



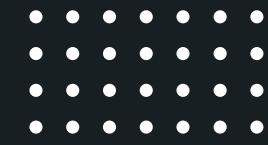
SMART ECONOMY

Yeison David Giraldo

05 de Junio 2025

Basic course on artificial intelligence

Udea



INTRODUCCIÓN DEL PROYECTO

El Producto Interno Bruto (PIB) per cápita es uno de los indicadores más utilizados para medir el nivel de desarrollo económico de los países, ya que refleja la producción de bienes y servicios en relación con la población. Su análisis permite comprender el bienestar económico promedio de los habitantes y es clave en el diseño de políticas públicas.

En la actualidad, gobiernos e instituciones enfrentan el reto de tomar decisiones informadas en entornos complejos y cambiantes. Gracias a los avances en ciencia de datos y aprendizaje automático, ahora es posible utilizar grandes volúmenes de información para anticipar tendencias económicas y sociales. En este contexto, surge la necesidad de explorar cómo diversas variables –como el acceso a la educación, la salud, la conectividad digital y la sostenibilidad ambiental– influyen en el desempeño económico de un país.

💡 DEFINICIÓN: ¿QUÉ ES EL PIB PER CÁPITA?

El PIB per cápita (Producto Interno Bruto por habitante) es una medida económica que calcula el valor total de todos los bienes y servicios producidos en un país en un año, dividido por la cantidad de habitantes.

Sirve como un indicador del nivel de vida promedio de una población y es comúnmente utilizado para comparar el desarrollo económico entre países o regiones.

Fórmula:

$$PIB\ pc = \frac{PIB}{Población}$$





¿Por qué es importante predecir el PIB per cápita?

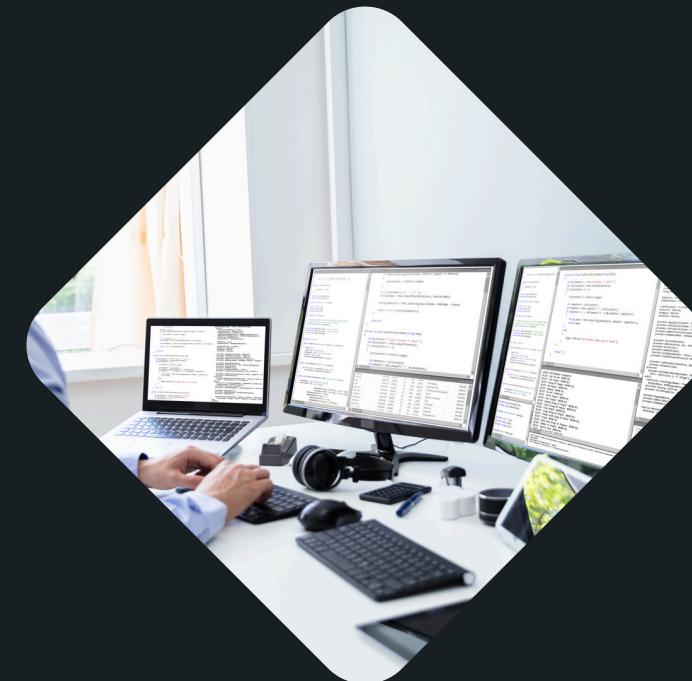
- Mide el nivel de vida: Indica cuánto produce un país por persona.
- Guía para políticas públicas: Ayuda a planificar inversiones en salud, educación y desarrollo.
- Detecta desigualdades: Identifica regiones o sectores con bajo crecimiento.
- Evalúa políticas económicas: Mide el impacto de decisiones gubernamentales.
- Atrae inversión: Proyecciones confiables generan confianza en el país.





PROBLEMA

¿Qué problema intenta resolver este proyecto?



Muchos países en desarrollo enfrentan desafíos para identificar qué factores sociales, económicos o tecnológicos tienen mayor impacto en su crecimiento económico. Esto limita la capacidad de diseñar políticas efectivas para mejorar la calidad de vida de su población.

Actualmente, el análisis tradicional del PIB per cápita no siempre logra capturar relaciones complejas entre variables como salud, educación, conectividad o medio ambiente.

🏛️ ¿A qué sector va dirigido este proyecto?

🎯 Sector objetivo: Sector Público y Organismos de Desarrollo
Este proyecto está dirigido principalmente a:



Gobiernos nacionales y locales,
especialmente de países en vías de
desarrollo.

Organizaciones internacionales
como el Banco Mundial, ONU y el
FMI .

**Instituciones de planeación
económica y social como el
DANE,DNP,Banco de la
República a nivel nacional**



¿POR QUÉ ES RELEVANTE?

- Permite anticipar cambios económicos según el comportamiento de indicadores clave.
- Ayuda a definir prioridades de inversión pública (salud, educación, conectividad).
- Apoya la formulación de políticas basadas en evidencia y datos históricos reales.
- Facilita la identificación de brechas entre países o regiones.

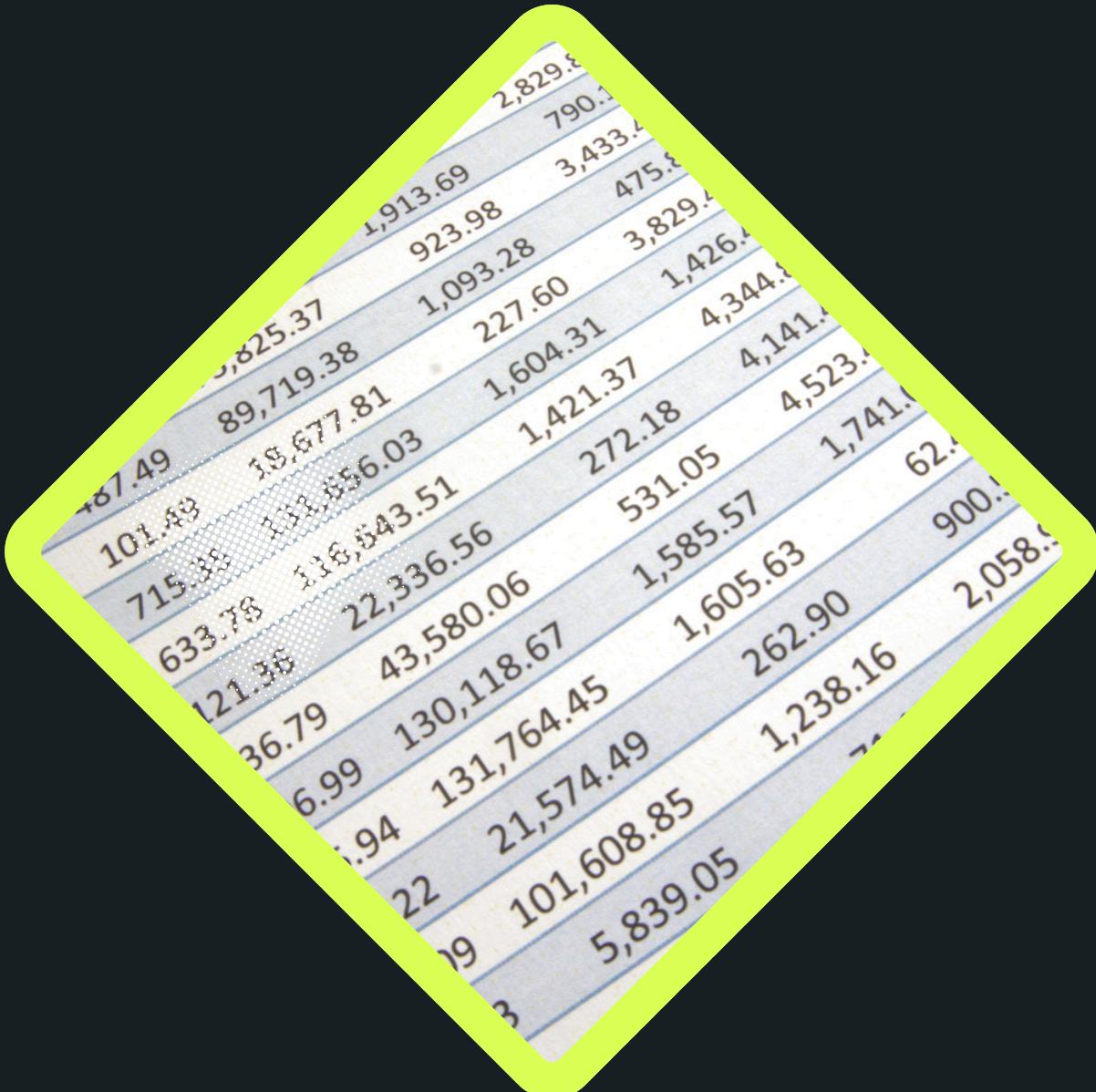


DATASET



Fuente:

- Kaggle: [Global Development Indicators \(2000-2020\)](https://www.kaggle.com/datasets/michaelmatta0/global-development-indicators-2000-2020cators_(2000-2020))
- Datos recopilados de organismos como el Banco Mundial, ONU, FMI, etc.



	Year	Indicator	Value
1	2000	GDP per capita (PPP)	1,913.69
2	2001	GDP per capita (PPP)	2,829.81
3	2002	GDP per capita (PPP)	790.51
4	2003	GDP per capita (PPP)	3,433.11
5	2004	GDP per capita (PPP)	475.81
6	2005	GDP per capita (PPP)	3,829.41
7	2006	GDP per capita (PPP)	1,426.41
8	2007	GDP per capita (PPP)	4,344.11
9	2008	GDP per capita (PPP)	4,141.11
10	2009	GDP per capita (PPP)	4,523.11
11	2010	GDP per capita (PPP)	1,741.11
12	2011	GDP per capita (PPP)	62.91
13	2012	GDP per capita (PPP)	900.11
14	2013	GDP per capita (PPP)	2,058.11
15	2014	GDP per capita (PPP)	1,605.63
16	2015	GDP per capita (PPP)	262.90
17	2016	GDP per capita (PPP)	1,238.16
18	2017	GDP per capita (PPP)	5,839.05
19	2018	GDP per capita (PPP)	101,608.85
20	2019	GDP per capita (PPP)	131,764.45
21	2020	GDP per capita (PPP)	21,574.49
22	2021	GDP per capita (PPP)	6.99
23	2022	GDP per capita (PPP)	121.36
24	2023	GDP per capita (PPP)	36.79
25	2024	GDP per capita (PPP)	715.35
26	2025	GDP per capita (PPP)	633.78
27	2026	GDP per capita (PPP)	118,556.03
28	2027	GDP per capita (PPP)	19,677.81
29	2028	GDP per capita (PPP)	89,719.38
30	2029	GDP per capita (PPP)	5,825.37
31	2030	GDP per capita (PPP)	101.83
32	2031	GDP per capita (PPP)	1,67.49

Contenido del dataset:

- Más de 200 países.
- Variables sociales, económicas, tecnológicas, ambientales, entre otras.
- Periodo cubierto: Años 2000 a 2020.

📅 Periodo de tiempo:

- Desde 2000 hasta 2020 (21 años de datos históricos por país)





🔑 VARIABLES CLAVE SELECCIONADAS:

- `internet_usage_pct` – Porcentaje de uso de internet
- Proporción de la población con acceso y uso activo de internet.
- `life_expectancy` – Esperanza de vida
- Número promedio de años que se espera que viva una persona.
- `health_development_ratio` – Índice de desarrollo en salud
- Indicador que refleja el acceso y calidad del sistema de salud en un país.
- `education_health_ratio` – Relación entre educación y salud
- Relación que evalúa el equilibrio entre los niveles de educación y salud.
- `co2_emissions_per_capita_tons` – Emisiones de CO₂ per cápita (toneladas)
- Cantidad de dióxido de carbono emitido por persona en un país.
- `school_enrollment_secondary` – Matrícula en educación secundaria
- Porcentaje de jóvenes inscritos en instituciones de nivel secundario.
- `digital_readiness_score` – Índice de preparación digital
- Mide qué tan preparado está un país para aprovechar las tecnologías digitales.
- `econ_opportunity_index` – Índice de oportunidades económicas
- Evalúa el acceso a oportunidades económicas y laborales.
- `income_group` – Grupo de ingreso
- Clasificación de los países según su nivel de ingreso (bajo, medio, alto).

Año	Código País	Nombre País	Población	PIB (USD)	PIB per Cápita	Inflación (%)	Ratio Educación/
2000	AFE	Africa Eastern an	398,113,044	283000000000	713.25	8.6	NaN
2001	AFE	Africa Eastern an	408,522,129	259000000000	633.61	5.84	NaN
2002	AFE	Africa Eastern an	419,223,717	265000000000	631.87	8.76	NaN
2003	AFE	Africa Eastern an	430,246,635	353000000000	819.74	7.45	NaN
2004	AFE	Africa Eastern an	441,630,149	439000000000	993.76	5.02	NaN

PREPROCESAMIENTO DE DATOS

Objetivo:

Preparar el dataset para el modelo de Machine Learning asegurando que no haya valores nulos y que las variables categóricas sean comprensibles por los algoritmos.

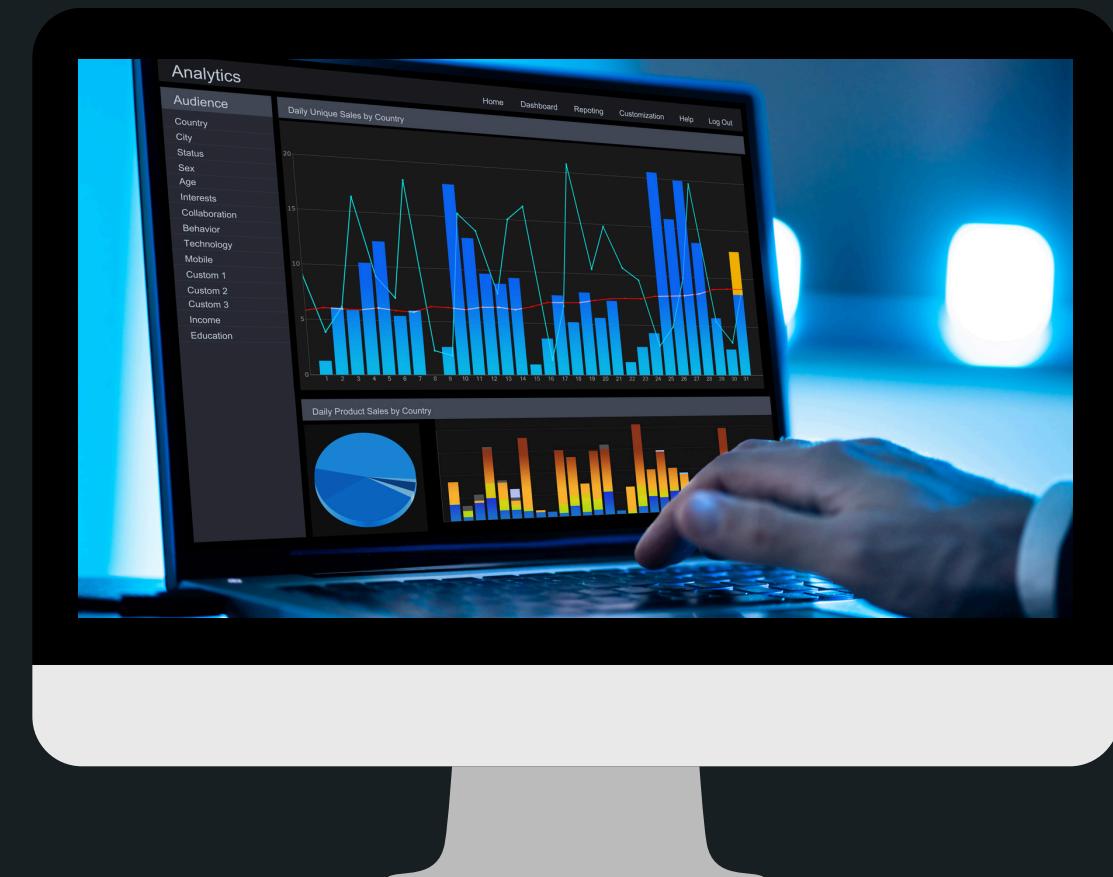
Imputación y Codificación Categórica:

- Se imputaron los valores faltantes en las variables region e income_group con la moda (valor más frecuente).
- Luego, se convirtieron estas variables a tipo categórico.
- Se aplicó One-Hot Encoding para transformarlas en variables numéricas binarias (0/1).



12 34 Imputación Numérica:

- Se identificaron las variables numéricas con valores faltantes.
- Los valores nulos se imputaron utilizando la media de cada columna





MODELOS UTILIZADOS

Regresión Lineal

¿Qué hace?

Encuentra una línea recta que mejor se ajusta a los datos, asumiendo una relación lineal entre las variables independientes y el PIB per cápita.

Resultados:

- R² Train: 0.61
- R² Test: 0.61
- MAE: 7,112 USD
- RMSE: 13,283 USD

Conclusión:

Tuvo un rendimiento limitado debido a su simplicidad y la incapacidad de capturar relaciones no lineales.

¿Por qué no funcionó tan bien?

Porque el PIB per cápita depende de múltiples factores que no se relacionan de forma estrictamente lineal. El modelo fue demasiado simple para la complejidad del problema.



Random Forest

¿Qué hace?

Usa múltiples árboles de decisión que votan juntos para hacer predicciones. Captura relaciones no lineales y complejas.

Resultados:

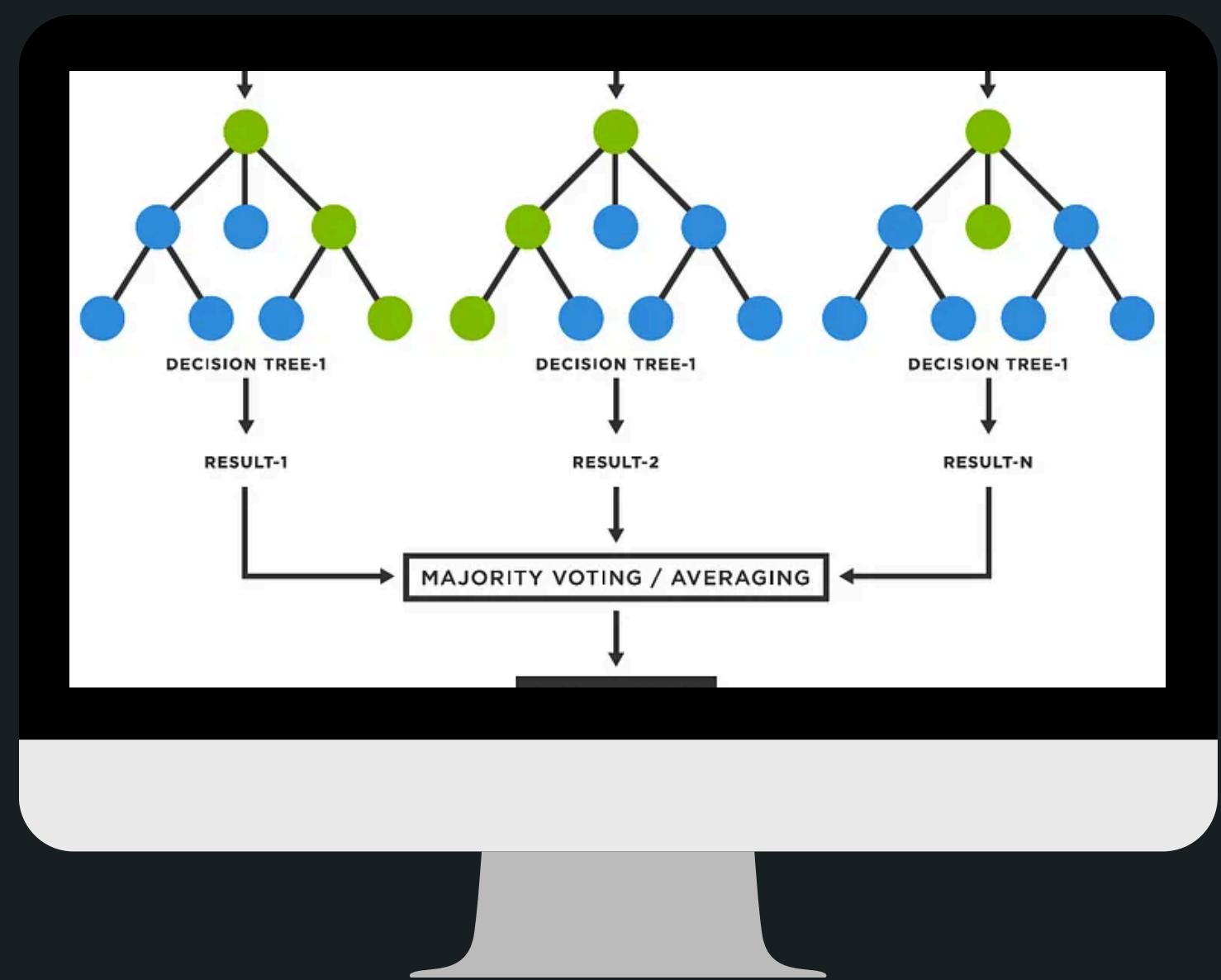
- R² Train: 0.99
- R² Test: 0.97
- MAE: ~1,400 USD
- RMSE: ~3,574 USD

Conclusión:

Mucho mejor que regresión lineal. Predicciones precisas con buen manejo de la complejidad del problema.

¿Por qué funcionó mejor?

Porque captura relaciones no lineales y complejas entre las variables sin necesidad de transformación previa.



Gradient Boosting

¿Qué hace?

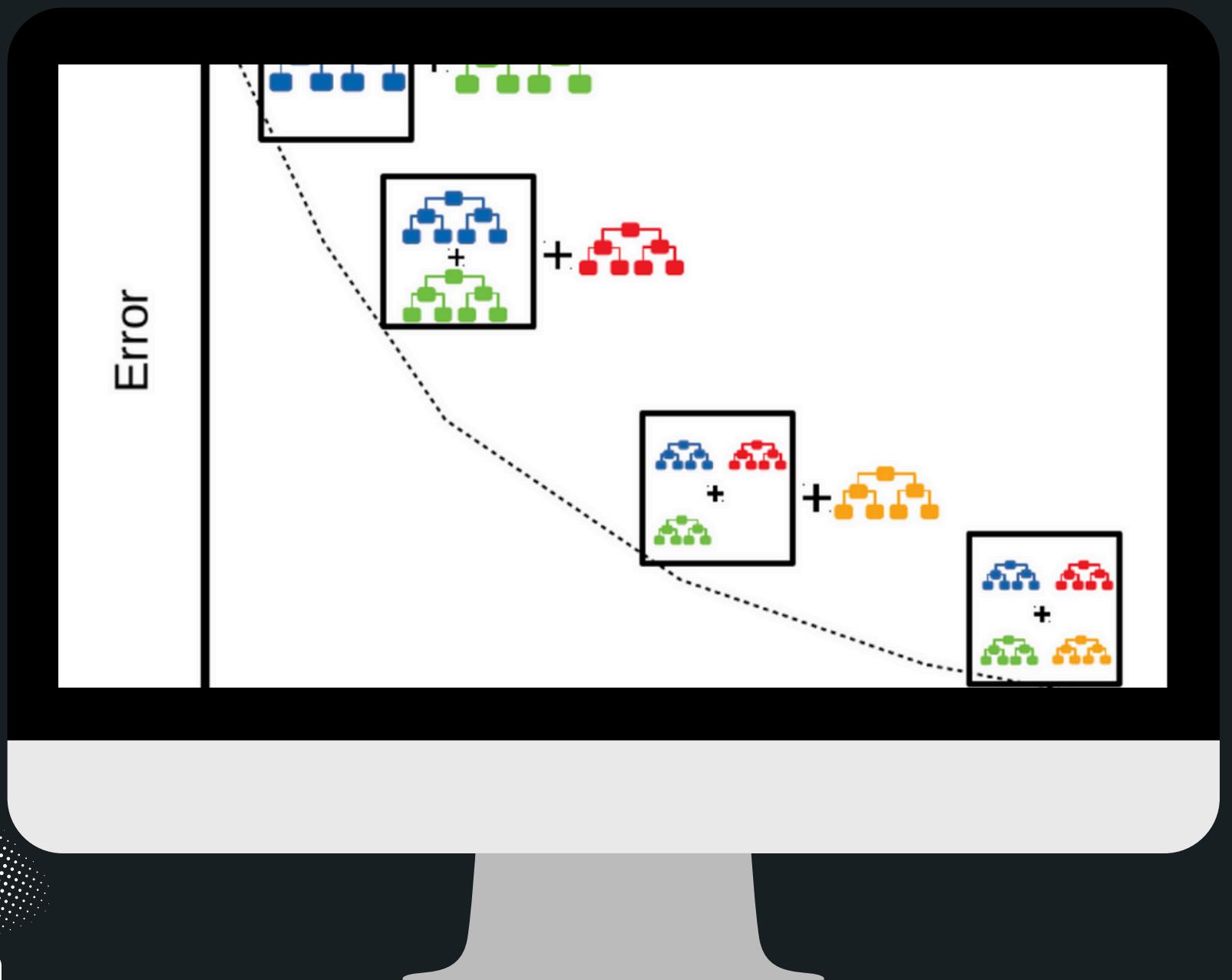
Construye árboles secuencialmente, donde cada nuevo árbol corrige los errores del anterior. Se enfoca más en los casos difíciles.

Resultados:

- R² Train: 0.9801
- R² Test: 0.9404
- MAE: 1,879.06 USD
- RMSE: 5,312.08 USD

Conclusión:

Muy buen rendimiento, aunque un poco menos preciso que XGBoost. Ideal para problemas donde los errores pequeños son importantes.



XGBoost

¿Qué hace?

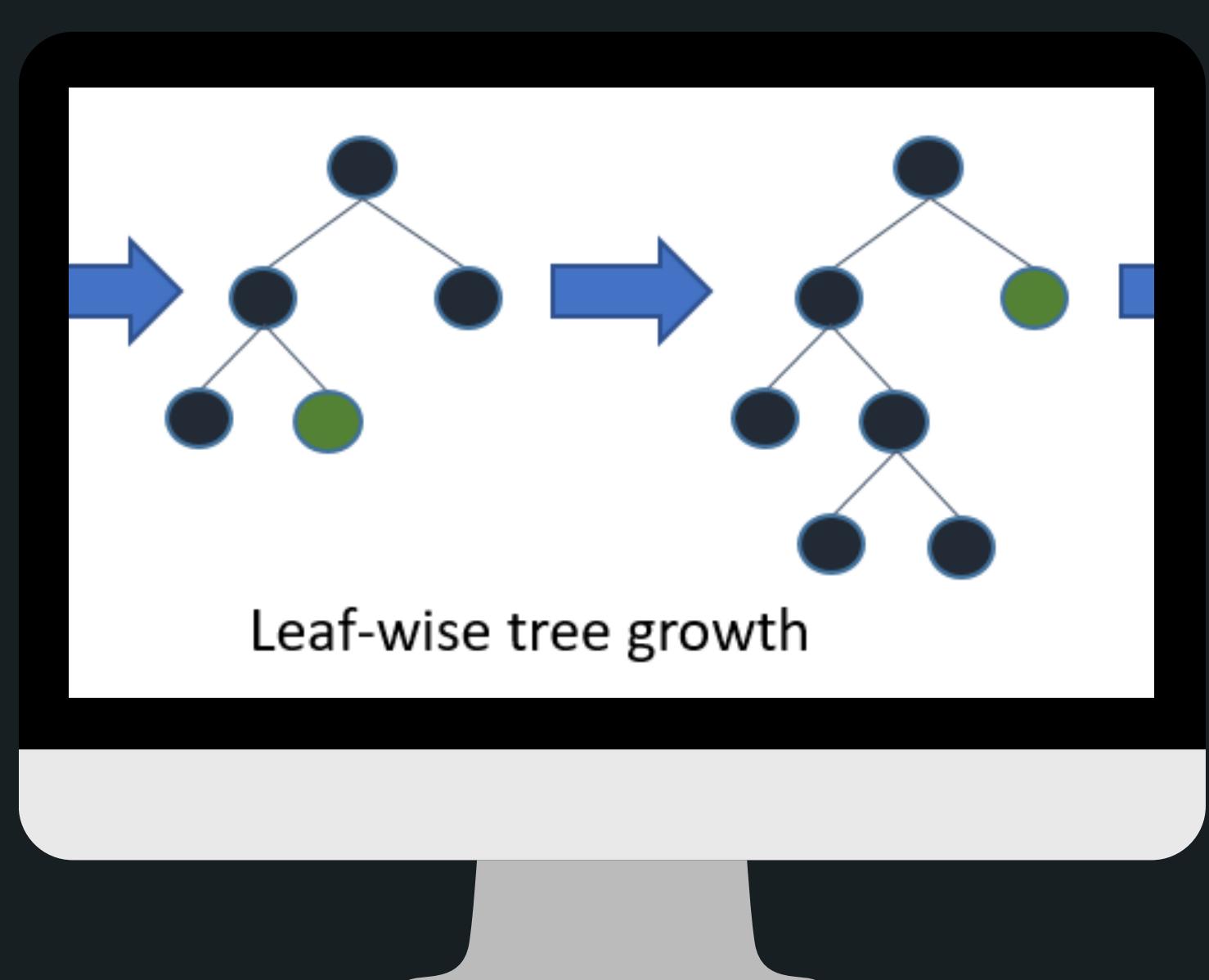
Es una versión mejorada de Gradient Boosting, más rápida y con regularización integrada, lo que reduce el sobreajuste.

Resultados:

- R² Train: 0.9940
- R² Test: 0.9687
- MAE: 1,341.73 USD
- RMSE: 3,846.01 USD

Conclusión:

El mejor modelo hasta ahora. Precisión alta y errores bajos.
Excelente para relaciones complejas como las del PIB per cápita





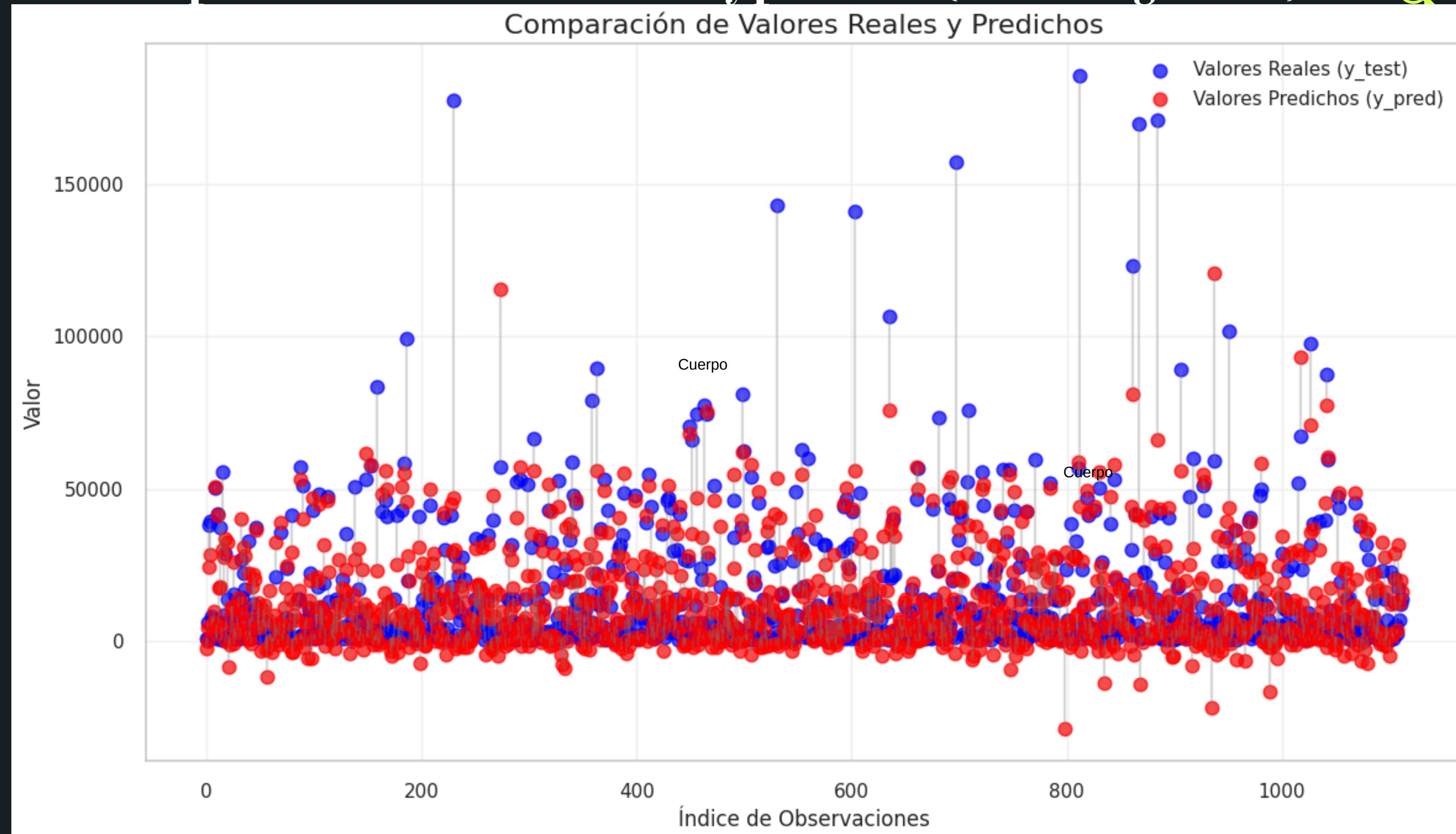
Resumen

Modelo	R ² Test	MAE (USD)	RMSE (USD)
Regresión Lineal	0.61	7,112	13,283
Random Forest	0.97	~1,400	~3,574
Gradient Boosting	0.94	1,879	5,312
XGBoost	0.97	1,341	3,846





Comparación entre valores reales y predichos (Linear Regression)



CONCLUSIONES



Conclusiones:

Alta dispersión en valores altos del PIB per cápita:

Se observa que, para los valores más altos (por encima de 50,000), el modelo no logra predecir con precisión, ya que los puntos rojos (valores predichos) están muy lejos de los puntos azules (valores reales). Esto indica que el modelo no captura bien los extremos del PIB per cápita.

Mejor rendimiento en rangos bajos y medios:

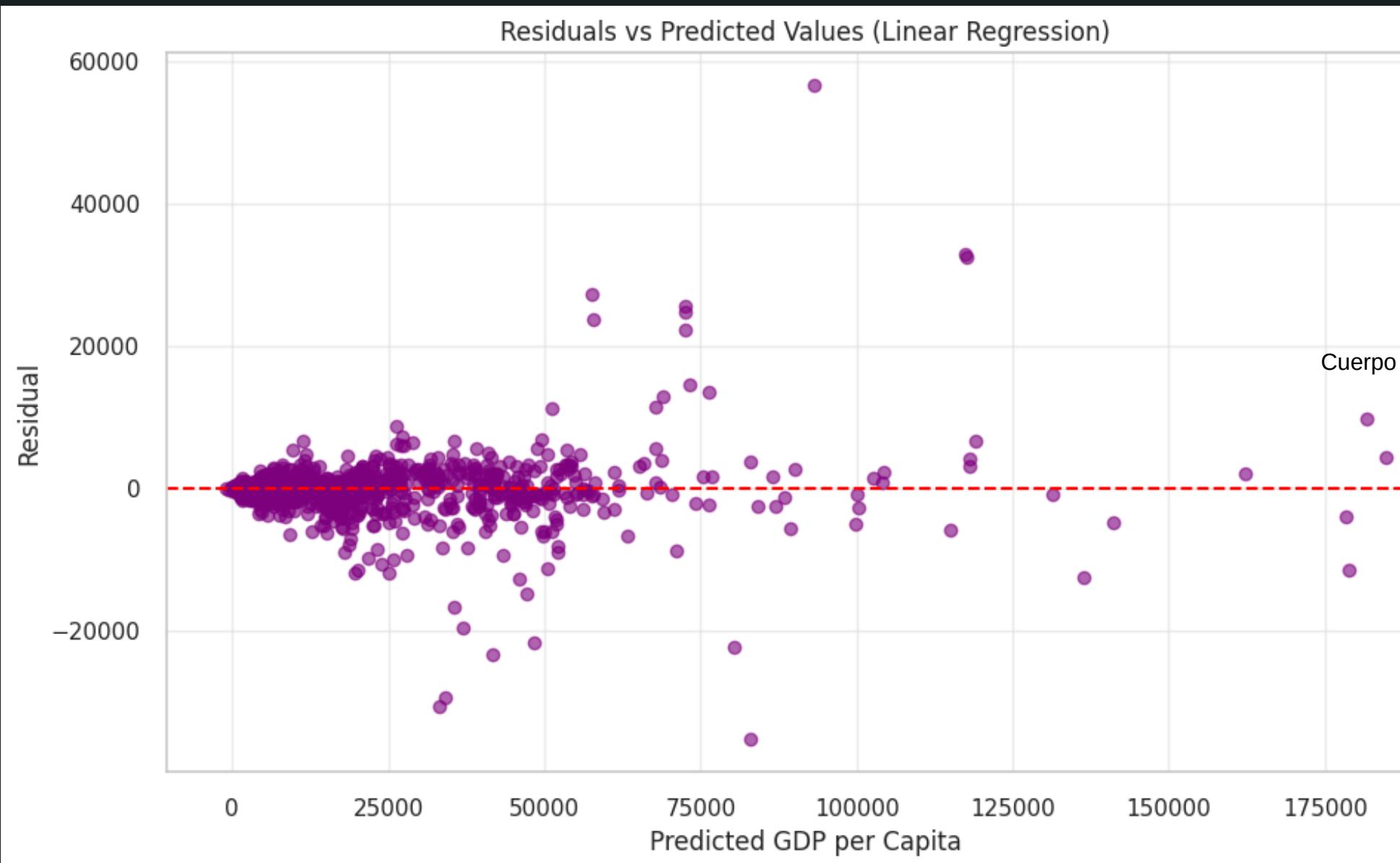
En valores más bajos o medios de PIB, hay mayor cercanía entre puntos reales y predichos, lo que sugiere que la regresión lineal tiene un mejor desempeño en ese rango.

Tendencia a subestimar el PIB en casos extremos:

Muchos valores reales altos (azules) tienen predicciones mucho menores (rojos), lo que implica una subestimación sistemática del modelo cuando el PIB es elevado.

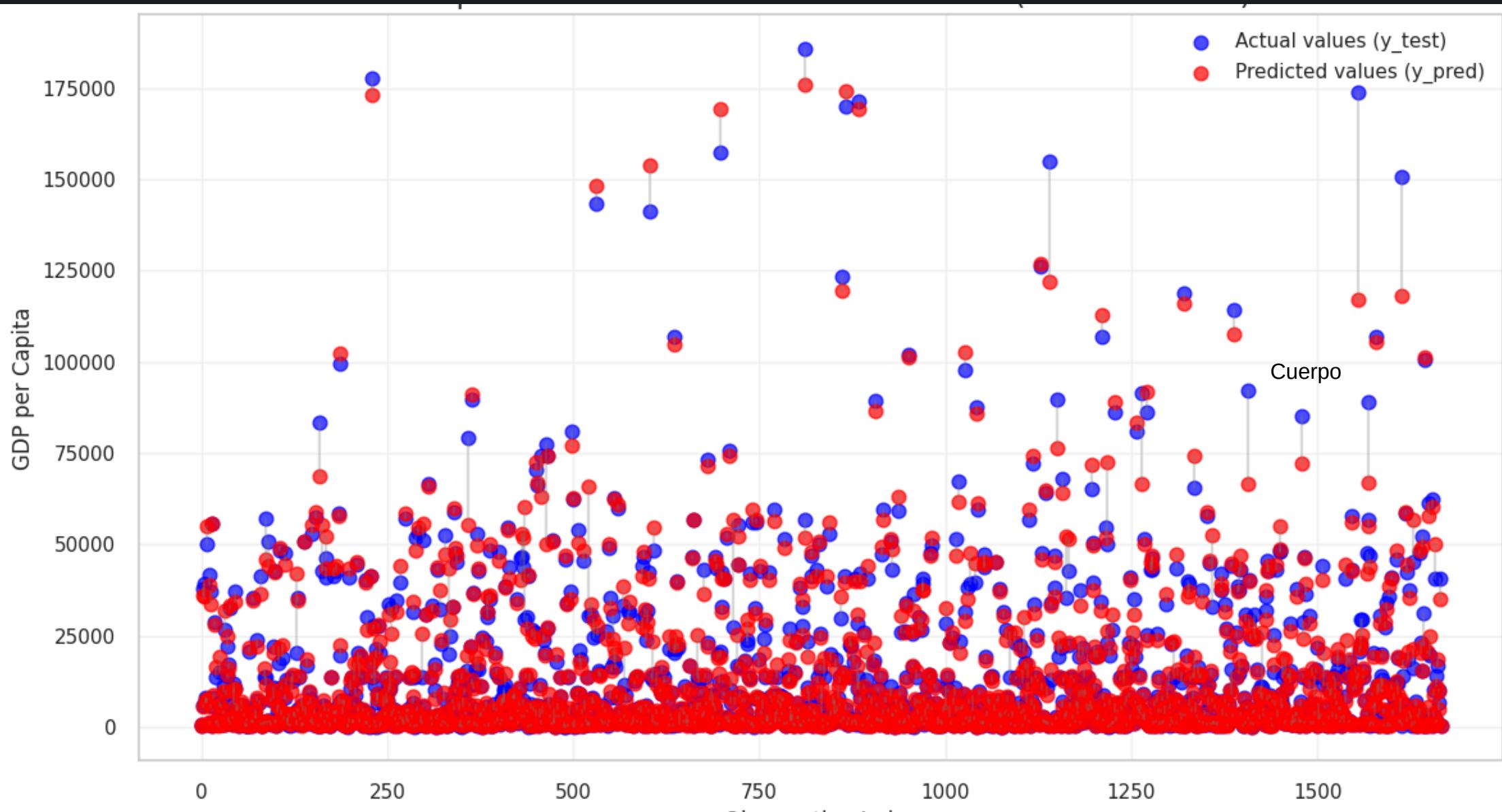
Cuerpo

GRAFICA RESIDUAL LINEAR REGRESION



- Los residuos se concentran alrededor de cero para predicciones bajas y medias, indicando buena precisión en esos rangos.
- A medida que aumentan los valores predichos, los residuos se dispersan más, mostrando que el modelo subestima o sobreestima en valores altos.
- Hay errores grandes en países con PIB per cápita muy alto, lo que confirma que la relación no es completamente lineal.
- En resumen, la regresión lineal funciona bien en rangos bajos/medios pero falla en extremos, por lo que modelos más complejos son necesarios.

Comparación entre valores reales y predichos (Random Forest)



Conclusiones:

Buena alineación general entre valores reales y predichos:

Los puntos rojos (predicciones) están bastante cercanos a los puntos azules (valores reales), lo cual indica que el modelo está haciendo predicciones bastante acertadas en la mayoría de los casos.

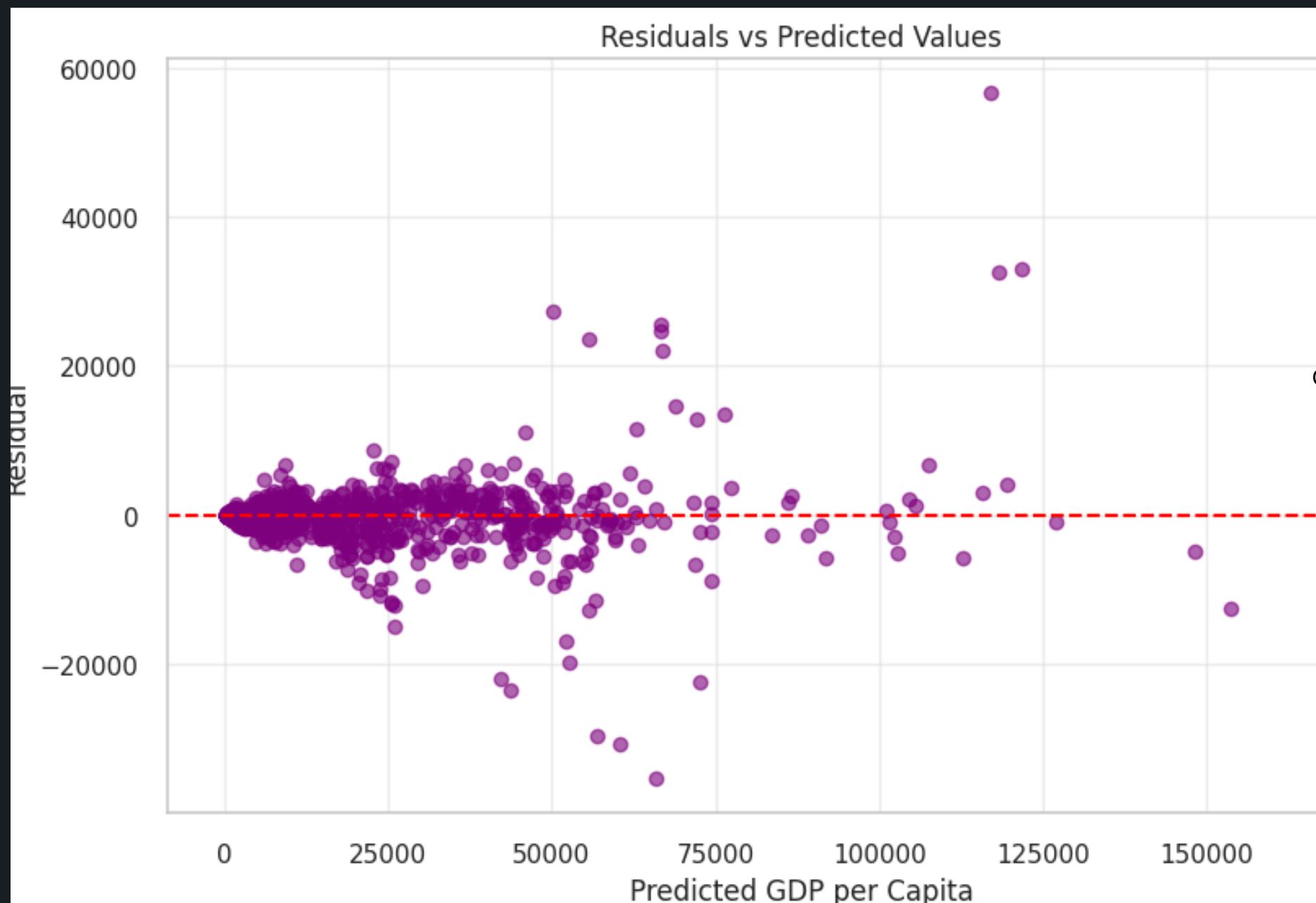
Errores más grandes en valores extremos:

Se observan diferencias más grandes entre predicho y real en los casos de PIB per cápita muy alto, lo cual sugiere que el modelo tiene más dificultad para predecir economías extremadamente desarrolladas o atípicas.

Consistencia en rangos medios y bajos:

En la gran mayoría de las observaciones (especialmente las más frecuentes y más densas, en la parte baja del gráfico), las predicciones son bastante cercanas a los valores reales.

GRAFICA RESIDUAL RANDOM FOREST REGRESOR



Conclusiones del gráfico de residuos (Random Forest)

Residuos centrados en cero:

La mayoría de los errores están cerca de la línea horizontal (0), lo que indica que el modelo no comete errores sistemáticos.

Mayor dispersión en valores altos de PIB per cápita:

A medida que el valor predicho aumenta, los errores también tienden a ser más grandes.

Esto sugiere que el modelo tiene más dificultad para predecir con precisión en países más desarrollados o con economías atípicas.

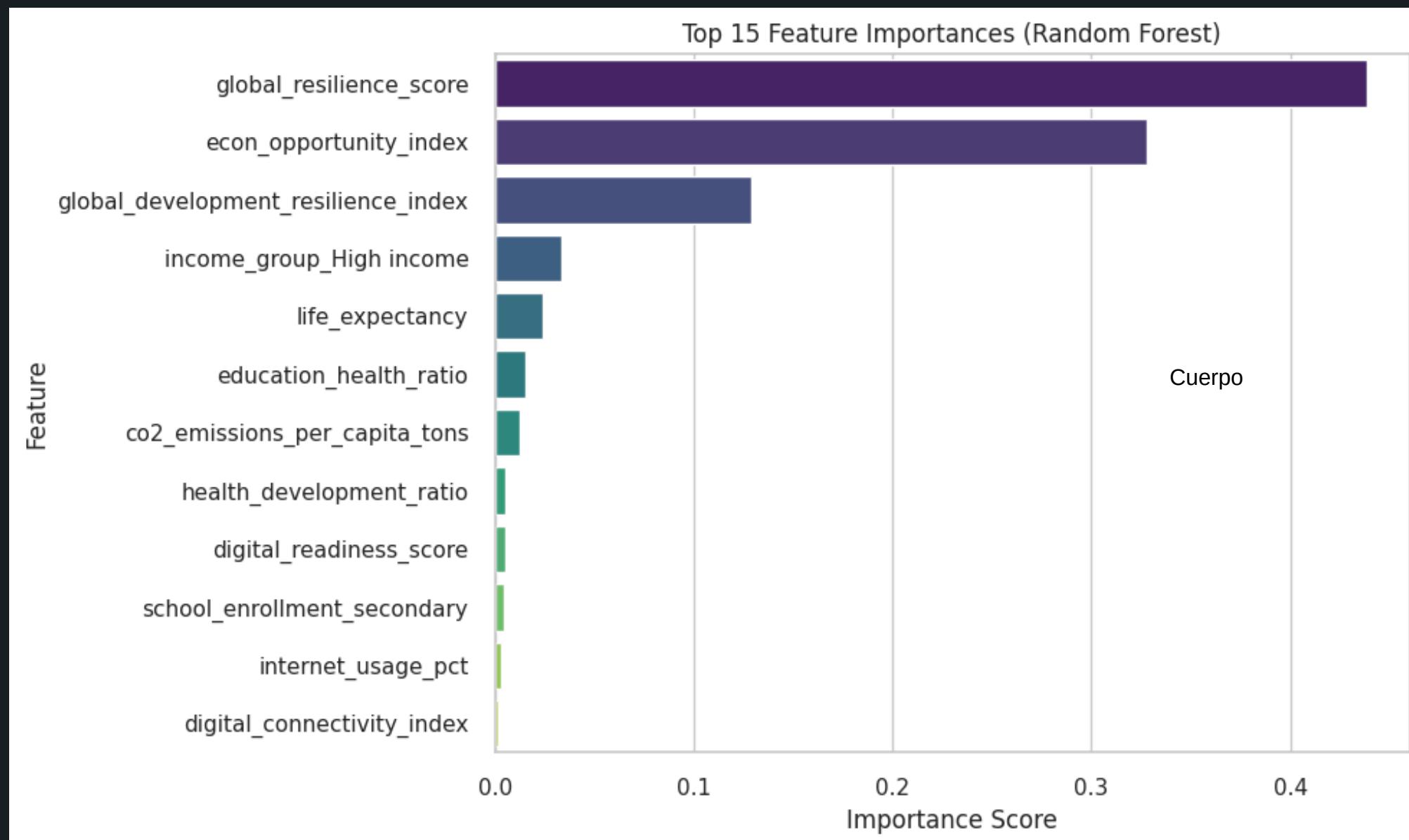
Buena precisión en rangos bajos y medios:

En países con PIB per cápita bajo o medio, el modelo hace buenas predicciones con errores pequeños.

✓ Conclusión general

El modelo Random Forest funciona bien en general, pero su precisión disminuye para países con un PIB per cápita muy alto. Aun así, es una opción sólida para predicción en contextos promedio.

Feature Importance Plot



Muestra qué variables fueron más importantes para predecir utilizando el algoritmo Random Forest Regressor

CONCLUSIONES

Objetivo cumplido

Se desarrolló un modelo de machine learning capaz de predecir el PIB per cápita a partir de indicadores sociales, económicos, ambientales y tecnológicos. Este modelo busca ser una herramienta útil para apoyar la toma de decisiones en políticas públicas.



Modelos aplicados y resultados

Se evaluaron diferentes modelos de regresión:

- Regresión lineal: mostró bajo desempeño con un R^2 test de 0.71, debido a su incapacidad para capturar relaciones no lineales entre las variables.
- Random Forest: logró un R^2 test de 0.94, mejorando significativamente la precisión y reduciendo los errores.
- Gradient Boosting: obtuvo un R^2 test de 0.94, con buenos resultados pero mayor error absoluto medio.
- XGBoost: fue el modelo más eficaz, con un R^2 test de 0.97, un MAE de 1,341.73 y un RMSE de 3,846.01, indicando gran precisión y generalización.



CONCLUSIONES



Variables más influyentes

Los factores que más influyeron en la predicción del PIB per cápita fueron:

- Índice de innovación
- Gasto en salud
- Esperanza de vida
- Nivel educativo
- Global resilience score
- Esto respalda la importancia de la inversión en capital humano, innovación y resiliencia para el crecimiento económico.

Aplicación práctica

El modelo puede ~~ayudar~~ a gobiernos y organismos internacionales a simular escenarios y prever el impacto de ciertas políticas públicas en el desarrollo económico, priorizando variables clave como salud, educación y  tecnología.



MUCHAS GRACIAS!



[Link repositorio github](#)

