Análisis Multifacético de Datos Socioeconómicos Mundiales: PCA, K-Means y Regresión Lineal

Claudio J. Gonzalez-Arriaga[1] and Gabriel E. Melendez-Zavala[2]

[1] *Tecnologico de Monterrey, Campus Guadalajara*

Publication date: 24/11/2023

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 implementó el algoritmo de agrupamiento 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.

Keywords— Análisis de Componentes Principales (PCA), Agrupación de Datos, K-Means, Reducción de Dimensionalidad, Correlación, Patrones de Similitud, Datos Socieconómicos, Clasificación

I. Introducción

El estudio de las interrelaciones socioeconómicas a nivel mundial ha sido siempre un desafío fascinante. 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 ha mejorado significativamente.

En un mundo donde la interconexión entre naciones es cada vez más intrincada, la comprensión de las dinámicas socioeconómicas globales se vuelve esencial para abordar desafíos y oportunidades emergentes. Este estudio surge de una profunda motivación: desentrañar los misterios detrás de los datos socioeconómicos de países de todo el mundo, utilizando técnicas matemáticas avanzadas como el Análisis de Componentes Principales (PCA) y K-Means para destilar patrones significativos y clasificar naciones en grupos coherentes.

La motivación central de este trabajo surge de la necesidad de superar la abrumadora complejidad de los conjuntos de datos socioeconómicos globales. Estos conjuntos, inherentemente multidimensionales, contienen una vasta cantidad de información que abarca desde indicadores económicos hasta variables sociales. Enfrentados a tal riqueza de datos, surge la pregunta crucial: ¿cómo podemos extraer conocimiento sustantivo de esta complejidad sin perder la riqueza informativa esencial?

La respuesta a esta interrogante encuentra su base en el uso de PCA, una herramienta poderosa que nos permite reducir la dimensionalidad de los datos sin sacrificar su sustancia. La motivación subyacente es clara: al aplicar PCA, buscamos identificar los componentes fundamentales que explican la mayor variabilidad en nuestros datos. Este enfoque no solo simplifica la representación, sino que también destaca las relaciones cruciales entre variables, abriendo la puerta a un análisis más profundo y significativo.

Una vez lograda la reducción de dimensionalidad mediante PCA, nos enfrentamos a otro desafío ambicioso: clasificar países en grupos significativos. Aquí entra en juego K-Means, un algoritmo de agrupamiento que nos permite discernir patrones y similitudes en la trama socioeconómica mundial. La motivación detrás de esta clasificación es proporcionar una perspectiva clara de las dinámicas compartidas y distintivas entre países, allanando el camino para un entendimiento más profundo de la realidad global.

En este contexto, nuestra investigación no solo se enmarca en la aplicación de técnicas matemáticas avanzadas, sino que encuentra su razón de ser en la necesidad de simplificar la complejidad inherente a los datos socioeconómicos globales. Al hacerlo, aspiramos no solo a descifrar patrones y clasificaciones, sino a ofrecer una ventana a la comprensión de las interrelaciones complejas que definen la economía mundial. Este enfoque, anclado en la motivación de hacer accesible la riqueza de datos, ilustra cómo la aplicación de herramientas matemáticas avanzadas puede desbloquear insights cruciales y proporcionar una visión más nítida del panorama socioeconómico global.

II. Metodología

Para llevar a cabo este trabajo, se seleccionó una base de datos con más de 300,000 observaciones de la gran mayoría de los países del mundo en diversas variables socioeconómicas, tales como el producto interno bruto, la tasa de mortalidad o las exportaciones, por mencionar algunas. Con el fin de realizar un análisis apropiado, es necesario llevar a cabo primeramente una limpieza de la base de datos, en la cual se eliminaron datos nulos y, más importante aún, se seleccionó

un año específico (2021) sobre el cual se iban a realizar todos los análisis. Asimismo, se llevó a cabo una selección de variables finales, las cuales se convertirían en la base sobre la cual aplicaríamos nuestro análisis.
Para la ejecución de la limpieza de datos y el desarrollo integral de este trabajo, se optó por la utilización de Python como lenguaje de programación, aprovechando la versatilidad y potencialidad que ofrece en el ámbito del análisis de datos. Todo el proceso fue llevado a cabo de manera eficiente y estructurada en entornos interactivos de Jupyter Notebooks, lo que permitió una exploración iterativa y una presentación clara de los resultados obtenidos.
Durante el desarrollo, se hizo un extenso uso de diversas bibliotecas especializadas, incluyendo pandas para la manipulación y limpieza eficiente de los conjuntos de datos, numpy para operaciones matriciales fundamentales, matplotlib y seaborn para la visualización gráfica de resultados, geopandas para análisis espaciales, warnings para el manejo de advertencias, y sklearn para implementar algoritmos de análisis como PCA, K-Means y regresión lineal. La sinergia entre estas herramientas proporcionó un entorno robusto y eficaz para abordar cada etapa del análisis socioeconómico mundial.
A continuación se describirán los procesos y metodología empleada para el desarrollo del análisis, junto con diversas secciones de código relevante, para el condigo completo, consultar el repositorio de GitHub cuyo link esta la sección de recursos.

Primeramente se llevó a cabo la lectura de la base de datos, utilizando la función readcsv de la biblioteca pandas para cargar la base de datos desde un archivo CSV llamado Çountry-data.csvz almacena los datos en un DataFrame llamado country, cabe resaltar que este archivo CSV, es el resultado después de haber seleccionado manualmente el año que se iba a utilizar y las variables a trabajar, proceso que se llevó a cabo en Excel, posteriormente se seleccionaron solo las variables numéricas y se aplicó el método describe de pandas para generar estadísticas descriptivas, como la media, la desviación estándar, los cuartiles, y otros, de las columnas presentes en el DataFrame original. La descripción resultante se almacena en la variable descripcionbasedatos.

```
plt.rcParams["figure.figsize"] = (15,8)
# Carga de la base de datos
country = pd.read_csv("Country-data.csv"
numeric_columns = country.select_dtypes
(include=[np.number]).columns
data = country[numeric_columns]
# Descripcion estad stica
descripcion_base_datos=country.describe()
```

Este fragmento de código establece una base sólida para la exploración inicial de la base de datos, permitiendo un entendimiento rápido de las características estadísticas clave de las variables numéricas.

a. Matriz de covarianza

Una vez que ya tenemos acceso a toda la base datos es posible comenzar a realizar una serie de análisis o de pruebas para ver que podemos hacer con la información disponible. Primeramente se realizó una serie de gráficas con toda la información, que visualiza las relaciones entre las variables

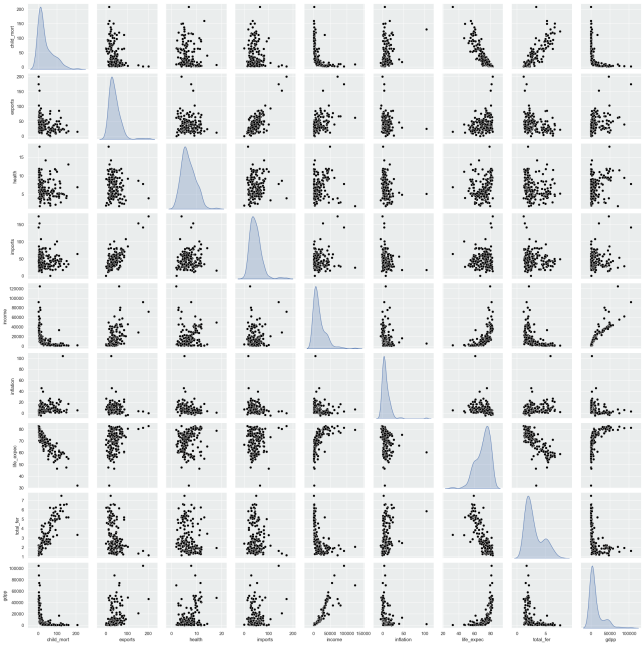

Fig. 1: Gráfico de pares que ilustra las relaciones entre diversas variables socioecnómicas

numéricas en un conjunto de datos.

A simple vista, realizar un análisis profundo a partir de la información brindada por la base de datos resulta bastante complejo. La Figura 1 evidencia la dispersión de la información entre los países, destacando la necesidad imperante de aplicar procesos analíticos para extraer conocimientos significativos. Esta visualización subraya la complejidad inherente de la estructura socioeconómica mundial y subraya la importancia de los procedimientos analíticos avanzados, como el Análisis de Componentes Principales (PCA) y K-Means, aplicados en secciones posteriores, para desentrañar patrones y relaciones subyacentes.

Un primer análisis que se puede llevar a cabo es obtener la matriz de covarianza entre las distintas variables presentes en la base de datos La covarianza es una medida utilizada al realizar PCA que describe la relación entre dos variables aleatorias. Una covarianza positiva indica que las variables tienden a aumentar o disminuir juntas mientras que una covarianza negativa señala que las variables se mueven en direcciones opuestas. Además, esto nos permite generar la matriz de covarianza entre múltiples variables aleatorias y así identificar correlaciones redundantes que puedan tener. Para dos vectores x_a y x_b la covarianza entre ellas es σ_{ab} y se calcula con la siguiente ecuación[1]:

$$\sigma_{ab} = \frac{1}{n-1}\sum_{i=1}^{n}(x_a^s - \mu_a)(x_b^s - \mu_b) \tag{1}$$

Una matriz de covarianza conteniendo los valores de covarianza siempre tendrá una forma de *n*x*n* de la siguiente manera:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

La obtención de la matriz de correlación para nuestra base de datos se realiza mediante el empleo del método corr,

disponible en la biblioteca pandas. Una vez calculada la correlación entre las variables, podemos crear y visualizar la matriz de correlación utilizando las bibliotecas Seaborn y Matplotlib en Python [2]. Estas herramientas nos permiten generar un mapa de calor que resalta la magnitud y dirección de las relaciones entre las diferentes dimensiones socioeconómicas presentes en el conjunto de datos. Este enfoque proporciona una representación gráfica eficaz de la interdependencia entre variables, facilitando la identificación de patrones y tendencias en la estructura de la información.

```
country_cor = country.drop('country',
                axis=1).corr()
mask = np.triu(np.ones_like(country_cor,
        dtype=bool))
fig, ax = plt.subplots(figsize=(15, 8))
cmap = sns.color_palette("coolwarm",
        as_cmap=True)
sns.heatmap(country_cor, mask=mask,
        cmap=cmap, vmax=.3,
        center=0,square=True,
        linewidths=.5, cbar_kws
        ={"shrink": .5}, annot=True)
plt.show()
```

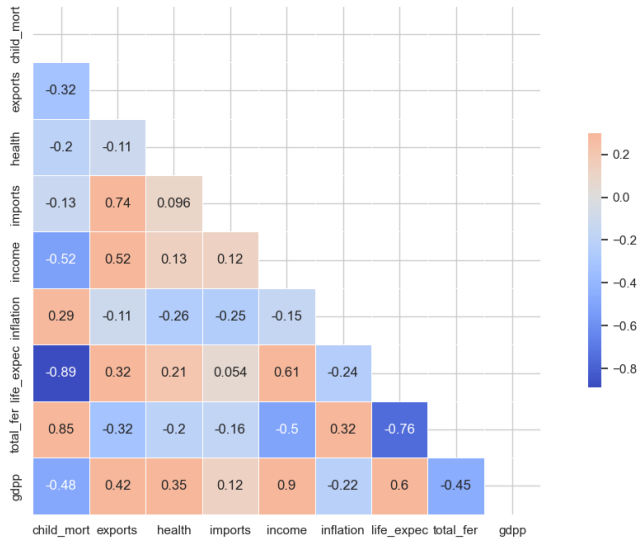

Fig. 2: Mapa de calor que ilustra la corrrelación de distintas variables

En el mapa de calor, se observa que diversas variables presentan una correlación significativa, como es el caso de ïncomeçon "gdpp", donde la correlación alcanza casi el valor de 1, indicando una relación altamente positiva. Este tipo de correlaciones cercanas a 1 sugieren una fuerte dependencia lineal entre las variables y señalan la posibilidad y recomendación de reducir la dimensionalidad, es decir, disminuir el número de variables sin perder información sustancial. En este contexto, surge la necesidad de emplear el Análisis de Componentes Principales (PCA).

b. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica fundamental en estadística y análisis multivariado que se utiliza para la reducción de dimensionalidad y la identificación de patrones subyacentes en conjuntos de datos. Su objetivo principal es transformar un conjunto de variables originales en un conjunto nuevo de variables no correlacionadas llamadas çomponentes principales".
El procedimiento para obtener los componentes principales implica una serie de cálculos y transformaciones fundamentales que serán detalladas a continuación. Este proceso, esencial en el Análisis de Componentes Principales (PCA), abarca desde el cálculo de la matriz de covarianza hasta la determinación de los eigenvectores, eigenvalores, delineando una metodología rigurosa para la obtención de dimensiones fundamentales que capturan la variabilidad intrínseca en los datos originales.

c. Estandarizar datos

La estandarización de datos desempeña un papel fundamental en el contexto del Análisis de Componentes Principales (PCA). Este proceso resulta esencial por diversas razones. En primer lugar, elimina las diferencias de escala entre las variables, evitando que aquellas con magnitudes numéricas más grandes dominen la varianza total y distorsionen la identificación de los componentes principales. Al ajustar las variables para que tengan media cero y desviación estándar uno, se logra una base común que asegura que todas las variables contribuyan equitativamente al análisis, independientemente de sus unidades de medida.

La estandarización también equilibra la importancia relativa de las variables, facilitando la comparación y evaluación coherente de sus relaciones. Los coeficientes de los componentes principales, que representan las combinaciones lineales de las variables originales, se vuelven más interpretables al estar en la misma escala. Además, este proceso mitiga los efectos de los valores atípicos, reduciendo la influencia desproporcionada de datos extremos en la covarianza y, por ende, en PCA.

Así mismo la estandarización contribuye a la comparabilidad y reproducibilidad de los resultados. Al eliminar la sensibilidad a las unidades de medida específicas de las variables originales, los resultados de PCA se vuelven más fácilmente comparables entre diferentes conjuntos de datos y análisis. Dicha estandarización se lleva acabo de la siguiente manera:

$$x_k^s = \frac{x - \mu_x}{\sigma_x} \tag{2}$$

Donde μ_x es la media de la columna y σ_x es la desviación estándar muestral. El resultado de la estandarización contiene los mismos parámetros que una distribución normal.

Para el análisis a realizar se utilizó la librería sklearn.preprocessing, la cual proporciona herramientas para la estandarización de datos, proceso que se muestra a continuación:

```
min_max_scaler = MinMaxScaler()
country_scale = min_max_scaler.
                fit_transform(country.drop
                ('country',axis=1))
country_scale_df = pd.DataFrame
                (data = country_scale,
```

d. Eigenvectores, Eigenvalores

Para determinar los componentes principales de la matriz de covarianza es necesario recurrir a conceptos del álgebra lineal tal como eigenvectores y eigenvalores. Los eigenvalores en el contexto del PCA (Análisis de Componentes Principales) representan la varianza explicada por cada componente principal (eigenvector). Estos valores indican la importancia relativa de cada componente principal en la representación de la variabilidad total de los datos originales.

Cuando realizas PCA, obtienes un conjunto de eigenvalores asociados a los eigenvectores que representan las direcciones principales en las que los datos tienen más variabilidad. Estos eigenvalores están ordenados de mayor a menor, lo que significa que el primer eigenvalor es el más grande y así sucesivamente.

La suma de todos los eigenvalores es igual a la varianza total de los datos originales. Por lo tanto, puedes calcular la proporción de varianza explicada por cada componente principal dividiendo el eigenvalor correspondiente por la suma total de los eigenvalores. Esto te da una idea de cuánta información de variabilidad está capturando cada componente principal. Es decir los eigenvalores te dan información sobre la importancia relativa de cada componente principal en la representación de la variabilidad de tus datos.

Eigenvalores más grandes indican que el componente principal asociado captura más variabilidad en los datos originales. Los Componentes Principales son construidos como combinaciones lineales de las variables iniciales y en esencia reducen la mayor cantidad de información posible en el primer componente, luego la máxima restante en el segundo y así consecutivamente con el resto de la información; representan la dirección en la que los datos explican la mayor cantidad de varianza. La relación es que entre mayor varianza lleve la línea mayor, mayor dispersión de los datos a lo largo de ella, y cuanta mayor dispersión a lo largo de una línea más información contiene[3].

Supongamos que tenemos la siguiente matriz de covarianza:

$$A = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

Los eigenvectores (v) y los eigenvalores (λ) se encuentran resolviendo la ecuación característica:

$$|A - \lambda I| = \lambda^2 - \text{traza} \cdot \lambda + \text{determinante} = 0 \qquad (3)$$

Los eigenvectores son los que determinan la dirección del eje donde hay mayor varianza y los eigenvalores su magnitud correspondiente. Los eigenvalores λ_1 y λ_2 son:

$$\lambda_1, \lambda_2 = \text{eigenvalores de } A$$

Los eigenvectores v_1 y v_2 correspondientes son:

$$v_1 = \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix}, \quad v_2 = \begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix}$$

Una vez que tenemos los eigenpares los ordenamos de manera descendente de manera que podemos visualizar los componentes con mayor relevancia. Para calcular el peso de cada eigenvalor se divide cada eigenvalor por la suma de todos.

Este proceso es realizado con ayuda de la función eigh de NumPy para calcular los eigenvectores y eigenvalores de la matriz de covarianza (covmatrix). Los eigenvectores están almacenados en eigenvectors, y los eigenvalores en eigenvalues.

```
eigenvalues, eigenvectors = np.linalg.
                eigh(cov_matrix)
sorted_indices = np.argsort
                (eigenvalues)[::-1]
eigenvalues = eigenvalues
                [sorted_indices]
eigenvectors = eigenvectors
                [:, sorted_indices]
explained_variance_ratio = eigenvalues
                / np.sum(eigenvalues)
cumulative_explained_variance =
                np.cumsum(explained_variance
                _ratio)
```

Un uso inmediato de estos valores es la generación de un gráfico que ilustra la varianza acumulativa explicada por cada componente principal. Este gráfico proporciona una representación visual esclarecedora de cuánta varianza total está representada o absorbida a medida que se incorporan sucesivamente los componentes principales. La curva resultante permite tomar decisiones informadas sobre la cantidad óptima de componentes principales a retener en función de la proporción deseada de varianza explicada, contribuyendo así a la eficiente reducción de dimensionalidad y a la preservación de la información crucial en el conjunto de datos.

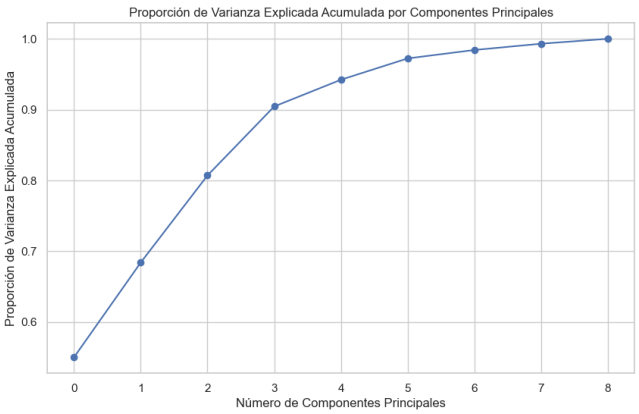

Fig. 3: Gráfico de Varianza Acumulativa: Ilustra cómo la incorporación sucesiva de componentes principales contribuye a la acumulación de varianza explicada en el conjunto de datos.

Este gráfico resulta esencial para determinar el número óptimo de componentes principales a retener, considerando la cantidad de varianza que estos explican. La curva traza cómo la varianza acumulada varía al incluir progresivamente más componentes principales. Es crucial destacar que el objetivo es seleccionar el menor número de componentes posible para facilitar el análisis subsiguiente, manteniendo al mismo tiem-

po la mayor cantidad de información. En este contexto, se ha optado por la selección de 5 componentes principales, ya que este número logra alcanzar una varianza del .98. Esto implica que el 98 por ciento de la información presente en los datos originales puede ser modelada por estos 5 componentes principales. Se tomó esta decisión considerando que seleccionar un número mayor de componentes apenas aumentaría la precisión, mientras que la complejidad para procesar los datos aumentaría significativamente.

e. Componentes Principales

Una vez determinada la cantidad óptima de componentes principales para el análisis, procedimos al cálculo de dichos componentes. En el entorno de programación Python, empleamos la clase PCA de la biblioteca scikit-learn[4], especializada en el análisis de componentes principales [5].
Este paso implica la identificación de los ejes principales que maximizan la varianza en los datos. Posteriormente, transformamos nuestros datos originales en un nuevo conjunto denominado countrypca, compuesto por los componentes principales calculados.
Una vez obtenidos los componentes principales, organizamos estos datos en un DataFrame llamado countrypcadf, cuyas columnas se etiquetan como 'principal component 0', 'principal component 1', y así sucesivamente, hasta el número de componentes especificado (en este caso, 5).
Este DataFrame, countrypcadf, representa de manera tabular nuestros datos en términos de los componentes principales calculados. Con este resultado, estamos preparados para realizar análisis avanzados, como visualizaciones, clustering, u otras exploraciones relevantes. Este enfoque nos permite trabajar de manera efectiva con la información resumida en los componentes principales, facilitando la realización de análisis más profundos en nuestro conjunto de datos.

El proceso mencionado anterior se puede resumir en las siguientes lineas de código:

```
#Componentes principales
country_pca = PCA(n_components=5).
              fit(country_scale).
              transform(country_scale)

country_pca=pd.DataFrame(data =
            country_pca,
            columns=[
            'principal component 0',
            'principal component 1',
            'principal component 2',
            'principal component 3',
            'principal component 4'])
country_pca.head()
```

Ahora bien, con el propósito de evaluar la efectividad de los componentes principales obtenidos, procedemos a calcular una matriz de correlación entre ellos. En esta etapa, comparamos los diferentes componentes entre sí, anticipando una correlación mínima. Este escenario se justifica por la naturaleza de los componentes principales, los cuales representan vectores ortogonales. La ortogonalidad implica que cada componente se esfuerza por captar la máxima información independiente de los demás, resultando en una matriz de correlación idealmente caracterizada por coeficientes cercanos a cero. Dicha baja correlación reflejaría la capacidad de cada componente principal para abordar aspectos únicos y no redundantes de la variabilidad en los datos, proporcionando así una visión más clara y desglosada de la estructura subyacente en nuestro conjunto de datos.

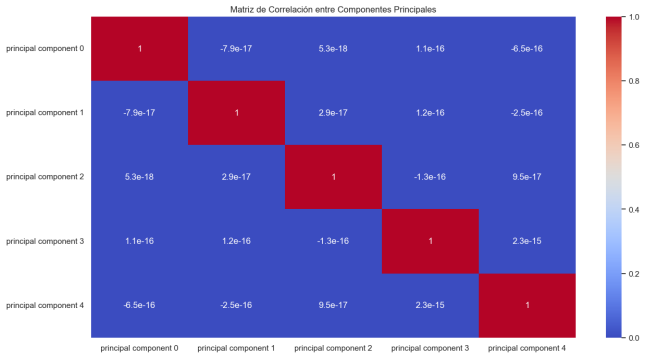

Fig. 4: Matriz de Correlación entre Componentes Principales.

Como se puede observar en el mapa de calor, existe una correlación prácticamente nula entre los diferentes componentes principales obtenidos. Este fenómeno se refleja en la matriz de correlación, donde cada uno de los valores presenta una tendencia a cero. La tonalidad azul en el mapa de calor indica una baja correlación entre los componentes principales. Esta evidencia respalda la conclusión de que los componentes principales obtenidos son efectivos, ya que capturan de manera independiente y significativa la información presente en los datos originales.

Cada variable en la base de datos original ejerce una influencia específica en cada componente principal. Esto implica que ciertos componentes principales tienden a captar más información sobre algunas variables en comparación con otras. La selección del número adecuado de componentes principales es crucial, ya que optar por un número insuficiente podría resultar en una representación menos precisa de algunas variables originales. Este fenómeno se conoce como "loadings", que representa el peso de cada variable en cada componente principal. A continuación, se presentan los loadings de los componentes principales, destacando la importancia relativa de cada variable en la formación de los componentes principales.

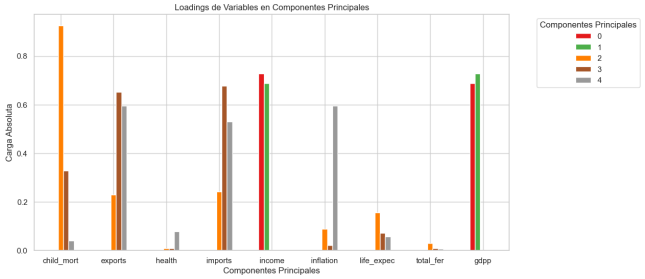

Fig. 5: "Loadings de Variables en Componentes Principales: Distribución de Pesos

En la figura 5, se destaca la influencia de ciertas variables en cada componente principal. Por ejemplo, se observa que la tasa de mortalidad infantil (childmort) ejerce un impacto significativo en el cálculo del componente principal número 2. Asimismo, la variable de salud (health) muestra una con-

tribución relativamente baja a todos los componentes principales.

Con esta información (loadings) es posible llevar acabo un análisis más detallado, por ejemplo analizando, podemos observar que el componente principal 4 está notablemente afectado por los pesos asociados a las variables de exportaciones, importaciones e inflación. Esto sugiere que el componente principal 4 podría ser particularmente útil para realizar un análisis más específico sobre aspectos financieros o la dinámica económica en general.

f. Uso de componentes principales para clasificación de paises

Para llevar acabo el agrupamiento de paises con ciertas similutudes y de tal manera descubrir patrones ocultos se utilizo k-means clustering. El algoritmo K-means identifica k número de centroides y luego asigna cada punto de datos al grupo más cercano, manteniendo los centroides lo más pequeños posible. Este centroide se refiere a la ubicacion real o imaginaria de un grupo de datos..
En el contexto del algoritmo K-means, se busca realizar una partición de un conjunto de datos en **K** clústeres, minimizando la suma de cuadrados de las distancias entre los puntos y los centroides de sus clústeres correspondientes.

Inicialmente, se selecciona el número de clústeres **K** y se eligen centroides iniciales. Posteriormente, se realiza la asignación de cada punto al clúster cuyo centroide está más cercano, utilizando la distancia Euclidiana. Los centroides se actualizan recalculándolos como la media de los puntos asignados a cada clúster.

Este proceso de asignación y actualización se repite iterativamente hasta que los centroides de los clústeres convergen o se alcanza un número máximo de iteraciones. La función objetivo que se busca minimizar es la suma de cuadrados de las distancias al cuadrado entre los puntos y sus centroides.

Es importante destacar que el algoritmo K-means es sensible a la inicialización, y la elección adecuada del número de clústeres **K** es crucial. La convergencia del algoritmo se determina cuando los centroides de los clústeres dejan de cambiar significativamente entre iteraciones. [6].

En la elección del número de clústeres, se empleó el método del codo. Este enfoque se fundamenta en la observación de que, al aumentar el número de clústeres, la reducción en la suma de cuadrados tiende a disminuir, dando lugar a la formación de un característico çodo.[en] el gráfico. El punto en el cual se aprecia este quiebre o codo se considera como el número óptimo de clústeres. Este criterio de selección se basa en minimizar la suma de cuadrados de las distancias entre los puntos y los centroides de los clústeres **K** , proporcionando así una guía intuitiva para la determinación del número adecuado de agrupamientos en el conjunto de datos analizado.

Realizando el análisis de codo se obtienen los siguientes resultados:

Con base en el análisis previo, se concluye que el número óptimo de clústeres **K** es 4. Este resultado establece la fundamentación para llevar a cabo el proceso de K-means, permi-

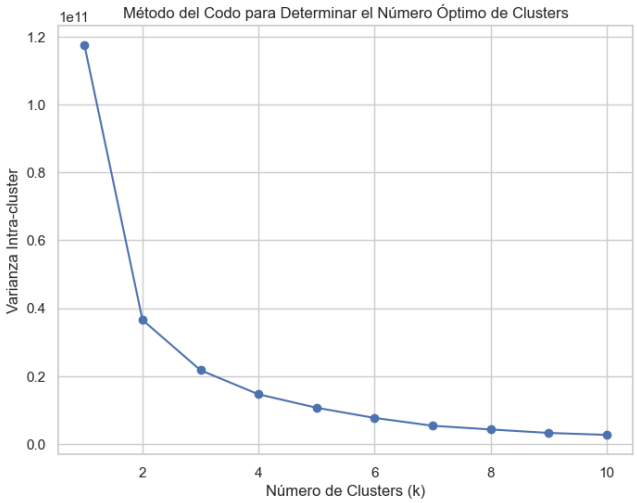

Fig. 6: Método de codo para determinar el número óptimo de Clusters.

tiendo la agrupación efectiva de los datos en cuatro clústeres distintos.Para realizar dicho proceso, se utiliza el metodo fit-predict de Sklearn como se muestra a continuacion:

```
num_clusters = 4
# Inicializar el modelo de KMeans
kmeans = KMeans(n_clusters=num_clusters,
        random_state=42)
# Aplicar KMeans a los datos despu s de PCA
clusters = kmeans.fit_predict(country_pca)
# Agregar la columna 'cluster'
country['cluster'] = clusters
```

La clasificación de todos los paises se puede consultar, haciendo uso del notebook de Jupyter, adjuntado siguiendo el link a GitHub presente en la sección de recursos adicionales.

g. Regresión lineal para modelar el comportamiento de una variable

Después de realizar la clasificación de países utilizando el método K-Means, surgió el interés en explorar la versatilidad de los componentes principales obtenidos mediante el PCA. Buscando una comprensión más profunda de las relaciones socioeconómicas, se llevó a cabo un análisis adicional que expande a otros usos fuera de la mera clasificación. Se eligió el análisis de regresión lineal para examinar cómo estos componentes principales podrían predecir el comportamiento de una métrica socioeconómica original.

La regresión lineal es un modelo estadístico que busca establecer la relación lineal entre una variable dependiente (o de respuesta) y una o más variables independientes (o predictoras). En su forma más simple, con una variable independiente, se expresa como

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde β_0 es la intercepción, β_1 es la pendiente, y ε es el término de error.

La regresión lineal con componentes principales implica realizar un análisis de componentes principales (PCA) sobre el conjunto de datos original. Esto transforma las variables

originales en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Luego, estos componentes se utilizan como predictores en un modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k + \varepsilon$$

El objetivo de la regresión lineal es encontrar los valores deB_0 y B_1 que minimizan la suma de los cuadrados de los errores.

La aplicación de la regresión lineal con componentes principales implica primero realizar un análisis de componentes principales (PCA) sobre el conjunto de datos original. PCA transforma las variables originales en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Luego, estos componentes se utilizan como predictores en un modelo de regresión lineal. La idea aquí es que los componentes principales capturan la variabilidad más importante en los datos originales, permitiendo así una representación más eficiente de las relaciones lineales subyacentes.

Al implementar la regresión lineal con componentes principales, es esencial seleccionar el número adecuado de componentes que retengan la mayor parte de la variabilidad. Esto se puede lograr mediante análisis de varianza explicada o métodos que seleccionen un número óptimo de componentes. Para el proceso completo realizado de la regresión lineal, consultar el link de GitHub ubicado en la sección de recursos adicionales.

Esta elección se basó en la capacidad del análisis de regresión para revelar relaciones lineales o casi lineales entre las variables predictivas y la métrica de interés, ofreciendo una visión clara y cuantitativa de su influencia. Además, esta técnica apuntaba no sólo a predecir el comportamiento futuro de la métrica elegida sino también a comprender cómo los componentes principales cuantificaban esa relación, más allá de la mera agrupación o clasificación de países. Los resultados proporcionaron una comprensión más amplia de cómo estas variables socioeconómicas se relacionan con la realidad de las exportaciones, mostrando cómo la combinación de métodos analíticos puede conducir a una comprensión más profunda de estos contextos complejos.

III. Análisis de Resultados

La reducción de dimensionalidad desempeña un papel crucial en el análisis de conjuntos de datos complejos y de gran escala. Al abordar conjuntos de datos que contienen numerosas variables, la reducción de dimensionalidad permite simplificar la complejidad al transformar el conjunto de datos original en un espacio de menor dimensión. Esta técnica no solo facilita la visualización y comprensión de los datos, sino que también puede mejorar significativamente el rendimiento de los modelos de aprendizaje automático y las técnicas estadísticas.

En el contexto de clasificación de países y regresión lineal, la reducción de dimensionalidad se vuelve especialmente relevante. Al reducir el número de variables, se puede mitigar el riesgo de sobreajuste, mejorar la eficiencia computacional y, en última instancia, lograr un mejor rendimiento predictivo. Además, la selección adecuada de componentes principales o características es esencial para capturar de manera efectiva la información esencial contenida en los datos.

En este contexto, se optó por realizar dos análisis haciendo uso de los componetes principales previamente calculados.

- Clasificación de paises (clusters)
- Regresión Lineal para modelar el comportamiento de una variable especifica

A continuación se muestrán los resultados obtenidos, despues de llevar acabo el respectivo análisis, haciendo uso de componentes principales.

a. Clasificación de paises (clusters)

Una vez completada la clasificación de los países en cuatro categorías distintas (clusters), es posible visualizar e interpretar esta agrupación de diversas maneras. En este caso, resulta especialmente interesante representar los países utilizando colores diferentes según el cluster al que pertenecen, en un mapa mundial. Sin embargo, la peculiaridad de este análisis radica en que se llevó a cabo la clasificación de los países tanto antes como después de aplicar PCA.

Los países fueron sometidos al proceso de clasificación mediante el algoritmo k-means, como se explicó previamente en la sección de metodología. Este procedimiento se llevó a cabo tanto antes como después de aplicar el análisis de componentes principales (PCA). En ambos casos, el algoritmo agrupa los países que comparten similitudes entre sí. Al utilizar 4 grupos o clusters, los países se colorearán con uno de los 4 colores según el grupo al que pertenezcan. Países del mismo color indican que comparten similitudes en términos generales, considerando todas las variables socioeconómicas analizadas y no solo una característica específica.

A continuación, se presentan dos mapas que muestran todos los países incluidos en la base de datos y su respectiva asignación a los clusters antes y después de aplicar PCA. Cabe destacar que los países que no están coloreados en los mapas no están presentes en la base de datos. Una ausencia notable es la de México, razón por la cual no aparece en el mapa ni fue clasificado en esta instancia.

Analizando ambos mapas, se destacan una serie de diferencias notables. En primer lugar, al observar la Figura 6 (primer mapa), donde se clasifican los países antes de aplicar el Análisis de Componentes Principales (PCA), se aprecia que prácticamente todos los países se distribuyen en tres clusters distintos, no en cuatro. Es evidente que la mayoría de los países pertenecientes al bloque occidental, comúnmente denominados "Países de primer mundo", se encuentran concentrados en el cluster rojo. Además, la gran mayoría de los países que pertenecieron al antiguo bloque comunista están categorizados conjuntamente en el

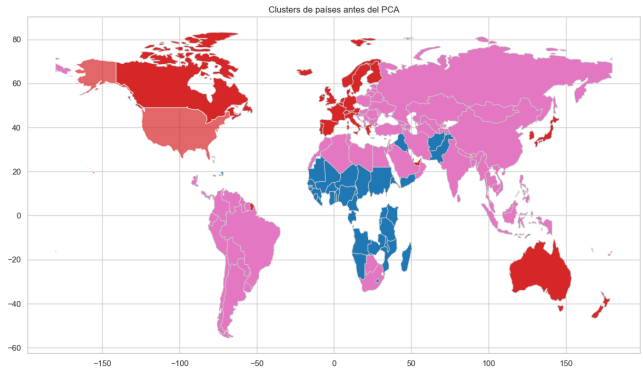

Fig. 7: Clasificación de paises antes de aplicar PCA.

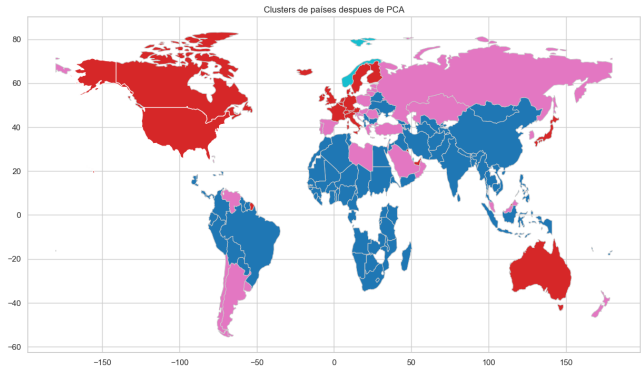

Fig. 8: Clasificación de paises despues de aplicar PCA.

cluster rosa. Por otro lado, aquellos países que suelen ser catalogados como menos desarrollados o más pobres están agrupados en el cluster de color azul. Estas observaciones revelan patrones interesantes en la distribución de los países antes de la aplicación del PCA, proporcionando insights valiosos sobre la agrupación de naciones en función de variables socioeconómicas clave.[7]

Una vez aplicado el Análisis de Componentes Principales (PCA), en la figura 7, se revelan patrones interesantes que enriquecen nuestra comprensión de la distribución de los países. En primer lugar, se observan ahora cuatro clusters distintos. El cluster representado en 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. El cluster rojo, que anteriormente agrupaba mayormente a los países del bloque occidental, ahora muestra una subdivisión, incluyendo a España, Portugal y Grecia en un cluster separado. Estos países, a pesar de ser parte del bloque occidental, han enfrentado recientemente desafíos económicos y migratorios que podrían haber impactado su bienestar general. 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.

Este análisis post-PCA proporciona una visión más matizada de la distribución de los países, permitiendo identificar matices y tendencias que no eran tan evidentes en la clasificación previa. Estos resultados contribuyen significativamente a nuestro entendimiento de la variabilidad socioeconómica a nivel global.

Gracias a la aplicación de PCA, el análsis realizado resulta más preciso. Una vez aplicado el Análisis de Componentes Principales (PCA), se revelan patrones interesantes que enriquecen nuestra comprensión de la distribución de los países. En primer lugar, se observan ahora cuatro clusters distintos. Este análisis post-PCA proporciona una visión más matizada de la distribución de los países, permitiendo identificar matices y tendencias que no eran tan evidentes en la clasificación previa. Estos resultados contribuyen significativamente a nuestro entendimiento de la variabilidad socioeconómica a nivel global.

b. Regresión Lineal para modelar el comportamiento de una variable especifica

Despues de haber llevado acabo la clasificación de los paises, para ilustrar la versatilidad y gran aplicabilidad de los compoentes principales, se llevo acabo un análisis secundario en el cual se opto por implementar una regresión lineal para predecir el comportamiento de una de las metricas socieconomicas originales.

Una vez completada la clasificación de países mediante el algoritmo K-Means, nuestro interés se volcó hacia la ilustración de la versatilidad y la aplicabilidad de los componentes principales obtenidos a través del Análisis de Componentes Principales (PCA). Con el objetivo de profundizar aún más en la comprensión de las interrelaciones socioeconómicas, se llevó a cabo un análisis secundario que se apartó de la clasificación pura para explorar la capacidad predictiva de los componentes principales.

En este análisis secundario, optamos por implementar una regresión lineal, una herramienta estadística poderosa que nos permitió prever el comportamiento de una de las métricas socioeconómicas originales. La elección de esta métrica específica se basó en su relevancia y su capacidad para proporcionar información crucial sobre la realidad económica de los países.

La elección de utilizar regresión lineal en el contexto de las exportaciones se sustenta en varios motivos. En primer lugar, la regresión lineal se selecciona cuando existe una aparente relación lineal o aproximadamente lineal entre las variables predictoras y la variable de respuesta, en este caso, las exportaciones. Además, la interpretación sencilla de los coeficientes y la claridad del modelo hacen que sea una elección lógica si se busca una explicación directa de cómo las variables predictoras afectan las exportaciones.

La regresión lineal no solo busca prever el comportamiento futuro de la métrica seleccionada, sino que también arroja luz sobre la relación cuantitativa entre los componentes principales y la variable socioeconómica en cuestión. Este análisis adicional no solo es una extensión natural de la aplicación de PCA, sino que también demuestra cómo las técnicas matemáticas avanzadas pueden ser empleadas para ir más allá de la simple clasificación y ofrecer una visión más profunda y predictiva de la realidad socioeconómica.

Una vez realizado el proceso de regresión lineal, explicado

en la sección de metología, los resultados de la regresión lineal con componentes principales (PCA) proporcionan una visión significativa sobre la relación entre variables socioeconómicas y las exportaciones de los países. A continuación se muestrán los coeficientes correspondientes a la regresión lineal, para hacer su posterior análisis:

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	$9{,}5403$

TABLE 1: RESULTADOS DE LA REGRESIÓN LINEAL CON COMPONENTES PRINCIPALES.

Los coeficientes obtenidos indican la influencia de cada componente principal en las exportaciones. Es notable que el quinto componente principal tiene un coeficiente positivo significativo (0.656), sugiriendo que un aumento en este componente está fuertemente asociado con un incremento en las exportaciones. Este hallazgo resalta la importancia de ciertos factores latentes no observados directamente en las variables originales.

El intercepto, con un valor de 41.03, representa el nivel

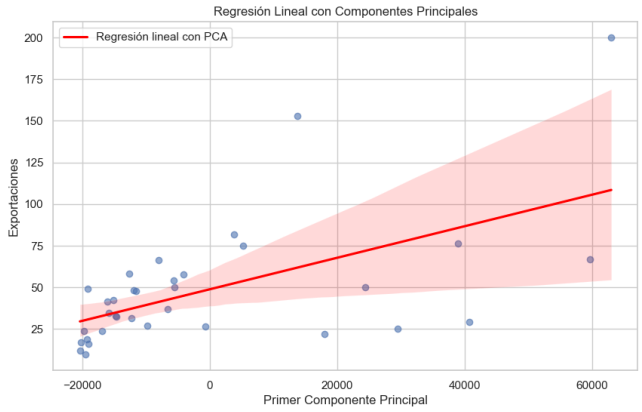

Fig. 9: Regresión Lineal usando 1 componente principal

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, proporcionando un punto de referencia esencial.

La evaluación del modelo reveló que la inclusión de cinco componentes principales mejoró significativamente la capacidad predictiva, evidenciado por un MSE más bajo (9.54)[8]. Esto sugiere que la información capturada por estos componentes adicionales es crucial para explicar la variabilidad en las exportaciones. Esto se puede apreciar gráficamente en la figura 9.

En términos generales, los resultados subrayan el papel esencial de PCA en la regresión lineal para comprender y predecir las exportaciones. La capacidad de PCA para reducir la dimensionalidad y resaltar patrones latentes en los datos se traduce en un modelo más preciso y robusto. La elección cuidadosa de componentes principales emerge como un aspecto crítico para un análisis detallado de las relaciones socioeconómicas y su impacto en las exportaciones a nivel global.

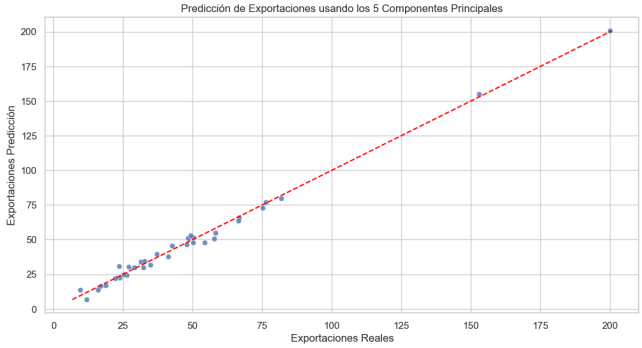

Fig. 10: Regresión Lineal usando 5 componentes principales

IV. CONCLUSIONS

A lo largo de este extenso análisis socioeconómico global, hemos empleado herramientas matemáticas avanzadas para abordar la complejidad intrínseca de los datos y desentrañar patrones significativos. La aplicación del Análisis de Componentes Principales (PCA) no solo ha permitido la reducción de dimensionalidad, sino que ha revelado la estructura subyacente que conecta variables aparentemente dispares. A través de este proceso, hemos logrado una comprensión más profunda y simplificada de las complejas interrelaciones que definen la realidad socioeconómica global.

La fase de clasificación mediante el algoritmo K-Means ha añadido otra capa de comprensión, proporcionando una visión clara de cómo los países se agrupan en función de sus perfiles socioeconómicos. La aplicación de K-Means antes y después de PCA ha destacado la eficacia de la reducción de dimensionalidad en la identificación de patrones y similitudes, revelando estructuras que podrían haber pasado desapercibidas en el análisis de datos originales.

La versatilidad de los componentes principales se ha destacado aún más a través de la implementación de una regresión lineal. Este paso ha ido más allá de la clasificación pura, permitiéndonos prever el comportamiento futuro de una métrica socioeconómica clave, en este caso, las exportaciones. La regresión lineal, basada en los componentes principales derivados del PCA, no solo ha ofrecido predicciones cuantitativas sino que también ha proporcionado una comprensión más profunda de la relación entre las variables y los componentes principales.

En conjunto, este trabajo representa un enfoque integral para el análisis de datos socioeconómicos a nivel global. Desde la preparación inicial de datos hasta la aplicación de técnicas avanzadas como PCA y K-Means, cada etapa ha sido guiada por la búsqueda de patrones significativos y la simplificación de la complejidad inherente a los conjuntos de datos multidimensionales.

La contribución fundamental de este estudio radica en la capacidad de desentrañar los hilos entrelazados de la realidad socioeconómica global, proporcionando una visión más clara y contextualizada. La aplicación de técnicas matemáticas avanzadas ha demostrado ser una herramienta invaluable para la exploración de complejidades, facilitando no solo la identificación de patrones, sino también la predicción del comportamiento futuro.

Es claro que la integración de PCA, K-Means y regresión lineal no solo ha ampliado nuestra comprensión de las interacciones socioeconómicas globales, sino que también ha sentado las bases para futuras investigaciones y aplicaciones prácticas. Este enfoque metodológico robusto no solo es aplicable al contexto específico de este estudio, sino que también sirve como un paradigma para el abordaje de complejidades similares en diversos campos. En última instancia, esta investigación destaca la importancia de la integración de herramientas matemáticas avanzadas para desentrañar los misterios que yacen en la intersección de datos socioeconómicos globales, contribuyendo así al continuo desarrollo del conocimiento en este campo dinámico.

a. Recursos Adicionales

En el siguiente enlace de GitHub, se encuentra disponible la base de datos previamente depurada, la cual fue empleada para llevar a cabo los análisis detallados a lo largo de este trabajo. Además, en este repositorio de GitHub, se incluye todo el código utilizado para los análisis descritos, presentado de manera integral en un cuaderno de Jupyter Notebooks. Este recurso facilita el acceso a la información subyacente y brinda transparencia sobre el proceso analítico llevado a cabo en el desarrollo de este trabajo. Puedes acceder al repositorio a través del siguiente enlace: Enlace de GitHub

REFERENCIAS

[1] W. H. Greub, *Linear algebra*, vol. 23. Springer Science & Business Media, 2012.

[2] E. Bisong, *Matplotlib and Seaborn*, pp. 151–165. Berkeley, CA: Apress, 2019.

[3] J. H. Wilkinson, F. L. Bauer, and C. Reinsch, *Linear algebra*, vol. 2. Springer, 2013.

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

[5] M. 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, p. 100, 2022.

[6] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*. Cengage Learning,https://doi.org/10.1007/978-1-4842-4470-8, 2015.

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

[8] E. Kreyszig *et al.*, "Matemáticas avanzadas para ingeniería," 2001.