PREDICCIÓN DE CONCENTRACIÓN DE SO2 EN EL AIRE USANDO MACHINE LEARNING

Forecasting of SO2 concentration in the air using machine learning

JOSE MANUEL GOMEZ JIMENEZ

Tesis

Asesor

Pablo Andres Saldarriaga Aristizabal

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2026

Predicción de concentración de SO2 en el aire usando machine learning

José Manuel Gómez Jiménez[1] Pablo Andres Saldarriaga Aristizabal[2]

[1] Universidad EAFIT, Medellín – Antioquia, Colombia
[2] Universidad EAFIT, Medellín – Antioquia, Colombia

Resumen: La calidad del aire es un tema de creciente preocupación a nivel mundial debido al impacto de la contaminación atmosférica en la salud y el medio ambiente. Los contaminantes criterio, como el material particulado (PM), el monóxido de carbono (CO), el ozono (O3), el dióxido de nitrógeno (NO2) y el dióxido de azufre (SO_2), representan un riesgo significativo, vinculándose con enfermedades respiratorias y cardiovasculares, además de contribuir al cambio climático. En el Valle de Aburrá, esta problemática se agrava debido a las características topográficas de la región, que limitan la dispersión de contaminantes e intensifican los episodios de mala calidad del aire. A pesar de los avances en monitoreo a través del Sistema de Alerta Temprana del Valle de Aburrá (SIATA), los esfuerzos predictivos se han enfocado principalmente en el $PM_{2.5}$, dejando un vacío en la predicción de otros contaminantes criterio. Este proyecto propone desarrollar modelos de machine learning que, utilizando datos históricos y variables ambientales, mejoren la predicción de contaminantes enfocada en el SO_2, proporcionando, mediante el uso de diferentes modelos de machine learning, herramientas para mejorar la gestión ambiental. Esto permitirá fortalecer las estrategias de mitigación y la toma de decisiones, reduciendo los impactos en la salud pública y el entorno natural de la región.

Palabras clave: Calidad del aire, Contaminantes criterio, Valle del Aburrá, SIATA, Machine Learning, Pronóstico de contaminantes

1 Introducción

La calidad del aire se ha convertido en un tema de creciente preocupación a nivel mundial debido al aumento sostenido de la contaminación atmosférica en las últimas décadas. Según la Organización Mundial de la Salud (OMS), más del 90% de la población mundial respira aire que excede los niveles de calidad recomendados [1], lo que pone en riesgo la salud de millones de personas. Este incremento se asocia principalmente al crecimiento industrial, la quema de combustibles fósiles, la agricultura, las emisiones de fábricas e industrias, los vehículos automotores y diversos fenómenos naturales [2].

Los denominados contaminantes criterio (material particulado (PM), monóxido de carbono (CO), ozono (O_3), dióxido de nitrógeno (NO_2) y dióxido de azufre (SO_2)) representan las principales amenazas para la salud pública [2]. Estos contaminantes están vinculados con un aumento en el número de casos de enfermedades respiratorias y cardiovasculares, así como con un incremento de la mortalidad prematura [3]. Además, contribuyen al cambio climático y al deterioro de los ecosistemas, afectando la biodiversidad y los recursos naturales.

En el Valle de Aburrá (Antioquia, Colombia), la contaminación atmosférica constituye un problema particularmente complejo. La combinación de alta densidad poblacional, tráfico vehicular intenso y actividades industriales genera un alto volumen de emisiones contaminantes. Sin embargo, el impacto se ve intensificado por la topografía cerrada del valle, que limita la circulación del aire y provoca fenómenos de estancamiento atmosférico, lo que dificulta la dispersión de los contaminantes [4]. Estas condiciones conducen a episodios recurrentes de mala calidad del aire, especialmente en temporadas específicas del año, lo que hace que la región sea especialmente vulnerable.

Ante esta situación, el Sistema de Alerta Temprana del Valle de Aburrá (SIATA) ha desempeñado un papel clave en el monitoreo constante de los principales contaminantes criterio. Este ha desarrollado diversas iniciativas de modelamiento y predicción, sin embargo, los esfuerzos se han centrado principalmente en el material particulado $PM_{2.5}$ [4], dejando un vacío importante en la predicción del resto de contaminantes. Esta limitación es preocupante, dado que dichos contaminantes también están regulados por organismos internacionales, como la OMS, y se asocian con impactos severos sobre la salud humana y el medio ambiente.

Dentro de estos otros contaminantes, el SO_2 cobra particular importancia por sus impactos tanto sobre la salud humana como sobre los ecosistemas. Según la United States Environmental Protection Agency (EPA), la inhalación de SO_2 en concentraciones elevadas puede irritar el sistema respiratorio y agravar enfermedades como el asma, especialmente en niños, adultos mayores y personas con enfermedades respiratorias preexistentes [5]. Incluso breves exposiciones, del orden de minutos u horas, pueden desencadenar broncoconstricción y síntomas respiratorios agudos. Además, las emisiones de SO_2 generan otros óxidos de azufre (SO_x) que, en presencia de otros compuestos atmosféricos, pueden transformarse en partículas finas capaces de penetrar profundamente en los pulmones y desencadenar afecciones cardiovasculares o respiratorias graves [5].

La evidencia reciente muestra que los efectos del SO_2 en la salud no son un hallazgo aislado sino un patrón consistente en distintos contextos y poblaciones. Un estudio de revisión sistemática realizado en 2021 identificó asociaciones positivas significativas entre la exposición ambiental a corto plazo al SO_2 y el aumento en las muertes tanto por cualquier causa como por problemas respiratorios [6]. Los autores destacaron, además, que no se identificó un umbral claramente seguro de exposición por debajo del cual no hubiera efecto adverso, lo que coincide con la postura de organismos como la OMS que advierte sobre los riesgos persistentes incluso en niveles bajos de contaminación [6].

Desde el punto de vista ambiental, el SO_2 también juega un papel importante en la acidificación atmosférica, este puede contribuir a la formación de lluvia ácida, alterar el pH de suelos y cuerpos de agua, y dañar la vegetación al deteriorar el follaje, reducir el crecimiento y disminuir la resistencia de los ecosistemas frente a otros factores de estrés [5]. La European Environment Agency (EEA) ha documentado que este contaminante figura entre los principales responsables de la acidificación y eutrofización de ecosistemas en Europa, con efectos extrapolables a zonas tropicales o de montaña [7]. De igual modo, las agencias de salud pública enfatizan los efectos indirectos que estos procesos tienen sobre la biodiversidad, la estabilidad de los ecosistemas acuáticos y la productividad agrícola [8].

Considerando estos múltiples impactos del SO_2, resulta esencial fortalecer las estrategias de monitoreo y predicción de este contaminante. Especialmente en regiones donde la presencia de fuentes industriales o de combustión fósil es significativa. La capacidad de anticipar deterioros en la calidad del aire y comunicar alertas tempranas permite proteger a poblaciones vulnerables, orientar políticas de salud pública, planificación ambiental y gestión territorial eficaz.

A nivel global, se estima que la contaminación del aire causa alrededor de 6.7 millones de muertes anuales, de las cuales cerca del 85% están relacionadas con enfermedades no transmisibles (ENT) como la enfermedad cardíaca isquémica, el accidente cerebrovascular, el cáncer de pulmón, el asma, la enfermedad pulmonar obstructiva crónica (EPOC) y la diabetes [9]. Esta situación convierte la contaminación del aire en la segunda causa principal de ENT a nivel mundial, después del consumo de tabaco. Debido a esta problemática, la OMS ha establecido pautas y límites para la concentración de contaminantes atmosféricos (Tabla 1) [3]. Sin embargo, la mayoría de las regiones del mundo aún enfrentan dificultades para mantener las concentraciones dentro de los valores recomendados, especialmente los países de ingresos bajos y medios, donde cerca del 99% de la población respira aire que excede los límites establecidos [1].

Tabla 1 Regulación concentración contaminantes del aire [3]

Contaminante	Tiempo	Niveles [$\mu g/m^3$]
$PM_{2.5}$	Anual	5
	24-horas	15
PM_{10}	Anual	15
	24-horas	45
O_3	Peak season	60
	8-horas	100
NO_2	Anual	10
	24-horas	25
SO_2	24-horas	40
CO	24-horas	4 mg/m^3

En el caso de Colombia, el Ministerio de Ambiente y Desarrollo Sostenible adoptó la Resolución 2254 de 2017 [10], mediante la cual se establecen los niveles máximos permisibles de calidad del aire para cada uno de los contaminantes criterio (Tabla 2). Esta normativa incorpora un esquema de cumplimiento progresivo que fija metas intermedias hasta el año 2030, año en el que el país deberá alcanzar los valores guía definidos por la OMS.

Dicho marco regulatorio representa un compromiso nacional con la reducción sostenida de la contaminación atmosférica y con la protección de la salud pública, especialmente en regiones urbanas e industriales donde las concentraciones de contaminantes superan con frecuencia los valores límite Resolución 2254 de 2017.

Tabla 2 Niveles máximos permitidos 2030 [10]

Contaminante	Niveles [$\mu g/m^3$]	Tiempo
PM_{10}	30	Anual
$PM_{2.5}$	15	Anual
NO_2	20	24 horas
SO_2	40	Anual

Considerando los impactos en la salud y el ambiente, así como las exigencias normativas nacionales e internacionales, se hace indispensable fortalecer las herramientas que permitan evaluar, anticipar y controlar la calidad del aire de manera oportuna y eficiente. En este marco, la predicción de la calidad del aire se consolida como una herramienta fundamental para la gestión urbana y ambiental. Gracias a los avances en sensores y tecnologías de análisis de datos, hoy es posible monitorear la calidad del aire en tiempo real, integrando información de tráfico, variables meteorológicas y condiciones urbanas. No obstante, el gran volumen y la complejidad de los datos plantean retos en su procesamiento y análisis, haciendo necesario el uso de técnicas avanzadas que permitan obtener predicciones más precisas y útiles para la toma de decisiones.

En este marco, el presente proyecto propone evaluar la capacidad de diferentes modelos de aprendizaje automático para predecir con precisión el comportamiento horario de las concentraciones de dióxido de azufre (SO_2) en el aire del Valle de Aburrá en horizontes de hasta 72 horas, a partir de datos históricos y variables meteorológicas, con el propósito de determinar su fiabilidad y utilidad en el análisis y la gestión de la contaminación atmosférica. Este enfoque busca fortalecer las capacidades locales de pronóstico, mejorar la planificación y la respuesta ante episodios críticos de contaminación, y contribuir al cumplimiento de los estándares nacionales e internacionales de calidad del aire establecidos por la Resolución 2254 de 2017 del Ministerio de Ambiente y Desarrollo Sostenible [10].

2 Estado del arte

El estudio y la predicción de la calidad del aire han incorporado, a lo largo de los años, una amplia variedad de enfoques metodológicos. En un inicio, los métodos estadísticos clásicos dominaron el panorama debido a su solidez teórica, su interpretabilidad y la disponibilidad limitada de datos. Sin embargo, conforme aumentó la cantidad y resolución de los registros históricos y se hizo evidente la necesidad de capturar relaciones no lineales entre las variables explicativas y las concentraciones de contaminantes, los modelos de aprendizaje automático comenzaron a emplearse de manera creciente para simular la variabilidad espaciotemporal de los contaminantes atmosféricos [11].

Dentro de los métodos tradicionales, uno de los enfoques más utilizados ha sido el modelo ARIMA, ampliamente reconocido por su capacidad para representar tendencias, patrones estacionales y componentes autoregresivos con relativa flexibilidad. Su popularidad se basa en su solidez estadística y la facilidad con la que puede ajustarse a diferentes procesos temporales [2]. motivo por el cual ha sido recurrentemente aplicado en estudios de calidad del aire. Un ejemplo destacado se encuentra en Surat, India [12], el modelo ARIMA demostró un desempeño destacado en la predicción de la calidad del aire, evidenciando un comportamiento superior frente a los distintos algoritmos de aprendizaje automático incluidos en el análisis .Para ello, se trabajó con registros históricos de los contaminantes criterio, recopilados desde enero de 2020 hasta abril de 2023, con el propósito de evaluar su comportamiento reciente y generar proyecciones confiables sobre su evolución futura. En el estudio se señala que las estimaciones generadas con ARIMA presentan un nivel elevado de precisión, lo que lo convierte en una herramienta adecuada para el análisis de datos de series de tiempo en este tipo de aplicaciones. Con el uso de este modelo, se logró predecir y pronosticar de manera efectiva los niveles de calidad del aire en la ciudad, cumpliendo el objetivo de anticipar la tendencia de los contaminantes y apoyar la toma de decisiones ambientales basadas en evidencia.

En un resultado similar se realizado en Bangladesh [13], se comparó el desempeño del modelo ARIMA frente a diferentes técnicas de aprendizaje automático, incluyendo evaluaciones detalladas de enfoques híbridos como ARIMA-ANN y ARIMA-SVM para la predicción de $PM_{2.5}$. El análisis se desarrolló a partir registros diarios recopilados entre enero de 2013 y mayo de 2019 en estaciones de Dhaka, Narayanganj y Gazipur. Acá el ARIMA permitió capturar de manera adecuada los patrones estacionales y la tendencia de fondo del contaminante, demostrando utilidad para anticipar su evolución temporal. Sin embargo, los autores señalan que su rendimiento fue inferior al de los modelos híbridos, destacando que el enfoque ARIMA-ANN incrementó la precisión en escenarios con alta variabilidad diaria, donde intervienen relaciones no lineales y multivariables. Esto indica que, aunque el ARIMA es eficaz para modelar la estructura temporal básica, su desempeño puede mejorarse notablemente cuando se integra con métodos más complejos.

A medida que se buscó superar las limitaciones de los modelos lineales, empezaron a emplearse métodos capaces de representar interacciones no lineales más complejas, entre ellos las máquinas de soporte vectorial en modo regresión (SVR). Estas técnicas, gracias al uso de funciones kernel, ofrecen una capacidad notable para aproximar relaciones no lineales sin perder garantías de generalización [14]. En un estudio publicado en la International Journal of Intelligence Science, desarrollado por un equipo de la Universidad Autónoma de Querétaro, México [14], aplicaron máquinas de

soporte vectorial (SVM) para modelar las concentraciones de O_3, NO_2 y PM_{10} en la Ciudad de México, aprovechando su capacidad para capturar relaciones no lineales a partir de datos diarios del año 2009. El trabajo destaca que la predicción de partículas contaminantes presenta un comportamiento dinámico no lineal; por lo tanto, su implementación no es un proceso trivial [14], lo que refuerza la pertinencia de recurrir a métodos capaces de representar estructuras complejas entre variables atmosféricas. Al comparar distintas funciones kernel, se observaron que algunas configuraciones permiten reproducir con mayor precisión la evolución de los contaminantes, mientras que otras, aunque menos costosas computacionalmente, tienden a mostrar un desempeño más variable. En conjunto, los resultados evidencian que las SVM pueden ajustarse de manera efectiva a la dinámica no lineal de los contaminantes, siempre que la selección del kernel se adapte al comportamiento específico de cada serie.

Esta versatilidad también fue observada en un estudio desarrollado en Malasia [15], donde se evaluó el desempeño de tres variantes de máquinas de soporte vectorial en regresión (SVR, linear SVR y libSVR) para predecir las concentraciones de contaminantes atmosféricos y estimar el Air Pollution Index (API) a partir de una serie diaria que abarcó 18 años de datos, desde 2002 hasta 2020. Los resultados mostraron que las SVM pueden aplicarse eficazmente incluso en series con distribuciones marcadamente asimétricas, como las de PM_{10} y $PM_{2.5}$, donde se observó ausencia de normalidad en los valores del contaminante. En la comparación entre modelos, linear SVR obtuvo los errores más bajos en la mayoría de los contaminantes, mientras que la variante SVR estándar presentó un ajuste superior en el caso de NO_2. El estudio concluye además que estas configuraciones permiten estimar el API con hasta tres días de anticipación, lo que las convierte en herramientas útiles para la predicción a corto plazo en contextos urbanos.

También, los métodos basados en árboles de decisión comenzaron a ganar relevancia por su capacidad para manejar grandes volúmenes de datos y capturar interacciones no lineales profundas. Random Forest, por ejemplo, ha sido ampliamente utilizado. Un estudio de revisión sistemática mostró que la mayoría de las publicaciones que utilizan Random Forest para predicción de $PM_{2.5}$ reportan coeficientes de determinación (R^2) mayores a 0.85, demostrando razonable confiabilidad y eficiencia en aplicaciones de predicción de contaminantes [16]. Su eficacia también fue demostrada en trabajo publicado en Atmospheric Environment [11] donde se desarrolló un modelo de predicción diaria de dióxido de azufre (SO_2) para toda China mediante un enfoque híbrido que integró Random Forest con un esquema de Kriging espacio temporal. El modelo empleó información satelital y variables geográficas para estimar las concentraciones del contaminante entre 2014 y 2015, obteniendo un rendimiento sólido con un R^2 de 0.62 y un RMSE de 10.36 microgramos por metro cúbico en una validación cruzada de diez pliegues. El uso de Random Forest permitió capturar relaciones no lineales entre los predictores y los niveles de SO_2, lo que demostró su capacidad para representar la variación espacio temporal del contaminante a escala nacional y servir como base para evaluaciones de riesgo en salud pública basadas en exposiciones diarias.

Por su parte, los algoritmos basados en boosting también han mostrado buenos desempeños en predicción de concentraciones de contaminantes. Diversos estudios demuestran el potencial de estos enfoques. Por el lado del XGBoost un estudio realizado en Shanghái desarrolló un modelo de predicción diaria de $PM_{2.5}$ utilizando XGBoost, integrando observaciones de contaminantes y variables meteorológicas con los pronósticos operativos del modelo numérico WRF-Chem. El enfoque mostró mejoras sustanciales frente al sistema tradicional, incrementando la correlación entre valores observados y estimados entre 50–100 % según el nivel de concentración, y reduciendo el error cuadrático medio en distintos rangos de contaminación. La combinación de información observacional y salidas del modelo químico permitió a XGBoost capturar patrones no lineales y corregir sesgos del pronóstico numérico, lo que derivó en estimaciones más estables, especialmente en episodios de alta concentración. En conjunto, los resultados respaldan el uso de XGBoost como un esquema eficaz para mejorar la predicción operativa de $PM_{2.5}$ en contextos urbanos complejos [17].

En este mismo ámbito, LightGBM ha destacado por su eficiencia computacional, su capacidad para manejar grandes volúmenes de datos y por ofrecer desempeños predictivos superiores. Un caso estudio desarrollado en China [18], donde se construyó un modelo nacional de predicción horaria de $PM_{2.5}$ a partir de más de 40 millones de observaciones

meteorológicas entre 2016 y 2019, logró un modelo altamente robusto con un desempeño notable: un R^2 de 0.80 y un RMSE de 19.80 µg/m³ en validación cruzada a escala horaria, que se incrementó hasta R^2 de 0.89 y 0.94 para escalas diaria y mensual, respectivamente. Incluso al predecir datos completamente no vistos de 2019, el modelo mantuvo un rendimiento sostenido con un R^2 de 0.75 en escala horaria y de 0.84 en escala diaria. Estos resultados demuestran la capacidad de LightGBM para capturar la variabilidad espaciotemporal del contaminante y generar predicciones estables en regiones con condiciones atmosféricas heterogéneas, consolidándose como una alternativa especialmente eficaz para la reconstrucción histórica y el pronóstico operacional de $PM_{2.5}$ a alta resolución temporal

La eficacia de los modelos impulsados por boosting también ha sido corroborada en el contexto colombiano. En un estudio realizado por el SIATA [4] se evaluó la capacidad de distintos modelos de aprendizaje automático para pronosticar concentraciones promedio de $PM_{2.5}$ con 24 horas de anticipación, integrando información de estaciones de calidad del aire, variables meteorológicas y datos satelitales. El trabajo comparó esquemas basados en Random Forest, Gradient Boosting y métodos lineales, encontrando que los modelos basados en árboles superaron ampliamente a la regresión lineal y que Gradient Boosting obtuvo los mejores resultados en la mayoría de las estaciones. Además, se observó una reducción sustancial del error respecto a los enfoques tradicionales y una mayor estabilidad en escenarios de alta variabilidad, lo que permitió anticipar episodios críticos con mayor confiabilidad.

En el ámbito de las redes neuronales recurrentes, los modelos LSTM (Long Short-Term Memory) han demostrado capacidad para aprender dependencias de largo plazo en series temporales ambientales. En Quintero, Chile, se implementó un modelo LSTM para pronosticar el máximo del promedio de las concentraciones de SO_2 para las primeras horas del siguiente día, obteniendo una precisión de 78 % comparado con un 52 % obtenido con un modelo de Random Forest [19]. Este modelo se alimentó de datos históricos de SO_2, velocidad y dirección del viento provenientes de varias estaciones de monitoreo del área industrial. La arquitectura incluyó capas convolucionales y LSTM en secuencia para extraer características espaciales y temporales, logrando una clasificación precisa de eventos críticos de contaminación por SO_2 en rangos de percentiles. Esta estrategia fue especialmente útil para anticipar episodios de concentración extrema durante las primeras horas del día, cuando los patrones de viento favorecían la acumulación de emisiones industriales.

De forma similar, en Taiwán se desarrolló un modelo LSTM para predecir concentraciones horarias de $PM_{2.5}$ utilizando variables meteorológicas y datos de estaciones cercanas. El modelo alcanzó coeficientes de correlación superiores a 0.90 para horizontes de predicción cortos, y mantuvo un rendimiento aceptable incluso en temporadas críticas como el invierno. Esto demuestra su capacidad para adaptarse a variaciones estacionales y condiciones locales complejas. Asimismo, en China, Jie et al. (2023) implementaron un modelo basado en LSTM con arquitectura autoregresiva (AR-LSTM) para predecir concentraciones de SO_2 en entornos urbanos industriales, logrando mejoras significativas en precisión frente a otros enfoques clásicos. Estos casos refuerzan la idoneidad de los modelos LSTM para capturar las dinámicas no lineales y dependencias temporales que caracterizan a los contaminantes atmosféricos en distintas regiones del mundo.

Además de la selección del modelo, otro elemento fundamental en el pronóstico de series de timepo es la estrategia empleada para generar predicciones a múltiples horizontes. La literatura reciente ha mostrado que el rendimiento no depende únicamente del algoritmo base, sino también del esquema mediante el cual se propagan las estimaciones hacia el futuro. En este sentido, se reconocen dos estrategias ampliamente utilizadas: el enfoque recursivo y el enfoque directo, cada uno con implicaciones distintas en términos de sesgo y varianza.

El enfoque recursivo entrena un único modelo para predecir un paso adelante y, posteriormente, reutiliza sus propias predicciones como entradas para generar los horizontes siguientes. Este esquema resulta eficiente en términos computacionales y mantiene coherencia estructural entre pasos sucesivos; sin embargo, presenta una limitación inherente: cuando el proceso generador de datos presenta componentes no lineales, los pronósticos recursivos se vuelven sistemá-

ticamente sesgados a partir del segundo horizonte debido a la curvatura de la dinámica interna. Se ha demostrado que los pronósticos recursivos sólo permanecen insesgados cuando el proceso subyacente es estrictamente lineal; en cualquier otro caso, el sesgo se acumula a medida que aumenta el horizonte [20]. Un estudio sobre predicción de $PM_{2.5}$ en doce estaciones de Beijing, se aplicó un modelo LSTM con estrategia recursiva para extender el horizonte de pronóstico hasta 24 horas, y se observó que a medida que aumenta el intervalo de predicción, el error se amplifica debido a la dependencia de predicciones previas y al carácter altamente no lineal del fenómeno [21]. En este caso de estudio, el uso de un esquema recursivo con un único modelo LSTM convencional produjo una degradación sustancial de la precisión en horizontes superiores a 12 horas, evidenciando que el error acumulado propio de las arquitecturas recursivas termina dominando la predicción cuando la dinámica es no lineal, como es habitual en procesos atmosféricos.

En contraste, el enfoque directo ajusta un modelo independiente para cada horizonte futuro. Al no depender de predicciones previas, evita la propagación de errores, dando lugar a estimaciones que pueden ser insesgadas si cada modelo posee suficiente flexibilidad. No obstante, esta independencia entre horizontes incrementa la varianza de las estimaciones, ya que cada modelo se entrena con un conjunto de datos efectivo reducido. Se ha documentado que el enfoque directo resulta especialmente ventajoso cuando el modelo subyacente presenta algún grado de mala especificación, por ejemplo, en presencia de autocorrelación residual o estructuras omitidas [20]. Un ejemplo de este comportamiento se aprecia en un estudio de pronóstico de $PM_{2.5}$ en Seúl, donde la aplicación de un esquema directo para horizontes de uno a cinco días evitó la acumulación progresiva de errores característica de los métodos iterativos [22]. En dicho análisis, la independencia entre horizontes permitió obtener predicciones más estables conforme aumentaba el plazo de pronóstico, incluso bajo relaciones no lineales y una marcada variabilidad estacional.

Ahora bien, los estudios también muestran que el rendimiento relativo entre estrategias no es uniforme. En un estudio aplicado al pronóstico de calidad del aire en la ciudad de Mashhad, Irán [23], en el cual se generaron predicciones diarias de $PM_{2.5}$ a 10 días utilizando estrategias recursivas, directas y variantes multi-salida, se observó que la estrategia recursiva, combinada con selección de características mediante LASSO en un modelo ARIMAX, alcanzó el mejor desempeño en la mayoría de los horizontes evaluados. Este resultado sugiere que, cuando la estructura temporal es capturada de forma adecuada por el modelo base, la recursividad puede ofrecer mayor estabilidad en horizontes extendidos. Además, el mismo estudio evidenció que, al emplear modelos no lineales o configuraciones híbridas, el comportamiento relativo entre las estrategias puede variar según el horizonte, mostrando en algunos casos un desempeño complementario entre los enfoques recursivo y directo [23].

Por lo tanto, la selección entre una estrategia recursiva o directa no es universal, sino que depende de la naturaleza del proceso, del horizonte de predicción y del tipo de modelo utilizado. En la práctica, ambas aproximaciones se consideran herramientas complementarias dentro de la metodológico para pronósticos multi-paso, especialmente en sistemas ambientales donde convergen dinámicas lineales, no lineales y múltiples fuentes de incertidumbre.

3 Metodología

Este proyecto se desarrolló utilizando CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual es una metodología robusta y probada para orientar trabajos de ciencia de datos [24]. Esta ofrece un enfoque estructurado y flexible que guía el desarrollo a través de seis fases iterativas las cuales se muestran en la Fig. 1

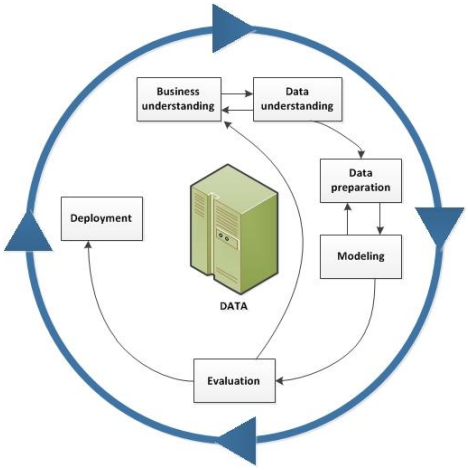

Fig. 1 Diagrama Crisp DM [24]

En la fase de Entendimiento del Negocio se centra en definir los objetivos y las restricciones del proyecto asegurando la alineación con las metas organizacionales. En la fase de Entendimiento de los Datos, se recopilan, exploran y analizan los datos disponibles para evaluar su calidad y relevancia con respecto al problema. La fase de Preparación de los Datos incluye la limpieza, transformación y preparación de los datos para optimizarlos en el proceso de modelado. Durante la fase de Modelado, se seleccionan y aplican algoritmos de aprendizaje automático o métodos estadísticos para desarrollar modelos alineados con los objetivos del proyecto. En la fase de Evaluación, se analiza el desempeño de los modelos para garantizar que cumplan con los criterios definidos y proporcionen información útil. Finalmente, la fase de Despliegue se centra en implementar los modelos en un entorno operativo y asegurar su accesibilidad para los usuarios finales.

Se eligió CRISP-DM debido a sus ventajas en flujos de trabajo de ciencia de datos. Su naturaleza iterativa permite ajustes y mejoras continuas, mientras que su flexibilidad facilita la adaptación a las particularidades que presentan los datos ambientales. Además, su enfoque estructurado facilita la transformación eficiente de datos en información valiosa, lo que lo hace ideal para proyectos de aplicación.

3.1 Entendimiento del negocio

El Sistema de Alerta Temprana del Valle de Aburrá (SIATA) es un componente clave dentro de la estrategia de gestión de riesgos del Área Metropolitana del Valle de Aburrá y la Alcaldía de Medellín, con el apoyo de EPM e ISAGEN. Según su descripción oficial, el SIATA alerta en tiempo real a los organismos de gestión de riesgos y a la comunidad frente a la posible ocurrencia de un fenómeno natural que pueda generar emergencias, monitoreando minuto a minuto las condiciones hidrometeorológicas de la región [25].

El proyecto utilizará los datos recolectados por el SIATA como base principal para desarrollar un modelo predictivo de la calidad del aire en el Valle de Aburrá, enfocado en los contaminantes criterio. Este modelo se alinea con el objetivo del SIATA de salvaguardar la vida de los habitantes de la región, ya que, sus resultados contribuirán a la toma de decisiones estratégicas y a la implementación de medidas que mejoren la capacidad de alerta temprana y mitiguen los riesgos asociados a la contaminación atmosférica, promoviendo así la salud pública y el bienestar regional.

En este contexto, la capacidad de anticipar el comportamiento futuro de los contaminantes se convierte en un elemento fundamental para el fortalecimiento operativo del SIATA. La predicción de la concentración de contaminantes en el aire permite anticipar episodios críticos de mala calidad del aire y generar alertas preventivas con la antelación necesaria para activar estrategias de mitigación. Contar con pronósticos confiables ofrece a las autoridades ambientales y de gestión del riesgo la posibilidad de planificar acciones oportunas, como la implementación de medidas temporales de restricción vehicular, la recomendación de reducción de actividades al aire libre o la comunicación temprana a la ciudadanía; con el fin de reducir la exposición de la población a niveles peligrosos de contaminación. Además, disponer de estimaciones confiables sobre la concentración de contaminantes, como el dióxido de azufre (SO_2), contribuye a fortalecer el seguimiento continuo de la calidad del aire y a evaluar el cumplimiento de los objetivos de reducción establecidos por la normativa ambiental vigente, en particular los compromisos adoptados para el año 2030 según la Resolución 2254 de 2017 del Ministerio de Ambiente y Desarrollo Sostenible. De este modo, el desarrollo de modelos predictivos se constituye como una herramienta estratégica que potencia las capacidades del SIATA para la gestión preventiva y la protección de la salud pública en el Valle de Aburrá.

3.2 Entendimiento de los datos

Los datos que se utilizarán para el desarrollo de los modelos corresponden a registros de cuatro estaciones ubicadas en el Valle de Aburrá, que incluyen mediciones horarias de la concentración en ug/m^3 de SO_2. Los datos abarcan el período comprendido entre enero de 2021 y febrero de 2025, con un total de 36480 registros por estación, lo que proporciona una base temporal suficiente para realizar un análisis exhaustivo de los patrones diarios, estacionales y meteorológicos asociados a la calidad del aire.

Para las variables meteorológicas, se cuenta con datos de temperatura, humedad relativa, presión atmosférica, velocidad y dirección del viento. En un principio se iba a trabajar de igual forma con los datos registrados por los sensores meteorológicos del SIATA en cada una de las estaciones; sin embargo, estos solo cubren desde enero de 2023. Por esta razón, para garantizar un conjunto de variables ambientales más completo, se optó por utilizar el conjunto de datos ERA5-Land. Este es un conjunto de datos que está disponible para uso público para el periodo comprendido entre 1950 y 5 días antes de la fecha actual que proporciona información horaria de alta resolución sobre variables de superficie [26].

Como primera etapa del procesamiento, se realizará una evaluación exhaustiva de la calidad de los datos, con especial énfasis en la detección, cuantificación y caracterización de los valores faltantes. Dado que los registros provienen de sensores automáticos, se puede anticipar la presencia de interrupciones en la medición, errores instrumentales, fallas de comunicación o condiciones ambientales extremas que podrían generar lagunas o inconsistencias temporales en las series. Por tanto, esta etapa buscará determinar la magnitud del problema y su impacto potencial sobre el análisis posterior.

En primer lugar, se calculará el porcentaje de datos ausentes para cada estación con el propósito de identificar como periodos prolongados de inactividad de los sensores o franjas horarias con mayor frecuencia de fallos en la medición. Este análisis permitirá dimensionar la magnitud del problema y reconocer posibles sesgos de disponibilidad asociados a la operación de las estaciones.

Posteriormente, se llevará a cabo un análisis de continuidad temporal, mediante el cual se evaluará la longitud de los intervalos consecutivos de ausencia de datos (gaps). Estos intervalos se clasificarán según su extensión con el fin de facilitar la posterior selección de estrategias de imputación adecuadas al tipo de vacío identificado. Este diagnóstico permitirá diferenciar el tratamiento que se dará a los vacíos de corta duración frente a aquellos más extensos.

Finalmente, se desarrollará un análisis de correlación general entre las variables meteorológicas y las concentraciones de SO_2, de manera que sea posible explorar las asociaciones lineales y no lineales más relevantes y comprender cómo las condiciones ambientales influyen en la dinámica del contaminante. Este conjunto de procedimientos permitirá obtener un diagnóstico sólido sobre la calidad, estructura y coherencia de los datos, sentando las bases para las etapas posteriores de imputación, depuración y modelado predictivo.

3.3 Preparación de los datos

La etapa de preparación de datos tendrá como objetivo garantizar la calidad, la coherencia temporal y la completitud de la información antes del modelado. Para ello, se realizará un proceso de limpieza de datos incluyendo detección y eliminación de valores atípicos y luego una imputación robusta de valores faltantes, aplicados de forma independiente en cada estación de monitoreo.

Detección y eliminación de outliers

Se comenzará con la detección de valores atípicos (outliers), ya que la presencia de estos puede distorsionar la distribución estadística de las variables, alterar las correlaciones entre los predictores y la variable objetivo, y comprometer la estabilidad y generalización de los modelos de predicción**.** Para el caso específico de datos obtenidos de sensores, los registros pueden verse afectados por errores instrumentales, interrupciones en la medición o fenómenos transitorios que no son representativos del comportamiento típico del sistema, que pueden derivar en observaciones atípicas que deben ser tratadas; por lo que es necesario manejar de una manera adecuada estos registros.

El proceso de detección se abordará desde una perspectiva mixta, combinando análisis univariado y multivariado, con el fin de identificar tanto anomalías individuales en la variable objetivo como patrones anómalos relacionados a las variables meteorológicas.

Detección univariada

En primer lugar, se aplicará una evaluación univariada sobre la variable objetivo. Este enfoque asume que las observaciones atípicas se manifiestan como desviaciones extremas respecto a la tendencia central de la distribución. Para su identificación se empleará el método de isolation forest, un método que es computacionalmente eficiente y ha demostrado ser muy efectivo en la detección de valores atípicos. [27]

Este método consiste en extraer submuestras aleatorias del conjunto de datos y procesarlas mediante una estructura de árboles de aislamiento construidos a partir de cortes aleatorios en los valores de características. Las muestras que requieren mayor número de divisiones tienden a ser observaciones comunes, mientras que aquellas que quedan aisladas en ramas más cortas se consideran más propensas a ser anómalas [27]. A partir de esto se emplea la longitud promedio del camino que va desde la raíz hasta la hoja del árbol para calcular una puntuación de anomalía para cada observación. Cuando la trayectoria es más corta, es decir, cuando el punto se aísla tras un menor número de divisiones, la puntuación de anomalía resulta más alta, lo que indica una mayor probabilidad de que dicho punto sea considerado anómalo.

Detección multivariada

Posteriormente, se implementará un enfoque multivariado orientado a identificar observaciones atípicas que, aunque no resulten anómalas al analizar cada variable de manera individual, muestran combinaciones inusuales dentro del espacio conjunto de las variables meteorológicas y de la variable objetivo. Este procedimiento se basa en el análisis de proyecciones de máxima curtosis, una técnica que permite resaltar aquellas direcciones del espacio de datos donde las observaciones extremas son más evidentes.

La detección de outliers mediante curtosis parte del principio de que los valores atípicos tienden a generar colas más pesadas y picos más agudos en la distribución. La curtosis mide el movimiento de la masa de probabilidad desde los hombros de una distribución hacia su centro y sus colas [28], reflejando tanto la concentración central como la frecuencia de valores extremos. En este enfoque, se buscan las proyecciones lineales que maximizan la curtosis, ya que dichas direcciones son las que mejor separan las observaciones anómalas del conjunto principal de los datos normales.

En el ámbito multivariado, esta metodología permite identificar combinaciones anómalas de variables que no serían evidentes al analizarlas individualmente. Este enfoque se basa en que, cuando los datos provienen de una mezcla de dos distribuciones normales con pesos muy diferentes e idénticas varianzas, la proyección que maximiza la curtosis coincide, salvo por cambios de ubicación y escala, con la proyección que mejor separa las medias de ambas distribuciones [28]. En términos prácticos, esto significa que la dirección en la que más se resaltan los valores extremos, es decir, la de mayor curtosis, coincide con la que mejor separa los datos normales de los atípicos. Por esta razón, el análisis de proyecciones de máxima curtosis se considera una técnica eficaz para detectar patrones inusuales y observar la estructura de los outliers en espacios de alta dimensionalidad.

A partir de este fundamento teórico, el procedimiento práctico consiste en aplicar dicha proyección a los datos con el fin de cuantificar el grado de anormalidad de cada observación. Este enfoque se basa en encontrar la dirección que maximiza la curtosis de una proyección lineal del vector de datos x tal como se expresa en la siguiente ecuación

$$\beta_{2M}(x) = max\beta_2(c^T x) \quad (1)$$

De manera práctica, esto se implementará generando múltiples proyecciones aleatorias y calculando la curtosis de cada una para identificar la que alcanza el valor máximo, correspondiente a la dirección de máxima curtosis. Una vez obtenida esta proyección óptima, se establecen umbrales basados en percentiles extremos (0.5 y 99.5) para clasificar las observaciones como normales o atípicas.

Imputación de valores faltantes.

La etapa de imputación de datos faltantes (gaps) se abordará mediante una combinación de técnicas determinísticas y modelos autorregresivos, con el propósito de preservar el comportamiento dinámico de la serie y evitar sesgos asociados a un rellenado arbitrario de los gaps. Esta fase constituirá un paso fundamental dentro de la preparación de la serie objetivo, ya que la continuidad y la coherencia temporal de los registros son condiciones esenciales para asegurar la estabilidad y la capacidad de generalización de los modelos predictivos.

Además, las etiquetas resultantes del proceso de detección de outliers, tanto del análisis univariado como del multivariado, serán tratadas como valores nulos, por lo que también deberán ser imputadas dentro de este mismo procedimiento. Esto permitirá mantener la integridad del conjunto de datos y garantizar que las observaciones anómalas no generen distorsiones en las etapas posteriores de modelado.

En primer lugar, se realizará una detección sistemática de los intervalos de datos faltantes presentes en los registros horarios de cada estación. A partir de este diagnóstico, se clasificará cada intervalo según su duración y frecuencia, con el fin de definir estrategias de imputación específicas para cada tipo de gap. Este análisis será clave para determinar

cómo abordar de forma diferenciada los vacíos cortos y los prolongados, asegurando que el proceso de reconstrucción mantenga la estructura temporal y la variabilidad inherente de la serie original.

Para los gaps de corta duración, de hasta aproximadamente 9 horas, se aplicará una interpolación lineal, bajo el supuesto de que en periodos breves la variación de la serie puede aproximarse de manera lineal sin alterar significativamente su comportamiento general. Esta técnica se restringirá a intervalos cortos con el objetivo de evitar distorsiones o sobreajustes en periodos más extensos.

Por su parte, los vacíos de mayor duración se imputarán mediante un modelo ARIMA, seleccionado debido a su capacidad para capturar dependencias internas, tendencias y patrones subyacentes del proceso generador de datos [29], lo que permitirá que la imputación de los estos conserve el comportamiento generador de la serie y mantenga la variabilidad natural de esta, garantizando que los valores reconstruidos no introduzcan sesgos que afecten el modelado. Para ello, el ARIMA será ajustado sobre un fragmento representativo de la serie, seleccionado en función de su extensión, completitud y estabilidad estadística, y se obtendrán sus coeficientes característicos: p, que representa el orden autoregresivo y cuantifica la influencia de los valores pasados de la serie; d, que indica el grado de diferenciación necesario para alcanzar la estacionariedad; y q, que corresponde al orden de la media móvil y describe la dependencia respecto a los errores pasados [29], de modo que estos reflejen adecuadamente la dinámica temporal de la serie y permitan reconstruirla de manera coherente con su comportamiento histórico.

Para identificar el fragmento representativo sobre el cual se ajustará el modelo se implementará un procedimiento sistemático. En este proceso se considerarán dos estrategias de selección: una estrategia simple, que priorizará el tramo más largo de datos continuos disponibles, y una estrategia híbrida, que ponderará varios criterios cuantitativos de selección, incluyendo la longitud relativa, la completitud del tramo, la densidad de datos válidos y la calidad del ajuste del modelo. De este modo, el fragmento elegido ofrecerá una base sólida para la calibración del modelo y contribuirá a mejorar la fiabilidad de las imputaciones generadas.

Una vez seleccionado el fragmento, se determinarán los parámetros y coeficientes asociados al modelo a partir este. Luego se aplica la diferenciación necesaria para alcanzar la estacionariedad, de acuerdo con el orden d obtenido del modelo. Sobre esta serie diferenciada se reconstruyen los residuos históricos, calculando en cada instante la diferencia entre el valor observado y el valor pronosticado por la combinación de las componentes autoregresiva y de medias móviles. Esta secuencia de errores permite estimar la media y la varianza del término de ruido, caracterizando la distribución estadística del componente aleatorio del proceso. Se asume que dichos errores siguen aproximadamente una distribución normal $\varepsilon_t \sim \mathcal{N}(\bar{\varepsilon}, \sigma_\varepsilon^2)$, donde ε_t representa el valor medio del residuo y σ_ε^2su varianza estimada.

Gracias a esto, cada nuevo término de error generado introducirá una variabilidad realista en la simulación, preservando la fluctuación natural de la serie y evitando que las imputaciones resulten artificialmente rígidas. Este aspecto será crucial para mantener la coherencia estadística de la serie reconstruida y asegurar su utilidad en el entrenamiento posterior de los modelos predictivos.

Con los parámetros del modelo ARIMA a partir del fragmento seleccionado, y la distribución del error, se lleva a cabo la imputación mediante una simulación paso a paso sobre cada intervalo de datos ausentes. Para cada gap, se toma como punto de partida la serie observada (incluyendo los tramos previamente imputados) hasta el instante inmediatamente anterior al inicio del gap. A partir de allí, cada nuevo valor faltante se genera utilizando la estructura ARIMA calibrada, la componente autoregresiva (AR) combina los valores pasados de la serie ponderados por los coeficientes ϕ_i, mientras que la componente de medias móviles (MA) incorpora la influencia de los errores pasados a través de los coeficientes θ_j [29]. Sobre la suma de ambas componentes se añade un nuevo término de ruido ε_t que será el error simulado a partir de la distribución $\mathcal{N}(\bar{\varepsilon}, \sigma_\varepsilon^2)$ estimada en la fase anterior. Cada valor simulado se incorpora de forma

secuencial a la serie, de modo que sirve como insumo para la imputación de los puntos siguientes dentro del mismo gap.

Este procedimiento permite reproducir la dinámica autoregresiva y estocástica del proceso, evitando imputaciones excesivamente suavizadas y preservando la variabilidad propia de la serie. En términos formales, cada nuevo valor imputado $y_t$se obtiene a partir de la siguiente expresión:

$$y_t = \sum_{i=1}^{p} \phi_i yt - i + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Donde ϕ_i y θ_j representan los coeficientes autoregresivos y de medias móviles respectivamente, y ε_t es el termino de error aleatorio al instante t.

Para mantener la consistencia física de los valores generados dado que las concentraciones de SO_2 no pueden asumir valores negativos, los datos imputados se restringirán a un rango mínimo definido por el menor valor observado en el fragmento base utilizado para el ajuste del modelo. Este control garantizará que los resultados conserven su plausibilidad ambiental y que las imputaciones no introduzcan valores incompatibles con las condiciones reales del fenómeno.

Finalmente, una vez simulados los valores en la escala diferenciada, se aplicará la operación inversa de integración con el fin de reconstruir la serie en su escala original. Este paso permitirá restablecer el nivel y la tendencia de la serie, asegurando que los valores imputados sean coherentes con el comportamiento histórico y que la estructura temporal global no se vea distorsionada.

A través de este proceso, cada intervalo de datos faltantes será reconstruido de manera probabilística y dinámica, aprovechando tanto la estructura temporal de la serie como las propiedades estadísticas del error histórico. En los casos donde se identifiquen gaps de gran extensión, se aplicará un esquema de imputación iterativa, en el cual cada fragmento prolongado será dividido en segmentos. En cada iteración, el modelo actualizará su estado interno con los valores recién imputados antes de continuar con el siguiente bloque. Este procedimiento permitirá minimizar la acumulación progresiva del error y mantener la estabilidad estadística del proceso de simulación, garantizando que las series reconstruidas conserven su coherencia temporal y su utilidad para las etapas de modelado predictivo.

3.4 Ingeniería de características

En la ingeniería de características se transformarán y enriquecerán las series originales para construir un conjunto de variables que describa de forma más completa el comportamiento temporal de la concentración de SO_2 y las condiciones meteorológicas asociadas. El propósito de esta etapa será generar variables derivadas que capten la estructura temporal y exógena del sistema, de modo que el modelo pueda identificar con mayor precisión tendencias locales, ciclos regulares y variaciones estacionales en diferentes escalas de tiempo. Durante esta etapa se buscará que las variables generadas mantengan coherencia con el tiempo y con el comportamiento físico real de los datos. Por ello, cada característica deberá derivarse únicamente de información disponible hasta el instante actual, evitando cualquier tipo de fuga de información (data leakage).

En primer lugar, se derivarán atributos temporales y astronómicos a partir del índice temporal, con el fin de incorporar explícitamente la estructura cíclica y estacional del comportamiento atmosférico. Se extraerán variables de calendario como mes, semana, día de la semana y hora del día, las cuales reflejan patrones regulares vinculados a actividades humanas y condiciones meteorológicas locales. Además, se incluirán variables solares calculadas a partir de la posición

geográfica del Valle de Aburrá, tales como la hora de amanecer, la hora de atardecer, la duración del día y una variable binaria que indique si el instante pertenece a horas de luz. Estas características permitirán capturar los efectos de radiación solar que influyen directamente en los procesos de dispersión y formación SO_2. Para preservar la naturaleza periódica de las variables temporales y solares, se aplicará una codificación cíclica mediante funciones seno y coseno, garantizando que valores adyacentes en el tiempo se representen como cercanos en el espacio de características y evitando discontinuidades artificiales en los límites de ciclo.

Posteriormente, se implementará un esquema de ventanas móviles (window features), con el propósito de capturar la memoria temporal de las series y proporcionar al modelo contexto sobre la tendencia local y la variabilidad reciente de las variables. Para cada variable se calcularán estadísticas móviles tales como la media, desviación estándar, mínimo, máximo y mediana, en distintos horizontes de tiempo. Las ventanas cortas permitirán capturar fluctuaciones rápidas y detectar incrementos o descensos repentinos en las concentraciones; las ventanas medias reflejarán la evolución diaria y los cambios entre los periodos diurnos y nocturnos; mientras que las ventanas largas representarán la variación de fondo o el comportamiento promedio a lo largo del tiempo. Este enfoque, aplicado tanto a la variable objetivo como a las variables meteorológicas, permitirá modelar dependencias cruzadas, como incrementos en SO_2 asociados a reducciones recientes de la velocidad del viento. Cada una de estas estadísticas se calculará de manera causal, utilizando exclusivamente valores pasados o presentes, garantizando la validez temporal de la información.

También, se incluirán variaciones temporales (deltas), calculadas como la diferencia entre el valor actual y los valores pasados de cada variable en distintos desfases temporales. Estos indicadores permiten describir la dirección y magnitud de los cambios recientes tanto de las series, ayudando a distinguir entre incrementos sostenidos y descensos marcados. La combinación de estos deltas con las estadísticas obtenidas mediante ventanas móviles permitirá al modelo reconocer con mayor precisión los episodios de cambio abrupto y los periodos de estabilidad.

Para representar de manera más precisa la estructura temporal de las series, se aplicará una descomposición estacional mediante STL (Seasonal-Trend Decomposition using Loess) sobre la serie del SO_2 y cada componente se agregará como una nueva variable. Esta técnica permite separar la señal original en tres componentes: tendencia, estacionalidad y residuo [30], lo que facilita analizar y modelar de forma independiente los patrones que actúan en distintas escalas temporales. En este contexto, la componente estacional tiene un papel central, ya que refleja los ciclos regulares vinculados a las variaciones diarias, semanales y estacionales que influyen directamente en la dinámica del contaminante. La tendencia resume la evolución de largo plazo, mostrando aumentos o disminuciones sostenidas en los niveles del contaminante, mientras que el residuo agrupa las fluctuaciones irregulares no explicadas por los componentes anteriores y resulta útil para identificar episodios anómalos o eventos de contaminación no periódicos.

Finalmente, se generarán interacciones polinomiales de segundo orden exclusivamente entre las variables originales del conjunto (la concentración de SO_2 y las variables meteorológicas). Estas combinaciones permitirán capturar relaciones no lineales y efectos conjuntos entre predictores, sin aumentar de forma excesiva la complejidad del modelo. Para reducir el riesgo de sobreajuste, no se considerarán interacciones sobre variables derivadas ni expansiones automáticas del espacio de características. Los valores faltantes que resulten de estas transformaciones se completarán mediante propagación temporal hacia adelante y hacia atrás, preservando la coherencia del conjunto de datos.

3.5 Selección de características

La selección de características se llevará a cabo con el propósito de reducir la dimensionalidad del conjunto de variables generadas durante la fase de ingeniería de características, conservando únicamente aquellas con mayor capacidad explicativa sobre la variabilidad de la concentración de SO_2. Este proceso permitirá eliminar redundancias, disminuir el riesgo de sobreajuste y mejorar la capacidad de generalización de los modelos de predicción.

El objetivo principal de esta etapa será identificar las variables más influyentes en la predicción de las concentraciones de SO_2, diferenciando tres grupos principales: los lags de la variable objetivo, las window features derivadas de estadísticas móviles y las variables exógenas. La selección buscará alcanzar un equilibrio complejidad y capacidad descriptiva, conservando el menor número posible de variables sin comprometer el desempeño del modelo. Este balance es especialmente importante en problemas de predicción ambiental, donde las altas correlaciones entre predictores meteorológicos pueden introducir redundancia y afectar la estabilidad del modelo.

Para ello, se implementará un enfoque sistemático de selección basado en criterios estadísticos y de desempeño predictivo. En particular, se aplicará el método de regularización lineal Lasso, de forma independiente a los datos de cada estación. Este enfoque permitirá identificar, dentro del conjunto de variables generadas en la etapa de ingeniería de características, aquellas con mayor capacidad explicativa sobre la variabilidad del contaminante (SO_2).

El método Lasso se fundamenta en una penalización L_1 que reduce a cero los coeficientes asociados a variables con baja relevancia, promoviendo modelos más simples y menos sensibles al ruido [31]. En el contexto de la selección de características, esta regularización actuará como un filtro orientado al desempeño del modelo. Para cada estación, las variables serán previamente estandarizadas y se explorará un rango de valores del parámetro de penalización (α) en escala logarítmica, aplicando validación cruzada con particiones temporales para determinar su valor óptimo. Una vez identificado dicho valor, se conservarán únicamente las variables cuyos coeficientes presenten valores absolutos superiores a un umbral mínimo ($|\beta| > 10^{-7}$), criterio que permitirá descartar predictores con escasa contribución. En los casos en que el número de variables seleccionadas resulte inferior al umbral mínimo establecido, se mantendrán las top-N equivalentes aproximadamente al 50 % del total de predictores con mayores coeficientes absolutos, garantizando así un nivel mínimo de representatividad del conjunto de variables.

Una vez obtenidos los conjuntos de características por estación, se analizará su grado de coincidencia y consistencia. Dado que se busca desarrollar un modelo capaz de generalizar su desempeño en diferentes ubicaciones, se seleccionará un conjunto único de variables comunes para todas las estaciones. Esta decisión permitirá reducir la dependencia de patrones locales y favorecer la comparabilidad y estabilidad del modelo en contextos atmosféricos variados.

El conjunto final de características se definirá considerando aquellas variables que presenten mayor frecuencia de selección y coherencia interpretativa entre estaciones, asegurando un equilibrio entre simplicidad, generalidad y capacidad predictiva. Estas variables constituirán la base definitiva para la etapa de modelado, proporcionando un conjunto compacto, consistente y representativo para la predicción.

3.6 Modelado y selección de hiperparámetros

En esta etapa se desarrollarán y evaluarán diversos modelos de predicción con el propósito de estimar las concentraciones horarias de dióxido de azufre (SO_2) en horizontes de hasta 72 horas. Se implementarán tanto modelos tradicionales basados en árboles de decisión como modelos neuronales capaces de capturar dependencias temporales más complejas, complementados con una línea base de comparación que servirá como punto de referencia mínimo para evaluar la ganancia real de los enfoques propuestos.

Los modelos basados en árboles incluirán LightGBM y XGBoost. El primero se caracteriza por su eficiencia y capacidad para manejar grandes volúmenes de datos mediante un esquema de gradient boosting optimizado [32], mientras que XGBoost implementa un sistema escalable de árboles potenciados mediante gradientes destinado a construir modelos robustos a partir de múltiples aprendices débiles [33]. Ambos enfoques permiten capturar relaciones no lineales y mejorar progresivamente el desempeño predictivo conforme se agregan nuevos árboles al ensamble.

De manera complementaria, se incorporará un modelo de tipo LSTM (Long Short-Term Memory), una red neuronal recurrente diseñada para aprender patrones temporales y dependencias de largo plazo en series cronológicas. Su arquitectura, basada en compuertas de memoria, permite controlar el flujo de información y mitigar problemas comunes como el desvanecimiento del gradiente, lo que la hace especialmente útil para modelar dinámicas no lineales y relaciones prolongadas entre observaciones sucesivas [34]. Este modelo se configurará bajo un esquema de salida directa (sequence-to-one), de modo que cada horizonte de predicción se entrene de manera independiente.

Como punto de comparación, se utilizará un modelo base (baseline) que se abordará desde un enfoque de desfase, en el cual la predicción para un determinado horizonte h corresponderá al valor observado h horas antes. Este método, aunque simple, constituye una referencia fundamental para medir el aporte real de los modelos propuestos, ya que representa el límite inferior del desempeño esperado en un sistema predictivo.

Para abordar el problema del pronóstico multi-paso se adoptarán dos estrategias: el enfoque recursivo y el enfoque directo. En el enfoque recursivo, se entrenará un único modelo para predecir un paso hacia adelante y, posteriormente, sus propias predicciones serán utilizadas como insumo para los pasos subsiguientes. Este método permite aprovechar la dinámica temporal del sistema, aunque puede acumular errores conforme aumenta el horizonte. En contraste, el enfoque directo consistirá en entrenar un modelo independiente para cada horizonte de predicción, de modo que cada uno aprenda las relaciones específicas entre las variables y la concentración futura correspondiente. Aunque este enfoque demanda mayor esfuerzo computacional, evita la propagación de errores entre pasos y proporciona una visión más clara del comportamiento del modelo a diferentes horizontes [20].

La evaluación de los modelos se fundamentará en esquemas de validación específicamente diseñados para series de tiempo, los cuales permiten reproducir de manera controlada las condiciones reales bajo las que operaría un sistema predictivo. A diferencia de la validación tradicional empleada en problemas independientes e idénticamente distribuidos (i.i.d.), en el análisis temporal resulta indispensable respetar el orden cronológico de los datos, evitando que la información futura influya en la estimación del modelo. Por ello, se adoptarán metodologías que simulan explícitamente el flujo natural del tiempo, evaluando la capacidad del modelo para generalizar hacia observaciones no vistas.

Primero se empleará un esquema de validación por ventanas temporales progresivas. En este procedimiento, la serie se divide en múltiples particiones ordenadas en el tiempo, cada iteración utiliza un bloque inicial de datos para el entrenamiento y evalúa el modelo sobre el bloque inmediatamente posterior. A medida que se avanza en los pliegues, la ventana de entrenamiento se expande y la ventana de validación se desplaza hacia adelante, permitiendo analizar cómo se comporta el modelo en diferentes periodos históricos y bajo diversos regímenes de variabilidad. Este enfoque resulta especialmente útil para capturar cambios estructurales, estacionalidades y patrones no estacionarios que pueden afectar la estabilidad del modelo.

Además, se implementará un procedimiento de backtesting por ventanas móviles, en el cual se generan predicciones sucesivas sobre múltiples tramos futuros de la serie. Este método evalúa el rendimiento del modelo en condiciones operativas reales, para cada instante de evaluación, el modelo se ajusta únicamente con datos disponibles hasta ese momento y luego produce pronósticos para el horizonte requerido. La comparación entre las predicciones y los valores observados permite calcular métricas de desempeño para cada ventana, lo que facilita estudiar la consistencia temporal del modelo, identificar periodos de degradación y estimar su comportamiento frente a distintos escenarios.

La optimización de hiperparámetros se llevará a cabo mediante la herramienta Optuna, un optimizador secuencial basado en técnicas de búsqueda bayesiana que ofrece una alternativa más eficiente frente a los métodos tradicionales de grid search o random search. A diferencia de estos últimos, que exploran el espacio de parámetros de manera exhaustiva o aleatoria, Optuna utiliza el algoritmo Tree-structured Parzen Estimator (TPE) para modelar la función de desempeño y seleccionar de forma inteligente las combinaciones más prometedoras, priorizando aquellas regiones del espacio que

históricamente han mostrado mejores resultados. Además, incorpora mecanismos de pruning, que interrumpen de manera anticipada las configuraciones poco prometedoras, reduciendo significativamente el tiempo de cómputo sin sacrificar calidad en la búsqueda. Con ello, será posible identificar de manera eficiente el conjunto de hiperparámetros que minimice el error de predicción en cada caso.

Las mallas de parámetros a explorar estarán conformadas por un conjunto amplio de configuraciones representativas para cada modelo, abarcando tanto parámetros estructurales como de regularización. Estas combinaciones constituyen el espacio inicial dentro del cual Optuna seleccionará y ajustará las alternativas más prometedoras durante el proceso de optimización.

Tabla 3Malla de hiperparámetros GXBoost

Categoría	**Hiperparámetro**	**Valores**
Estructura del modelo	n_estimators	300, 600, 1000, 1500
	learning_rate	0.10, 0.05, 0.03, 0.02
	num_leaves	31, 63, 127, 255
	max_depth	-1, 10, 15, 20
Criterios de división	min_child_samples	10, 20, 50, 100
	min_split_gain	0.0, 0.1, 0.3
Muestreo	feature_fraction	0.6, 0.8, 1.0
	bagging_fraction	0.6, 0.8, 1.0
	bagging_freq	0, 1, 5
Regularización	lambda_l1	0.0, 0.1, 1.0, 5.0
	lambda_l2	0.0, 0.1, 1.0, 5.0
Otros	extra_trees	True, False

Tabla 4 Malla de hiperparámetros LigthGBM

Categoría	**Hiperparámetro**	**Valores**
Estructura del modelo	n_estimators	300, 600, 1000, 1500
	learning_rate	0.10, 0.05, 0.03, 0.02
	max_depth	3, 5, 7, 10
	min_child_weight	1, 3, 5, 10
Muestreo	subsample	0.6, 0.8, 1.0
	colsample_bytree	0.6, 0.8, 1.0
	colsample_bylevel	0.6, 0.8, 1.0
Penalización / Regularización	reg_alpha	0.0, 0.1, 1.0, 10.0
	reg_lambda	0.1, 1.0, 10.0
	gamma	0.0, 0.1, 0.3, 1.0
Método de crecimiento	tree_method	"hist"
	grow_policy	"depthwise", "lossguide"

Tabla 5Malla de hiperparámetros LSTM

Categoría	Hiperparámetro	Valores
Arquitectura recurrente	recurrent_units	[128, 64], [256, 128], [64, 32]
Capa densa	dense_units	[64, 32], [128, 64], [32, 16]
Regularización	dropout	0.0, 0.1, 0.2, 0.3
Optimización	learning_rate	0.01, 0.005, 0.001
Entrenamiento	epochs	5, 10, 20
	batch_size	64, 128, 256
Otros	recurrent_dropout	0.0, 0.1

Para cada estación, se entrenarán 72 modelos independientes, uno para cada horizonte horario comprendido entre 1 y 72 horas. Este procedimiento permitirá analizar cómo evoluciona el error de predicción a medida que se incrementa el horizonte, identificando hasta qué punto las estimaciones mantienen un nivel de precisión aceptable y a partir de qué momento las predicciones dejan de ser rentables desde una perspectiva práctica. Este análisis de degradación progresiva del desempeño resultará fundamental para determinar la ventana temporal óptima de pronóstico, es decir, el número máximo de horas hacia el futuro en el cual el modelo conserva su capacidad predictiva con un grado de fiabilidad suficiente para su uso en sistemas de alerta temprana.

3.7 Evaluación

El desempeño de los modelos se evaluará mediante un conjunto de métricas cuantitativas diseñadas para capturar distintos aspectos del ajuste predictivo, garantizando una valoración integral de la precisión, estabilidad y robustez de las estimaciones. La métrica principal será el Error Porcentual Absoluto Medio Ponderado (WMAPE), debido a su capacidad para expresar el error relativo respecto a los niveles observados de la serie y su mayor estabilidad en escenarios con valores bajos de concentración. A diferencia del MAPE tradicional, cuyo denominador depende del valor puntual observado y puede generar distorsiones cuando las concentraciones tienden a cero, el WMAPE pondera las desviaciones por la suma total de los valores reales, evitando explosiones numéricas y proporcionando una medida más consistente del desempeño. Esta característica resulta especialmente relevante en series de SO_2, donde los niveles pueden presentar variaciones abruptas y periodos con valores reducidos, propios de contaminantes de baja concentración y fuerte dependencia meteorológica.

Complementariamente, se calculará el Error Absoluto Medio (MAE), que permite cuantificar la magnitud promedio de las desviaciones en las unidades originales del contaminante ($\mu g/m^3$). Esta métrica resulta fundamental para evaluar la utilidad práctica de las predicciones, dado que errores absolutos elevados pueden traducirse en interpretaciones equivocadas frente a umbrales regulatorios o decisiones operativas relacionadas con la gestión de episodios de contaminación. El MAE aportará así una referencia directa sobre el nivel de precisión que es esperable de cada modelo desde una perspectiva aplicada.

Asimismo, se empleará el Error Cuadrático Medio (RMSE) como medida complementaria de la dispersión del error. A diferencia del MAE, el RMSE penaliza de forma más severa las desviaciones grandes, otorgando mayor peso a los eventos donde el modelo presenta fallas significativas. Esto es especialmente relevante en el análisis de contaminación atmosférica, donde los picos de concentración suelen tener un impacto sanitario y operativo mayor que los valores de fondo. El RMSE permitirá identificar si los modelos son capaces de anticipar estos episodios críticos o si, por el contrario, tienden a subestimarlos, comprometiendo su utilidad en un sistema de alerta temprana.

El uso conjunto de estas métricas permitirá obtener una evaluación equilibrada del desempeño: mientras el WMAPE ofrecerá una estimación relativa estable y comparable entre horizontes y estaciones, el MAE permitirá interpretaciones directas en unidades ambientales y el RMSE identificará posibles debilidades frente a eventos extremos. Esta combinación garantizará que la valoración de los modelos no se limite a un único aspecto del error, sino que considere tanto la precisión promedio como la capacidad de generalización y la respuesta frente a variaciones abruptas.

4 Resultados

Como se mencionó en la sección de metodología, se comenzará por realizar un análisis de entendimiento de los datos. En la Fig. 2 se muestra el comportamiento a través del tiempo de las concentraciones horarias de SO_2 para cada una de las estaciones entre enero de 2021 y marzo de 2025. En todas ellas se observan interrupciones visibles en la continuidad de la serie, evidenciadas como tramos sin registro que corresponden a períodos de inactividad del sensor o fallas en la adquisición de los datos. Este patrón de vacíos aparece de manera recurrente en todo el periodo de estudio y constituye un aspecto relevante a considerar en las etapas de imputación y depuración. Además, es importante destacar que en las estaciones MED-FISC e ITA-CJUS la serie no inicia en 2021 sino en 2023, por lo que la ausencia de información previa no debe interpretarse como un gap, sino como una falta estructural de cobertura histórica vinculada al inicio efectivo de operación o disponibilidad del sensor.

En términos de la dinámica observada, las cuatro estaciones exhiben oscilaciones intensas y una variabilidad marcada entre periodos de baja y alta concentración. Adema se observa que la estación MED-FISC muestra un comportamiento particularmente distintivo, presenta picos que alcanzan valores cercanos a 400 $\mu g/m^3$, notablemente superiores a los demás registros de esa estación y los de las demás estaciones. Estos aumentos abruptos, claramente visibles en la serie, podrían corresponder a episodios reales de alta acumulación, pero también podrían representar mediciones atípicas.

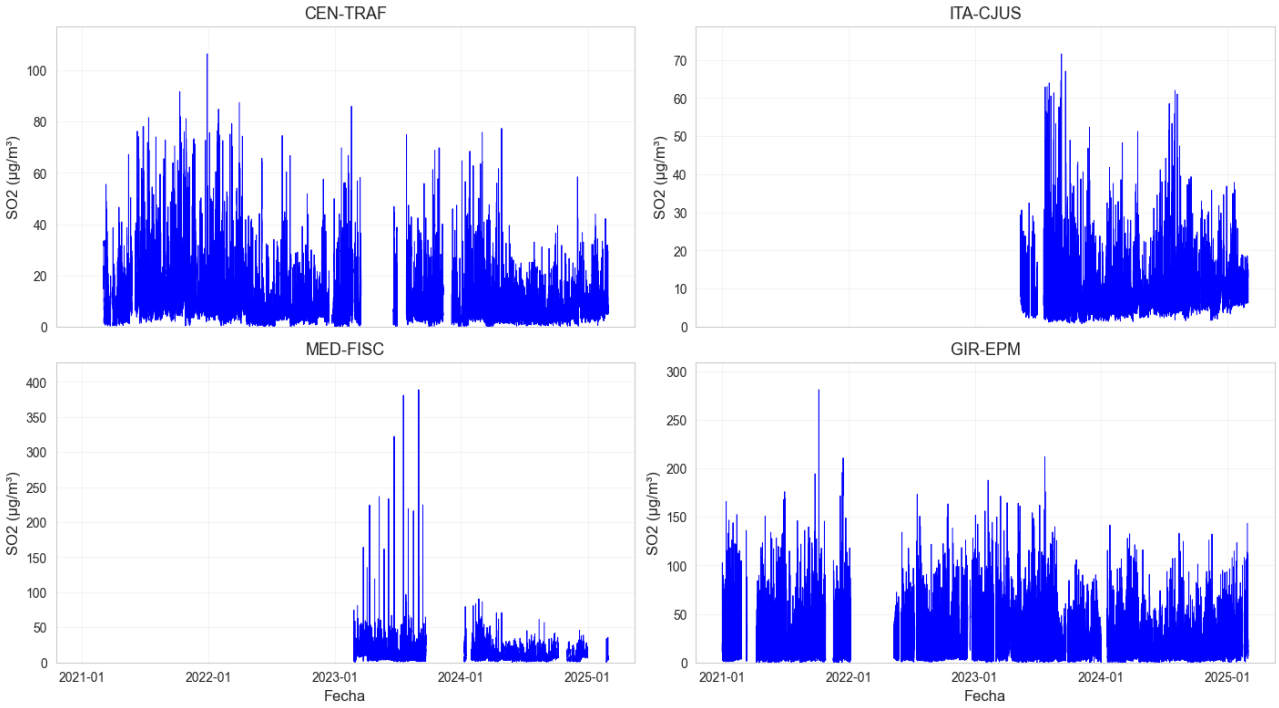

Fig. 2 Concentración de SO_2 en el tiempo por estación

La Tabla 6 resume la disponibilidad de datos de SO_2 para cada una de las estaciones, incluyendo la fecha inicial de registro, el número total de observaciones y la cantidad absoluta de valores faltantes. En esta se observa que CEN-TRAF y GIR-EPM cuentan con series extensas que comienzan en 2021, mientras que ITA-CJUS y MED-FISC disponen de registros únicamente desde 2023. En cuanto a la completitud de las mediciones, CEN-TRAF y GIR-EPM presentan volúmenes totales de datos significativamente mayores, pero también concentran una proporción elevada de valores ausentes, con 18 % y 21 %, respectivamente. Por su parte, ITA-CJUS registra el menor porcentaje de datos faltantes (7 %), mientras que MED-FISC alcanza un 33 %, el valor más alto entre las estaciones consideradas.

Tabla 6 Datos SO_2 por estación

Estación	Inicio datos	Datos totales	Datos nulos	Porcentaje datos nulos
CEN TRAF	2021-03-04	34992	6256	18 %
GIR EPM	2021-01-01	36480	7552	21 %
ITA CJUS	2023-05-12	15793	1105	7 %
MED FISC	2023-02-24	17652	5739	33 %

Profundizando en la caracterización de los faltantes, la Tabla 7 muestra el ranking de gaps consecutivos en cada una de las estaciones. En esta se observa que las interrupciones más extensas se concentran en CEN-TRAF, GIR-EPM y MED-FISC, donde se registran lapsos continuos de 2235, 2988 y 2667 horas sin datos, respectivamente. El caso de GIR-EPM representa el mayor intervalo individual del conjunto, mientras que en MED-FISC ese único tramo equivale a cerca del 46 % de todos sus faltantes (5739). En contraste, ITA-CJUS, pese a presentar el nivel más bajo de nulos en términos globales, también evidencia un vacío significativo de 432 horas, que por sí solo constituye aproximadamente el 39 % de sus datos ausentes (1105). Este patrón indica que, en estaciones como ITA-CJUS y MED-FISC, los faltantes no se distribuyen de forma dispersa, sino que se concentran en un episodio de gran extensión.

Tabla 7 Longitud de gaps consecutivos

Ranking	CEN TRAF	GIR EPM	ITA CJUST	MED FISC
1	2235	2988	432	2667
2	1488	614	147	1297
3	617	524	73	554
4	578	432	72	349
5	194	312	52	97

Las variables meteorológicas utilizadas en el análisis se presentan en la Fig. 3, donde se observa su evolución temporal para cada una de las estaciones. A diferencia del comportamiento registrado en la serie de SO_2, estas variables no presentan valores faltantes, lo que evidencia una mayor estabilidad y continuidad en la adquisición de los datos. Sus patrones exhiben dinámicas acordes con la variabilidad esperada de cada magnitud física de cada variable. No obstante, la variable de presión atmosférica muestra un comportamiento atípico en las estaciones CEN-TRAF y MED-FISC, caracterizado por un desplazamiento abrupto en el nivel medio de la serie. Este tipo de salto podría deberse a justes instrumentales, recalibraciones o cambios en la configuración del sistema satelital, más que a una variación física repentina.

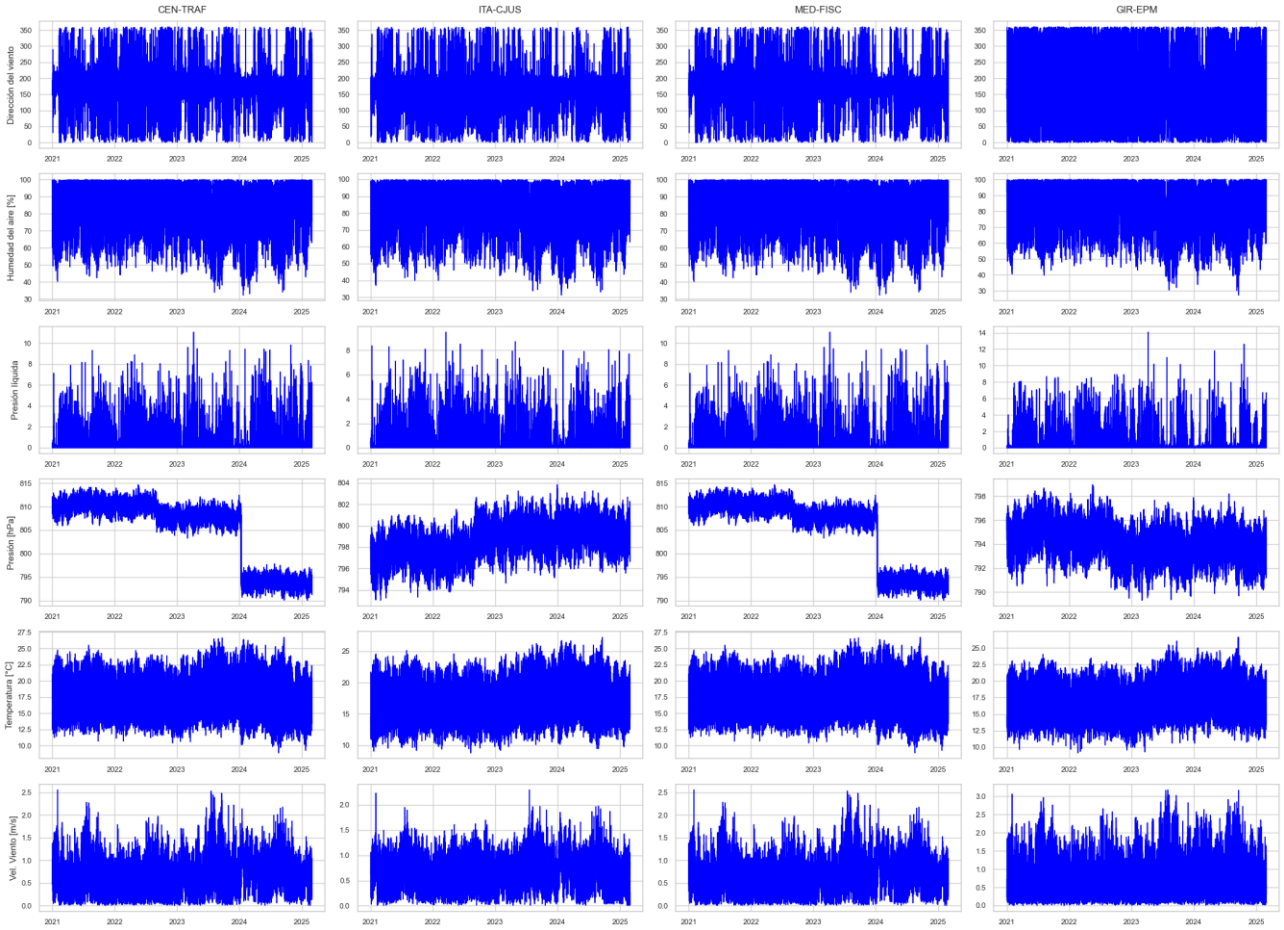

Fig. 3 Variables meteorológicas

La matriz de correlación Fig. 4 permite examinar la relación lineal entre las variables meteorológicas y las concentraciones de SO_2 en cada estación. En términos generales, no se identifican asociaciones fuertes entre las variables exógenas y el SO_2, lo que indica que, al menos de manera lineal, los factores meteorológicos incluidos no explican de forma directa la variabilidad del contaminante. Las correlaciones con SO_2 se mantienen en valores bajos y estables en todas las estaciones, sin superar magnitudes cercanas a 0.20. Por otro lado, se observan patrones internos más notorios entre algunas variables ambientales, como la correlación negativa consistente entre humedad relativa y temperatura, típica de climas urbanos con ciclos diarios pronunciados, así como correlaciones moderadas entre pares de variables asociadas a la dinámica atmosférica como lo son presión y temperatura o velocidad del viento y humedad. Estos resultados indican que, aunque las variables meteorológicas aportan información relevante sobre el estado general de la atmósfera, su influencia directa sobre las concentraciones de SO_2 parece limitada, por lo que su aporte al modelo podría manifestarse más claramente a través de efectos no lineales o interacciones que no se captan mediante una correlación simple.

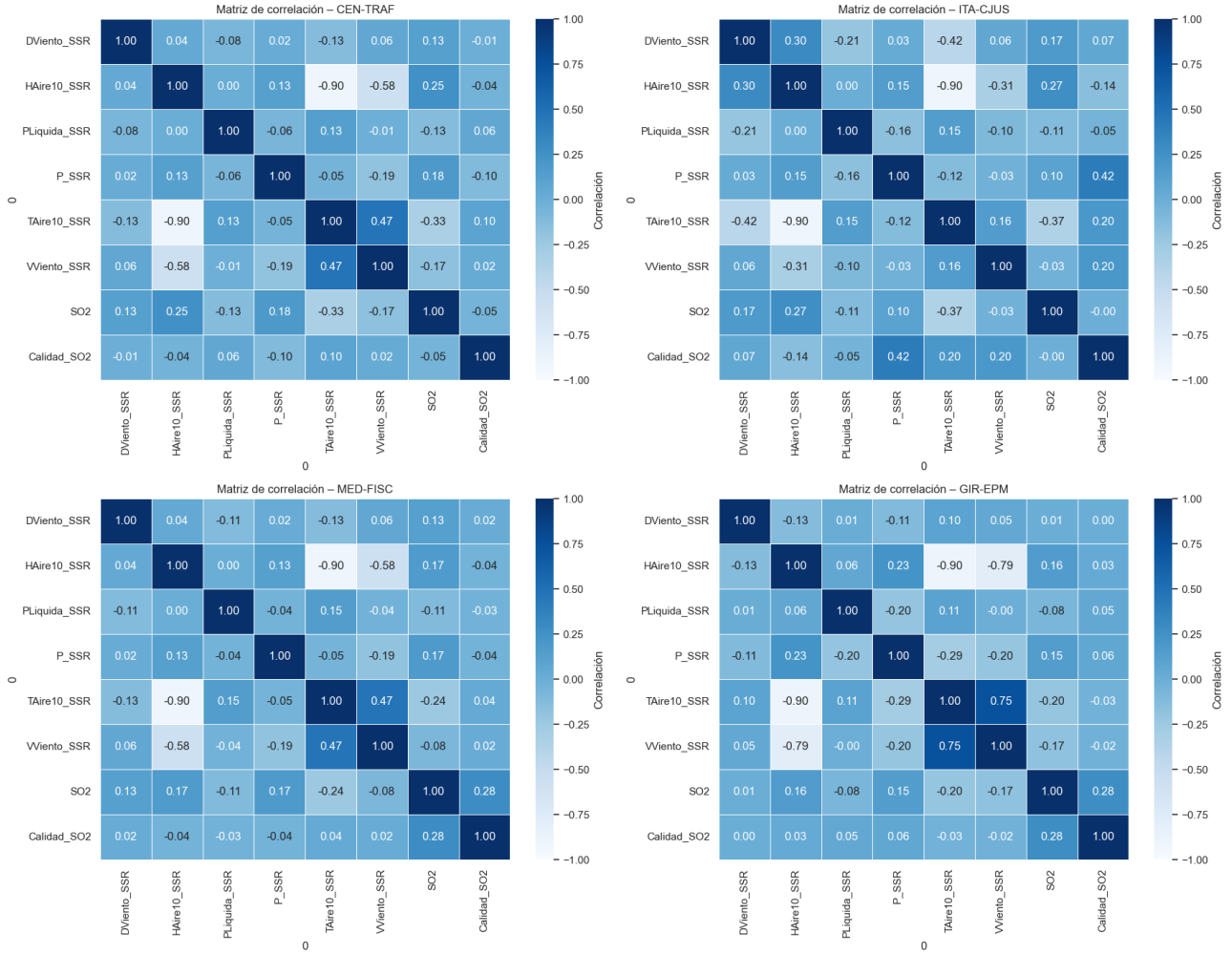

Fig. 4 Matriz de correlación

Las Fig. 5y Fig. 6, muestran los resultados de la detección de valores atípicos mediante los enfoques univariado y multivariado. En la inspección univariada, los outliers se identifican como picos aislados que se elevan de manera notable frente al comportamiento dominante de la serie de SO_2, reflejando episodios puntuales de concentraciones inusualmente altas. Estos puntos extremos aparecen de forma esporádica pero claramente diferenciada del resto de las observaciones, lo que sugiere la presencia de eventos atípicos bien delimitados a lo largo del tiempo. En el análisis multivariado (la Fig. 9 presenta únicamente la gráfica frente a la temperatura como un ejemplo representativo), se observa la misma tendencia general, donde los valores anómalos tienden a agruparse en la región donde el SO_2 alcanza niveles elevados, separándose visualmente de la nube principal de puntos y conformando un conjunto compacto de observaciones extremas.

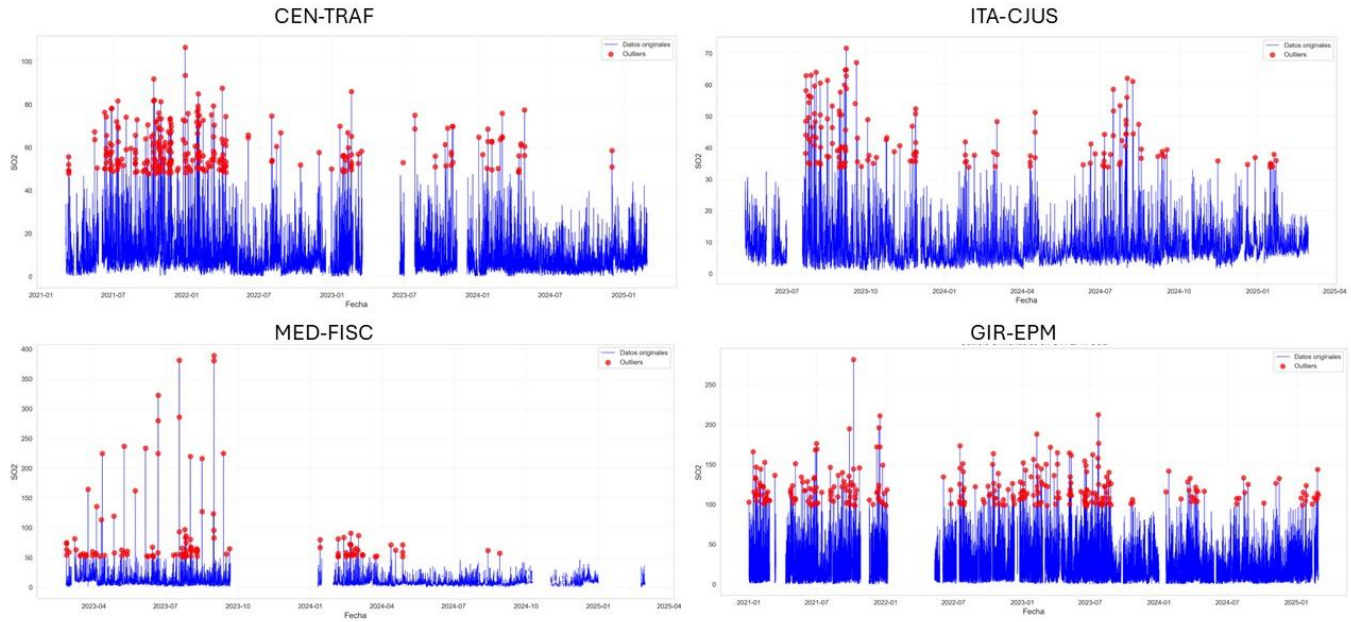

Fig. 5 Detección univariable de outliers

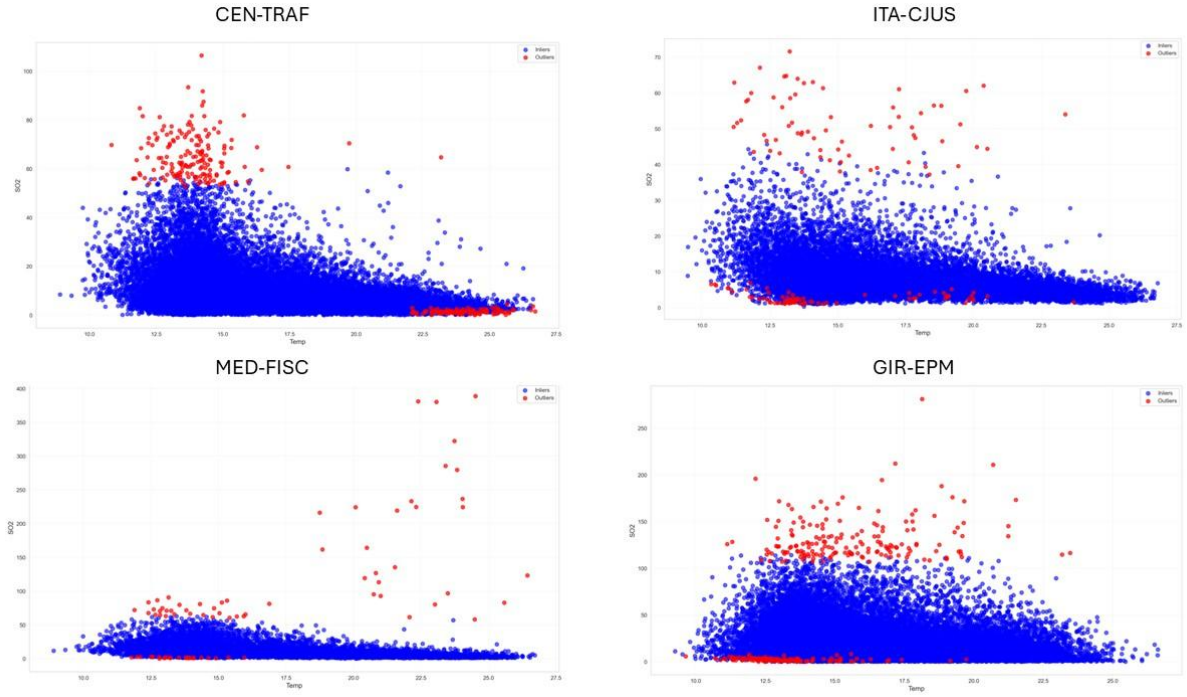

Fig. 6 Detección multivariable de outliers

Los resultados del proceso de imputación aplicado a las series de SO_2, se presentan en la Fig. 7. Para cada estación se muestran, en azul, los valores originales disponibles y, en rojo, los datos imputados que reemplazan los vacíos identificados previamente. Acá se puede observar que la serie resultante conserva la estructura temporal y la variabilidad general de los datos observados, mientras que los valores imputados se integran de manera suave dentro del patrón predominante de cada estación, sin introducir saltos abruptos ni inconsistencias visibles.

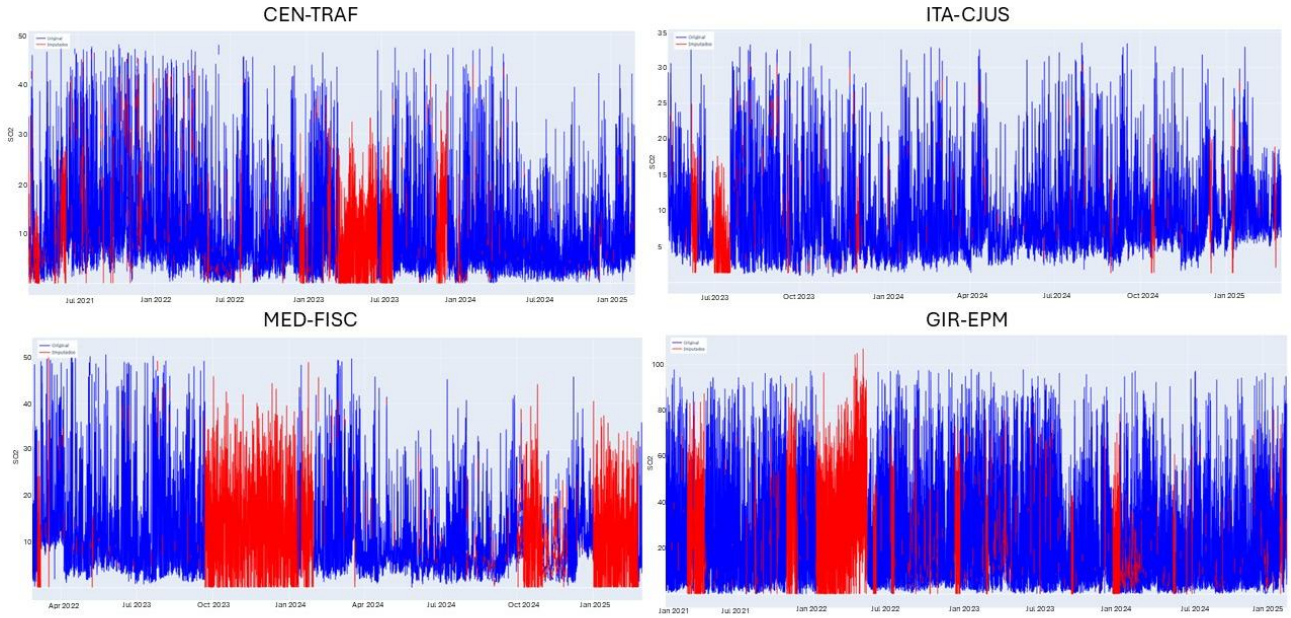

Fig. 7 Imputación datos faltantes

Los resultados de los mejores modelos para cada estación, obtenidos después del proceso de búsqueda de hiperparámetros, se presentan en las Fig. 8 y Fig. 9, donde se observa la evolución del error a lo largo del horizonte de predicción y la tabla 6 contiene el promedio de las métricas para todos los pasos de predicción. Este análisis permite identificar cómo se comportan los modelos cuando la distancia temporal del pronóstico aumenta. En ambas métricas, WMAPE y RMSE, se aprecia un incremento progresivo del error en los pasos más alejados, un comportamiento esperable dada la alta variabilidad del SO_2 y la pérdida gradual de información relevante conforme crece el horizonte.

Los modelos recursivos muestran un desempeño consistente, comienzan con errores considerablemente más bajos que el baseline en los primeros pasos, pero a medida que avanzan los horizontes sus curvas revelan un aumento continuo del error. Esto se explica por la naturaleza acumulativa del método recursivo, en el que cada predicción se alimenta de la estimación previa, amplificando pequeñas desviaciones que se vuelven más notorias en los horizontes lejanos. A pesar de esta acumulación, LightGBM y XGBoost recursivos mantienen un comportamiento más estable que el baseline y logran mejoras visibles a lo largo de la mayor parte del horizonte.

En contraste, los modelos directos ofrecen el comportamiento más favorable dentro del conjunto de enfoques evaluados. Tanto LightGBM como XGBoost en su versión directa sostienen valores bajos de error a lo largo de todo el horizonte y presentan curvas más suaves y homogéneas. La independencia entre los modelos diseñados para cada step evita la propagación de errores que caracteriza a los métodos recursivos y permite que el incremento del error con el horizonte sea menos abrupto. Este patrón se repite en todas las estaciones y convierte a ambos modelos directos en los que sistemáticamente se mantienen por debajo de las otras alternativas.

El modelo LSTM exhibe un comportamiento más irregular. Aunque en los primeros pasos ofrece resultados competitivos y en ocasiones similares a los modelos directos, su estabilidad disminuye en horizontes intermedios y largos. Las curvas muestran oscilaciones pronunciadas y picos esporádicos que indican una mayor sensibilidad frente a la complejidad de las series de contaminantes y a la extensión del horizonte de predicción. Esto sugiere que, pese a su capacidad

para capturar relaciones no lineales, su desempeño se vuelve más errático a medida que se alejan los pasos de pronóstico.

Al comparar entre estaciones, además de las tendencias generales, se aprecian diferencias notables en las magnitudes absolutas del error, las cuales responden a las particularidades propias que adopta la serie en cada estación. En GIR-EPM se identifican los valores de error más altos del conjunto, tanto en WMAPE como en RMSE. Aunque la jerarquía entre modelos se mantiene y los enfoques directos continúan siendo los más precisos, las métricas alcanzan magnitudes considerablemente mayores que en las demás estaciones, lo que indica un entorno mucho más complejo y variable donde los modelos no logran desempeños tan competitivos. En cambio, en MED-FISC la situación es distinta. Las líneas de todos los modelos aparecen mucho más próximas entre sí y las diferencias absolutas son menores. Aunque las tendencias relativas no cambian, la brecha entre los modelos avanzados y el baseline se reduce, lo que sugiere una serie menos volátil y más uniforme, en la que incluso los métodos simples logran aproximaciones razonables y los beneficios de modelos más sofisticados son menos marcados.

CEN-TRAF e ITA-CJUS presentan comportamientos intermedios. En ambas estaciones, los modelos directos mantienen una ventaja sostenida sobre los demás, los recursivos muestran un deterioro gradual conforme avanza el horizonte y LSTM conserva una estabilidad relativa que posteriormente se vuelve fluctuante. Sin embargo, en estas estaciones las magnitudes de error son más moderadas que en GIR-EPM y las diferencias entre modelos están más claramente definidas que en MED-FISC, lo que sugiere condiciones predictivas de dificultad media donde los modelos más potentes expresan ventajas de forma visible.

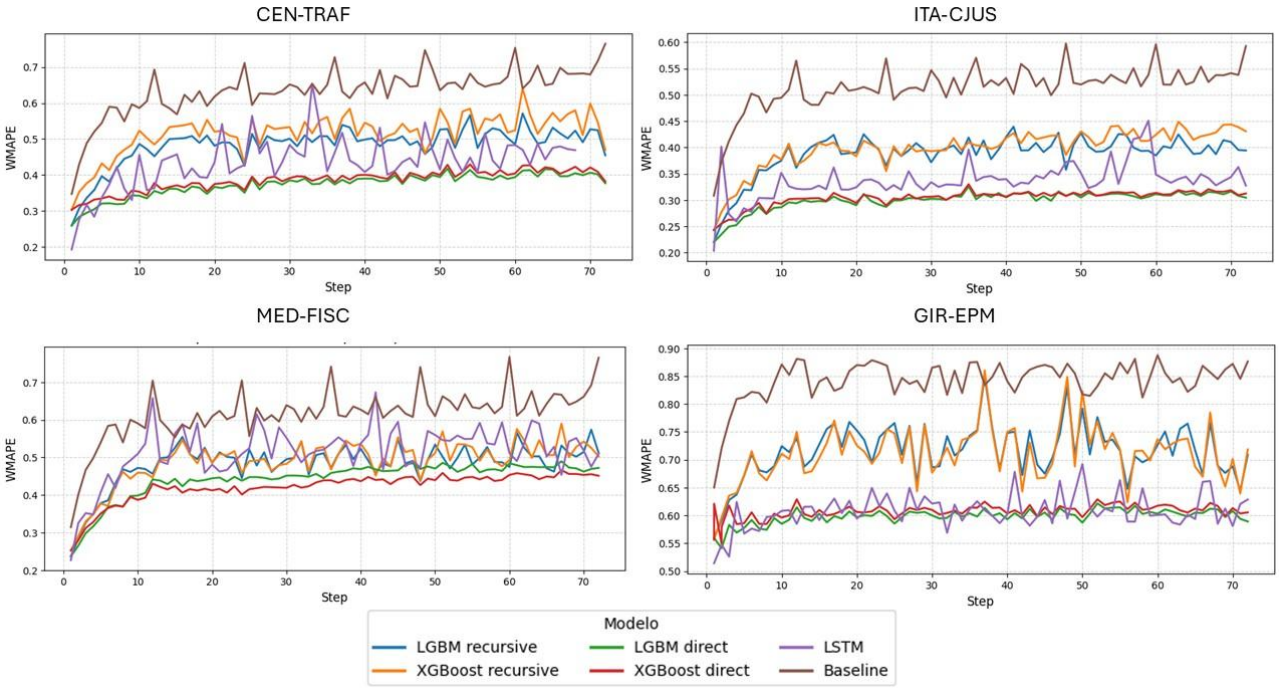

Fig. 8 Evolución del WMAPE por horizonte de predicción en validación

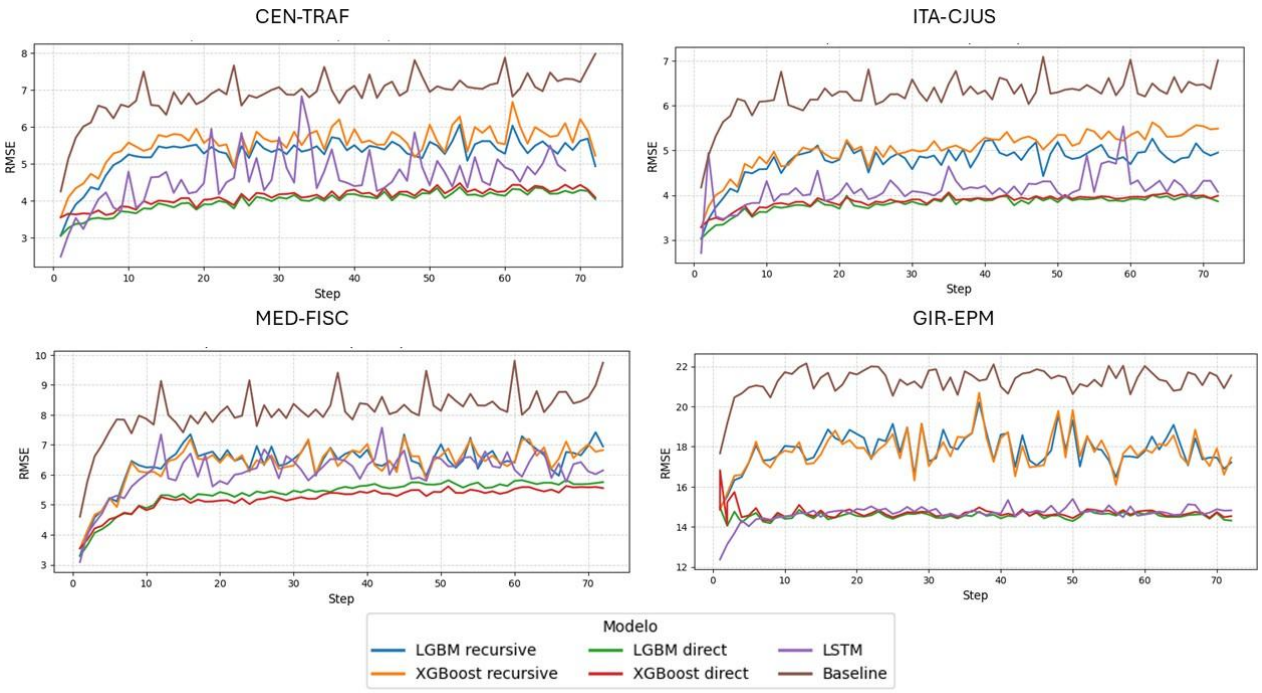

Fig. 9 Evolución del RMSE por horizonte de predicción en validacion

La Tabla 8 presenta los valores promedio del error en todo el horizonte de predicción para cada estación y modelo. En WMAPE se observa que LGBM-direct es el modelo que obtiene los valores más bajos en la mayoría de los casos, con resultados que van desde 0.29 en ITA-CJUS hasta 0.59 en GIR-EPM. En contraste, el baseline muestra valores considerablemente más altos, entre 0.51 y 0.84, de modo que la reducción relativa lograda por LGBM-direct frente a esta referencia varía según la estación, aproximadamente entre 43 % y 30 %. El modelo XGBoost-direct presenta valores ligeramente superiores, entre 0.39 y 0.71, manteniendo aun así mejoras claras frente al baseline. Los modelos recursivos se ubican en un nivel intermedio, mientras que el LSTM exhibe un comportamiento más variable según la estación, con valores entre 0.48 y 0.62.

Los valores de RMSE muestran una tendencia similar el modelo LGBM-direct alcanza errores promedio que oscilan entre 3.8 y 5.3, dependiendo de la estación, claramente por debajo del baseline, cuyos valores varían entre 6.3 y 21.2. Esto se traduce en reducciones que oscilan aproximadamente entre 40 % en las estaciones donde el baseline presenta errores moderados y hasta alrededor de 75 % en aquellas donde el baseline tiene errores particularmente altos. El modelo XGBoost-direct mantiene un buen comportamiento, aunque con RMSE ligeramente superiores (entre 5.1 y 6.4). Los modelos recursivos muestran valores intermedios, mientras que el LSTM conserva errores entre 6.1 y 8.1, reflejando su comportamiento más inestable.

En conjunto, LGBM-direct es el modelo con mejor desempeño general, este obtiene los menores valores promedio tanto en WMAPE como en RMSE en tres de las cuatro estaciones y presenta reducciones claras frente al baseline, con magnitudes que varían, pero se mantienen sustanciales en todos los casos. Su consistencia y su capacidad para adaptarse a diferencias entre estaciones lo consolidan como la opción más sólida para la predicción multi-horizonte de SO_2.

Tabla 8 comparación de modelos en el conjunto de validación

Modelo	CEN-TRAF		GIR-EPM		ITA-CJUS		MED FISC	
	WMAPE	RMSE	WMAPE	RMSE	WMAPE	RMSE	WMAPE	RMSE
Baseline	0.63	6.9	0.84	21.2	0.51	6.3	0.62	8.1
LGBM recursive	0.48	5.3	0.71	17.9	0.38	4.8	0.48	6.4
Xgboost recursive	0.52	5.6	0.71	17.8	0.39	5.1	0.49	6.4
LGBM direct	0.37	3.9	0.59	14.6	0.29	3.8	0.44	5.3
Xgboost direct	0.39	4.1	0.61	14.7	0.31	3.9	0.42	5.1
LSTM	0.44	4.6	0.61	14.6	0.33	4.1	0.52	6.1

Dado que el modelo LGBM-direct fue el que obtuvo el mejor desempeño durante la validación, se seleccionó como el modelo final para la evaluación en el conjunto de prueba. Las Figuras Fig. 10 y Fig. 11 muestran la evolución del WMAPE y del RMSE para cada horizonte de predicción utilizando únicamente este modelo, lo que permite analizar su comportamiento frente a datos completamente nuevos. En estas curvas se observa un patrón muy similar al registrado durante la validación, el error aumenta gradualmente a medida que se amplía el horizonte, pero se mantiene en niveles claramente inferiores a los del baseline en todos los horizontes y en todas las estaciones. Además, la estabilidad del LGBM-direct se mantiene, con curvas suaves, poca dispersión entre pasos consecutivos y una tendencia general que reproduce fielmente su desempeño previo, lo que sugiere que el modelo generaliza adecuadamente y conserva su capacidad predictiva incluso fuera del conjunto de entrenamiento y validación.

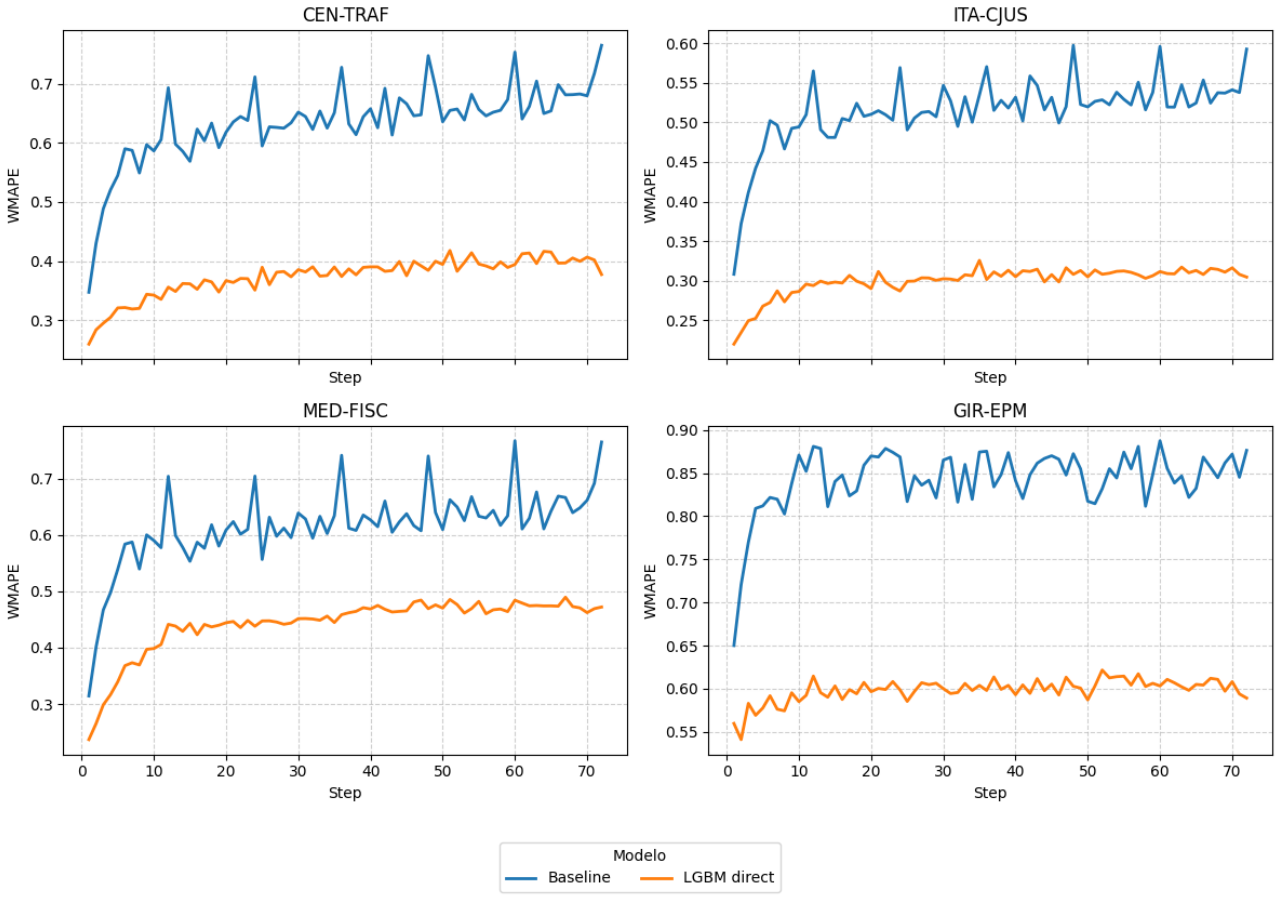

Fig. 10 Evolución del WMAPE por horizonte de predicción en test

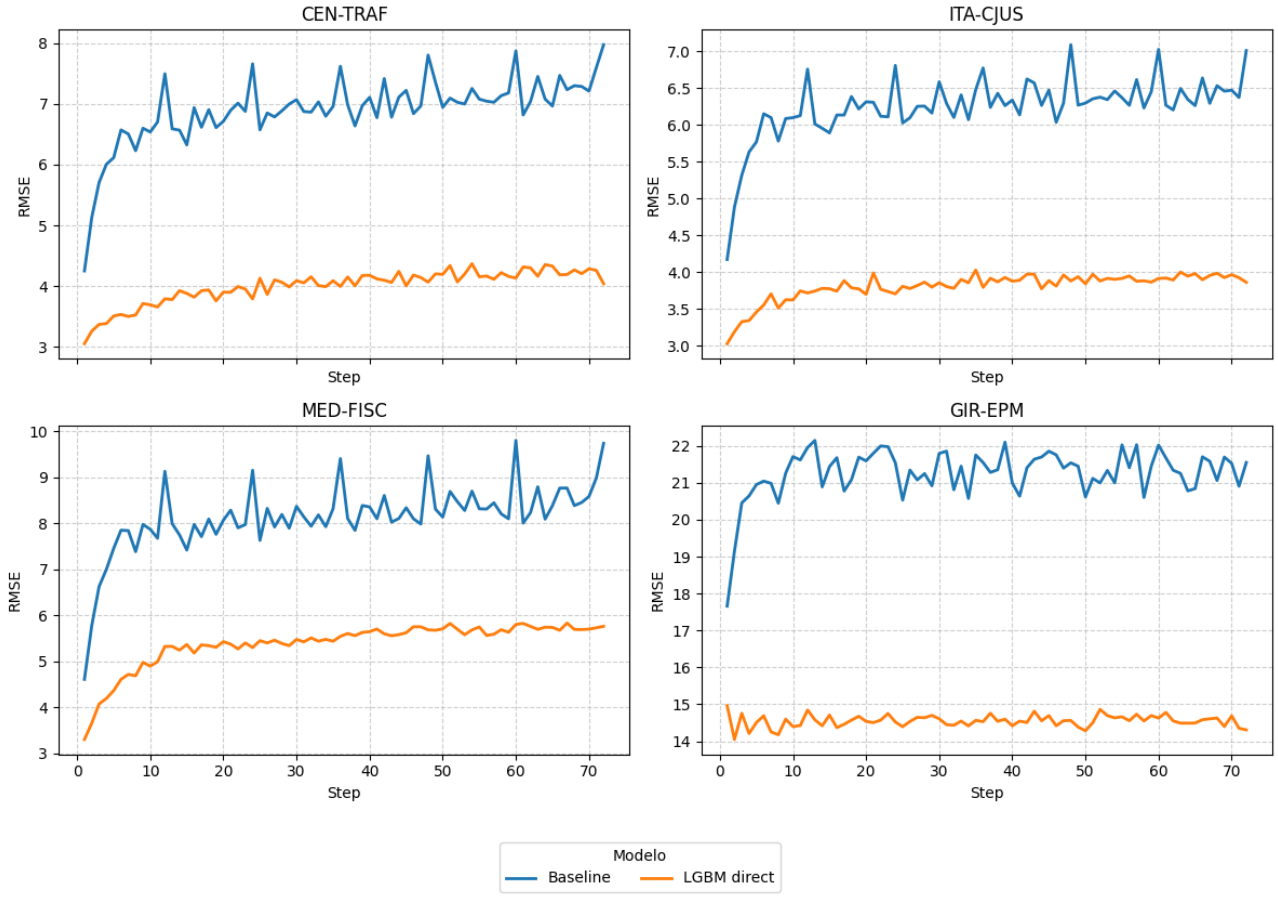

Fig. 11 Evolución del RMSE por horizonte de predicción en test

Las Fig. 12, Fig. 13, Fig. 14 muestran el desempeño del modelo LGBM-direct en el conjunto de prueba para horizontes diferentes horizontes a cercano, mediano y largo plazo (6, 24 y 72 horas, respectivamente), permitiendo analizar cómo evoluciona su capacidad de ajuste a medida que aumenta la distancia temporal del pronóstico. En los horizontes más cortos (6 horas), el modelo reproduce con gran precisión tanto la forma general de la serie como la variación horaria del contaminante; las líneas de predicción siguen de cerca las oscilaciones reales y capturan adecuadamente los picos y descensos más pronunciados, especialmente aquellos de carácter recurrente dentro del ciclo diario. A 24 horas, la correspondencia entre valores reales y predichos sigue siendo visible, aunque las predicciones se suavizan ligeramente, el modelo mantiene una buena capacidad para anticipar la secuencia relativa de aumentos y disminuciones, preservando la estructura temporal de la serie. En los horizontes más largos (72 horas), la predicción tiende a alisarse con mayor intensidad, lo que se traduce en una menor amplitud respecto a los valores reales; aun así, el modelo conserva la dirección general de los cambios y reproduce la tendencia de fondo, aunque pierde parte de la capacidad para ajustarse a los picos más altos que caracterizan a algunas estaciones. En general se observa que el modelo mantiene un nivel de ajuste coherente y estable incluso en horizontes extendidos, este logra capturar el comportamiento global de la serie, respetando la dinámica temporal del SO_2, mientras que la pérdida gradual de sensibilidad a los máximos y mínimos extremos es consistente con el comportamiento esperado en pronósticos de mayor alcance.

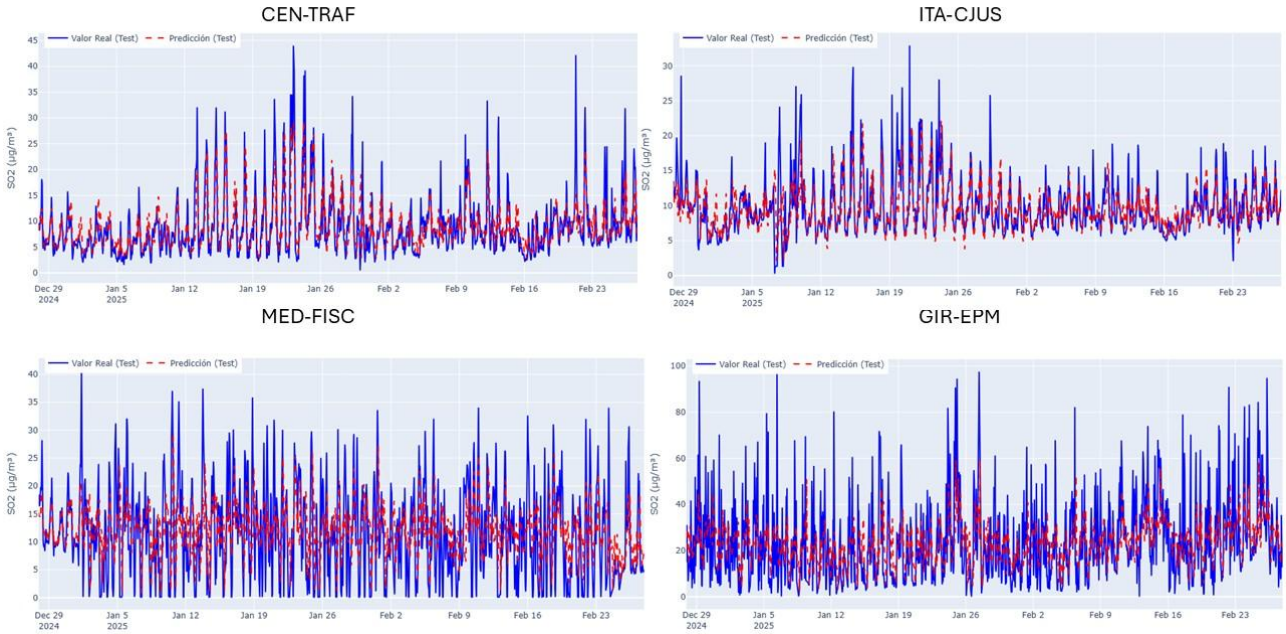

Fig. 12 Desempeño en test horizonte de predicción 6 horas

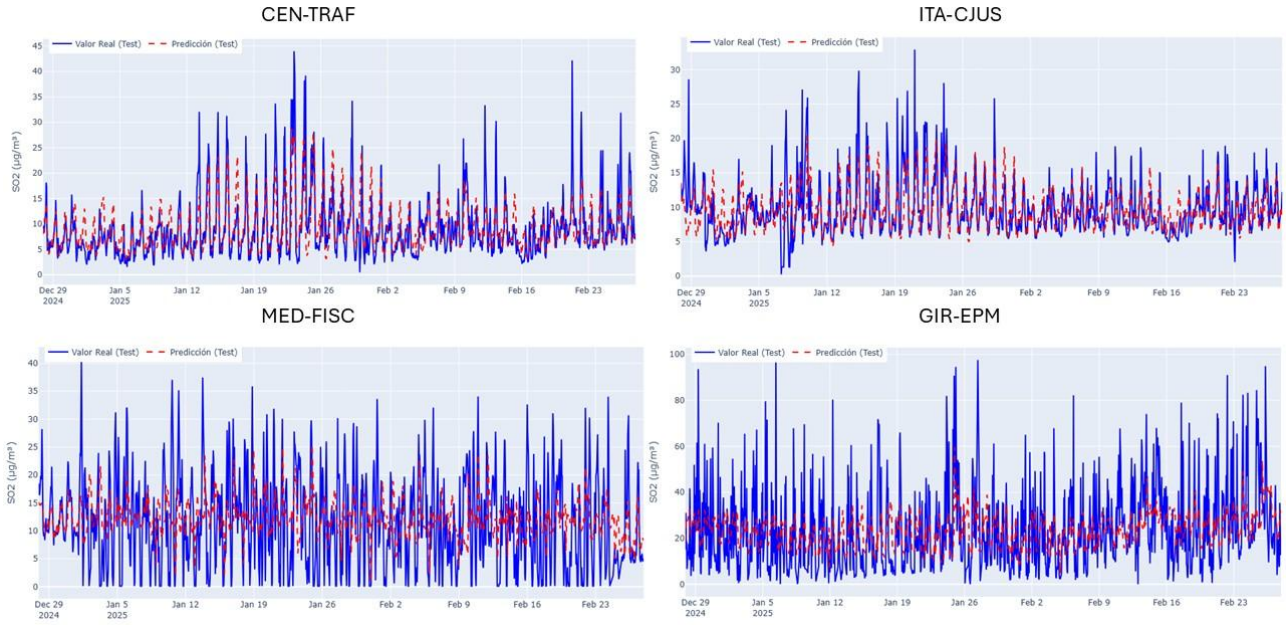

Fig. 13 Desempeño en test horizonte de predicción 24 horas

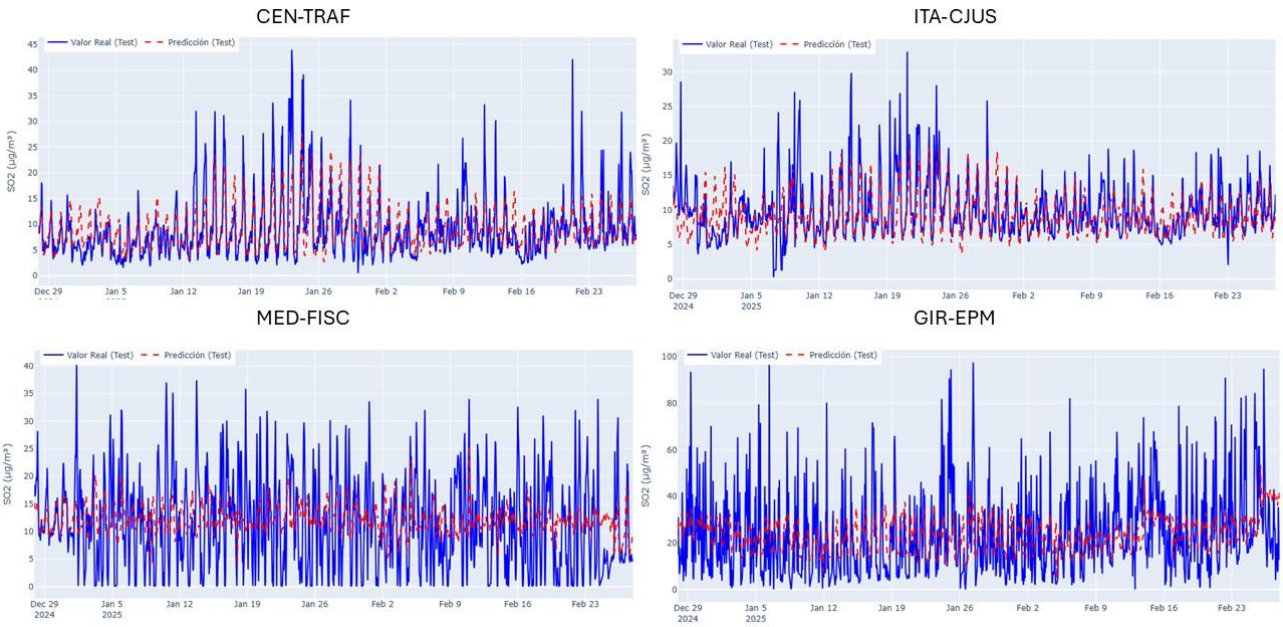

Fig. 14 Desempeño en test horizonte de predicción 72 horas

Los resultados numéricos presentados en la Tabla 9 complementan esta observación, mostrando que, si bien el error aumenta progresivamente de 6 a 24 y 72 horas, LGBM-direct mantiene en todas las estaciones un desempeño claramente superior al baseline. Las diferencias entre estaciones son, no obstante, importantes. ITA-CJUS se posiciona como la estación con los errores más bajos en todo el horizonte, con valores alrededor de 0.27–0.30 en 6 horas y aún cifras moderadas en 72 horas, lo que refleja una dinámica del contaminante más estable. CEN-TRAF exhibe un comportamiento intermedio, con errores algo mayores, pero igualmente muy por debajo del baseline. En contraste, MED-FISC presenta niveles de error más elevados, en especial a 72 horas (0.47), aunque la mejora frente al baseline sigue siendo sustancial. Finalmente, GIR-EPM es la estación con el ajuste más desafiante: presenta los WMAPE más altos en los tres horizontes y la mayor distancia respecto a las demás estaciones; aun así, el modelo reduce de manera considerable los valores del baseline (de 0.88 a 0.59 en 72 horas), evidenciando que, incluso en escenarios de alta variabilidad, LGBM-direct aporta una mejora significativa.

La Tabla 10 confirma este patrón al analizar el RMSE, donde la magnitud absoluta del error también crece con el horizonte y las diferencias entre estaciones se vuelven más marcadas en predicciones de 72 horas. ITA-CJUS vuelve a mostrar los valores más bajos (3.5–3.9), evidenciando poca dispersión del error incluso en horizontes largos. CEN-TRAF se mantiene en niveles intermedios (3.5–4.0), reflejando un comportamiento similar al observado en el WMAPE. MED-FISC alcanza valores más altos (hasta 5.8), aunque todavía muy inferiores a los del baseline. GIR-EPM, por su parte, presenta nuevamente los errores más altos del conjunto, con RMSE notoriamente mayores que los de las demás estaciones; pese a ello, la reducción frente al baseline sigue siendo importante (de más de 21 a alrededor de 14–15). En todas las estaciones y en todos los horizontes, LGBM-direct supera de forma sistemática al baseline, tanto en términos relativos como absolutos.

En conjunto, las métricas de prueba y las visualizaciones demuestran que, aunque el error aumenta de manera natural con el horizonte de predicción, el modelo mantiene un desempeño estable y consistentemente mejor al baseline. Además, las diferencias entre estaciones indican que la predictibilidad del SO_2 depende fuertemente de la dinámica local: es

mayor en estaciones con variabilidad moderada y más desafiante en aquellas con picos recurrentes o cambios abruptos, como en el caso evidente de GIR-EPM.

Tabla 9 WMAPE en test por horizonte de predicción

Modelo	CEN-TRAF			GIR-EPM			ITA-CJUS			MED FISC		
	6 h	24 h	72 h	6 h	24 h	72 h	6 h	24h	72h	6 h	24 h	72 h
Baseline	0.59	0.71	0.76	0.82	0.86	0.88	0.50	0.56	0.59	0.58	0.70	0.76
LGBM direct	0.32	0.35	0.37	0.59	0.60	0.59	0.27	0.29	0.30	0.36	0.43	0.47

Tabla 10 RMSE en test por horizonte de predicción

Modelo	CEN-TRAF			GIR-EPM			ITA-CJUS			MED FISC		
	6 h	24 h	72 h	6 h	24 h	72 h	6 h	24h	72h	6 h	24 h	72 h
Baseline	6.6	7.7	8.0	21.0	21.5	21.6	6.2	6.8	7.0	7.8	9.2	9.7
LGBM direct	3.5	3.8	4.0	14.7	14.5	14.3	3.5	3.7	3.9	4.6	5.3	5.8

5 Conclusiones

El análisis desarrollado evidencia que la calidad y completitud de los datos de SO_2 constituye una limitación estructural para el modelado predictivo en el Valle de Aburrá. La presencia de vacíos extensos, con interrupciones superiores a 2000 horas consecutivas en varias estaciones y un gran porcentaje de datos faltantes afectan de manera directa la representación de la variabilidad real del contaminante. Si bien los métodos de imputación permiten reconstruir series continuas útiles para el modelado conservando el comportamiento generador de la serie, estas deficiencias introducen incertidumbre adicional en todas las etapas del proceso. Contar con datos más completos, sensores con menor tiempo de inactividad y un sistema más robusto de control y mantenimiento podría mejorar significativamente la capacidad de los modelos y reducir la dependencia de correcciones posteriores.

Adicionalmente, el modelado del SO_2 se ve condicionado por la propia naturaleza de su comportamiento, asociada a transformaciones químicas y procesos secundarios que no se reflejan de forma directa en las variables disponibles, así como por las condiciones locales del Valle de Aburrá. La presencia de dinámicas altamente variables, acumulaciones puntuales y cambios abruptos en las concentraciones, favorecidos por la topografía cerrada del valle y sus condiciones particulares de ventilación, introduce una complejidad adicional que dificulta la captura de relaciones estables entre las variables explicativas disponibles y la evolución del contaminante. Estos procesos, marcadamente no lineales y en gran medida no observados de forma directa en las variables meteorológicas incluidas, constituyen una fuente adicional de incertidumbre que limita la capacidad de los modelos para representar de manera consistente la dinámica real del SO_2.

En consecuencia, es esperable que los modelos presenten mayores errores relativos y una capacidad restringida para reproducir picos abruptos o episodios extremos, particularmente en horizontes de predicción más largos. La ausencia de variables que representen explícitamente estos fenómenos limita la capacidad de los modelos para describir adecuadamente la dinámica del SO_2, lo cual se traduce en un incremento progresivo del error y en una mayor dispersión de las estimaciones.

No obstante, los resultados muestran que, aun bajo estas condiciones desafiantes, el modelo seleccionado (LGBM-direct) ofrece mejoras sustanciales frente al baseline y se posiciona como la alternativa más estable y precisa en todos los horizontes evaluados. Los valores de WMAPE, aunque elevados en términos absolutos debido a la fuerte variabilidad y presencia de picos en las series, se reducen entre un 30 % y un 43 % respecto al baseline, mientras que en RMSE las mejoras alcanzan hasta un 75 % en estaciones donde la referencia presenta errores muy pronunciados. Esto indica que el modelo no solo mantiene coherencia en sus predicciones, sino que captura de manera consistente la dirección de los cambios y evita la amplificación de errores, incluso en escenarios con ruido, discontinuidad y heterogeneidad entre estaciones. Su estabilidad entre validación y prueba evidencia una buena capacidad de generalización y confirma que puede ofrecer pronósticos útiles sin requerir arquitecturas más complejas.

En conjunto, estos resultados demuestran que, aunque los modelos no alcanzan niveles de error particularmente bajos en términos relativos, sí ofrecen predicciones suficientemente precisas en magnitud absoluta, especialmente reflejado en los valores de RMSE, que permanecen en rangos pequeños y estables incluso en horizontes amplios. Esto es lo que realmente permite anticipar con utilidad operativa las concentraciones futuras de SO_2, pues proporciona estimaciones numéricas razonablemente cercanas a los valores reales, necesarias para activar medidas preventivas y detectar episodios de contaminación elevada. En este sentido, más que aspirar a una precisión perfecta, los modelos cumplen su función principal: ofrecer señales tempranas de riesgo que permitan actuar antes de que los episodios críticos se materialicen, aun en un contexto caracterizado por alta variabilidad, discontinuidades en los datos y dinámicas locales complejas.

Adicionalmente, el comportamiento estable de los modelos a lo largo del horizonte de predicción refuerza su utilidad práctica en escenarios reales de gestión ambiental. La capacidad de mantener errores acotados y coherentes, incluso cuando aumenta la incertidumbre asociada a horizontes más largos, permite que los pronósticos sean utilizados como una referencia confiable para la anticipación de tendencias y cambios relevantes en la calidad del aire. De esta manera, los modelos no solo aportan información puntual, sino que contribuyen a una comprensión prospectiva del comportamiento del SO_2, facilitando la planificación y priorización de acciones de control y mitigación.

Contar con predicciones consistentes y numéricamente confiables respalda la labor de seguimiento continuo de la calidad del aire, contribuye a evaluar el avance hacia las metas establecidas por la Resolución 2254 de 2017 para 2030, y fortalece la capacidad institucional para gestionar riesgos, informar oportunamente a la ciudadanía y tomar decisiones basadas en evidencia. En este marco, los modelos predictivos no deben entenderse como herramientas determinísticas, sino como instrumentos estratégicos que permiten al SIATA anticipar escenarios adversos, optimizar la planificación operativa y mejorar la efectividad de las acciones de mitigación en el Valle de Aburrá, complementando los sistemas de monitoreo existentes y apoyando una gestión más preventiva de la contaminación atmosférica.

De cara a desarrollos posteriores, mejorar el desempeño de los modelos requerirá avanzar hacia series de SO_2 más completas, con menor discontinuidad y mayor estabilidad en las mediciones, de modo que el entrenamiento no dependa de procesos extensos de imputación y los modelos puedan capturar patrones más finos de variabilidad temporal. Asimismo, se abre un campo prometedor en el uso de métodos de ensamble, combinando modelos directos, recursivos y redes neuronales para aprovechar ventajas complementarias y continuar reduciendo el error tanto en WMAPE como en RMSE. Finalmente, resulta relevante explorar arquitecturas diseñadas específicamente para iniciar la predicción desde horizontes más amplios en lugar de depender exclusivamente del horizonte inmediato, lo que podría traducirse en mejoras adicionales en estabilidad y utilidad operativa para escenarios de anticipación temprana. Estas líneas de trabajo permitirían seguir fortaleciendo la calidad de los pronósticos y consolidar su aporte al sistema de gestión de la calidad del aire en la región.

6 Referencias

[1] World Health Organization, “Air pollution.” Accessed: Nov. 17, 2024. [Online]. Available: https://www.who.int/health-topics/air-pollution#tab=tab_1

[2] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8049504.

[3] World Health Organization, “Ambient (outdoor) air pollution,” 2024, Accessed: Jun. 06, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] J. S. Pérez-Carrasquilla, P. A. Montoya, J. M. Sánchez, K. S. Hernández, and M. Ramírez, “Forecasting 24h averaged PM2.5 concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information,” *Adv. Stat. Climatol. Meteorol. Oceanogr.*, vol. 9, no. 2, pp. 121–135, Dec. 2023, doi: 10.5194/ascmo-9-121-2023.

[5] United States Enviromental Protection Agency (EPA), “Sulfur Dioxide Basics,” 2025, Accessed: Nov. 03, 2025. [Online]. Available: https://www.epa.gov/so2-pollution/sulfur-dioxide-basics#effects#:~:text=What%20are%20the%20health%20effects,2

[6] P. Orellano, J. Reynoso, and N. Quaranta, “Short-term exposure to sulphur dioxide (SO2) and all-cause and respiratory mortality: A systematic review and meta-analysis,” *Environ. Int.*, vol. 150, May 2021, doi: 10.1016/j.envint.2021.106434.

[7] World Health Organization, “What are the WHO Air quality guidelines?” [Online]. Available: https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines?utm_source=chatgpt.com

[8] United States Enviromental Protection Agency (EPA), “Sulfur dioxide EPA”, Accessed: Oct. 19, 2025. [Online]. Available: https://www.epa.vic.gov.au/sulfur-dioxide

[9] World Health Organization, “Health consequences of air pollution,” 2024.

[10] Ministerio de ambiente y desarrollo sostenible, “Resolucion-2254-de-2017”, Accessed: Nov. 03, 2025. [Online]. Available: https://www.minambiente.gov.co/wp-content/uploads/2021/10/Resolucion-2254-de-2017.pdf

[11] R. Li, L. Cui, Y. Meng, Y. Zhao, and H. Fu, “Satellite-based prediction of daily SO2 exposure across China using a high-quality random forest-spatiotemporal Kriging (RF-STK) model for health risk assessment,” *Atmos. Environ.*, vol. 208, pp. 10–19, Jul. 2019, doi: 10.1016/j.atmosenv.2019.03.029.

[12] H. N. Mahendra *et al.*, “Assessment and Prediction of Air Quality Level Using ARIMA Model: A Case Study of Surat City, Gujarat State, India,” *Nature Environment and Pollution Technology*, vol. 22, no. 1, pp. 199–210, Mar. 2023, doi: 10.46488/NEPT.2023.V22I01.018.

[13] S. A. Shahriar *et al.*, “Potential of arima-ann, arima-svm, dt and catboost for atmospheric pm2.5 forecasting in bangladesh,” *Atmosphere (Basel).*, vol. 12, no. 1, pp. 1–21, Jan. 2021, doi: 10.3390/atmos12010100.

[14] A. Sotomayor-Olmedo, M. A. Aceves-Fernández, E. Gorrostieta-Hurtado, C. Pedraza-Ortega, J. M. Ramos-Arreguín, and J. E. Vargas-Soto, “Forecast Urban Air Pollution in Mex-

ico City by Using Support Vector Machines: A Kernel Performance Approach," *Int. J. Intell. Sci.*, vol. 03, no. 03, pp. 126–135, 2013, doi: 10.4236/ijis.2013.33014.
[15] S. U. Solehah *et al.*, "ENHANCING ECOSYSTEM BIODIVERSITY THROUGH AIR POLLUTION CONCENTRATIONS PREDICTION USING SUPPORT VECTOR REGRESSION APPROACHES," *International Journal of Conservation Science*, vol. 14, no. 4, pp. 1619–1626, Oct. 2023, doi: 10.36868/IJCS.2023.04.24.
[16] S. Babu and B. Thomas, "A survey on air pollutant PM2.5 prediction using random forest model," *Environmental Health Engineering and Management*, vol. 10, no. 2, pp. 157–163, Mar. 2023, doi: 10.34172/EHEM.2023.18.
[17] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the xgboost machine learning method in pm2.5 prediction: A case study of shanghai," *Aerosol Air Qual. Res.*, vol. 20, no. 1, pp. 128–138, Jan. 2020, doi: 10.4209/aaqr.2019.08.0408.
[18] J. Zhong *et al.*, "Robust prediction of hourly PM2.5 from meteorological data using LightGBM," *Natl. Sci. Rev.*, vol. 8, no. 10, Oct. 2021, doi: 10.1093/nsr/nwaa307.
[19] P. Perez, F. Gomez, C. Menares, and Z. L. Fleming, "Sulfur dioxide concentrations forecasting using a deep learning model in Quintero, Chile," *Atmos. Pollut. Res.*, vol. 16, no. 8, Aug. 2025, doi: 10.1016/j.apr.2025.102534.
[20] S. Ben Taieb and R. J. Hyndman, "Recursive and direct multi-step forecasting: the best of both worlds," 2012. [Online]. Available: http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/
[21] W. Wang, W. Mao, X. Tong, and G. Xu, "A novel recursive model based on a convolutional long short-term memory neural network for air pollution prediction," *Remote Sens. (Basel).*, vol. 13, no. 7, Apr. 2021, doi: 10.3390/rs13071284.
[22] S. Park and M. J. Kim, "Forecasting Ultrafine Dust Concentrations in Seoul: A Machine Learning Approach," *Atmosphere (Basel).*, vol. 16, no. 3, Mar. 2025, doi: 10.3390/atmos16030239.
[23] I. Kalate Ahani, M. Salari, and A. Shadman, "Statistical models for multi-step-ahead forecasting of fine particulate matter in urban areas," *Atmos. Pollut. Res.*, vol. 10, no. 3, pp. 689–700, May 2019, doi: 10.1016/j.apr.2018.11.006.
[24] IBM, "Conceptos básicos de ayuda de CRISP-DM." Accessed: Nov. 19, 2024. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
[25] SIATA, "Quienes Somos." Accessed: Nov. 19, 2024. [Online]. Available: https://siata.gov.co/sitio_web/index.php/nosotros
[26] ECMWF, "ERA5-Land", Accessed: Nov. 03, 2025. [Online]. Available: https://www-ecmwf-int.translate.goog/en/era5-land?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc
[27] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended Isolation Forest," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1479–1489, Apr. 2021, doi: 10.1109/TKDE.2019.2947676.
[28] N. Loperfido, "Kurtosis-based projection pursuit for outlier detection in financial time series," *European Journal of Finance*, vol. 26, no. 2–3, pp. 142–164, Feb. 2020, doi: 10.1080/1351847X.2019.1647864.
[29] IBM, "What are ARIMA models? Introducing ARIMA models Hello! How can we help you?" [Online]. Available: https://www.ibm.com/think/topics/arima-model

[30] R. J. , Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. 2021.
[31] T. , Hastie, R. , Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009.
[32] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." [Online]. Available: https://github.com/Microsoft/LightGBM.
[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[34] I. , Goodfellow, A. Courville, and Y. Bengio, "Deep Learning," 2016.