Máster en Inteligencia Artificial UMU

PRÁCTICA 2

Machine Learning

Ignacio Ruiz Chicano

Juan Jesús Torralba Mateos

Ana Gil Molina

ÍNDICE

ÍNDICE

Preprocesamiento

- **DATASET:** Aljorrahorarias2017_2022.csv
- Datos de las medidas horarias de una estación sensora situada en la Alojorra.
- Incluye 16 columnas:
  - Óxido Nítrico (NO)
  - Dióxido de Nitrógeno (NO2)
  - Disulfuro de Azufre (SO2)
  - Ozono (O3)
  - Óxidos de Nitrógeno (NOx)
  - PM10
  - Benceno (C6H6)
  - Tolueno (C7H8)
  - Xileno (XIL)
  - Temperatura media (TMP)
  - Humedad Relativa (HR)
  - Dirección del viento
  - Presión atmosférica (PRB)
  - Velocidad del Viento (W)
  - Radiación Solar (RS)

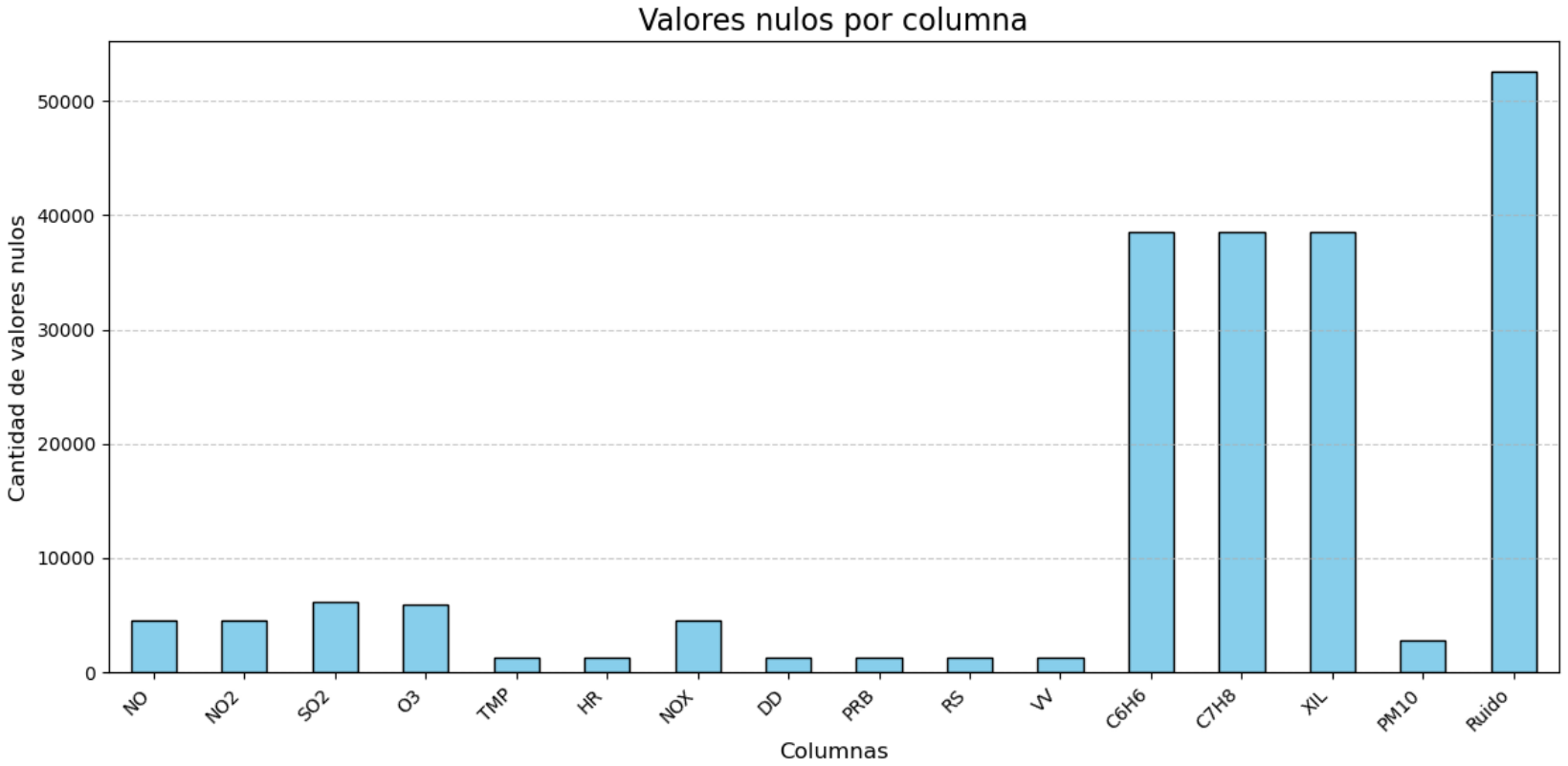
Calidad del aire

Datos meteorológicos

- **OBJETIVO:** Se pide la predicción a 7 días del Óxido Nítrico (NOX).

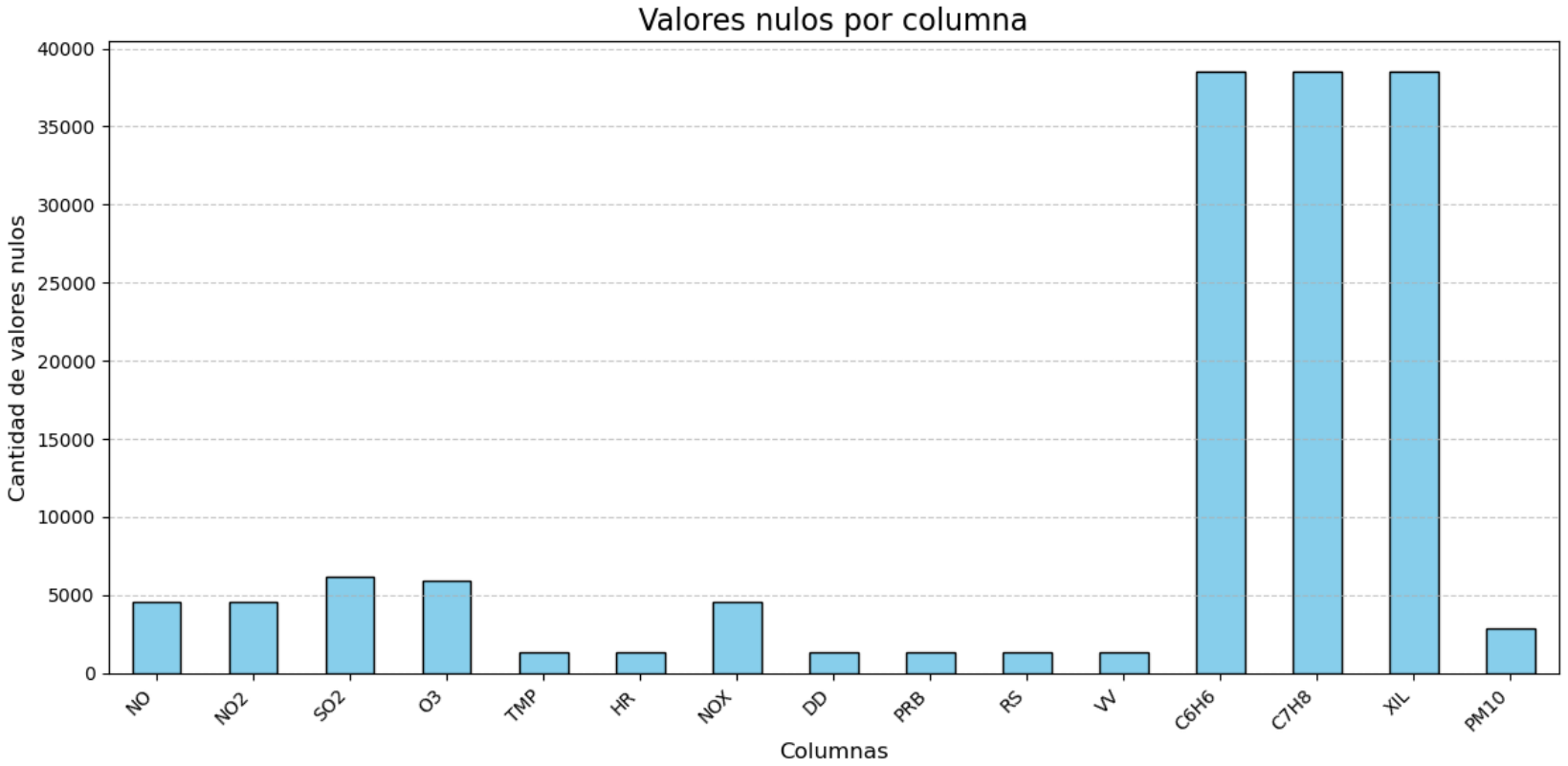
ÍNDICE

Preprocesamiento: Análisis de los valores nulos



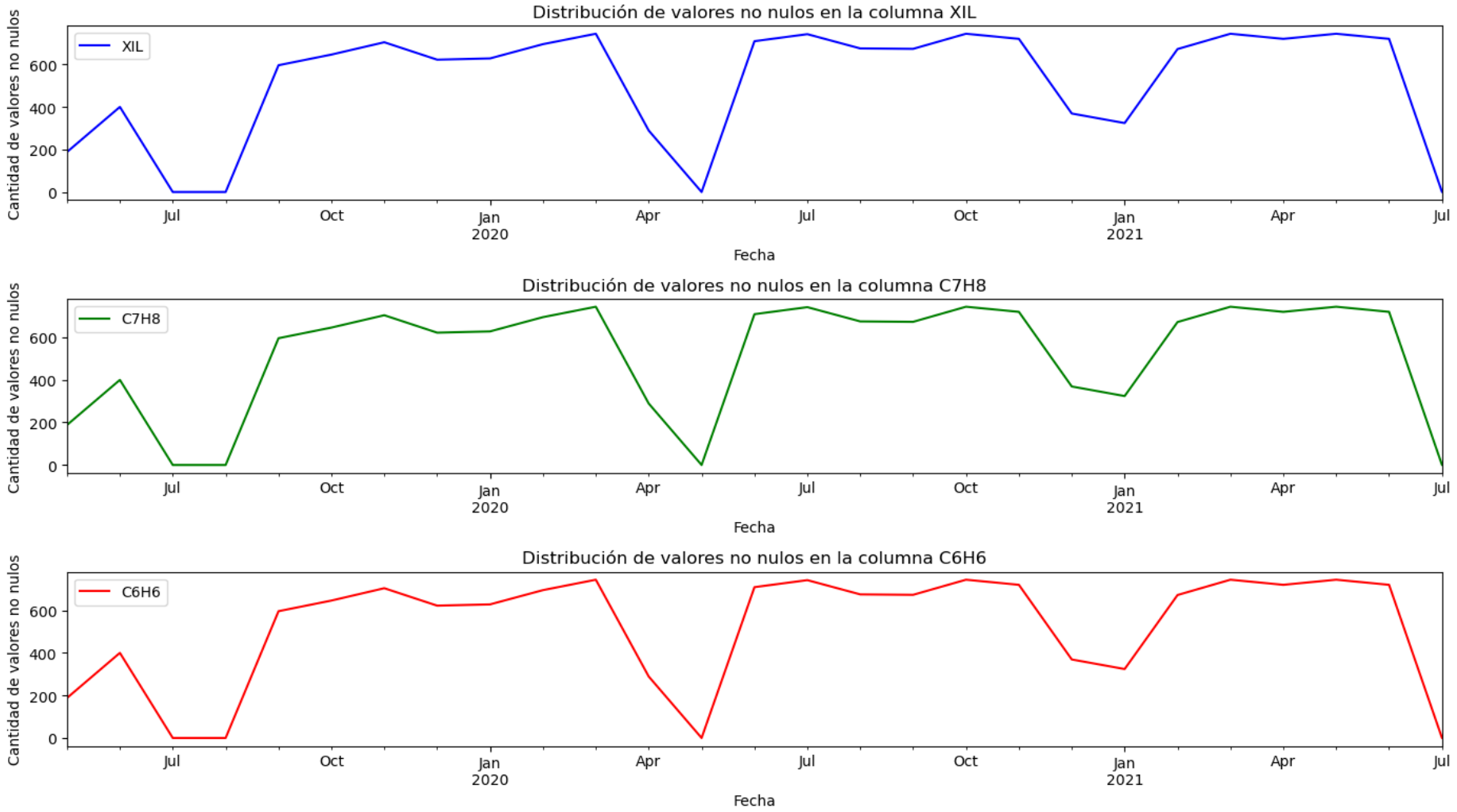
Columna Ruido → Posee 0 valores NO nulos, es decir, solo tiene NaN

Preprocesamiento: Análisis de los valores nulos

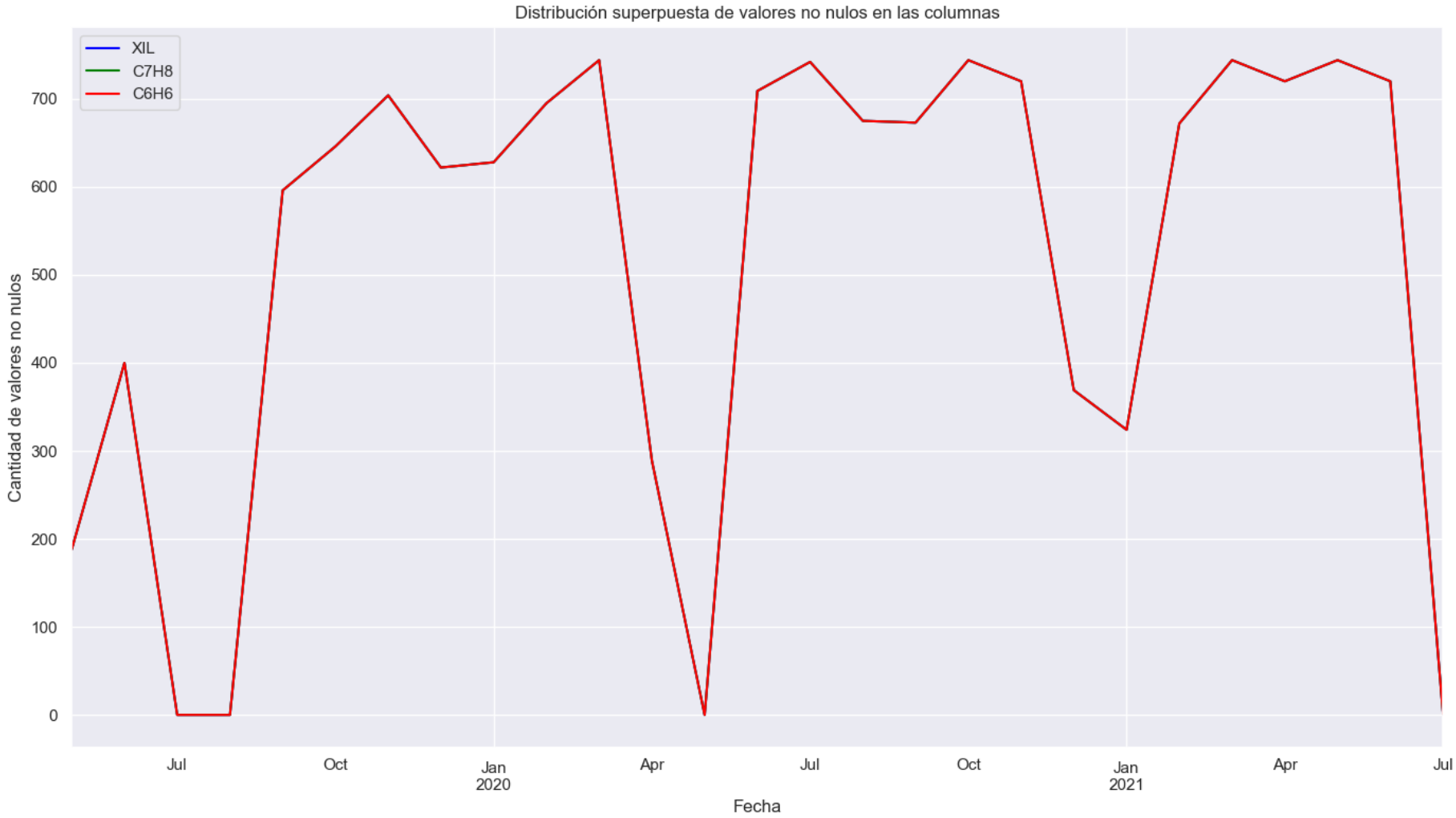


Columnas C6H6, C7H8, XIL → Tienen el mismo número de valores nulos

Preprocesamiento: Análisis de los valores nulos

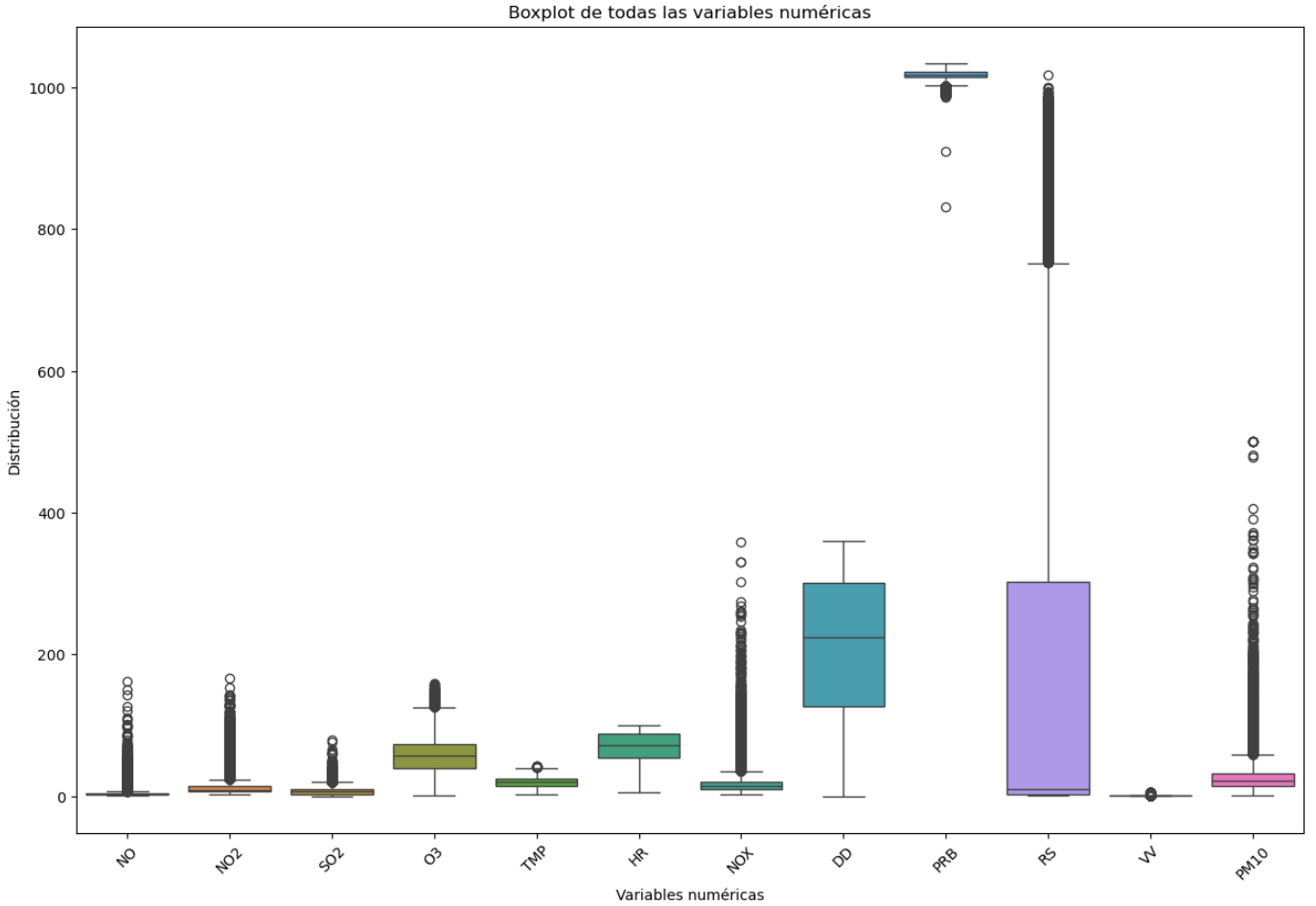


Preprocesamiento: Análisis de los valores nulos



ÍNDICE

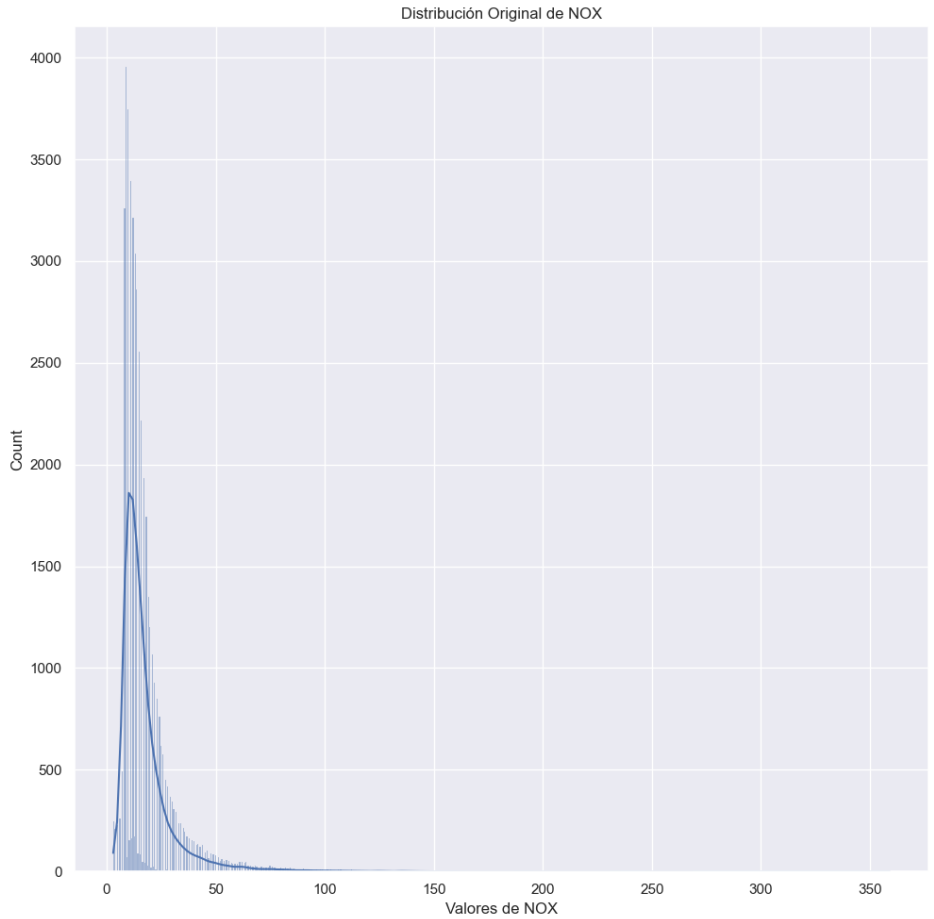
Preprocesamiento: Visualización

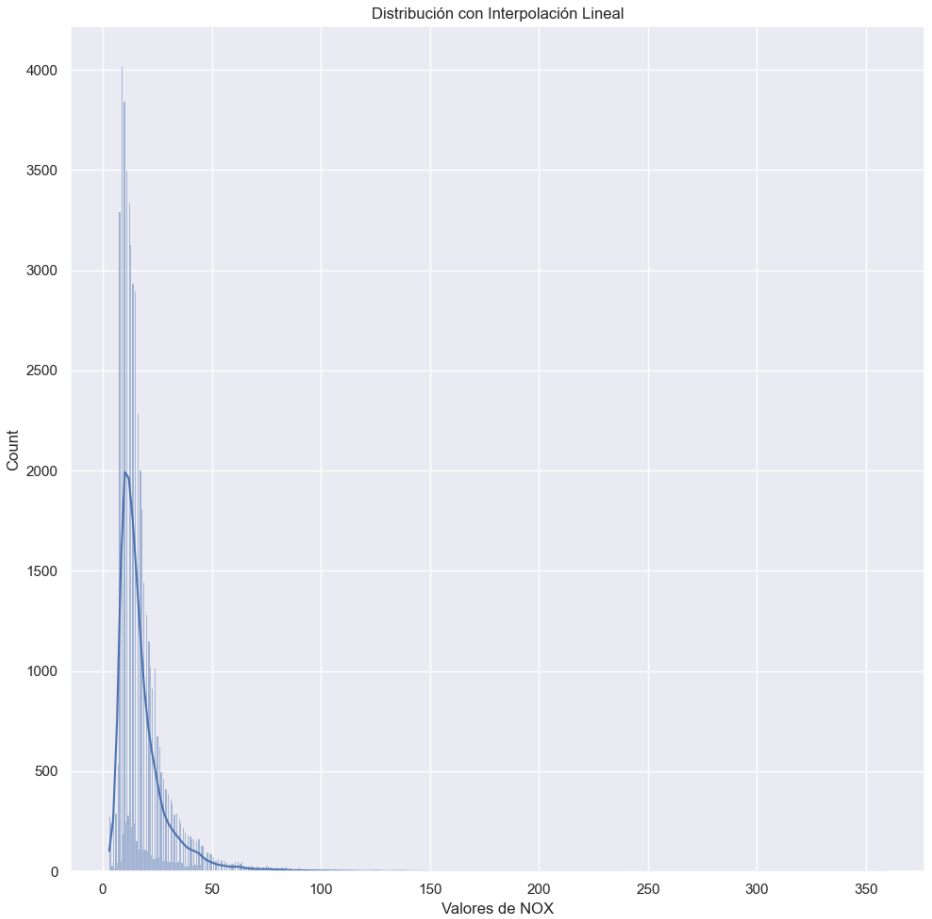


- Rangos de valores distintos
- Gran cantidad de outliers

ÍNDICE

Preprocesamiento: Interpolación lineal

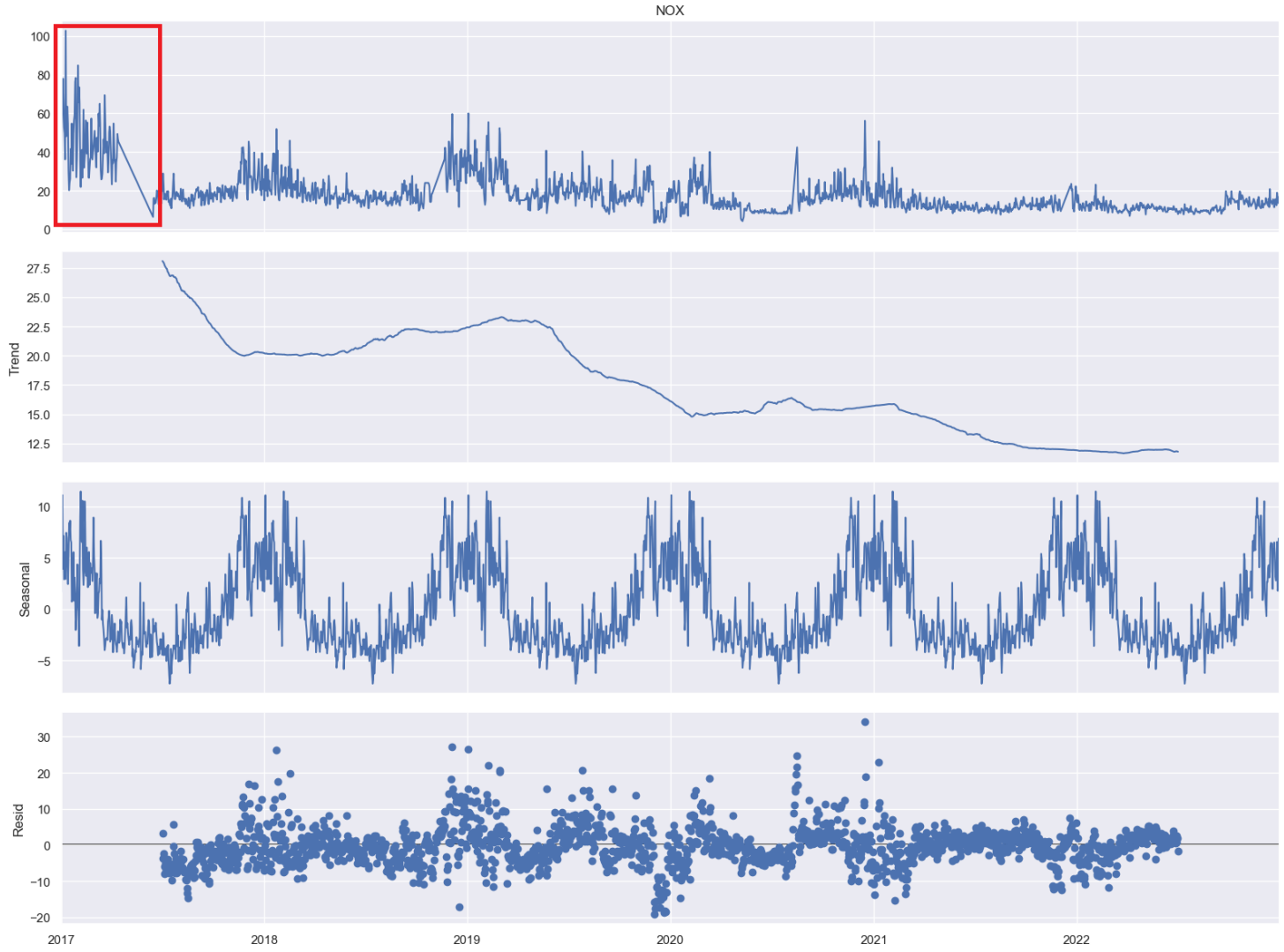




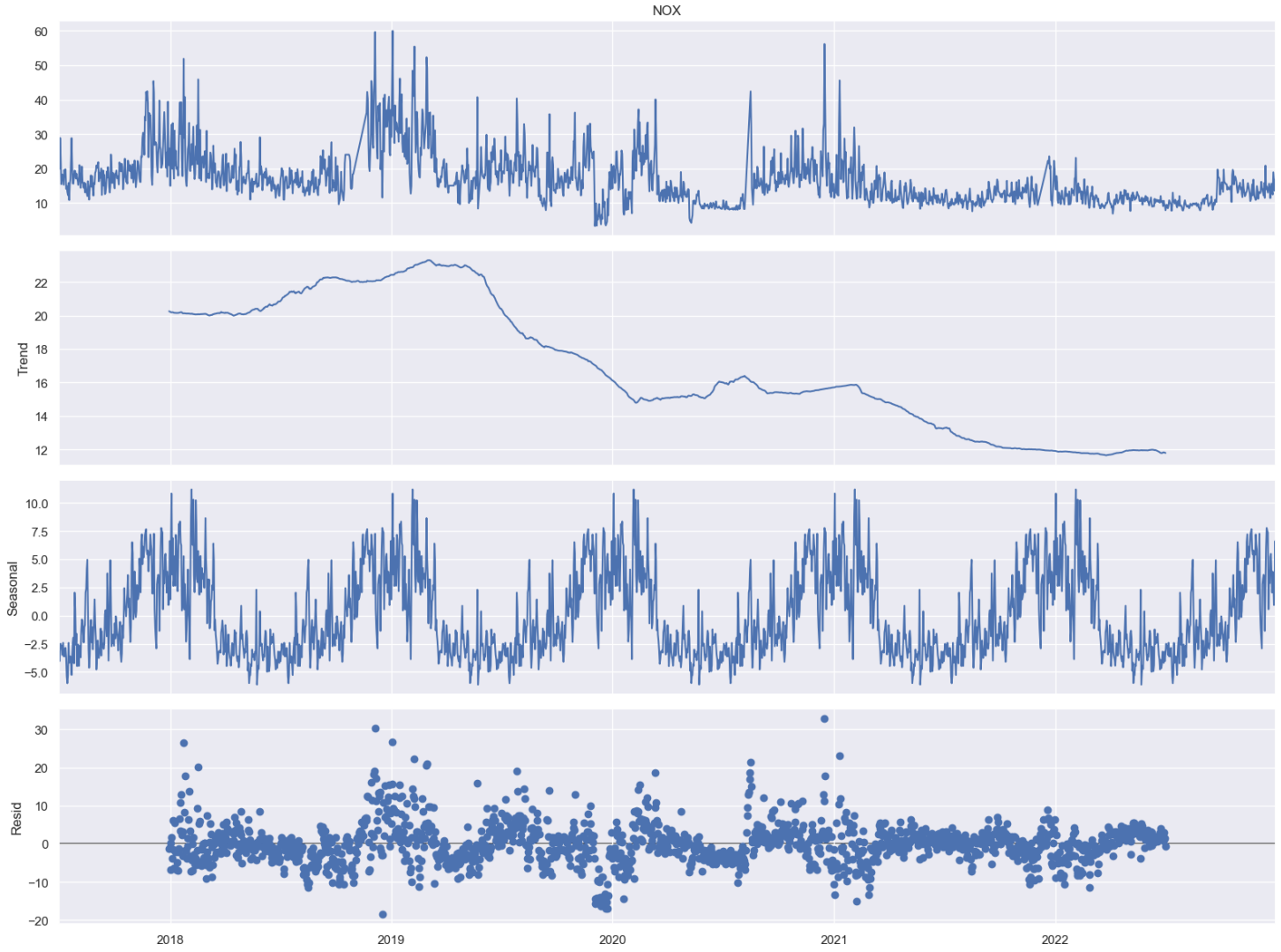
- La varianza apenas varía.

ÍNDICE

Preprocesamiento: Descomposición de la serie

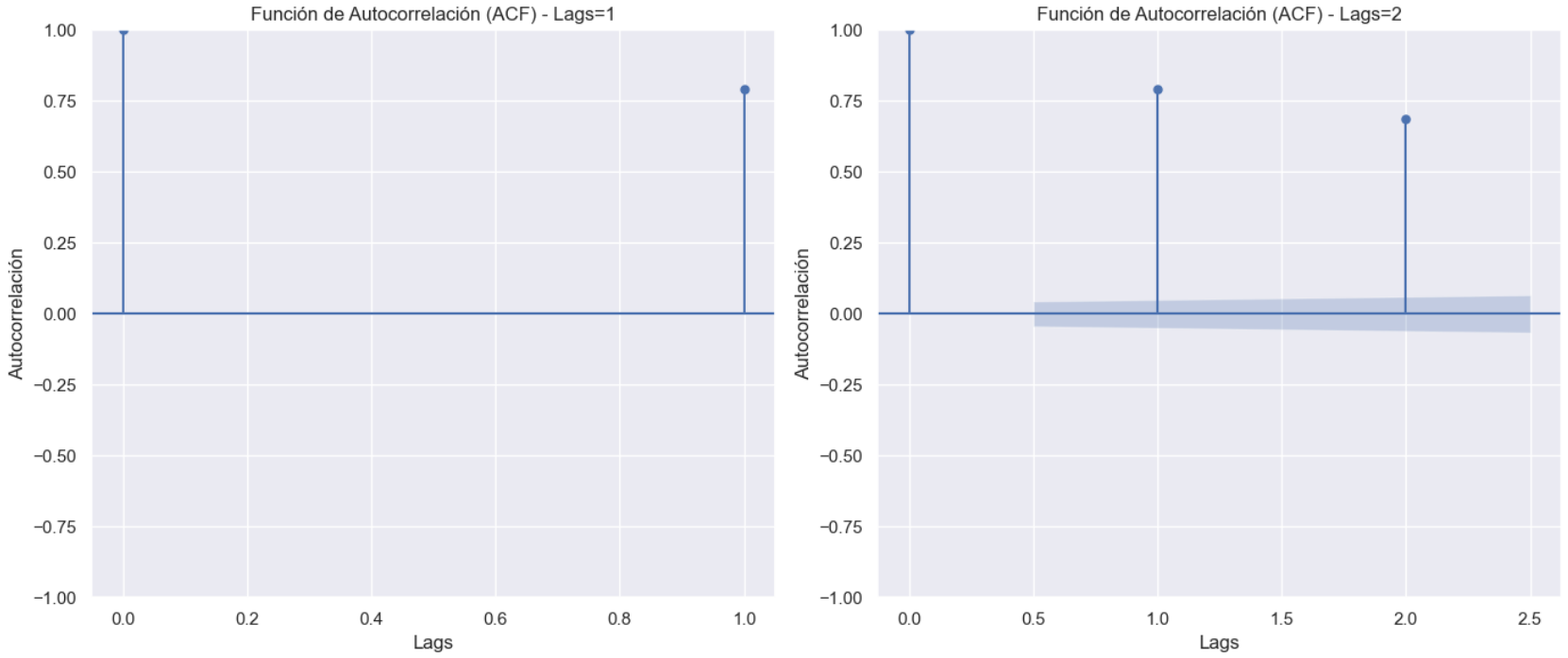


Preprocesamiento: Descomposición de la serie



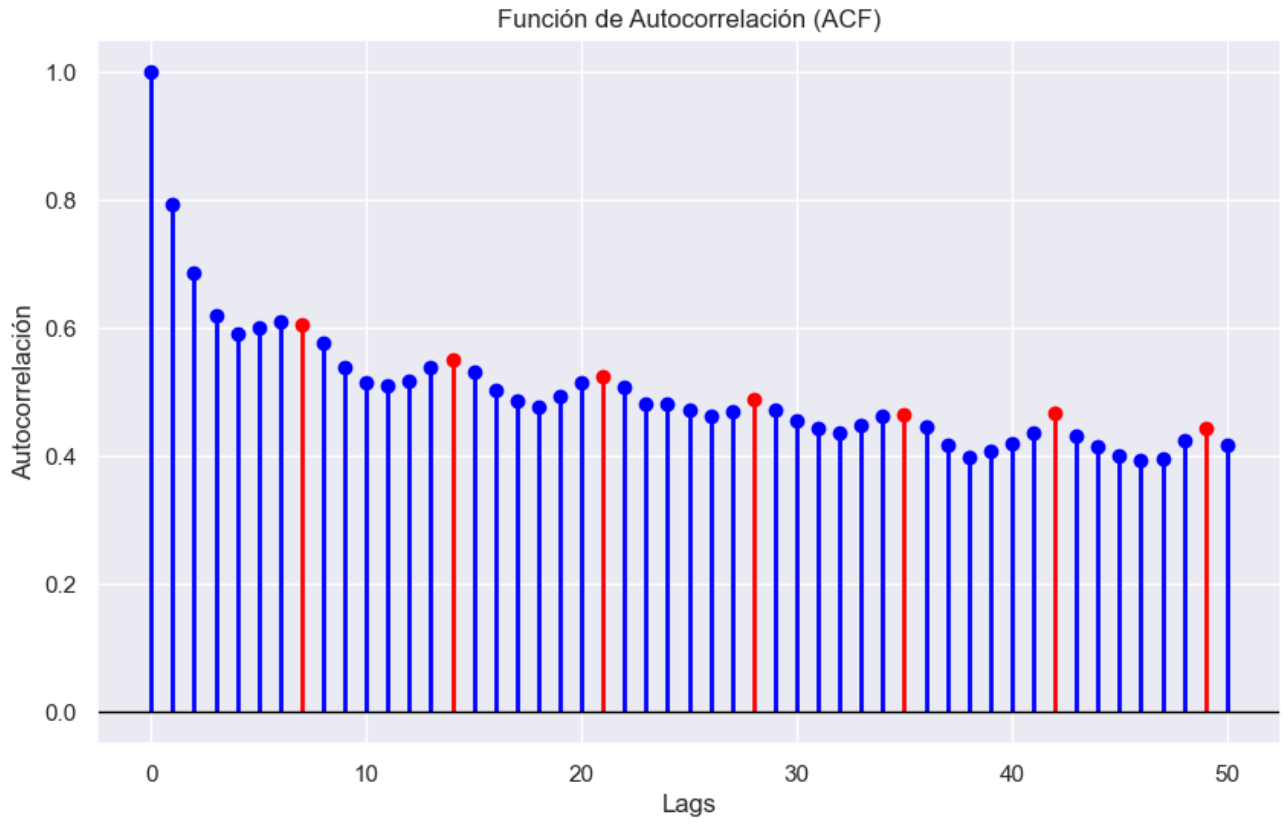
ÍNDICE

Preprocesamiento: Visualización de la Autocorrelación



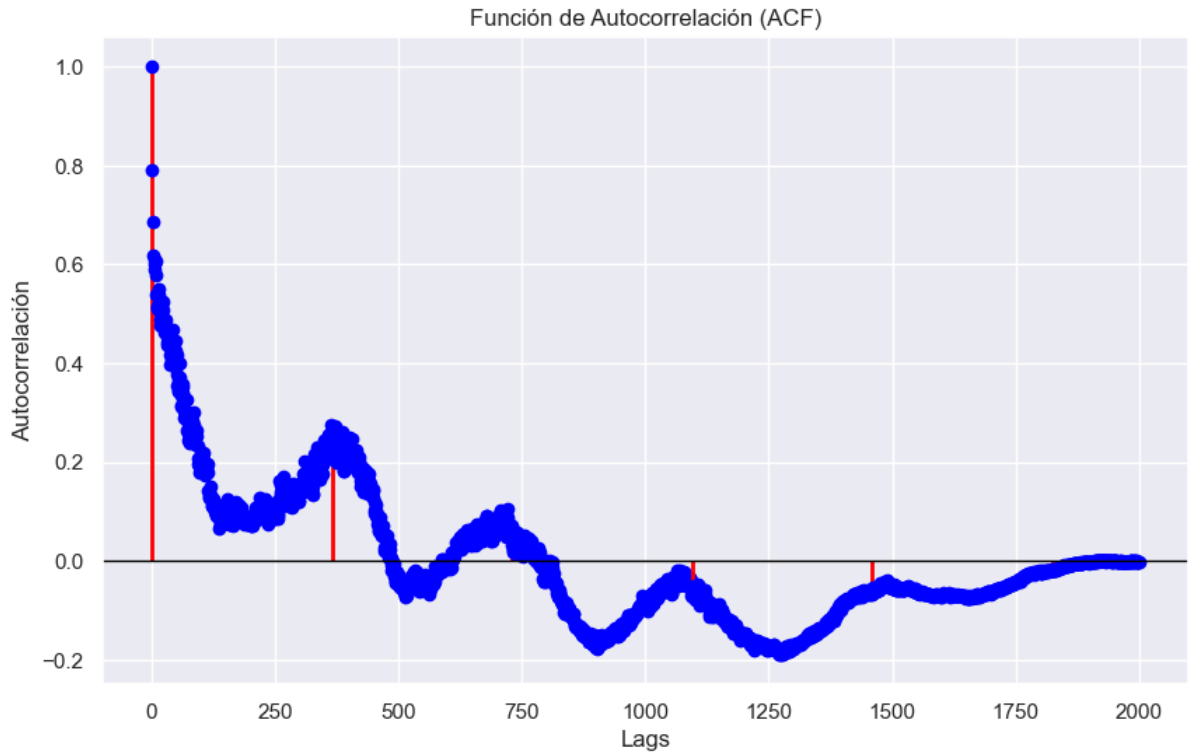
Valores altos en lags bajos → Componente autoregresiva

Preprocesamiento: Visualización de la Autocorrelación



Componente estacional con picos locales de correlación positiva cada 7 lags aproximádamente, que tiende a disminuir con el tiempo.

Preprocesamiento: Visualización de la Autocorrelación



Cada 365 lags existe un pico en el que la correlación es más positiva mostrando cierta estacionalidad anual, aunque la tendencia es a disminuir.

ÍNDICE

ÍNDICE

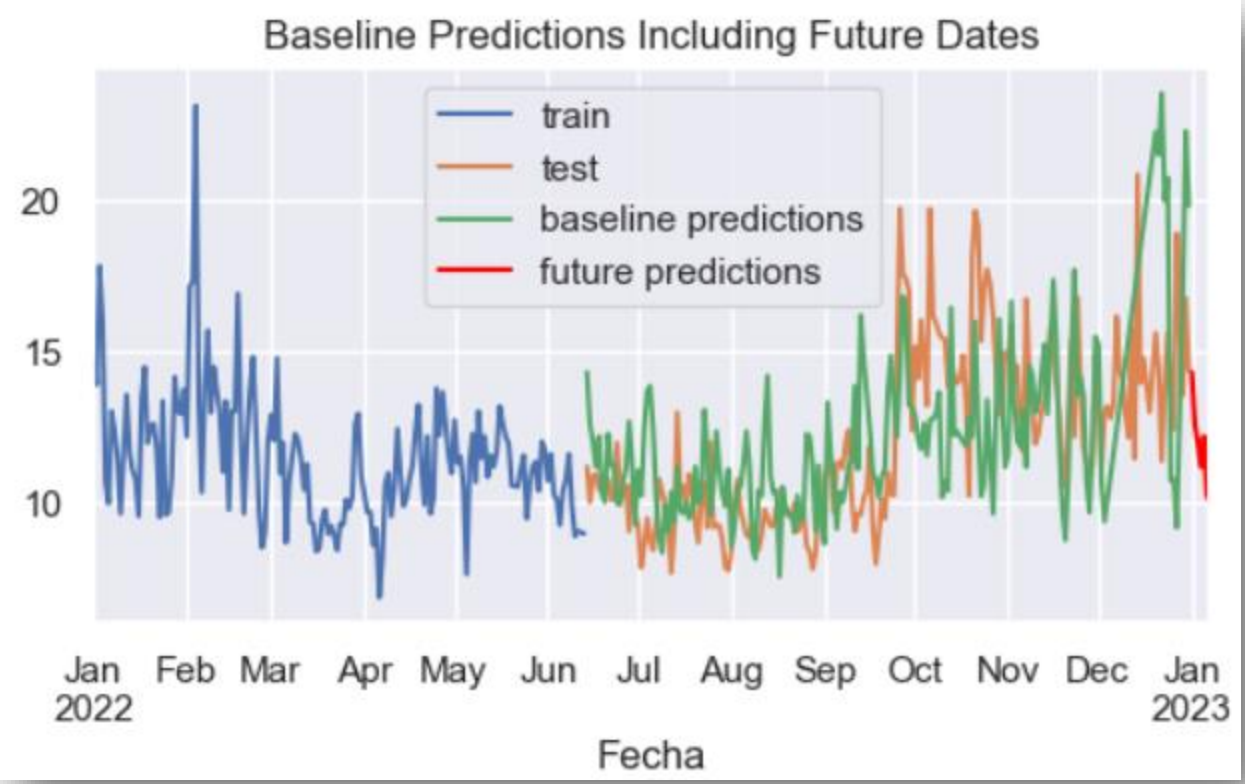
Modelos únicamente usando la variable NOX

MODELO BASELINE: ForecasterEquivalentData

- Hay **estacionalidad anual** en los datos.
- Se asume que el comportamiento de la serie en el mismo período de tiempo en el año anterior es una buena aproximación para el futuro.
- Este modelo predice el **valor para un día** como el valor del **mismo día en el año anterior**.
- Transformaciones de los datos:
  - Sin transformación
  - Transformación raíz cuadrada
  - Transformación logarítmica
  - Transformación Box-Cox
- Métricas:
  - Mean Absolute Percent Error (MAPE)
  - Root Mean Squared Error (RMSE)

PREDICCIONES: ForecasterEquivalentData

- Mismos resultados con y sin transformaciones.



Predicciones a 7 días

Fecha	Predicciones
2023-01-01	14.291667
2023-01-02	12.583333
2023-01-03	11.958333
2023-01-04	11.166667
2023-01-05	12.166667
2023-01-06	10.250000
2023-01-07	10.000000

ERRORES: ForecasterEquivalentData

- Error en el test:
  - **MAPE** = 19.49%
  - **RMSE** = 3.0975238105985543
- Error Backtesting:
  - Número de observaciones de entrenamiento inicial: 1000
  - Número de observaciones para el backtesting: 809
  - Número de folds para CV: 116
  - Número de pasos por fold: 7
  - **MAPE** Backtesting = 49.114614%
  - **RMSE** Backtesting = 8.479244

ÍNDICE

Modelos únicamente usando la variable NOX

MODELO CON VENTANA DESLIZANTE: RandomForestRegressor

Fecha	NOX_lag_4	NOX_lag_3	NOX_lag_2	NOX_lag_1	NOX
2017-07-01	NaN	NaN	NaN	NaN	23.083333
2017-07-02	NaN	NaN	NaN	23.083333	19.875000
2017-07-03	NaN	NaN	23.083333	19.875000	28.875000
2017-07-04	NaN	23.083333	19.875000	28.875000	22.934295
2017-07-05	23.083333	19.875000	28.875000	22.934295	15.419872
2017-07-06	19.875000	28.875000	22.934295	15.419872	17.875000
2017-07-07	28.875000	22.934295	15.419872	17.875000	18.250000

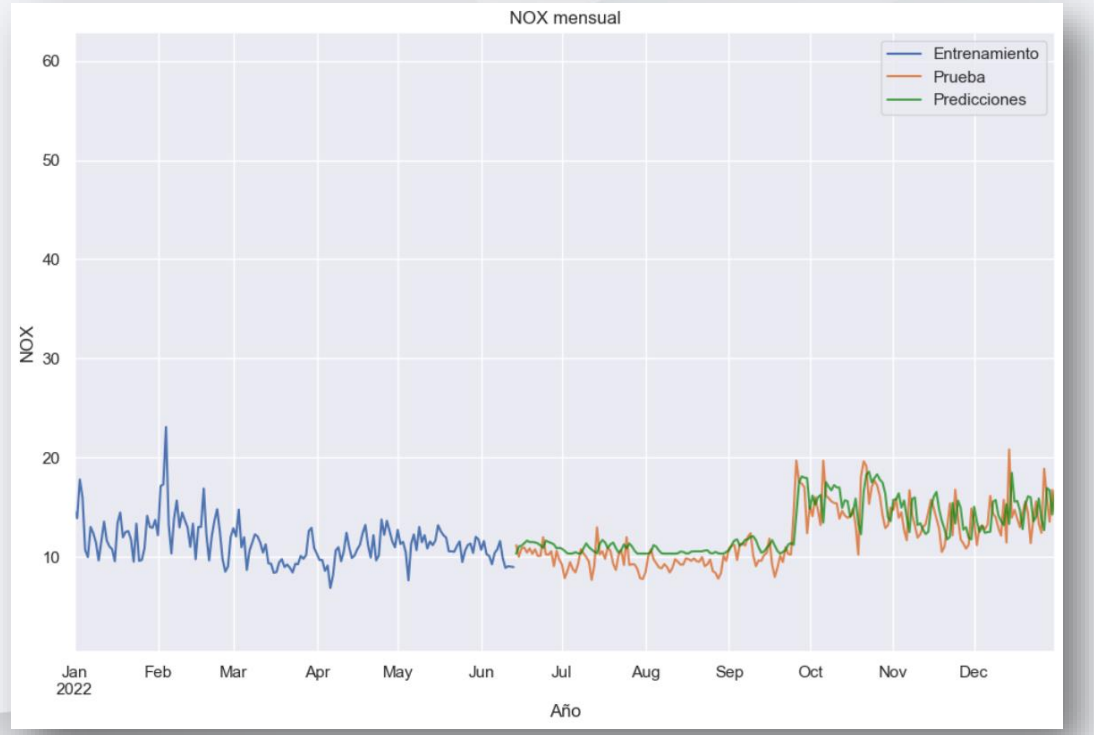
Se eliminan las 4 primeras filas

MODELO CON VENTANA DESLIZANTE: RandomForestRegressor

Sin transformación

- Parametros óptimos obtenidos mediante CV:
  - n_estimators = 100
  - min_samples_split = 10
  - min_samples_leaf = 10
  - max_features = 'log2'
  - max_depth = 3
  - bootstrap = True

- MAPE = 27.21%
- RMSE = 2.0466



Predicciones a 7 días

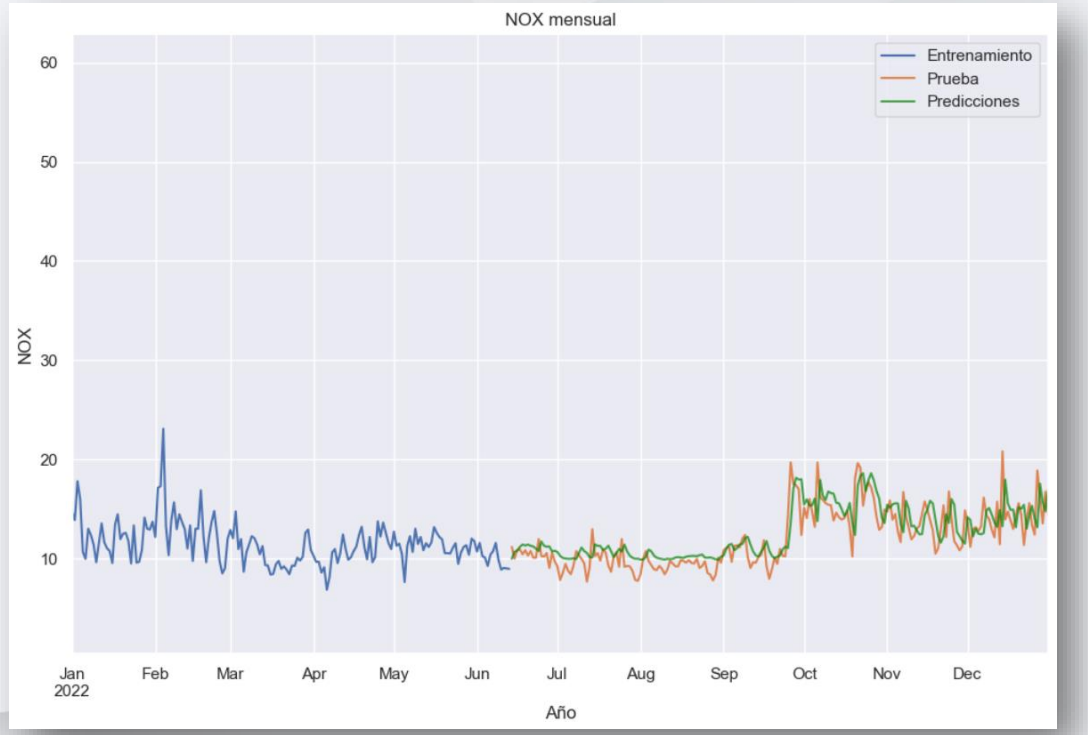
Fecha	Predicciones
2023-01-01	16.733804
2023-01-02	17.228232
2023-01-03	17.631295
2023-01-04	18.224915
2023-01-05	18.454472
2023-01-06	18.591535
2023-01-07	18.757378

MODELO CON VENTANA DESLIZANTE: RandomForestRegressor

Transformación de raíz cuadrada

- Parametros óptimos obtenidos mediante CV:
  - n_estimators = 100
  - min_samples_split = 10
  - min_samples_leaf = 10
  - max_features = 'log2'
  - max_depth = 3
  - bootstrap = True

- MAPE = 26.77%
- RMSE = 1.9121



Predicciones a 7 días

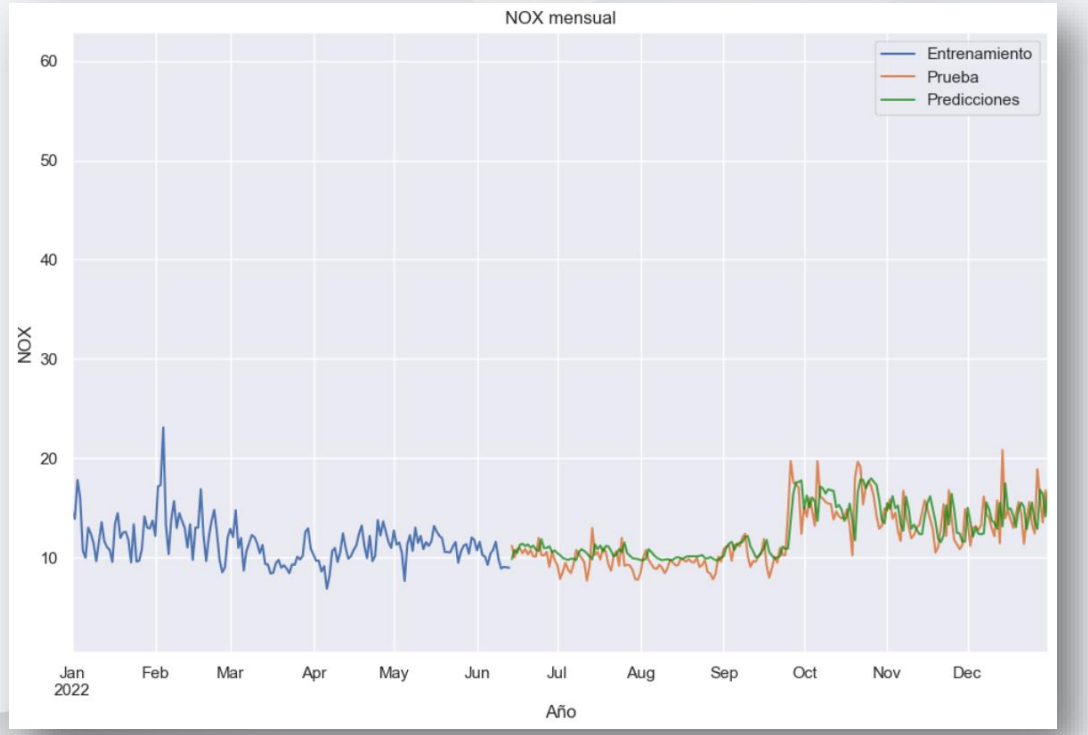
Fecha	Predicciones
2023-01-01	17.002808
2023-01-02	17.661727
2023-01-03	18.161260
2023-01-04	18.819571
2023-01-05	18.846402
2023-01-06	18.846402
2023-01-07	18.846402

MODELO CON VENTANA DESLIZANTE: RandomForestRegressor

Transformación logarítmica

- Parametros óptimos obtenidos mediante CV:
  - n_estimators = 100
  - min_samples_split = 10
  - min_samples_leaf = 10
  - max_features = 'log2'
  - max_depth = 3
  - bootstrap = True

- MAPE = 26.45%
- RMSE = 1.9083



Predicciones a 7 días

Fecha	Predicciones
2023-01-01	16.891637
2023-01-02	17.402468
2023-01-03	17.515466
2023-01-04	17.851892
2023-01-05	18.023211
2023-01-06	18.126589
2023-01-07	18.126589

MODELO CON VENTANA DESLIZANTE: RandomForestRegressor

Transformación Box-Cox

- Parametros óptimos obtenidos mediante CV:
  - n_estimators = 300
  - min_samples_split = 10
  - min_samples_leaf = 4
  - max_features = 'log2'
  - max_depth = 3
  - bootstrap = True

- MAPE = 25.77%
- RMSE = 1.8945

Predicciones a 7 días

Fecha	Predicciones
2023-01-01	16.589740
2023-01-02	16.868352
2023-01-03	16.973343
2023-01-04	17.445434
2023-01-05	17.650590
2023-01-06	17.826011
2023-01-07	17.939082

ÍNDICE

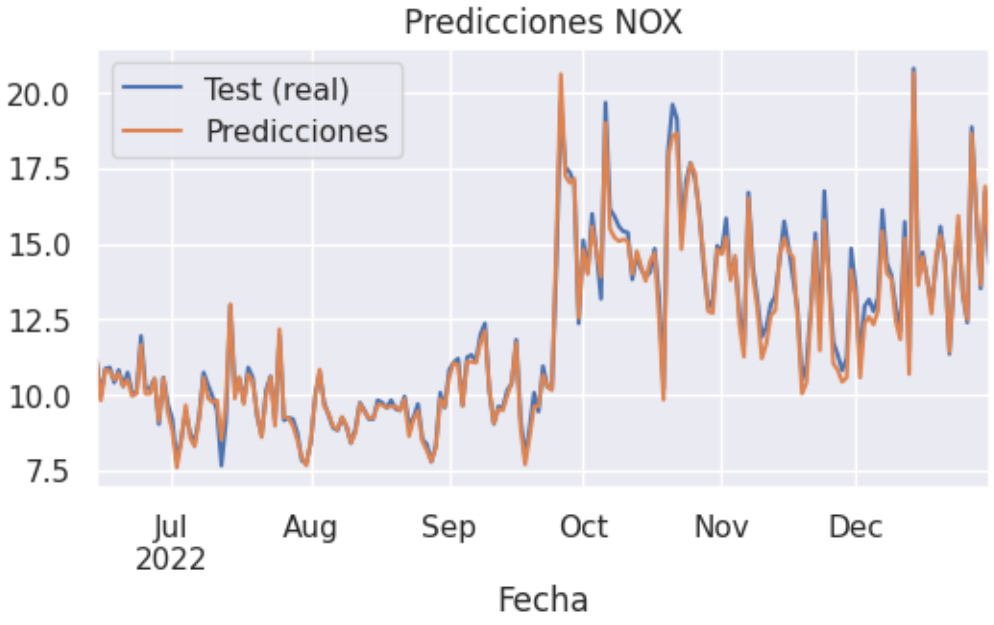
Modelos con variables endógenas

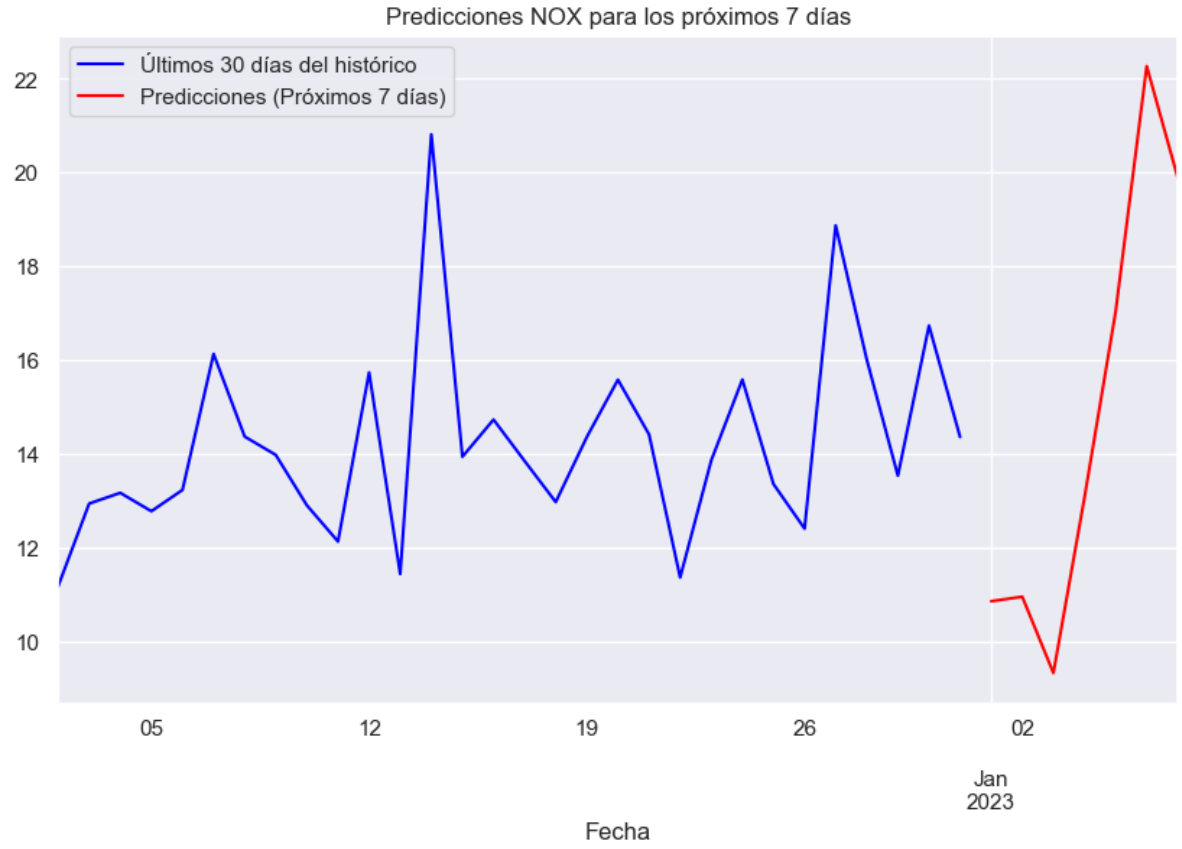
- Modelos con solo datos de calidad del aire:
  - RandomForest: Sin transformación
  - RandomForest: Transformación raíz cuadrada
  - RandomForest: Transformación logarítmica
  - RandomForest: Transformación Box-Cox
- Modelos con datos de calidad del aire y meteorológicos:
  - RandomForest: Sin transformación
  - RandomForest: Transformación raíz cuadrada
  - RandomForest: Transformación logarítmica
  - RandomForest: Transformación Box-Cox

ÍNDICE

Modelos con solo datos de calidad del aire: RandomForest

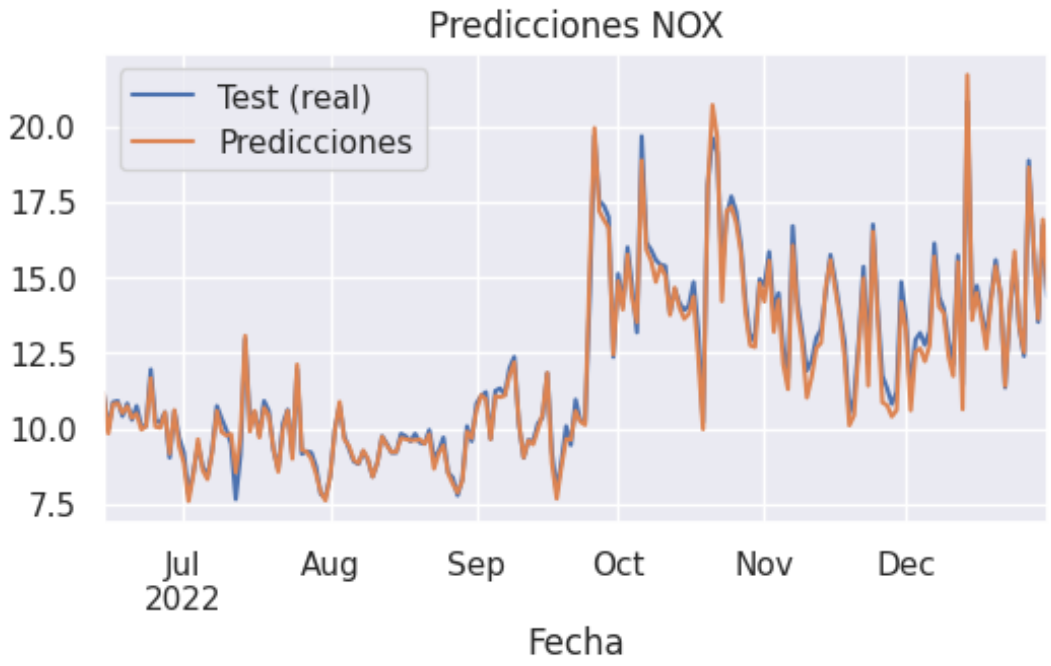
- Sin transformación:
  - MAPE: 1.99%
  - RMSE: 0.3292

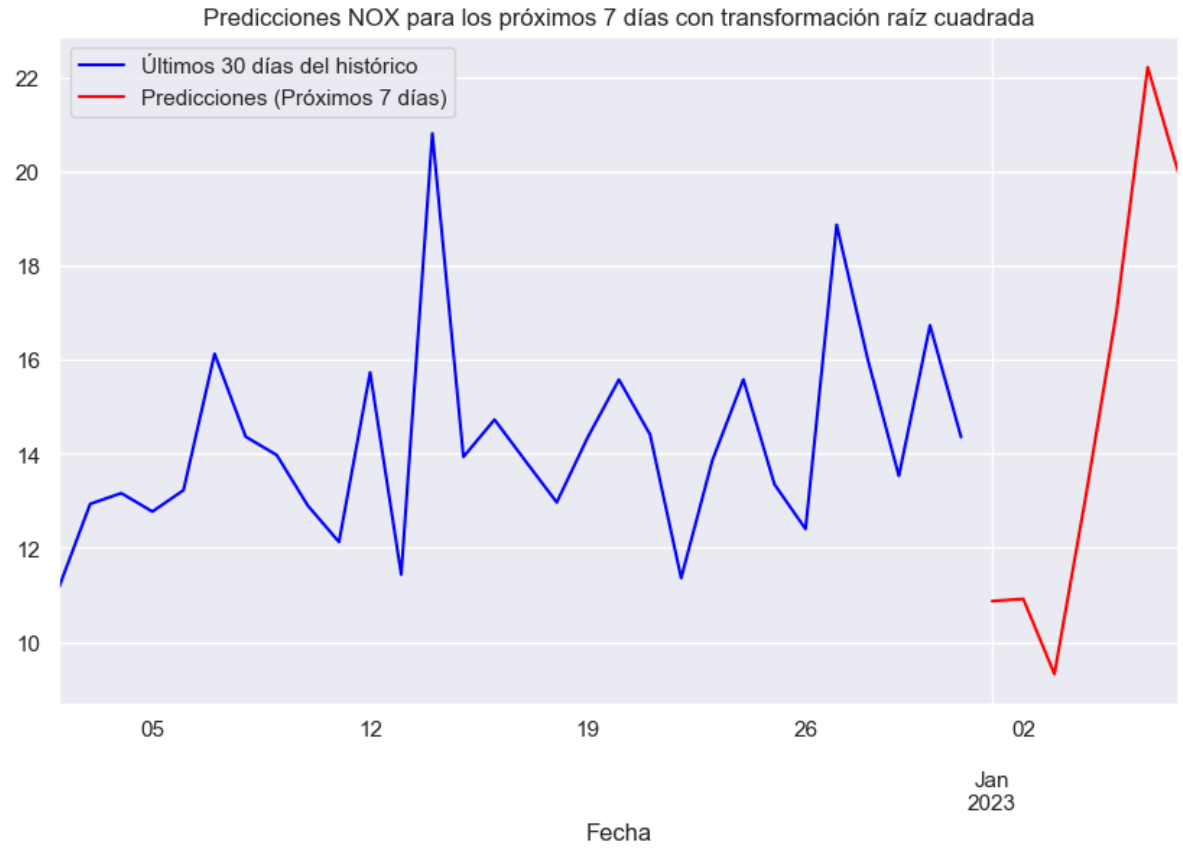




Modelos con solo datos de calidad del aire: RandomForest

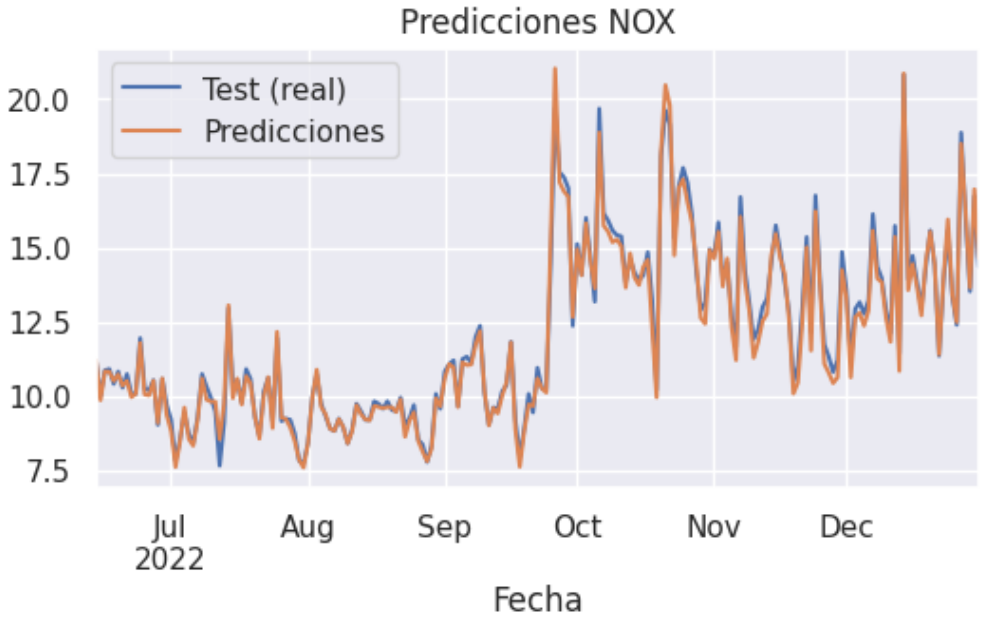
- Transformación raíz cuadrada:
  - MAPE: 2.05%
  - RMSE: 0.3395

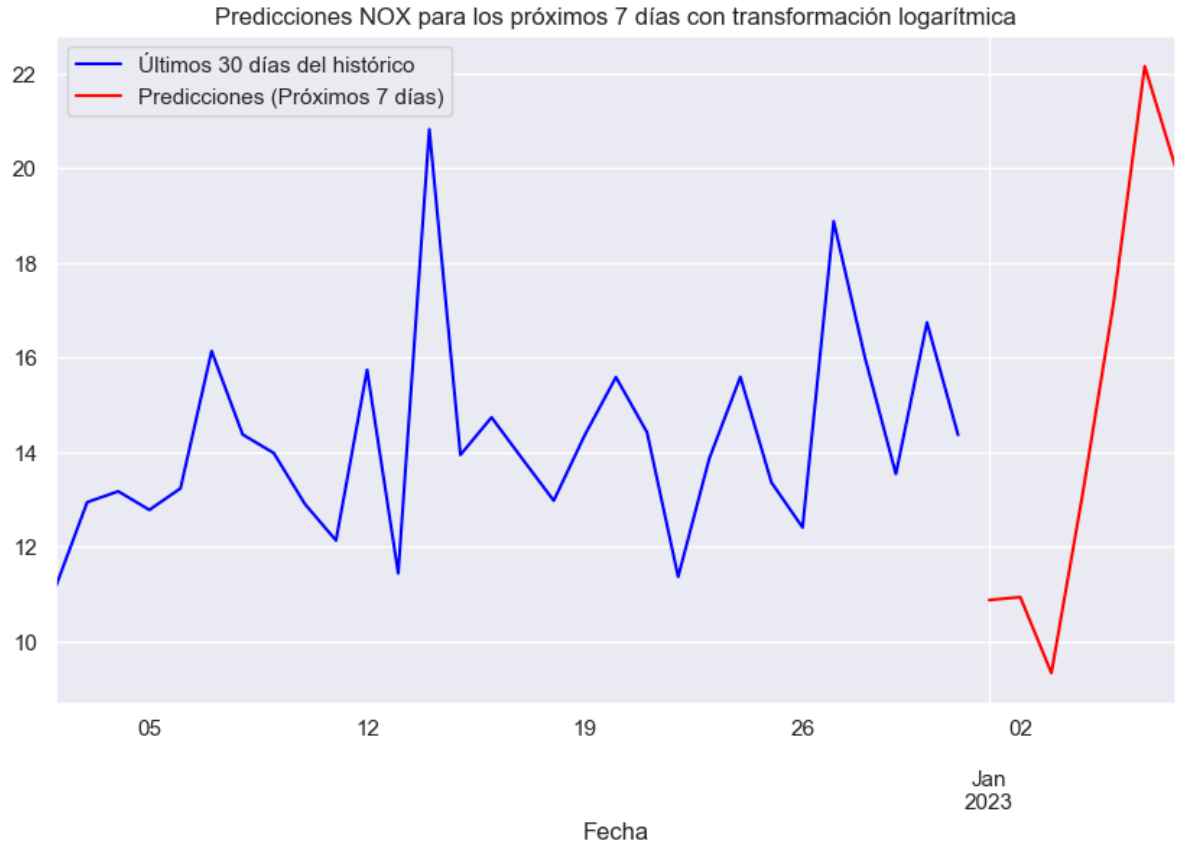




Modelos con solo datos de calidad del aire: RandomForest

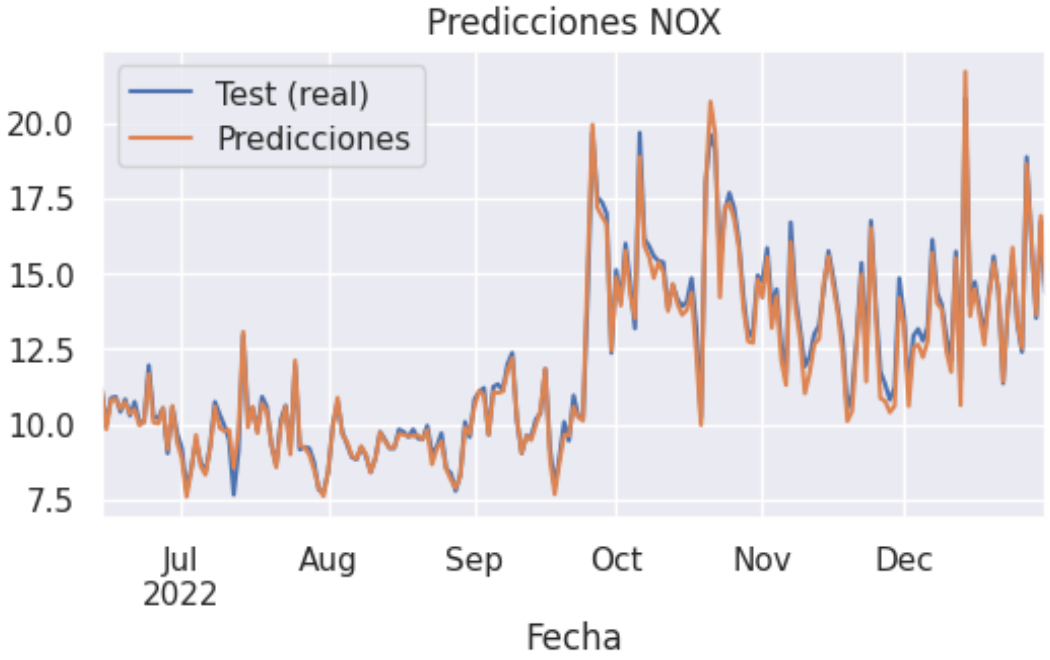
- Transformación logarítmica:
  - MAPE: 1.95%
  - RMSE: 0.3230

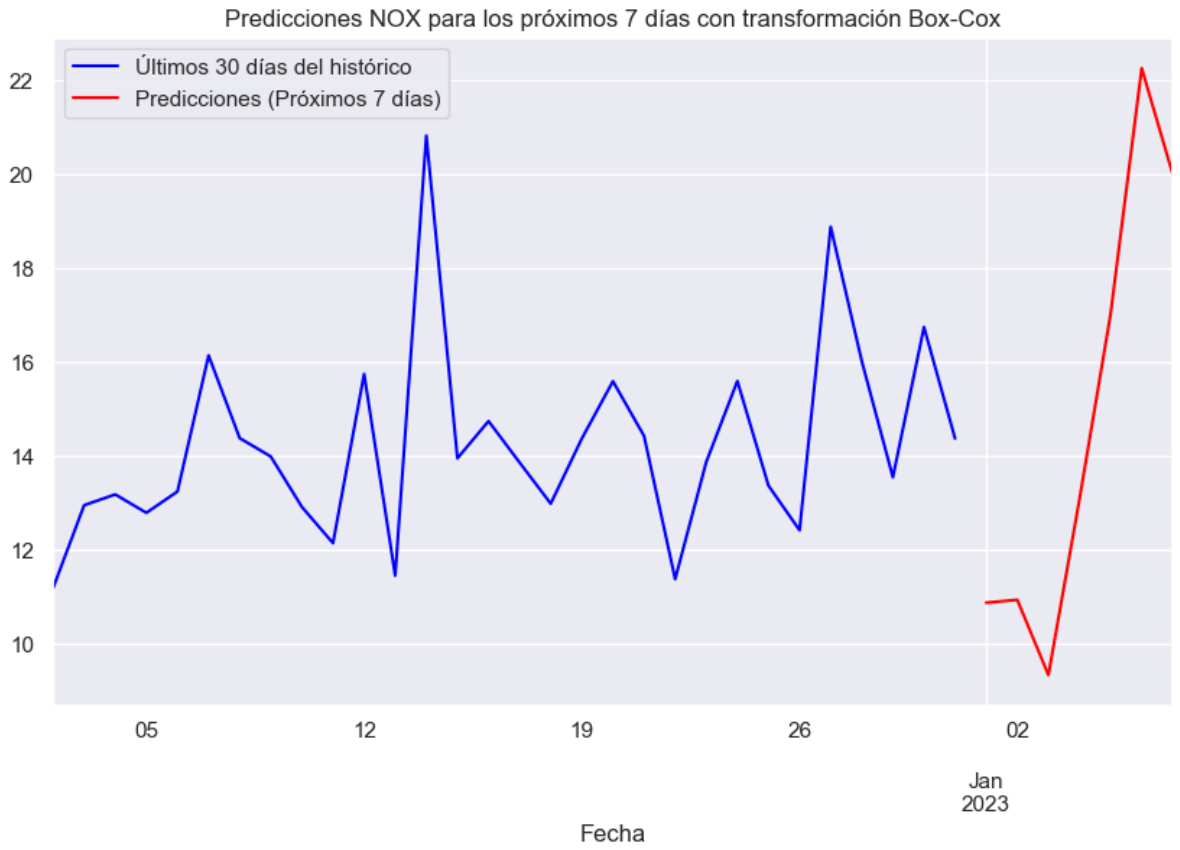




Modelos con solo datos de calidad del aire: RandomForest

- Transformación Box-Cox:
  - MAPE: 1.89%
  - RMSE: 0.3092

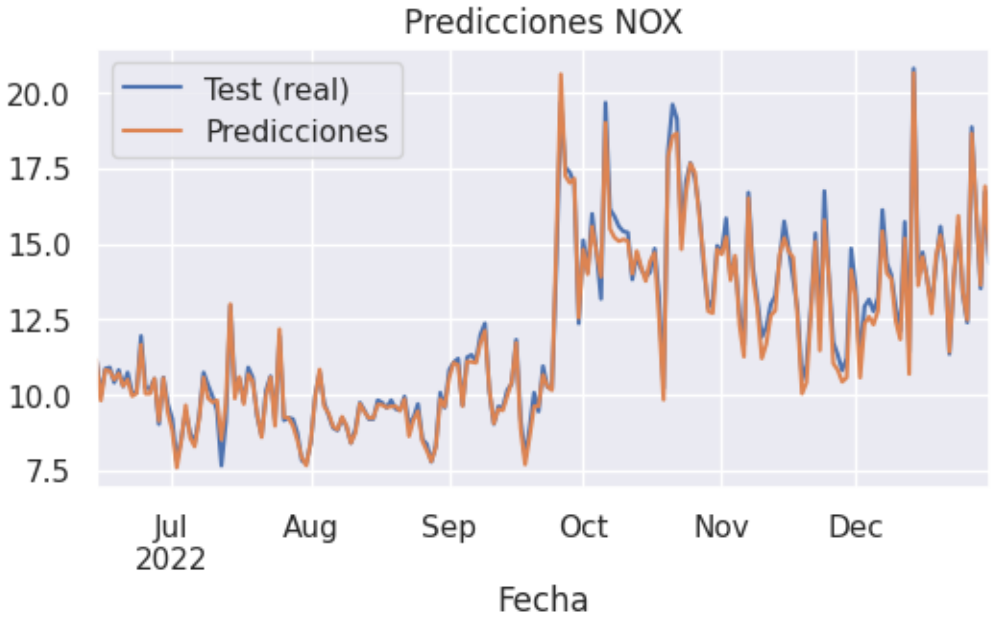


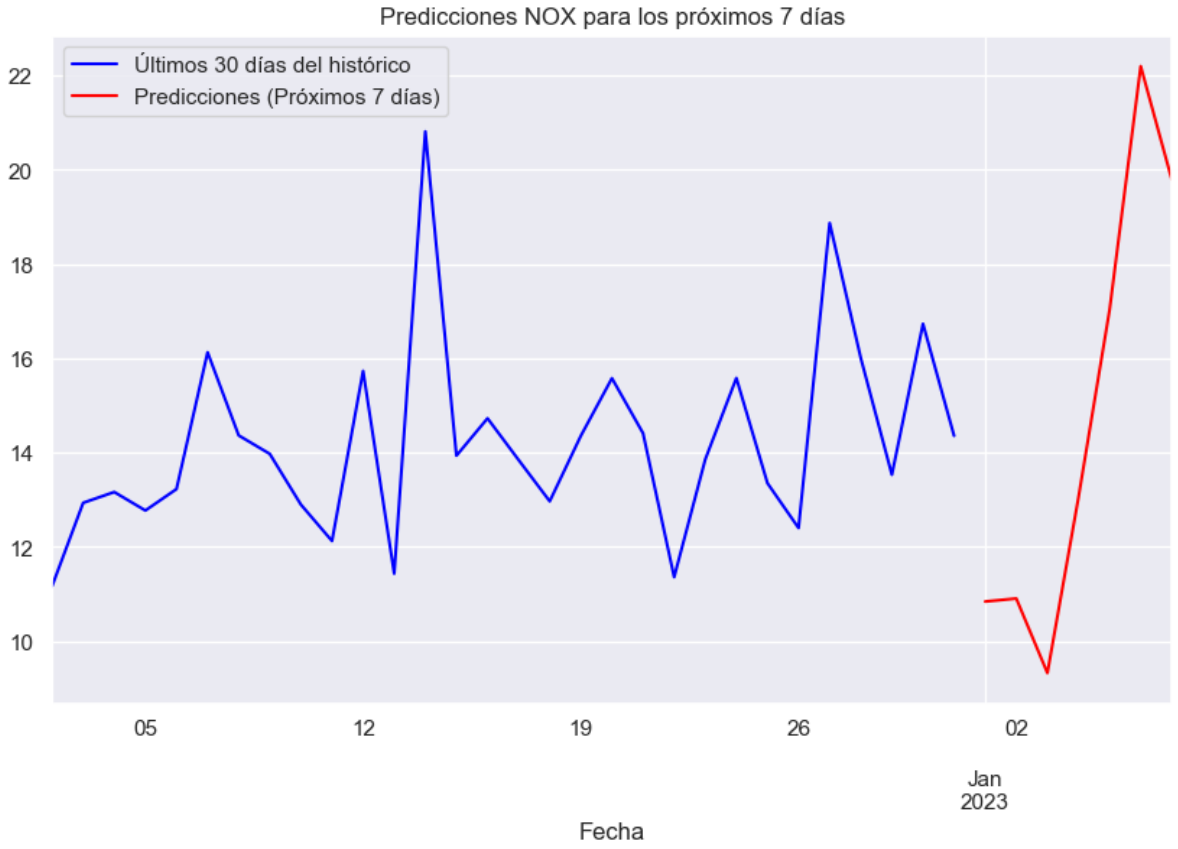


ÍNDICE

Modelos con datos de calidad del aire y meteorológicos: RandomForest

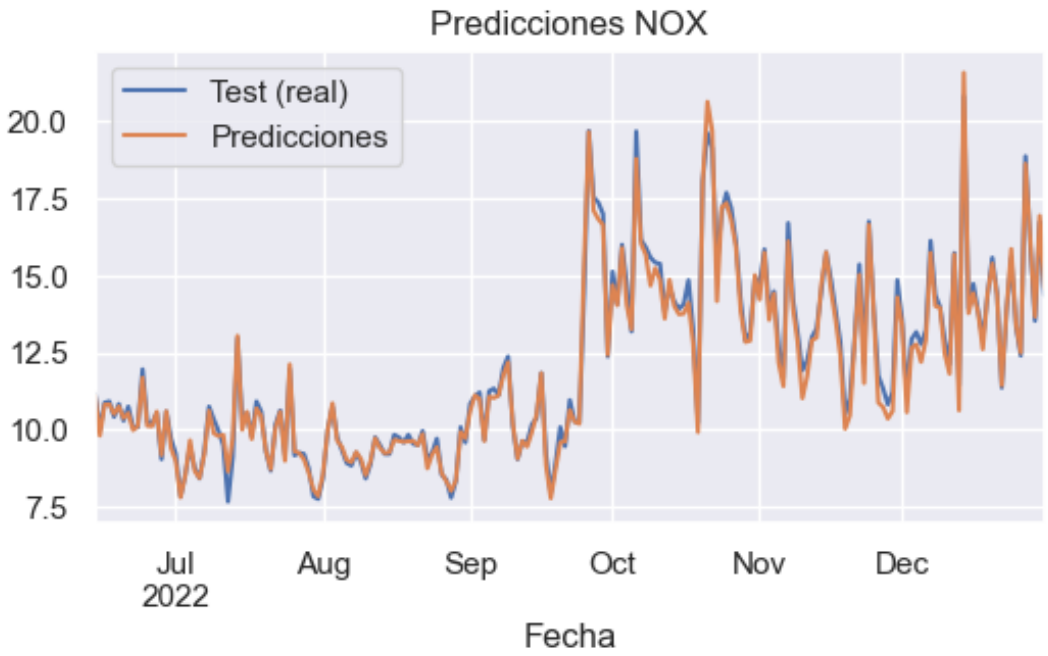
- Sin transformación:
  - MAPE: 1.90%
  - RMSE: 0.3242

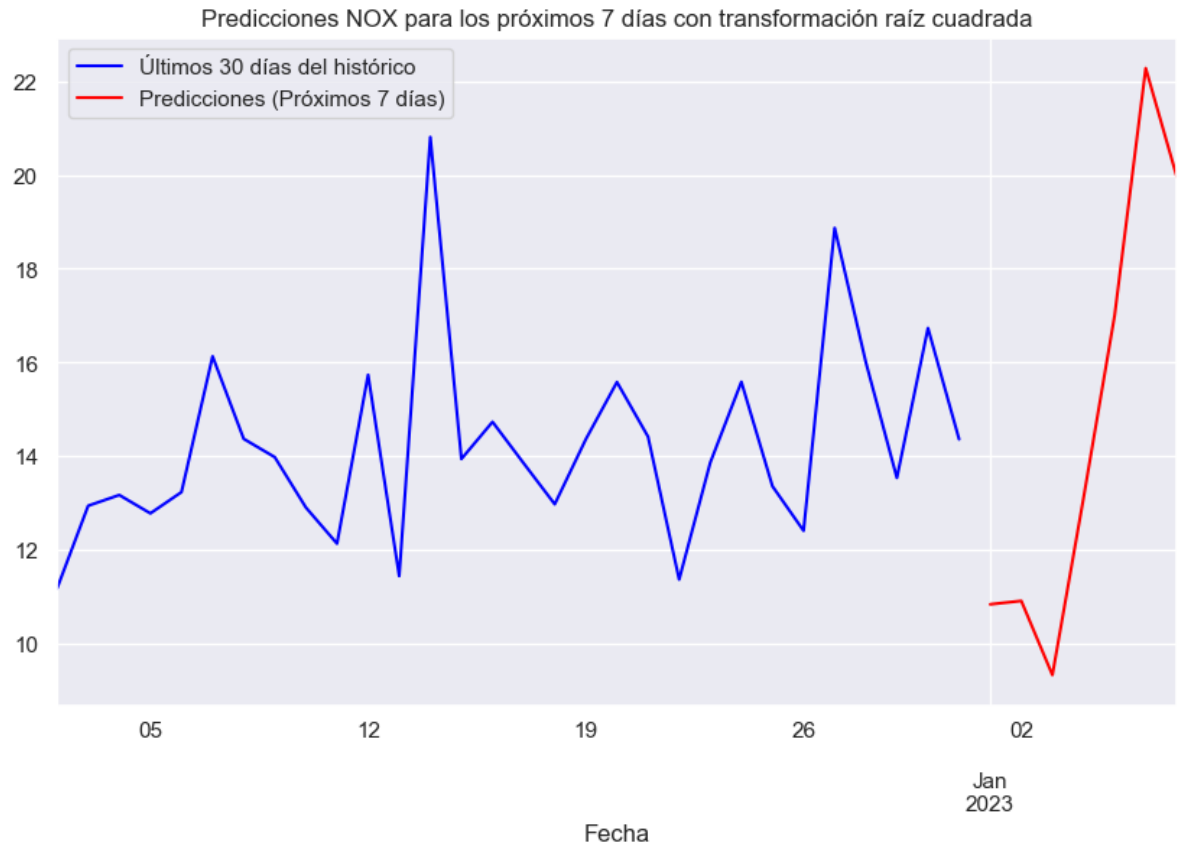




Modelos con datos de calidad del aire y meteorológicos: RandomForest

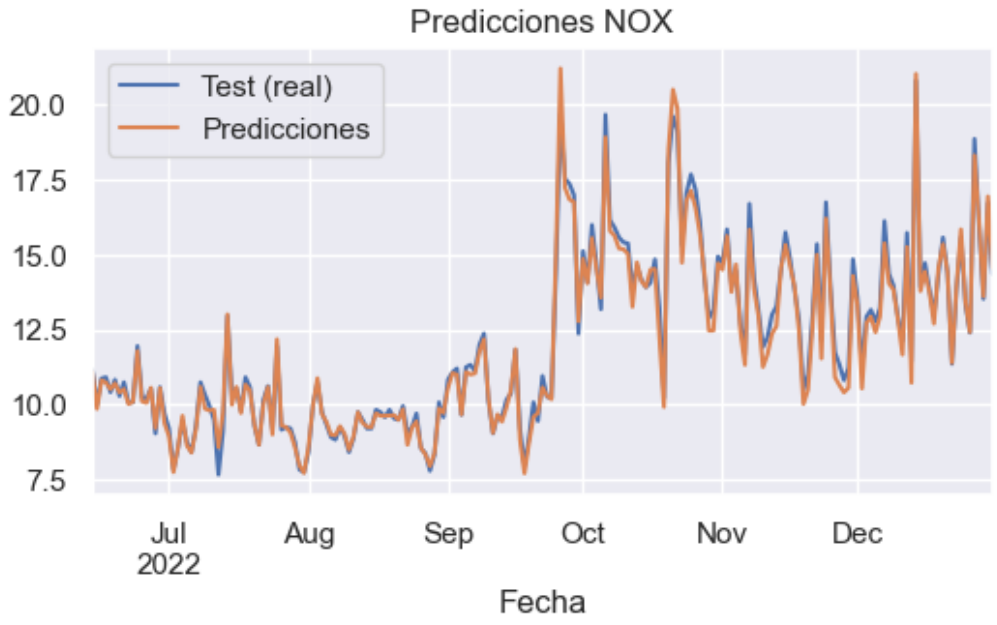
- Transformación raíz cuadrada:
  - MAPE: 1.92%
  - RMSE: 0.3290

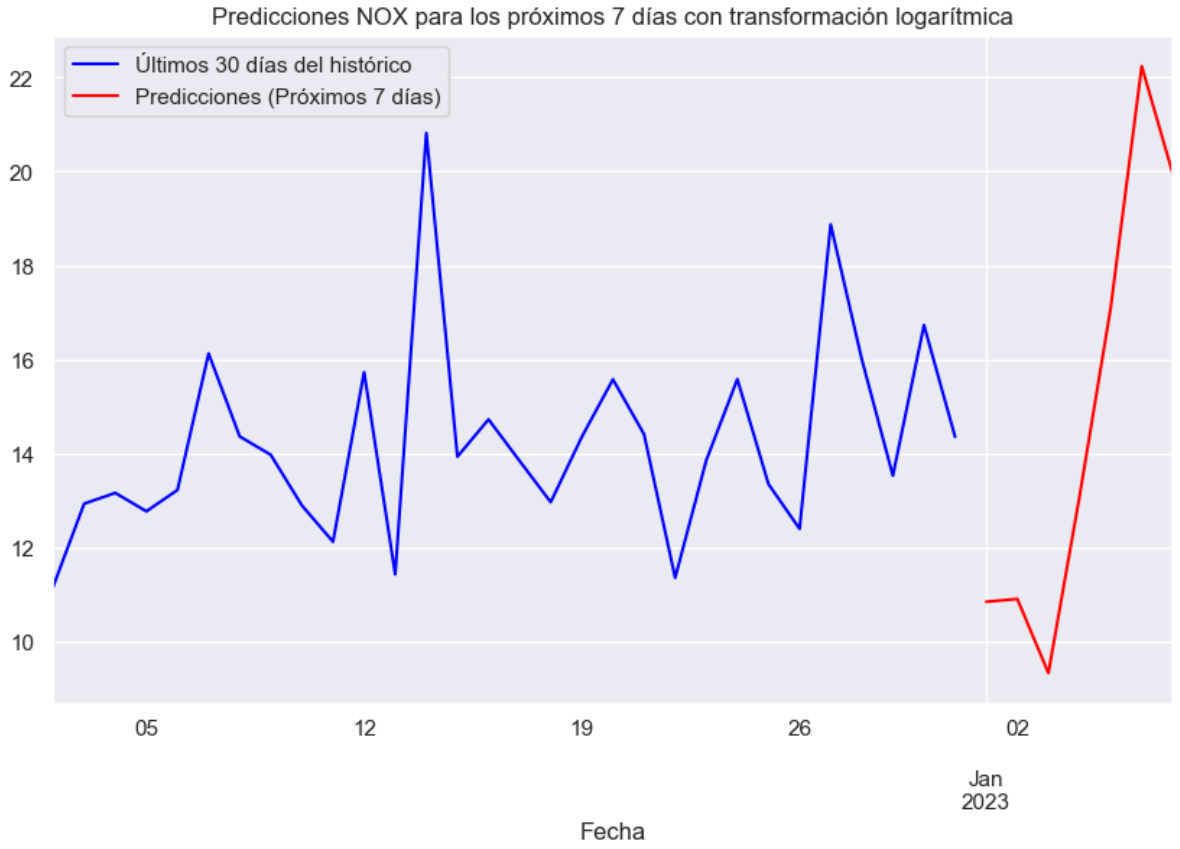




Modelos con datos de calidad del aire y meteorológicos: RandomForest

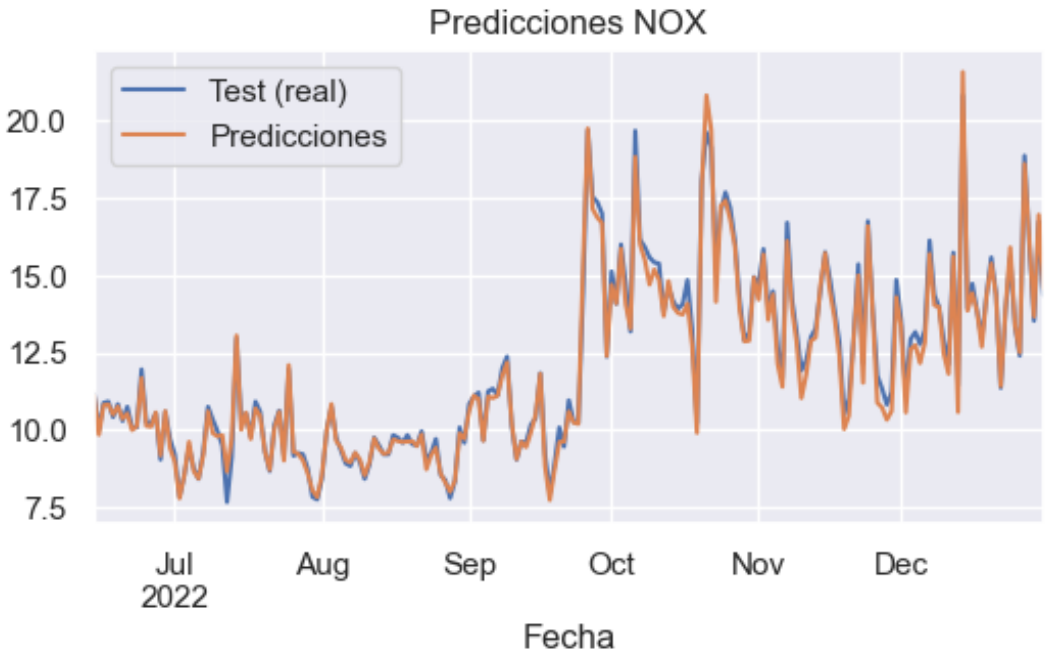
- Transformación logarítmica:
  - MAPE: 2.03%
  - RMSE: 0.3445

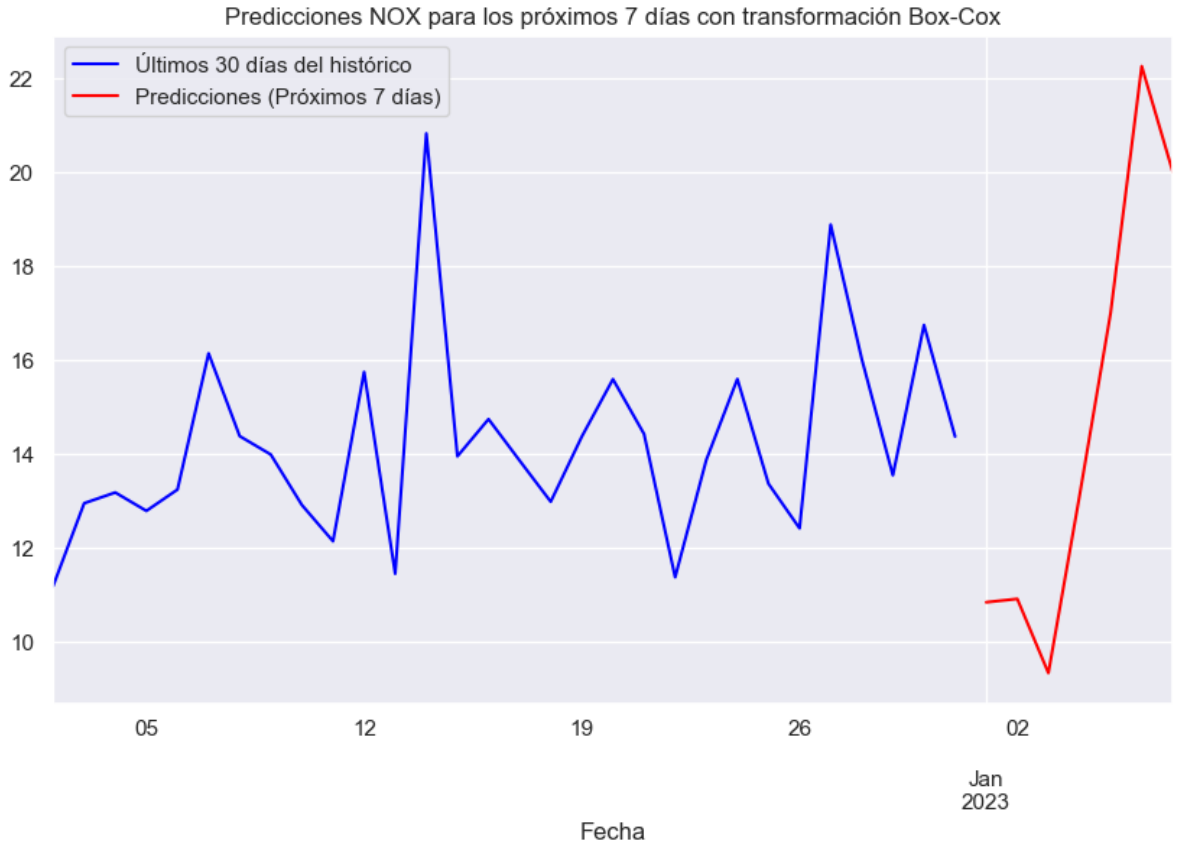




Modelos con datos de calidad del aire y meteorológicos: RandomForest

- Transformación Box-Cox:
  - MAPE: 1.93%
  - RMSE: 0.3337





ÍNDICE

Conclusiones

- El modelo baseline no da resultados demasiado buenos debido a su simplicidad.
- Los modelos con ventana deslizante dan mejores resultados en cuanto al RMSE, aunque podrían estar presentando algo de overfitting.
- Los modelos con variables endógenas dan bastantes buenos resultados.
  - Se debe tener en cuenta que esto solo es posible si en la predicción contamos con el valor de dichas variables, ya sea debido a que se nos proporciona o a que lo podemos predecir.