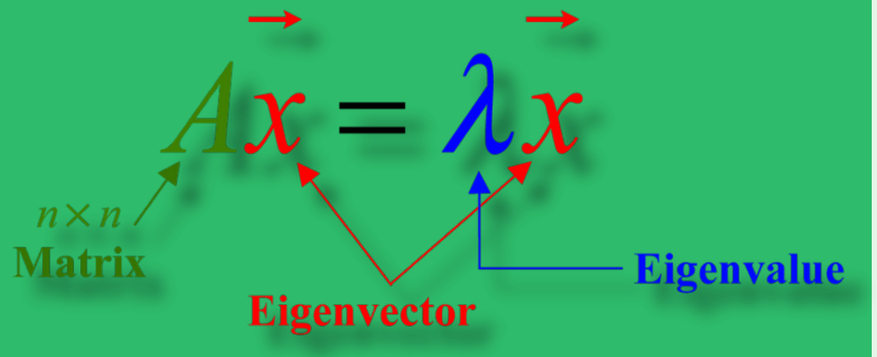


Problema C

Encontrar relaciones entre indicadores principales para comprender las dinamicas entre ellos

Marco teórico

$$\begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nm}^2 \end{pmatrix}$$



Matriz de Covarianza: Describe relaciones entre variables.

Análisis de Componentes Principales (PCA): Reduce la dimensionalidad manteniendo la variabilidad.

Estandarización de Datos: Normaliza la escala de las variables.

Eigenvectores y Eigenvalores: Define direcciones y cantidad de variabilidad.

Clasificación de Países: Identifica patrones y grupos comunes.

Regresión Lineal: Modela comportamientos basados en patrones identificados.

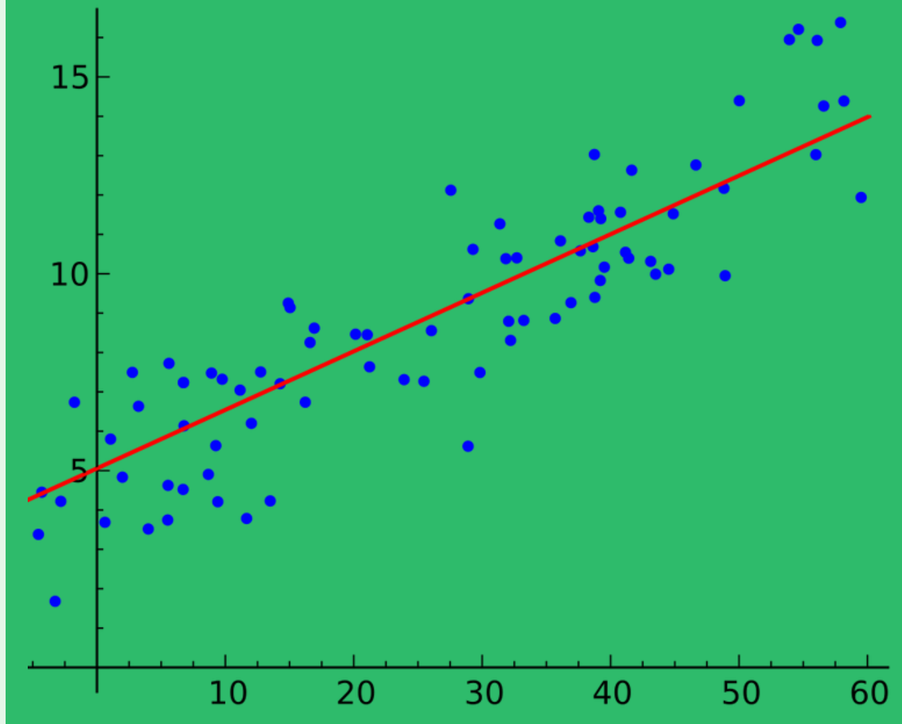
Marco teórico

Clasificación de Países (Clusters): Identifica grupos de comportamientos similares.

Regresión Lineal Específica: Modela cómo una variable se relaciona con otras.

Conclusiones

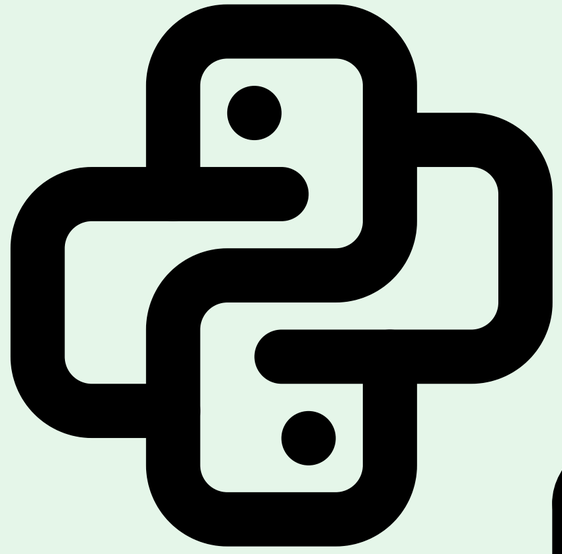
El análisis multivariable, incluyendo PCA y regresión lineal, ha permitido entender patrones y comportamientos nacionales. Estas técnicas ofrecen insights valiosos para la toma de decisiones.

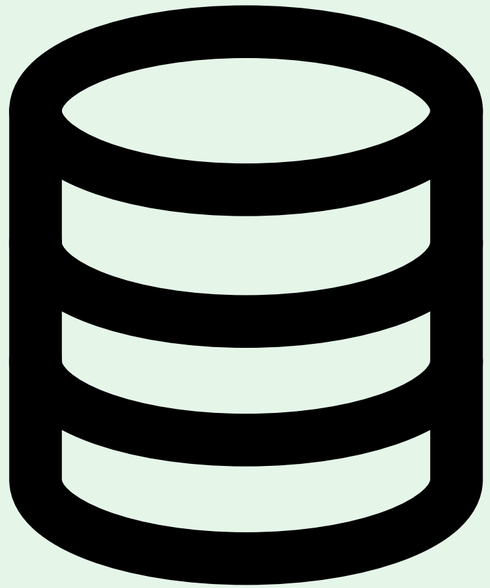


METODOLOGIA

Metodología de Análisis Socioeconómico Global

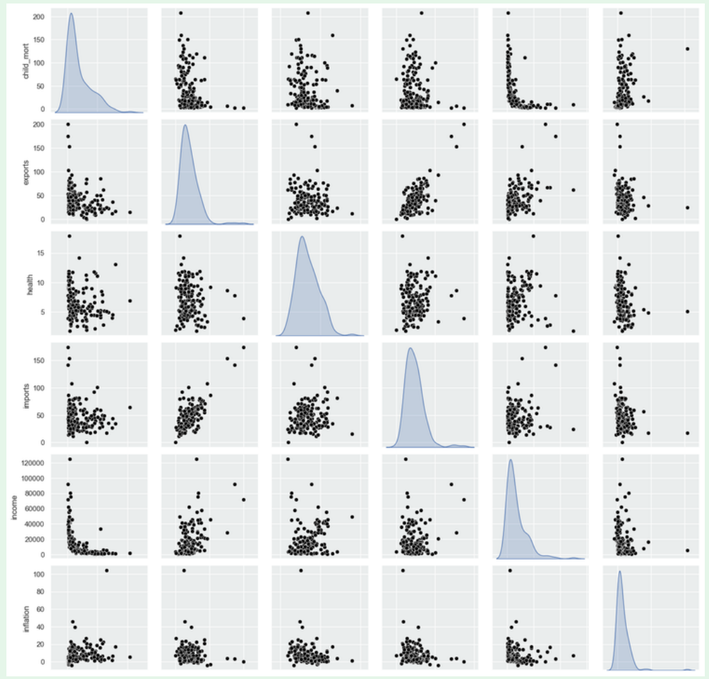
- Descripción de la base de datos socioeconómicos con más de **300,000** observaciones.
- Proceso de limpieza de datos: **eliminación de nulos** y selección de un **año específico** (2021).
- Uso de **Python**, **Jupyter Notebooks** y diversas bibliotecas para análisis de datos.

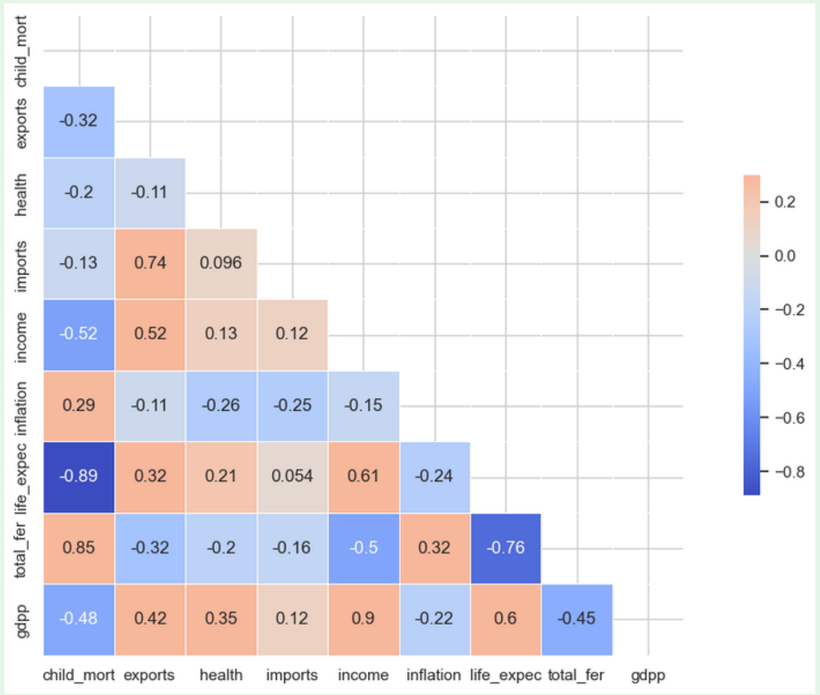




Análisis Inicial y Exploración

- Obtención de **estadísticas descriptivas** de las variables numéricas.
- **Visualización** de relaciones entre variables socioeconómicas a través de gráficos.
- **Matriz de covarianza** y su importancia en la comprensión de la relación entre variables.

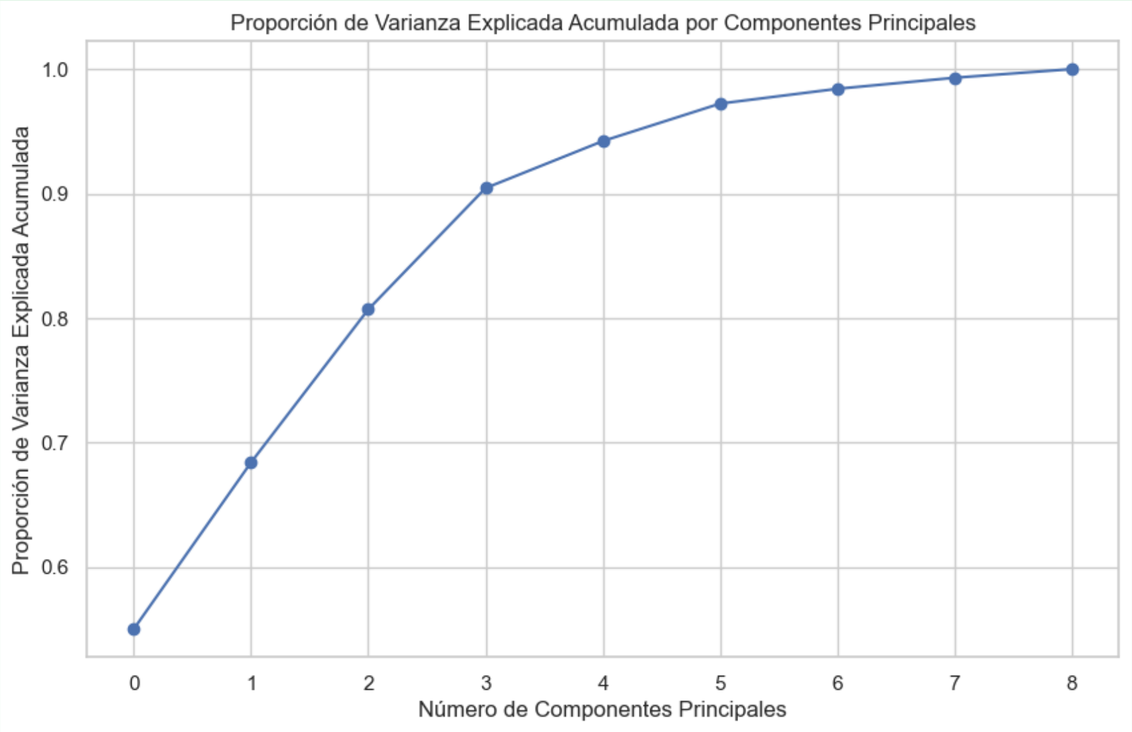




Análisis de Componentes Principales

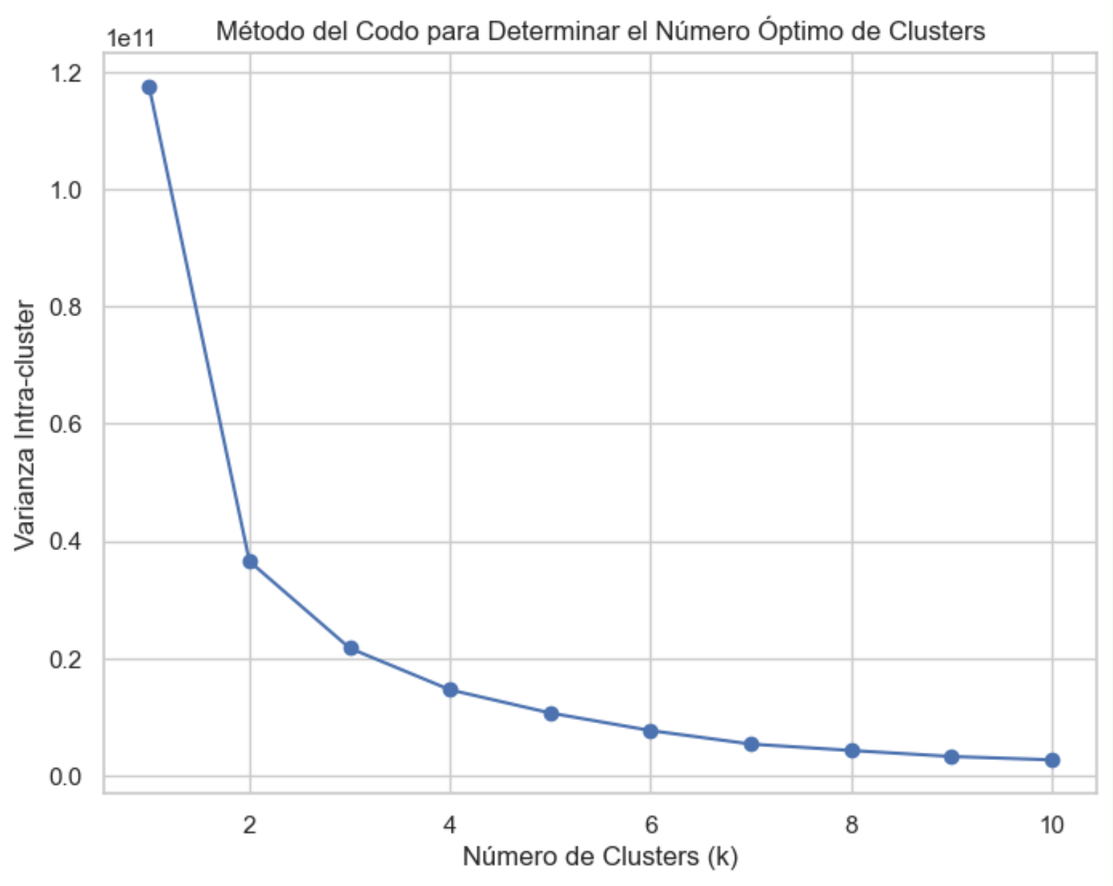
$$x_k^s = \frac{x - \mu_x}{\sigma_x}$$

- Importancia del **PCA** en la reducción de dimensionalidad y la identificación de patrones.
- Proceso de **estandarización** de datos para el PCA.
- Cálculo de **eigenvectores** y **eigenvalores**, exploración de la varianza acumulativa y selección de componentes.



$$|A - \lambda I| = \lambda^2 - \text{traza} \cdot \lambda + \text{determinante} = 0$$

Clasificación con Componentes Principales

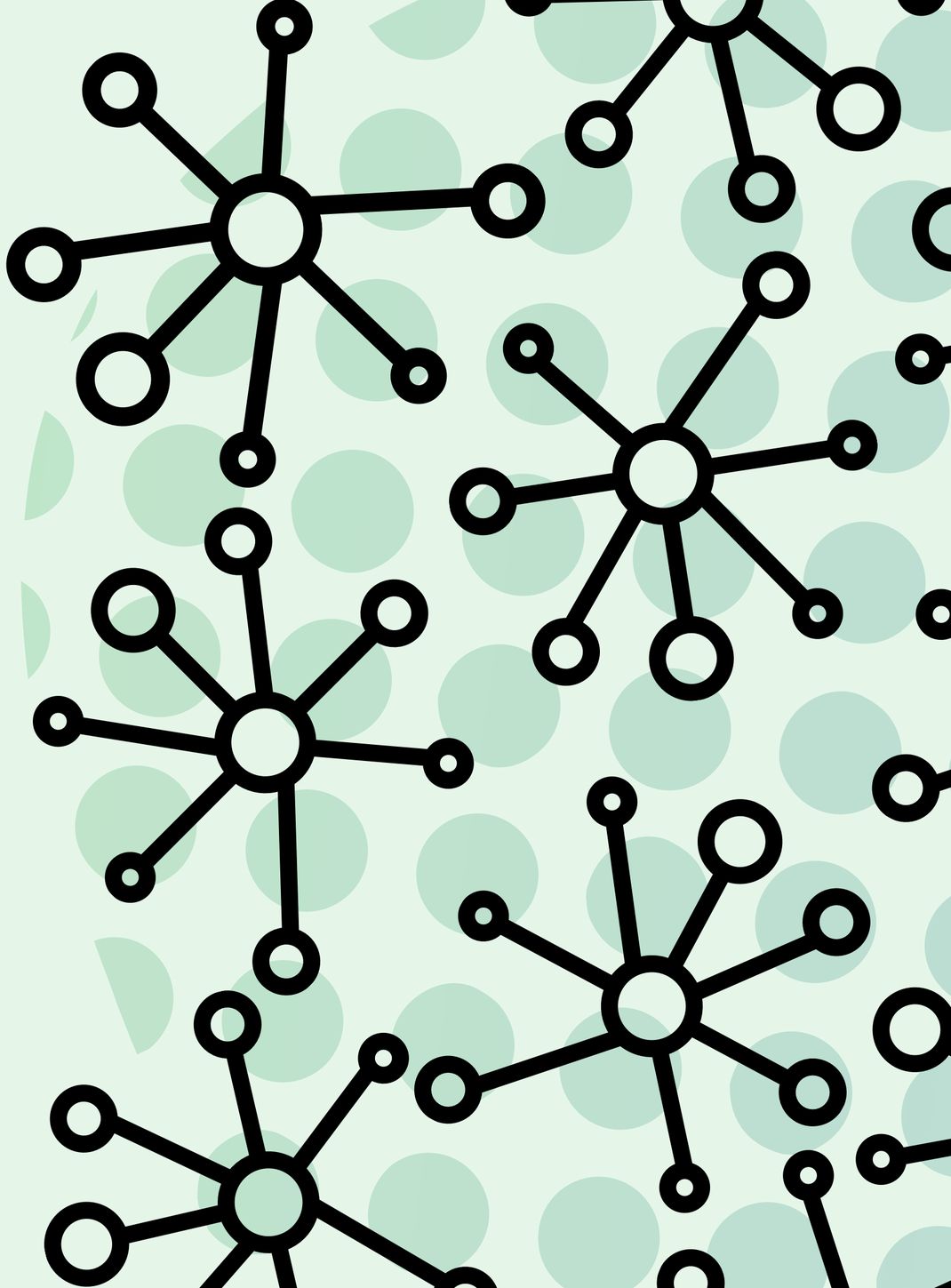


- Uso de **K-means** para agrupar países con similitudes.
- Determinación del número óptimo de clústeres mediante el **método del codo**.
- Implementación del algoritmo K-means para **clasificación** de países.

ÁNALISIS DE RESULTADOS

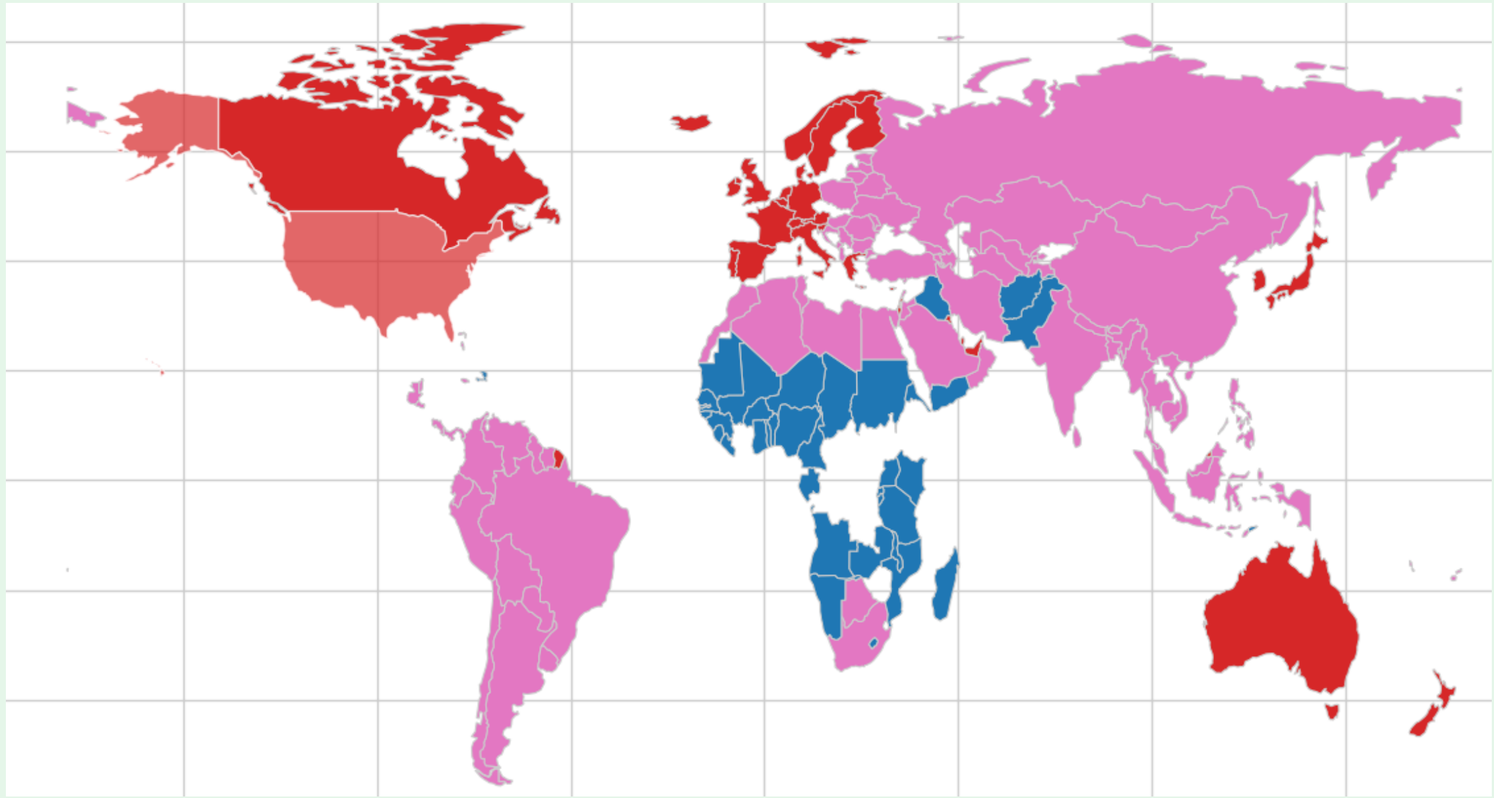
Clasificación de Clusters

- Clasificación de países en **cuatro categorías** distintas mediante k-means.
- **Visualización** de la agrupación utilizando colores diferentes en un mapa mundial.
- Aplicación de clasificación antes y después de PCA para **comparar resultados**.
- Algoritmo **k-means** agrupa países con similitudes entre sí en 4 clusters.
- Coloreo de países según su grupo, reflejando **similitudes generales** en variables socioeconómicas.



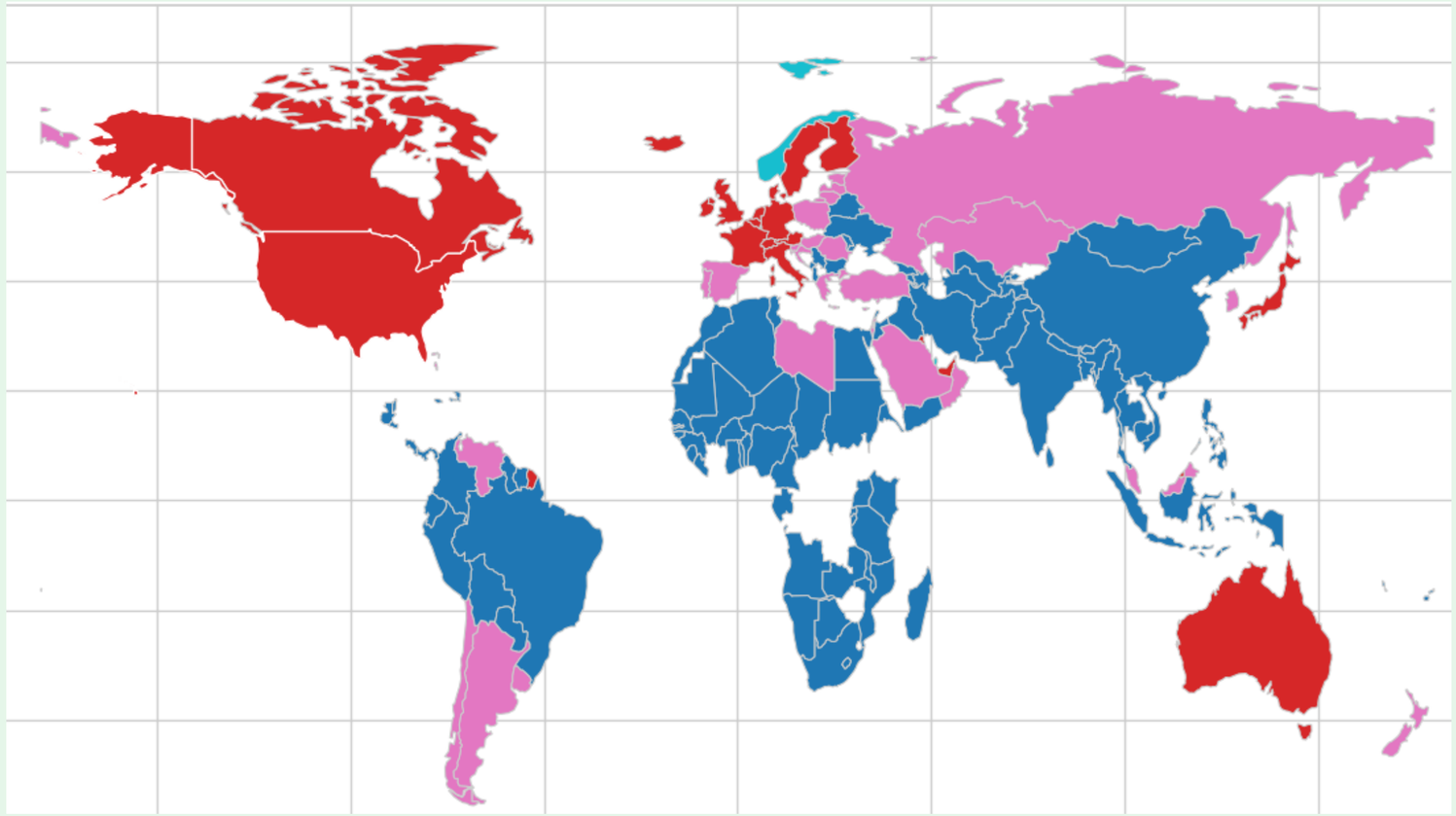
Pre PCA

Se observo que antes de aplicar PCA los países se distribuyen en **tres clusters** distintos, no en cuatro. Es evidente que la mayoría de los **países pertenecientes al bloque occidental**, denominados "Países de primer mundo", se encuentran concentrados en el **cluster rojo**.



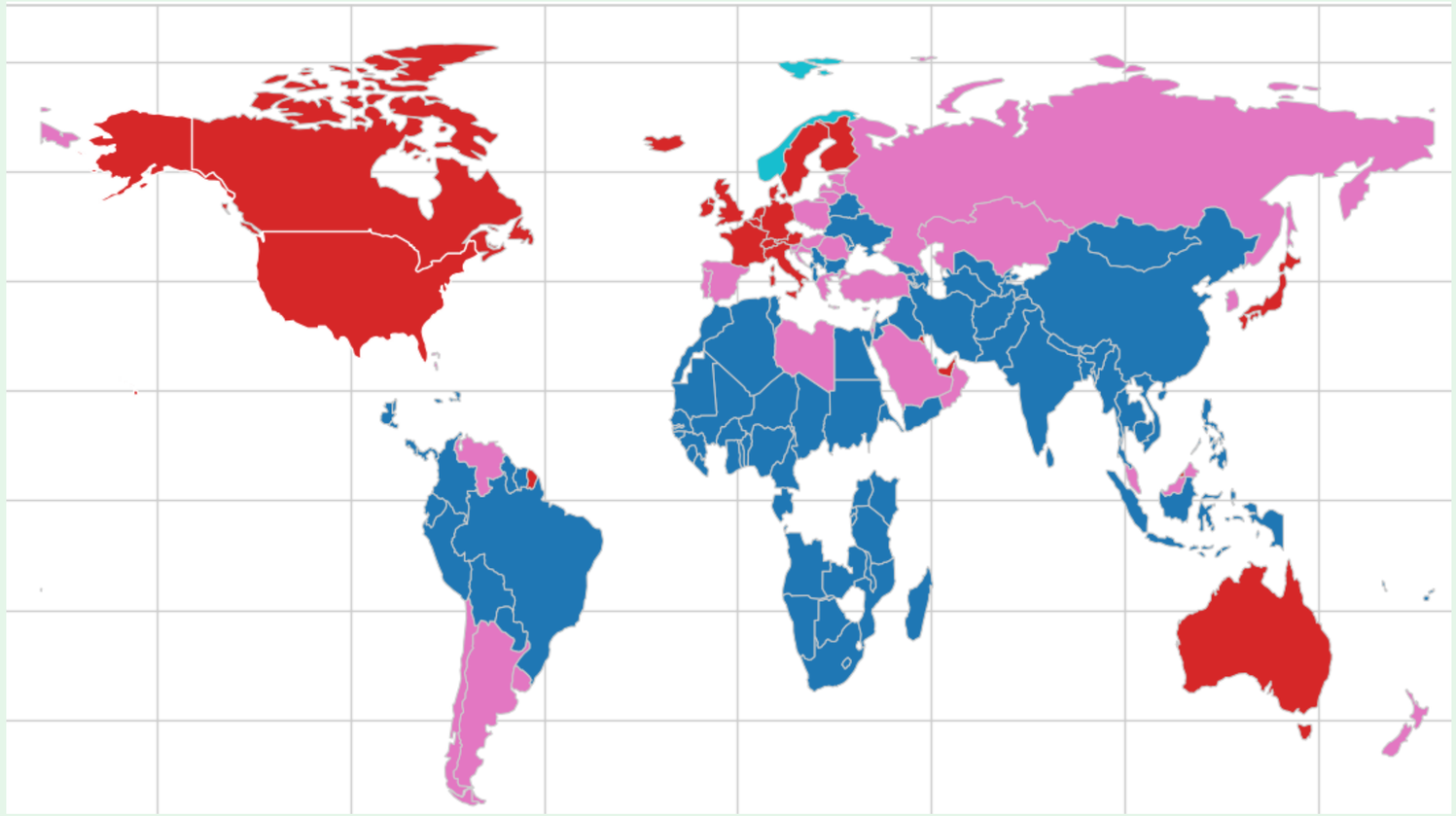
Además, la mayoría de los países que pertenecieron al **antiguo bloque comunista** están categorizados en el **cluster rosa**. Por otro lado, aquellos países catalogados como **menos desarrollados** están agrupados en el **cluster azul**.

Post PCA



El **cluster rojo**, que anteriormente agrupaba mayormente a los países del bloque occidental, muestra una **subdivisión**, incluyendo a España, Portugal y Grecia en un cluster separado. Estos países, han enfrentado **recientes desafíos económicos** y migratorios que podrían haber impactado su bienestar general a pesar de ser parte del bloque occidental.

Una vez aplicado el PCA, se observan ahora **cuatro clusters distintos**. El **cluster azul celeste** destaca por albergar a los **países más desarrollados** y con una elevada calidad de vida, entre los cuales se encuentran Noruega, Qatar y Luxemburgo.



Post PCA

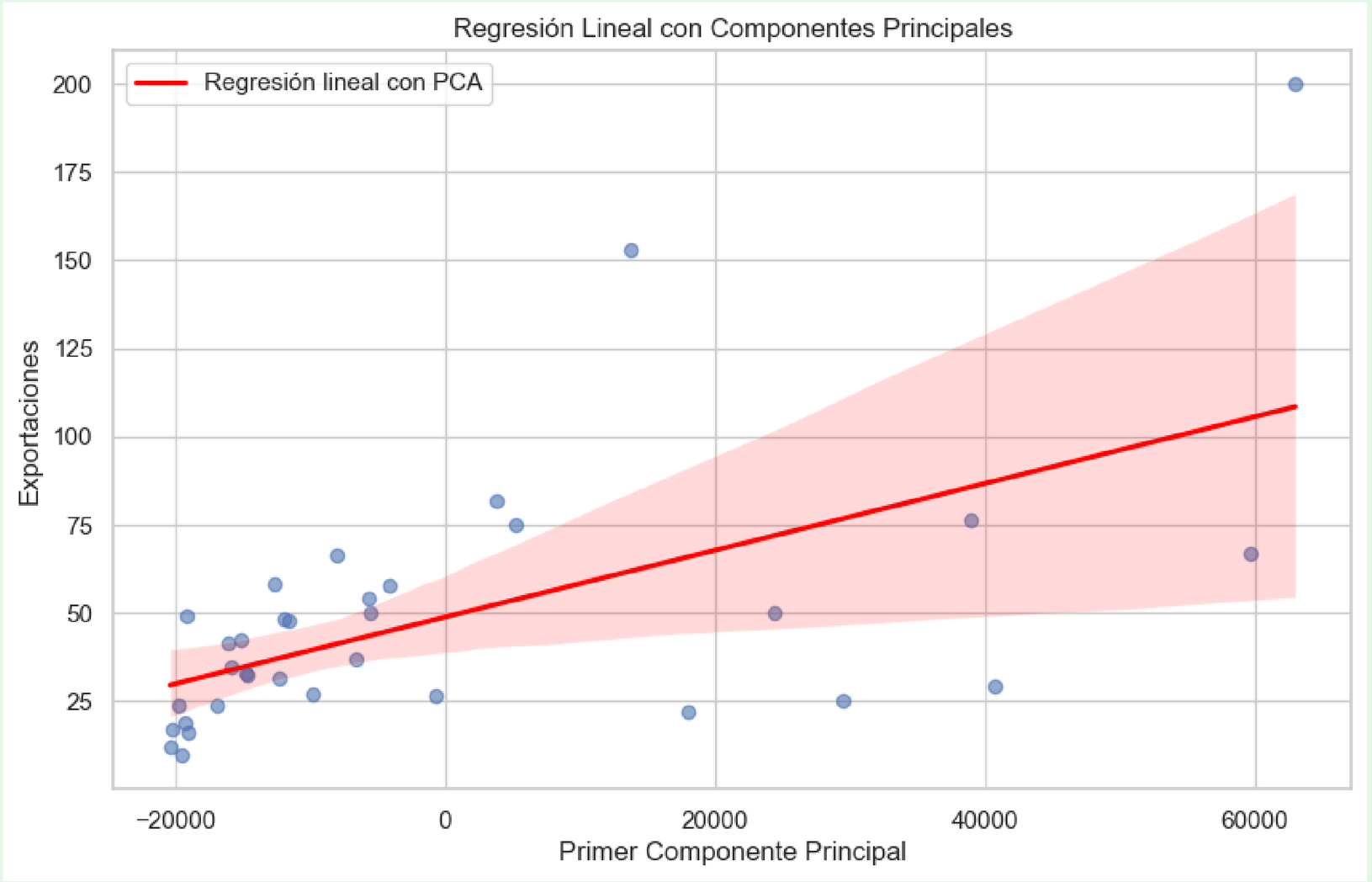
Posteriormente, encontramos a los **países en vías de desarrollo** agrupados en el **cluster rosa**, junto con naciones de Oriente Medio y Europa Oriental que presentan cierto nivel de desarrollo y bienestar.

Finalmente, en el **cluster azul** se ubican los **países menos desarrollados**, afectados por la violencia, inseguridad y diversos problemas económicos como la inflación y la desigualdad.

ÁNALISIS DE RESULTADOS

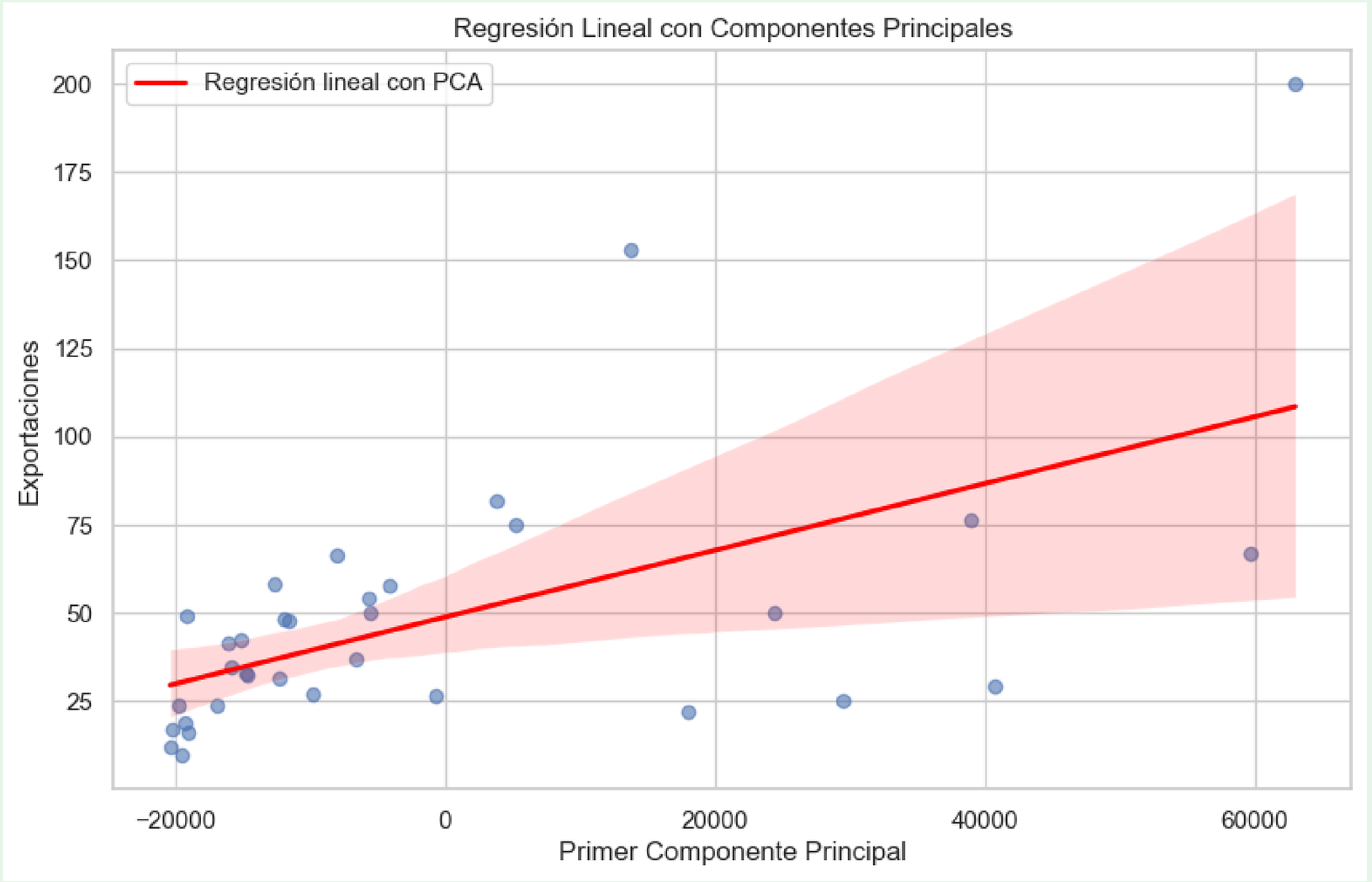
Regresion Linear

- Elección de regresión lineal basada en la **interpretación de coeficientes** y su **capacidad explicativa** en las exportaciones.
- La regresión lineal no solo predice el comportamiento futuro, sino que revela la **relación cuantitativa** entre componentes principales y variables socioeconómicas.
- Resultados obtenidos ofrecen una visión significativa de la relación entre variables socioeconómicas y exportaciones.
- **Coeficientes resultantes** son fundamentales para un análisis detallado de estas relaciones.



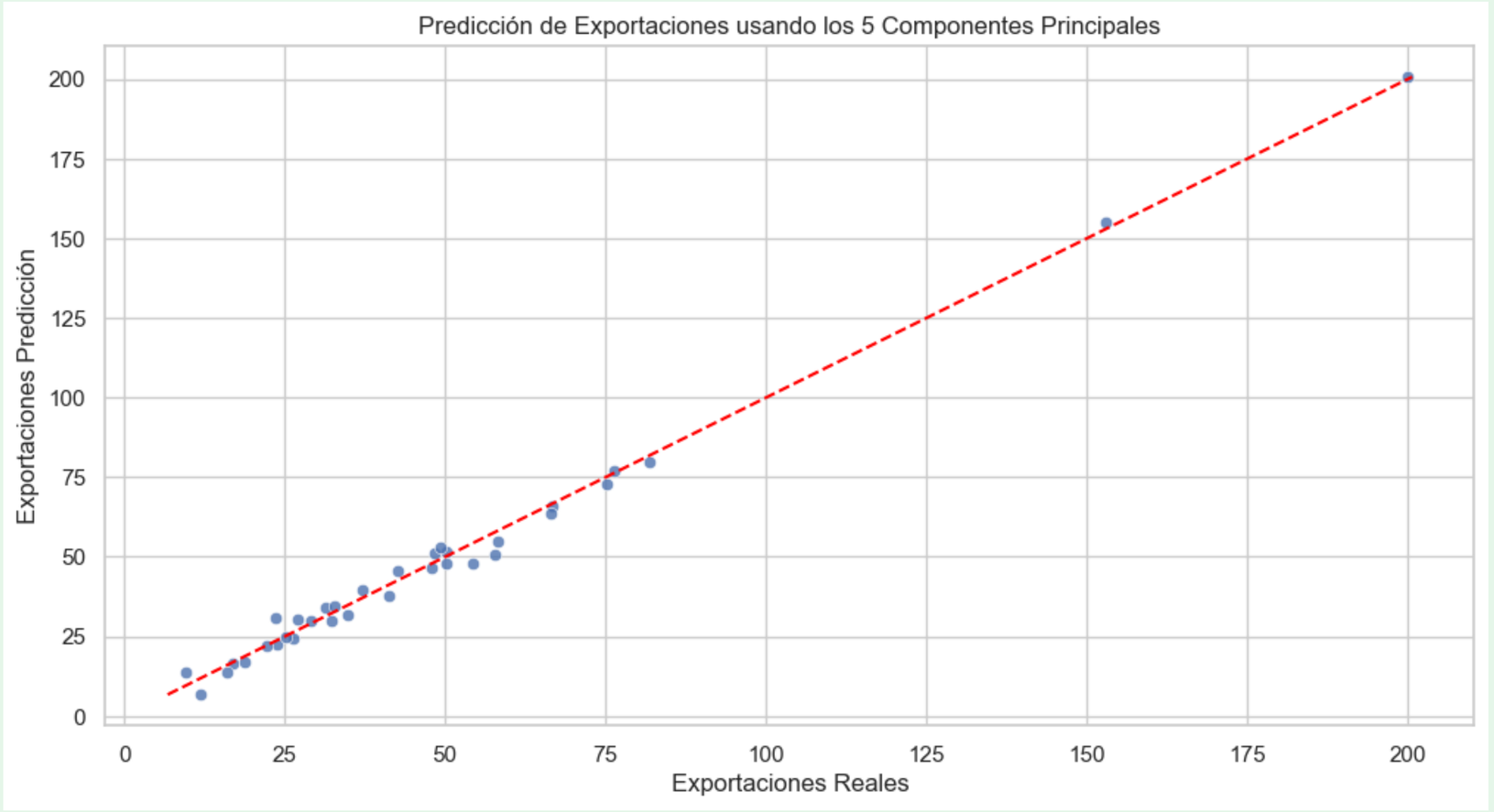
Coeficientes	
Coeficiente 1	$5{,}1244 \times 10^{-4}$
Coeficiente 2	$9{,}2685 \times 10^{-4}$
Coeficiente 3	−0,2342
Coeficiente 4	0,6563
Coeficiente 5	0,5611
Intercepto	41,0305
MSE con PCA	9,5403

Los coeficientes indican la influencia de cada componente principal en las exportaciones. Es notable que el cuarto componente principal tiene un coeficiente positivo significativo (O.656), sugiriendo que un aumento en este componente está fuertemente asociado con un incremento en las exportaciones.



Coeficientes	
Coeficiente 1	$5{,}1244 \times 10^{-4}$
Coeficiente 2	$9{,}2685 \times 10^{-4}$
Coeficiente 3	−0,2342
Coeficiente 4	0,6563
Coeficiente 5	0,5611
Intercepto	41,0305
MSE con PCA	9,5403

El intercepto, con un valor de 41.03, representa el nivel esperado de exportaciones cuando todas las variables principales son cero. Esto puede interpretarse como la base de exportaciones sin la influencia de las componentes principales.



Coeficientes	
Coeficiente 1	$5{,}1244 \times 10^{-4}$
Coeficiente 2	$9{,}2685 \times 10^{-4}$
Coeficiente 3	−0,2342
Coeficiente 4	0,6563
Coeficiente 5	0,5611
Intercepto	41,0305
MSE con PCA	9,5403

La evaluación del modelo reveló que la inclusión de cinco componentes principales mejoró significativamente la capacidad predictiva. Esto sugiere que la información capturada de estos componentes adicionales es crucial para explicar la variabilidad en las exportaciones.

Conclusion

PCA
Permitio reduccion de dimensiones

K-Means
Determino clusters para clasificación de paises

Regresión lineal
Predicción y comprension de C.P

Interrelaciones variables
Conexiones complejas entre variables

Patrones datos socioeconómicos
Detección de tendencias

Referencias

01 W. H. Greub, Linear algebra, vol. 23. Springer Science & Business Media, 2012.

02 E. Bisong, Matplotlib and Seaborn, pp. 151–165. Berkeley, CA: Apress, 2019.

03 J. H. Wilkinson, F. L. Bauer, and C. Reinsch, Linear algebra, vol. 2. Springer, 2013.

04 K. Jolly, Machine learning with scikit-learn quick start guide: classification, regression, and clustering techniques in Python. Packt Publishing Ltd, 2018.

05 Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," Nature Reviews Methods Primers, vol. 2, no. 1,

06 J. L. Devore, Probability and Statistics for Engineering and the Sciences.

07 K. R. Žalik, "An efficient k-means clustering algorithm," Pattern Recognition Letters, vol. 29, no. 9, pp. 1385–1391, 2008.

08 E. Kreyszig et al., "Matemáticas avanzadas para ingeniería," 2001.

Gracias