Análisis multivariado de emisión de CO2 de modelos de IA

Juan Felipe Gallo Rendón* *Facultad de Ingeniería*
Universidad de Antioquia
Medellín, Colombia

***Abstract*—Este estudio aborda la creciente preocupación por la sostenibilidad ambiental en la inteligencia artificial, centrándose en la huella de carbono de los modelos de aprendizaje automático. A pesar de la popularidad de plataformas como Hugging Face, existe un conocimiento limitado sobre cómo los desarrolladores miden y reportan las emisiones generadas durante el entrenamiento de sus modelos, dificultando la adopción de prácticas de "Green AI". Para analizar esta problemática, se emplearon dos técnicas de análisis multivariado: el Análisis de Componentes Principales (ACP) y el Análisis Factorial (AF). y a partir de estos datos se pretende encontrar alguna relación entre las variables reportadas y la emisión de gases de efecto invernadero**

***Index Terms*—Estadística multivariada, PCA, PFA, regresión, GreenAI.**

I. Introducción

La creciente relevancia de la inteligencia artificial (IA), particularmente en el ámbito de la IA Verde (GreenAI) [1], ha impulsado la necesidad de comprender y mitigar su impacto ambiental. Este análisis aborda la escasa información disponible en fuentes abiertas sobre este tema. Para ello, se ha utilizado la base de datos HCO2.csv, proveniente de un repositorio asociado a un estudio previo enfocado en la plataforma de Hugging Face [2]. Ese estudio inicial buscaba responder a la pregunta de investigación: ¿Cómo los usuarios de Hugging Face reportan su huella de carbono? Los hallazgos revelaron que la mayoría de los usuarios no documentan con detalle las emisiones de carbono de sus modelos. En ese contexto, incluso tecnologías emergentes y dominantes como GPT-2 no reportaban datos específicos sobre su entrenamiento y producción. A partir de esta base de datos, el presente trabajo se centra en un análisis multivariado para identificar las interrelaciones entre las variables clave. La metodología aplicada incluyó el cálculo de las matrices de varianza-covarianza y correlación. Posteriormente, se calcularon los valores y vectores propios de cada técnica para determinar la varianza explicada y los factores significativos. Los resultados fueron visualizados y comparados mediante biplots.

II. Metodología

El primer paso para ejecutar el análisis se centró en el entendimiento general de la base de datos: cuáles eran sus variables categóricas y su frecuencia, revisar los datos nulos y los que no podían ser analizados, posteriormente incluyó el cálculo de las matrices de varianza-covarianza y correlación para identificar las interrelaciones entre las variables clave del conjunto de datos. Después, se obtuvieron los valores y vectores propios de cada técnica para determinar la varianza explicada y los factores significativos. Finalmente, los resultados fueron visualizados a través de biplots, permitiendo una comparación gráfica de ambos métodos. Se usó la tecnología de *Jupyter Notebooks* con *Python 3.3*, se desarrolló una serie de scripts para cada fase del análisis, y unos cuantos scripts útiles que se guardaron en una carpeta **src**, el código se encuentra disponible en el siguiente enlace: https://github.com/jufegare000/hugging-face-co2-multivariate-analysis

A. Análisis de datos

La base de datos **HFCO2.csv** es el producto del trabajo expuesto en el estudio anteriormente mencionado, en el se encuentran las siguientes variables:

text (without indent)

1) `co2 eq emissions`: Reporte de emisiones de CO2 equivalentes
2) `downloads`: Cantidad total de descargas del modelo.
3) `likes`: El número de “me gusta” (calificación por usuario) del modelo
4) `training type`: El tipo de entrenamiento, si es preentrenado o fine tuning.
5) `geographical location`: El lugar donde fue entrenado.
6) `domain`: the Tipo de dominio del modelo
7) `size`: Tamaño del modelo, en KB.
8) `auto`: if the model is AutoTrained.
9) `source`: Cuál es la fuente de donde se reporta

las variables `modelId`, `datasets`, `co2 reported`, `createdat`, `libraryname`, ya que no aportan información de valor para el análisis. Las variables categóricas nominales son `source`, `geographical location`, `training type` y `domain`, para entender su comportamiento general, se muestran gráficas de frecuencia. Para las variables cuantitativas, se realiza un análisis para encontrar cuáles observaciones contienen datos en nulo, indeterminado o con un formato no adecuado.

B. Limpieza de datos

Después de revisar las particularidades asociadas al conjunto de datos, como datos nulos o incorrectos, se procede con la limpieza de datos, para lograr esto, se realizó una imputación

de basada en regresión, luego se retiraron aquellos datos que presentaban problemas en el formato y finalmente, se realizó una estandarización de datos para facilitar la visualización de los datos en cada etapa del proceso.

C. Análisis de varianza y covarianza

Se usó la librería de scikitlearn para escalar y ver el comportamiento de los datos a nivel de variabilidad, el detalle de la implementación puede verse en la ruta del pruecto `notebooks/2_variance_covariance.ipynb`.

D. Análisis de componentes principales

El análisis de componentes principales se realizó usando las librerías `sklearn.decomposition`, `sklearn.preprocessing`, adicionalmente se ejecutaron algunas operaciones sobre las matrices resultantes para revelar información de las matrices de eigenvalues y eigenvectors, también se usó gráficas de dispersión y biplots.

E. Análisis de factores principales

El análisis de factores principales se realizó usando la librería `factor_analyzer`, adicionalmente se ejecutaron algunas operaciones sobre las matrices resultantes para revelar información de las matrices de eigenvalues y eigenvectors, también se usó gráficas de dispersión y biplots.

F. Análisis biplots

Con base la información revelada en las gráficas de biplots, se hizo una comparativa sobre cuál podría explicar mejor la variabilidad de los datos y cuál se ajustaba más al conjunto de datos

III. Resultados y discusión

Después de El proceso del análisis se divide principalmente en los anteriormente mencionados en la metodología, cada paso del análisis tiene asociado uno o más scripts de python, el proyecto se divide en 3 carpetas principales: *assets*, *notebooks*, y *src*, en ellos se encuentran recursos de código y base de datos usados para el análisis. `assets` contiene la base de datos *HCO2.csv* y en ella se encuentra la versión original de los datos recolectados en el estudio de hugging face [2], en este directorio, se almacenará de igual forma: las gráficas, los csvs procesados, y estructuras de datos auxiliares para los análisis. Por otro lado, *notebooks* contiene los archivos de los jupyter notebooks que corresponden a cada de análisis, para legibilidad, se tiene un notebook por cada fase del análisis. Finalmente *src* contiene scripts comunes para apoyar cada fase del análisis, y son importados en cada notebook.

A. Análisis general de los datos

Tras entender de qué se compuso la base de datos extraida del estudio de hugging face, lo siguiente es encontrar la información que brindan estos datos. Para ello, se realizó una gráfica de frecuencias para las variavles categóricas: `source`, `geographical location`, `training_type` y `domain`, se realiza entonces una gráfica de frecuencias para cada variable categórica. La figura 1 muestra la tabla de frecuencias de la variable `source`

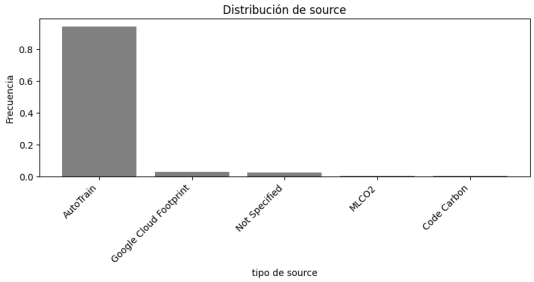

Fig. 1: Gráfico de frecuencia para `source`

La gráfica revela que hay un desbalance respecto a la fuente de donde se reporta la emisión de CO2 del modelo, en este caso del autotrain, y es algo consecuente, ya que hugging face ofrece la opción de autoentrenar el modelo una vez se ha subido a la plataforma. La proporción por cada uno de los tipos de reporte se puede ver la gráfica 2:

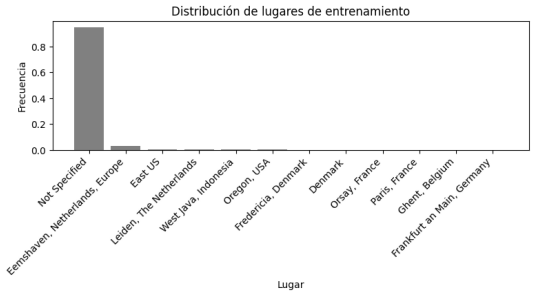

Fig. 2: Gráfico de frecuencia para `geographical_location`

Para la zona geográfica, se puede evidenciar que se la mayoría de modelos no reporta desde dónde se hizo el entrenamiento. Esta es una variable importante para medir la huella de carbono, por lo que no tener este dato no permitiría saber con certeza la fuente de energía que alimenta al datacenter donde se despliega el modelo.

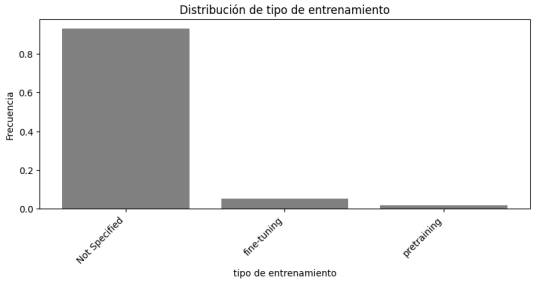

Fig. 3: Gráfico de frecuencia para tipo de entrenamiento

En la gráfica 3, muestra cuál fue el tipo de entrenamiento del modelo, la plataforma de hugging face, permite subir modelos entrenados y no entrenados, se puede ver que la muestra usada para los modelos carece de información, y

por otro lado la se puede evidenciar que la proporción de tipos de entrenamiento sopesa aquellos que se les hizo fine tunning, donde a los modelos se les aplica un ajuste de hiper parámetros, mientras que en un pretraining, se puede cambiar los datos de entrenamiento para obtener resultados diferentes. La gráfica 4 muestra los tipos de dominios de los modelos, se puede evidenciar que las muestras se centran en NLP *(Natural language processing)*

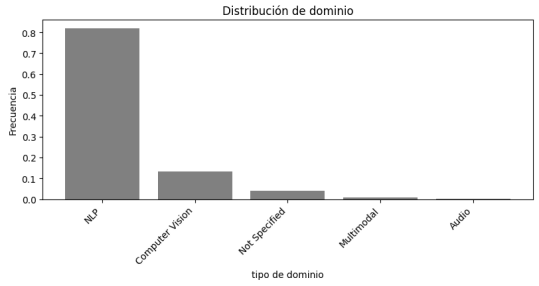

Fig. 4: Gráfico de frecuencia para tipo de dominio

Finalmente, verificó cuáles columnas tenían datos en nulo, cuántos datos correspondían a true o false para la variable *auto* y cuáles tenían valores no legibles por la herramienta, el resultado se presenta en las tablas I, II y III:

size	0.079634
size efficency	0.079634

TABLE I: variables con su respectivo porcentaje de nulos

True	0.909796
False	0.085976
1	0.004228

TABLE II: Cantidad de verdaderos y falsos en la variable auto

size_efficency	0.070472

TABLE III: proporción de valores infinitos de size_eficiency

Para ver el detalle de los resultados de cómo se implementó, en el proyecto se encuentran los scripts en la ruta `"notebooks/1_1_data_analysis.ipynb`

B. Limpieza de datos

En el anterior apartado se realizó un análisis general de los resultados y se encontró que las variables `size` y `size_efficiency` poseen algunos datos nulos, cerca del 8%, por lo tanto, es necesario aplicar una regresión, en la ruta del proyecto `notebooks/1_2_data_cleaning.ipynb` se encuentra la implementación de la regresión y la limpieza de datos, se escogió la regresión, porque es la técnica que menos sesgos podría generar en el conjunto de datos, a diferencia de las otras formas de imputación. Luego de la imputación de variables, se limpiaron los valores infinitos, luego se realizó una estandarización, para realizar el análisis de varianza y de covarianza, en este punto se excluyen las variables categóricas

C. Análisis de covarianza y correlación

Para realizar el análisis de covarianza y correlación, lo primero es estandarizar los datos usando `StandardScaler` para presentarlos en la gráfica de manera más amigable, dado que las variables tienen una dispersión muy alta, a continuación se muestra la gráfica de correlación, ya que la de covarianza no permite ver las relaciones de manera tan clara

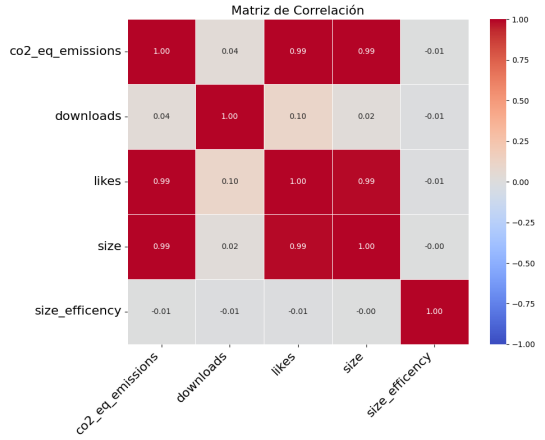

Fig. 5: Gráfico de correlación

Para interpretar las correlaciones, se toma de manera univariada las relaciones entre las variables del conjunto de datos

1) `co2_eq_emissions` vs. `likes`: Con una correlación de 0.99, esta es una correlación positiva extremadamente fuerte. Sugiere que los modelos de IA que tienen un mayor número de likes (`likes`) también tienden a tener una emisión de CO2 equivalente mucho más alta.

2) `co2_eq_emissions` vs. `size`: Con una correlación de 0.99, similar al caso anterior, existe una correlación positiva extremadamente fuerte. Esto implica que los modelos de IA más grandes en términos de tamaño (`size`) son los que emiten una mayor cantidad de CO2.

3) `co2_eq_emissions` vs. `downloads`: Con una correlación de 0.04, La correlación es prácticamente cero (muy débil). Esto indica que no hay una relación lineal significativa entre el número de descargas (`downloads`) y la cantidad de emisiones de CO2. Un modelo puede tener muchas descargas independientemente de su huella de carbono.

4) `co2_eq_emissions` vs. `size_efficency`: Con una correlación de -0.01, La correlación es también cercana a cero. Esto sugiere que la eficiencia del tamaño (`size_efficency`) no está linealmente relacionada con las emisiones de CO2. Es posible que esta variable no esté capturando la eficiencia en términos de emisiones, o que la relación sea no lineal.

D. Análisis de componentes principales (PCA) y factores principales (PFA)

Para este apartado, se excluyeron las variables cualitativas, y se conservó de las ejecuciones anteriores la base de datos de imputados, limpios y estandarizados, luego de estandarizar los datos, se le aplicó la función `fit_transform` los detalles de la implementación se presentan en el notebook que se

encuentra en la ruta: `3_analisis_factores.ipynb` del proyecto.

1) Análisis de componentes principales: El primer resultado que se encuentra es acerca de la varianza explicada, los resultados se encuentran consolidados en la siguiente tabla:

TABLE IV: Resultados del Análisis de Componentes Principales

Componente	Valor propio	Varianza Explicada	Varianza Acumulada
PC1	2.989	59.74%	59.74%
PC2	1.012	20.23%	79.97%
PC3	0.988	19.75%	99.72%
PC4	0.010	0.21%	99.93%
PC5	0.003	0.07%	100.00%

2) Interpretación de los resultados:

- **Componente 1**: Con un valor propio de `2.989`, este componente es el más importante. Explica casi el 60% de la varianza total de los datos por sí solo. Esto significa que la mayor parte de la información está concentrada en esta primera dimensión.
- **Componente 2**: Explica un `20.23%` de la varianza. Al combinarlo con el Componente 1, ambos componentes juntos explican casi el `80%` de la varianza acumulada.
- **Componente 3**: Este componente explica casi un 20% de la varianza. Al incluirlo, la varianza acumulada asciende al `99.72%`.
- **Componentes 4 y 5**: Estos componentes tienen valores propios muy pequeños y explican una varianza insignificante (menos del 1% entre ambos).

3) Selección de componentes: Para decidir cuántos componentes conservar, se aplicó el criterio de Kaiser, que sugiere mantener los componentes con un valor propio mayor a 1. Según este criterio, se deberías considerar conservar solo los dos primeros componentes principales, ya que sus valores propios son `2.989` y `1.012`. Estos dos componentes juntos resumen el `79.97%` de la varianza de tus datos, conservarlos permite reducir la dimensionalidad del conjunto de datos de cinco variables a solo dos, perdiendo muy poca información esencial. A continuación, se presenta la matriz de cargas factoriales para los componentes principales. Cada valor indica la correlación de la variable original con el componente correspondiente.

TABLE V: Resultados del Análisis de Componentes Principales

Variable Original	Carga en PC1	Carga en PC2
`co2_eq_emissions`	0.577	0.047
`downloads`	0.033	-0.695
`likes`	-0.031	0.715
`size`	-0.126	0.052
`size_efficency`	0.805	0.031

4) Interpretación de los Componentes Principales Clave:

- **Componente 1 (PC1)**: Este componente, que explica la mayor parte de la varianza del conjunto de datos, está fuertemente asociado con las variables 4 y 0. Ambas contribuyen de manera positiva a este componente, sugiriendo que el PC1 captura un factor común de estas dos variables.
- **Componente 2 (PC2)**: Este componente representa un contraste entre la variable 2 y la variable 1. La alta carga positiva de la variable 2 y la alta carga negativa de la variable 1 indican que el PC2 describe una relación inversamente proporcional entre ellas."

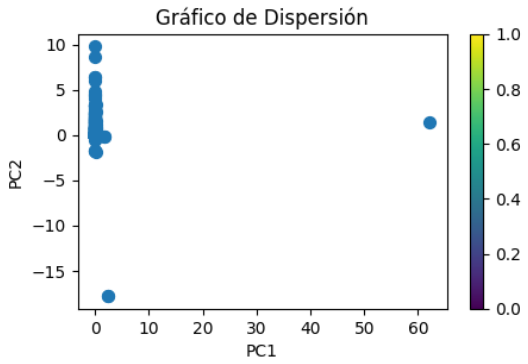

Fig. 6: Gráfico de dispersión de componentes

- **Varianza en el PC1**: La mayoría de los puntos se agrupan cerca de la coordenada 0 en el eje horizontal, con la notable excepción de un punto que se encuentra muy alejado, alrededor de PC1 = 60. Este punto atípico (o outlier) es una observación con un valor extremadamente alto en el primer componente principal. La gran dispersión a lo largo del eje horizontal refuerza lo que vimos en el análisis de los valores propios: el PC1 captura la mayor parte de la varianza en tus datos.
- **Varianza en el PC2**: En el eje vertical (PC2), los puntos también están bastante concentrados cerca de la coordenada 0, aunque hay una mayor dispersión vertical en este grupo principal. Hay un par de puntos que se alejan del grupo, pero la variabilidad es considerablemente menor en comparación con el PC1. Esto confirma que el PC2 explica menos varianza que el PC1.
- **Presencia de un Outlier**: La observación aislada en la esquina inferior derecha es un caso extremo. Este punto tiene un valor muy alto en el PC1 y un valor moderadamente bajo en el PC2. Es probable que esta observación represente un modelo de IA con características extremas en las variables originales que más contribuyen al PC1 (por ejemplo, un tamaño o una emisión de CO2 excepcionalmente grandes).

5) Análsis de factores principales (PFA): Para el Análisis Factorial de Principales Componentes (PFA), los números que ha obtenido son los valores propios (eigenvalues) de los factores. A diferencia del PCA, donde los valores propios explican la varianza total de los datos, en el PFA estos valores explican solo la varianza común (o compartida) entre las variables, lo cual es la varianza que puede atribuirse a factores subyacentes. La interpretación de los resultados se presenta a continuación en la siguiente tabla:

TABLE VI: Interpretación de Factores Principales

Factor	Valor propio	% de Varianza Común Explicada	% Acumulado
F1	2.987	74.49%	74.49%
F2	1.012	25.26%	99.75%
F3	0.988	0.22%	99.97%
F4	0.010	0.03%	100.00%
F5	0.003	0.00%	100.00%

6) Interpretación de factores (PFA):

- **Factor 1 (Valor Propio: 2.987)**: Este factor captura la mayor parte de la varianza compartida, explicando un notable 74.49% de ella por sí solo. Es, con diferencia, el factor más importante.
- **Factor 2 (Valor Propio: 1.012)**: Este factor también es muy significativo, ya que explica otro 25.26% de la varianza común. Al combinar el Factor 1 y el Factor 2, se explica casi el 99.75% de toda la variabilidad compartida de tu conjunto de datos.
- **Factores Restantes (3, 4 y 5)**: Los valores propios de estos factores son muy bajos, lo que indica que explican una cantidad insignificante de la varianza común

Para determinar cuantos factores retener, se usa el criterio de kaiser, el detalle de la implementación puede verse en la ruta: `notebooks/3_analisis_factores.ipynb`

7) Análisis de cargas factoriales: El Factor 1 es el más significativo, tal como lo vimos en el análisis de los valores propios. Las variables con las cargas más altas en este factor son:

- `co2_eq_emissions`: `0.999200`
- `likes`: `0.995705`
- `size`: `0.993753`

Estos valores son extremadamente altos, lo que significa que el Factor 1 es, en esencia, un resumen de estas tres variables. Todas ellas tienen una correlación positiva muy fuerte con el factor, lo que sugiere que los modelos con un gran tamaño y un alto número de likes tienden a tener una alta emisión de CO2. Este factor puede ser interpretado como un factor de "Complejidad y Popularidad del Modelo" que está directamente relacionado con la huella de carbono. Las variables `downloads` y `size_efficency` tienen cargas muy bajas en el Factor 1, lo que indica que no contribuyen significativamente a este factor latente. El Factor 2 explica la varianza común restante. La variable con la carga más alta en este factor es `downloads`: `0.718192`. Esto sugiere que el Factor 2 está definido principalmente por el número de descargas. Las otras variables tienen cargas muy bajas, lo que indica que no están fuertemente correlacionadas con este factor. Por lo tanto, el Factor 2 puede ser interpretado como un factor de "Alcance o Adopción del Modelo", que es conceptualmente distinto del primer factor. Los resultados confirman que las variables de emisiones de CO2, likes y tamaño están altamente correlacionadas y se agrupan en el mismo constructo subyacente, mientras que las descargas representan un concepto separado el conjunto datos.

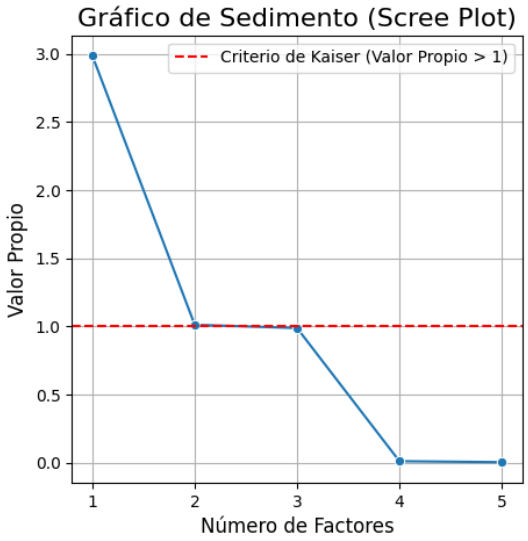

Fig. 7: Gráfico de sedimento

El gráfico de sedimento confirma visualmente lo que los valores numéricos indicaban: los Factores 1 y 2 son los más significativos y los únicos que vale la pena conservar para el análisis. Los factores 3, 4 y 5 tienen valores propios muy bajos, lo que significa que explican una cantidad trivial de varianza y pueden ser descartados. A continuación se presentan los biplots de los métodos PCA y PFA

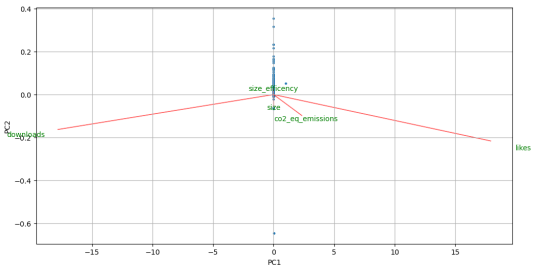

Fig. 8: Biplots PCA

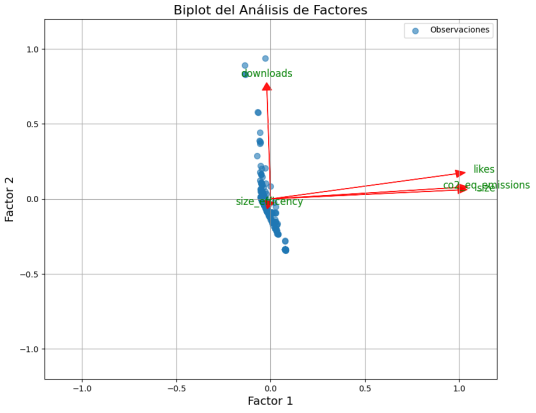

Fig. 9: Biplots PFA

8) Interpretación de Biplots: **Para el biplot de factores principales**: **Factor1 popularidad y tamaño:** Este factor,

representado por el eje horizontal, está fuertemente correlacionado con las variables likes, `co2_emissions` y `size`. Las aplicaciones con valores altos en este factor tienden a ser más populares, con una mayor cantidad de 'me gusta', un tamaño de archivo más grande y, posiblemente, un mayor impacto ambiental. **Factor 2** Capacidad de Descarga: este factor, representado por el eje vertical, se relaciona principalmente con la variable downloads. Indica que las aplicaciones con altas puntuaciones en este factor son las que tienen un alto número de descargas, independientemente de su popularidad o tamaño. La mayoría de las observaciones se agrupan en el centro, lo que sugiere que la mayoría de las aplicaciones tienen características promedio. Sin embargo, se pueden identificar dos grupos distintos de aplicaciones:

- Las que se caracterizan por una alta popularidad y tamaño (Factor 1).
- Las que se distinguen por un alto número de descargas (Factor 2).

La variable size_efficiency muestra una correlación muy baja con ambos factores, lo que indica que no es un buen predictor de la popularidad, el tamaño o las descargas de una aplicación. En cuanto al biplot de análisis de componentes principales. El PC1, que explica la mayor parte de la varianza en los datos, establece un contraste directo entre likes y downloads. Esto indica una fuerte correlación negativa: las aplicaciones con un alto número de "me gusta" tienden a tener pocas descargas, mientras que aquellas con muchas descargas reciben menos "me gusta". Esto podría sugerir dos estrategias distintas de éxito. Además, el PC2 no es un factor dominante, ya que las variables size, size_efficiency y co2_eq_emissions tienen una baja correlación con él. Esto significa que estas variables no aportan valor para explicar la principal variabilidad del conjunto de datos. y En síntesis: El biplot revela que la varianza de los datos se explica principalmente por una sola dimensión: el equilibrio entre la popularidad social y el volumen de descargas. Las otras variables tienen un impacto mínimo en la forma en que tus datos se agrupan.

IV. Conclusión

El análisis comparativo de las técnicas multivariadas aplicadas al conjunto de datos, específicamente el Análisis de Componentes Principales (PCA) y el Análisis de Factores Principales (PFA), ha revelado información crucial sobre la estructura subyacente de las variables.

A. Justificación del Método Seleccionado

El Análisis de Factores Principales (PFA) se identifica como el método más apropiado para este estudio. Aunque el PCA logró identificar una dimensión de varianza predominante, su principal objetivo es la reducción de dimensionalidad y no la interpretación de constructos latentes. En contraste, el PFA, diseñado para tal fin, proporcionó un modelo más robusto y conceptualmente interpretable.

B. Hallazgos encontrados

El modelo de PFA extrajo satisfactoriamente dos factores latentes que explican las interrelaciones observadas entre las variables. Estos factores se interpretan de la siguiente manera:

Factor 1: Popularidad y Escala. Este factor se correlaciona fuertemente con las variables likes, co2_emissions y size. Dicho factor representa un constructo de éxito que agrupa atributos de popularidad y tamaño.

Factor 2: Capacidad de Adquisición. Este factor está casi exclusivamente definido por la variable downloads. Su independencia del Factor 1 sugiere que la capacidad de una aplicación para ser descargada es una dimensión distinta y no redundante del éxito.

C. Implicaciones

Los resultados demuestran que el comportamiento de las aplicaciones no se puede explicar adecuadamente por una sola dimensión, sino que está influenciado por al menos dos constructos subyacentes. La identificación de estos factores permite una comprensión más profunda de la dinámica del mercado y proporciona una base sólida para la toma de decisiones estratégicas, orientadas a optimizar el rendimiento en cada una de estas dimensiones independientes.

D. Conclusión final

Aunque el conjunto de datos es funcional para un análisis exploratorio, su limitada granularidad y la ausencia de estandarización en las métricas de evaluación impiden un análisis riguroso de la huella de carbono. La variabilidad en el hardware utilizado, la ubicación geográfica de los centros de datos y la falta de consistencia en los informes de consumo energético y emisiones de CO2 representan obstáculos significativos para identificar con certeza las variables más influyentes. Estos factores comprometen la capacidad de extraer conclusiones definitivas sobre la relación entre las configuraciones de los modelos de inteligencia artificial y su impacto ambiental. Por lo tanto, se enfatiza la necesidad de futuros estudios que se beneficien de conjuntos de datos más detallados, actualizados y estandarizados. Un enfoque más riguroso en la recopilación de datos permitirá una comprensión más profunda del impacto ambiental de la IA y facilitará el desarrollo de prácticas más sostenibles en la industria.

References

[1] UST, "What is green ai?" https://www.ust.com/en/ust-explainers/what-is-green-ai, 2023, accessed: 11 de septiembre de 2025.

[2] ——, "Exploring the carbon footprint of hugging face's ml models: A repository mining study," https://arxiv.org/pdf/2305.11164, 2023, accessed: 11 de septiembre de 2025.