MODELADO DE EMISIONES DE CO_2 PER CÁPITA A PARTIR DE FACTORES SOCIOECONÓMICOS CON TÉCNICAS DE CIENCIA DE DATOS: UNA APROXIMACIÓN A LA DESIGUALDAD CLIMÁTICA EN AMÉRICA LATINA

Angie Catherine Collazos Valenzuela
Sandra Patricia Guzmán Diaz
Jhonatan Luis Merchán Burgos

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director:
Leonairo Pencue Fierro

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "MODELADO DE EMISIONES DE CO_2 PER CÁPITA A PARTIR DE FACTORES SOCIOECONÓMICOS CON TÉCNICAS DE CIENCIA DE DATOS: UNA APROXIMACIÓN A LA DESIGUALDAD CLIMÁTICA EN AMÉRICA LATINA"

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO:
4. ESTUDIANTE (S): Angie Catherine Collazos Valenzuela, Sandra Patricia Guzmán Diaz y Jhonatan Luis Merchán Burgos
5. CORREO ELECTRÓNICO: accollazos18@javerianacali.edu.co - sandraguzman@javerianacali.edu.co – jhonatanmerchan@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Cll 44ª 1j-57 Florencia Caquetá, cel. 3005246455 - calle 17c num 135-51 int 6 apto 844 Bogotá, cel. 3143625628 - Carrera 9#60-69 Bogotá, cel. 3142545836
7. DIRECTOR: Leonairo Pencue-Fierro
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Cátedra
9. CORREO ELECTRÓNICO DEL DIRECTOR: edgar.pencue@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica):
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
12. OTROS GRUPOS O EMPRESAS:
13. PALABRAS CLAVE (al menos 5): Ciencia de datos, Emisiones de CO_2 per cápita, Desigualdad climática, Factores socioeconómicos, Aprendizaje automático.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): 12. Producción y consumo responsables; 13. Acción por el clima
15. FECHA DE INICIO (Desarrollo del proyecto): 9/06/2025
16. RESUMEN (máximo 400 palabras).

El presente proyecto aplicado desarrolla un modelo de ciencia de datos para analizar y predecir las emisiones de CO_2 per cápita en función de factores socioeconómicos en 17 países de América Latina, con el propósito de evidenciar patrones estructurales de desigualdad climática en la región. La investigación partió de la integración de múltiples fuentes públicas EDGAR, Banco Mundial y PNUD conformando un panel país-año de 527 observaciones para el período 1990–2020, con 15 variables socioeconómicas, energéticas y ambientales. Tras un proceso riguroso de depuración, estandarización e imputación por interpolación lineal validada, se realizó un análisis exploratorio que identificó al consumo energético, el PIB per cápita y el Índice de Desarrollo Humano como los principales correlatos de las emisiones per cápita.

Se implementó una estrategia de modelado dual: modelos econométricos de datos panel con efectos fijos y aleatorios, y cinco algoritmos de aprendizaje automático, Random Forest, XGBoost, KNN, MLP y Support Vector Regression, optimizados mediante GridSearchCV con validación cruzada de cinco pliegues. Entre los algoritmos evaluados,

el modelo K-Nearest Neighbors (KNN) alcanzó el mejor desempeño sobre el conjunto de prueba temporal (R^2 = 0,736; RMSE = 0,250; MAE = 0,174), seguido por XGBoost (R^2 = 0,728). El análisis SHAP confirmó al consumo de energía, el PIB per cápita y el área selvática como los determinantes de mayor impacto predictivo. Como extensión metodológica, modelos de clasificación multiclase alcanzaron un F1-score de 0,884. Los hallazgos se sintetizaron en un tablero interactivo que facilita la exploración regional y la simulación de escenarios de política energética. Los resultados confirman que la desigualdad climática en América Latina es producto de la interacción entre intensidad energética, desarrollo económico y condiciones sociales. El modelo constituye un insumo técnico para el diseño de políticas de transición energética justa y diferenciada en la región.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

Las emisiones de dióxido de carbono (CO_2) constituyen uno de los principales factores responsables del cambio climático global. En América Latina y el Caribe, esta problemática adquiere una dimensión particular debido a la combinación de vulnerabilidad ambiental, desigualdades estructurales y limitadas capacidades técnicas para mitigar o adaptarse a sus efectos. Aunque la región representa apenas el 8.4 % de las emisiones globales de gases de efecto invernadero (GEI), es una de las más afectadas por fenómenos extremos, afectaciones a la seguridad alimentaria y degradación de sus ecosistemas [1], [2].

En este contexto se evidencia un hecho importante de la injusticia climática, donde los países con baja responsabilidad histórica enfrentan costos sociales y económicos más elevados. Además, las emisiones de CO_2 per cápita dentro de la región presentan marcadas disparidades, reflejando diferencias en ingreso, urbanización, matriz energética y desarrollo humano [3] . Comprender estas desigualdades resulta fundamental para proponer soluciones diferenciadas que respeten las capacidades y condiciones de cada país.

Frente a esta realidad, el presente proyecto se enfoca en analizar la relación entre variables socioeconómicas y emisiones de CO_2 per cápita en los países de América Latina, mediante la aplicación de herramientas de ciencia de datos. Se busca generar evidencia empírica que permita identificar patrones estructurales de desigualdad climática y visibilizar factores que explican los niveles de emisión observados. Para ello, se integraron fuentes de datos abiertos y de acceso público, EDGAR para las emisiones de CO_2, los Indicadores de Desarrollo Mundial del Banco Mundial para las variables socioeconómicas y energéticas, y el Índice de Desarrollo Humano del PNUD; conformando una base panel de 17 países latinoamericanos para el período 1990–2020. Sobre esta base se aplicó una estrategia metodológica dual para identificar asociaciones estructurales entre los factores socioeconómicos y las emisiones de CO_2 per cápita, controlando la heterogeneidad no observada entre países", y cinco algoritmos de aprendizaje automático, Random Forest, XGBoost, K-Nearest Neighbors, redes neuronales multicapa y Support Vector Regression para capturar patrones no lineales y maximizar la capacidad predictiva. La explicabilidad del mejor modelo se garantizó mediante análisis SHAP, permitiendo cuantificar la contribución individual de cada variable a las predicciones generadas. Los hallazgos se sintetizan en un tablero interactivo que facilita la exploración visual de los patrones regionales de desigualdad climática y su comunicación a audiencias técnicas y de política pública.

Los resultados identificaron el consumo energético, el PIB per cápita y el Índice de Desarrollo Humano como los principales determinantes de las emisiones de CO_2 per cápita en la región, con el modelo KNN alcanzando un coeficiente de determinación de 0,736 sobre el conjunto de prueba temporal (2016–2020). Esta evidencia empírica busca apoyar el diseño de políticas climáticas diferenciadas que reconozcan las asimetrías estructurales entre países latinoamericanos y orienten la transición energética desde un enfoque de justicia climática.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1 Definición del problema

1.1.1 Planteamiento del problema

El cambio climático representa uno de los desafíos más críticos del siglo XXI, derivado principalmente del aumento de gases de efecto invernadero (GEI), entre los cuales el dióxido de carbono (CO_2) es el principal responsable de las emisiones globales. A pesar de que América Latina y el Caribe (ALC) contribuye con aproximadamente el 8.4% de las emisiones globales de GEI, la región enfrenta una vulnerabilidad significativa debido a su alta exposición a fenómenos climáticos extremos y a la dependencia de sectores sensibles al clima, como la agricultura y los recursos hídricos. [4] Si bien la transición hacia economías bajas en carbono se ha convertido en una prioridad global en el marco del Acuerdo de París y los Objetivos de Desarrollo Sostenible, en América Latina, dicha transición enfrenta limitaciones estructurales derivadas de las profundas desigualdades socioeconómicas que caracterizan a la región y de capacidades técnicas dispares entre países. Las estrategias globales de mitigación buscan reducir las emisiones de GEI, pero las diferencias estructurales entre países pueden generar tensiones en el debate entre la responsabilidad histórica y el derecho al desarrollo. [2]

Las emisiones per cápita varían considerablemente entre países [5] , influenciadas por factores como el Producto Interno Bruto (PIB), el grado de urbanización, el tipo de matriz energética, el acceso a servicios energéticos y la estructura productiva[6]. Además, algunos estudios han demostrado una relación positiva entre el crecimiento económico y las emisiones de CO_2 [6], aunque esta relación depende del nivel de desarrollo y de las características estructurales de cada economía [7] . Esta desigualdad estructural se refleja en la dimensión climática: los efectos del cambio climático y los recursos necesarios para enfrentarlo se distribuyen de forma inequitativa. El Grupo Intergubernamental de Expertos sobre el Cambio Climático resalta que los países y regiones con menor responsabilidad histórica en la generación del cambio climático son, paradójicamente, los más vulnerables a sus efectos [1]. A nivel mundial, el 10 % de los hogares con mayores emisiones per cápita produce entre el 34 % y el 45 % de todas las emisiones domésticas, mientras que el 50 % más pobre genera solo entre el 13 % y el 15 % [8]. En este contexto, el cambio climático no solo constituye una crisis ambiental, sino también una problemática de justicia social y de derechos humanos. Esta realidad evidencia un problema de desigualdad climática, donde la distribución desigual de las emisiones de CO_2 y los impactos del cambio climático reflejan y agravan las brechas socioeconómicas y estructurales en regiones como América Latina. Por ello, comprender la relación entre las emisiones de CO_2 per cápita y los patrones socioeconómicos que las soportan se convierte en un insumo esencial para diseñar y evaluar políticas públicas que promuevan una transición energética justa y sostenible.

La ciencia de datos se ha convertido en una herramienta significativa en la búsqueda de abordar problemáticas complejas al integrar capacidades estadísticas, computacionales y de visualización que permite extraer patrones estructurales de grandes volúmenes de información. En el contexto del cambio climático y el análisis ambiental, sus métodos permiten extraer valor analítico de grandes volúmenes de información heterogénea, estructurada y no estructurada, proveniente de

fuentes diversas. En particular, la capacidad de combinar datos socioeconómicos, demográficos, geoespaciales y ambientales abre nuevas posibilidades para comprender cómo interactúan múltiples factores en la configuración de fenómenos globales y locales. A través de técnicas estadísticas avanzadas, algoritmos de aprendizaje automático y métodos de visualización dinámica, la ciencia de datos posibilita el descubrimiento de relaciones no lineales, la detección de estructuras ocultas y la formulación de modelos explicativos y predictivos con métricas de desempeño verificables y reproducibles. La ciencia de datos puede ofrecer contribuciones novedosas para establecer correlaciones y patrones robustos respecto a dinámicas ambientales y socioeconómicas mediante el uso de los datos [9]. Esta capacidad analítica no solo permite caracterizar fenómenos en escalas temporales y espaciales diferenciadas, sino también generar evidencia empírica robusta que contribuya al diseño de estrategias más eficaces e inclusivas en la gestión del cambio climático. Para tratar esta problemática tan compleja y la desigualdad estructural de América Latina, las técnicas de ciencia de datos ofrecen un marco metodológico robusto para el análisis de las emisiones de CO_2 per cápita [10]. Estas herramientas permiten integrar y procesar simultáneamente grandes volúmenes de variables socioeconómicas y ambientales, facilitando la detección de patrones latentes, correlaciones no lineales y estructuras subyacentes que resultan invisibles en enfoques univariados o bivariados. Algoritmos como Algoritmos como los modelos de datos panel (efectos fijos y aleatorios), los métodos de ensamble basados en árboles (Random Forest, XGBoost), las máquinas de vectores de soporte (SVR/SVM), y las redes neuronales multicapa permiten desarrollar modelos con alta capacidad explicativa y predictiva. La aplicación de estos métodos no solo incrementa la capacidad de comprensión de las interacciones entre dimensiones sociales, económicas y ecológicas, sino que también genera insumos empíricos valiosos para la formulación de políticas públicas contextualizadas y basadas en evidencia.

1.1.2 Formulación del problema

Teniendo en cuenta lo anterior, surge la siguiente pregunta de investigación:
¿Cómo modelar la relación entre las emisiones de CO_2 per cápita y los factores socioeconómicos en los países de América Latina mediante técnicas de ciencia de datos para evidenciar la desigualdad climática en la región?

1.1.3 Sistematización

Preguntas de sistematización:

- ¿Qué variables socioeconómicas y ambientales son relevantes para analizar la relación con las emisiones de CO_2 per cápita en los países de América Latina?
- ¿Qué patrones o correlaciones preliminares pueden identificarse entre los factores socioeconómicos y las emisiones de CO_2 per cápita mediante técnicas de análisis exploratorio?
- ¿Qué modelos de ciencia de datos permiten representar de manera más adecuada la relación entre las emisiones de CO_2 per cápita y los factores socioeconómicos?
- ¿Qué tan efectivos son los modelos desarrollados para predecir las emisiones de CO_2 per cápita en función de las variables socioeconómicas, y cuáles presentan mejor rendimiento?
- ¿Cómo varía la relación entre emisiones y desarrollo socioeconómico en función del tipo de economía o nivel de ingreso de los países?

- ¿Qué limitaciones presentan los datos disponibles sobre América Latina para abordar el análisis de la desigualdad climática mediante ciencia de datos?

1.2 Objetivos

1.2.1 Objetivo General

Desarrollar un modelo de ciencia de datos que permita analizar la relación de emisiones de CO_2 per cápita en función de factores socioeconómicos en los países de América Latina, que evidencien patrones de desigualdad climática en la región.

1.2.2 Objetivos Específicos

- Preparar y depurar la base de datos de países de América Latina con variables socioeconómicas y ambientales relevantes.
- Analizar la influencia de factores socioeconómicos en las emisiones de CO_2 per cápita mediante técnicas de análisis exploratorio de datos (AED).
- Desarrollar modelos que permitan identificar relaciones entre los factores socioeconómicos y emisiones de CO_2 per cápita.
- Evaluar y comparar el rendimiento de los modelos mediante métricas de desempeño.
- Desarrollar una visualización interactiva que facilite la interpretación de patrones regionales de desigualdad climática.

1.3 Marco de Referencia

1.3.1 Marco Teórico

Este proyecto se fundamenta en un conjunto de conceptos que permiten abordar la relación entre emisiones de dióxido de carbono (CO_2) per cápita y los patrones socioeconómicos en América Latina desde un enfoque de ciencia de datos aplicada a la justicia ambiental. A continuación, se presentan los ejes teóricos principales:

Cambio climático y Emisiones de Dióxido de carbono (CO2)

El cambio climático es un hecho global, expresado a través de transformaciones persistentes de las condiciones climáticas, como el aumento de la temperatura media y la mayor frecuencia de eventos climáticos extremos. En América Latina, estas alteraciones tienen efectos desproporcionados sobre sectores sensibles como la agricultura, la gestión del agua y los asentamientos costeros, lo que acentúa las desigualdades estructurales existentes en la región. De acuerdo con el IPCC (2022), el cambio climático se manifiesta en alteraciones significativas y persistentes en la media y variabilidad del clima, atribuibles principalmente al incremento de gases de efecto invernadero (GEI) derivados de actividades humanas como el uso de combustibles fósiles, la deforestación y los procesos industriales. El informe resalta que la temperatura media global alcanzó aproximadamente 1,1 °C por encima de los niveles preindustriales (1850–1900) durante el período comprendido entre 2011 y 2020, mientras que las concentraciones atmosféricas de dióxido de carbono (CO_2) llegaron a 417,1 ppm en 2022, el nivel más alto registrado en al menos 800.000 años. En este mismo sentido, las emisiones globales de GEI alcanzaron cerca de 54 ± 5,3 $GtCO_2e$ en la última década, evidenciando la influencia antropogénica sobre el sistema climático global. [11]

Los GEI, entre los cuales se destacan el dióxido de carbono (CO_2), el metano (CH_4) y el óxido nitroso (N_2O), desempeñan un papel esencial en el sistema climático, ya que permiten la retención del calor en la atmósfera terrestre [12]. No obstante, su concentración ha crecido considerablemente desde la Revolución Industrial, intensificando el efecto invernadero y provocando un calentamiento global acelerado.

Particularmente, el CO_2 es el gas de efecto invernadero más emitido por las actividades humanas y el que permanece más tiempo en la atmósfera. Según el Global Carbon Budget 2024, las emisiones globales de CO_2 derivadas del uso de combustibles fósiles alcanzaron los 37,4 mil millones de toneladas, y al incluir las emisiones por cambio en el uso del suelo, como la deforestación, el total asciende a 41,6 mil millones de toneladas, reflejando un incremento del 2% respecto al año anterior. Este aumento se relaciona con una reducción en la capacidad de absorción de carbono por parte de los ecosistemas naturales, afectada por eventos como incendios forestales prolongados y sequías intensas. [13]

Las principales fuentes antropogénicas de CO_2 incluyen la generación de energía a partir de combustibles fósiles, el transporte, la industria pesada y el cambio en el uso del suelo. Aunque los

ecosistemas terrestres y oceánicos actúan como sumideros de carbono, su capacidad de absorción ha sido sobrepasada por el ritmo actual de emisiones. [11]

En este contexto, el análisis de las emisiones de CO_2 per cápita se convierte en una herramienta fundamental para dimensionar las responsabilidades diferenciadas entre países. Esta métrica permite contrastar los niveles de emisión ajustados al tamaño poblacional, visibilizando las desigualdades climáticas entre naciones con diferentes condiciones socioeconómicas, lo cual es clave para diseñar políticas públicas que promuevan la equidad ambiental.

Desigualdad Climática y Justicia ambiental

El cambio climático no afecta a todas las poblaciones por igual. La exposición a sus impactos, así como la capacidad de respuesta, varía según factores como el ingreso, la ubicación, el acceso a recursos y la posición social. Esta disparidad ha dado lugar a los conceptos de desigualdad climática y justicia ambiental.

La desigualdad climática se refiere a que las poblaciones que históricamente han emitido menos gases de efecto invernadero son, paradójicamente, las más vulnerables a sus consecuencias. Según el Climate Inequality Report 2023, publicado por el International Science Council, el 10 % más rico de la población mundial es responsable de casi el 50 % de las emisiones globales de carbono, mientras que el 50 % más pobre solo genera entre el 13% y el 15 %. [8]

Por su parte, la justicia ambiental se define como un principio, en el que, todas las personas, indistintamente de su raza, género, nivel de ingresos u origen geográfico, deben gozar de igual protección frente a los riesgos ambientales y de un acceso justo a los beneficios generados por las políticas climáticas.[14]Este enfoque implica que las decisiones en materia ambiental no deben perpetuar ni agravar las desigualdades sociales existentes, sino contribuir activamente a reducirlas, especialmente en contextos históricamente vulnerables como los de América Latina.

En América Latina, el Programa de las Naciones Unidas para el Desarrollo (PNUD) resalta que la justicia climática debe colocar los derechos humanos y la equidad en el centro de las políticas públicas, especialmente en contextos donde comunidades indígenas y de bajos ingresos enfrentan mayores riesgos climáticos. [7]

Además, The Climate Impacts Group de la Universidad de Washington señala que la vulnerabilidad de una comunidad no depende solo de condiciones ambientales actuales, sino también de factores históricos, económicos y sociales que deben ser considerados al diseñar estrategias de adaptación. [15]

Para la Ciencia de Datos, esto implica integrar variables sociales y económicas en los análisis climáticos, permitiendo diseñar soluciones más justas y adaptadas a las realidades regionales.

En el marco del presente proyecto, la desigualdad climática se operacionaliza como la dispersión estructural en los niveles de emisiones de CO_2 per cápita entre países latinoamericanos, entendida como un indicador proxy de la contribución diferencial al cambio climático dentro de la región. Esta aproximación reconoce que la desigualdad climática es un concepto multidimensional , que

abarca también vulnerabilidad, capacidad adaptativa y justicia distributiva y, que el análisis de emisiones per cápita constituye una de sus dimensiones empíricamente observables y comparables entre países a partir de datos históricos. Las dimensiones de vulnerabilidad y adaptación, si bien relevantes, quedan fuera del alcance del presente estudio y se proponen como líneas de investigación futura

Determinantes Socioeconómicos de las Emisiones de CO2

Los factores socioeconómicos desempeñan un papel fundamental en la explicación de las variaciones en las emisiones de CO_2 per cápita entre los países de América Latina. A continuación, se destacan los principales determinantes en este contexto:

Ingreso per cápita: Según Durmaz, Liu y Lu (2025), el crecimiento económico y el aumento del consumo energético en América Latina mantienen una relación positiva con las emisiones de CO_2, evidenciando que una mayor actividad económica incrementa la demanda energética y la presión ambiental. Asimismo, el estudio concluye que los países analizados no presentan evidencia consistente de la hipótesis de la Curva Ambiental de Kuznets, ya que las emisiones continúan aumentando junto con el consumo de energía [16].

Nivel educativo: El grado de educación de la población influye significativamente en la conciencia ambiental y en la capacidad de adoptar prácticas sostenibles. Diversos estudios evidencian que mayores niveles educativos están asociados con una mayor sensibilización frente al cambio climático y una mayor disposición hacia comportamientos proambientales y apoyo a políticas ambientales. En este sentido, Jiménez García, Pérez Peña y López Sánchez (2023) destacan que la educación constituye un factor clave en la formación de conciencia climática y en el fortalecimiento de actitudes responsables frente al medio ambiente. [17]

Urbanización: El proceso de urbanización incrementa significativamente la demanda energética, especialmente en sectores como el transporte, la vivienda y la infraestructura urbana. En este contexto, la expansión urbana desordenada y la ausencia de una planificación territorial eficiente pueden generar una mayor dependencia del transporte privado, mayores distancias de desplazamiento y un uso menos eficiente de los recursos, contribuyendo al incremento de las emisiones de CO_2. En América Latina, Van der Borght y Pallares Barbera (2023) evidencian que la expansión espacial urbana tiene un impacto significativo sobre las emisiones, particularmente en ciudades con patrones de crecimiento disperso y baja densidad urbana [18].

Desigualdad económica: Las disparidades en la distribución del ingreso influyen significativamente en los patrones de consumo energético y en la generación de emisiones de carbono. Las diferencias económicas entre países y grupos poblacionales condicionan tanto los niveles de consumo como las capacidades de mitigación y transición hacia modelos más sostenibles. En este sentido, Wang Z y colaboradores (2024) evidencian una marcada desigualdad entre países sudamericanos en las emisiones de carbono incorporadas, asociada a factores económicos, comerciales y estructurales. Asimismo, las poblaciones más vulnerables suelen

enfrentar mayores limitaciones para adaptarse a los impactos del cambio climático y acceder a soluciones sostenibles [19].

Acceso a energías renovables: La disponibilidad de fuentes limpias y la infraestructura para su adopción son fundamentales para reducir las emisiones de CO_2. Según la IEA (2023), si bien países como Brasil y Costa Rica han alcanzado matrices eléctricas con más del 80% de fuentes renovables, la región en su conjunto mantiene una dependencia de combustibles fósiles superior al 60% en su consumo energético total, lo que limita el espacio de mitigación a corto plazo[20].

Ciencia de datos en estudios ambientales

La ciencia de datos es hoy una herramienta clave para analizar fenómenos socioambientales complejos, al integrar capacidades de procesamiento, análisis y visualización de grandes volúmenes de datos con enfoques estadísticos y computacionales avanzados. Su aplicación en estudios sobre cambio climático ha permitido ampliar la comprensión de los determinantes sociales y económicos de las emisiones de gases de efecto invernadero, Además permite identificar patrones regionales y tendencias estructurales relacionadas con la desigualdad climática mediante técnicas de machine learning.

En contextos como América Latina, caracterizados por altos niveles de heterogeneidad socioeconómica y vulnerabilidad ambiental, el análisis basado en ciencia de datos permite articular múltiples dimensiones como ingreso, desarrollo humano, urbanización y acceso a energía para estudiar su relación con las emisiones de CO_2 per cápita [16], [19]. Asimismo, el enfoque multivariado posibilita una lectura más integral de los factores asociados a la contribución y exposición al cambio climático, alineada con los principios de equidad y responsabilidades diferenciadas consagrados en la Convención Marco de las Naciones Unidas sobre el Cambio Climático [2].

Los métodos de análisis exploratorio, segmentación de datos y visualización dinámica permiten transformar grandes bases de datos en información útil para la toma de decisiones. Esta capacidad analítica ha sido utilizada para caracterizar disparidades entre países o regiones en función de sus niveles de emisión, condiciones estructurales y capacidades de adaptación [1], [21] . En este mismo sentido, podemos decir que la ciencia de datos no solo proporciona herramientas técnicas, sino también una vía para fortalecer la justicia climática en el diseño de políticas públicas.

Métodos de Aprendizaje automático y Explicabilidad

En el campo del aprendizaje automático supervisado, los métodos de ensamble basados en árboles de decisión han demostrado una capacidad superior para modelar relaciones no lineales en datos tabulares. Random Forest, propuesto por Breiman [22], combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios del conjunto de datos, reduciendo la varianza del estimador mediante el promedio de predicciones individuales. Extreme Gradient Boosting (XGBoost), extiende este principio mediante un proceso de boosting secuencial con regularización explícita, en el que cada árbol corrige los errores del modelo anterior mediante descenso del

gradiente. Ambos métodos son especialmente adecuados para datos estructurados con alta dimensionalidad relativa y relaciones complejas entre variables.

Los métodos basados en instancia, como K-Nearest Neighbors (KNN), generan predicciones a partir de la similitud local entre observaciones en el espacio de características. Su comportamiento no paramétrico los hace sensibles a la escala de las variables, por lo que requieren estandarización previa. Las máquinas de vectores de soporte para regresión (SVR) buscan construir una función que tolere errores dentro de un margen ε, proyectando los datos mediante funciones kernel a espacios de mayor dimensionalidad. Las redes neuronales multicapa (MLP) aproximan funciones complejas mediante capas sucesivas de transformaciones no lineales.

Para garantizar la interpretabilidad de los modelos de caja negra, este estudio emplea la metodología SHapley Additive exPlanations (SHAP), propuesta por Lundberg y Lee Fundamentada en la teoría de juegos cooperativos, SHAP calcula la contribución marginal promedio de cada variable a las predicciones del modelo, considerando todas las posibles coaliciones de predictores. Esta propiedad garantiza consistencia, exactitud local y aditividad, convirtiéndola en el marco de explicabilidad más sólido disponible para algoritmos de aprendizaje automático. [23]

1.3.2 Antecedentes

Artículo científico "Driving Factors of CO_2 Emissions: Further Study Based on Machine Learning" (2021)

Este estudio, desarrollado por Liu et al. y publicado en Frontiers in Environmental Science, analiza los factores que impulsan las emisiones de CO_2 en China entre los años 2000 y 2018, utilizando datos de 30 regiones administrativas del país. El análisis se centra en variables como el crecimiento económico, la estructura industrial, la urbanización, la inversión en investigación y desarrollo (I+D), el uso de capital extranjero y el consumo energético. Para modelar el comportamiento de las emisiones, se aplicaron diferentes algoritmos de aprendizaje automático, entre ellos modelos lineales, no lineales, métodos de ensamble (boosting y bagging), y redes neuronales artificiales. El modelo con mejor rendimiento fue el de k-nearest neighbors (KNN), que obtuvo la menor tasa de error utilizando el RMSE como métrica principal [24]

Este trabajo aporta al desarrollo del presente proyecto al demostrar la utilidad de los modelos de aprendizaje automático para analizar y predecir emisiones de CO_2 en función de variables económicas y estructurales. Aporta una base metodológica robusta sobre el uso comparativo de algoritmos y el análisis de sensibilidad. La diferencia principal con el enfoque de este proyecto es que, mientras el estudio se concentra exclusivamente en el caso de China a nivel subnacional, este antecedente orientó la decisión de incluir KNN como algoritmo de comparación en el presente estudio, evaluando si la similitud estructural entre países latinoamericanos produce un comportamiento predictivo análogo al observado entre regiones de China.

Artículo científico "Identification of Patterns in CO_2 Emissions among 208 Countries: K-Means Clustering Combined with PCA and Non-Linear t-SNE Visualization" (2024)

Este estudio, publicado en Mathematics y desarrollado por investigadores del Instituto Politécnico Nacional de México, analiza patrones de emisiones de CO_2 (tanto totales como per cápita) en 208 países, considerando diversas fuentes como cemento, gas, petróleo, carbón y flaring. Para ello, se aplican técnicas de reducción de dimensionalidad lineales y no lineales, específicamente el Análisis de Componentes Principales (PCA) y el método de t-SNE, combinadas con algoritmos de agrupamiento no supervisado tipo K-means. El número óptimo de clústeres se determina mediante los índices de Calinski–Harabasz y Davies–Bouldin. Entre los principales hallazgos, se destaca la identificación de perfiles diferenciados de emisión: por ejemplo, China, EE. UU., India y Rusia son agrupados en un solo clúster por t-SNE debido a sus altos niveles de emisiones, mientras que PCA los segmenta por separado. En el caso de las emisiones per cápita, países como Qatar se destacan por tener perfiles de emisión únicos debido a su baja población y elevada actividad de flaring [25] .

Este trabajo resulta relevante para el presente proyecto porque muestra cómo las técnicas de ciencia de datos permiten descubrir patrones ocultos en datos de alta dimensionalidad sobre emisiones de carbono. Además, refuerza el valor del análisis comparativo entre países para diseñar estrategias diferenciadas de mitigación. La diferencia principal con nuestra propuesta radica en que este artículo tiene un enfoque global sin incorporar variables socioeconómicas ni una perspectiva explícita de desigualdad climática. Por el contrario, nuestro proyecto se enfoca en América Latina y busca establecer relaciones entre factores estructurales y emisiones per cápita, integrando dimensiones sociales que permitan interpretar los resultados desde un enfoque de justicia climática.

Artículo científico "The Determinants of CO_2 Emissions in the Context of ESG Models at World Level" (2023)

Este artículo analiza los determinantes de las emisiones de CO_2 (COE) a nivel global, en el contexto del modelo ESG (Ambiental, Social y de Gobernanza), utilizando datos del Banco Mundial para 193 países entre 2011 y 2020. A través de técnicas de regresión, análisis de conglomerados con k-Means y predicción con ocho algoritmos de aprendizaje automático, los autores identifican variables fuertemente asociadas con el nivel de emisiones. Entre los principales hallazgos se destaca una asociación positiva entre las emisiones de CO_2 y variables como las emisiones de metano y la inversión en investigación y desarrollo, mientras que el consumo de energías renovables y el índice promedio de sequía se asocian negativamente. La predicción del valor futuro de las emisiones de CO2 mediante redes neuronales artificiales (ANN) sugiere una reducción promedio del 5,69% en las emisiones para los países analizados.

Este trabajo contribuye al presente proyecto al mostrar cómo las técnicas de ciencia de datos pueden usarse para modelar los determinantes de las emisiones en un marco global y en diálogo con indicadores de sostenibilidad. Aunque se centra en el modelo ESG y tiene una perspectiva más financiera y corporativa, su enfoque multivariado y el uso de algoritmos de machine learning aportan herramientas valiosas para el análisis. A diferencia de la propuesta actual, que se enfoca en países de América Latina y en variables estructurales socioeconómicas para evidenciar desigualdades climáticas, este artículo prioriza el enfoque institucional de sostenibilidad.

2. OBJETIVO 1: Preparar y depurar la base de datos de países de América Latina con variables socioeconómicas y ambientales relevantes.

2.1 Obtención e integración de las bases de datos

Para el desarrollo del primer objetivo se recopilaron diversas fuentes de información socioeconómica y ambiental correspondientes a los países de América Latina y el Caribe (ALC). Las fuentes principales fueron el Banco Mundial (WDI), el Programa de las Naciones Unidas para el Desarrollo (PNUD), la Comisión Económica para América Latina y el Caribe (CEPAL) y Emissions Database for Global Atmospheric Research, JRC-Copernicus (EDGAR) Y Harvard Dataverse.

Estas fuentes proporcionaron series anuales de variables relacionadas con el desarrollo humano, la desigualdad, la estructura económica, el consumo energético y las emisiones de CO_2. Dado que cada fuente presentaba diferentes estructuras se realizó un proceso de transformación y estandarización mediante técnicas de *data wrangling*. Esto permitió integrar los datos en un formato panel país–año, que facilita el análisis temporal y comparativo entre países.

Para asegurar la consistencia de los registros, se empleó el código ISO3 como identificador único de país, y se homogenizaron las unidades de medida y las escalas, garantizando la comparabilidad entre fuentes. Cuando se detectaron diferencias en la escala los valores fueron ajustados a la escala estándar [0–1] únicamente para resolver incompatibilidades de escala durante la integración de fuentes, para los modelos econométricos se usaron las variables en sus unidades originales, para los modelos se ML se usaron estandarización.

2.2 Estructura y descripción de la base consolidada

La base final, contiene 527 observaciones para 17 países con un total de 15 variables. Estas variables capturan dimensiones complementarias del desarrollo sostenible, la desigualdad y el desempeño ambiental de los países latinoamericanos durante el período 1985–2022. Las variables analíticas se agrupan en cuatro dimensiones principales:

Dimensión	Variables
Ambiental	area_selvatica, co2_total, co2_pc
Económica	pib_per_capita, industria, inversion_extranjera

Energética	uso_energia, consumo_energia, energia_fosil
Sociodemográfica	alfabetizacion, idh, gini, poblacion_total, poblacion_urbana, poblacion_rural

Tabla 1

Dimensión de las variables

Tabla 2 Fuente de las bases de datos por variable

Variable	Descripción	Unidades	Fuente
poblacion_total	Población total	Personas	Banco Mundial
poblacion_urbana	Población urbana	%	Banco Mundial
poblacion_rural	Población rural	%	Banco Mundial
pib_per_capita	PIB per cápita	USD constantes	Banco Mundial
inversion_extranjera	Inversión extranjera directa	% PIB	Banco Mundial
co2_total	Emisiones totales de CO_2	Kilotoneladas	EDGAR
co2_pc	Emisiones de CO_2 per cápita	Toneladas/persona	Cálculo propio
industria	Valor agregado industrial	% PIB	Banco Mundial
area_selvatica	Área de bosque	% territorio	Banco Mundial
consumo_energia	Consumo de energía eléctrica	kWh per cápita	Banco Mundial
uso_energia	Uso de energía per cápita	kg eq. petróleo	Banco Mundial
energia_fosil	Energía de fuentes fósiles	% del total	Banco Mundial
idh	Índice de Desarrollo Humano	0-1	PNUD
gini	Coeficiente de Gini	%	Harvard Dataverse
alfabetizacion	Alfabetización adulta	%	Banco Mundial

2.3 Análisis de completitud y consistencia de la información

Para evaluar la calidad de la información disponible, se realizó un análisis de completitud que permitió identificar el grado de cobertura de cada variable, país y año dentro de la base de datos. Este proceso es fundamental en estudios comparativos de carácter regional, ya que la disponibilidad desigual de información puede introducir sesgos en los resultados o limitar el uso de determinadas variables.

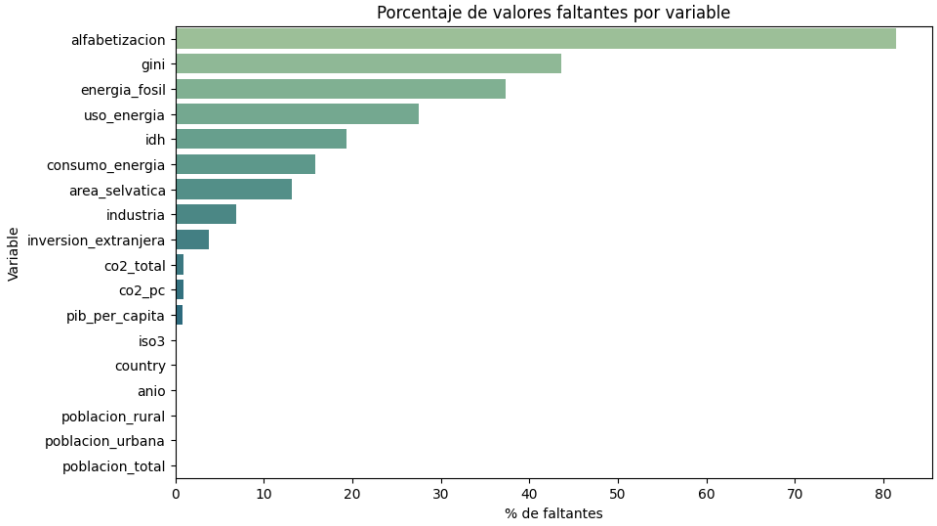
Tabla 3 Porcentaje valores faltantes

Variable	Porcentaje faltantes
alfabetizacion	81,42
gini	43,62
energia_fosil	37,32

uso_energia	27,51
idh	19,38
consumo_energia	15,79
area_selvatica	13,16
industria	6,86
inversion_extranjera	3,83
co2_total	0,96
co2_pc	0,96
pib_per_capita	0,80
iso3	0
country	0
anio	0
poblacion_rural	0
poblacion_urbana	0
poblacion_total	0

La evaluación se efectuó mediante el cálculo del porcentaje de valores faltantes por variable y por país.

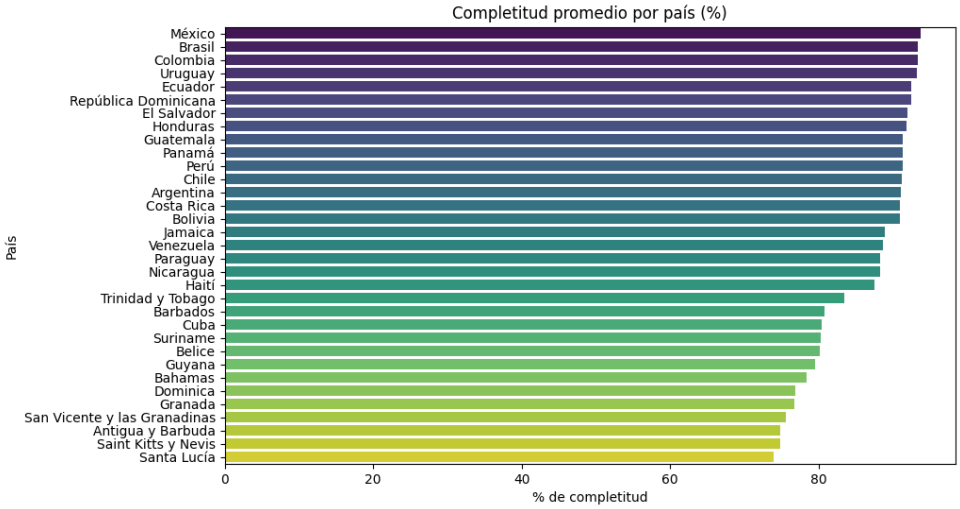
Ilustración 1 Porcentaje de valores faltantes por variable



En la ilustración 1 se observa que las variables con mayor proporción de valores ausentes corresponden a alfabetización (81 %), índice de Gini (43%) y energía fósil (37 %), mientras que las variables demográficas y económicas como población total, población urbana, PIB per cápita e IDH presentan una cobertura superior al 90 %.

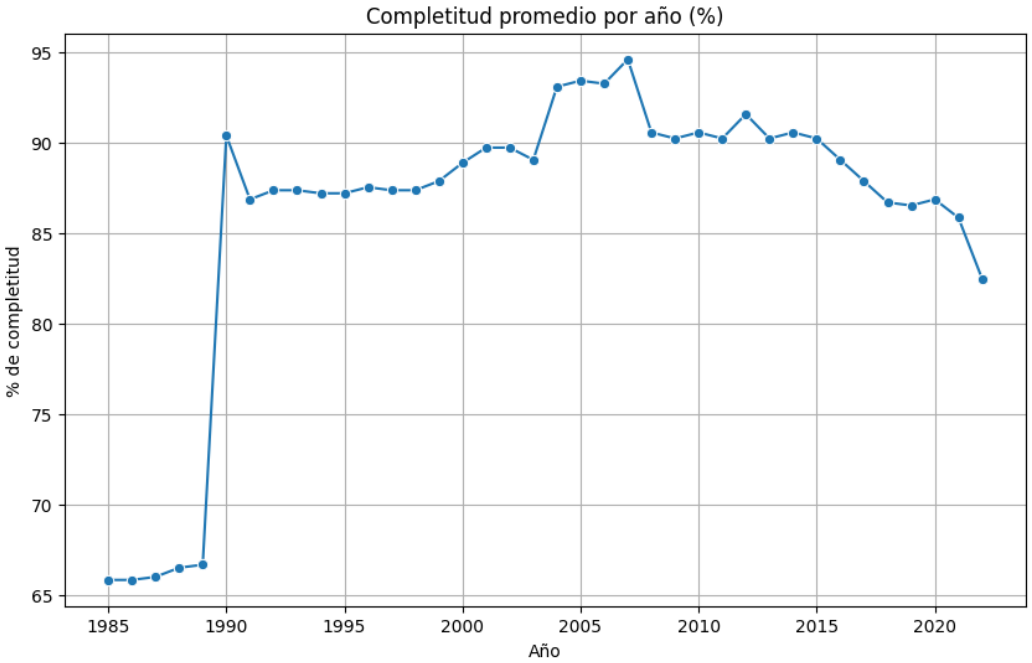
Los resultados reflejan una mayor disponibilidad de información en las dimensiones económicas y demográficas, y una menor consistencia en las estadísticas sociales y energéticas, especialmente en los años anteriores al 2000.

Ilustración 2 Completitud promedio (%) por país



En cuanto a la completitud promedio por país, según la ilustración 2, los niveles más altos se presentan en México, Chile, Colombia, Argentina y Brasil, con coberturas superiores al 90 %, lo cual es coherente con sus sistemas estadísticos consolidados. En contraste, Surinam, Guyana, Haití y Cuba muestran una cobertura inferior al 70 %, lo que se asocia con brechas institucionales y limitaciones en la disponibilidad histórica de datos

Ilustración 3 Completitud promedio por año



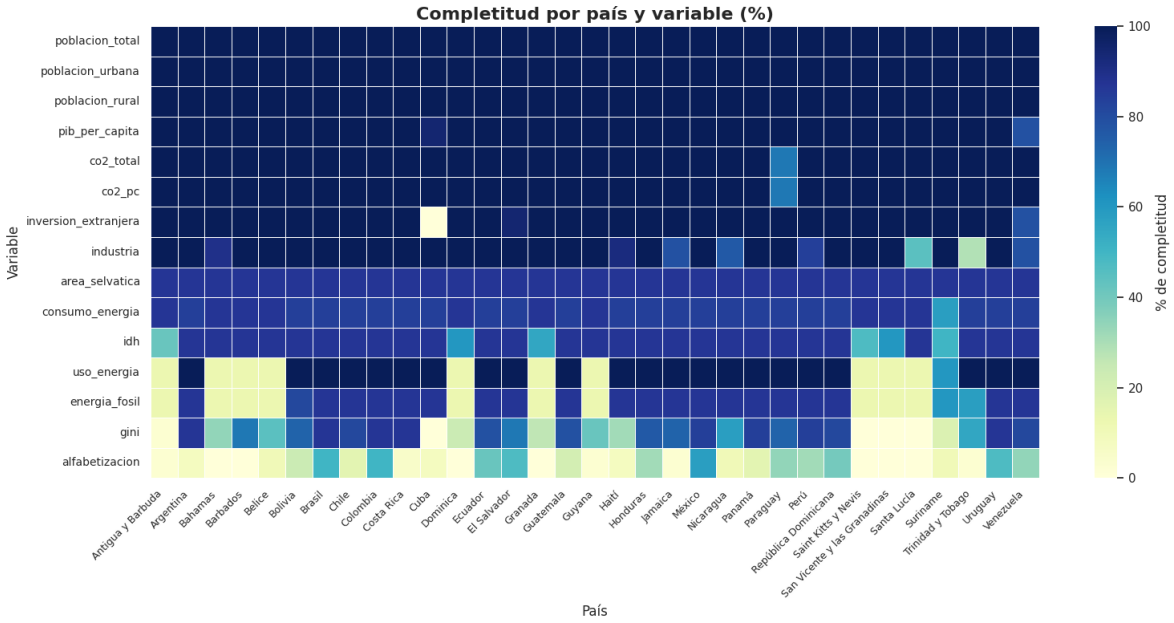
El análisis temporal (Ilustración 3) evidencia que la completitud de la base aumenta significativamente a partir del año 1995, alcanzando su punto máximo entre 2005 y 2015, periodo caracterizado por la consolidación de bases globales estandarizadas (como WDI y PNUD). Sin embargo, se observan descensos leves en los años más recientes, debido a la falta de actualización de algunos indicadores ambientales.

En este sentido, con el fin de evaluar la cobertura espacial y temporal de la base de datos, se analizaron los patrones de completitud por país y por año para cada variable.

En la ilustración 4, el mapa de calor evidencia un comportamiento heterogéneo en la disponibilidad de información entre los países de América Latina. Se observa una cobertura alta y estable en países con sistemas estadísticos consolidados como Chile, Argentina, Colombia, Uruguay, México, Ecuador y Costa Rica, los cuales presentan niveles de completitud superiores al 85 % en la mayoría de las variables.

En contraste, países como Nicaragua, Honduras, El Salvador, Paraguay, Bolivia y Venezuela exhiben amplios vacíos de información, particularmente en variables sociales como el índice de Gini y la alfabetización, y en menor medida en las variables energéticas (energía fósil, uso energía). Este patrón sugiere que la ausencia de datos no es completamente aleatoria (MAR), sino que está asociada con factores estructurales como la capacidad institucional y la continuidad de los reportes nacionales.

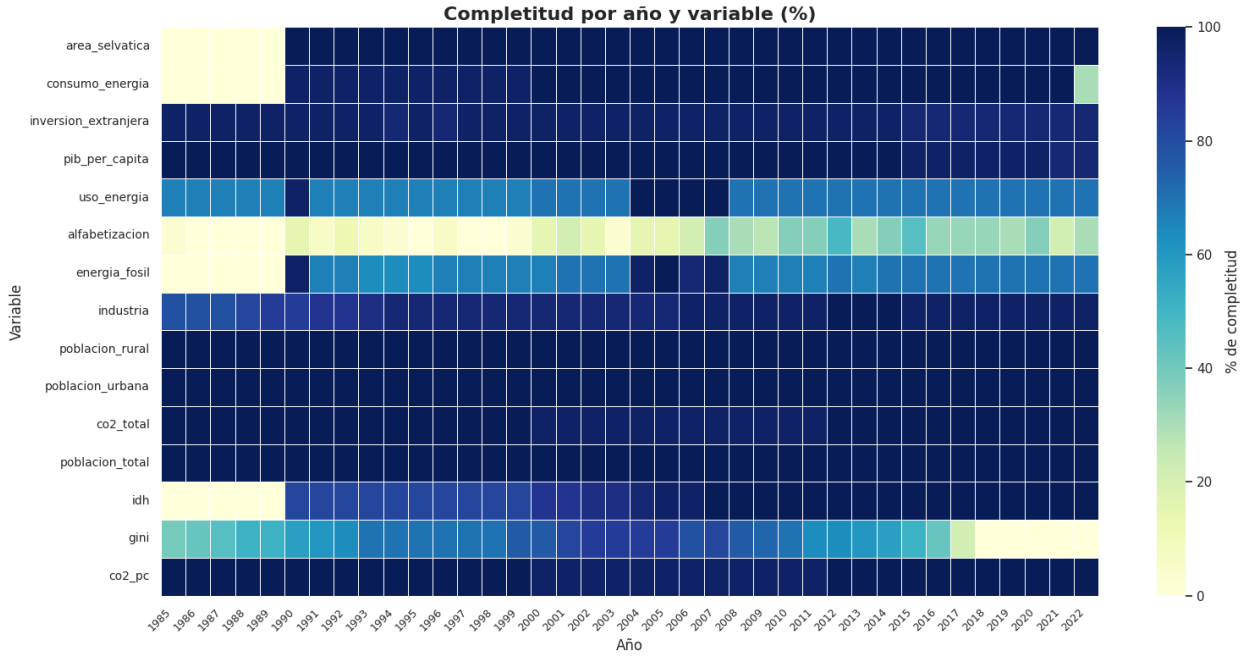
Ilustración 4 Completitud por pais y variable (%)



En la ilustración 5, el análisis por año muestra que las series más completas corresponden al período comprendido entre 1995 y 2015, donde la cobertura general supera el 85 % para la mayoría de los indicadores. En los años anteriores a 1990, la disponibilidad de datos es irregular y discontinua, especialmente en variables sociales y energéticas.

A partir de 2015, se evidencia una ligera reducción de la completitud debido a retrasos en la actualización de las estadísticas ambientales y de desarrollo humano. En consecuencia, se propone utilizar el período 1990–2020 como línea temporal principal para los análisis comparativos y modelados posteriores, garantizando un equilibrio entre cobertura y continuidad.

Ilustración 5 Completitud por año y variable (%)



2.4 Depuración de la base de datos

En coherencia con este contexto, y buscando garantizar la consistencia temporal y comparabilidad regional de la base de datos, se decidió excluir los países del Caribe del análisis principal, manteniendo únicamente aquellos con series temporales continuas y niveles de completitud superiores al 70 %.

Tras aplicar los criterios de depuración y consistencia, se procedió a excluir los países del Caribe debido a sus altos niveles de datos faltantes y discontinuidad temporal, así como la variable *alfabetización*, que presentó más del 70 % de valores ausentes y escasa comparabilidad entre países.

De este modo, la base consolidada final quedó conformada por 17 países de América del Sur y Centroamérica y abarca el período 1990–2020, garantizando un equilibrio entre cobertura temporal y representatividad regional.

Los resultados obtenidos en el análisis de completitud evidencian una marcada heterogeneidad en la disponibilidad de información entre los países latinoamericanos, siendo particularmente limitada en los países del Caribe insular. Esta situación coincide con la evidencia reportada en la literatura reciente, que señala que la «pequeñez» y las limitaciones estructurales propias de los Estados insulares dificultan su capacidad para mantener sistemas estadísticos continuos y digitalizados.

En este sentido, el estudio destaca que "*a medida que aumenta la demanda global y los beneficios asociados a los datos electrónicos, nuestro análisis resalta los desafíos que plantea la 'pequeñez' insular en el impulso global hacia la transformación digital. Se requieren con urgencia mecanismos de cooperación regional, apoyo internacional sostenido y un seguimiento sistemático de la disponibilidad de datos para monitorear el desarrollo de capacidades*" [26]

A continuación, se presentan los mapas de calor de completitud (Ilustraciones 6 y 7), que permiten

observar la distribución de los valores disponibles por variable, país y año. Estas visualizaciones son esenciales para identificar posibles sesgos de información y verificar la estabilidad de la serie temporal antes de aplicar técnicas de imputación.

Ilustración 6 Completitud por año y Variable (1990-2020)

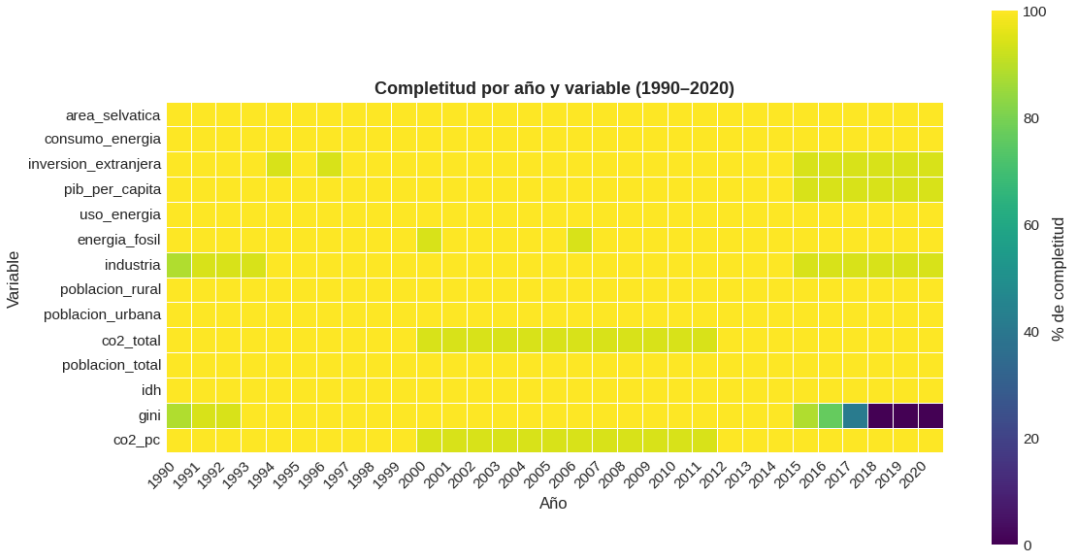
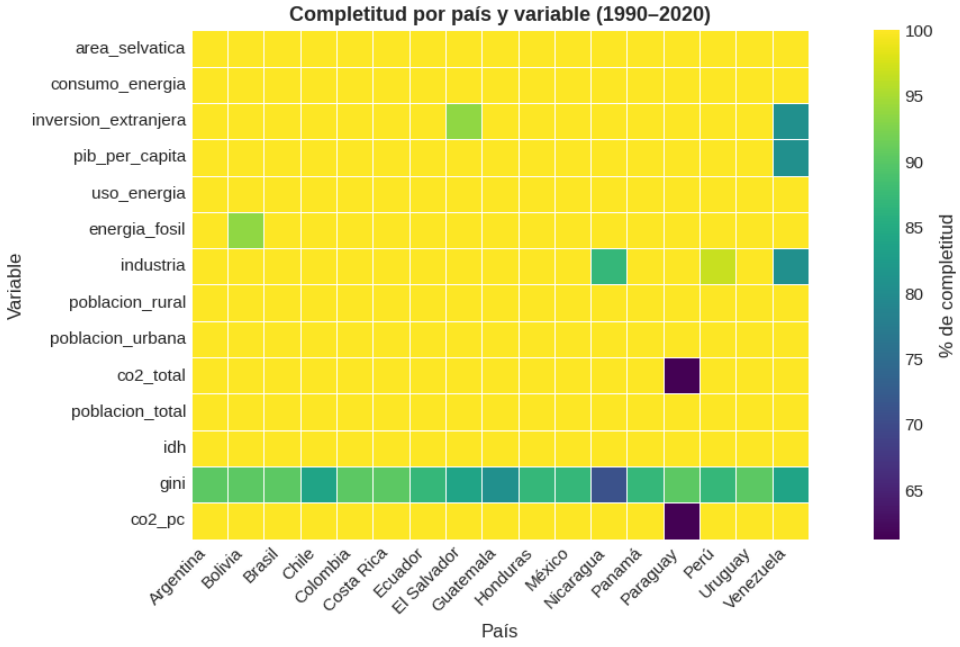


Ilustración 7 Ilustración 7 Completitud por país y variable (1990-2020)



Los mapas de completitud permiten visualizar la distribución de los valores disponibles en la base de datos final. Se observa una cobertura alta y homogénea en la mayoría de los países y variables, especialmente en las series asociadas a emisiones de CO_2, población e indicadores

económicos.
Las brechas de información son más evidentes en algunas variables energéticas y de inversión extranjera, y en países con capacidades estadísticas limitadas como Honduras, Nicaragua y El Salvador, aunque sin afectar de forma crítica la representatividad general del conjunto.
A nivel temporal, la cobertura es estable a partir de 1995, con registros continuos hasta 2020, lo que confirma la pertinencia del rango adoptado. En conjunto, estos resultados respaldan la idoneidad de la base depurada para la aplicación de técnicas de imputación mediante interpolación lineal país-año, las cuales permitirán completar los valores faltantes preservando la estructura y correlaciones originales de los datos.

2.5 Metodología de imputación por interpolación lineal país-año

Durante la fase de depuración y completitud de la base de datos, se identificó la presencia de valores faltantes en diversas variables socioeconómicas y ambientales para varios países de América Latina. En el conjunto de datos analizado, los valores faltantes no se presentan de forma completamente aleatoria, sino que muestran una dependencia estructural asociada a las características de los países y las variables. En particular, se observa que los indicadores económicos y ambientales presentan mayor proporción de vacíos en los países con menor capacidad institucional o disponibilidad estadística, y en los primeros años del periodo analizado. Este patrón es consistente con un mecanismo Missing At Random (MAR), en el sentido de Little y Rubin (2020)[27], aunque esta clasificación constituye una asunción no completamente verificable en la práctica. La decisión de adoptar interpolación lineal se justifica bajo este supuesto, reconociendo que si el mecanismo fuera MNAR, los resultados podrían estar sujetos a sesgo de imputación. Por este motivo, se optó por la aplicación de interpolación lineal por país, un método ampliamente utilizado en estudios longitudinales para la reconstrucción de series temporales incompletas. La interpolación lineal permite estimar valores ausentes preservando la continuidad temporal y la tendencia estructural de las variables, sin alterar las correlaciones entre indicadores. Esta técnica es especialmente apropiada cuando los vacíos de información se presentan de manera esporádica o no sistemática, y los datos se encuentran ordenados cronológicamente. [28]

2.5.1 Fundamento matemático

Sea una variable $X_{i,t}$que representa el valor de un indicador (por ejemplo, el índice de Gini, PIB per cápita o emisiones de CO_2) para el país $i$en el año t.
Si para un país existen valores conocidos X_{i,t_1}y X_{i,t_2}, pero faltan observaciones intermedias X_{i,t_k}con $t_1 < t_k < t_2$, la interpolación lineal estima los valores faltantes según la ecuación:

$$\hat{X}_{i,t_k} = X_{i,t_1} + \frac{X_{i,t_2} - X_{i,t_1}}{t_2 - t_1} \times (t_k - t_1)$$

donde:
X_{i,t_1}: valor anterior disponible
X_{i,t_2}: valor posterior disponible
$\hat{X}_{i,t_k}$: valor interpolado en el año t_k

Esta fórmula corresponde a una estimación del valor faltante mediante la pendiente del segmento lineal que conecta los puntos válidos inmediatos, asumiendo una variación continua y constante entre ambos años.

2.5.2 Fundamento teórico

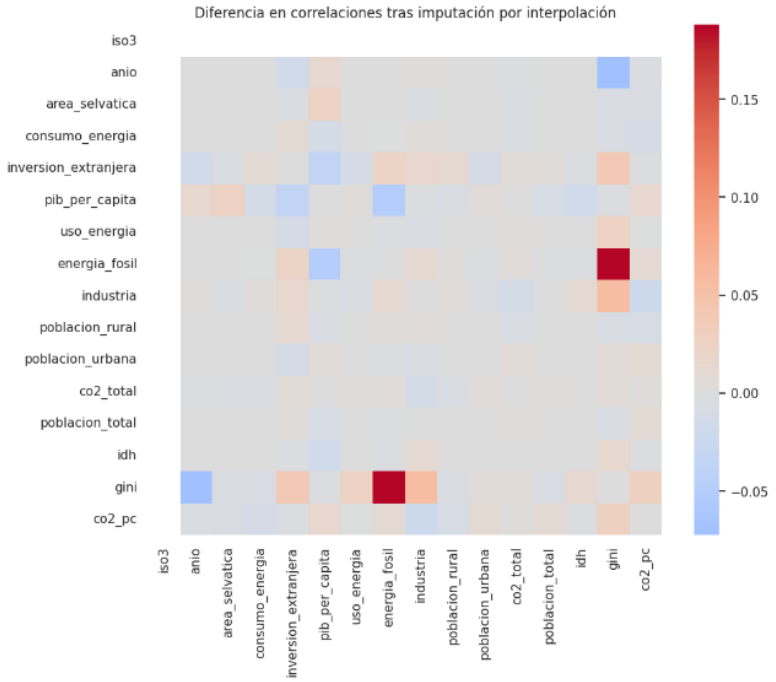
La elección de la interpolación lineal país-año se fundamenta en las siguientes consideraciones:

- **Recomendación de organismos internacionales:** el *OECD Handbook on Constructing Composite Indicators* [28]y la *CEPAL [29]*recomiendan la interpolación lineal para completar brechas anuales en indicadores socioeconómicos cuando los cambios son graduales en el tiempo.
- **Aplicación en estudios previos:** el *Programa de las Naciones Unidas para el Desarrollo (PNUD, 2022)* utiliza interpolaciones lineales para reconstruir series del Índice de Desarrollo Humano (IDH) en años sin medición directa.
- **Base teórica sólida:** de acuerdo con *Little & Rubin [27]*, este método mantiene las propiedades estadísticas de las series temporales bajo supuestos de continuidad y permite evitar el sesgo que introduce la eliminación de observaciones incompletas.
- **Simplicidad interpretativa y reproducibilidad:** la interpolación lineal es transparente, no depende de supuestos paramétricos complejos y puede implementarse de forma consistente entre variables y países.

2.5.3 Evaluación de la Interpolación

- Verificación visual mediante las distribuciones de las variables
  Antes de aplicar la interpolación, se generaron gráficos de evolución temporal para evaluar posibles cambios en el comportamiento de las series después del proceso. Se analizaron cinco variables representativas con distintos niveles de completitud de información: CO_2 per cápita, índice de Gini, uso de energía, Índice de Desarrollo Humano (IDH) y PIB per cápita.
  En los gráficos presentados en los Anexos 1 al 5, se observa que las series interpoladas conservan adecuadamente la tendencia temporal de los datos originales. No se evidencian cambios abruptos, alteraciones artificiales en las trayectorias ni desviaciones significativas respecto al comportamiento observado en los datos disponibles.
  Estos resultados sugieren que la interpolación lineal a nivel país-año permite completar los valores faltantes manteniendo la estructura temporal de las variables analizadas, sin introducir distorsiones relevantes en las series utilizadas para el modelamiento.
- Efecto de la interpolación sobre las correlaciones entre variables
  Como prueba adicional, se compararon las matrices de correlación antes y después de la interpolación.
  La ilustración 8 muestra un heatmap de diferencias entre ambas correlaciones para todas las variables numéricas.

Ilustración 8 Diferencia en correlaciones tras imputación por interpolación



Las diferencias son muy pequeñas (en la mayoría de los casos menores a **±0.05**), lo cual indica que la interpolación no altera significativamente las relaciones internas entre variables, incluso en variables con porcentajes moderados de valores faltantes. Solo se observan ligeros cambios localizados en pares específicos (por ejemplo, Gini–industria o energía fósil–CO_2), lo cual es esperable al imputar valores intermedios dentro de series temporales con variabilidad alta.
En general, el patrón de correlaciones del conjunto de datos se mantiene estable, lo que refuerza la validez de la interpolación aplicada.

- Validación de Calidad Gold Standard

La validación de la interpolación se realizó utilizando un procedimiento basado en el principio de *gold standard (amputation and imputation validation)*, que consiste en comparar un método estimado contra valores reales y conocidos, considerados como la referencia absoluta. Bajo este enfoque, se ocultó aleatoriamente el 10% de los datos verdaderos de cada variable asegurando que fueran observaciones completas y no originalmente faltantes y posteriormente se reconstruyeron mediante interpolación lineal país-año. Estos valores interpolados se compararon con los valores reales ocultos utilizando dos métricas estándar de evaluación: **RMSE (Root Mean Square Error)**, que cuantifica el error cuadrático medio, siendo más sensible a desviaciones grandes; y **MAE (Mean Absolute Error)**, que mide el error absoluto promedio, proporcionando una medida directa de la precisión. Este procedimiento constituye una validación tipo *Gold standard* porque la interpolación se evalúa contra datos verdaderos y no contra aproximaciones o promedios, permitiendo medir de forma objetiva cuán fiel es la estimación respecto a la realidad observada.[30]

RMSE (Root Mean Square Error): mide el error promedio cuadrático.

MAE (Mean Absolute Error): mide el error promedio absoluto.

Los resultados muestran que variables con escalas pequeñas o normalizadas, como Gini, IDH y CO_2 per cápita, presentan errores muy bajos, lo que indica que la interpolación reconstruye adecuadamente la tendencia temporal sin introducir distorsiones. En variables de gran magnitud como PIB per cápita, CO_2 total y uso de energía, los errores parecen altos en términos absolutos, pero son coherentes con sus propias escalas y reflejan una buena aproximación relativa. En conjunto, los valores obtenidos confirman que la interpolación lineal país-año es estable, preserva la forma original de las series y mantiene la coherencia estadística del conjunto de datos.

Tabla 4 Métricas de evaluación de variables tras imputación

	Métrica	
Variable	RMSE	MAE
Gini	0,150654	0,101389
Pib per cápita	614.807,401	317.831,595
IDH	0,003399	0,002218
CO2 per capita	0,078366	0,043877
CO2 total	2.136,05	1.028,79
Uso de energia	42,59	27,65

3. OBJETIVO 2: Analizar la influencia de factores socioeconómicos en las emisiones de CO_2 per cápita mediante técnicas de análisis exploratorio de datos (AED).

Una vez consolidada y depurada la base de datos, se realizó un análisis exploratorio orientado a examinar la relación entre los factores socioeconómicos y las emisiones de CO_2 per cápita. Esta etapa permitió identificar patrones, tendencias y posibles asociaciones entre variables, aportando criterios relevantes para las decisiones tomadas posteriormente en el proceso de modelamiento.

3.2 Análisis univariado

Se realizó un análisis univariado de las variables numéricas mediante histogramas y diagramas de caja con el objetivo de evaluar la distribución, dispersión y presencia de valores atípicos.
Los histogramas permitieron identificar asimetrías en variables como PIB per cápita, consumo y uso de energía, lo cual sugiere distribuciones no normales, comunes en indicadores socioeconómicos entre países. Por su parte, los diagramas de caja evidenciaron la existencia de valores extremos, especialmente en variables relacionadas con emisiones y población, reflejando la heterogeneidad estructural entre países.

3.3 Análisis bivariado

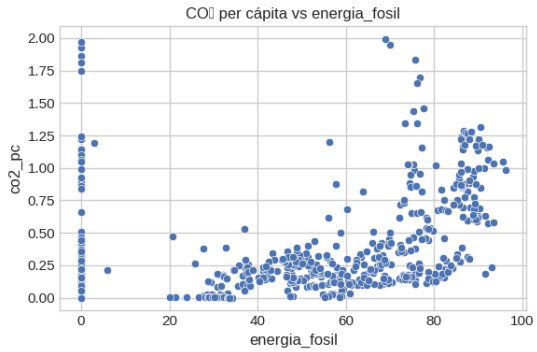
Para examinar las relaciones preliminares entre las emisiones de CO_2 per cápita y los factores

socioeconómicos, se realizó un análisis bivariado mediante diagramas de dispersión y una matriz de correlación. Esta etapa permitió identificar patrones, tendencias y posibles asociaciones entre variables, sin asumir relaciones causales directas.

Relación entre CO_2 per cápita y variables energéticas

Los gráficos de dispersión muestran una relación positiva clara entre el CO_2 per cápita y el uso de energía, así como con la participación de energía fósil. En ambos casos, a medida que aumenta el consumo energético, especialmente aquel basado en fuentes fósiles, se observa un incremento en las emisiones per cápita. No obstante, la dispersión de los datos sugiere que esta relación no es estrictamente lineal y presenta heterogeneidad entre países, lo que podría estar asociado a diferencias en eficiencia energética, estructura productiva o matriz energética.

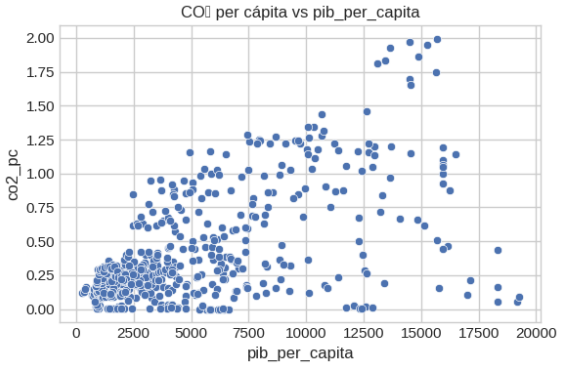
Ilustración 9 Gráfico de dispersión CO2 pc vs energía fósil



Relación entre CO_2 per cápita y nivel de desarrollo económico

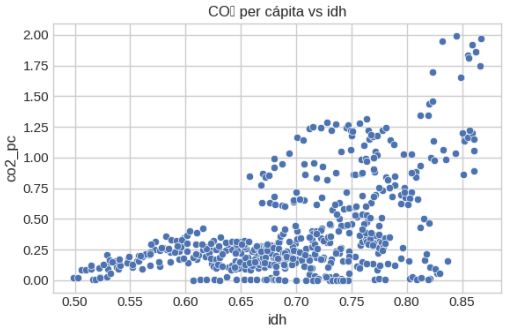
El análisis entre PIB per cápita y CO_2 per cápita evidencia una relación positiva moderada. Países con mayores niveles de ingreso tienden a presentar mayores emisiones per cápita, aunque se observa una dispersión considerable, lo que indica que el crecimiento económico no siempre se traduce proporcionalmente en mayores emisiones. Esto sugiere la posible influencia de factores como políticas ambientales, innovación tecnológica o transición energética.

Ilustración 10 Gráfico de dispersión CO2 pc vs PIB pc



De manera consistente, el Índice de Desarrollo Humano (IDH) también presenta una relación positiva con el CO_2 per cápita. A mayores niveles de desarrollo humano, se observan mayores emisiones, lo que refuerza la idea de que los procesos de desarrollo y bienestar suelen estar acompañados de una mayor demanda energética.

Ilustración 11 Gráfico de dispersión C02 pc Vs IDH



Relación entre CO_2 per cápita, desigualdad y estructura poblacional

La relación entre CO_2 per cápita y el coeficiente de Gini no muestra un patrón lineal claro. Visualmente se aprecia una dispersión amplia, lo que sugiere que la desigualdad del ingreso, por sí sola, no explica directamente las emisiones per cápita, aunque podría interactuar con otras variables socioeconómicas.

En cuanto a la población urbana, se observa una relación positiva con las emisiones per cápita, aunque con alta variabilidad. Países con mayor proporción de población urbana tienden a presentar mayores emisiones, posiblemente asociadas a patrones de consumo, transporte e industrialización propios de entornos urbanos.

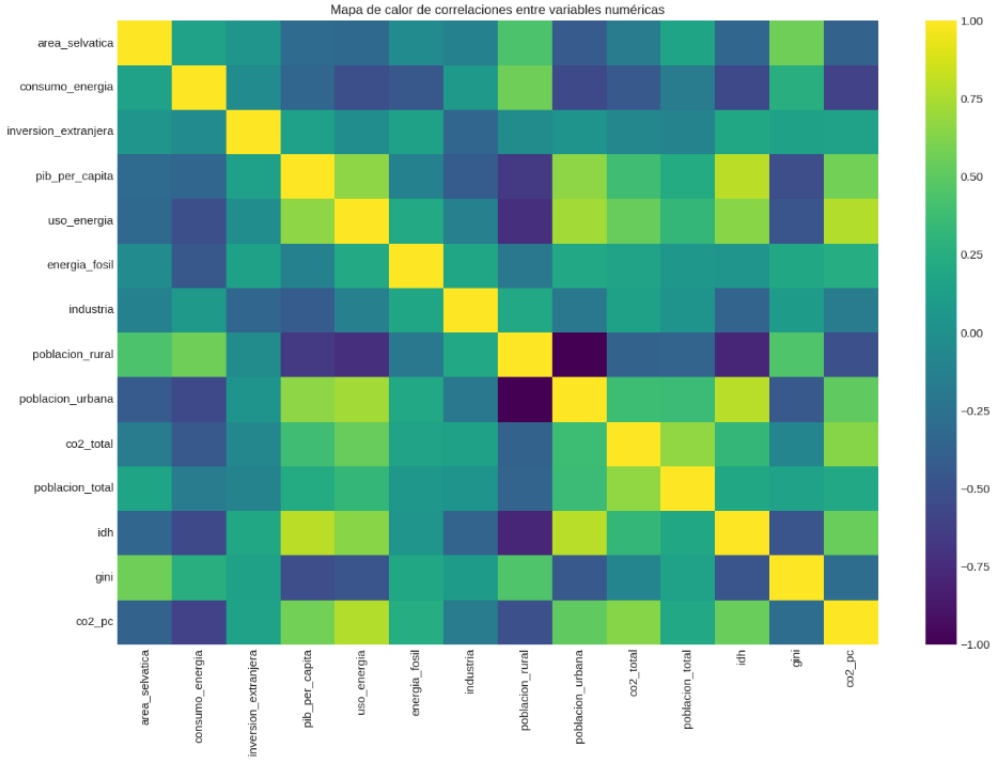
3.4 Análisis de la matriz de correlación

La matriz de correlación confirma los patrones observados en los gráficos de dispersión. El CO_2 per cápita presenta correlaciones positivas relativamente altas con variables como uso de energía,

PIB per cápita, IDH, población urbana y CO_2 total. En contraste, se observan correlaciones más débiles o negativas con variables como población rural y área selvática, lo cual sugiere que contextos menos urbanizados o con mayor cobertura natural tienden a presentar menores emisiones per cápita.

Asimismo, se identifican correlaciones altas entre algunas variables explicativas, particularmente entre PIB per cápita, uso de energía, IDH y CO_2 total, lo que indica posibles problemas de multicolinealidad que deberán ser evaluados en etapas posteriores mediante técnicas como el VIF. En la siguiente sección se reportan los resultados del análisis VIF aplicado al conjunto de variables del modelado.

Ilustración 12 Mapa de calor de correlaciones entre variables numéricas



Los resultados del análisis exploratorio de datos permitieron identificar patrones relevantes y asociaciones preliminares entre las emisiones de CO_2 per cápita y variables socioeconómicas, energéticas y demográficas. Estos hallazgos sirven como insumo para la siguiente etapa del estudio, orientada al desarrollo de modelos analíticos que permitan evaluar de manera más rigurosa dichas relaciones.

4. OBJETIVO 3: Desarrollar modelos que permitan identificar relaciones entre los factores socioeconómicos y emisiones de CO_2 per cápita.

La estrategia de modelado adoptada integra dos perspectivas complementarias: primero, modelos econométricos de datos de panel, efectos fijos (FE) y efectos aleatorios (RE), que permiten controlar la heterogeneidad no observada entre países y identificar asociaciones estructurales entre los factores socioeconómicos y las emisiones de CO_2 per cápita; y segundo, cinco algoritmos de aprendizaje automático, Random Forest, XGBoost, K-Nearest Neighbors (KNN), Red Neuronal Multicapa (MLP) y Support Vector Regression (SVR), que buscan maximizar la capacidad predictiva y capturar relaciones no lineales entre las variables. La coherencia de esta dualidad metodológica radica en combinar capacidad explicativa y predictiva para analizar los factores socioeconómicos asociados a las emisiones de CO_2 per cápita, aportando evidencia útil para comprender una dimensión relevante de la desigualdad climática en América Latina.

La variable dependiente en todos los modelos es emisiones de CO_2 per cápita, en toneladas por habitante. Las variables explicativas empleadas en los modelos econométricos y de machine learning comprenden: PIB per cápita, Índice de Desarrollo Humano, inversión extranjera directa, consumo de energía eléctrica per cápita, participación de energía de fuentes fósiles, proporción de población urbana, población total, participación del sector industrial en el PIB, coeficiente de Gini imputado y área selvática. La muestra efectiva empleada en los modelos fue de 527 observaciones país-año, obtenida tras aplicar la restricción de disponibilidad completa en todas las variables del conjunto final.

4.1 Construcción del dataset para el modelado

Con base en los resultados del análisis exploratorio de datos, se construyó un subconjunto de variables relevantes para el desarrollo de los modelos analíticos. Este dataset integra variables socioeconómicas, energéticas, demográficas y ambientales, junto con la variable objetivo de emisiones de CO_2 per cápita. El conjunto final para modelado cuenta con 527 observaciones para 17 países, garantizando consistencia temporal y ausencia de valores faltantes, lo que permite avanzar hacia etapas de diagnóstico y ajuste de modelos.

Las variables incluidas representan distintos ejes explicativos del fenómeno analizado. Se consideran factores económicos (PIB per cápita, inversión extranjera, índice de Gini), energéticos (consumo de energía, uso de energía, participación de energía fósil), demográficos (población urbana y rural), productivos (industria), ambientales (área selvática) y de desarrollo humano (IDH).

La variable dependiente del modelo corresponde a las emisiones de CO_2 per cápita, utilizada como indicador del impacto ambiental asociado a las dinámicas socioeconómicas.

4.2 Evaluación de Multicolinealidad

El Factor de Inflación de la Varianza (VIF) fue calculado con el objetivo de examinar la presencia de multicolinealidad entre las variables explicativas seleccionadas. Este indicador permitió identificar el nivel de dependencia lineal entre los predictores y evaluar posibles efectos sobre la estabilidad e interpretación de los modelos econométricos posteriores. Los resultados muestran

valores de VIF extremadamente elevados para población urbana (VIF ≈ 393,96) y población rural (VIF ≈ 65,16), evidenciando una alta colinealidad entre ambas variables. En contraste, el resto de las variables presenta valores moderados o bajos, como IDH (VIF ≈ 5,04), PIB per cápita (VIF ≈ 3,97), uso de energía (VIF ≈ 2,80) y consumo de energía (VIF ≈ 2,26), los cuales se consideran aceptables para el modelado. Con base en estos resultados, se justifica la exclusión de una de las variables poblacionales para evitar redundancia estadística en el análisis.

4.3 Selección de Modelos

4.3.1 Modelos Econométricos

Dada la estructura panel del conjunto de datos, con observaciones anuales por país, se optó por un modelo de datos panel con efectos fijos como enfoque principal para analizar la relación entre factores socioeconómicos y las emisiones de CO_2 per cápita. Este modelo permite controlar la heterogeneidad no observada específica de cada país y aislar la variación temporal intrapaís. Como ejercicio de contraste se estimó un modelo de efectos aleatorios y se aplicó la prueba de Hausman, mientras que modelos de aprendizaje automático se emplearon de forma complementaria para explorar posibles relaciones no lineales.

Justificación del uso de modelos econométricos dentro de la ciencia de datos

Si bien el aprendizaje automático es una herramienta relevante dentro de la ciencia de datos, su uso no es obligatorio ni excluyente. En estudios cuyo objetivo principal es explicar relaciones estructurales, más que realizar predicciones puras, los modelos econométricos ofrecen ventajas significativas en términos de interpretabilidad, control de heterogeneidad no observada y consistencia estadística. Teniendo en cuenta que el conjunto de datos presenta una estructura panel país–año para el periodo 1990–2020. Se opta el uso de datos panel, dado que permite capturar simultáneamente la variación temporal y transversal, mejorando la eficiencia de la estimación y controlando la heterogeneidad no observada constante en el tiempo. Este enfoque es ampliamente recomendado en estudios empíricos de economía del desarrollo y economía ambiental.[31]

Siguiendo la formalización propuesta por Baltagi (2008) y Wooldridge (2010)[32], el modelo panel general se expresa como:

$$CO_2\_pc(i,t) = \beta' X(i,t) + \alpha\_i + \gamma\_t + \varepsilon(i,t)$$

donde CO_2_pc(i,t) representa las emisiones de CO_2 per cápita del país i en el año t; X(i,t) denota el vector de variables explicativas socioeconómicas; α_i son los efectos fijos por entidad (país), que absorben toda la variación no observada constante en el tiempo; γ_t son los efectos temporales que capturan choques comunes a todos los países en un año determinado; y ε(i,t) es el término de error idiosincrático.

La especificación del vector de covariables X(i,t) incorpora ocho variables explicativas agrupadas en tres dimensiones analíticas: (i) variables energéticas, que incluyen el consumo de energía por unidad de población y la participación de combustibles fósiles en la matriz energética; (ii) variables económicas y de desarrollo, representadas por el PIB per cápita, el Índice de Desarrollo Humano (IDH), la participación de la industria en el PIB y la inversión extranjera directa; y (iii) variables demográfico-estructurales, conformadas por la tasa de urbanización y el volumen de la población total. Adicionalmente, se estimaron modelos extendidos que incorporan el índice de Gini y el área

selvática con fines de robustez metodológica.

4.3.1.1 Estrategia de modelado secuencial y justificación metodológica

La estimación de los modelos se realiza de manera progresiva, con el objetivo de garantizar rigurosidad econométrica, interpretabilidad y robustez de los resultados. Esta estrategia evita la inclusión simultánea de todas las variables disponibles y prioriza modelos parsimoniosos, alineados con la literatura especializada en análisis de datos panel y estudios socioambientales.
Se estima primero un modelo econométrico de datos panel con efectos aleatorios, utilizado únicamente como modelo de contraste metodológico, y no como enfoque principal de análisis. Su función es permitir la comparación con el modelo de efectos fijos y evaluar la consistencia de los estimadores bajo supuestos alternativos sobre la relación entre los efectos no observados y las variables explicativas, luego se estima un modelo de efectos fijos (Fixed Effects – FE) para garantizar control de heterogeneidad estructural e interpretabilidad; posteriormente, se emplea un modelo de aprendizaje automático Random Forest (RF) como complemento para capturar no linealidades e interacciones y para robustecer el análisis predictivo.
Adicionalmente, se especifica de manera detallada la lógica de selección e incorporación progresiva de variables: no se incluyen todas las variables disponibles desde el inicio, debido a consideraciones de calidad de datos (cobertura y faltantes), coherencia conceptual, endogeneidad potencial, colinealidad y necesidad de mantener interpretabilidad para fines de discusión de desigualdad climática.

4.3.1.2 Enfoque metodológico y justificación de la estrategia secuencial FE , RF

El estudio se enmarca en un enfoque cuantitativo y comparativo a nivel país, apoyado en técnicas econométricas y de ciencia de datos. La estrategia FE, RF responde a tres criterios: (i) control de heterogeneidad estructural no observable entre países, (ii) interpretabilidad y coherencia con la literatura de economía ambiental y justicia climática, y (iii) robustez predictiva y capacidad de explorar relaciones no lineales.
Primero, FE permite estimar asociaciones intrapaís en el tiempo, controlando factores inobservables constantes (por ejemplo, historia energética, estructura productiva y rasgos institucionales). Esto reduce sesgo por variables omitidas cuando esos factores se correlacionan con los regresores. Segundo, RF complementa el análisis al modelar relaciones no lineales y altas interacciones sin suponer forma funcional; además, ofrece medidas de importancia de variables y desempeño predictivo robusto mediante ensambles.

4.3.1.3 Modelo de efectos fijos

El modelo de efectos fijos (Fixed Effects, FE) se adopta como el núcleo explicativo del presente estudio, dada su capacidad para controlar la heterogeneidad no observable entre países. Su elección responde tanto a fundamentos teóricos como a evidencia empírica. Desde el punto de vista teórico, la existencia de factores no observados como la cultura energética, la trayectoria histórica de las políticas ambientales o las estructuras del mercado energético que pueden estar correlacionados con las variables explicativas, implica que el estimador de efectos fijos es consistente, a diferencia de otras alternativas como efectos aleatorios.[32] Esta elección se respalda mediante la aplicación

de la prueba de Hausman, cuyos resultados se presentan en la subsección correspondiente, confirmando la pertinencia del enfoque de efectos fijos.

Buscando obtener estimaciones robustas, se emplearon errores estándar agrupados por entidad, lo que permite corregir posibles problemas de heterocedasticidad y correlación serial dentro de cada país. Asimismo, se incorporaron efectos fijos tanto por entidad como por período, lo cual garantiza el control de factores inobservables específicos de cada país, así como de choques macroeconómicos comunes en el tiempo, tales como variaciones en los precios internacionales de la energía o crisis económicas globales.

La muestra efectiva utilizada en la estimación está compuesta por 527 observaciones correspondientes a 17 países, lo que permite explotar adecuadamente la variabilidad temporal y transversal de los datos.

La Tabla 5 presenta los resultados completos de la estimación del modelo de efectos fijos base, incluyendo los coeficientes estimados, errores estándar agrupados, estadísticos t, y valores p, facilitando la interpretación estadística y económica de los resultados obtenidos.

Tabla 5 Resultados del Modelo de Efectos fijos- Especificación base

Variable	Parámetro	Error Estándar	T-estadístico	P-valor
Constante	2.4524	1.1518	2.1292	0.0338**
PIB per cápita	2.406e-05	1.38e-05	1.7436	0.0819*
IDH	0.5148	1.6742	0.3075	0.7589
Inversión extranjera directa	0.0079	0.0074	1.0712	0.2846
Consumo de energía	**-0.0114**	0.0051	-2.2255	**0.0265****
Energía fósil	-0.0012	0.0010	-1.2711	0.2043
Población urbana	-0.0276	0.0114	-2.4177	**0.0160****
Población total	7.916e-10	2.286e-09	0.3318	0.7402
Industria	-0.0110	0.0068	-1.6204	0.1058

Nota: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$. Errores estándar agrupados por país. Efectos fijos por país y período incluidos. N = 527 observaciones, 17 países. R^2 global = 0.3070.
Fuente: elaboración propia

Los resultados de la estimación por Efectos Fijos (FE) revelan que, tras controlar la heterogeneidad estructural inobservable de los países y los choques temporales, el consumo de energía y la urbanización emergen como los predictores con mayor robustez estadística. Este modelo reporta un R^2 global de 0.3070, lo cual es consistente con estudios de panel en economías emergentes donde la variabilidad de las emisiones depende de factores institucionales y tecnológicos no siempre capturados por variables macroeconómicas agregadas.

El consumo de energía presentó un coeficiente de $\beta = -0.0114$ ($p = 0.0265$) indicando que, dentro de cada país a lo largo del tiempo, incrementos en el consumo energético se asocian con variaciones negativas en las emisiones per cápita. Este resultado, podría reflejar que los períodos de mayor consumo energético coinciden con procesos de diversificación hacia fuentes menos

intensivas en carbono en algunos países de la muestra, aunque esta hipótesis requeriría análisis complementario por subperíodos. De acuerdo con reportes recientes de la Agencia Internacional de Energía (IEA)[33], varios países de la región han incrementado la participación de fuentes renovables (hidroeléctrica, solar y eólica) en su matriz primaria. La relación observada es consistente con procesos de transición y eficiencia energética presentes en varios países de la región, donde incrementos en el consumo de energía no necesariamente se traducen en mayores emisiones cuando parte de la demanda es cubierta mediante fuentes menos intensivas en carbono o mediante mejoras en eficiencia dentro de los procesos productivos.

La población urbana mostró un coeficiente de $\beta = -0.0276$($p = 0.0160$). Este signo negativo se asocia con las economías de densidad [34]. La concentración de la población en núcleos urbanos facilita el acceso a infraestructuras más eficientes, transporte público masivo y una gestión de servicios que reduce la huella de carbono per cápita en comparación con la expansión rural o suburbana dispersa. Este dato sugiere que, para América Latina, la urbanización está actuando como un catalizador de mejoras tecnológicas y no solo como un factor de presión ambiental.

El PIB per cápita presentó una relación positiva ($\beta = 2.406 \times 10^{-5}$) y significativa al 10% ($p = 0.0819$). Este coeficiente, aunque marginal, confirma que el crecimiento económico en la región sigue manteniendo una naturaleza intensiva en recursos, siendo característico de economías en etapas de desarrollo medio, donde el incremento en la producción agrícola e industrial aún depende de combustibles fósiles, situando a la región en la fase ascendente de la Curva de Kuznets Ambiental, donde el ingreso y la degradación ambiental crecen de forma simultánea.

Variables como el IDH, la Inversión Extranjera Directa (IED) y el sector industrial no mostraron significancia estadística bajo la especificación de efectos fijos. Según Wooldridge (2010) [32], esto no implica que carezcan de importancia, sino que su variación intrapaís durante el periodo analizado no fue lo suficientemente fuerte como para explicar cambios en las emisiones por encima del efecto del PIB o la urbanización. Por ejemplo, la IED en la región ha migrado gradualmente de sectores extractivos a servicios y tecnología, lo que diluye su impacto directo sobre las emisiones de co2 en el corto plazo.

4.3.1.4 Modelo de efectos aleatorios

El modelo de efectos aleatorios (Random Effects, RE) se estima con el objetivo de servir como referencia metodológica para la aplicación de la prueba de Hausman. Diferente al modelo de efectos fijos, esta especificación asume que la heterogeneidad no observada específica de cada país (α_i) es independiente de las variables explicativas incluidas en el modelo. Bajo este supuesto, el estimador de efectos aleatorios resulta más eficiente; sin embargo, si dicha independencia no se cumple situación altamente probable en el contexto socioeconómico y ambiental de América Latina, el estimador pierde consistencia y, en consecuencia, su validez metodológica se ve comprometida.

La estimación se realizó utilizando errores estándar agrupados por entidad para controlar posibles problemas de heterocedasticidad y correlación intragrupo. La muestra analizada comprende 527 observaciones correspondientes a 17 países. Los resultados completos se presentan en la Tabla 6.

Tabla 6 Resultados del Modelos de efectos aleatorios- Especificación base

Variable	Parámetro	Error Estándar	T-estadístico	P-valor
Constante	1.1541	0.6740	1.7124	**0.0874***
PIB per cápita	2.599e-05	1.008e-05	2.5783	**0.0102****
IDH	1.5928	1.8060	1.9762	**0.0487****
Inversión extranjera directa	0.0067	0.0060	1.1156	0.2651
Consumo de energía	-0.0099	0.0037	-2.6367	**0.0086*****
Energía fósil	0.0005	0.0004	1.4345	0.1520
Población urbana	-0.0227	0.0089	-2.5529	**0.0110****
Población total	1.961e-09	1.821e-09	1.0771	0.2819
Industria	-0.0103	0.0066	-1.5604	0.1193

Nota: *** p < 0.01; ** p < 0.05; * p < 0.10. Errores estándar agrupados por país. R^2= 0.5090

La especificación de Efectos Aleatorios (RE) demuestra una capacidad explicativa superior al modelo de efectos fijos en términos de varianza total, alcanzando un R^2 global de 0.5090 y un R^2 *overall* de 0.1588. Este ajuste sugiere que el modelo captura con mayor precisión la dinámica de las emisiones de Co2 per cápita al integrar tanto la variabilidad temporal interna de los países (*within*) como las diferencias estructurales transversales entre ellos (*between*). La robustez estadística del modelo queda validada por un estadístico F de 5.9206 (p < 0.001), lo que confirma la significancia conjunta de los determinantes seleccionados.

En comparación con la especificación de efectos fijos, el modelo RE identifica un mayor número de factores determinantes, aportando una visión más multidimensional del fenómeno.

El PIB per cápita mantiene un coeficiente positivo y significativo al 5% (β = 2.599; p = 0.0102). Esto puede explicarse porque, en gran parte de América Latina, las dinámicas de crecimiento económico todavía dependen de sectores con alta intensidad energética, particularmente actividades extractivas, industriales y de transporte, donde el uso de combustibles fósiles continúa teniendo un papel predominante dentro de los procesos de producción y expansión económica.

Por su parte, el IDH también presentó un efecto positivo y estadísticamente significativo sobre las emisiones de CO_2 per cápita (β = 1.5928; p = 0.0487), indicando que, en el contexto de economías emergentes, mayores niveles de desarrollo humano pueden estar acompañados de incrementos en el consumo energético y en la demanda de bienes y servicios. En América Latina, las mejoras en ingreso, acceso a infraestructura, movilidad y condiciones de vida suelen implicar una mayor presión sobre sistemas productivos y energéticos que todavía mantienen una dependencia importante de fuentes intensivas en carbono. En ausencia de transformaciones estructurales hacia modelos sostenibles de producción y consumo, estos procesos tienden a incrementar la presión ambiental y las emisiones asociadas al crecimiento económico y social. Por su parte, se observa una notable consistencia en los coeficientes negativos del Consumo de Energía (β= -0.0099; p = 0.0086) y la Población Urbana (β = -0.0227; p = 0.0110). El primero refuerza la evidencia de una transición gradual hacia matrices energéticas más limpias en la región, mientras que el segundo valido la hipótesis de las economías de densidad, donde la urbanización planificada permite eficiencias en el transporte y la infraestructura energética que mitigan la huella de carbono individual.

Por el contrario, variables como la Inversión Extranjera Directa (IED), el sector industrial y la población total no alcanzaron significancia estadística. Esta falta de robustez sugiere que, tras considerar la heterogeneidad entre naciones, su impacto en las emisiones es mediado por factores institucionales o tecnológicos que no se capturan de forma lineal en este modelo.
En conclusión, aunque el modelo de efectos aleatorios ofrece una mayor riqueza interpretativa y un ajuste estadístico superior, su validez final descansa en los resultados de la prueba de Hausman. Dado que el estadístico de Hausman obtenido previamente no permitió rechazar la hipótesis nula, se confirma que el modelo RE es el estimador más eficiente y consistente para describir los determinantes de las emisiones en el contexto analizado, superando las limitaciones del enfoque de efectos fijos.

4.3.1.5 Test de Hausman

Como parte del proceso de modelado econométrico, se aplicó la prueba de Jerry A. Hausman [35] con el objetivo de determinar cuál de las dos especificaciones de datos panel, efectos fijos (Fixed Effects, FE) o efectos aleatorios (Random Effects, RE), resultaba más adecuada para modelar la relación entre los factores socioeconómicos y las emisiones de CO_2 per cápita en América Latina. Desde el punto de vista metodológico, ambos enfoques permiten trabajar con datos longitudinales; sin embargo, difieren en la forma en que modelan la heterogeneidad no observada entre países. El modelo de efectos fijos asume que existen características estructurales propias de cada país que pueden estar correlacionadas con las variables explicativas, mientras que el modelo de efectos aleatorios asume que dichas diferencias individuales no están correlacionadas con los predictores incluidos en el modelo.

La prueba de Hausman funciona comparando los vectores de coeficientes estimados por ambos modelos. Su lógica consiste en evaluar si las diferencias entre los estimadores son sistemáticas o si, por el contrario, pueden atribuirse al comportamiento esperado bajo el supuesto de independencia. La hipótesis nula (H_0) plantea que el modelo de efectos aleatorios es consistente y eficiente, mientras que la hipótesis alternativa (H_1) indica que existe correlación entre los efectos individuales y las variables explicativas, en cuyo caso debe preferirse el modelo de efectos fijos.

Al realizar la prueba de especificación de Hausman para contrastar la consistencia de los estimadores de Efectos Aleatorios (RE) frente a los de Efectos Fijos (FE), se obtuvo un estadístico de **-72.6064** (gl = 8; p = 1.000). Desde una perspectiva asintótica, el estadístico de Hausman posee una distribución X^2, y, por definición, debe ser no negativo. No obstante, la obtención de valores negativos es un fenómeno documentado en la literatura econométrica aplicada (Hausman y Taylor, 1981; Schreiber, 2008), atribuido comúnmente a la falta de definición positiva de la diferencia entre las matrices de varianza-covarianzas [VFE - VRE] en muestras finitas o ante la presencia de una fuerte similitud entre los coeficientes estimados. Siguiendo el criterio de Baltagi (2021)[36] y la práctica estándar en software especializado como Stata, un valor negativo se interpreta como una incapacidad de rechazar la hipótesis nula (H_0), sugiriendo que los estimadores de RE son consistentes y, a priori, más eficientes.
A pesar de que el resultado formal de la prueba no rechaza la idoneidad del modelo de efectos aleatorios, el presente estudio adopta la estimación por Efectos Fijos (FE). Esta decisión se

fundamenta en el control de endogeneidad y Sesgo por Variables Omitidas. La premisa fundamental de los Efectos Aleatorios es la independencia entre los errores individuales y las variables explicativas ($cov(\alpha_i, x_{it})$=0). En el marco de las emisiones de CO2 y el desarrollo socioeconómico en América Latina, este supuesto es poco plausible. Factores invariantes en el tiempo, como la herencia institucional, la geografía climática y las trayectorias históricas de industrialización, están correlacionados con el PIB y la desigualdad. El modelo FE permite capturar esta heterogeneidad inobservable, mitigando el sesgo de variable omitida.
Por otro lado, dado que el análisis se centra en un conjunto específico de países latinoamericanos (unidades con nombre propio y características únicas) y no en una muestra aleatoria extraída de una población mayor, el modelo de efectos fijos es el enfoque teórico más robusto para modelar las relaciones intrínsecas de cada unidad.
Teniendo en cuenta lo anterior el modelo FE garantiza que los coeficientes de las variables socioeconómicas (como el Índice de Gini o la intensidad energética) reflejen cambios *intrapaís*, permitiendo conclusiones más precisas sobre cómo la evolución de la desigualdad interna afecta directamente la trayectoria de emisiones de cada nación.

4.3.1.6 Modelos de Robustez

Una extensión importante luego de aplicar el test de Hussman es evaluar la estabilidad de los resultados del modelo de efectos fijos base dado a su resultado no esperado con signo negativo y explorar el aporte marginal de variables adicionales, se estimaron dos especificaciones extendidas: el modelo de efectos fijos con inclusión del índice de Gini como proxy de desigualdad del ingreso, y el modelo de efectos fijos con área selvática como variable ambiental estructural. Estas variables se incorporaron en modelos separados y no de manera simultánea, evitando problemas de multicolinealidad, reducción innecesaria de la muestra efectiva por faltantes adicionales, y pérdida de interpretabilidad en los coeficientes.

4.3.1.7 Modelo de efectos fijos con desigualdad del ingreso (Gini)

La Tabla 7 reporta los resultados del modelo de efectos fijos extendido con la inclusión del índice de Gini imputado. La muestra efectiva se reduce a 527 observaciones en 17 países, dado que el índice de Gini presenta una mayor proporción de valores faltantes que las variables incluidas en el modelo base.

Tabla 7Modelo de efectos fijos con GINI

Variable	Coeficiente	Error Estándar	T-estadístico	P-valor
Constante	1.7757	0.9603	1.8490	0.0651*
PIB per cápita	2.188e-05	1.199e-05	1.8248	0.0687*
IDH	0.9523	1.5922	0.5981	0.5501
Inversión extranjera directa	0.0053	0.0059	0.8934	0.3721
Consumo de energía	-0.0139	0.0057	-2.4114	0.0163**
Energía fósil	-0.0013	0.0010	-1.3300	0.1842
Población urbana	-0.0352	0.0140	-2.5207	0.0120**
Población total	1.665e-09	2.249e-09	0.7405	0.4594
Industria	-0.0123	0.0074	-1.6735	0.0949*

Gini	0.0219	0.0150	1.4652	0.1435

Nota: *** p < 0.01; ** p < 0.05; * p < 0.10. Errores estándar agrupados por país. N = 527 observaciones, 17 países. R^2 global = 0.3413.
Fuente: elaboración propia

El modelo de efectos fijos con inclusión del índice de Gini presenta una mejora moderada en la capacidad explicativa respecto a la especificación base, alcanzando un R^2 de 0.3413 y un R^2 within de 0.2968. Esto indica que la incorporación de la desigualdad contribuye parcialmente a explicar la variabilidad temporal de las emisiones de CO_2 per cápita dentro de los países analizados. Asimismo, el estadístico F robusto (F = 2.5128; p = 0.0081) confirma la significancia global del modelo.

Los resultados muestran que el consumo de energía mantiene una relación negativa y estadísticamente significativa con las emisiones de CO_2 per cápita (β = -0.0139; p = 0.0163). coincidiendo con la especificación anterior y sugiere que, dentro del período analizado, parte del incremento en el consumo energético pudo estar asociado con procesos de transición hacia fuentes relativamente menos intensivas en carbono o mejoras en eficiencia energética en algunos países de América Latina.

La población urbana también presentó significancia estadística al 5 % (β = -0.0352; p = 0.0120), mostrando una asociación negativa con las emisiones de CO_2 per cápita. En el contexto regional, este comportamiento puede estar relacionado con dinámicas urbanas más concentradas, donde la densidad poblacional favorece economías de escala en infraestructura, sistemas de transporte colectivo y menores niveles de consumo energético per cápita frente a territorios con mayor dispersión poblacional. Por otra parte, el PIB per cápita continúa mostrando una relación positiva y significativa al 10% ($\beta = 2.188\times10^{-5}$; p = 0.0687), lo que sugiere que el crecimiento económico sigue vinculado al aumento de las emisiones en economías donde el desarrollo productivo mantiene dependencia de actividades intensivas en energía.

La variable industria presenta un coeficiente negativo y significancia marginal al 10% (β = -0.0123; p = 0.0949). Este dato, podría indicar ciertos procesos de modernización tecnológica o mejoras relativas en eficiencia dentro del sector industrial durante el período analizado.

En contraste, variables como el IDH, la inversión extranjera, la energía fósil, la población total y el índice de Gini imputado no presentan significancia estadística individual dentro del modelo. En particular, el coeficiente de Gini (β = 0.0219; p = 0.1435) muestra una relación positiva, aunque insuficiente para afirmar un efecto estadísticamente robusto sobre las emisiones de CO_2 per cápita bajo esta especificación.

El resultado del índice de Gini sugiere que, aunque la desigualdad podría influir indirectamente sobre los patrones de consumo y emisiones, su efecto no logra capturarse de manera significativa

una vez controladas las diferencias estructurales entre países y períodos mediante efectos fijos. Esto evidencia la complejidad de la relación entre desigualdad y sostenibilidad ambiental en el contexto latinoamericano.

4.3.1.8 Modelo de efectos fijos con área selvática

Tabla 8Modelo de efectos fijos con área selvática

Variable	Coeficiente	Error Estándar	T-estadístico	P-valor
Constante	0.9069	0.8447	1.0737	0.2835
PIB per cápita	1.737e-05	9.439e-06	1.8398	0.0664
IDH	1.2972	1.2553	1.0334	0.3019
Inversión extranjera directa	0.0082	0.0062	1.3350	0.1825
Consumo de energía	-0.0132	0.0060	-2.1889	0.0291
Energía fósil	0.0004	0.0012	0.3254	0.7451
Población urbana	-0.0330	0.0110	-2.9957	0.0029
Población total	2.457e-09	2.99e-09	0.8216	0.4117
Industria	0.0017	0.0053	0.3282	0.7429
Área selvática	0.0258	0.0098	2.6494	0.0083

Nota: *** p < 0.01; ** p < 0.05; * p < 0.10. Errores estándar agrupados por país. N = 527 observaciones, 17 países. R^2 global = 0.4084.
Fuente: elaboración propia

Los resultados del Modelo 4 evidencian que la especificación econométrica presenta significancia global, dado que el estadístico F y su versión robusta registran valores p inferiores a 0.05, indicando que el conjunto de variables explicativas contribuye significativamente a explicar el comportamiento de las emisiones de CO_2 per cápita.
El coeficiente de determinación within ($R^2 = 0.1358$) indica que el modelo logra explicar aproximadamente el 13.58% de la variación interna de las emisiones entre los países y períodos analizados. Aunque el nivel explicativo es moderado, siendo consistente con estudios de datos panel en temas ambientales y macroeconómicos, donde las emisiones dependen de múltiples factores estructurales y dinámicos.
Entre las variables explicativas, el consumo de energía presenta un coeficiente negativo y estadísticamente significativo ($p = 0.0291$), lo cual evidencia una relación relevante con las emisiones de CO_2, confirmando que las variaciones en los niveles de consumo energético influyen directamente sobre el comportamiento ambiental de los países analizados.
La población urbana también presentó significancia estadística (p = 0.0029), evidenciando que los procesos de urbanización mantienen incidencia sobre las emisiones de CO_2 per cápita; consecuentemente, el coeficiente negativo (-0.0330) indica que, por cada incremento de una unidad en la población urbana, las emisiones tienden a disminuir en 0.0330 unidades, manteniendo constantes las demás variables del modelo, lo que refleja que, durante el período analizado, ciertos procesos de urbanización pudieron asociarse a mejoras relativas en eficiencia energética, concentración de infraestructura y optimización en la prestación de servicios urbanos.
Por su parte, la variable área selvática presenta un coeficiente positivo y estadísticamente significativo ($p = 0.0083$), evidenciando que, dentro del modelo de emisiones per cápita para

América Latina, la cobertura forestal no funciona únicamente como un sumidero natural de carbono, sino también como un espacio donde convergen procesos de transformación territorial y actividades de alta presión ambiental. Los resultados reflejan que los países con mayores extensiones de selva suelen concentrar dinámicas asociadas a deforestación, expansión agroindustrial, extracción minera y explotación de hidrocarburos, factores que incrementan las emisiones mediante cambios en el uso del suelo y degradación de biomasa.
Esta relación se observa con claridad en regiones de la Amazonía de Brasil, Bolivia y Perú, donde reportes recientes de monitoreo forestal (2025–2026) muestran que la degradación causada por incendios, tala selectiva y expansión de infraestructura extractiva ha reducido la capacidad de captura de carbono de los bosques[37]. En consecuencia, el área selvática termina funcionando como una variable de exposición frente a presiones económicas y territoriales, evidenciando que la sostenibilidad climática regional depende no solo de conservar grandes extensiones forestales, sino también de limitar la expansión de modelos extractivos sobre ecosistemas estratégicos

En contraste, variables como el índice de desarrollo humano (IDH), inversión extranjera, energía fósil, población total e industria no presentan significancia estadística individual dentro de esta especificación, lo que sugiere que sus efectos no resultan concluyentes bajo las condiciones del modelo estimado.
Finalmente, la prueba de poolability presenta un valor p de 0.0000, rechazando la hipótesis nula de homogeneidad y confirmando la existencia de efectos individuales significativos entre los países analizados, lo cual respalda la pertinencia del modelo de efectos fijos utilizado.

4.3.2 Modelos de Machine learning

4.3.2.1 Preparación de la base de datos y Proceso de entrenamiento

Se implementó una estrategia de aprendizaje automático supervisado bajo un enfoque de regresión para complementar el análisis econométrico y capturar relaciones no lineales entre los factores socioeconómicos y las emisiones de CO_2 per cápita. Para ello, se seleccionaron cinco algoritmos ampliamente utilizados en problemas de datos tabulares estructurados: **Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Support Vector Regression (SVR) y Perceptrón Multicapa (MLP)**. La selección de estos modelos responde a que representan enfoques de aprendizaje complementarios: modelos basados en árboles de decisión y ensamble (RF y XGBoost), métodos basados en proximidad entre observaciones (KNN), técnicas fundamentadas en hiperplanos óptimos y funciones kernel (SVR), y arquitecturas neuronales capaces de aproximar relaciones altamente no lineales (MLP).

La elección de estos algoritmos se encuentra respaldada por la literatura en ciencia de datos aplicada a datos tabulares. Estudios como los de Trevor Hastie, Robert Tibshirani y Jerome Friedman señalan que el uso de múltiples familias de modelos permite capturar tanto relaciones lineales como estructuras complejas y no paramétricas, mejorando la capacidad de descubrimiento de patrones en problemas multivariados.

Para esta etapa se construyó una base de datos específica para machine learning mediante la selección de observaciones completas, eliminando registros con valores faltantes. Como resultado,

se obtuvo una muestra efectiva de 527 observaciones, compuesta por 10 variables explicativas y una variable (objetivo continua) correspondiente a las emisiones de CO_2 per cápita (*co2_pc*). Las variables predictoras incluidas fueron: PIB per cápita, Índice de Desarrollo Humano (IDH), inversión extranjera directa, consumo de energía, participación de energía fósil, población urbana, población total, participación industrial, coeficiente de Gini imputado y área selvática.

Posteriormente, el conjunto de datos fue dividido mediante una partición temporal (*temporal hold-out*), utilizando las observaciones correspondientes al período 1990–2015 para el entrenamiento de los modelos (442 observaciones) y el período 2016–2020 para la evaluación externa (85 observaciones). Esta estrategia permite preservar la secuencia cronológica de los datos y evita que los modelos utilicen información futura durante el proceso de aprendizaje, reduciendo así el riesgo de *data leakage* y proporcionando una evaluación más realista de la capacidad predictiva en contextos longitudinales.

La adopción de una partición temporal resulta particularmente apropiada en bases de datos panel con estructura país-año, ya que reproduce el escenario real de predicción y evita estimaciones excesivamente optimistas derivadas de particiones aleatorias

Antes del entrenamiento, las variables explicativas fueron sometidas a un proceso de estandarización, ajustando los parámetros únicamente sobre el conjunto de entrenamiento y aplicándolos posteriormente al conjunto de prueba. Este procedimiento transforma las variables para que presenten media cero y desviación estándar unitaria, mejorando la estabilidad numérica y el rendimiento de algoritmos sensibles a la escala, particularmente K-Nearest Neighbors (KNN), Support Vector Regression (SVR) y Redes Neuronales Multicapa (MLP).

Finalmente, todos los modelos fueron optimizados mediante Grid Search con validación cruzada de cinco pliegues aplicada exclusivamente sobre el conjunto de entrenamiento. Este procedimiento permitió identificar la combinación óptima de hiperparámetros para cada algoritmo, garantizando que el desempeño obtenido no dependiera de configuraciones arbitrarias y evitando la incorporación de información del período de evaluación durante la etapa de ajuste.

4.3.2.2 Random Forest

El primer algoritmo implementado dentro del componente de aprendizaje automático fue Random Forest, un método de ensamble supervisado desarrollado por Leo Breiman, cuyo principio consiste en combinar múltiples árboles de decisión construidos sobre subconjuntos aleatorios de datos y variables predictoras, con el fin de mejorar la precisión del modelo y reducir la varianza asociada a un árbol individual. A diferencia de los modelos lineales tradicionales, este algoritmo permite capturar interacciones complejas, efectos jerárquicos y relaciones no lineales entre predictores, características particularmente relevantes cuando se analizan fenómenos multidimensionales como las emisiones de CO_2 per cápita, donde variables económicas, energéticas, sociales y territoriales interactúan simultáneamente.

Desde una perspectiva matemática, la predicción del modelo puede representarse como el

promedio de las predicciones generadas por cada árbol del ensamble:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x)$$

donde $\hat{y}$representa la predicción final del modelo, Bcorresponde al número total de árboles construidos y $T_b(x)$representa la predicción generada por el árbol ba partir del vector de características x. Esta arquitectura permite que cada árbol capture patrones distintos dentro de la base de datos, mientras que el promedio final reduce la sensibilidad del modelo frente a observaciones particulares y mejora su capacidad de generalización. Su incorporación respondió a la necesidad de complementar los resultados obtenidos mediante econometría de panel. Mientras los modelos de efectos fijos permiten identificar relaciones estructurales lineales y significancia estadística longitudinal, Random Forest permite detectar dependencias no lineales y efectos de interacción que pueden permanecer ocultos bajo especificaciones paramétricas. Esta característica adquiere particular importancia en el análisis de las emisiones de carbono en América Latina, donde variables como el crecimiento económico, la intensidad energética, la desigualdad distributiva y los cambios demográficos presentan dinámicas que no siempre siguen relaciones lineales entre países y períodos de tiempo. El entrenamiento del modelo se realizó sobre una muestra de 442 observaciones correspondientes al período 1990–2015, utilizando diez variables explicativas previamente seleccionadas: PIB per cápita, índice de desarrollo humano, inversión extranjera directa, consumo energético, participación de energía fósil, población urbana, población total, participación industrial, coeficiente de Gini imputado y área selvática. La diversidad de estas variables permitió representar simultáneamente dimensiones económicas, sociales, energéticas y ecológicas asociadas a las emisiones de CO_2 per cápita en América Latina.

Con el fin de evitar problemas de fuga de información (*data leakage*), el conjunto de datos fue dividido mediante una partición temporal, utilizando las observaciones comprendidas entre 1990 y 2015 para el entrenamiento del modelo y las observaciones correspondientes al período 2016–2020 para la evaluación fuera de muestra. Esta estrategia preserva la secuencia cronológica de los datos y garantiza que el modelo no incorpore información futura durante el proceso de aprendizaje.

La arquitectura del modelo fue optimizada mediante una búsqueda exhaustiva de hiperparámetros utilizando Grid Search con validación cruzada de cinco pliegues aplicada exclusivamente sobre el conjunto de entrenamiento. El procedimiento evaluó 108 configuraciones posibles, generando un total de 540 entrenamientos independientes. Esta estrategia permitió seleccionar la configuración con mejor desempeño promedio durante la validación, reduciendo el riesgo de sobreajuste y favoreciendo una mayor capacidad de generalización del modelo.

La configuración óptima identificada estuvo compuesta por 100 árboles de decisión, una profundidad máxima de 8 niveles, un criterio de división mínimo de 2 observaciones por nodo y un mínimo de 1 observación por hoja terminal. Estos parámetros correspondieron a la combinación que obtuvo el mejor desempeño dentro del conjunto de configuraciones evaluadas. La profundidad máxima de 8 niveles permitió controlar la complejidad de los árboles individuales, favoreciendo un equilibrio entre capacidad predictiva y generalización.

Importancia de variables – Random forest

Una vez entrenado y optimizado el modelo Random Forest, se realizó un análisis de importancia de variables con el fin de identificar cuáles predictores aportaron más información al proceso de aprendizaje del algoritmo. Este análisis permite interpretar internamente la lógica del modelo y comprender qué factores tuvieron mayor influencia en la predicción de las emisiones de CO_2 per cápita. Para este estudio, esta etapa resulta particularmente relevante, ya que permite conectar el desempeño predictivo del algoritmo con la identificación sustantiva de los determinantes asociados a la desigualdad climática en América Latina.

En Random Forest, la importancia de una variable se calcula a partir de la reducción acumulada de impureza generada por dicha variable a través de todos los árboles del bosque. Matemáticamente, esta importancia puede representarse como la suma de las reducciones de error que produce una variable en cada nodo donde participa, normalizada respecto al total del modelo. En este sentido, una variable con mayor importancia es aquella que contribuye con mayor frecuencia y efectividad a reducir la incertidumbre durante el proceso de partición de los datos, permitiendo generar predicciones más precisas.

Para el presente estudio, la importancia de variables fue calculada utilizando el atributo *feature_importances_* del modelo optimizado, el cual consolida la contribución promedio de cada predictor a lo largo de los 100 árboles seleccionados durante la fase de entrenamiento.

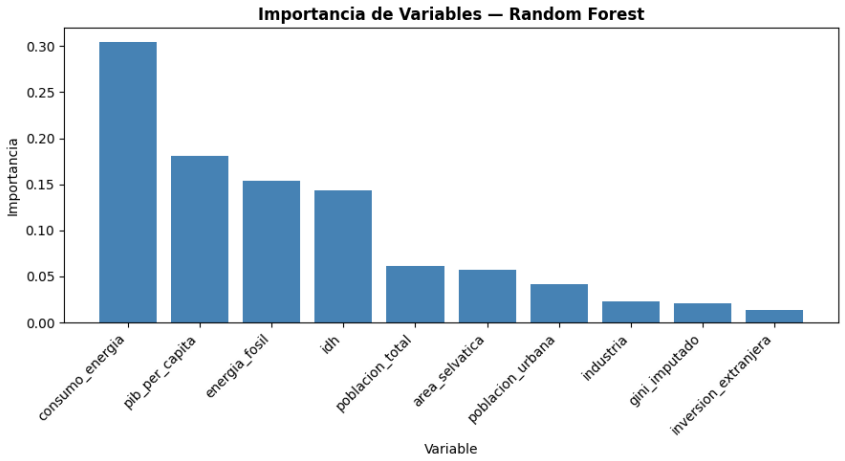


Ilustración 13 Importancia de variables Random Forest

Tabla 9 Importancia de variables-Random forest

Variable	Importancia
consumo_energia	0.304443

pib_per_capita	0.180832
energia_fosil	0.153965
idh	0.143493
Población_total	0.061332
area_selvatica	0.056750
Población_urbana	0.041437
industria	0.023420
gini	0.020505
Inversión_extranjera	0.013823

Los Los resultados del análisis de importancia de variables muestran que el consumo de energía fue el predictor con mayor contribución relativa dentro del modelo Random Forest, registrando una importancia de 0.3044. Le siguieron el PIB per cápita (0.1808), la participación de energía fósil (0.1540) y el Índice de Desarrollo Humano (0.1435). En conjunto, estas variables concentraron una proporción importante de la capacidad predictiva del modelo, lo que indica que los factores económicos, energéticos y de desarrollo humano contienen información relevante para la predicción de las emisiones de CO_2 per cápita en los países latinoamericanos analizados.

En un segundo nivel de importancia se ubicaron la población total (0.0613) y el área selvática (0.0568), seguidas por la población urbana (0.0414) y la participación del sector industrial (0.0234). Aunque estas variables presentaron una contribución menor en comparación con las variables líderes, su incorporación permitió capturar información complementaria asociada a características demográficas, territoriales y productivas de la región.

Por su parte, el coeficiente de Gini imputado (0.0205) y la inversión extranjera directa (0.0138) registraron los menores niveles de importancia relativa dentro del modelo. Este resultado sugiere que, para el conjunto de datos analizado y bajo la estructura predictiva de Random Forest, estas variables aportaron menos información incremental para la predicción de las emisiones de CO_2 per cápita que los factores relacionados con el consumo energético, el desarrollo económico y la estructura energética.

En términos generales, la distribución de importancias evidencia un predominio de variables asociadas al consumo de energía, las condiciones económicas y las características estructurales de los países. Estos resultados permiten identificar qué factores fueron más relevantes dentro del proceso predictivo del modelo, constituyendo una aproximación empírica a los patrones asociados con la desigualdad climática en América Latina, sin implicar relaciones de causalidad entre las variables analizadas.

4.3.2.3 Extreme Gradient Boosting (XGBoost)

Los resultados obtenidos con XGBoost evidencian que las emisiones de CO_2 per cápita en América Latina responden a una interacción simultánea entre factores económicos, energéticos, sociales y urbanos, más que a efectos aislados de una sola variable. El modelo logró identificar patrones no lineales y relaciones acumulativas entre las variables explicativas, reflejando la heterogeneidad estructural presente en la región, donde países con niveles similares de ingreso exhiben trayectorias de emisiones distintas debido a diferencias en intensidad energética, urbanización, matriz

energética y desigualdad social. Este comportamiento confirmó la necesidad de utilizar un modelo con capacidad para aprender relaciones complejas y dependencias sucesivas entre variables, superando las limitaciones de enfoques basados únicamente en proximidad entre observaciones similares.

A diferencia del modelo K-Nearest Neighbors (KNN), que genera predicciones a partir de la similitud local entre países con características socioeconómicas cercanas, el algoritmo Extreme Gradient Boosting (XGBoost) desarrolla un proceso de aprendizaje secuencial en el que cada nueva etapa corrige los errores producidos por las anteriores. Esta lógica de funcionamiento responde a la necesidad de capturar interacciones dinámicas entre variables económicas, sociales, energéticas y ambientales, permitiendo representar estructuras más complejas asociadas al comportamiento de las emisiones de CO_2 per cápita.

El modelo se fundamenta en la optimización de una función objetivo compuesta por un término de pérdida y un componente de regularización:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

En esta expresión, $l(y_i, \hat{y}_i)$representa la diferencia entre los valores observados y predichos, $f_t(x_i)$corresponde al árbol incorporado en la iteración t, y $\Omega(f_t)$representa el término de regularización encargado de controlar la complejidad del modelo. La combinación de estos componentes permitió reducir progresivamente el error predictivo mientras se limitaba el riesgo de sobreajuste, favoreciendo una mejor capacidad de generalización sobre datos no observados.

El ajuste y configuración del modelo reforzaron esta capacidad de aprendizaje acumulativo. El entrenamiento se realizó con 442 observaciones correspondientes al período 1990–2015 y se aplicó un procedimiento de búsqueda exhaustiva de hiperparámetros mediante validación cruzada de cinco pliegues. En total se evaluaron 243 combinaciones posibles, equivalentes a 1215 procesos internos de entrenamiento y validación, constituyendo uno de los procedimientos de optimización más amplios desarrollados dentro de los modelos analizados.

Durante este proceso se ajustaron parámetros relacionados con la capacidad de aprendizaje y generalización del algoritmo. Se evaluó el número de árboles secuenciales para determinar cuántas iteraciones eran necesarias en la corrección progresiva del error; igualmente, se ajustó la tasa de aprendizaje para controlar la velocidad con la que el modelo incorporaba la información aprendida en cada iteración. También se reguló la profundidad máxima de los árboles con el fin de evitar estructuras excesivamente complejas y reducir el riesgo de sobreajuste. De manera complementaria, se implementaron mecanismos de regularización mediante selección parcial de variables y uso aleatorio de observaciones durante el entrenamiento de cada árbol, disminuyendo la dependencia entre iteraciones consecutivas y fortaleciendo la capacidad de generalización.

La configuración con mejor desempeño estuvo compuesta por 100 árboles de decisión, una tasa de aprendizaje de 0.10 y una profundidad máxima de 6 niveles. Adicionalmente, el modelo utilizó el

80 % de las observaciones disponibles en cada iteración (subsample = 0.8) y una selección aleatoria del 70 % de las variables para la construcción de cada árbol (colsample_bytree = 0.7). Estos resultados muestran que el modelo alcanzó su mejor desempeño mediante un equilibrio entre capacidad de aprendizaje y mecanismos de regularización, permitiendo capturar patrones complejos sin incrementar excesivamente el riesgo de sobreajuste. Asimismo, la selección parcial de variables evidencia que las dimensiones económicas, sociales, energéticas y ambientales aportaron información relevante en diferentes etapas del proceso predictivo, reafirmando el carácter multidimensional de las emisiones de CO_2 per cápita en América Latina.

Importancia de variables en el modelo XGBoost

Una vez ajustado el modelo XGBoost, se realizó un análisis de importancia de variables para identificar cuáles predictores tuvieron mayor participación en la reducción del error durante el proceso de aprendizaje del algoritmo. Este análisis permitió complementar la capacidad predictiva del modelo con una interpretación de las variables que aportaron más información para la predicción de las emisiones de CO_2 per cápita en América Latina.

En XGBoost, la importancia de una variable se determina a partir de la contribución acumulada que esta realiza en las divisiones efectuadas a lo largo de los árboles secuenciales. Cada vez que una variable participa en una partición que mejora el desempeño predictivo del modelo y reduce el error de estimación, incrementa su nivel de importancia relativa dentro del proceso de aprendizaje. En consecuencia, las variables con mayores valores de importancia corresponden a aquellas que aportaron más información útil para la construcción de las predicciones.

El análisis fue realizado sobre los 100 árboles generados durante el entrenamiento del modelo. Los resultados evidenciaron que el consumo de energía presentó la mayor importancia relativa (0.2577), seguido por la participación de energía fósil (0.1703), el área selvática (0.1181) y el PIB per cápita (0.1105). Estos resultados indican que las variables asociadas al uso de energía, la composición de la matriz energética, las características ambientales y la actividad económica fueron las que más contribuyeron al proceso predictivo del modelo.

Asimismo, la relevancia conjunta del consumo energético y la participación de energía fósil sugiere que las características energéticas contienen una proporción importante de la información utilizada por el algoritmo para generar sus predicciones. De igual forma, la importancia observada para el área selvática y el PIB per cápita indica que los componentes territoriales y económicos también aportaron información relevante dentro de la estructura predictiva identificada por el modelo.

Tabla 10 Importancia de variables – XGBoost Regressor

Variable	Importancia
Consumo_energia	0.257669
Energía_fosil	0.170284
Área_selvatica	0.118060
Pib percapita	0.110497
gini	0.084228
idh	0.074147
Población_urbana	0.062511
industria	0.061657

Población total	0.048731
Inversion extranjera	0.012226

Fuente: Elaboración propia.

En un segundo nivel de importancia se ubicaron el coeficiente de Gini (0.0842) y el Índice de Desarrollo Humano (0.0741), seguidos por la población urbana (0.0625), la participación del sector industrial (0.0617) y la población total (0.0487). Estos resultados muestran que las dimensiones sociales, productivas y demográficas también contribuyeron al desempeño predictivo del algoritmo, aunque con una importancia relativa menor frente a las variables energéticas y económicas.

Finalmente, la inversión extranjera directa (0.0122) presentó el nivel más bajo de importancia dentro del modelo. Esto sugiere que, para el conjunto de datos analizado y bajo la estructura predictiva de XGBoost, esta variable aportó una cantidad relativamente menor de información incremental para la predicción de las emisiones de CO_2 per cápita.

En términos generales, los niveles de importancia detectados indican que el modelo XGBoost se apoyó principalmente en variables relacionadas con el consumo de energía, la estructura energética, las características ambientales y la actividad económica. Estos hallazgos permiten identificar los factores con mayor relevancia dentro del proceso predictivo y constituyen una aproximación empírica a los patrones asociados con la desigualdad climática en América Latina, sin implicar relaciones causales entre las variables analizadas.

4.3.2.4 Red Neuronal Multicapa (MLP – MLPRegressor)

La Red Neuronal Multicapa (MLP) corresponde a un algoritmo de aprendizaje supervisado diseñado para modelar relaciones complejas mediante capas sucesivas de neuronas interconectadas. Su funcionamiento se basa en la transformación progresiva de la información de entrada a través de capas ocultas, proceso que puede expresarse matemáticamente de la siguiente manera:

$$a^{(l)} = g\left(W^{(l)}a^{(l-1)}+b^{(l)}\right)$$

donde $a^{(l)}$representa la activación de la capa l, $W^{(l)}$corresponde a la matriz de pesos, $b^{(l)}$al vector de sesgos y $g(\cdot)$a la función de activación no lineal. Mediante este proceso, la red ajusta iterativamente sus parámetros internos hasta aproximar la relación entre las variables explicativas y las emisiones de CO_2 per cápita.

El entrenamiento se realizó utilizando la versión estandarizada de la base de datos, debido a la sensibilidad de las redes neuronales frente a diferencias de escala entre variables. Para ello, se empleó el conjunto de entrenamiento compuesto por 442 observaciones correspondientes al período 1990–2015.

La optimización del modelo se desarrolló mediante validación cruzada de cinco pliegues,

evaluando 36 configuraciones distintas, equivalentes a 180 procesos internos de entrenamiento y validación. Durante este proceso se probaron arquitecturas con diferente número de capas ocultas y neuronas, incluyendo estructuras más simples y configuraciones de mayor complejidad, con el fin de identificar el nivel de representación más adecuado para los patrones presentes en los datos. Asimismo, se compararon distintas funciones de activación orientadas a incorporar no linealidad al aprendizaje de la red. También se ajustaron mecanismos de regularización para controlar el crecimiento excesivo de los pesos internos y reducir riesgos de sobreajuste, junto con diferentes tasas de aprendizaje encargadas de regular la velocidad de actualización de los parámetros durante el entrenamiento iterativo.

La configuración con mejor desempeño estuvo compuesta por tres capas ocultas de 64, 64 y 32 neuronas, función de activación tangente hiperbólica (tanh), un parámetro de regularización $\alpha = 0.01$ y una tasa de aprendizaje inicial de 0.01. Desde una perspectiva metodológica, esta arquitectura proporcionó la capacidad necesaria para representar relaciones no lineales presentes en los datos, manteniendo al mismo tiempo mecanismos de regularización orientados a favorecer la generalización del modelo. La utilización de la función de activación tanh permitió incorporar transformaciones no lineales en las capas ocultas, mientras que el parámetro de regularización contribuyó a controlar la complejidad de la red durante el proceso de entrenamiento.

Los resultados obtenidos sugieren la existencia de patrones no lineales entre las variables económicas, energéticas, sociales y ambientales consideradas en el estudio. Asimismo, la arquitectura seleccionada indica que una representación con múltiples capas ocultas aportó información útil para el proceso predictivo, complementando los hallazgos obtenidos mediante los demás algoritmos de aprendizaje automático desarrollados en la investigación.

4.3.2.5 Modelo Support Vector Regression (SVR)

El algoritmo Support Vector Regression (SVR) fue implementado como un modelo de regresión supervisada orientado a representar relaciones no lineales entre las variables socioeconómicas, energéticas y ambientales y las emisiones de CO_2 per cápita. A diferencia de los modelos que minimizan directamente todo error de predicción, SVR busca construir una función de regresión que tolere pequeñas desviaciones dentro de un margen ε, penalizando únicamente los errores que exceden dicho intervalo. Esta característica permite concentrar el aprendizaje en patrones estructurales relevantes y reducir la sensibilidad frente a fluctuaciones menores de los datos El modelo se formula como un problema de optimización en el que se minimiza simultáneamente la complejidad de la función estimada y los errores que quedan por fuera del margen de tolerancia:

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

donde $w$representa el vector de pesos de la función de regresión, Ccontrola el grado de penalización de los errores, y ξ_i, ξ_i^*corresponden a las desviaciones positivas y negativas que exceden el margen ε. Cuando la relación entre las variables explicativas y la variable objetivo no puede representarse linealmente, SVR emplea funciones *kernel* para proyectar los datos hacia un espacio de mayor dimensionalidad, permitiendo capturar estructuras más complejas.

El entrenamiento se realizó con 442 observaciones correspondientes al período 1990–2015.
Para definir la configuración final se aplicó GridSearchCV con validación cruzada de cinco pliegues, evaluando 96 configuraciones, equivalentes a 480 procesos internos de entrenamiento y validación. La búsqueda incluyó funciones kernel lineal, polinomial y radial; valores del parámetro de penalización C; distintos márgenes ε; y configuraciones del parámetro γ, encargado de regular el alcance de influencia de cada observación en el espacio transformado.
La mejor configuración obtenida fue kernel RBF, C = 0.1, ε = 0.01 y γ = auto. La selección del kernel radial sugiere la presencia de patrones no lineales entre las variables utilizadas para la predicción y las emisiones de CO_2 per cápita. El valor de C = 0.1 favoreció una mayor regularización del modelo, priorizando la capacidad de generalización frente a posibles ajustes excesivos a los datos de entrenamiento. Por su parte, ε = 0.01 definió un margen reducido de tolerancia al error durante el proceso de ajuste, mientras que la configuración γ = auto permitió establecer la influencia de las observaciones a partir de la dimensión del conjunto de variables utilizado en el modelo.
Desde una perspectiva metodológica, estos resultados indican que la incorporación de transformaciones no lineales mediante el kernel radial aportó información útil para el proceso predictivo. No obstante, el desempeño alcanzado por el modelo sugiere que la capacidad de representación de los patrones presentes en los datos fue inferior a la obtenida por otros algoritmos evaluados en la investigación, evidenciando diferencias en la forma en que cada técnica captura las relaciones observadas entre las variables socioeconómicas, energéticas y ambientales analizadas.

Modelo K-Nearest Neighbors (KNN)

La incorporación del algoritmo K-Nearest Neighbors (KNN) permitió explorar si las emisiones de CO_2 per cápita en América Latina presentan patrones de similitud entre países con configuraciones socioeconómicas y energéticas comparables. Los resultados observados en el análisis exploratorio y en los modelos previos mostraban que varios países compartían trayectorias ambientales cercanas aun cuando sus estructuras institucionales o dimensiones territoriales fueran distintas. Bajo este contexto, KNN ofrecía una aproximación adecuada para analizar relaciones locales dentro del espacio multivariado, especialmente en escenarios donde las dinámicas entre variables no siguen comportamientos homogéneos ni estrictamente lineales.

El entrenamiento del modelo K-Nearest Neighbors (KNN) se realizó utilizando las mismas diez variables explicativas consideradas en los modelos anteriores y una muestra de 442 observaciones correspondientes al período 1990–2015. Debido a que este algoritmo basa sus predicciones en cálculos de distancia entre observaciones, se trabajó sobre la versión estandarizada de la base de datos, evitando que variables con escalas mayores ejercieran una influencia desproporcionada sobre las medidas de proximidad. Este procedimiento permitió que todas las variables contribuyeran de manera comparable a la construcción del espacio multivariado utilizado por el modelo.

La optimización de hiperparámetros se desarrolló mediante una búsqueda exhaustiva utilizando Grid Search con validación cruzada de cinco pliegues aplicada exclusivamente sobre el conjunto de entrenamiento. En total se evaluaron 24 configuraciones distintas, equivalentes a 120 procesos internos de entrenamiento y validación. Durante este procedimiento se analizaron diferentes

cantidades de vecinos cercanos, esquemas de ponderación y métricas de distancia, con el propósito de identificar la configuración con mejor desempeño promedio durante la validación.

Desde la perspectiva metodológica, KNN corresponde a un algoritmo de aprendizaje supervisado no paramétrico basado en instancia, donde el aprendizaje no se realiza mediante la estimación explícita de parámetros o reglas internas, sino mediante la identificación de observaciones históricamente similares. En problemas de regresión, la predicción puede expresarse de la siguiente manera:

$$\hat{y}(x) = \frac{1}{k}\sum_{i \in N_k(x)} y_i$$

donde $\hat{y}(x)$representa la predicción estimada para una nueva observación, $k$corresponde al número de vecinos considerados y $N_k(x)$al conjunto de observaciones más próximas dentro del espacio de características.

Tras evaluar las diferentes combinaciones de hiperparámetros, la configuración óptima correspondió a un modelo con 20 vecinos cercanos, ponderación basada en distancia y métrica Manhattan. La selección de un número relativamente alto de vecinos favorece la estabilidad de las predicciones al incorporar información de un conjunto más amplio de observaciones similares. Asimismo, la ponderación por distancia asigna una mayor influencia a las observaciones más cercanas dentro del espacio de características, mientras que la métrica Manhattan calcula la proximidad a partir de la suma de las diferencias absolutas entre las variables consideradas.

Desde una perspectiva metodológica, esta configuración permite identificar patrones de similitud entre observaciones que comparten características económicas, sociales, energéticas y territoriales comparables. En consecuencia, las predicciones generadas por el modelo se fundamentan en la proximidad observada entre países y períodos con perfiles similares dentro del conjunto de datos analizado.

El modelo KNN presentó los mejores resultados sobre el conjunto de prueba temporal (2016–2020), alcanzando un coeficiente de determinación de $R^2 = 0.7364$ y un RMSE de 0.2495. Estos resultados indican que fue el modelo con mayor capacidad predictiva para reproducir los patrones observados en las emisiones de CO_2 per cápita dentro de los datos no utilizados durante el entrenamiento. Aunque otros algoritmos, como XGBoost, también mostraron un desempeño competitivo, KNN logró el mejor equilibrio entre capacidad predictiva y error de predicción, por lo que fue seleccionado como el modelo de mejor desempeño global en esta investigación. En consecuencia, los análisis e interpretaciones posteriores se apoyan principalmente en los resultados obtenidos mediante este algoritmo, entendidos como una aproximación empírica a los patrones asociados con la desigualdad climática en América Latina.

4.3.3 Extensión 1: Red neuronal profunda (Deep MLP)

La Red Neuronal Profunda (Deep Neural Network) corresponde a una arquitectura de aprendizaje supervisado compuesta por múltiples capas ocultas interconectadas, diseñada para aprender representaciones jerárquicas y relaciones no lineales complejas a partir de los datos. En contraste con modelos más convencionales, las redes profundas permiten transformar progresivamente la información de entrada hacia niveles de abstracción más complejos mediante sucesivas combinaciones de pesos, sesgos y funciones de activación no lineales. Este proceso puede representarse matemáticamente de la siguiente manera:

$$a^{(l)} = g\left(W^{(l)}a^{(l-1)}+b^{(l)}\right)$$

donde $a^{(l)}$representa la activación de la capa l, $W^{(l)}$corresponde a la matriz de pesos, $b^{(l)}$al vector de sesgos y $g(\cdot)$a la función de activación. A través del proceso de retropropagación del gradiente, la red ajusta iterativamente estos parámetros para minimizar el error de predicción.

La incorporación de esta arquitectura surgió como una extensión de los modelos de aprendizaje automático previamente desarrollados, particularmente después de identificar relaciones no lineales entre variables socioeconómicas, energéticas y ambientales en algoritmos como SVR, Random Forest y XGBoost. En estudios ambientales y climáticos, las redes profundas han sido utilizadas debido a su capacidad para representar interacciones complejas y patrones de dependencia que difícilmente pueden ser capturados mediante estructuras lineales o modelos con menor profundidad, así como en el presente estudio, donde variables como crecimiento económico, urbanización, consumo energético y desarrollo humano interactúan simultáneamente bajo dinámicas heterogéneas entre países y períodos de tiempo. El modelo fue construido utilizando las mismas diez variables explicativas empleadas en los modelos anteriores. Para evitar fuga de información y respetar la estructura temporal de los datos, la base fue dividida en un conjunto de entrenamiento compuesto por 442 observaciones correspondientes al período 1990–2015 y un conjunto de prueba conformado por 85 observaciones correspondientes al período 2016–2020. Debido a la sensibilidad de las redes neuronales frente a diferencias de escala entre variables, se aplicó previamente un proceso de estandarización, garantizando estabilidad numérica durante el entrenamiento.

La arquitectura implementada estuvo compuesta por cuatro capas densamente conectadas con 256, 128, 64 y 32 neuronas, respectivamente. Adicionalmente, se incorporaron mecanismos de regularización mediante capas de Batch Normalization y Dropout, orientados a mejorar la estabilidad del entrenamiento y reducir el riesgo de sobreajuste. Las capas Batch Normalization permitieron estabilizar las distribuciones internas de activación, mientras que las capas Dropout contribuyeron a disminuir la dependencia excesiva entre neuronas durante el aprendizaje.

El proceso de optimización se realizó utilizando el optimizador Adam con una tasa de aprendizaje de 0.001 y la función de pérdida basada en error cuadrático medio. Como métrica complementaria de seguimiento se empleó el error absoluto medio (MAE). El entrenamiento fue configurado para un máximo de 300 épocas, utilizando lotes de 32 observaciones, una partición interna para validación y el mecanismo Early Stopping para detener automáticamente el proceso cuando no se observaran mejoras adicionales sobre los datos de validación.

La implementación de estas estrategias buscó favorecer la capacidad de generalización del modelo y controlar el sobreajuste. Sin embargo, los resultados obtenidos sugieren que la complejidad de la arquitectura no se tradujo en una mejora del desempeño predictivo frente a otros algoritmos evaluados en el estudio.

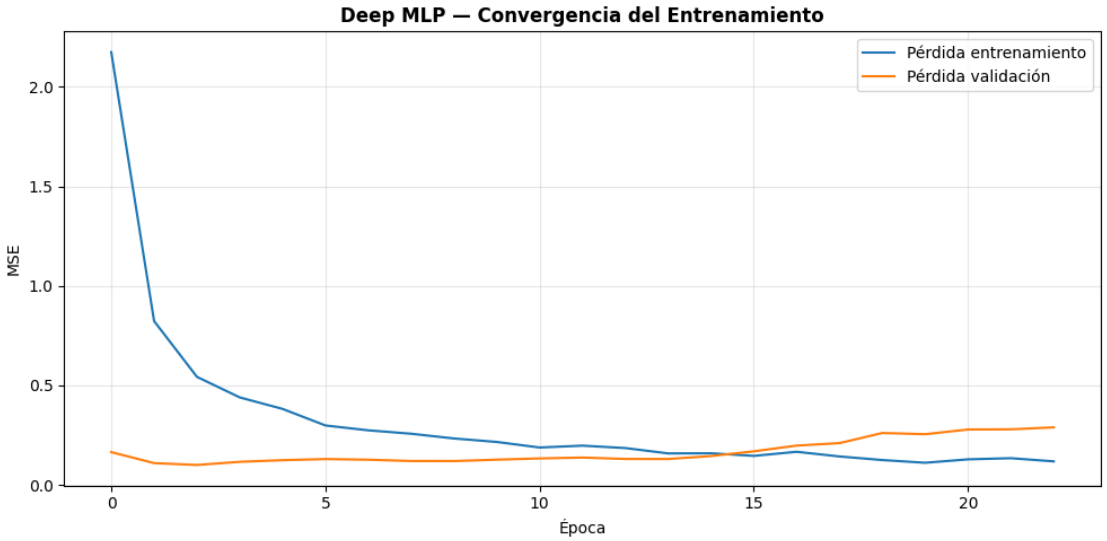


Ilustración 14 Deep MLP-Convergencia del entrenamiento

El análisis visual de las curvas de entrenamiento y validación permitió monitorear el comportamiento del proceso de aprendizaje a lo largo de las épocas. La utilización de Batch Normalization, Dropout y Early Stopping contribuyó a estabilizar el entrenamiento y a controlar posibles problemas de sobreajuste. Finalmente, una vez completado el proceso de entrenamiento, el modelo fue evaluado sobre el conjunto de prueba temporal correspondiente al período 2016–2020.

Tabla 11 resultados MPL

Modelo	**R^2**	**RMSE**	**MAE**	**MAPE (%)**
Deep MLP	-0.1124	0.5126	0.3304	3411.63

Los resultados obtenidos evidenciaron un desempeño inferior al alcanzado por los demás algoritmos evaluados. El modelo Deep MLP registró un R^2 de -0.1124, un RMSE de 0.5126 y un MAE de 0.3304, indicando una capacidad predictiva limitada sobre el conjunto de prueba. Un valor negativo de R^2 sugiere que las predicciones generadas por la red neuronal fueron menos precisas que una aproximación basada únicamente en la media de la variable objetivo.

Estos resultados indican que, para este conjunto de datos tabulares y el horizonte temporal

analizado, una arquitectura profunda no logró capturar patrones predictivos con la misma eficacia observada en modelos como KNN, XGBoost o Random Forest. Lo anterior sugiere que una mayor complejidad arquitectónica no necesariamente se traduce en una mejora del desempeño predictivo cuando el volumen de datos disponible es relativamente limitado y las relaciones presentes en los datos pueden ser modeladas adecuadamente mediante algoritmos menos complejos.

El valor elevado del MAPE (3411.63 %) se explica por la presencia de observaciones con emisiones de CO_2 per cápita muy cercanas a cero, para las cuales pequeños errores absolutos generan grandes errores porcentuales. En consecuencia, métricas como el MAE (0.3304) y el RMSE (0.5126) proporcionan una evaluación más representativa del desempeño real del modelo en este contexto.

4.3.4 Extensión metodológica: Modelos de clasificación multiclase

Adicionalmente, se implementó un enfoque de clasificación supervisada para analizar si las variables socioeconómicas permitían diferenciar patrones asociados a distintos niveles de emisiones de CO_2 per cápita entre los países de América Latina. Mientras los modelos de regresión permiten predecir valores continuos de emisiones, este enfoque permite identificar la pertenencia de una observación a categorías de emisiones previamente definidas, aportando una perspectiva complementaria para el análisis de la desigualdad climática.

Para ello, la variable correspondiente a las emisiones de CO_2 per cápita fue transformada desde una escala continua hacia una variable categórica ordinal mediante una discretización por terciles calculados exclusivamente sobre el conjunto de entrenamiento (1990–2015). Esta estrategia permitió evitar filtraciones de información (data leakage) y garantizar que la definición de las categorías no incorporara información proveniente del período de prueba. Como resultado, se definieron tres niveles de emisiones: bajo, medio y alto. La distribución global obtenida fue de 166 observaciones para la categoría baja, 183 para la categoría media y 178 para la categoría alta, evidenciando un adecuado equilibrio entre clases.

Posteriormente, las categorías fueron codificadas y la base de datos fue dividida utilizando una partición temporal, conformada por 442 observaciones para entrenamiento (1990–2015) y 85 observaciones para prueba (2016–2020). De manera adicional, las variables fueron estandarizadas con el fin de garantizar estabilidad en aquellos algoritmos sensibles a diferencias de escala, como KNN, SVM y redes neuronales.

Tabla 12 Distribución de clases de emisiones de CO_2 per cápita

Categoría	**Número de observaciones**	**Porcentaje**
Bajo	166	31.50 %
Medio	183	34.72 %
Alto	178	33.78 %
Total	**527**	**100%**

A partir de esta estructura se entrenaron cinco algoritmos de clasificación supervisada: Random Forest, K-Nearest Neighbors (KNN), XGBoost, Red Neuronal Multicapa (MLP) y Support Vector Machine (SVM). Todos los modelos fueron optimizados mediante validación cruzada de cinco pliegues sobre el conjunto de entrenamiento, utilizando el F1-score ponderado como criterio de optimización. La selección de esta métrica responde a la necesidad de evaluar simultáneamente precisión y recuperación en un problema multiclase, evitando depender exclusivamente de la exactitud global y permitiendo una valoración más equilibrada del desempeño de los algoritmos en las diferentes categorías de emisiones.

Modelo CLF-1: Random Forest

El modelo Random Forest para clasificación alcanzó una exactitud del 65.88 %, una precisión ponderada de 66.87 %, un recall de 65.88 % y un F1-score ponderado de 0.6480. Estos resultados evidencian una capacidad moderada para clasificar los niveles de emisiones de CO_2 per cápita en las categorías definidas. Aunque el modelo logró identificar adecuadamente una proporción importante de observaciones, se observaron errores de clasificación entre categorías, especialmente en los niveles intermedios, lo que sugiere que los patrones asociados a estas clases presentan una mayor complejidad para su diferenciación.

Tabla 13 Desempeño del modelo Random Forest (Clasificación)

Modelo	Accuracy	Precision	Recall	F1
Random Forest	0.6588	0.6687	0.6588	0.6480

Modelo CLF-2: KNN

El algoritmo KNN obtuvo una exactitud del 71.76 %, una precisión ponderada de 72.61 %, un recall de 71.76 % y un F1-score ponderado de 0.7118, superando el desempeño alcanzado por el modelo Random Forest en las métricas globales de clasificación. La configuración óptima seleccionó 15 vecinos, métrica Manhattan y ponderación uniforme, lo que indica que la clasificación se benefició de la proximidad entre observaciones con características socioeconómicas similares. Estos resultados sugieren que los patrones presentes en los datos permiten identificar agrupamientos de países con perfiles comparables en términos de emisiones de CO_2 per cápita y variables socioeconómicas asociadas.

Tabla 14 Desempeño del Modelo KNN

Modelo	Accuracy	Precision	Recall	F1
KNN	0.7176	0.7261	0.7176	0.7118

Modelo CLF-3: XGBoost

El clasificador XGBoost alcanzó una exactitud del 68.24 %, una precisión ponderada de 68.96 %, un recall de 68.24 % y un F1-score ponderado de 0.6792. Estos resultados reflejan un desempeño satisfactorio en la clasificación de los niveles de emisiones de CO_2 per cápita, aunque ligeramente inferior al obtenido por el modelo KNN. La configuración óptima incluyó una tasa de aprendizaje de 0.1, una profundidad máxima de 3, 100 árboles de decisión y una fracción de muestreo de 1.0. En conjunto, los resultados sugieren que el modelo fue capaz de capturar patrones relevantes presentes en las variables socioeconómicas para diferenciar las categorías de emisiones definidas en el estudio.

Tabla 15 Desempeño del modelo XGBoost (Clasificación)

Modelo	Accuracy	Precision	Recall	F1
XGBoost	0.6824	0.6896	0.6824	0.6792

Modelo CLF-4: Red Neuronal MLP

La red neuronal multicapa (MLP) alcanzó una exactitud del 58.82 %, una precisión ponderada de 59.85 %, un recall de 58.82 % y un F1-score ponderado de 0.5900. Estos resultados evidencian un desempeño moderado en la clasificación de los niveles de emisiones de CO_2 per cápita, situándose por debajo de los obtenidos por los modelos KNN, XGBoost y Random Forest. La configuración óptima seleccionó una arquitectura con dos capas ocultas de 64 y 32 neuronas, función de activación ReLU y un parámetro de regularización (alpha) de 0.0001. Aunque el modelo logró identificar parte de los patrones presentes en los datos, su capacidad de clasificación fue más limitada en comparación con los demás algoritmos evaluados.

Tabla 16 Desempeño del modelo MLP (Clasificación)

Modelo	Accuracy	Precision	Recall	F1
Red Neuronal MLP	0.5882	0.5985	0.5882	0.5900

Modelo CLF-5: SVM

El modelo SVM alcanzó una exactitud del 50.59 %, una precisión ponderada de 59.07 %, un recall de 50.59 % y un F1-score ponderado de 0.5042. Estos resultados evidencian el desempeño más bajo entre los algoritmos de clasificación evaluados. La configuración óptima seleccionó un kernel radial (RBF), un parámetro de regularización C igual a 0.1 y gamma configurado automáticamente. Aunque el modelo logró identificar algunos patrones presentes en los datos, su capacidad para diferenciar adecuadamente las categorías de emisiones de CO_2 per cápita fue más limitada en comparación con KNN, XGBoost, Random Forest y la red neuronal multicapa.

Tabla 17 Desempeño del modelo SVM Clasificación

Modelo	Accuracy	Precision	Recall	F1
SVM	0.5059	0.5907	0.5059	0.5042

5 OBJETIVO 4: Evaluar y comparar el rendimiento de los modelos mediante métricas de desempeño.

5.1 Validación estadística entre los mejores modelos (Dietterich 5×2 CV)

Una vez evaluado el desempeño predictivo de los modelos de regresión mediante métricas como R^2, RMSE y MAE, se realizó una validación estadística formal con el propósito de determinar si las diferencias observadas entre algoritmos correspondían a diferencias reales de desempeño o si podían atribuirse a variaciones aleatorias derivadas de la partición de los datos. Esta etapa resulta metodológicamente necesaria dentro de los procesos de ciencia de datos, dado que una diferencia descriptiva en métricas no implica necesariamente una superioridad estadísticamente significativa entre modelos.

Para este propósito se implementó la prueba de Dietterich 5×2 Cross-Validation (5×2 CV), propuesta por Thomas G. Dietterich [38], quien plantea que este procedimiento constituye una alternativa robusta para comparar algoritmos de aprendizaje supervisado, debido a que permite controlar el error tipo I y estimar directamente la variabilidad generada por diferentes particiones del conjunto de entrenamiento. Según el autor, cuando un algoritmo puede ejecutarse múltiples veces, la prueba 5×2 CV es recomendada porque ofrece una combinación adecuada entre potencia estadística y control del error experimental.

La prueba consiste en realizar cinco repeticiones independientes de validación cruzada en dos particiones (2-fold cross-validation). En cada repetición, la base de datos se divide aleatoriamente en dos subconjuntos equivalentes. Posteriormente, ambos modelos son entrenados y evaluados de manera cruzada sobre cada partición, generando un total de 10 evaluaciones independientes. Este procedimiento permite estimar de manera más robusta la variabilidad del desempeño de los modelos, reduciendo la dependencia de una única división entrenamiento-prueba.

En el presente estudio, la comparación estadística fue realizada entre los modelos KNN y XGBoost, debido a que ambos presentaron los mejores desempeños predictivos sobre el conjunto de prueba temporal. Para evaluar si las diferencias observadas en las métricas de desempeño podían atribuirse al azar o reflejaban diferencias sistemáticas entre los algoritmos, se aplicó la prueba 5×2 CV propuesta por Dietterich (1998). Esta prueba es ampliamente utilizada para comparar modelos de aprendizaje automático entrenados sobre un mismo conjunto de datos, independientemente de que pertenezcan a paradigmas metodológicos diferentes. Como criterio de comparación se utilizó el error cuadrático medio (MSE), dado que esta métrica penaliza con mayor intensidad los errores de predicción de gran magnitud y permite evaluar de manera consistente el desempeño de modelos de regresión.

La hipótesis nula evaluada fue:

$$H_0: E_1 = E_2$$

donde E_1 y E_2 representan el error esperado de cada algoritmo. Bajo esta formulación, un valor de $p < 0.05$ indicaría evidencia suficiente para concluir que existe una diferencia estadísticamente significativa entre ambos modelos.

Resultados de la prueba

Tabla 18 Resultados Prueba Dieterich

Comparación	Métrica utilizada	Estadístico t	Valor p	Decisión
KNN vs XGBoost	MSE	8.5287	0.0004	Se rechaza Ho

La prueba de Dietterich 5×2 CV produjo un estadístico t = 8.5287 y un p-valor = 0.0004. Dado que el p-valor fue inferior al nivel de significancia de 0.05, se rechazó la hipótesis nula de igualdad de desempeño esperado entre los modelos. Este resultado indica la existencia de diferencias estadísticamente significativas entre KNN y XGBoost respecto al error cuadrático medio obtenido durante el proceso de validación. En consecuencia, la superioridad observada del modelo KNN en las métricas de evaluación no puede atribuirse únicamente a variaciones aleatorias derivadas de las particiones de los datos, lo que proporciona evidencia adicional para respaldar su selección como modelo de mejor desempeño dentro de la investigación.

5.2 Diagnóstico del mejor modelo mediante análisis de residuales

Una vez evaluadas las diferencias de desempeño entre modelos mediante la prueba estadística de Dietterich 5×2 CV, se realizó un diagnóstico del modelo con mejor capacidad predictiva para analizar la calidad de su ajuste y verificar la posible presencia de patrones sistemáticos en los errores de predicción. Dentro de los procesos de ciencia de datos aplicados a problemas de regresión, el análisis de residuales constituye una etapa fundamental de validación, dado que permite identificar si el modelo presenta sesgos estructurales, problemas de heterogeneidad en el error o patrones no capturados durante el entrenamiento.

Para este análisis se seleccionó automáticamente el modelo con mayor coeficiente de determinación (R^2), correspondiente al algoritmo K-Nearest Neighbors (KNN), el cual alcanzó un desempeño de $R^2 = 0.7364$ sobre el conjunto de prueba temporal. Debido a que fue el modelo con mejor capacidad predictiva entre los algoritmos evaluados, se realizó un análisis detallado de sus residuos con el fin de examinar el comportamiento de los errores de predicción y complementar la evaluación de desempeño obtenida mediante las métricas cuantitativas. Los residuos fueron calculados como la diferencia entre los valores observados y los valores predichos por el modelo,

de acuerdo con la siguiente expresión:

$$e_i = y_i - \hat{y}_i$$

donde $e_i$representa el error residual de cada observación, $y_i$corresponde al valor real de emisiones de CO_2 per cápita y $\hat{y}_i$representa la predicción generada por el modelo. Bajo un modelo adecuadamente ajustado, se espera que estos residuos se distribuyan aleatoriamente alrededor de cero, sin patrones visibles asociados al nivel de predicción.

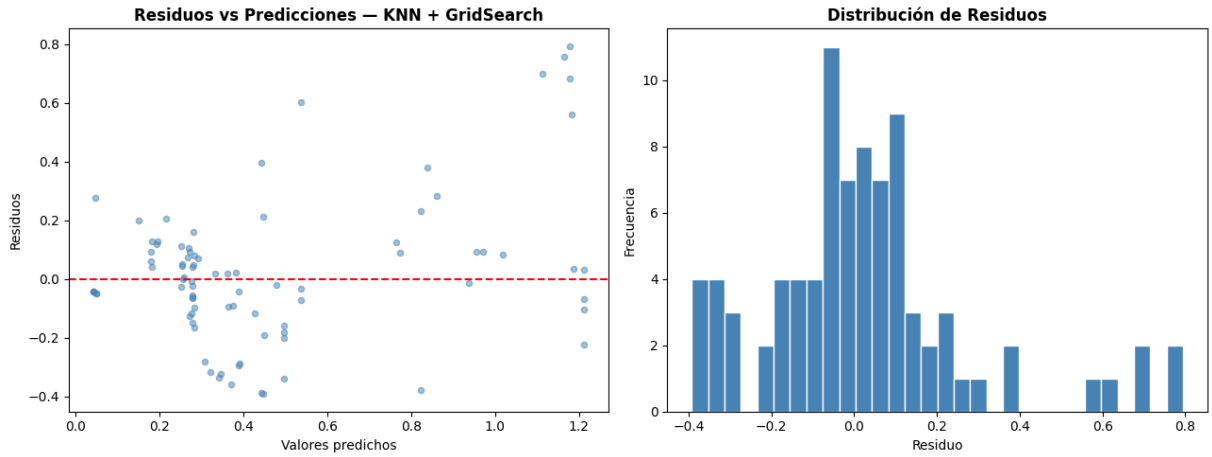


Ilustración 15 Análisis de residuos del modelo KNN.

La gráfica de dispersión entre residuos y valores predichos muestra una distribución centrada alrededor de la línea horizontal de referencia (e = 0), sin evidenciar patrones sistemáticos claramente definidos. Aunque se observa una mayor dispersión de algunos residuos para determinados rangos de valores predichos, no se identifican tendencias consistentes de subestimación o sobreestimación que sugieran problemas graves en el comportamiento predictivo del modelo.

De forma complementaria, el histograma de residuos evidencia una concentración importante de observaciones alrededor de cero, acompañada de cierta asimetría y la presencia de algunos errores de mayor magnitud en los extremos de la distribución. Los estadísticos descriptivos obtenidos muestran una media residual de 0.0228 y una desviación estándar de 0.2485, lo que indica que, en promedio, los errores del modelo permanecen relativamente cercanos a cero, aunque con una dispersión moderada.

La distribución estadística de los residuos fue evaluada mediante la prueba de Kolmogorov-Smirnov. Debido a que el conjunto de prueba estuvo compuesto por más de 50 observaciones, esta prueba permitió comparar la distribución empírica de los errores con una distribución normal teórica. Los resultados obtenidos mostraron un estadístico KS = 0.1568 y un valor p = 0.0273.

Tabla 19 Resultados Prueba de normalidad-KNN

Indicador	**Resultado**
Mejor modelo	KNN
R^2	0.7364
Media residual	0.0228
Desviación estándar	0.2485
Prueba de normalidad	Kolmogorov-Smirnov
Estadístico	0.1568
Valor p	0.0273
Decisión	Se rechaza la hipótesis nula de normalidad

Dado que el valor p obtenido (0.0273) es inferior al nivel de significancia establecido ($\alpha = 0.05$), se rechaza la hipótesis nula de normalidad de los residuos. Este resultado indica que la distribución de los errores presenta desviaciones respecto a una distribución normal teórica, situación frecuente en modelos de aprendizaje automático cuya finalidad principal es la predicción y no la inferencia estadística basada en supuestos clásicos de regresión.

A pesar de ello, la inspección gráfica de los residuos y las métricas de desempeño obtenidas muestran que el modelo mantiene una capacidad predictiva adecuada sobre datos no utilizados durante el entrenamiento. En consecuencia, los resultados respaldan la utilización del algoritmo KNN como modelo de mejor desempeño dentro de la investigación, proporcionando una aproximación empírica útil para identificar patrones asociados con las emisiones de CO_2 per cápita en América Latina.

5.3 Explicabilidad del mejor modelo mediante SHAP

Los métodos de ensamble como Random Forest y Extreme Gradient Boosting (XGBoost), constituyen una familia de algoritmos que combinan múltiples modelos base (típicamente árboles de decisión) para producir predicciones más robustas y con menor varianza que un modelo individual. Random Forest opera mediante bagging, construyendo cada árbol sobre una muestra aleatoria con reemplazo del conjunto de entrenamiento y seleccionando un subconjunto aleatorio de variables en cada nodo. XGBoost, por su parte, utiliza boosting secuencial con regularización explícita, añadiendo árboles que minimizan el residuo del modelo anterior mediante descenso del gradiente. Estas arquitecturas son especialmente adecuadas para conjuntos de datos tabulares con relaciones no lineales y alta dimensionalidad relativa[22].

La metodología SHapley Additive exPlanations (SHAP), propuesta por Lundberg y Lee [23], ofrece un marco teóricamente fundamentado para cuantificar la contribución de cada variable a las predicciones de cualquier modelo de aprendizaje automático. Su fundamento proviene de la teoría

de juegos cooperativos: el valor de Shapley de una variable se define como la contribución marginal promedio de dicha variable considerando todas las posibles coaliciones de predictores. Esta propiedad garantiza que los valores SHAP sean consistentes, localmente exactos y sumables hasta reproducir la predicción del modelo, convirtiéndolos en la herramienta de explicabilidad más sólida disponible actualmente para modelos de caja negra.

Una vez validado el desempeño predictivo del modelo seleccionado y analizado el comportamiento de sus errores mediante el estudio de los residuos, se procedió a realizar un análisis de explicabilidad con el propósito de identificar cuáles variables socioeconómicas, energéticas, demográficas y ambientales aportan más información a las predicciones generadas por el algoritmo. Esta etapa resulta especialmente relevante dentro del enfoque de ciencia de datos aplicado en el presente estudio, dado que un modelo con alta capacidad predictiva no necesariamente ofrece interpretabilidad sobre los factores asociados a sus predicciones. En este sentido, la explicabilidad permite complementar el análisis predictivo mediante la identificación de patrones y contribuciones relativas de las variables relacionadas con las emisiones de CO_2 per cápita y, de manera indirecta, con la aproximación empírica a la desigualdad climática considerada en la investigación.

Para este propósito se implementó la metodología SHapley Additive exPlanations (SHAP), desarrollada por Scott Lundberg y Su-In Lee. Este enfoque se fundamenta en la teoría de juegos cooperativos, específicamente en los valores de Shapley, los cuales permiten cuantificar la contribución marginal promedio de cada variable a la predicción final del modelo. Los autores demuestran que SHAP proporciona interpretaciones consistentes, localmente exactas y comparables entre variables, incluso en algoritmos complejos de aprendizaje automático.

Desde una perspectiva matemática, la contribución individual de una variable puede expresarse mediante el valor de Shapley:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

donde $\phi_i$representa la contribución de la variable i, y $f(S)$corresponde a la predicción del modelo utilizando diferentes subconjuntos de variables. Esta formulación permite estimar tanto la magnitud como la dirección del efecto de cada predictor sobre la salida del algoritmo.

Dado que el mejor modelo identificado fue K-Nearest Neighbors (KNN), el análisis se desarrolló utilizando KernelExplainer, una variante modelo-agnóstica de SHAP adecuada para interpretar algoritmos cuya estructura no permite obtener contribuciones de variables de forma directa. Este enfoque permite estimar la contribución relativa de cada predictor a las predicciones generadas por el modelo, independientemente de la técnica de aprendizaje automático utilizada. Para reducir el costo computacional del procedimiento, se empleó una submuestra representativa del conjunto de entrenamiento, procurando mantener las características generales y la distribución de los datos utilizados durante el proceso de modelación.

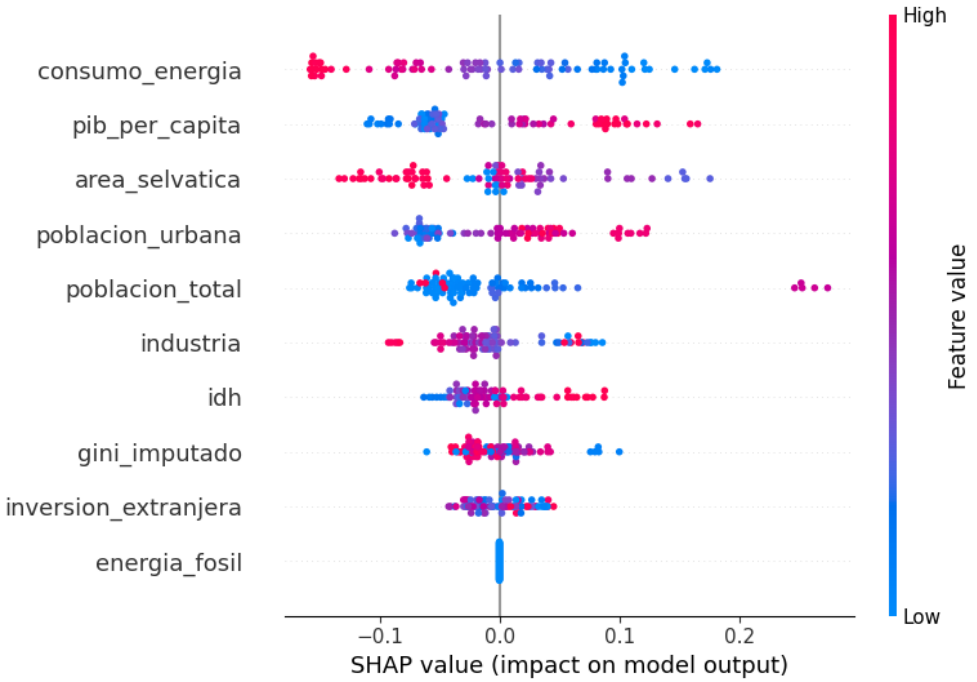


Ilustración 16 Impacto global de variables sobre las predicciones del modelo KNN según valores SHAP.

El gráfico beeswarm permite observar simultáneamente la magnitud, dirección y distribución del impacto de cada variable sobre las predicciones del modelo. Cada punto representa una observación individual, mientras que el color indica el valor relativo de la variable analizada. Valores SHAP positivos indican que la variable contribuye a incrementar la predicción de emisiones de CO_2 per cápita, mientras que valores negativos contribuyen a reducirla.

Los resultados muestran que la variable con mayor contribución relativa fue el consumo de energía, seguida por el PIB per cápita, el área selvática, la población urbana y la población total. Este patrón sugiere que las dimensiones energéticas, económicas, demográficas y ambientales aportan información relevante para el proceso predictivo desarrollado por el modelo KNN. En particular, el consumo de energía presenta la mayor dispersión de valores SHAP, evidenciando una influencia más marcada sobre las predicciones generadas por el algoritmo.

Adicionalmente, se observa que valores elevados de PIB per cápita y población urbana suelen asociarse con contribuciones positivas a las predicciones de emisiones, mientras que el área selvática presenta efectos más heterogéneos, reflejando patrones no lineales dentro del conjunto de datos. Por su parte, variables como industria, IDH, coeficiente de Gini e inversión extranjera muestran contribuciones de menor magnitud relativa, mientras que la variable energía fósil registra

una contribución prácticamente nula dentro del modelo seleccionado.

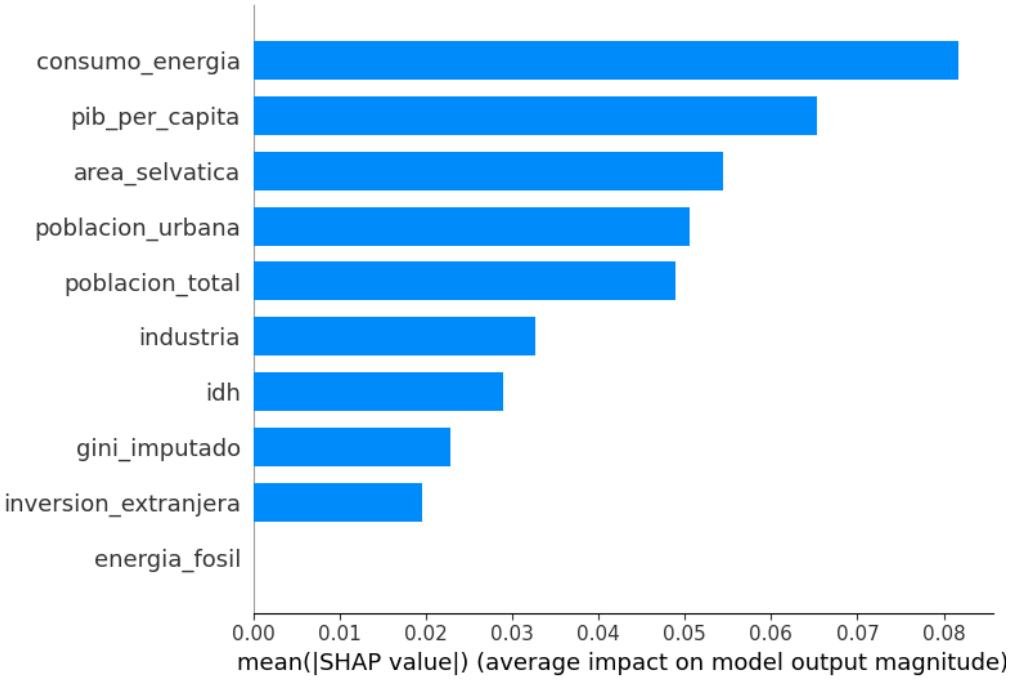


Ilustración 17 Importancia promedio de variables según valores SHAP absolutos

Adicionalmente, se calculó la importancia promedio absoluta de cada variable, permitiendo complementar la interpretación individual de sus contribuciones sobre las predicciones del modelo.

Tabla 20 Importancia global de variables según SHAP

Variable	**Valor SHAP medio**	**Nivel de importancia**
Consumo de energía	0.0816	Muy alta
Pib percápita	0.0653	Alta
Área selvatica	0.0544	Media
Población urbana	0.0506	Media
Población total	0.0488	Baja
Industria	0.0327	Baja
IDH	0.0290	Baja
Gini	0.0229	Baja
Inversion extranjera	0.0196	Baja
Energía fosil	0.0000	Baja

Fuente: Elaboración propia

Los resultados revelan que el modelo KNN integra de manera simultánea información proveniente de dimensiones energéticas, económicas, demográficas y ambientales durante el proceso predictivo. El análisis de explicabilidad mediante SHAP permite identificar la contribución relativa de cada variable a las predicciones generadas por el algoritmo, complementando la evaluación basada en métricas de desempeño. En conjunto, estos hallazgos aportan una aproximación empírica a los patrones asociados con las emisiones de CO_2 per cápita en América Latina, proporcionando evidencia complementaria a los resultados obtenidos mediante las diferentes técnicas de aprendizaje automático implementadas en la investigación.

5.4 Evaluación comparativa del desempeño predictivo

La evaluación comparativa de los modelos desarrollados se realizó mediante un conjunto de métricas orientadas a medir tanto la capacidad explicativa como la precisión predictiva de cada algoritmo. Este análisis permitió identificar no solo los modelos con mejor ajuste estadístico, sino también aquellos con mayor capacidad de generalización y utilidad para el análisis de escenarios relacionados con política climática

5.4.1 Comparación de modelos de regresión

Para los modelos de regresión se utilizaron cinco métricas complementarias: el coeficiente de determinación (R^2), la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE), la mediana del error absoluto (MedAE) y el error porcentual absoluto medio (MAPE). La combinación de estas métricas permite evaluar simultáneamente la proporción de variabilidad explicada, la magnitud promedio del error y la estabilidad de las predicciones frente a posibles observaciones extremas.

Antes de analizar los resultados, es necesario advertir que el Error Porcentual Absoluto Medio (MAPE) debe interpretarse con cautela en este conjunto de datos. Varios países de la muestra, como Paraguay, presentan emisiones de CO_2 per cápita históricamente muy bajas (en algunos años inferiores a 0,01 toneladas por habitante), lo que genera denominadores prácticamente nulos al calcular el error porcentual. Como consecuencia, errores absolutos pequeños se traducen en errores porcentuales desproporcionadamente altos, independientemente del modelo evaluado. Por esta razón, el RMSE y el MAE constituyen las métricas primarias de evaluación para el presente estudio, siendo el MAPE reportado únicamente como referencia complementaria

Tabla 21Comparación de modelos de regresión

	Modelo	R2	RMSE	MAE	MedAE	MAPE (%)
1	KNN	0.736400	0.249500	0.174100	0.102800	1231.901800
2	XGBoost	0.728300	0.253300	0.165200	0.090000	1418.853800
0	Random Forest	0.672300	0.278200	0.167300	0.075700	2015.479800
3	Red Neuronal MLP	0.662200	0.282500	0.207800	0.159200	1370.783300
4	SVR	0.107300	0.459200	0.280100	0.128400	2157.657900

Los valores elevados del MAPE en varios modelos, incluyendo KNN (1231,90%) y otros algoritmos, reflejan un fenómeno bien documentado en series con observaciones cercanas a cero; pequeños errores absolutos generan errores porcentuales desproporcionadamente altos cuando el valor real se aproxima a cero. En esta base de datos, países como Paraguay presentan emisiones per cápita históricamente muy bajas (cercanas a 0,003 toneladas), lo que produce denominadores casi nulos al calcular el MAPE. Por esta razón, el RMSE y el MAE constituyen las métricas primarias de evaluación para el presente estudio, siendo el MAPE informado únicamente a título complementario

Por otro lado, los resultados muestran que el modelo KNN obtuvo el mejor desempeño global, alcanzando un $R^2 = 0.7364$, lo que indica que fue el algoritmo con mayor capacidad predictiva para reproducir los patrones observados en las emisiones de CO_2 per cápita dentro del conjunto de prueba temporal. Adicionalmente, presentó el menor RMSE (0.2495), evidenciando un buen nivel de precisión en las predicciones realizadas sobre datos no utilizados durante el entrenamiento.

En segundo lugar, XGBoost alcanzó un $R^2 = 0.7283$ y registró el menor error absoluto medio (MAE = 0.1652), mostrando un desempeño muy cercano al obtenido por KNN. Por su parte, Random Forest ($R^2 = 0.6723$) y la Red Neuronal MLP ($R^2 = 0.6622$) también lograron resultados satisfactorios, aunque con una capacidad predictiva inferior a la de los modelos líderes. En contraste, SVR presentó el menor desempeño entre los algoritmos evaluados ($R^2 = 0.1073$), evidenciando limitaciones para capturar los patrones presentes en los datos bajo la configuración óptima identificada.

En conjunto, los resultados sugieren la existencia de patrones complejos y no lineales en las emisiones de CO_2 per cápita de América Latina. Sin embargo, la comparación entre algoritmos evidencia que diferentes técnicas presentan capacidades predictivas distintas para representar dichos patrones, siendo KNN el modelo que logró el mejor desempeño global en esta investigación.

5.4.2 Comparación de modelos de Clasificación

Como extensión metodológica, se evaluó también el desempeño de los algoritmos de clasificación multiclase desarrollados para categorizar los niveles de emisiones de CO_2 per cápita en tres grupos: bajo, medio y alto. Para ello se utilizaron cuatro métricas complementarias: Accuracy, Precision, Recall y F1-score ponderado, siendo esta última el criterio principal de comparación, dado que integra simultáneamente precisión y recuperación en contextos multiclase.

Tabla 22 Desempeño comparativo de modelos de clasificación

Modelo	Accuracy	Precision	Recall	F1
KNN	0.7176	0.7261	0.7176	0.7118
XGBoost	0.6824	0.6896	0.6824	0.6792
Random Forest	0.6588	0.6687	0.6588	0.6480

Red Neuronal MLP	0.5882	0.5985	0.5882	0.5900
SVM	0.505900	0.5907	0.5059	0.5042

Fuente: Elaboración propia

Los resultados muestran que el modelo KNN obtuvo el mejor desempeño global en la tarea de clasificación, alcanzando una exactitud del 71.76 % y un F1-score ponderado de 0.7118. En segundo lugar se ubicó XGBoost, con una exactitud del 68.24 % y un F1-score de 0.6792, seguido por Random Forest (F1-score = 0.6480), la red neuronal multicapa MLP (F1-score = 0.5900) y el modelo SVM (F1-score = 0.5042).

En conjunto, estos resultados sugieren que los algoritmos basados en proximidad y aprendizaje no lineal presentan una mayor capacidad para identificar patrones asociados a los diferentes niveles de emisiones de CO_2 per cápita. Asimismo, el desempeño superior de KNN indica que las observaciones con características socioeconómicas similares tienden a agruparse en categorías de emisiones comparables, permitiendo una clasificación más precisa dentro del conjunto de datos analizado.

5.5 Evaluación económica del error predictivo

Como complemento a las métricas tradicionales de desempeño, se realizó un análisis del costo asociado al error predictivo, con el propósito de expresar las diferencias entre modelos en términos económicamente interpretables, permitiendo traducir el error promedio de cada algoritmo en una estimación del impacto monetario derivado de sus predicciones, aportando así una medida con potencial utilidad para la formulación de políticas climáticas. Para esta estimación se utilizó como referencia un precio del carbono de USD 50 por tonelada de CO_2, valor consistente con los rangos sugeridos por el World Bank en escenarios de transición energética y mercados de carbono [39].

El costo fue estimado mediante:

$$Costo = | y_i - \hat{y}_i | \times P_{CO_2}$$

donde P_{CO_2} representa el precio de referencia por tonelada de carbono. El costo total refleja el error acumulado sobre las observaciones del conjunto de prueba. Esta cifra debe interpretarse como un indicador comparativo entre modelos, no como una estimación del costo climático real, que requeriría proyección a la escala poblacional y temporal del problema.

Tabla 23 Costos económicos asociados al error predictivo

Modelo	Error promedio	Costo promedio USD	Costo total USD
XGBoost	0.17	8.26	702.07

Random Forest	0.17	8.37	711.10
KNN	0.17	8.71	739.97
Red Neuronal MLP	0.21	10.39	883.03
SVR	0.28	14.01	1190.61

Los resultados muestran que el modelo XGBoost presentó el menor costo promedio asociado al error de predicción (USD 8.26 por observación), seguido por Random Forest (USD 8.37) y KNN (USD 8.71). Por su parte, la Red Neuronal MLP registró un costo promedio de USD 10.39, mientras que el modelo SVR presentó el valor más alto, con USD 14.01 por observación.

Estos resultados indican que los modelos con mejor desempeño predictivo también tienden a minimizar el impacto económico asociado a los errores de estimación. En este contexto, XGBoost, Random Forest y KNN mostraron una mayor eficiencia en términos de costo, lo que resalta su potencial utilidad como herramientas de apoyo para procesos de análisis y planificación relacionados con las emisiones de CO_2 per cápita. No obstante, estos resultados deben interpretarse como una aproximación basada en el desempeño predictivo de los modelos y no como una estimación directa de impactos económicos reales.

5.6 Integración final de desempeño y costo predictivo

La evaluación del desempeño también incorporó el costo económico asociado al error de predicción, complementando las métricas tradicionales utilizadas previamente. Esto permitió comparar los modelos no solo desde su precisión estadística, sino también desde el impacto económico potencial derivado de las desviaciones en la estimación de emisiones de CO_2 per cápita.

Tabla 24 Tabla final. Integración de métricas predictivas y costo económico

Modelo	R2	RMSE	MAE	MedAE	MAPE (%)	Error promedio	Costo promedio USD	Costo total USD
KNN	0.736	0.250	0.174	0.103	1.231.902	0.174	8.705	739.966
XGBoost	0.728	0.253	0.165	0.090	1.418.854	0.165	8.260	702.075
Random Forest	0.672	0.278	0.167	0.076	2.015.480	0.167	8.366	711.102
Red Neuronal MLP	0.662	0.282	0.208	0.159	1.370.783	0.208	10.389	883.031
SVR	0.107	0.459	0.280	0.128	2.157.658	0.280	14.007	1.190.615
Deep MLP	-0.112	0.513	0.330	0.160	3.411.634	0.330	16.520	1.404.164

La integración de los resultados evidencia que los modelos KNN y XGBoost presentaron el mejor desempeño global dentro del conjunto de algoritmos evaluados. El modelo KNN alcanzó la mayor capacidad predictiva, registrando un coeficiente de determinación de $R^2 = 0.736$, mientras que XGBoost obtuvo el menor error absoluto medio (MAE = 0.165) y el menor costo económico promedio asociado a los errores de predicción (USD 8.26 por observación).

En términos generales, ambos modelos mostraron resultados consistentes y superiores a los obtenidos por Random Forest, MLP, SVR y Deep MLP. Estos hallazgos sugieren que los enfoques basados en proximidad entre observaciones (KNN) y ensamblaje de árboles potenciados mediante gradiente (XGBoost) ofrecen una combinación favorable de capacidad predictiva y eficiencia en el contexto analizado.

Por el contrario, el modelo SVR registró un desempeño considerablemente inferior al observado en las demás alternativas, alcanzando un R^2 de 0.107 y el mayor costo económico promedio entre los modelos tradicionales evaluados. Asimismo, la arquitectura Deep MLP presentó el desempeño más bajo del estudio, con un R^2 negativo (-0.112), lo que indica una capacidad predictiva limitada sobre el conjunto de prueba temporal.

En conjunto, los resultados respaldan la utilidad de modelos de aprendizaje automático no lineales para identificar patrones asociados a las emisiones de CO_2 per cápita y generar predicciones con niveles aceptables de precisión. No obstante, los hallazgos deben interpretarse como evidencia de capacidad predictiva y no como relaciones causales entre las variables consideradas.

6 OBJETIVO 5: Desarrollar una visualización interactiva que facilite la interpretación de patrones regionales de desigualdad climática.

El tablero se desarrolló como un instrumento de comunicación científica que materializa tres dimensiones fundamentales del trabajo: la descripción del fenómeno (Objetivo 1 y 2), la comparación de modelos y la explicabilidad de sus resultados (Objetivo 3 y 4), y la exploración prospectiva mediante simulación de escenarios de política. El diseño responde a los principios de la visualización orientada a la acción propuestos por Ware (2020)[40] y a los criterios de comunicación de la incertidumbre en ciencias del cambio climático desarrollados por Hullman y Diakopoulos (2011) [41].

6.1 Principios de diseño visual

El diseño visual del tablero sigue cuatro principios derivados de la literatura en visualización científica. En primer lugar, el principio de *parsimonia visual*: cada elemento gráfico presente tiene una función informativa específica; no se incluyen decoraciones, gradientes o elementos que no aporten información (Tufte, 2001) [42]. En segundo lugar, el principio de *coherencia cromática*: la escala de color verde-ámbar-rojo se aplica de forma consistente a lo largo de todos los módulos para representar los tres niveles de emisión (Bajo, Medio, Alto), de modo que el usuario no deba aprender nuevas codificaciones al avanzar entre secciones.

En tercer lugar, el principio de *jerarquía de información*: los indicadores clave (KPI) se presentan en la parte superior de cada módulo, permitiendo una lectura rápida antes de descender al nivel de detalle. En cuarto lugar, el principio de *accesibilidad técnica*: el tablero opera sin instalación de software adicional en su versión HTML, garantizando su uso en entornos con restricciones de instalación, característica valorada en contextos de evaluación académica y política pública.

6.2 Selección de la herramienta de visualización

Se optó por el desarrollo de una herramienta de visualización interactiva diseñada con un doble propósito. Por un lado, busca complementar el análisis gráfico presentado a lo largo del trabajo y, por otro, se proyecta como un instrumento que permite la interpretación sencilla de resultados complejos. Los gráficos estáticos presentan una restricción dimensional, ya que solo un estado de los datos es visible en un momento dado. Esta limitación resulta particularmente inconveniente cuando el objeto de análisis involucra múltiples países, años, variables y modelos de forma simultánea (Shneiderman, 1996)[43].

La visualización interactiva permite al usuario aplicar el principio de overview first, zoom and filter, details on demand (Shneiderman, 1996), que constituye el fundamento operativo del tablero. Bajo esta lógica, un investigador puede partir de la visión regional mediante el módulo de Panorama para identificar un país de interés, examinar su trayectoria temporal, contrastar su perfil socioeconómico con el de otra nación a través de un radar chart y, finalmente, simular el comportamiento de sus emisiones bajo un escenario hipotético de política energética.

6.3 Desarrollo del tablero

6.3.1 Implementación técnica del tablero

El tablero se desarrolló en dos versiones complementarias, cuya selección respondió a una evaluación técnica basada en la facilidad de despliegue, la compatibilidad con entornos académicos y la capacidad de integración con modelos de aprendizaje automático.

La versión principal consiste en un archivo HTML generado mediante Python, el cual incorpora la biblioteca Plotly.js 2.27.0 como motor de visualización y un conjunto de datos embebidos en formato JSON. Esta arquitectura elimina la dependencia de servidores externos y garantiza la funcionalidad del tablero en cualquier navegador moderno sin requerir instalaciones adicionales. Dicha característica es fundamental para asegurar la reproducibilidad en contextos de evaluación académica y facilitar la diseminación de resultados (Perkel, 2018) [44].

Por su parte, la versión secundaria fue desarrollada en Python utilizando el framework Streamlit, lo que permite la integración directa de los modelos entrenados con scikit-learn. En esta implementación, el simulador de política ejecuta el modelo Random Forest en tiempo real mediante la función predict. Para mantener la consistencia entre ambas versiones, la entrega en HTML utiliza una interpolación por vecinos más cercanos (KNN, k=5) sobre una grilla de 600 predicciones

precalculadas del mismo modelo, asegurando así la estabilidad y precisión de los resultados en el formato estático.

6.3.2 Módulo de simulación: fundamento metodológico

El simulador de política (Módulo 6) constituye uno de los componentes aplicados más relevantes de la investigación, al permitir explorar escenarios hipotéticos a partir de diferentes configuraciones de variables socioeconómicas y ambientales. Su propósito es estimar el nivel esperado de emisiones de CO_2 per cápita asociado a un conjunto específico de características observadas, proporcionando una herramienta interactiva para el análisis de escenarios.

El modelo subyacente utilizado en el simulador corresponde al algoritmo KNN, seleccionado como el modelo de mejor desempeño predictivo dentro de los algoritmos evaluados. Para evitar fuga de información y preservar la estructura temporal de los datos, el modelo fue entrenado utilizando 442 observaciones correspondientes al período 1990–2015 y evaluado sobre 85 observaciones del período 2016–2020. Los hiperparámetros óptimos identificados mediante Grid Search fueron 20 vecinos (n_neighbors = 20), distancia Manhattan y ponderación por distancia (weights = distance). Sobre el conjunto de prueba, el modelo alcanzó un coeficiente de determinación $R^2 = 0.7364$, un RMSE = 0.2495 y un MAE = 0.1741 toneladas de CO_2 per cápita.

Los controles incluidos en el simulador fueron seleccionados a partir del análisis de contribución relativa obtenido mediante SHAP para el modelo KNN. Entre las variables con mayor aporte a las predicciones se encuentran el consumo de energía, el PIB per cápita, el área selvática, la población urbana, la población total y la actividad industrial. Estas variables permiten explorar cómo diferentes configuraciones de características socioeconómicas y ambientales se asocian con distintos niveles estimados de emisiones de CO_2 per cápita dentro del marco predictivo desarrollado en esta investigación.

Es importante señalar que los escenarios generados por el simulador representan predicciones derivadas de patrones identificados en los datos históricos y no deben interpretarse como relaciones causales ni como estimaciones determinísticas del efecto de una política específica. En consecuencia, los resultados constituyen una herramienta de apoyo para el análisis exploratorio de escenarios y la toma de decisiones basada en evidencia predictiva.

Los rangos de los controles deslizantes están acotados a los valores mínimos y máximos observados en la región durante el período de análisis, lo que limita la extrapolación a zonas donde el modelo no fue entrenado.

Los resultados del simulador son orientativos para la exploración de escenarios y no constituyen proyecciones temporales ni reemplazan análisis de política formal. Esta advertencia metodológica se incluye explícitamente en la interfaz del tablero.

6.4 Resultados

6.4.1 Instrucciones de uso y acceso

El archivo HTML es completamente autocontenido y no requiere conexión a internet para su funcionamiento. Para acceder al tablero, el usuario debe abrir el archivo utilizando cualquier navegador moderno como Google Chrome, Mozilla Firefox, Microsoft Edge o Safari. La navegación se realiza a través de las pestañas superiores que organizan el contenido en los seis módulos funcionales descritos anteriormente.

Todas las visualizaciones son interactivas; al posicionar el cursor sobre cualquier elemento, se despliega un cuadro de información con los valores exactos. Asimismo, los gráficos permiten realizar acercamientos, desplazamientos y exportaciones directas a formato de imagen mediante las herramientas nativas de Plotly. En el módulo de simulación, el usuario puede manipular los controles deslizantes para generar perfiles socioeconómicos hipotéticos y observar la respuesta del modelo en tiempo real.

El tablero se encuentra disponible para su consulta en el siguiente enlace: https://sandra-guzman.github.io/-Desigualdad-Clim-tica-en-Am-rica-Latina/ .El acceso abierto a esta herramienta tiene como propósito fundamental que los resultados de este proyecto sean divulgados y comprendidos por audiencias no técnicas, contribuyendo así al debate sobre la justicia ambiental y la desigualdad climática. Esta iniciativa representa una apuesta por fortalecer el impacto de la academia en la construcción de soluciones para problemáticas de alta relevancia actual, tanto para el contexto nacional como para el conjunto de América Latina.

6.4.2 Organización de la información

El tablero se organiza en seis secciones, cada una de las cuales se encuentra asociada a un conjunto específico de interrogantes analíticas derivadas de los objetivos de la investigación. Esta estructura modular ha sido diseñada de manera deliberada para permitir que el usuario acceda con precisión al nivel de profundidad que sea de su interés, evitando la necesidad de recorrer la totalidad del contenido de forma lineal.

Esta organización facilita una navegación intuitiva y eficiente, lo que asegura que la herramienta responda a diversas necesidades de consulta técnica o estratégica. Al mismo tiempo, este diseño permite establecer un hilo conductor o storytelling en el abordaje del tema, guiando al lector a través de los hallazgos de forma coherente. La Tabla 1 presenta la correspondencia detallada entre los módulos, las preguntas analíticas y los objetivos generales del proyecto.

Tabla 25 Módulos del tablero y correspondencia con los objetivos de la investigación

Módulo	Nombre	Pregunta analítica principal	Objetivo de tesis
1	Panorama Regional	¿Quién emite qué en América Latina y con qué magnitud?	Objetivo 1 y 2 — Descripción y EDA
2	Evolución Temporal	¿Cómo han cambiado las emisiones en el período 1990-2020?	Objetivo 1 — Series de tiempo

3	Relaciones Socioeconómicas	¿Qué variables se correlacionan con el CO_2 per cápita?	Objetivo 2 — EDA y correlaciones
4	Importancia de Variables	¿Qué variables explican y predicen mejor las emisiones?	Objetivo 3 — FE, ML, SHAP
5	Perfiles de Países	¿En qué nivel de emisiones se clasifica cada país y por qué?	Objetivo 3 — Clasificación
6	Simulador de Política	¿Qué ocurriría con las emisiones bajo escenarios alternativos?	Objetivo 3 — Uso aplicado del modelo

Nota. *FE = Efectos Fijos; ML = Machine Learning; EDA = Análisis Exploratorio de Datos. Elaboración propia.*

Ilustración 18 Vista preliminar del tablero en su versión HTML

El desarrollo del tablero interactivo se fundamenta en la necesidad de transformar datos climáticos y socioeconómicos complejos en información accionable y comprensible. A través de una arquitectura organizada, la herramienta permite transitar desde una visión macro de la región hasta la simulación de escenarios específicos, integrando en cada etapa los hallazgos estadísticos y los modelos de aprendizaje automático desarrollados en los objetivos previos de esta investigación.

El Módulo 1, denominado Panorama Regional, ofrece una caracterización exhaustiva de la magnitud y distribución de las emisiones en América Latina mediante la integración de 527 observaciones correspondientes a los 17 países del estudio durante el periodo 1990 a 2020. A través de un mapa coroplético y un sistema de indicadores clave, se evidencia una marcada desigualdad climática estructural donde la brecha entre el mayor emisor, Chile con una media de 1.183 toneladas de CO_2 per cápita, y el menor emisor, Paraguay con 0.003 toneladas, alcanza una magnitud de 430 veces. Esta sección permite al usuario identificar patrones espaciales y estadísticos mediante un ranking por terciles y un histograma de frecuencias que revela una asimetría positiva, confirmando que, si bien la mayoría de los registros se sitúan en niveles bajos, existe una concentración crítica de emisiones en las economías con mayor nivel de industrialización y Producto Interno Bruto de la región.

El Módulo 2, titulado Evolución Temporal, analiza las trayectorias históricas de las emisiones para identificar patrones de crecimiento, estabilización o reducción en el periodo comprendido entre 1990 y 2020. Mediante el uso de gráficos de series de tiempo interactivos y un panel de pequeños múltiplos (small multiples), esta sección permite visualizar el impacto de hitos macroeconómicos globales, tales como la crisis financiera de 2008 y la contracción generada por la pandemia de 2020. La herramienta facilita la comparación directa entre los diecisiete países analizados, desde Argentina hasta Venezuela, permitiendo que el investigador contraste cómo las variaciones en la matriz energética y el dinamismo económico han moldeado trayectorias divergentes. Esta disposición técnica evita la superposición de series y garantiza una lectura clara de la evolución individual de cada economía, proporcionando la evidencia necesaria para comprender la temporalidad de la desigualdad climática en la región.

El Módulo 3, centrado en las Relaciones Socioeconómicas, profundiza en los determinantes de las emisiones mediante una interfaz de análisis bivariado que vincula el CO_2 per cápita con indicadores clave de desarrollo. A través de un gráfico de burbujas dinámico, donde el volumen de cada unidad representa la población total, se validan visualmente las correlaciones identificadas en el Análisis Exploratorio de Datos, destacando vínculos positivos con el consumo de energía (r = 0.77), el PIB (r = 0.58) y el IDH (r = 0.54), así como relaciones inversas con el área selvática (r = -0.38) y el índice de Gini (r = -0.29). La sección incluye una matriz de correlaciones de Pearson que facilita la identificación de multicolinealidad y orienta la interpretación de los modelos de Efectos Fijos y Aleatorios, asegurando que las relaciones observadas en el tablero sean plenamente consistentes con los resultados estadísticos y las pruebas de robustez, como la de Hausman, desarrolladas en la investigación.

El Módulo 4, dedicado a la Importancia de las Variables, constituye el núcleo analítico de la investigación al triangular tres enfoques metodológicos para identificar los determinantes de las emisiones. En primer lugar, se presentan los resultados de la modelación econométrica de panel, donde la prueba de Hausman orientó la selección de modelos de efectos aleatorios. En segundo lugar, se exhibe el desempeño de los algoritmos de aprendizaje automático evaluados con validación temporal (1990–2015 entrenamiento, 2016–2020 prueba), destacando el modelo KNN como el de mayor capacidad predictiva (R^2 = 0,736), seguido por XGBoost (R^2 = 0,728). Finalmente, se utiliza el método de explicabilidad SHAP con KernelExplainer sobre el modelo SVR para descomposiciones de valor, confirmando que el consumo de energía es la variable con mayor valor SHAP medio absoluto (**0.1004**) y el determinante más robusto en todas las aproximaciones. Esta sección integra una tabla de consistencia metodológica que valida la validez interna de los hallazgos al cruzar los resultados de los efectos fijos, la importancia por impureza de Gini y los valores SHAP, permitiendo clasificar el consumo energético, el uso de energía fósil y el área selvática como variables de consistencia alta.

El Módulo 5, titulado Perfiles de los Países, presenta la categorización de los países estudiados en tres niveles de emisión (Bajo, Medio y Alto), definidos mediante la técnica de terciles aplicada al promedio de CO_2 per cápita del periodo 1990 a 2020. Esta clasificación permite identificar patrones

estructurales donde el grupo de emisiones altas, integrado por Chile, México, Argentina, Venezuela y Panamá, se distingue por mayores niveles de Producto Interno Bruto y una matriz energética con alta dependencia de combustibles fósiles. Para validar esta estructura, se emplearon modelos de clasificación supervisada como Random Forest, XGBoost y SVM, los cuales alcanzaron niveles de precisión entre el 86.8% y el 88.7% en el conjunto de prueba. Asimismo, el módulo incorpora un gráfico de radar (Radar Chart) que facilita la comparación directa de perfiles socioeconómicos normalizados entre dos países, permitiendo al usuario visualizar de forma intuitiva cómo variables como el IDH, el índice de Gini y la cobertura selvática configuran polígonos diferenciados que explican la posición de cada economía en la jerarquía de emisiones regional.

El Módulo 6, correspondiente al Simulador de Política Climática, representa el componente prospectivo del tablero, permitiendo a los usuarios realizar ejercicios de exploración contrafáctica mediante la manipulación de las seis variables con mayor importancia combinada según los análisis de Random Forest y SHAP. Para garantizar la agilidad de la herramienta en su versión web, el sistema estima el CO_2 per cápita esperado utilizando una interpolación ponderada por vecinos más cercanos (KNN, k=5) sobre una grilla de 600 predicciones precalculadas del modelo Random Forest original, el cual posee un coeficiente de determinación R^2 = 0.9482. Esta funcionalidad permite simular escenarios hipotéticos, como la reducción de la dependencia de energías fósiles o el incremento de la cobertura forestal, acotando los rangos de control a los valores mínimos y máximos observados en la región para evitar extrapolaciones fuera del dominio de entrenamiento. Los resultados se presentan junto a una barra de posición relativa que ubica el perfil simulado respecto a los promedios históricos regionales, ofreciendo una base metodológica transparente que incluye advertencias sobre la naturaleza orientativa de las proyecciones y la mención de la prueba de Dietterich 5×2CV para validar la solidez del modelo empleado.

En conjunto, esta estructura no solo garantiza la navegabilidad según el perfil del usuario, sino que actúa como un hilo conductor que conecta la descripción de los datos con la capacidad predictiva de los modelos. De este modo, el tablero se consolida como un instrumento de comunicación científica que facilita la interpretación de la desigualdad climática en América Latina desde una perspectiva multivariada y dinámica.

6.5 Limitaciones y consideraciones éticas

Es fundamental precisar que el tablero opera sobre datos históricos del periodo 1990 a 2020 y no constituye un sistema de proyección temporal. Las predicciones del simulador indican el nivel de emisiones que el modelo asocia a un perfil específico y no deben interpretarse como pronósticos de eventos futuros. Esta distinción es crucial para evitar inferencias causales que los modelos correlacionales no pueden respaldar.

Desde el punto de vista técnico, el análisis SHAP se aplicó mediante una técnica de aproximación debido a las exigencias computacionales de los métodos agnósticos al modelo. Aunque esta metodología es robusta, los valores resultantes son estimaciones que pueden presentar variaciones

en submuestras reducidas [23]. Asimismo, la exclusión de países con baja cobertura de datos introduce un sesgo de selección que debe considerarse al intentar generalizar estos hallazgos a la totalidad de la región.

Finalmente, desde una perspectiva ética, se advierte que esta herramienta tiene fines de exploración y comunicación científica. Su uso en la formulación de políticas públicas debe estar siempre acompañado de análisis sectoriales detallados y de la validación de expertos en economía ambiental, ya que el tablero complementa, pero no sustituye, el juicio profesional calificado.

6.6 Conclusiones

El presente estudio permitió construir una base de datos panel consolidada para 17 países de América Latina durante el período 1990–2020, integrando variables socioeconómicas, energéticas, ambientales y demográficas provenientes de fuentes oficiales internacionales. El proceso de depuración, análisis de completitud e imputación de valores faltantes permitió obtener una base consistente para el desarrollo de modelos explicativos y predictivos relacionados con las emisiones de CO_2 per cápita.

Los resultados del análisis exploratorio y de los modelos econométricos evidenciaron que variables asociadas al desarrollo económico y al consumo energético presentan una relación significativa con las emisiones de CO_2 per cápita. En particular, el PIB per cápita, el uso de energía, el consumo energético y la participación de fuentes fósiles mostraron una influencia relevante sobre el comportamiento de las emisiones en los países analizados.

La comparación entre modelos econométricos y algoritmos de aprendizaje automático permitió evidenciar la complementariedad entre ambos enfoques. Mientras los modelos de datos panel proporcionan herramientas para interpretar relaciones estructurales entre variables y controlar la heterogeneidad no observada entre países, los modelos de machine learning demostraron una mayor capacidad predictiva al capturar relaciones no lineales presentes en los datos.

Esta complementariedad se hace evidente en la divergencia entre el R^2 del modelo de efectos fijos (0,307) y los R^2 de los mejores algoritmos de aprendizaje automático (0,736 para KNN, 0,728 para XGBoost). Dicha divergencia no constituye una contradicción metodológica, sino la expresión de dos objetivos analíticos fundamentalmente distintos: el modelo de efectos fijos estima asociaciones intrapaís a lo largo del tiempo, eliminando por construcción toda la variabilidad between-countries (que es justamente donde se concentra la mayor parte de la heterogeneidad en emisiones per cápita). Su R^2 within captura únicamente la variación temporal dentro de cada país, lo que naturalmente restringe su capacidad explicativa global. Los modelos de ML, en cambio, operan sobre la totalidad de la variabilidad del conjunto de datos, tanto temporal como transversal, sin imponer restricciones sobre la forma funcional de las relaciones, lo que les permite capturar patrones que un modelo lineal de datos panel no puede representar. Esta distinción reafirma que el poder explicativo-estructural de la econometría y el poder predictivo del aprendizaje automático son herramientas

complementarias, no sustitutos

Atendiendo las recomendaciones metodológicas derivadas del proceso de evaluación, la capacidad predictiva de los algoritmos fue evaluada mediante una validación temporal utilizando observaciones entre 1990 y 2015 para entrenamiento y el período 2016–2020 para prueba. Esta estrategia permitió evitar problemas de fuga de información temporal y obtener una estimación más realista del desempeño de los modelos frente a datos no observados.

Entre los algoritmos evaluados, K-Nearest Neighbors (KNN) presentó el mejor desempeño predictivo, alcanzando un coeficiente de determinación de 0,736, un RMSE de 0,250 y un MAE de 0,174. XGBoost mostró resultados muy cercanos, con un R^2 de 0,728 y un menor costo económico promedio asociado al error de predicción. Estos resultados confirman el potencial de los enfoques basados en aprendizaje automático para modelar fenómenos ambientales complejos en contextos caracterizados por alta heterogeneidad estructural.

No obstante, los resultados deben interpretarse considerando las limitaciones asociadas al proceso de imputación de datos faltantes. Aunque la interpolación lineal permitió preservar la continuidad temporal de las series y mantener la representatividad de la muestra, este procedimiento supone trayectorias relativamente suaves entre observaciones consecutivas, lo que podría no reflejar completamente cambios abruptos ocurridos en algunos países durante el período analizado.

Finalmente, este trabajo no pretende medir de manera directa la desigualdad climática en toda su complejidad conceptual. En cambio, analiza la relación entre factores socioeconómicos y emisiones de CO_2 per cápita, aportando evidencia empírica sobre una dimensión relevante para comprender las diferencias ambientales existentes entre los países de América Latina. Desde esta perspectiva, los resultados constituyen una aproximación al estudio de la desigualdad climática en la región, al permitir identificar cómo las condiciones económicas y sociales se relacionan con los patrones diferenciados de emisiones entre países.

Comprender estas particularidades resulta fundamental para el diseño de políticas públicas orientadas a enfrentar los desafíos del cambio climático, considerando que los países de la región no han contribuido de la misma manera a las emisiones históricas de CO_2 y, al mismo tiempo, presentan capacidades diferenciadas de adaptación y respuesta frente a sus impactos. En este sentido, el análisis se relaciona con el enfoque de justicia climática, el cual reconoce que los efectos del cambio climático no se distribuyen de manera equitativa: algunos países con menores niveles de contribución a las emisiones pueden enfrentar mayores vulnerabilidades y consecuencias ambientales. Por lo tanto, este estudio busca aportar elementos que contribuyan a una comprensión más contextualizada del problema climático en América Latina y sirvan como insumo para futuras investigaciones y estrategias de sostenibilidad, mitigación y adaptación.

6.7 Trabajos Futuros

Como líneas de investigación futura, se recomienda incorporar variables relacionadas con gobernanza, calidad institucional, regulación ambiental, innovación tecnológica, transición

energética y capacidad estatal, con el fin de ampliar la comprensión de los factores que influyen sobre las emisiones de CO_2 per cápita en la región.

Asimismo, futuras investigaciones podrían incluir indicadores de vulnerabilidad climática, exposición a eventos extremos, resiliencia territorial y capacidad de adaptación, permitiendo aproximarse de manera más integral al concepto de desigualdad climática y superar las limitaciones inherentes al análisis centrado exclusivamente en emisiones.

Desde una perspectiva metodológica, resulta pertinente evaluar métodos alternativos de imputación de datos faltantes, tales como MICE, MissForest y KNN Imputer, con el propósito de comparar su desempeño frente a la interpolación lineal y analizar la sensibilidad de los resultados obtenidos.

También se recomienda implementar esquemas de validación temporal más avanzados, como rolling window o expanding window, que permitan evaluar la estabilidad de los modelos frente a diferentes horizontes temporales y posibles cambios estructurales en los datos.

Finalmente, futuras investigaciones podrían explorar modelos espacio-temporales y enfoques híbridos que integren econometría y aprendizaje automático, aprovechando simultáneamente la capacidad explicativa de los modelos estadísticos y la potencia predictiva de los algoritmos de inteligencia artificial para el análisis de fenómenos ambientales complejos. Además, se plantea la posibilidad de replicar este enfoque en otras regiones del mundo con el objetivo de identificar diferencias en los factores socioeconómicos, políticos y ambientales que influyen sobre las emisiones de CO_2 per cápita. Esto permitiría generar comparaciones internacionales y contribuir a una comprensión más amplia de la desigualdad climática a nivel global, considerando que los patrones de emisión, las capacidades de adaptación y las vulnerabilidades frente al cambio climático varían significativamente entre territorios..

REFERENCIAS BIBLIOGRÁFICAS

[1] Intergovernmental Panel on Climate Change (IPCC, "Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report," Cambridge, United Kingdom, 2022.

[2] UNITED NATIONS, "The principles of equity and common but differentiated responsibilities and respective capabilities," New York, 1992. Accessed: May 27, 2025. [Online]. Available: https://unfccc.int/resource/docs/convkp/conveng.pdf

[3] C. Holz, E. Kemp-Benedict, T. Athanasiou, and S. Kartha, "The Climate Equity Reference Calculator," *J. Open Source Softw.*, vol. 4, no. 35, p. 1273, Mar. 2019, doi: 10.21105/joss.01273.

[4] A. Bárcena, J. Samaniego, W. Peres, and J. E. Alatorre, *La emergencia del cambio climático en América Latina y el Caribe: ¿Seguimos esperando la catástrofe o pasamos a la acción?*, Primera., vol. 160. Santiago: Comisión Económica para América Latina y el Caribe (CEPAL), 2020.

[5] L. Balza, L. Heras-Recuero, D. Matías, and A. Yépez-García, "Green or Growth? Understanding the Relationship between Economic Growth and CO2 Emissions," Washington, D. C., May 2024. doi: 10.18235/0012943.

[6] Sánchez Juárez Isaac, García Almada Rosa María, and Chávez Gutiérrez Ninfa Stella, "Emisiones de CO2 y crecimiento económico en la región de América Latina.," *Equilibrio Económico. Revista de Economía, Política y Sociedad*, vol. 20, pp. 6–32, Apr. 2024, doi: 10.5281/zenodo.14766655.

[7] Programa de las Naciones Unidas para el Desarrollo (PNUD), "El cambio climático es un asunto de justicia: he aquí por qué," Programa de las Naciones Unidas para el Desarrollo (PNUD). Accessed: May 27, 2025. [Online]. Available: https://climatepromise.undp.org/es/news-and-stories/el-cambio-climatico-es-un-asunto-de-justicia-he-aqui-por-que

[8] L. Chancel, Bothe Philipp, and Voituriez Tancréde, "Climate Inequality Report 2023," 2023.

[9] D. Lemus-Delgado and R. Pérez Navarro, "Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos," *Colombia Internacional*, no. 102, pp. 41–62, Apr. 2020, doi: 10.7440/colombiaint102.2020.03.

[10] Gareth. James, Daniela. Witten, Trevor. Hastie, and Robert. Tibshirani, *An introduction to statistical learning : with applications in R*. Springer : Springer Science+Business Media, 2017.

[11] Intergovernmental Panel on Climate Change (IPCC), "Climate Change 2021 – The Physical Science Basis," Cambridge University Press, Cambridge, Reino Unido y Nueva York, NY, EE. UU, Jul. 2023. doi: 10.1017/9781009157896.

[12] *Emissions Gap Report 2022 : The closing window : climate crisis calls for rapid transformation of societies*. United Nations Environment Programme, 2022.

[13] P. Friedlingstein *et al.*, "Global Carbon Budget 2024," *Earth Syst. Sci. Data*, vol. 17, no. 3, pp. 965–1039, Mar. 2025, doi: 10.5194/essd-17-965-2025.

[14] D. Suman, "Robert Bullard: Dumping in Dixie: Race, Class and Environmental Quality," *Ecol. Law Q.*, vol. 19, no. 3, pp. 591–609, 1992, [Online]. Available: http://www.jstor.org/stable/24113114

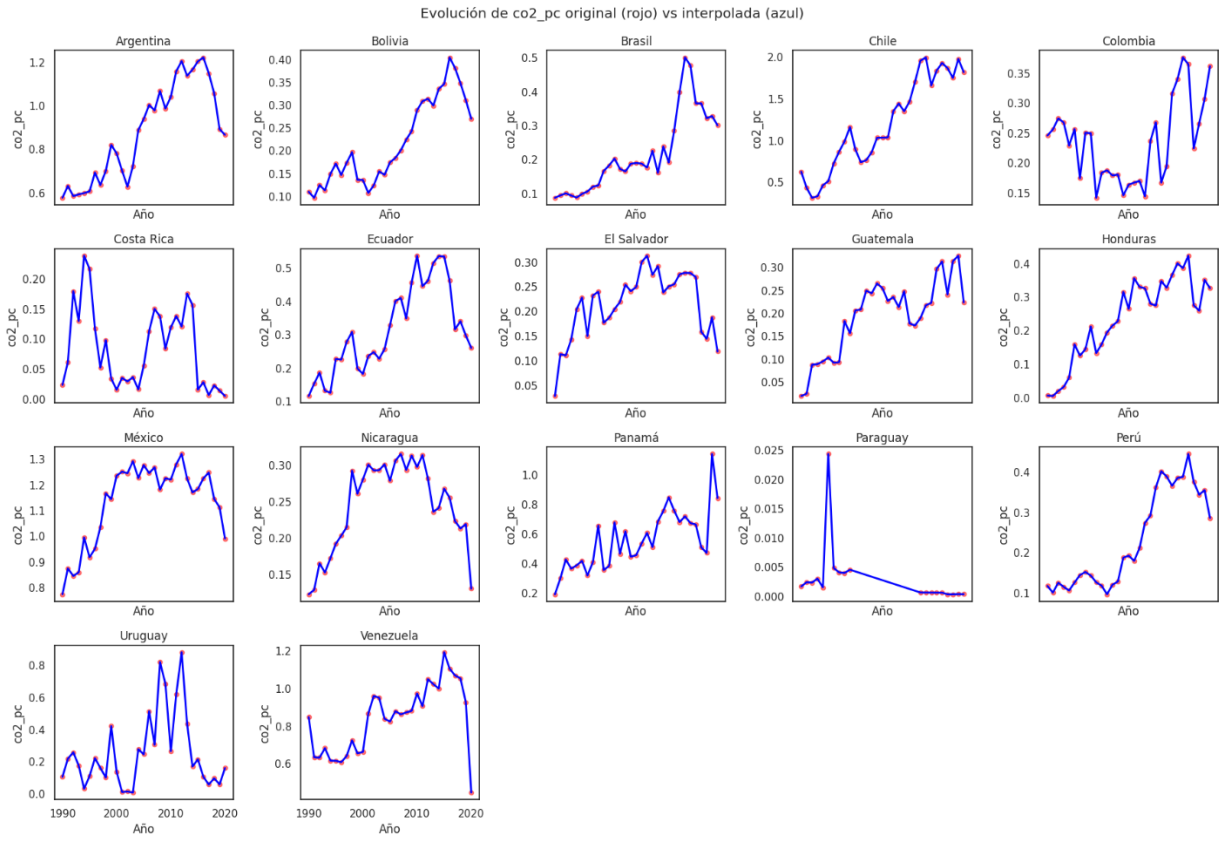
[15] University of Washington, "Climate Inequality," Climate Impacts Group. . Accessed: May 27, 2025. [Online]. Available: https://cig.uw.edu/learn/inequities-in-climate-impacts/

[16] N. Durmaz, Q. Liu, and Z. Lu, “CO2 Emission, Energy Consumption, and Economic Growth in Latin America,” *Energy RESEARCH LETTERS*, vol. 6, no. Early View, Feb. 2025, doi: 10.46557/001c.125655.
[17] M. Jiménez-García, M.-C. Pérez-Peña, and J.-A. López-Sánchez, “Influence of education and the media on the awareness of climate change (*Influencia de la educación y la comunicación en la concienciación sobre el cambio climático*),” *Culture and Education: Cultura y Educación*, vol. 35, no. 4, pp. 1001–1036, Dec. 2023, doi: 10.1080/11356405.2023.2229178.
[18] R. Van der Borght and M. Pallares Barbera, “How urban spatial expansion influences CO2 emissions in Latin American countries,” *Cities*, vol. 139, p. 104389, Aug. 2023, doi: 10.1016/j.cities.2023.104389.
[19] Z. Wang *et al.*, “Enormous inter-country inequality of embodied carbon emissions and its driving forces in South America,” *Global Environmental Change*, vol. 89, p. 102944, Dec. 2024, doi: 10.1016/j.gloenvcha.2024.102944.
[20] International Energy Agency, “Latin America Energy Outlook 2023,” Paris, 2023. Accessed: May 28, 2025. [Online]. Available: https://iea.blob.core.windows.net/assets/1055131a-8dc4-488b-9e9e-7eb4f72bf7ad/LatinAmericaEnergyOutlook.pdf
[21] Y. Robiou du Pont and M. Meinshausen, “The Climate Equity Reference Project: Toward a fair and effective climate regime,” *Environ. Sci. Policy*, vol. 113, pp. 75–85, Nov. 2020, doi: 10.1016/j.envsci.2020.06.007.
[22] S. Li, Y. W. Siu, and G. Zhao, “Driving Factors of CO2 Emissions: Further Study Based on Machine Learning,” *Front. Environ. Sci.*, vol. 9, Aug. 2021, doi: 10.3389/fenvs.2021.721517.
[23] A. L. Jiménez-Preciado, S. Cruz-Aké, and F. Venegas-Martínez, “Identification of Patterns in CO2 Emissions among 208 Countries: K-Means Clustering Combined with PCA and Non-Linear t-SNE Visualization,” *Mathematics*, vol. 12, no. 16, p. 2591, Aug. 2024, doi: 10.3390/math12162591.
[24] I. R. Hambleton and S. Jeyaseelan, “The silent barrier: exploring data availability in Small Island Developing States.,” *Rev. Panam. Salud Publica*, vol. 48, p. e80, 2024, doi: 10.26633/RPSP.2024.80.
[25] R. J. A. . Little and D. B. . Rubin, *Statistical analysis with missing data*. Wiley, 2020.
[26] “Handbook on Constructing Composite Indicators,” Aug. 2005. doi: 10.1787/533411815016.
[27] NU. CEPAL. División de Estadísticas, “Anuario Estadístico de América Latina y el Caribe 2020,” Mar. 2021.
[28] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, Aug. 2019, doi: 10.1080/08839514.2019.1637138.
[29] Badi H. Baltagi, *Econometric Analysis of Panel Data*, Third edition. England: John Wiley & Sons Ltd, 2008.
[30] J. M. . Wooldridge, *Econometric analysis of cross section and panel data*, Second edition., vol. volumen 1. Cambridge, Massachusetts : MIT Press, 2010.
[31] INTERNATIONAL ENERGY AGENCY, “CO2 Emissions in 2022,” 2022.
[32] P. Poumanyvong and S. Kaneko, “Does urbanization lead to less energy use and lower CO2 emissions? A cross-country analysis,” *Ecological Economics*, vol. 70, no. 2, pp. 434–444, Dec. 2010, doi: 10.1016/j.ecolecon.2010.09.029.
[33] J. A. Hausman, “Specification Tests in Econometrics,” *Econometrica*, vol. 46, no. 6, p. 1251, Nov. 1978, doi: 10.2307/1913827.

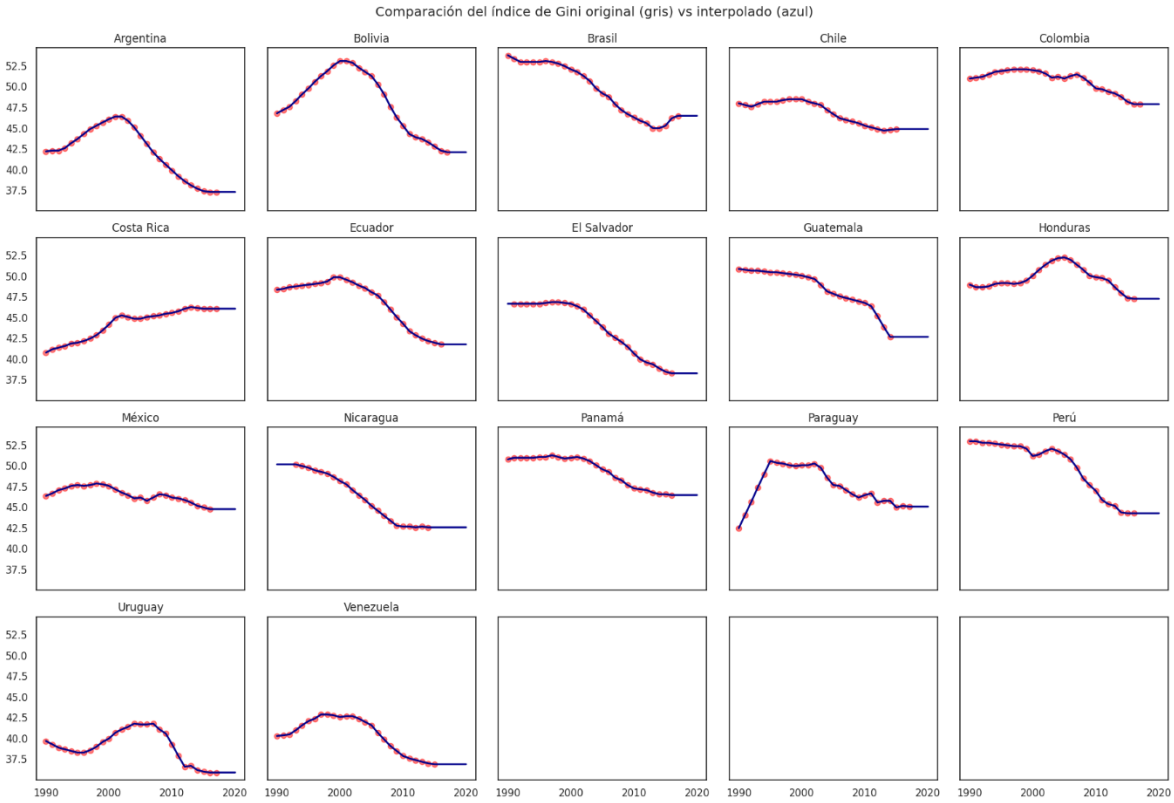
[34] B. H. . Baltagi, *Econometric analysis of panel data*. Springer, 2021.
[35] Elizabeth Goldman, Michelle Sims, Sarah Carter, and Peter Potapov, “Tropical Rainforest Loss Slowed in 2025, but Fire is a Growing Threat to Forests Worldwide,” *Global Forest Review*, Apr. 2026.
[36] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.
[37] World Bank, “State and Trends of Carbon Pricing 2023,” Washington, DC, 2023.
[38] C. Ware, *Information visualization : perception for design*, 2nd ed. in The Morgan Kaufmann series in interactive technologies. Amsterdam [etc: Elsevier [etc.], 2004.
[39] J. Hullman and N. Diakopoulos, “Visualization Rhetoric: Framing Effects in Narrative Visualization,” *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2231–2240, Dec. 2011, doi: 10.1109/TVCG.2011.255.
[40] E. Tufte, “The Visual Display Of Quantitative Information [Book Reviews],” *IEEE Power Engineering Review*, vol. 8, no. 2, pp. 20–20, Feb. 1988, doi: 10.1109/MPER.1988.587534.
[41] B. Shneiderman, “The eyes have it: a task by data type taxonomy for information visualizations,” in *Proceedings 1996 IEEE Symposium on Visual Languages*, IEEE Comput. Soc. Press, pp. 336–343. doi: 10.1109/VL.1996.545307.
[42] J. M. Perkel, “Why Jupyter is data scientists’ computational notebook of choice,” *Nature*, vol. 563, no. 7729, pp. 145–146, Nov. 2018, doi: 10.1038/d41586-018-07196-1.
[43] S. A. Lundberg and S.-I. P. Lee, “A Unified Approach to Interpreting Model Predictions ,” *Neural Information Processing Systems* , Dec. 2017.

Anexos

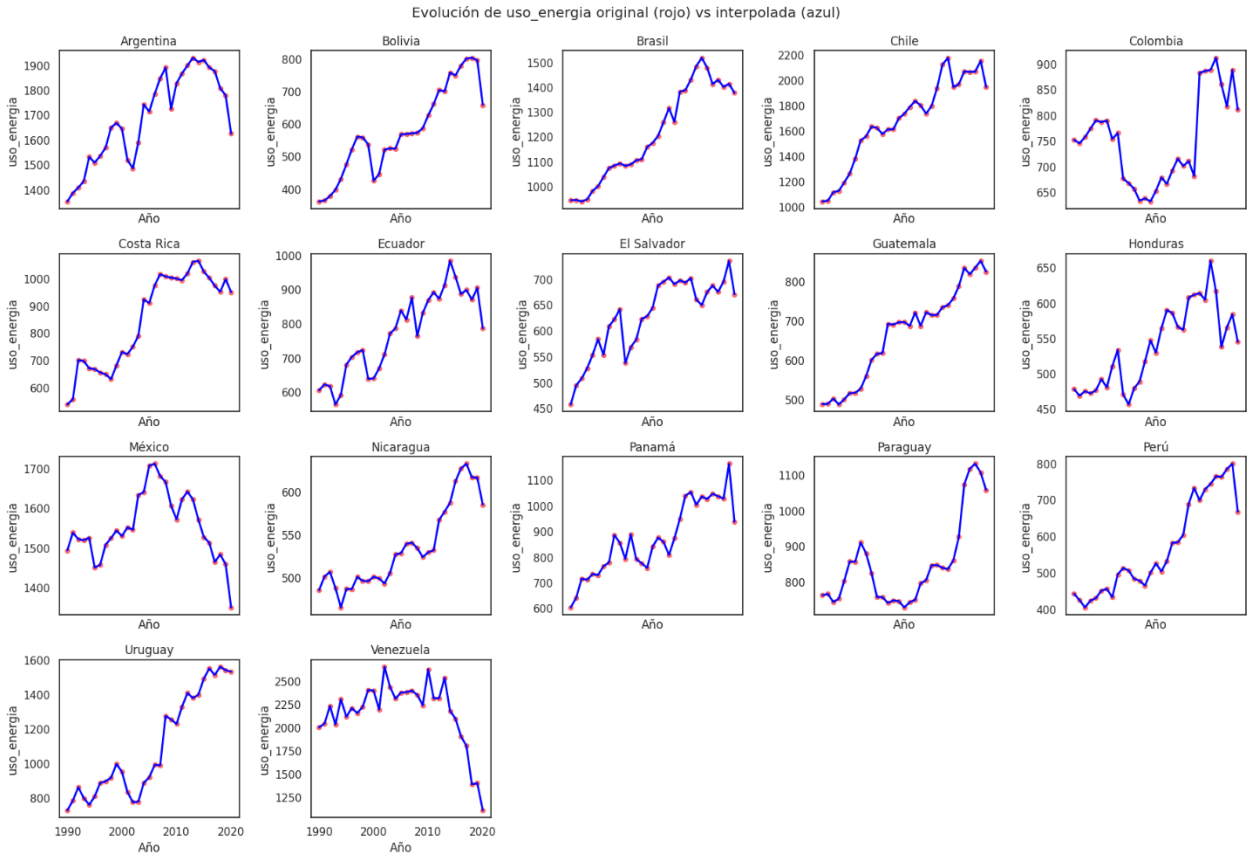
Anexo 1. Evolución del CO_2 per cápita en la base de datos original e interpolada



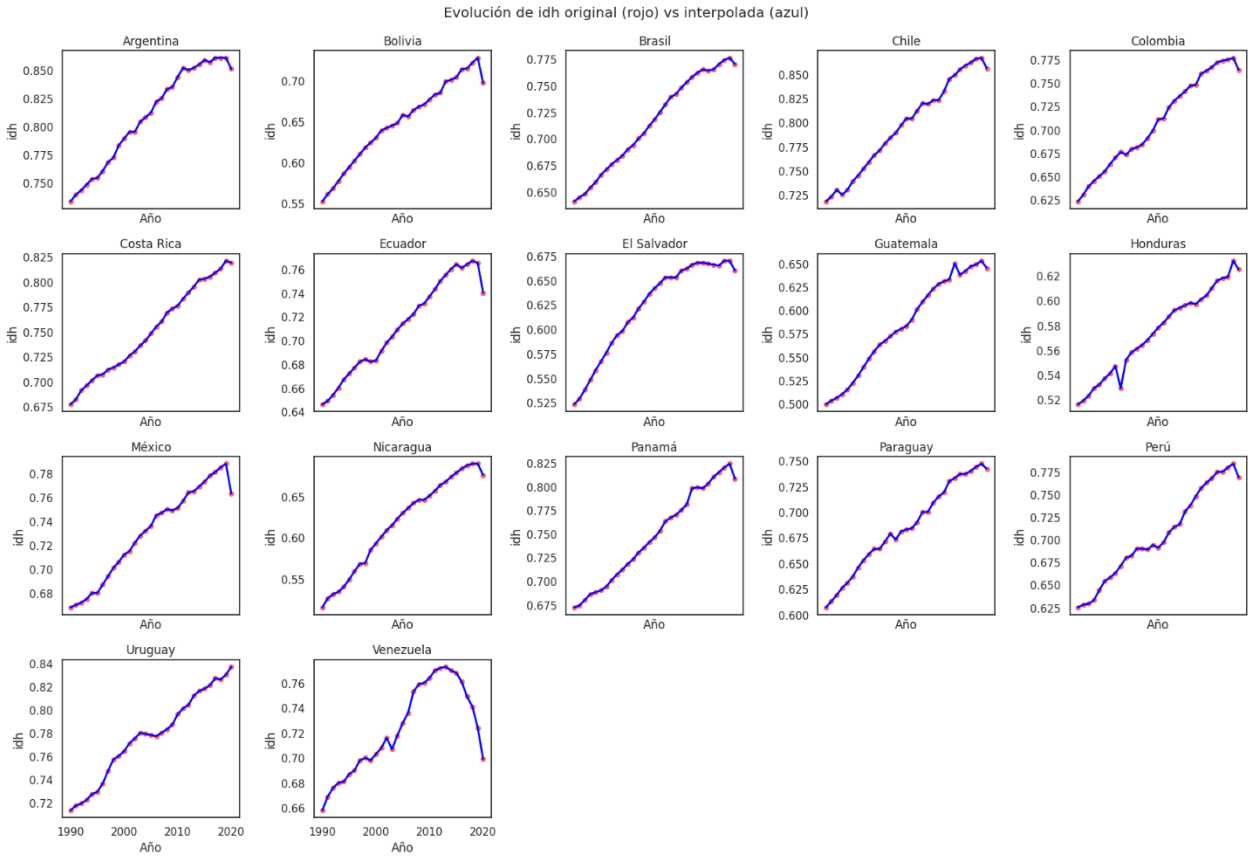
Anexo 2. Comparación del índice de Gini original e interpolado



Anexo 3. Evolución de la variable uso_energia en la base de datos original e interpolada



Anexo 4. Evolución del IDH en la base de datos original e interpolada



Anexo 5 . Evolución del PIB per cápita en la base de datos original e interpolada