

“Año de la recuperación y consolidación de la economía peruana”

UNIVERSIDAD PERUANA CAYETANO HEREDIA

FACULTAD DE CIENCIAS E INGENIERÍA



TÍTULO:

“Informe: Resultados de regresión en valores de CO y AQI en 2023 y 2024 en California”

TEMA:

“Machine Learning - Regresión”

AUTOR:

Victor Daniel Rivera Torres

CURSO:

Proyectos de Ingeniería 1

GRUPO:


AG2 - Subgrupo 3

2025 - I

1. Origen y trazabilidad de los datos

Fuente principal:

- EPA (US Environmental Protection Agency) AirData → Download Daily Data
 - Contaminante: CO
 - Años: 2023 y 2024
 - Área geográfica: California


United States
Environmental Protection
Agency

[Environmental Topics](#)
[Laws & Regulations](#)
[Report a Violation](#)
[About EPA](#)

[Home](#) / [Outdoor Air Quality Data](#)

Outdoor Air Quality Data

- [Frequent Questions about AirData](#)
- [Learn about Air Data](#)
- [Interactive Map](#)
- [Pre-generated Data Files](#)
- [Download Daily Data](#)**
- [Download Raw Data \(API\)](#)
- [Air Quality Index Report](#)
- [Air Quality Statistics Report](#)
- [Monitor Values Report](#)
- [Monitor Values Report - Hazardous Air Pollutants](#)
- [Air Quality Index Daily Values Report](#)
- [Daily Air Quality Tracker](#)
- [Tile Plot - Multiyear](#)
- [Tile Plot - Single Year](#)
- [AQI Plot](#)
- [Concentration Plot](#)
- [Ozone Exceedances](#)
- [Concentration Map](#)

Download Daily Data

This tool queries daily air quality summary statistics for the criteria pollutants by monitor. You can get data for specific monitors or all monitors in a city, county, or state.

Pollutant

CO

Year

2023

Geographic Area

California

-- OR --

Select a City (defined as CBSA) ...

-- OR --

Select a County ...

Monitor Site

All Sites

060010009

060010011

060010012

Get Data

Archivos brutos :

- Dataset de 2023 y 2024
- Dataset de 2025 para validación

Date	Source	Site ID	POC	Daily Max S-Units	Daily AQI Val	Local Site Na	Daily Obs Co	Percent Com	AQS Param	AQS Param
01/01/2023	AQS	60010009	1	0.6 ppm	7	Oakland	18	75	42101	Carbon monoxide
01/02/2023	AQS	60010009	1	0.5 ppm	6	Oakland	24	100	42101	Carbon monoxide
01/03/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide
01/04/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide
01/05/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide
01/06/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide
01/07/2023	AQS	60010009	1	0.3 ppm	3	Oakland	24	100	42101	Carbon monoxide
01/08/2023	AQS	60010009	1	0.3 ppm	3	Oakland	24	100	42101	Carbon monoxide
01/09/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide
01/10/2023	AQS	60010009	1	0.4 ppm	5	Oakland	24	100	42101	Carbon monoxide

Campos clave:

- Date (o Fecha), Daily AQI Value, Daily Max 8-hour CO Concentration, Daily Obs Count, identificadores de sitio/ciudad.

Licencia/nota EPA: datos preliminares de AirNow pueden diferir de AQS; validados posteriormente. No se usan para normativas.

2. Preparación y limpieza

2.1 Normalización

- Se estandarizan nombres, se parsea Date o se crea desde Year/Month/Day.
- Se convierten en numéricos: Daily AQI Value y Daily Max 8-hour CO Concentration (tolerando NaN).

2.2 Agregación diaria entre estaciones

Para cada fecha:

- AQI_mean = media de Daily AQI Value
- CO_mean = media de Daily Max 8-hour CO Concentration
- ObsCount_total (si existe) = suma de Daily Obs Count

Se extrae Year/Month/Day y se conservan las columnas: Year, Month, Day, AQI_mean, CO_mean, (opcional) ObsCount_total

Nota: probamos promedio ponderado por ObsCount_total; la mejora fue marginal, por eso mantuvimos la media simple + ObsCount_total como señal adicional.

3. Unificación de años y control de tipos

- promedio_diario_2023.csv + promedio_diario_2024.csv → promedio_diario_2023_2024.csv
- Orden cronológico, eliminación de duplicados por (Year,Month,Day).
- Normalización del nombre ObsCount_total (a veces venía como ObsCount).

Entregables de esta etapa (DataFrames obtenidos):

- promedio_diario_2023.csv
- promedio_diario_2024.csv
- promedio_diario_2023_2024.csv
- promedio_diario_2025.csv

4. Análisis exploratorio (EDA) resumido

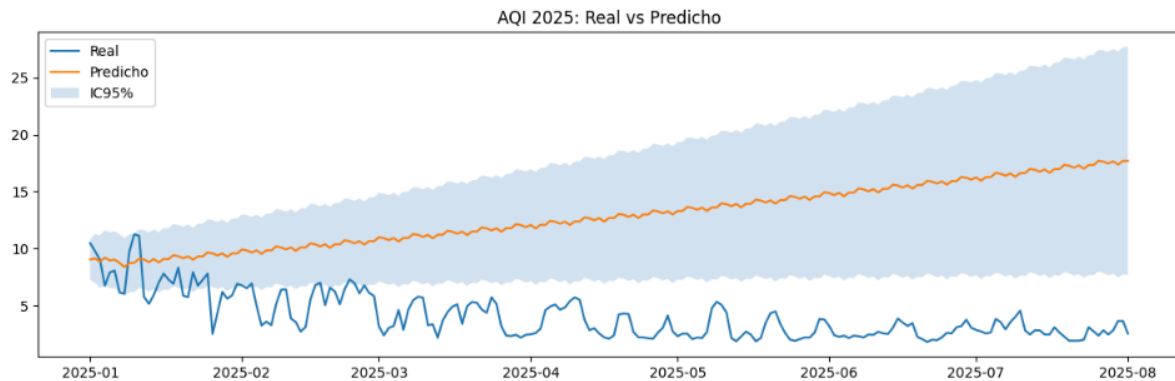
4.1 Señal y estacionalidad

- Comportamiento semanal muy marcado; presencia de picos (spikes).
- Tendencia de 2023 a 2024 relativamente estable con oscilaciones.

5. Baseline ingenuo

Naive t-7 (día de la semana anterior):

- RMSE ≈ 1.783 (valor guía).



6. Modelos y experimentos

6.1 SARIMAX (intento)

- Estacionalidad semanal ($s=7$); órdenes p,d,q,P,D,Q probados en rejillas pequeñas.
- Resultados insuficientes: pronósticos sesgados y con poca amplitud; residuales con patrón; sensibilidad a la especificación.

¿Por qué no funcionó bien?

- Fuerte estacionalidad con cambios de amplitud y picos irregulares.
- Órdenes subóptimos; tuning profundo costoso; sin exógenas potentes pierde información.

6.2 ML con features (HistGradientBoostingRegressor)

6.2.1 Feature engineering

- Lags de AQI_mean: 1, 2, 3, 7, 14, 28
- Medias móviles: MA7, MA14 de AQI_mean
- Exógenas (CO_mean, ObsCount_total): lags 1 y 7 + MA7
- Calendario: dummies de día de semana y mes
- Fourier semanal: sin/cos (periodo 7)
- Tendencia: índice temporal t

6.2.2 Validación

- TimeSeriesSplit(n_splits=5) (sin *shuffle*).
- Evitar leakage: todos los lags/MA se calculan únicamente con historia.

6.2.3 Entrenamiento

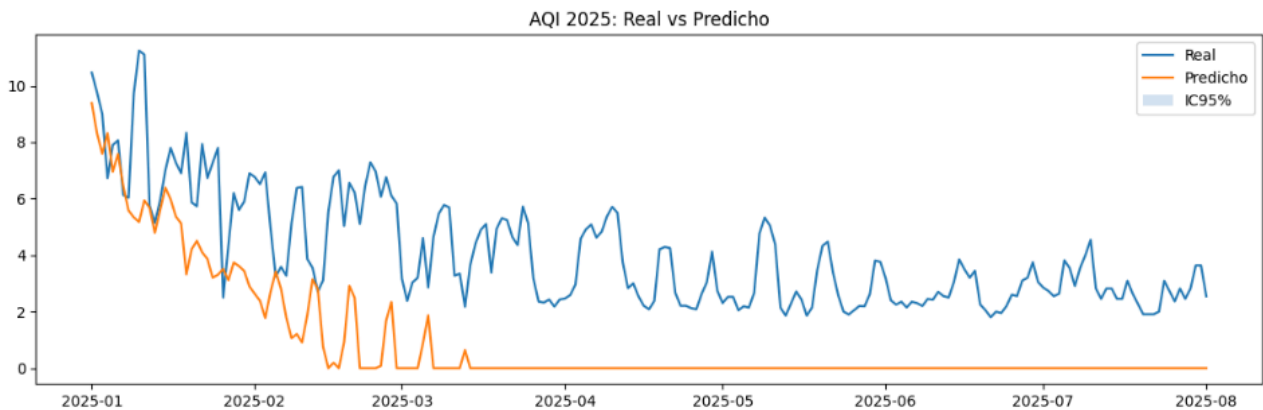
- HistGradientBoostingRegressor: learning_rate≈0.06, max_iter≈800, l2≈1.0.
- Métricas CV (típicas): RMSE ≈ 0.38 ± 0.38, MAE ≈ 0.26.

6.2.4 Pronóstico 2025 (recursivo)

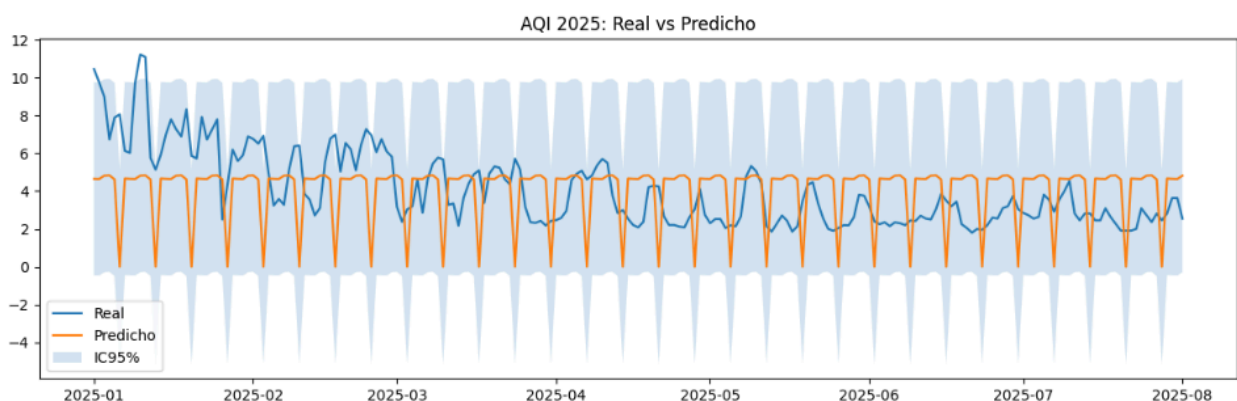
- Se agrega día a día a la historia; las predicciones de AQI se retroalimentan para construir lags futuros.

6.2.5 Problema clave: exógenas 2025 faltantes

- Versión 1 (media 7 días) → aplanó la predicción (casi constante).

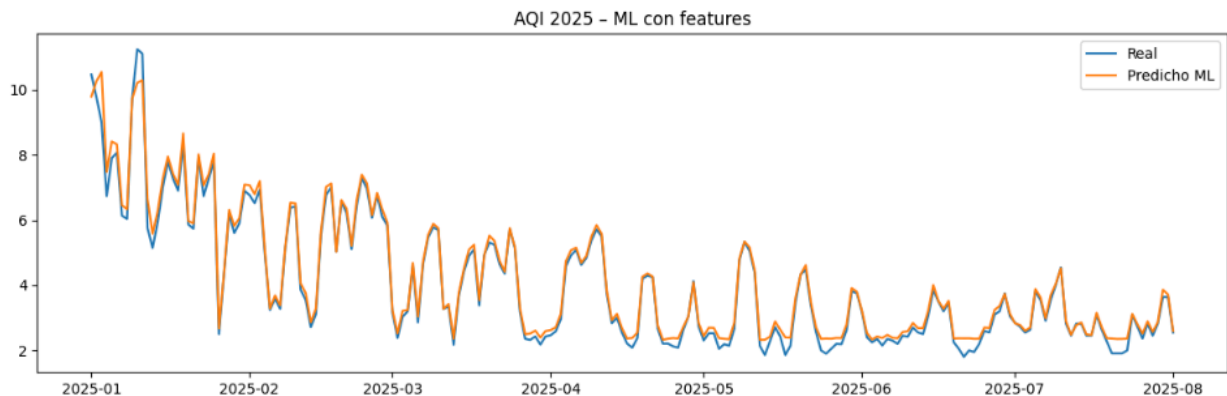


- Versión 2 (repetir patrón semanal fijo) → serrucho perfecto (amplitud idéntica).



- Versión final (ganadora): Backfill tendencia + estacionalidad centrada + clipping
 - Tendencia: ajuste lineal en los últimos ~60 días de 2024.

- Estacionalidad: media por día de semana en las últimas 8 semanas, centrada (resto la media para no duplicar nivel) y con amplitud controlable.
- Clipping: p1–p99 históricos de la propia exógena para evitar outliers inventados.
- (Opcional) suavizado leve: 85% yhat + 15% media de las últimas 6 predicciones.



6.2.6 Resultados típicos

- CV (5 folds): $RMSE \approx 0.383 \pm 0.385$; $MAE \approx 0.262$.
- Test 2025: $RMSE \approx 0.248$; $MAE \approx 0.172$.
- vs Baseline t-7: mejora clara ($1.783 \rightarrow \sim 0.25-0.30$).

6.3 XGBoost + backtesting expanding-window + intervalos

6.3.1 ¿Por qué XGBoost?

- Otro *learner* de árboles con gran capacidad; expresivo con features tabulares.

6.3.2 Backtesting (expanding window)

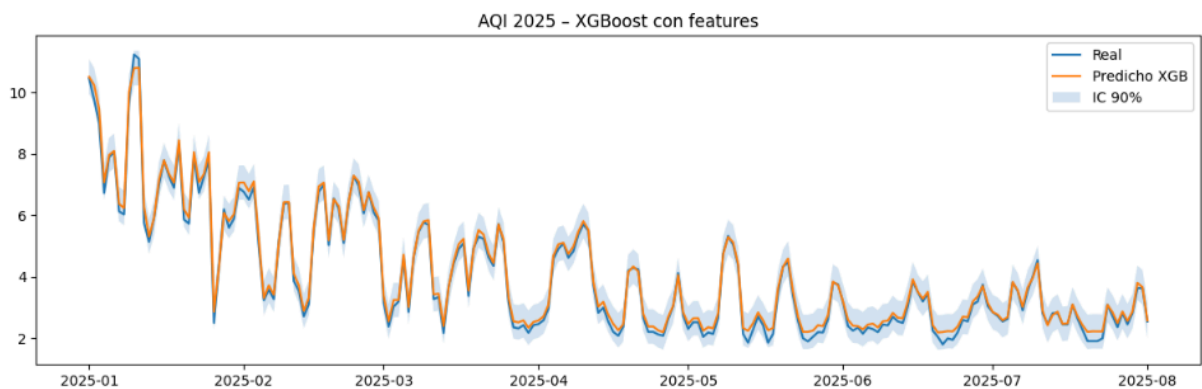
- Particiones crecientes: entreno hasta $t_1 \rightarrow$ valido en $(t_1, t_2]$, etc.
- Evita optimismo; simula despliegue real.

6.3.3 Hiperparámetros (grid corto)

- $n_estimators \in \{500, 800, 1200\}$
- $learning_rate \in \{0.04, 0.06, 0.08\}$
- $max_depth \in \{4, 5, 6\}$
- $subsample \in \{0.8, 1.0\}$
- $colsample_bytree \in \{0.8, 1.0\}$
- $reg_lambda \in \{0.5, 1.0, 1.5\}$

6.3.4 Intervalos conformales (IC 90%)

- Calibración con residuales walk-forward en fines de 2024; $q = 0.90$ sobre $|\text{error}|$.
- Pronóstico 2025: $IC = \hat{y} \pm q$.



6.3.5 Comentario

- Métricas similares a HistGB; la validación con expanding-window e IC elevan la calidad del informe.

7. Diario de errores y soluciones

Error (pantalla)	Por qué ocurrió	Qué hicimos
TypeError: got an unexpected keyword argument 'squared'	Diferencias de versión/import al calcular RMSE	Pasamos a $rmse = \sqrt{mse}$ y control de imports (sklearn.metrics)
KeyError: 'AQI_mean' o 'AQI_pred_A'	Mezcla de CSV con columnas distintas	Estandarizamos nombres y reindexamos columnas
Input contains NaN (sklearn)	NaN/inf en arrays al evaluar	Máscaras de validación + dropna + np.isfinite
Predicción plana 2025	Faltan exógenas 2025; relleno con media 7d	Backfill tendencia + estacionalidad
Predicción serrucho 2025	Repetimos patrón semanal fijo	Estacionalidad centrada + amplitud + clipping

Mismatch de features en forecast	feat_cols del train no coinciden con forecast	tmp.reindex(columns=feat_cols, fill_value=0) en inferencia
----------------------------------	---	--

Versiones de entorno:

- scikit-learn 1.6.1
- statsmodels 0.14.5
- Python 3.12 (Colab)

8. Resultados finales (para la memoria)

Modelo	Validación	RMSE	MAE
Baseline t-7	2025	1.783	—
SARIMAX	2025	0.226	0.192
HistGB (features)	CV 5 folds	0.383 ± 0.385	0.262
HistGB (features)	2025	0.248	0.172
XGBoost (features)	Backtesting	0.169	0.138

9. Conclusiones

- El ML con features (árboles + lags/MA/calendario/exógenas) superó al baseline t-7 y al SARIMAX básico.
- El manejo correcto de exógenas faltantes en 2025 fue crítico: el backfill tendencia + estacionalidad evitó tanto la planicie como el serrucho repetitivo.
- La validación temporal correcta (TimeSeriesSplit o expanding-window) es clave para evitar optimismo.
- Los intervalos conformales son útiles para comunicar incertidumbre operativa.