Análisis de sistemas de infraestructura parcial 1 parte 2

Karol Rivera - 201815009

11 de octubre de 2023

Parte 1 - PCA

Para comenzar se deben importar los datos. Originalmente se contaba con un archivo de excel (P1-CIndexN.xlsx), el cual fue dividido en dos archivos:

- P1-CIndexN.xlsx: El archivo contiene la información de los indicadores para cada país. En las filas están los indicadores, en las columnas los países.
- P1-CIndex_Score.xlsx: Contiene el puntaje que obtiene cada país. En la fila está el Score, en las columnas los países.

Antes de realizar cualquier procesamiento sobre los datos es necesario eliminar columnas con información poco relevante, así como acomodar el dataframe. Para esto, se cargan los datos de ambos exceles usando la función *read_excel* de *pandas*. En el excel con la información de los indicadores se elimina la columna 'Series units', se define la columna 'Series name' como el index, se transpone la base de datos y se define como nuevo index la columna 'Country Name'. En el excel con el puntaje para el segundo pilar de infraestructura se elimina también la columna 'Series units', se define 'Series name' como los index, se transpone el arreglo de datos, y se define como nuevo index 'Country Name'. De este modo el excel importado queda con los países como filas, y los indicadores o el score como columnas. Para tener una sola base de datos se une el dataframe con los indicadores con el dataframe con el score. Esto se hace mediante la función *merge* de *pandas*. Se unen mediante el index dado por 'Country Name' y se define la función con el parámetro 'how=outer' de modo que se incluya toda la información contenida en ambos arreglos de datos. Vale la pena mencionar que algunos indicadores de sistemas férreos cuentan con 'Nan', en este caso se rellenaron los 'Nan' con '0' asumiendo que si no se tiene información es porque no se tiene este sistema (se usó *fillna(0,inplace=True)*). Luego de todo este procedimiento se cuenta con un arreglo de 140 países, 18 indicadores y el score.

A continuación se debe realizar el preprocesamiento de los datos. Dado que no se trata de una serie de tiempo, sino que se tienen datos asociadas a varias variables, se deben eliminar los países con datos faltantes. Para esto se usa la función *dropna* ajustando el parámetro 'how=any' de modo que se elimine cualquier país con información faltante. Luego de este proceso se obtiene un arreglo con 85 países y 19 columnas entre indicadores y el score.

Para realizar PCA inicialmente se deben estandarizar los datos. Este proceso convierte la media igual a cero y la varianza igual a 1, además de que hace los datos más comparables entre sí. El proceso es realizado creando la instancia *StandardScaler()*, y solicitándole a la instancia realizar un fit sobre los datos. Con los datos estandarizados se puede realizar el análisis de componentes principales instanciando *PCA()* de la librería *sklearn*. Finalmente, se pueden obtener los λ de cada componente, y normalizar para obtener la varianza explicada.

De la gráfica de varianza explicada por cada componente se puede concluir que el primer componente almacena más del 50 % de la información contenida en los datos, el segundo aproximadamente el 10 % de la información, y los otros componentes almacenan menos de 10 % cada uno. Se observa además que la cantidad de información almacenada por componente tiene una tendencia decreciente. Por otra parte, de la gráfica de varianza explicada acumulada se puede corroborar el hecho que la cantidad de información almacenada decrece conforme se avanza en los componentes, motivo por el cual la pendiente de la curva va disminuyendo. En esta gráfica se tiene además una línea horizontal que marca el límite en el que se tiene 70 % de la información, de modo que se evidencia la necesidad de usar 4 componentes para conservar esta cantidad de información (se tendría el 74 % de la información). En caso de usar 3 componentes solo se tendría el 68.78 % de la información por lo que serían insuficientes para superar el límite.

Del análisis anterior se puede concluir que tomando los primeros componentes se va a abarcar la mayor parte de la información contenida en los datos. Para mostrar la utilidad de este análisis se puede, por ejemplo, tomar los primeros 3 componentes y transformar los datos a un espacio 3D compuesto por estos componentes (usando *PCA.transform()*).

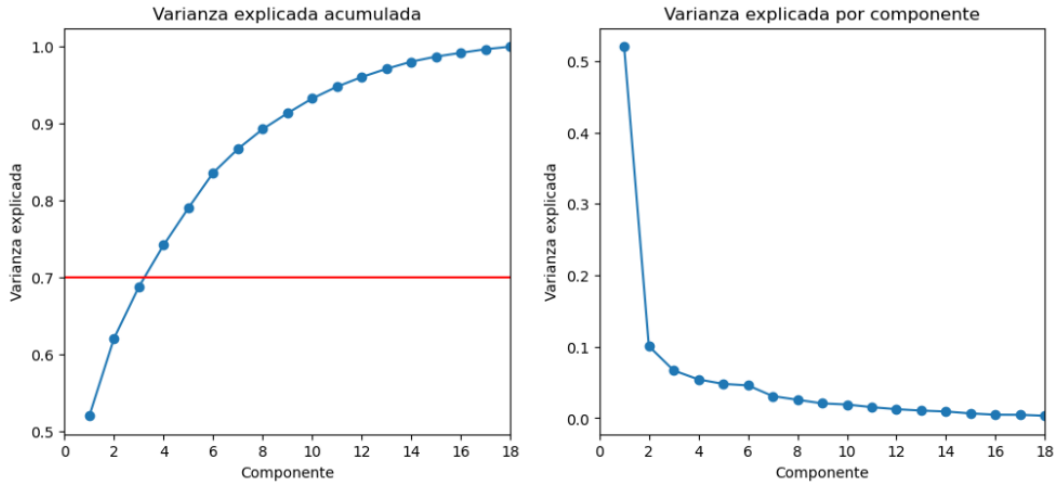

Figura 1: Resultado del PCA tomando todos los componentes principales. La gráfica de la izquierda representa la varianza explicada acumulada. La gráfica de la derecha explica la varianza explicada por cada componente.

Además, sería interesante buscar tendencias sobre los datos en este espacio transformado. Por lo anterior se seleccionan países de diferentes agupaciones:

- América Latina: Colombia, Argentina, Peru, Chile y Ecuador.
- OCDE: Alemania, Italia, España, Estados Unidos y Canadá.
- Africa: Nigeria, Tanzania, Sur África, Kenia y Yemen.
- Asia: India, China, Indonesia, Japón y Tailandia.

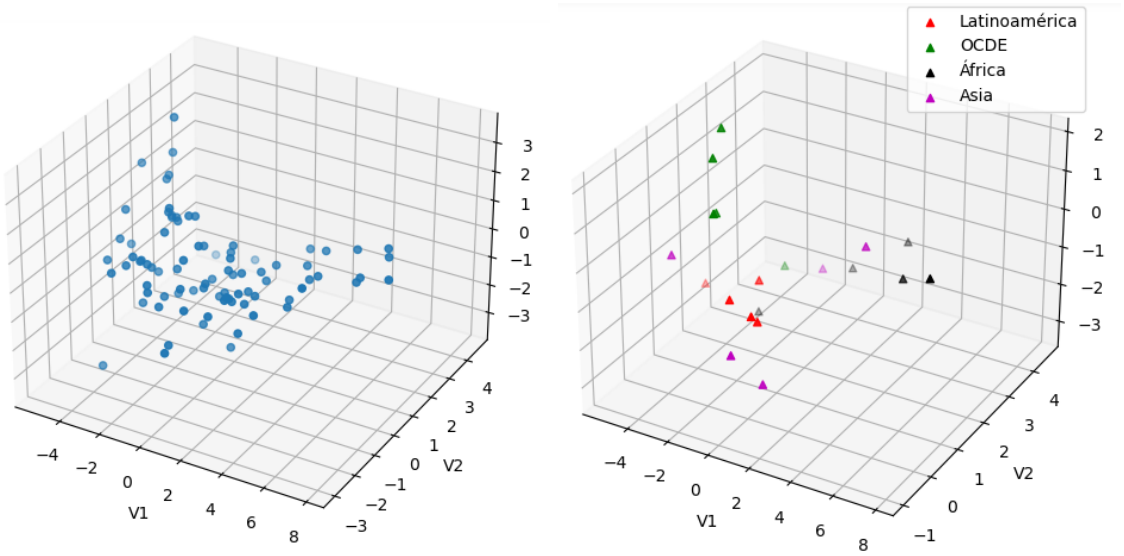

Figura 2: Datos proyectados a un espacio 3D formado por los primeros tres componentes principales. A la izquierda se observa la distribución de todos los datos, a la derecha solo las agrupaciones.

Se pueden sacar algunas conclusiones con respecto a los países agrupados. Por ejemplo, se observa que los países que hacen parte de la OCDE se encuentran agrupados en una zona dónde no hay datos de las otras agrupaciones. Esto puede responder al hecho de que los países usados son reconocidos por tener un nivel de desarrollo muy significativo, sobre todo en términos de infraestructura (Todos son países considerados como desarrollados). En este grupo hay un país que se encuentra separado del grupo sin coíncidir con otros, sin embargo, los otros están marcadamente agrupados. También se observa que los países latinoamericanos se encuentran concentrados en una zona distinta a la de los países de la OCDE.

Esta diferencia puede responder al hecho de que estos países son reconocidos por estar en un nivel de desarrollo intermedio (países en vía de desarrollo). Los países africanos se encuentran un poco más distribuidos, pero igualmente ubicados en una zona diferente a los países latinoamericanos y los de la OCDE. Solo se ubica un país cerca al grupo de los latinoamericanos. Esto puede responder al hecho de que los países africanos son, en su mayoría países subdesarrollados, por lo que son diferentes a los primeros dos grupos. Finalmente, los países asiáticos se encuentran mucho más distribuidos, no se concentran en una zona específica del diagrama. Esto se relaciona con el hecho de que los países asiáticos cuentan con variaciones en nivel de desarrollo. Vale la pena aclarar que se habló de nivel de desarrollo, el cual puede relacionarse con la infraestructura con la que cuenta cada país, y por lo tanto, con los indicadores.

Hablando a nivel de la posición con respecto a los 3 componentes, se observa que los desarrollados se encuentran en la región negativa del componente 1, en los valores más altos del componente 2 y en los valores más altos del componente 3. Por parte de los países latinoamericanos, estos se ubican en la parte negativa del componente 1, en la parte positiva del componente 2 y en los valores más negativos del componente 3. Preliminarmente se puede decir que la principal diferencia entre los países de la OCDE y los latinoamericanos es la posición en el componente 3. Por su parte, los países africanos se encuentran en valores positivos del componente 1, en valores muy positivos del componente 2 y en valores negativos intermedios del componente 3. Por lo tanto, difieren de los países de la OCDE en el componente 1 y en el componente 3, mientras que difieren de los países latinoamericanos en el componente 1. Finalmente, los países asiáticos se encuentran distribuidos por lo que no hay diferencias marcadas con respecto a las otras agrupaciones.

Teniendo en cuenta que estos componentes son los que almacenan la mayor cantidad de información, es razonable que existan diferencias en las posiciones de países de características diferentes, y evidentemente estas diferencias van a estar relacionadas con temas de infraestructura, que es la información de los indicadores usados para el PCA. Sin embargo, el hecho de que los componentes muestren diferencias entre agrupaciones de países no significa que estos representen medidas reales de alguna característica de los países. Los componentes principales en realidad pueden interpretarse como variables nuevas construidas a partir de los indicadores ingresados al lagoritmo de PCA. En ese orden de ideas los componentes pueden interpretarse como dimensiones sobre las cuales se asignan valores a partir de una combinación lineal de los indicadores iniciales.

Con el análisis de componentes principales es posible construir variables nuevas. En este caso se plantea la contrucción de un indicador que obedece a la siguiente formulación matemática:

$$I_p = \sum_{j=1}^{m} \omega_j c_{pj} \qquad \sum_{i=1}^{m} \omega_i = 1$$

donde m es la cantidad de componentes principales usados y c_{pj} es el valor de cada país con respecto al componente principal. Además, ω son los pesos de cada componente principal, y se desea realizar una normalización dependiendo de la cantidad de componentes tomados. Para calcular este indicador se crea una función que permite variar la cantidad de componentes principales tomados. Además, la función devuelve el indicador sin normalizar y el indicador normalizado. El sentido de obtener el indicador normalizado es poder comparar con el score porporcionado por el banco mundial. Se decide calcular el indicador con 2, 4, 6, 8, 10 y todos los componentes principales. El primer resultado que se puede observar del cálculo del indicador es que se obtienen valores negativos y valores positivos. Una forma de comparar el indicador al tomar diferente cantidad de componentes es calculando el rango del indicador. Esta gráfica se presenta a continuación:

Se observa entonces una disminución del rango conforme aumenta la cantidad de componentes tomados. Esta variación indica que el valor del indicador obtenido para cada país va a variar dependiendo de la cantidad de componentes. Además, la pendiente de la gráfica va disminuyendo conforme aumenta la cantidad de indicadores, de modo que la variación del rango es menor a medida que se usan más componentes. La disminución de la varianza de los datos del indicador también se puede observar en los histogramas superpuestos del indicador con 2 componentes y con todos los componentes.

Para comparar el indicador a medida que se cambian los componentes también se puede usar el indicador normalizado. Vale la pena aclarar que para la normalización de los indicadores se usa la siguiente expresión:

$$I_{normj} = \frac{I_j - min(I)}{max(I) - min(I)}$$

Bajo esta normalización el indicador va a variar entre 0 y 1. Con el indicador normalizado se quiere analizar si varían las posiciones de los países dependiendo de la cantidad de componentes incluida. Para ello se construye una tabla. Se

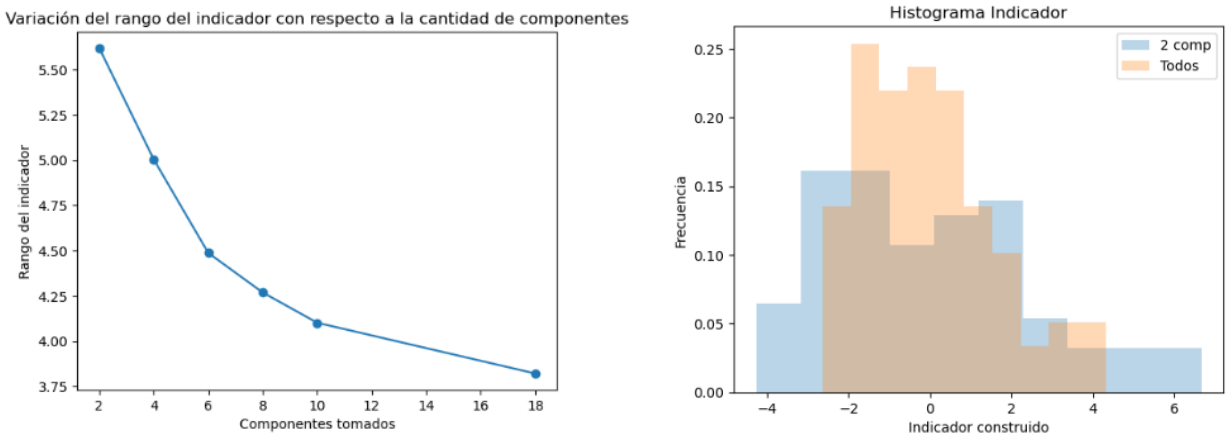

Figura 3: Variación del rango del indicador con respecto a la cantidad de componentes e histograma del indicador para 2 y todos los componentes.

puede concluir que la posición de los países va variando dependiendo de la cantidad de componentes incluidos, aunque pareciera que de los 8 componentes en adelante se mantienen las posiciones. Con estos componentes se estaría conservando casi el 90 % de la información (89.27 %). Es lógico que contando con casi toda la información, la posición de los países se estabilice para el indicador construido.

Componentes	Primeros cuatro	Últimos cuatro
2	Japón, Paises bajos, Finlandia, Suecia	Nigeria, Mozambique, Angola, Yemen
4	Japón, Emiratos Árabes, Paises bajos, Finlandia	Nigeria, Angola, Mozambique, Yemen
6	Japón, Paises bajos, Emiratos árabes, Finlandia	Nigeria, Angola, Mozambique, Yemen
8	Japón, Emiratos Árabes, Paises bajos, Finlandia	Nigeria, Angola, Mozambique, Yemen
10	Japón, Emiratos Árabes, Paises bajos, Finlandia	Nigeria, Angola, Mozambique, Yemen
Todos	Japón, Emiratos Árabes, Paises bajos, Finlandia	Nigeria, Angola, Mozambique, Yemen

Cuadro 1: Ordenamiento de países de menor a mayor dependiendo de la cantidad de componentes principales.

Estos indicadores también puede ser comparado con el score propuesto por el banco mundial para cada país. En este caso la comparación tendrá que ser entre el score normalizado y el indicador construido normalizado. Para una primera comparación se puede revisar el órden de los países. Se comparará el ordenamiento del score con el ordenamiento del indicador incluyendo todos los componentes (ya que las posiciones tienden a estabilizarse).

	Country Name	Indicador WBF norm
55	Netherlands	1.000000
41	Japan	0.984203
18	Germany	0.962919
26	France	0.961665
80	United States	0.950418
..	...	...
54	Nigeria	0.149108
14	Cameroon	0.144860
0	Angola	0.122400
51	Mozambique	0.064177
83	Yemen	0.000000

	Country Name	Ind_todos
83	Yemen	1.000000
51	Mozambique	0.946972
0	Angola	0.934450
54	Nigeria	0.871083
14	Cameroon	0.858292
..	...	...
71	Sweden	0.057765
25	Finland	0.047895
55	Netherlands	0.012653
2	United Arab Emirates	0.011207
41	Japan	0.000000

Figura 4: Odenamiento de los países. A la izquierda para el score normalizado, a la derecha para el indicador normalizado usando todos los componentes.

De la comparación del ordenamiento se observa que los países punteros en el score son los últimos en el indicador construido. Esto indica que la interpretación en términos de infraestructura del indicador debe ser que a menor puntaje

obtenido, mayor su nivel de infraestructura. Para continuar con la comparación se podrian construir los histogramas de frencuencia relativa para el score y para el indicador construido. El resultado de los histogramas muestra que el comportamiento de las variables es inverso, es decir, dónde el histograma del indicador construido presenta la mayor frecuencia, el histograma del score presenta la mínima (aproximadamente).

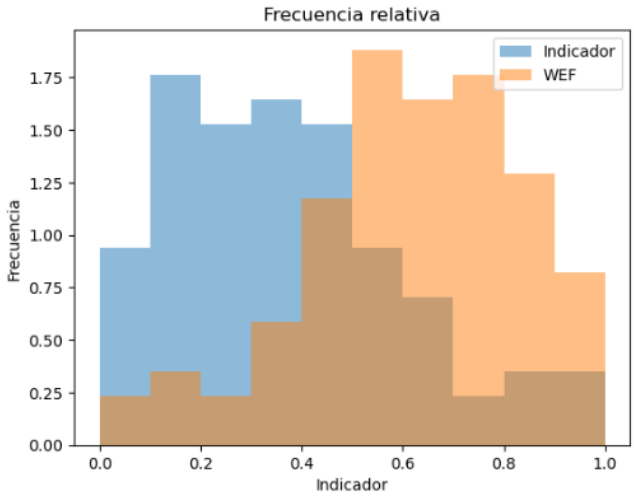

Figura 5: Histogramas de frencuencia relativa para el score propuesto por el banco mundial y para el indicador construido considerando todos los componentes principales.

Para finalizar, también se podría revisar la relación existente entre el indicador y el score. Esto con el objetivo de verificar si realmente el indicador tiene un comportamiento inversamente proporcional al score (relación lineal negativa) como indican los otros análisis. Para ello se calcula la matriz de correlación y se realizan algunas gráficas.

	Ind_2	Ind_4	Ind_6	Ind_8	Ind_10	Ind_todos	Indicador WBF norm
Ind_2	1.000000	0.998388	0.997661	0.997497	0.997442	0.997413	-0.956493
Ind_4	0.998388	1.000000	0.999271	0.999108	0.999052	0.999024	-0.959280
Ind_6	0.997661	0.999271	1.000000	0.999836	0.999781	0.999752	-0.960019
Ind_8	0.997497	0.999108	0.999836	1.000000	0.999945	0.999916	-0.960100
Ind_10	0.997442	0.999052	0.999781	0.999945	1.000000	0.999971	-0.960552
Ind_todos	0.997413	0.999024	0.999752	0.999916	0.999971	1.000000	-0.960381
Indicador WBF norm	-0.956493	-0.959280	-0.960019	-0.960100	-0.960552	-0.960381	1.000000

Figura 6: Matriz de correlación lineal para los indicadores construidos a partir de los componentes principales y el score.

En este caso se observa que, independientemente de la cantidad de componentes tomados, el coeficiente de correlación entre el score y el indicador es mayor a 0.95, además de que en todos los casos es negativo. Esto indica una correlación lineal fuerte entre las variables, además de que indica que mientras una incrementa su valor, la otra lo va a disminuir. Adicionalmente, se observa que el coeficiente de correlación aumenta conforme incrementa la cantidad de componentes incluidos. La matriz de correlación permite también concluir con respecto a la relación entre los indicadores variando los componentes. Se observa que tienen una relación lineal directa, lo que tiene sentido por la construcción del indicador (linealidad).

De las gráficas se puede concluir que efectivamente existe una relación lineal inversa. Además, se observa que en el caso con dos componentes se tiene una mayor dispersión de los datos, de modo que se disminuye el coeficiente de correlación.

Parte 2 - Kmeans e histogramas

Para comenzar se deben importar los datos del excel nombrado 'datos_emisiones.xlsx' utilizando *read_excel* de la librería *pandas*. Se tiene un arreglo de datos con 300 registros para emisiones de CO_2 y NO_x. De una revisión de la base de

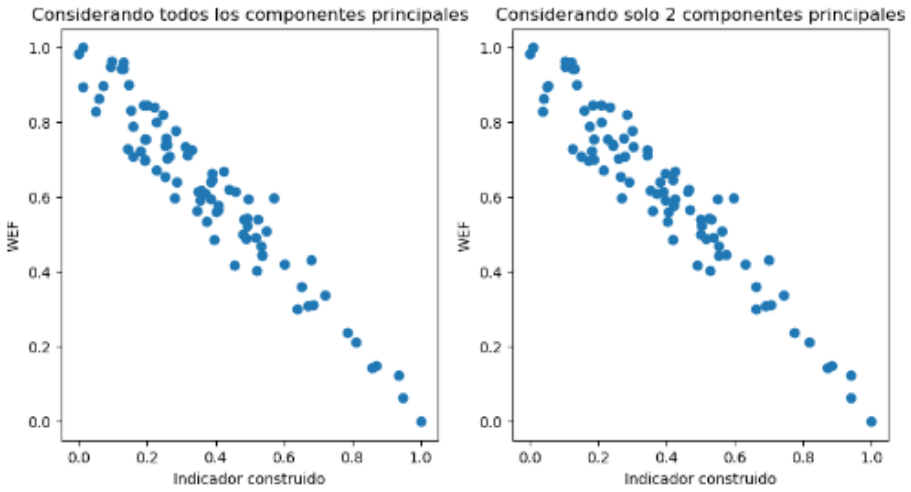

Figura 7: Gráficas Score vs indicador.

datos se identifica que esta no cuenta con datos incompletos por lo que no se requiere realizar un preprocesamiento de los datos. A continuación se debe realizar la normalización de manera que se tengan variables sin unidades y que conserven las posiciones relativas. Para ello se usa la siguiente expresión:

$$Z_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Para continuar, el algoritmo de k-means se ejecuta al instanciar la función, definirle la cantidad de clusters y pasarle los datos a clasificar. Como se quiere clasificar en 3 categorías, en el algoritmo de clusterización se define la cantidad de clusters como 3, mientras que los datos de ingreso son los datos normalizados de las emisiones. Vale la pena mencionar que k-means es un algoritmo de aprendizaje no supervisado para clasificar en k categorías basándose en la similitud entre datos, generalmente basándose en la distancia a los clusters. Una vez entrenado el algoritmo de clusterización, se le pasan nuevamente los datos de emisiones para que realice la predicción del cluster al que pertenece cada país. Posteriormente, se almacenan los index de cada categoría. Con los index almacenados se pueden graficar cada una de las fuentes con un color diferente.

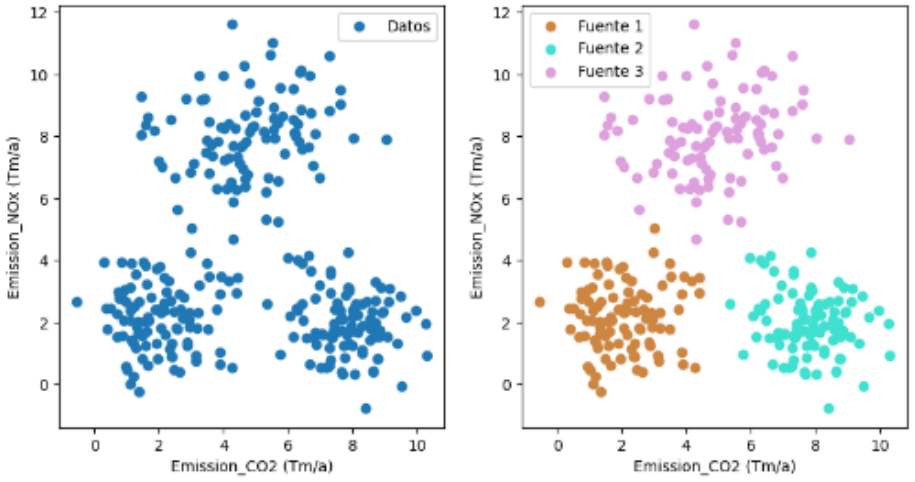

Figura 8: A la izquierda datos originales, a la derecha clasificación construida con k-means.

De la clasificación realizada se pueden sacar algunas conclusiones, la primera es que el algoritmo es muy útil para identificar fuentes diferentes ya que realiza una clasificación lógica para los datos ingresados, además de que determina una clasificación para los datos más problemáticos de clasificar manualmente. Sin embargo, la clasificación podría variar

Medida	F2	F2 sin -	F3	F3 sin -	F1	F1 sin -
Máximo	1276.34	1276.34	3466.61	3466.61	1507.02	1507.02
Promedio	588.69	602.66	2401.30	2401.30	652.83	659.94
Mínimo	-221.83	104.24	1404.25	1404.25	-65.22	6.83

Cuadro 2: Emisiones máxima mínima y promedio para cada fuente considerando y sin considerar los negativos.

dependiendo del algoritmo de clusterización utilizado por lo que trabajar con esta solución tiene implícito un sesgo. Los resultados muestran que 102 observaciones son de la fuente 1, 103 de la fuente 2 y 95 de la fuente 3. Por lo tanto, aproximadamente 1/3 de los datos provienen de cada fuente. Sobre la clusterización se pueden realizar algunas conclusiones preliminares: se podría decir que la fuente 3 sea probablemente la más contaminane ya que presenta las mayores emisiones de NO_x, mientras que las fuentes 1 y 2 serán diferentes entre sí ya que tienen marcadas diferencias en las emisiones de CO_2.

Sobre los datos de emisiones, y teniendo en cuenta los resultados de la clusterización se puede tratar de identificar la fuente más contaminante. Para ello se debe generar una sola variable para las emisiones utilizando algún factor de conversión para convertir a emisiones de CO_2 equivalentes:

$$Emisiones_{eq} = Emisiones_{CO_2} + Emisiones_{NO_x} * F_{eq}$$

En este caso el factor de conversión utilizado para pasar de emisiones de NO_x a CO_{2eq} es 298. En este caso se usan los index para generar dataframes separados para cada una de las fuentes, y sobre ellos se calcula el máximo, el mínimo y el valor promedio. Vale la pena mencionar que, al revisar los dataframes se identifica que algunas fuentes tienen valores negativos de emisiones, lo que se puede considerar como datos errados o ocasiones en las que la fuente absorbió emisiones. No se considera adecuado eliminar directamente los datos negativos ya que no se conoce la historia de los datos, por lo tanto se decide realizar el análisis incluyendo los datos negativos y sin incluirlos.

Considerando o sin considerar los valores de emisiones equivalentes negativos la fuente más contaminante es la fuente 3. Evidentemente estos valores corresponden a la fuente marcada en rosa en la figura previmanete presentada. Claramente la fuente 3 debe ser una industria mucho más contaminate que las otras dos, esto atendiendo a que es la que tiene más emisiones de NO_x, emisiones con un factor de equivalencia muy significativo. De hecho, el promedio de la fuente 3 es 4 veces superior al de las otras dos fuentes. Por otra parte, las otras dos fuentes tienen promedios similares, aunque la fuente 1 resulta ser la segunda más contaminante, seguida de la fuente 2.

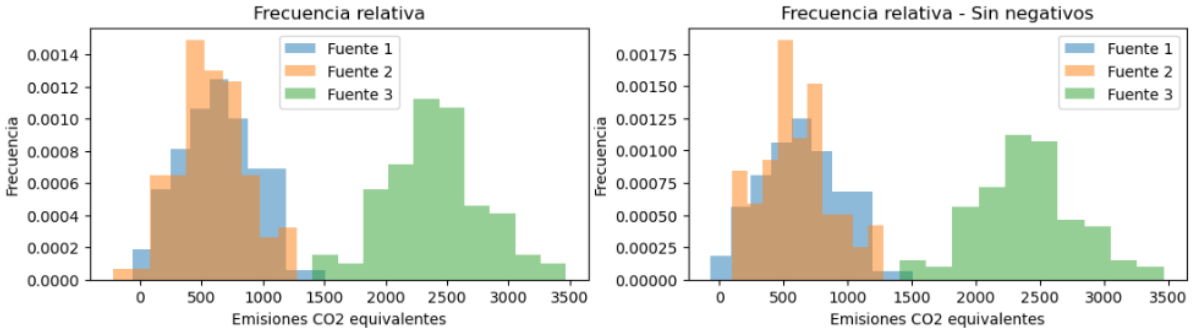

Figura 9: Histogramas de frecuencia acumulada para las fuentes de emisiones.

Para finalizar con el análisis de las fuentes se puede construir el histograma de frencuencia relativa para cada una. Estos histogramas confirman que la fuente 3 es la más contaminante, mientras que las fuentes 1 y 2 son similares entre sí en la cantidad de emisiones. Los histogramas también permiten ver que la fuente 3 se encuentra distribuida en un rango más amplio. Para obtener unos parámetros más exactos con respecto a la distribución de los datos se puede ajustar una distribución de densidad de probabilidad de tipo normal a los histogramas normalizados. Este tipo de distribución es adecuada ya que, por teorema de límite central, las muestras de datos provenientes de variables aleatorias que cuentan con muchos datos son bien descritas por distribuciones normales. La ecuación para la distribución normal es la siguiente:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

en la cual se deben ajustar los parámetros μ y σ que representan la media y la desviación estándar. Para realizar este ajuste sobre los histogramas se crea una función la cual recibe los datos y usa la función *norm* del módulo *stats* para realizar un fit sobre los datos y estimar los parámetros de la distribución. En este la distribución se ajusta a los histogramas incluyendo valores negativos ya que esto no cambia el ordenamiento de las fuentes con respecto a cual es más contaminante.

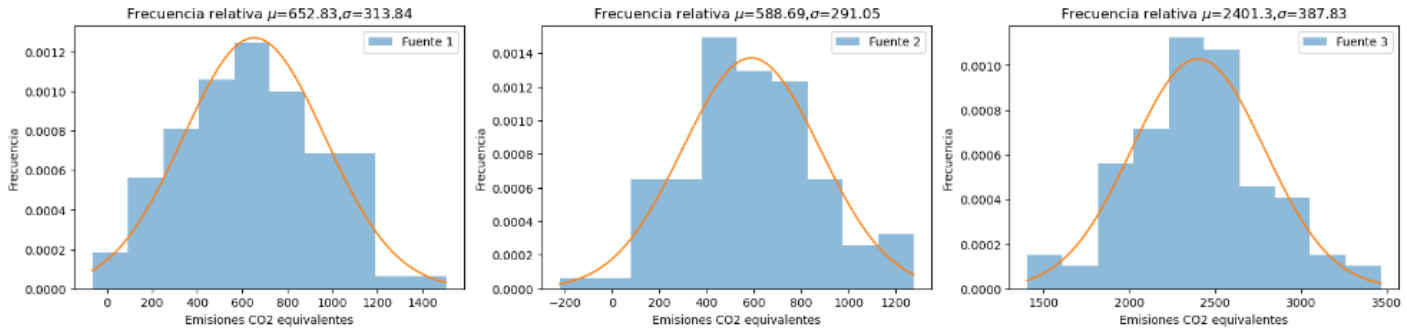

Figura 10: Histogramas ajustados a una distribución normal para cada fuente. En el título se presentan los parámetros de la distribución.

El resultado del ajuste muestra que la distribución normal ajusta bien el histograma (como era de esperarse), además de que uno de los parámetros de la distribución ya había sido calculado previamente ya que para la distribución normal μ es el promedio. Otra variable que se estima con la distribución es la desviación estándar. Como se había mencionado previamente, la fuente 3 tiene datos distribuidos en un rango más amplio lo que se traduce en una desviación estándar mayor a la de las otras fuentes. Sin embargo, la diferencia en esta variable no es tan significativa como en el promedio.

Parte 3 - Regresión logística

Para comenzar se deben importar los datos del excel 'datos_emisiones_rl_ent.xlsx'. En este archivo se cuenta don datos de emisiones de CO_2 y NO_x, además de la información de la fuente. Para realizar el entrenamiento de la regresión logística se deben definir la variables X que entrenan la regresión y la variable Y que contiene la clasificación correcta. En este caso las X van a ser las emisiones (por lo que se tienen dos variables), y la Y será la información de la fuente. A continuación se deben separar los datos utilizando la función *train_test_split* para tener un set de entrenamiento y un set de prueba. En este caso se utiliza el 80 % de los datos para entrenamiento al ajustar el parámetro 'test_size = 0.2'. Para realizar la regresión nuevamente se toman el 80 % de los datos para entrenamiento y el restante paratestear. Para entrenar la red se crea la instancia del modelo de regresión logística usando la función *LogisticRegression()* y se realiza un fit sobre los datos de entrenamiento. Adicionalmente, usando la matriz de confusión se calcula el accuracy. Luego con el modelo entrenado se realiza la predicción sobre los datos para testear y se calcula nuevamente el accuracy.

```
The model performance for training set
Accuracy of 0.95
--------------------------------------
The model performance for testing set
Accuracy of 0.9
--------------------------------------
Rendimiento del modelo
              precision    recall  f1-score   support

           0       0.92      0.92      0.92        36
           1       0.88      0.88      0.88        24

    accuracy                           0.90        60
   macro avg       0.90      0.90      0.90        60
weighted avg       0.90      0.90      0.90        60
```

Figura 11: Resultados para la regresión logística.

En lo que respecta al accuracy, el modelo en entrenamiento obtuvo 0.95, mientras que sobre los datos para testear obtuvo 0.9. Teniendo en cuenta que el accuracy se define como las predicciones acertadas sobre todas las predicciones, el resultado obtenido para el modelo generado es aceptable. Evidentemente al usar datos nuevos el accuracy va a disminuir con respecto a los datos de entrenamiento. Para continuar, sería interesante analizar por separadas las dos clases. Para ello se utilizarán las métricas: precisión, recall y el f1-score. La presición se refiere a cuantas predicciones para una clase resultan ser correctas. En el caso de la fuente 1 (clase 0), el 92 % de las predicciones hechas son correctas; mientras que para la fuente 2 (clase 1), el 88 % de las predicciones fueron correctas. Por otra parte, el recall se refiere a la cantidad de casos reales identificados para cada clase. En este caso para la fuente 1 la regresión logística identificó correctamente el 92 % de los casos reales; mientras que para la fuente 2 identificó el 88 % de los casos reales. El recall indica que el modelo generado es bueno ya que supera el 50 % para ambas clases. Finalmente, f1-score trata de equilibrar la precisión con el recall. En este caso se obtienen valores de 0.92 y 0.88 para la fuente 1 y 2 respectivamente, lo que indica que hay un buen equilibrio entre la precisión y el recall (entre los falsos positivos y los falsos negativos).

Para entender mejor los resultados de la regresión logística se pueden señalar los datos provenientes de cada fuente de acuerdo con la base de datos original, y trazar la línea de decisión que genera la regresión logística. Para ello es necesario recuperar los coeficientes que entrega la regresión. La ecuación de la recta generada por la regresión logística es:

$$m_1 * X_{CO_2} + m_2 X_{NO_x} + b = 0$$

sin embargo lo que se quiere es tener una variable en función dela otra para graficar en un plano cartesiado, por lo que, si se selecciona como variable dependiente X_{NO_x}, la ecuación de la recta, construida a partir de los coeficientes obtenidos, es:

$$X_{NO_x} = \frac{-m_1}{m_2} X_{CO_2} - \frac{b}{m_2}$$

De este modo, se llega a la siguiente gráfica:

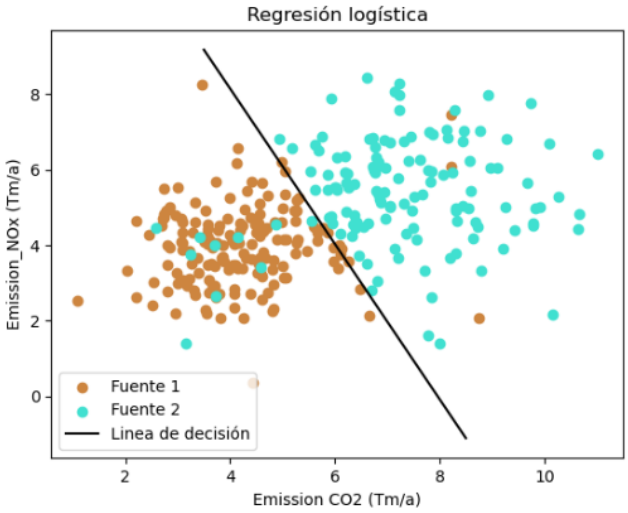

Figura 12: Emisiones de NO_x vs emisiones de CO_2. Incluye la recta de decisión generada por la regresión logística.

Evidentemente el rendimiento del modelo no es perfecto ya que se identifican varios puntos de la otra fuente que se encuentran en el lado incorrecto de la recta de decisión. En particular, se identifican más puntos pertenecientes a la fuente 2 en el lado correspondiente a la fuente 1, lo que explica por qué el rendimiento del modelo para clasificar la fuente 2 es menor que para la fuente 1. Sin emabrgo, la recta permite generar una clasificación correcta y adecuada en la mayoría de los casos, por lo que se considera un buen modelo de clasificación.

Finalmente, para testear el modelo de regresión, se tienen datos nuevos provenientes del excel 'datos_emisiones_rl_nuevos.xlsx' (50 datos). Estos se ingresan al modelo y se calcula la matriz de confusión y las métricas.

$$\text{Matriz de confusión} = \begin{bmatrix} 24 & 3 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} T_n & F_n \\ F_p & T_p \end{bmatrix}$$

Los resultados generales del modelo disminuyeron de forma significativa. Para comenzar el accuracy pasó de 0.9 a 0.72, lo que indica que el 72 % de las predicciones resultaron ser correctas para el conjunto de datos nuevos. Este cambio se

	precision	recall	f1-score	support
0	0.69	0.89	0.77	27
1	0.80	0.52	0.63	23
accuracy			0.72	50
macro avg	0.74	0.71	0.70	50
weighted avg	0.74	0.72	0.71	50

Figura 13: Resultados de la regresión logística para los datos nuevos ingresados.

evidencia en la matriz de confusión ya que se tienen 11 falsos positivos y 3 falsos negativos, es decir, se tienen 14 errores. Lo anterior influencia significativamente la precisión, sobre todo para la fuente 1 (0.69) que resulta ser menos precisa que la fuente 2 (0.8). Adicionalmente, se ve también afectado el recall, aunque en este caso la fuente 1 tiene un mejor resultado que la fuente 2. Para la fuente 1 se identificaron el 89 % de los pertenecientes, mientras que para la fuente 2 solo se identificaron el 52 %. Sin embargo, como ambos valores son mayores a 0.5 se consideran buenos.