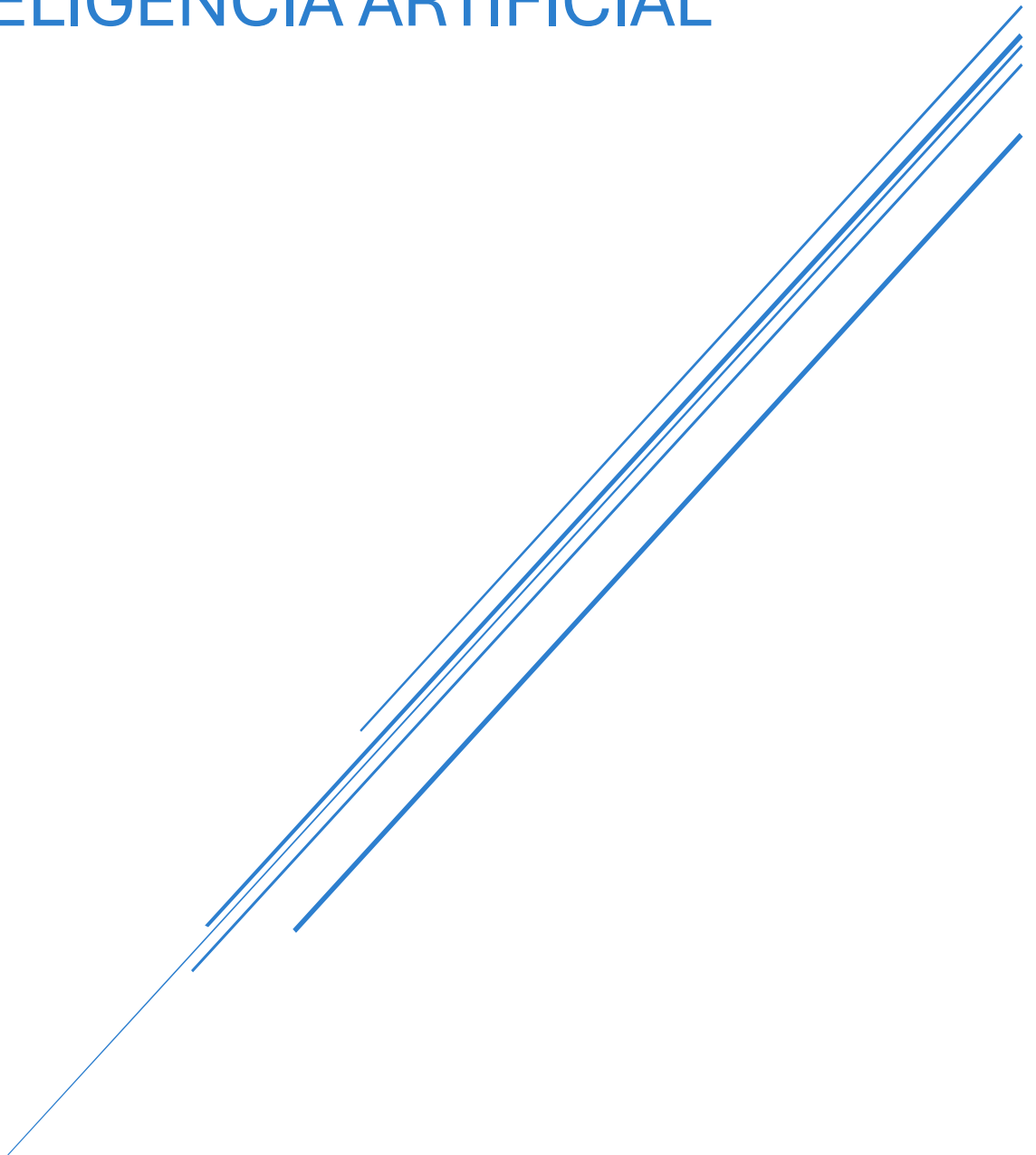


# ANÁLISIS Y PREDICCIÓN DE LA CALIDAD DEL AIRE EN VALENCIA MEDIANTE OPEN DATA Y MODELOS DE INTELIGENCIA ARTIFICIAL



Raúl Sancha Llamosí

## Contenido

Resumen Ejecutivo.....	2
Contexto y Motivación .....	2
Objetivos del Proyecto .....	3
Descripción de los Datos Utilizados .....	4
Metodología.....	4
Resultados Obtenidos .....	6
Impacto Social y Potencial de Reutilización .....	11
Conclusiones y Futuros Desarrollos .....	11
Relación con los Objetivos del Premio .....	12

## Resumen Ejecutivo

El presente proyecto analiza la influencia combinada de la **meteorología** y la **ubicación urbana de las estaciones de medición** en los niveles de **calidad del aire en la ciudad de València**, integrando más de una década de datos abiertos municipales, autonómicos y europeos. A través del uso de **técnicas avanzadas de ciencia de datos e inteligencia artificial**, se desarrollan modelos predictivos capaces de anticipar episodios de contaminación, con aplicaciones prácticas en gestión urbana y prevención sanitaria.

Partiendo de datos abiertos procedentes del **Portal de Transparencia del Ayuntamiento de València**, la Generalitat Valenciana, AEMET, AVAMET y el programa europeo Copernicus, se ha construido un conjunto de datos integrados que combina información sobre contaminantes atmosféricos, condiciones climáticas y características del entorno urbano (alturas edificatorias, vegetación, red viaria, etc.).

Mediante algoritmos de clasificación no supervisada se identifican **arquetipos meteorológicos y tipologías de entorno urbano**, y se analizan sus efectos sobre los principales contaminantes. A continuación, se desarrollan modelos predictivos de concentración para  $\text{NO}_2$ ,  $\text{O}_3$ ,  $\text{PM}_{10}$  y  $\text{PM}_{2.5}$  utilizando **Random Forest**, optimizados y validados. Finalmente, se implementa una **aplicación web interactiva**, desarrollada con tecnología open-source, que permite realizar predicciones personalizadas a partir de condiciones introducidas por el usuario.

Este trabajo no solo proporciona herramientas prácticas de análisis y predicción de la contaminación atmosférica, sino que también **promueve la cultura del open data**, la reutilización responsable de la información pública y el desarrollo de soluciones tecnológicas orientadas al bien común.

## Contexto y Motivación

La **contaminación atmosférica** es uno de los principales problemas ambientales de las ciudades contemporáneas. Numerosos estudios han demostrado su impacto directo sobre la **salud pública**, con especial incidencia en personas con enfermedades respiratorias, niños y personas mayores. En ciudades como València, el tráfico rodado, las condiciones meteorológicas y la morfología urbana interactúan de forma compleja, afectando los niveles de concentración de contaminantes como el dióxido de nitrógeno ( $\text{NO}_2$ ), el ozono ( $\text{O}_3$ ) y las partículas en suspensión ( $\text{PM}_{10}$  y  $\text{PM}_{2.5}$ ).

La **ubicación de las estaciones de medida** de la calidad del aire no es neutra. Elementos como la orientación de la calle, la altura de los edificios o la presencia

de vegetación pueden condicionar significativamente los registros. Además, la meteorología local (vientos, temperaturas, humedad, presión) también afecta la dispersión o acumulación de contaminantes.

Frente a esta complejidad, el **acceso a datos abiertos** permite abordar el problema desde una perspectiva analítica avanzada. En particular, el **portal de datos abiertos del Ayuntamiento de València** ofrece una oportunidad única para acceder a información geoespacial clave (ubicación de estaciones, red viaria, edificios...), la cual, combinada con datos de calidad del aire, meteorología y teledetección, facilita la construcción de modelos interpretables y útiles para la ciudad.

El proyecto se inscribe dentro de la iniciativa PREDATICS, pero trasciende sus límites para crear una propuesta autónoma, centrada en València, con vocación de servicio público y abierta a la comunidad. Esta motivación impulsa la creación de una herramienta replicable y transparente que pueda integrarse en procesos de planificación urbana, campañas de salud pública o iniciativas ciudadanas por el aire limpio.

## Objetivos del Proyecto

- Analizar la relación entre condiciones meteorológicas y niveles de contaminación en Valencia.
- Evaluar la influencia del entorno urbano inmediato de cada estación en los datos registrados.
- Desarrollar modelos predictivos de calidad del aire mediante aprendizaje automático.
- Implementar una aplicación interactiva que permita realizar predicciones según distintos escenarios.
- Promover el uso y visualización de datos abiertos en herramientas útiles para la ciudadanía.

## Descripción de los Datos Utilizados

- **Calidad del aire:** Datos de la Generalitat Valenciana (2010–2024) por estación y contaminante.
- **Meteorología:** Estaciones de AEMET y AVAMET en el municipio de València.
- **Datos urbanos:** Datos abiertos del Ayuntamiento de València:
  - Localización de estaciones de calidad del aire.
  - Red viaria, altura de edificios.
  - Uso del suelo (Copernicus).
  - Índices geoespaciales (NDVI, NDBI, UI) calculados desde Sentinel-2 (2016–2024).

## Metodología

El proyecto sigue un enfoque multidisciplinar que combina técnicas de análisis estadístico, procesamiento de datos geoespaciales y modelado predictivo mediante inteligencia artificial, todo ello aplicado sobre datos abiertos de calidad del aire, meteorología y entorno urbano de la ciudad de València.

### 6.1. Recopilación y tratamiento de datos

Se recopilaron datos desde 2010 hasta 2024, procedentes de:

- **Portal de datos abiertos del Ayuntamiento de València** (ubicación de estaciones, red viaria, edificios, capas urbanas).
- **Generalitat Valenciana** (datos de calidad del aire por estación).
- **AEMET y AVAMET** (variables meteorológicas: temperatura, humedad, viento, precipitación, etc.).
- **Copernicus – Sentinel-2** (uso del suelo, índices de vegetación y construcción mediante NDVI, NDBI y UI).

Todos los datos fueron estandarizados, limpiados, convertidos a formatos homogéneos y georreferenciados. Se realizó un emparejamiento entre estaciones meteorológicas y de calidad del aire mediante análisis de proximidad (radio  $\leq 1.500$  m).

### 6.2. Procesamiento geoespacial

Con herramientas como QGIS y Google Earth Engine (GEE), se delimitaron zonas de influencia de 150 m alrededor de cada estación. En estas áreas se extrajeron variables como:

- Densidad de edificación y altura media de edificios.
- Longitud y tipo de vías (por velocidad).
- Cobertura vegetal y urbana (NDVI, NDBI, UI).
- Tipos de uso del suelo (urbano, agrícola, forestal...).

### 6.3. Análisis exploratorio y clasificación

Se calculó la correlación entre variables meteorológicas y contaminantes atmosféricos mediante el coeficiente de Pearson, tanto a nivel global como por estación y temporada. Además, se exploró el efecto de las condiciones meteorológicas del **día anterior** sobre los niveles de contaminación.

Se identificaron **arquetipos meteorológicos** usando clustering (KMeans, DBSCAN y GMM), agrupando días con condiciones atmosféricas similares, para estudiar cómo se comportan los contaminantes bajo distintos regímenes climáticos.

También se agruparon las estaciones en función de su **entorno urbano** mediante técnicas no supervisadas de clustering, generando **arquetipos urbanos** que ayudan a interpretar la representatividad espacial de cada estación.

### 6.4. Modelos predictivos

Se desarrollaron dos enfoques de modelado:

- **Regresión lineal múltiple**, para evaluar relaciones lineales entre contaminantes y variables predictoras.
- **Random Forest**, para modelar relaciones complejas y no lineales. Se desarrollaron:
  - Modelos individuales por contaminante.
  - Modelos multisalida.
  - Modelos con ajuste de hiperparámetros mediante validación cruzada.

Los modelos se entrenaron y validaron con división 80/20 (entrenamiento/prueba) y se evaluaron usando métricas como **R<sup>2</sup>, MAE y MSE**.

### 6.5. Herramienta interactiva

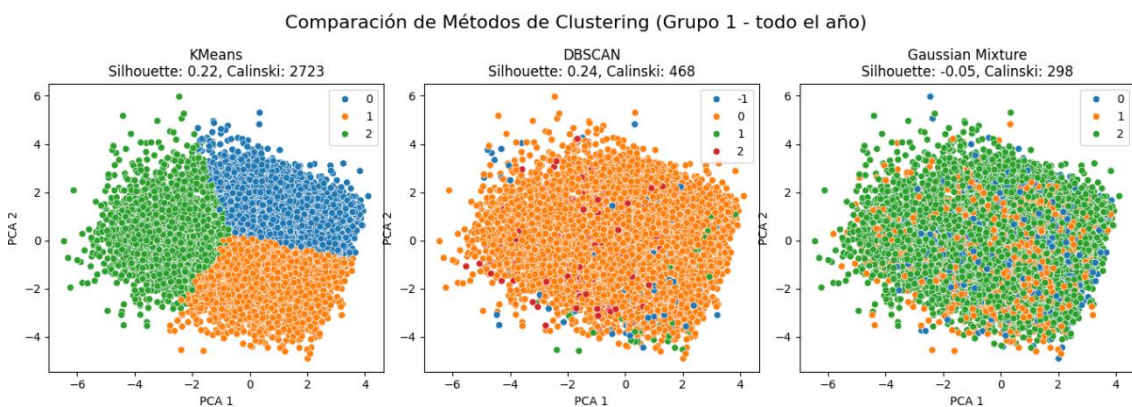
Finalmente, se integraron los modelos en una aplicación web funcional, desarrollada con **Streamlit**, que permite:

- Introducir variables meteorológicas y urbanas.
- Seleccionar arquetipos meteorológicos y urbanos.
- Obtener predicciones de concentración para cada contaminante.
- Visualizar el resultado y su fiabilidad ( $R^2$ ) de forma intuitiva.

## Resultados Obtenidos

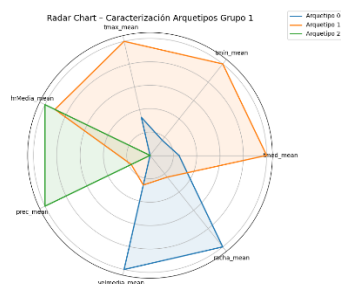
Los resultados obtenidos destacan por su solidez técnica y su aplicabilidad directa al análisis ambiental y la toma de decisiones en la ciudad de València.

### Arquetipos meteorológicos

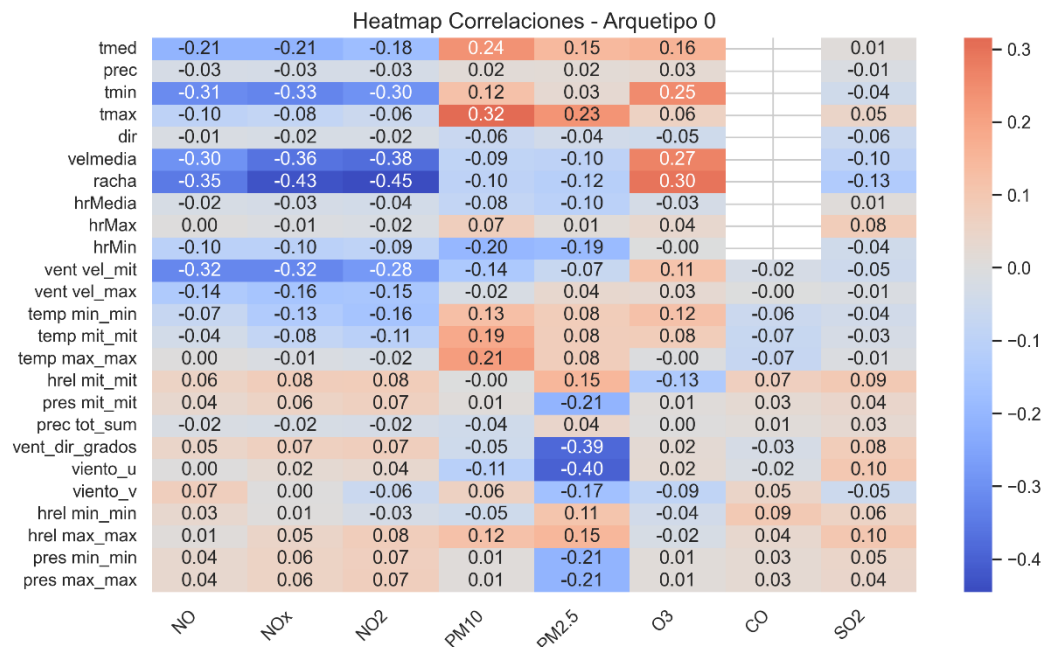


Se identificaron tres patrones meteorológicos principales:

- **Arquetipo 0 (Patrón estival):** Altas temperaturas, baja humedad, viento moderado. Relacionado con niveles altos de  $O_3$ .
- **Arquetipo 1 (Transicional):** Condiciones mixtas, frecuentes en primavera y otoño.
- **Arquetipo 2 (Régimen invernal):** Baja temperatura, alta estabilidad atmosférica, poca ventilación. Asociado a picos de  $NO_2$  y PM.

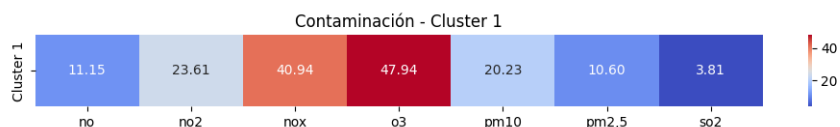
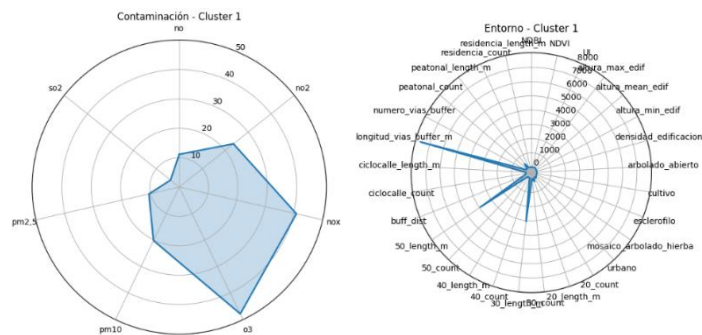


Estos arquetipos permiten interpretar los comportamientos estacionales de la contaminación y anticipar sus valores con mayor precisión.



## 7.2. Influencia del entorno urbano

Se comprobó que factores como la **altura de edificios**, la **proximidad a vías rápidas**, o la **escasez de vegetación** influyen directamente en la acumulación de contaminantes, especialmente NO<sub>2</sub> y PM<sub>2.5</sub>. Algunas estaciones, por su localización, **no representan fielmente el aire de su entorno inmediato**, lo que podría distorsionar estudios urbanos si no se corrige.

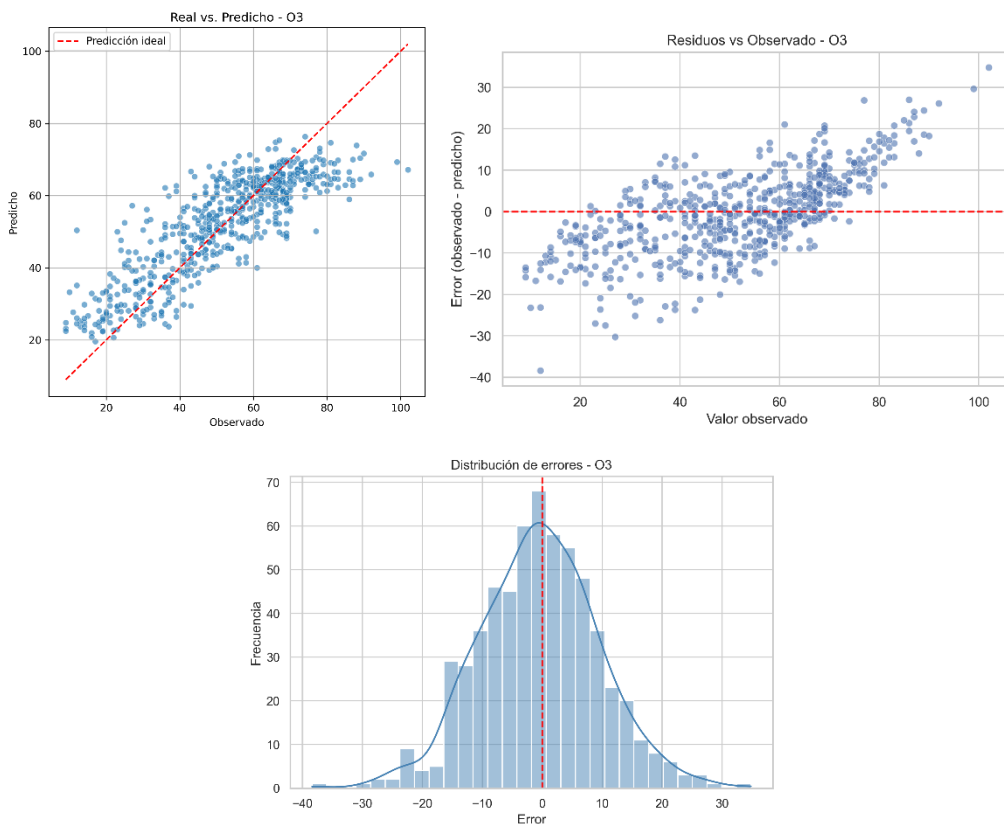


### 7.3. Modelos predictivos

- Los modelos de **Random Forest** superaron a la regresión lineal en precisión:
  - **R<sup>2</sup> superior a 0.85** para contaminantes como NO<sub>2</sub> y PM<sub>10</sub>.
  - Precisión ligeramente menor en O<sub>3</sub>, debido a su formación secundaria más compleja.
- La incorporación de **arquetipos meteorológicos y urbanos** como variables categóricas mejoró el rendimiento general de los modelos.

Además, el análisis de importancia de variables mostró que:

- Para NO<sub>2</sub> y NOx: influyen viento, presión, densidad urbana y uso del suelo.
- Para O<sub>3</sub>: dominan temperatura, radiación solar (estimada), y estabilidad atmosférica.
- Para partículas: la humedad y la vegetación urbana tienen un peso significativo.



## 7.4. Herramienta web interactiva

La aplicación final permite:

- Ingresar condiciones meteorológicas y urbanas.
- Obtener la predicción diaria para NO<sub>2</sub>, PM<sub>10</sub>, PM<sub>2.5</sub>, O<sub>3</sub>, etc.
- Visualizar si la concentración prevista supera o no los valores recomendados por la OMS.
- Consultar el nivel de fiabilidad del modelo para cada contaminante.

Esto representa una **aportación directa a la ciudadanía** y a la toma de decisiones técnicas en urbanismo y salud pública.

The screenshot shows a web browser window with the URL localhost:8501. The page features logos for DADES VALÈNCIA, AJUNTAMENT DE VALÈNCIA, and PREDATICS. The main heading is 'Predicción de Contaminantes Atmosféricos'. Below it, there is a section titled 'Predicción de Contaminantes con Random Forest' with a dropdown menu set to 'SO2'. A section titled 'Variables meteorológicas' displays a grid of 18 input fields for meteorological data, each with a numerical value and a +/- control.

Variables meteorológicas	
Altitud	Dirección del viento (°)
6,00	25
Temperatura media (°C)	Velocidad media viento (m/s)
12,20	5,30
Precipitación (mm)	Racha máxima (m/s)
0,00	20,60
Temperatura mínima (°C)	Hora Racha (min)
9,70	870
Hora Tmin (min desde medianoche)	Humedad relativa media (%)
290	51
Temperatura máxima (°C)	HR máxima (%)
14,70	60
Hora Tmax (min desde medianoche)	Hora HR Max (min)
835	671
	HR mínima (%)
	43
	Hora HR Min (min)
	735

## Arquetipos

Arquetipo

Estival

Estival

Transicional

Invernal

Arquetipo 3 – Urbano compacto

## Arquetipos ↔

Arquetipo

Estival

Época del año

Invierno

Arquetipo entorno

Arquetipo 3 – Urbano compacto

Arquetipo 3 – Urbano compacto

Arquetipo 2 – Natural o rural

Arquetipo 1 – Urbano consolidado

Arquetipo 0 – Mixto verde-residencial

## Variables urbanas/ambientales

NDBI	NDVI	UI	altura_max_edif
0,00 - +	0,00 - +	0,00 - +	0,00 - +
altura_mean_edif	altura_min_edif	densidad_edificacion	arbolado_abierto
0,00 - +	0,00 - +	0,00 - +	0,00 - +
cultivo	esclerofilo	mosaico_arbolado_hierba	urbano
0,00 - +	0,00 - +	0,00 - +	0,00 - +
20_count	20_length_m	30_count	30_length_m
0,00 - +	0,00 - +	0,00 - +	0,00 - +
40_count	40_length_m	50_count	50_length_m
0,00 - +	0,00 - +	0,00 - +	0,00 - +
ciclocalle_count	ciclocalle_length_m	longitud_vias_buffer_m	numero_vias_buffer
0,00 - +	0,00 - +	0,00 - +	0,00 - +
peatonal_count	peatonal_length_m	residencia_count	residencia_length_m
0,00 - +	0,00 - +	0,00 - +	0,00 - +

Predecir

Predicción para O3:

**77.78  $\mu\text{g}/\text{m}^3$**

Fiabilidad del modelo ( $R^2$ ): 0.619

Predicción para SO2:

**3.21  $\mu\text{g}/\text{m}^3$**

Fiabilidad del modelo ( $R^2$ ): 0.064

## Impacto Social y Potencial de Reutilización

- Contribuye a una **ciudad más transparente**, permitiendo entender cómo y por qué varía la calidad del aire.
- Permite a ciudadanos, técnicos y responsables públicos **anticipar episodios de contaminación**.
- Ofrece una **base replicable** para otras ciudades con datos abiertos similares.
- Refuerza el papel del **portal de datos abiertos del Ayuntamiento de València** como fuente útil y accesible para investigación e innovación.

## Conclusiones y Futuros Desarrollos

El presente trabajo demuestra que es posible **reutilizar eficazmente datos abiertos** de calidad del aire, meteorología y entorno urbano para generar **conocimiento aplicable y herramientas predictivas** al servicio de la ciudad.

Entre las principales conclusiones destacan:

- La combinación de condiciones meteorológicas y características del entorno urbano **explica una parte significativa de la variabilidad en los niveles de contaminación atmosférica** registrados en València.
- La clasificación en **arquetipos meteorológicos y arquetipos urbanos** permite interpretar mejor los datos registrados y ajustar los modelos predictivos con mayor precisión local.
- El uso de técnicas de inteligencia artificial, en particular **Random Forest**, ha permitido obtener modelos robustos, con una **fiabilidad superior al 80 % en algunos contaminantes**, lo que los hace viables para aplicaciones prácticas.
- La herramienta web desarrollada **acerca la predicción de la calidad del aire a la ciudadanía**, de forma intuitiva y transparente, demostrando que el open data puede convertirse en servicios digitales con impacto real.
- Se ha verificado que algunas estaciones presentan **entornos no representativos**, lo cual puede afectar la interpretación de los datos de fondo urbano. Esto pone de manifiesto la necesidad de revisar su distribución para mejorar la representatividad espacial del monitoreo.

### Futuros desarrollos

Este trabajo abre la puerta a múltiples líneas de mejora y ampliación, entre las que destacan:

- **Actualización automática** con datos en tiempo real para generar alertas inmediatas y análisis en vivo.
- Integración con **otros datos municipales**: tráfico, movilidad, ruido, datos de salud pública.
- Extensión del sistema a nivel **metropolitano** o a otras ciudades con portales de datos abiertos similares.
- Desarrollo de una **versión móvil** de la herramienta para mayor accesibilidad.
- Publicación del código y datasets en un repositorio abierto, fomentando la **reproducibilidad científica y la colaboración ciudadana**.

## Relación con los Objetivos del Premio

Este proyecto cumple plenamente los criterios establecidos:

- ✓ Promueve la **transparencia** y el uso del portal de datos abiertos del Ayuntamiento de València.
- ✓ Aplica buenas prácticas de **reutilización de datos** públicos.
- ✓ Genera **herramientas innovadoras** accesibles y replicables.
- ✓ Contribuye al **conocimiento de la ciudad** y a su gestión ambiental.
- ✓ Fomenta la **cultura del open data** entre investigadores, técnicos y ciudadanía.