

Herramienta de previsión de demanda en Urgencias para el Hospital Puerta de Hierro



Candidatura a los XIV Premios a la
Innovación y la Mejora de la
Competitividad Energética a través de la
Digitalización



Introducción

Una necesidad, una colaboración, un proyecto, una realidad.

En los meses de septiembre y octubre de 2025, Alhambra IT en su sección salud, Alhambra Health planteó al grupo de investigación GHEODE de la Universidad de Alcalá, con el que ya ha colaborado previamente, la necesidad del Hospital Puerta de Hierro de mejorar su actual sistema de predicción de entrada de personas en urgencias, tanto en su totalidad como particularizado en tres categorías “Pediatría”, “Ginecología y Obstetricia” y “General” mediante técnicas avanzadas de inteligencia artificial. Pequeñas mejoras porcentuales en la estimación de la demanda implican beneficios sustanciales tanto en la planificación de recursos humanos y materiales como en la optimización de consumos energéticos y, sobre todo en la mejora asistencial de las personas.

Tras una fase de estudio previo se definió un proyecto que comenzó en noviembre del año 2025 y que actualmente está en su fase de últimas pruebas en pre-producción y paso a producción, por lo que es ya una realidad y un caso de éxito en la aplicación de las tecnologías de la información y comunicaciones (TIC) en la transformación digital de una entidad pública.

Tanto en la memoria introducida en la plataforma como en este documento se puede comprobar la mejora sustancial en la predicción, llegando a errores de únicamente el 5%, que están en el límite intrínseco de ruido de la señal de entrada, es decir, los datos de asistencia a urgencias en el periodo 2022-2025.

Este documento, complementario al formulario presentado en la plataforma, muestra de manera abreviada parte de la experimentación y resultados realizados, así como el manual de la aplicación, donde se podrá comprobar tanto la utilidad como la usabilidad de la herramienta.

Equipo de Trabajo

Como llevar a la realidad la colaboración público-privada.

Cuando se habla de colaboración público-privada en la innovación siempre se tratan tres pilares fundamentales: La academia, las empresas y las administraciones públicas. Este proyecto, esta realidad que está presta a ser implementada en el hospital se puede considerar el paradigma de esta colaboración. Las personas que han compuesto el equipo en este proyecto son

- MDr. Miguel Cuchi, Director Médico y Director de Innovación Hospital Universitario de Puerta de Hierro Majadahonda.
- Claudio Andrés Paradela, Director de Desarrollo e Integración de datos y procesos · Alhambra IT.
- Juan José Gómez Borrallo, Responsable del Area de Salud, Alhambra IT
- Dr. Antonio Portilla, Catedrático de Universidad, Universidad de Alcalá.
- Dr. Sancho Salcedo, Catedrático de Universidad, Universidad de Alcalá.
- Dra. Silvia Jiménez, Profesora Titular de Universidad, Universidad de Alcalá
- Dra Laura Cornejo, Profesora Permanente Laboral, Universidad de Alcalá
- Dr. Jorge Pérez, Profesor Ayudante Doctor, Universidad de Alcalá
- Dr. Alberto Palomo, Profesor Sustituto, Universidad de Alcalá.
- Jorge Ruiz Morales, Investigador, Universidad de Alcalá

Breve descripción de los algoritmos

La herramienta es una aplicación web autónoma que estima la demanda diaria del Servicio de Urgencias con un horizonte de siete días. A partir del registro histórico completo de episodios del hospital (778.472 episodios entre 2022 y 2026, con una media de unas 533 llegadas al día), predice, para cada uno de los próximos siete días, las **llegadas totales**, su desglose **por área asistencial** (Pediatría, Ginecología-Obstetricia y Resto) y los **ingresos esperados** para la planificación de camas. Cada cifra se acompaña de una **banda de fiabilidad** (intervalo P10–P90) y de un **margen de seguridad α** ajustable, y puede consultarse en pantalla o descargarse como parte de previsión en PDF. La aplicación funciona sin conexión a internet (salvo para actualizar los índices de búsqueda de Google Trends) y permite incorporar datos nuevos y reentrenar los modelos.

El motor combina varios modelos de aprendizaje automático entrenados sobre la misma serie temporal diaria, empleando una ventana de 28 días de historia para generar los 7 días siguientes:

- **Transformer** (red neuronal encoder–decoder autorregresiva, con decodificador causal de tipo GPT): aprende la estacionalidad semanal (pico los lunes, valle el fin de semana) y anual (mayor demanda en invierno). Recibe como entrada la serie pasada junto con el calendario futuro conocido (día de la semana, festivos), variables meteorológicas y señales de búsqueda de Google Trends (gripe, fiebre, urgencias, etc.).
- **XGBoost** (gradient boosting sobre árboles de decisión), en dos variantes —ventana completa y versión resumida por retardos y medias móviles— como modelos complementarios.
- **Transformer + fiabilidad** (predicción por cuantiles): el modelo más preciso. Sobre el residuo del transformer entrena una cabeza monótona de cuantiles ($P10 \leq P50 \leq P90$, con función de pérdida *pinball*) que produce el intervalo de confianza. El pronóstico operativo se calcula con la fórmula $P50 + \alpha \cdot (P90 - P50)$: con $\alpha = 0$ se emplea la predicción media (se queda corta aproximadamente la mitad de los días) y con $\alpha = 1$ la predicción garantista (solo se queda corta en torno a 1 de cada 10 días), de modo que el gestor elige el margen según su tolerancia al riesgo.

La previsión por áreas y por ingresos reentrena esta misma arquitectura para cada objetivo. En la validación con datos reales del primer semestre de 2026, el modelo de fiabilidad obtiene un error porcentual medio (MAPE) en torno al 5–7 % en un horizonte temporal de 1 a 7 días, reduciendo prácticamente a la mitad el error del modelo actualmente en uso en el hospital (error absoluto medio de unas 57 a unas 29 llegadas al día) y situándose cerca del suelo de ruido intrínseco del propio dato (unos 17 episodios al día).

Manual de la Herramienta

Manual de la herramienta

Previsión de demanda en Urgencias

Hospital Universitario Puerta de Hierro Majadahonda (Madrid)

Introducción

La aplicación es una herramienta web autónoma que predice la demanda diaria del Servicio de Urgencias a siete días vista: llegadas totales, desglose por área asistencial e ingresos, cada uno con su banda de fiabilidad. Se ejecuta en local y funciona sin conexión a internet, salvo para actualizar los índices de Google Trends.

En el modo preproducción, para ejecutarla basta con hacer doble clic en «UrgenciasApp.exe»: se inicia un servidor local y se abre el navegador en la dirección <http://127.0.0.1:5000>. La barra superior da acceso a las ocho secciones que se describen a continuación. Los detalles de los modelos y algoritmos empleados se recogen en el documento de descripción de la herramienta.

En el modo de producción se instalará la herramienta en los servidores de Alhambra IT que, como se ha indicado en el apartado “Cuantificación/Estimación reducción emisiones CO2” utilizan energía con garantía de origen 100 % verde y neutralidad de carbono, reduciendo la huella de mercado asociada a la operación digital.

1. Inicio

Es la página de entrada de la aplicación. Presenta el objetivo de la herramienta (a partir de la serie histórica de llegadas a urgencias y su contexto se entrenan varios modelos de predicción) y muestra las cifras clave: número de episodios del periodo 2022–2026, días de la serie, y el error porcentual medio (MAPE) del modelo de fiabilidad en validación y en el test real del primer semestre de 2026.

Recuerda además los dos patrones que dominan la demanda —la estacionalidad semanal (pico los lunes, valle el fin de semana) y la anual (más urgencias en invierno)— y el llamado «suelo de ruido» de unos 17 episodios al día, que marca el límite práctico de precisión de cualquier modelo. Desde aquí se enlaza al resto de secciones.

Predicción de afluencia a urgencias (agregada y por áreas asistenciales) y de ingresos

URGENCIAS · HOSPITAL PUERTA DE HIERRO

INICIO

PROCESADO

DATASET

URGENCIAS

ÁREA

INGRESOS

INFORME

CARGAR DATOS

SISTEMA DE PREVISIÓN DE DEMANDA ASISTENCIAL

Predicción de llegadas diarias a urgencias para Hospital Puerta de Hierro

Modelos de forecasting entrenados sobre el registro completo de episodios del hospital (media de ~533 urgencias/día). La aplicación predice las llegadas diarias a +1...+7 días, cuantifica su fiabilidad y permite ampliar la base con datos nuevos.

778.472

episodios (2022–2026)

1642

días en la serie

4.82%

MAPE validación (transformer + fiabilidad)

5.23%

MAPE test 2026-H1 (semestre real)

Objetivo

Partiendo de la serie temporal de llegadas a urgencias y la información asociada a ella (datos asistenciales del paciente y del ingreso) se entrenan distintos modelos de predicción.

Es importante señalar que existe **estacionalidad semanal** (hay un pico de llegadas a urgencias los lunes y un valle durante los fines de semana) y **anual** (siendo las llegadas superiores durante el invierno). Adicionalmente, existe un **suelo de ruido** de unos 17 episodios al día que va a limitar la precisión alcanzable.

Se han entrenado distintos modelos, que pueden consultarse en las pestañas [Urgencias](#), [Área](#) e [Ingresos](#).

01

Cómo se procesa el dato (SQL a modelo)

02

Explorar el dataset diario

03

Urgencias: probar un modelo y medir su acierto

04

Predicción por área asistencial

05

Previsión de ingresos para camas

06

Cargar datos y ampliar la base

Hospital Universitario Puerta de Hierro Majadahonda · Madrid · 778.472 episodios (2022–2026) · Alhambra — Universidad de Alcalá · v2

2. Procesado

Explica de forma visual cómo se transforma el dato desde el volcado original del hospital hasta la serie que alimentan los modelos: los ficheros de episodios y de codificación clínica se unifican en un único conjunto por episodio y, a partir de él, se construye la serie temporal diaria con sus variables de calendario, clima e índices de búsqueda de Google Trends.

Es una sección documental (no requiere interacción) pensada para que cualquier persona entienda el recorrido del dato «del SQL al modelo».

Predicción de afluencia a urgencias (agregada y por áreas asistenciales) y de ingresos

URGENCIAS · HOSPITAL PUERTA DE HIERRO

INICIO **PROCESADO** DATASET URGENCIAS ÁREA INGRESOS INFORME CARGAR DATOS

Procesado de los datos

1. Datos extraídos de la base de datos del hospital

Se dispone de una **base de datos** que recoge las llegadas a urgencias registradas desde el **1 de enero de 2022** hasta el **30 de junio de 2026**.

A cada llegada de una persona al hospital y los consiguientes datos demográficos, asistenciales, del triaje, etc. se le denomina “**episodio**”. En total se dispone de unos **778.000 episodios**.

La información se recoge en **dos tablas** distintas: episodios y diagnósticos.

- **Tabla de episodios:** una fila por visita, con la fecha y hora de entrada, la de salida, la edad del paciente, etc.
- **Tabla de diagnósticos:** qué enfermedades se le diagnosticaron a cada paciente (una visita puede tener varios diagnósticos, así que aquí hay varias filas por episodio).

No se usa el fichero “informe médico” de cada visita, dado que contiene **datos personales sensibles** y los modelos no necesitan su utilización.

2. Pasos seguidos por la aplicación

En la pestaña [Cargar datos](#) se pueden actualizar o añadir los registros de episodios, así como los códigos clínicos asociados a cada episodio.

En primer lugar, dado que cada visita tiene un **identificador único** (como el número de expediente), se **emparejan** las visitas a los diagnósticos obtenidos. Es importante notar que a una visita se pueden asociar **varios diagnósticos**.

Asimismo, para cada visita se calculan una serie de **variables** que posteriormente se emplean para ajustar los modelos:

- **Edad** del paciente.
- **Tiempo que estuvo** en urgencias.
- **Número de diagnósticos** que tuvo y **de qué tipo** eran (código CIE-10). Es importante notar que, debido a la gran cantidad de códigos CIE, estos se han agrupado por **capítulos**.

3. Resumen diario

Adicionalmente, se realiza un resumen diario que se puede consultar en la pestaña [Dataset](#).

El resumen consta de la siguiente información:

- **Cuántas personas llegaron a urgencias** ese día. **Es la variable a predecir**.
- Cuántas personas llegaron a **urgencias pediátricas o infantiles** ese día. Es otra de las variables a predecir.
- Cuántas personas llegaron a **urgencias ginecológicas u obstétricas** ese día. Es otra de las variables a predecir.
- La **media** de los datos proporcionados: edad media de los pacientes del día, número medio de diagnósticos, etc.
- El **calendario**. Añadimos información externa sobre qué día de la semana era, qué mes o si era fin de semana o festivo.
- Las **búsquedas que se registraron en Google** esos días de ciertos términos. Es importante notar que estos términos son **programables**. En esta prueba de concepto se han elegido términos como fiebre, gripe, covid, etc.

4. Datos de entrada a los modelos

Un modelo **aprende patrones de los datos pasados** para predecir cuál será el valor futuro. Para poder hacer la predicción es importante trabajar con “**ventanas**”: después de X días como estos, suele venir un día así (que luego resultará en la predicción que haga).

Se han probado distintos valores de ventana:

- **4 semanas**. Ésta es la ventana que **mejores predicciones** proporciona.

- 5 semanas.
- 6 semanas.

También es importante que todos los datos que emplea el modelo estén en la **misma escala**, de forma que todas las variables tengan el mismo peso sobre el resultado. Esta **normalización** solo se aplica sobre datos de entrenamiento.

Las variables empleadas son **internas / endógenas** (las extraídas de la información proporcionada por el hospital) o **externas / exógenas** (las tendencias de Google, la información del calendario).

Dentro de las variables internas / endógenas, también se emplean **variables calculadas** (tiempo transcurrido en la urgencia, tiempo transcurrido en el triaje, agrupamiento por capítulos de los códigos CIE, etc.).

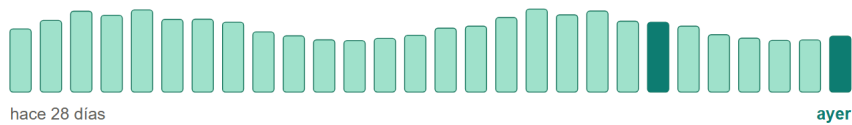
Modelos desarrollados

Con el fin de abordar el problema desde distintos paradigmas, se han implementado y comparado modelos de predicción pertenecientes a familias metodológicas heterogéneas, abarcando desde enfoques estadísticos clásicos de series temporales (SARIMAX) y modelos lineales generalizados (Poisson-GLM), hasta métodos de *ensemble* (Random Forest, XGBoost) y arquitecturas basadas en atención (Transformers). De todos estos, los que mejores resultados han proporcionado son:

- **XGBoost completo.** Arquitectura basada en **árboles**, que van generándose con la respuesta que se obtiene de una serie de preguntas encadenadas. A este modelo se le pasa la **ventana completa** de episodios y sus diagnósticos correspondientes.
- **XGBoost resumido.** En lugar de recibir todas las variables de entrada correspondientes a la ventana, recibe un **resumen**: el valor de todas las variables ayer, hace 7 días, el valor medio de cada variable en la última semana y el valor medio de cada variable en el último mes.
- **Transformer.** Arquitectura novedosa que suele captar bien los **ritmos** (las pautas semanales o las olas estacionales).
- **Transformer + Fiabilidad.** Además de proporcionar la predicción, se genera un **margen de confianza** de dicho valor.

Ventana de 28 días

Cada barra representa el número de episodios de urgencias de un día.



Transformer y «XGBoost completo»

Emplean los valores de todas las variables en los 28 días de la ventana.



28 valores de entrada

Captan mejor los patrones finos.

«XGBoost resumido»

Emplea un resumen de los datos incluidos en la ventana:

valor de ayer

hace 7 días

media de la última semana

media del último mes

4 valores de entrada

Pierde el patrón fino, pero al ser más simple, a veces también es más robusto.

3. Dataset

Es el visor de la serie diaria unificada (2022–2026). Se elige un rango de fechas y, al pulsar «Ver dataframe», se muestra día a día la tabla con todas las variables de entrada de los modelos: llegadas totales, desglose por área (Pediatría, Ginecología, Resto), ingresos, edad media, estancia, número de diagnósticos, capítulos CIE, etc.

Sirve para inspeccionar los datos y comprobar que los días cargados por la sección «Cargar datos» se han incorporado correctamente.

Predicción de afluencia a urgencias (agregada y por áreas asistenciales) y de ingresos

URGENCIAS · HOSPITAL PUERTA DE HIERRO

INICIO

PROCESADO

DATASET

URGENCIAS

ÁREA

INGRESOS

INFORME

CARGAR DATOS

Explorar el dataset diario

Serie diaria unificada (2022 a 2026) con las variables de entrada de los modelos. Elige un rango de fechas:

DESDE

01/01/2026

HASTA

30/06/2026

VER DATAFRAME

181 días · 2026-01-01 → 2026-06-30

date	episodes	episodes_pediatria	episodes_ginecologia	episodes_resto	ingresos	edad_med	los_hours	n_diagnostico	cie_chap_H60_95	cie_ch
2026-01-01	491.0	86.0	38.0	367.0	65.0	45.286	4.929	1.658	0.028	
2026-01-02	629.0	76.0	63.0	490.0	88.0	49.060	6.114	1.631	0.021	
2026-01-03	610.0	79.0	47.0	484.0	54.0	47.523	7.033	1.630	0.046	
2026-01-04	567.0	88.0	67.0	412.0	65.0	45.219	5.392	1.563	0.019	
2026-01-05	527.0	67.0	54.0	406.0	72.0	49.185	7.048	1.651	0.042	
2026-01-06	423.0	58.0	32.0	333.0	45.0	46.972	6.048	1.728	0.024	
2026-01-07	554.0	57.0	63.0	434.0	48.0	48.198	5.380	1.690	0.034	
2026-01-08	475.0	54.0	51.0	370.0	50.0	47.393	5.663	1.619	0.032	
2026-01-09	564.0	62.0	60.0	442.0	60.0	49.927	5.032	1.754	0.034	
2026-01-10	527.0	90.0	38.0	399.0	51.0	45.572	5.375	1.584	0.030	
2026-01-11	530.0	104.0	33.0	393.0	51.0	44.970	5.435	1.681	0.062	
2026-01-12	585.0	103.0	62.0	420.0	64.0	44.907	5.030	1.838	0.029	
2026-01-13	550.0	71.0	50.0	429.0	69.0	48.163	5.838	1.693	0.047	
2026-01-14	497.0	65.0	46.0	386.0	54.0	47.980	6.481	1.771	0.032	
2026-01-15	491.0	71.0	48.0	372.0	58.0	46.816	5.468	1.813	0.035	

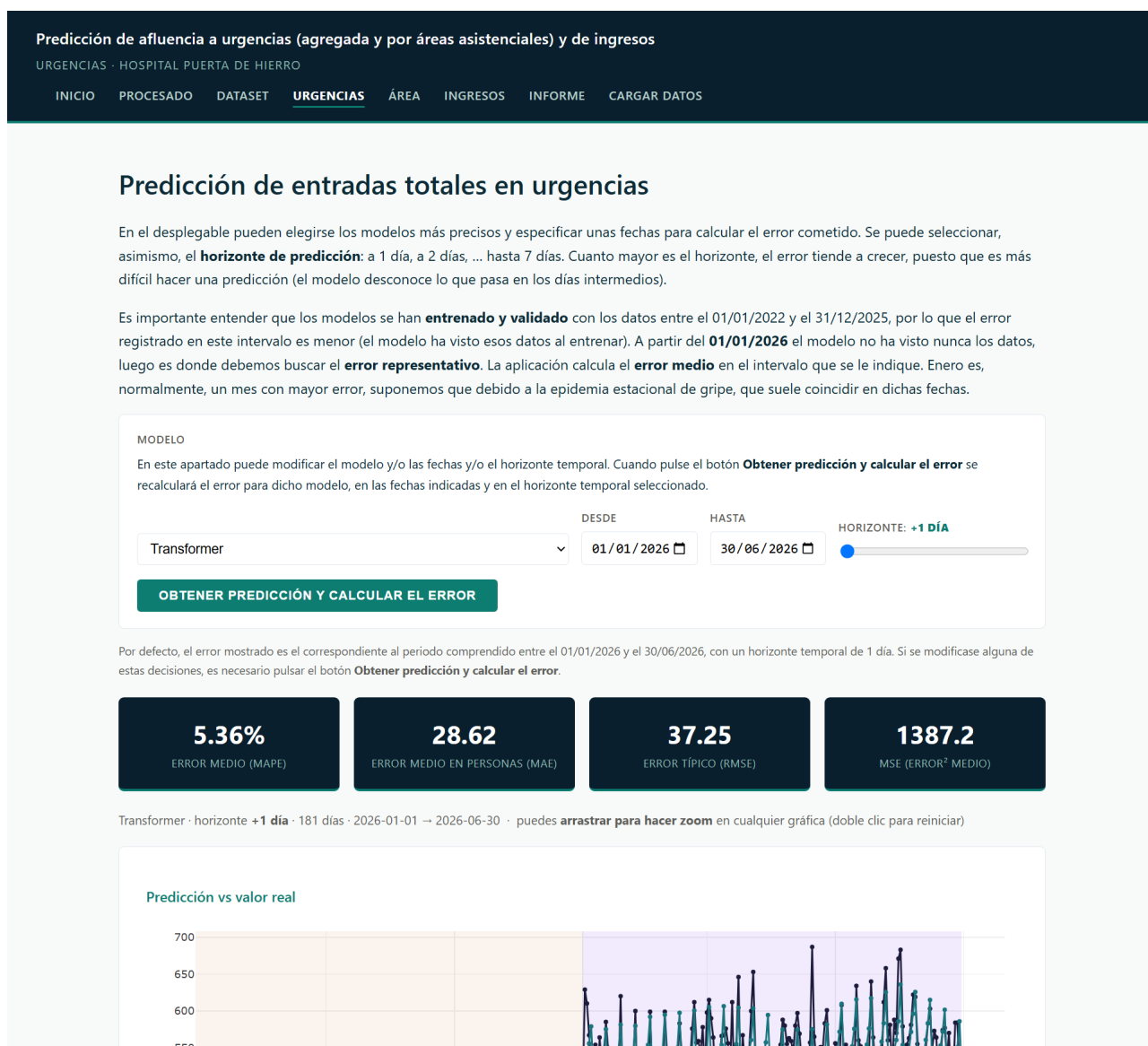
Hospital Universitario Puerta de Hierro Majadahonda · Madrid · 778.472 episodios (2022–2026) · Alhambra — Universidad de Alcalá · v2

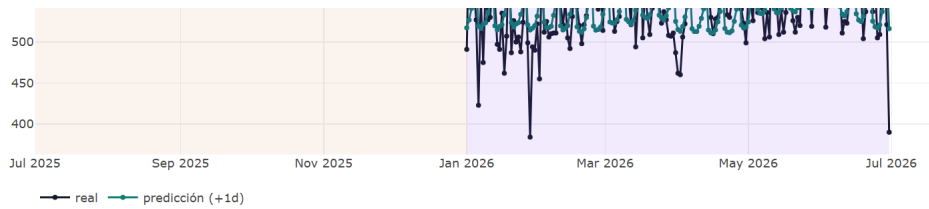
4. Urgencias

Es la sección central para evaluar el modelo de llegadas totales. Permite elegir el modelo (Transformer, XGBoost de ventana completa, XGBoost resumido o Transformer + fiabilidad), el rango de fechas y el horizonte de predicción (de +1 a +7 días). Al pulsar «Obtener predicción y calcular el error» los modelos se ejecutan en vivo sobre el periodo indicado.

Muestra las métricas de error (MAPE, MAE, RMSE y MSE) y varias gráficas interactivas: predicción frente a valor real, error con signo (en naranja los días en que se predijo de menos y en verde los de más) y las series diarias de MAPE, MAE y RMSE. Al pie se recogen las fórmulas de cada métrica.

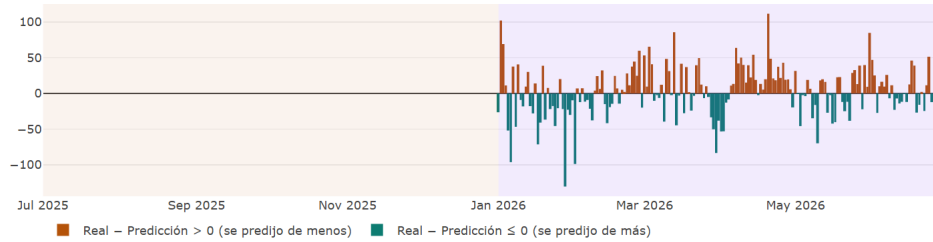
Conviene recordar que el error es menor dentro del periodo de entrenamiento (2022–2025, que el modelo sí ha visto) y que el error representativo es el de 2026 en adelante, datos que el modelo no conocía; enero suele ser el mes de mayor error por la epidemia estacional de gripe.





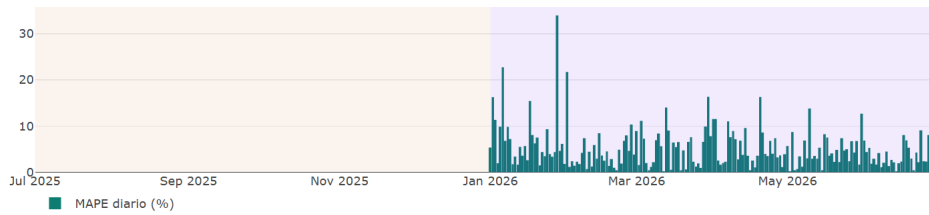
Error con signo

Aquí se muestra la **diferencia diaria entre la predicción y el valor real**. Un valor **positivo (mostrado en color naranja)** quiere decir que la predicción se quedó por debajo del valor real (habré planificado menor personal que el que habría sido necesario). Un valor **negativo (mostrado en color verde)** quiere decir que la predicción fue excesiva.

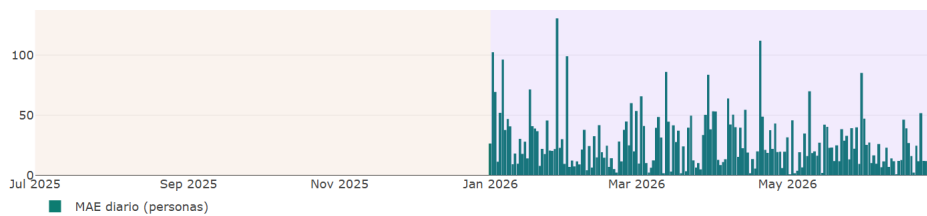


Error diario (magnitud)

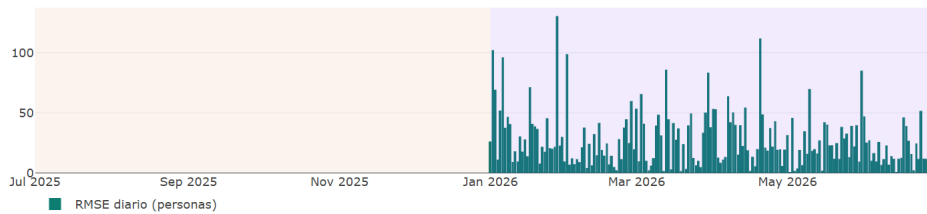
$$MAPE_{día} = \frac{|predicción - real|}{real} \times 100 (\%)$$



$$MAE_{día} = |predicción - real| \text{ (personas)}$$



$$RMSE_{\text{día}} = \sqrt{(predicción - real)^2} \text{ (personas)}$$



¿Cómo se calculan? Sobre los días del rango elegido:

- **Error con signo:** positivo = la predicción se quedó por debajo del valor real; negativo = la predicción fue excesiva.

$$error = real - predicción$$

- **MAPE:** error medio en **porcentaje**.

$$MAPE = media\left(\frac{|predicción - real|}{real}\right) \times 100 \text{ (\%)}$$

- **MAE:** error medio en **personas**.

$$MAE = media(|predicción - real|)$$

- **MSE:** penaliza mucho los fallos grandes (unidades: personas²).

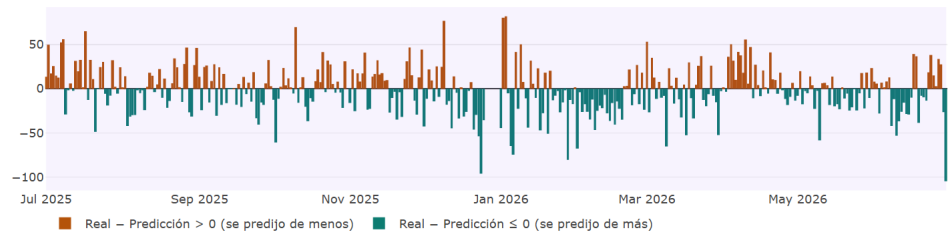
$$MSE = media((predicción - real)^2)$$

- **RMSE:** vuelve a "personas"; es el error típico.

$$RMSE = \sqrt{MSE}$$

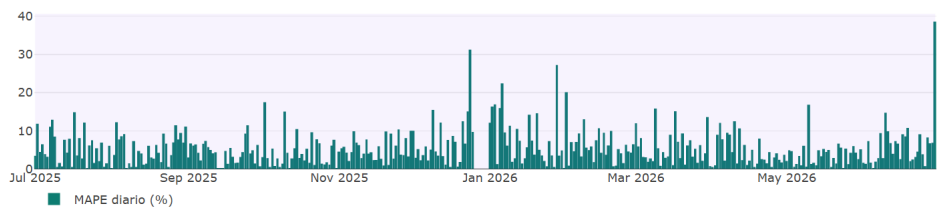
Error con signo

Aquí se muestra la **diferencia diaria entre la predicción y el valor real**. Un valor **positivo (mostrado en color naranja)** quiere decir que la predicción se quedó por debajo del valor real. Un valor **negativo (mostrado en color verde)** quiere decir que la predicción fue excesiva.

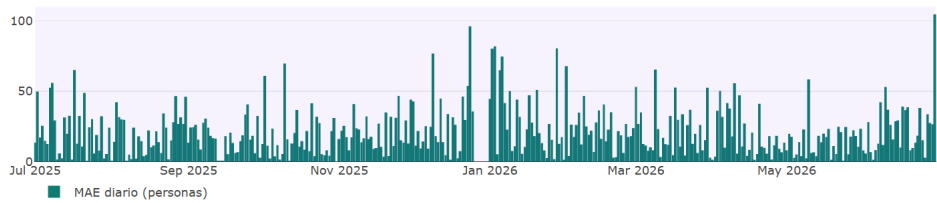


Error diario (magnitud)

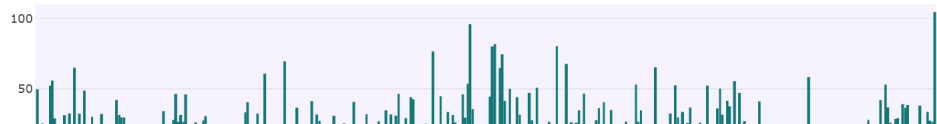
$$MAPE_{día} = \frac{|predicción - real|}{real} \times 100 (\%)$$



$$MAE_{día} = |predicción - real| \text{ (personas)}$$



$$RMSE_{día} = \sqrt{(predicción - real)^2} \text{ (personas)}$$





Predicción según margen de seguridad

Esta predicción permite ajustar el **margen de seguridad** para minimizar el número de días que el modelo se queda corto (hará predicciones en exceso):

$$\text{predicción} = P50 + \alpha \cdot (P90 - P50)$$

$\alpha = 1.00$

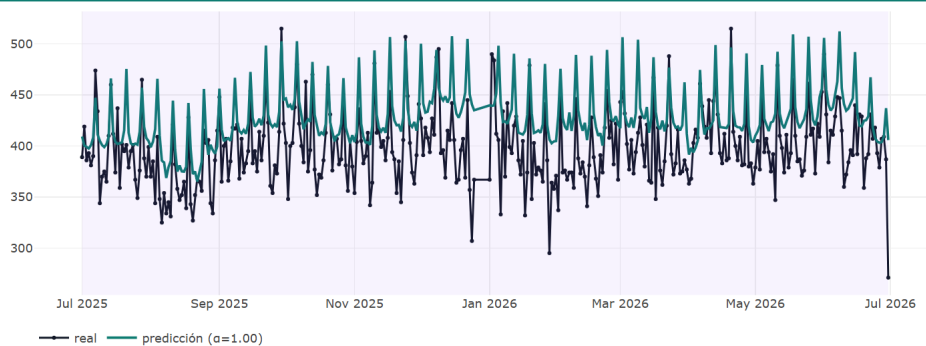


9.05%

ERROR (MAPE) CON ESTE α

11%

% DÍAS QUE TE QUEDAS CORTO



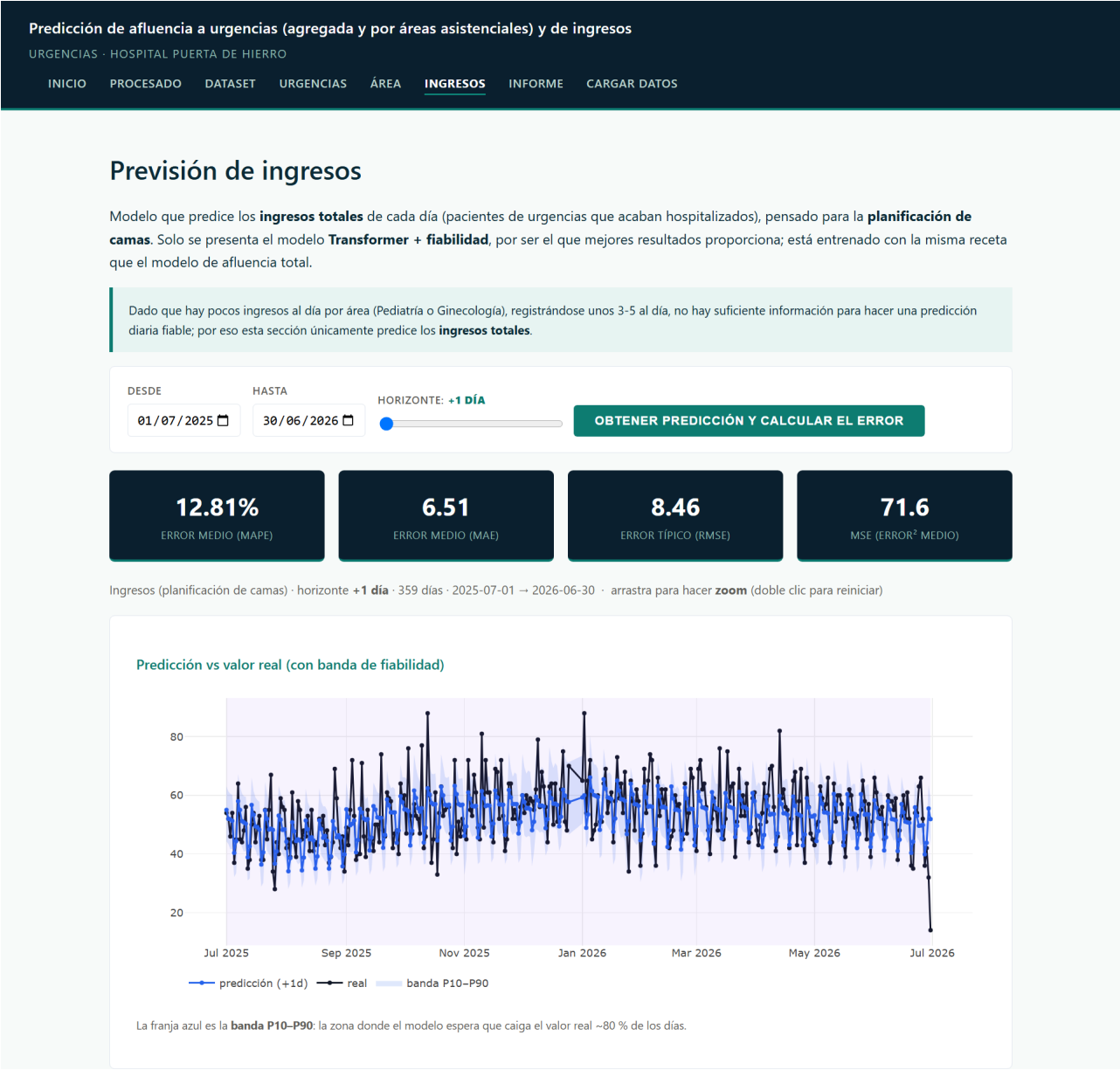
Con $\alpha = 0$ la predicción coincide con la predicción central del modelo Transformer + fiabilidad de la gráfica "Predicción vs valor real" de la parte superior.

Las áreas de bajo volumen (Gine-Obstetricia ~40/día) tienen un error **relativo** (MAPE) mayor que el total, aunque su error absoluto sea pequeño.

6. Ingresos

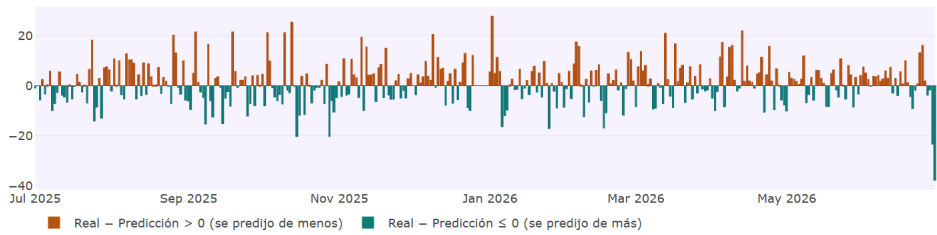
Ofrece la previsión de ingresos diarios —pacientes de urgencias que acaban hospitalizados—, pensada para la planificación de camas. Se presenta el modelo Transformer + fiabilidad, con su banda P10–P90, su horizonte y sus métricas, igual que en las secciones anteriores.

Se modela el ingreso total (no por área): los ingresos de Pediatría o Ginecología son demasiado escasos (3–5 al día) para una predicción diaria fiable.



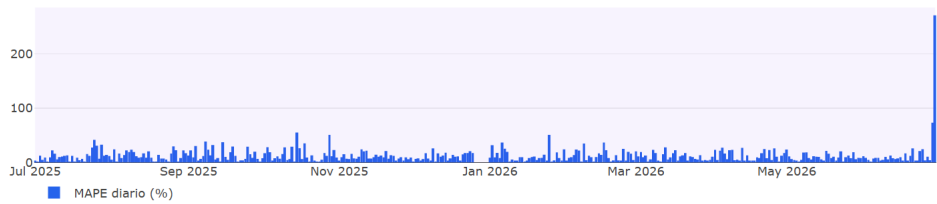
Error con signo

Aquí se muestra la **diferencia diaria entre la predicción y el valor real**. Un valor **positivo (mostrado en color naranja)** quiere decir que la predicción se quedó por debajo del valor real. Un valor **negativo (mostrado en color verde)** quiere decir que la predicción fue excesiva.

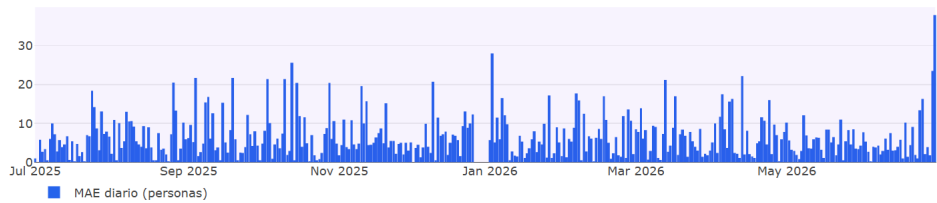


Error diario (magnitud)

$$MAPE_{día} = \frac{|predicción - real|}{real} \times 100 \text{ (\%)}$$

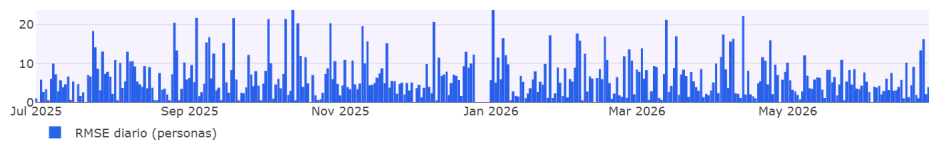


$$MAE_{día} = |predicción - real| \text{ (personas)}$$



$$RMSE_{día} = \sqrt{(predicción - real)^2} \text{ (personas)}$$





Predicción según margen de seguridad

Esta predicción permite ajustar el **margen de seguridad** para minimizar el número de días que el modelo se queda corto (hará predicciones en exceso):

$$\text{predicción} = P50 + \alpha \cdot (P90 - P50)$$

$\alpha = 1.00$

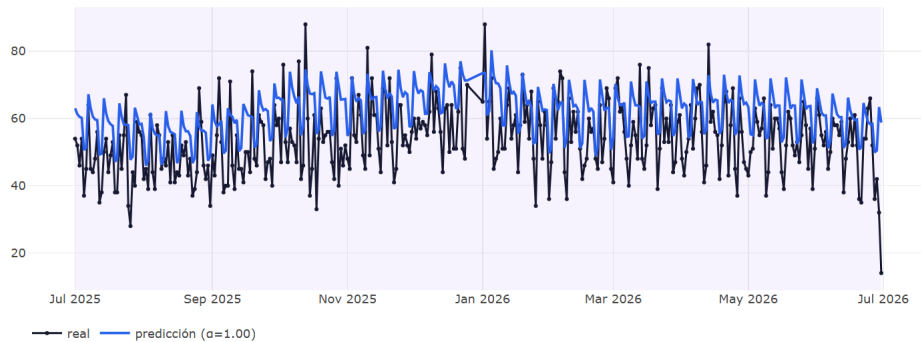


20.75%

ERROR (MAPE) CON ESTE α

11%

% DÍAS QUE TE QUEDAS CORTO



Con $\alpha = 0$ la predicción coincide con la predicción central del modelo Transformer + fiabilidad de la gráfica "Predicción vs valor real" de la parte superior.

7. Informe

Genera el parte de previsión a 7 días a partir de una fecha de emisión, con pronóstico hacia delante: predice la semana siguiente a esa fecha. Reúne en un solo documento las llegadas totales, el desglose por área asistencial y los ingresos esperados, cada uno con su banda de fiabilidad P10–P90.

Los cinco primeros días se agrupan en un recuadro del color de cada sección (verde, azul y naranja) y los dos últimos quedan aparte, marcados como opcionales. El margen de seguridad α es ajustable con la barra o el número: con $\alpha = 0$ se muestra la predicción media (50 % de probabilidad de quedarse corto) y con $\alpha = 1$ la predicción garantista (solo un 10 %). El parte es imprimible (Ctrl+P) y descargable en PDF con la α elegida.

Predicción de afluencia a urgencias (agregada y por áreas asistenciales) y de ingresos

URGENCIAS · HOSPITAL PUERTA DE HIERRO

INICIO

PROCESADO

DATASET

URGENCIAS

ÁREA

INGRESOS

INFORME

CARGAR DATOS

Informe de Previsión

Parte de previsión a 7 días, generado con el modelo de fiabilidad para: llegadas totales, por área asistencial e ingresos en urgencias, cada uno con su banda de fiabilidad. Puede elegirse el margen de seguridad α y puede descargar el informe en formato PDF.

FECHA DE EMISIÓN

Margen de seguridad $\alpha = 1.00$ (techo P90)

24/06/2026

1,00

GENERAR

DESCARGAR PDF

$\alpha = 0$, predicción media (50 % de probabilidad de infraestimación). $\alpha = 1$, predicción garantista (10 % de probabilidad de infraestimación). El PDF se genera con la α elegida.

Pacientes en Urgencias

Hospital U. Puerta de Hierro Majadahonda · Madrid

Emitido el

24/06/2026

Previsión a 7 días generada con los modelos. Documento de apoyo a la planificación.

Banda de fiabilidad P10–P90 (~80 %). Cifra mostrada = $P50 + \alpha(P90 - P50)$ con $\alpha = 1.00$.

Llegadas totales esperadas

25/06/2026

Jueves

545

fiab. 478–545

26/06/2026

Viernes

544

fiab. 480–544

27/06/2026

Sábado

563

fiab. 506–563

28/06/2026

Domingo

602

fiab. 535–602

29/06/2026

Lunes

637

fiab. 563–637

DÍAS 6–7 · OPCIONALES

30/06/2026

Martes

574

fiab. 507–574

01/07/2026

Miércoles

551

fiab. 485–551

Llegadas por area asistencial

25/06/2026	Jueves	26/06/2026	Viernes	27/06/2026	Sábado	28/06/2026	Domingo	29/06/2026	Lunes
Resto	404	Resto	401	Resto	402	Resto	415	Resto	431
Pediatría	93	Pediatría	89	Pediatría	109	Pediatría	122	Pediatría	105
Gine-Obs	52	Gine-Obs	56	Gine-Obs	49	Gine-Obs	47	Gine-Obs	59

DÍAS 6-7 · OPCIONALES

30/06/2026	Martes	01/07/2026	Miércoles
Resto	406	Resto	399
Pediatría	89	Pediatría	84
Gine-Obs	55	Gine-Obs	60

Ingresos esperados (planificación de camas)

25/06/2026	Jueves	26/06/2026	Viernes	27/06/2026	Sábado	28/06/2026	Domingo	29/06/2026	Lunes
58		57		49		48		61	
fiab. 42-58		fiab. 41-57		fiab. 34-49		fiab. 33-48		fiab. 46-61	

DÍAS 6-7 · OPCIONALES

30/06/2026	Martes	01/07/2026	Miércoles
59		59	
fiab. 43-59		fiab. 41-59	

Generado por el sistema de predicción de demanda · Alhambra — Universidad de Alcalá. Cifras por área e ingresos de los modelos multitarget; no incluye desglose por turno ni por enfermedad.

8. Cargar datos

Permite ampliar la base de datos con información nueva del hospital y mantener los modelos al día. Está organizada en cinco pasos y una acción de seguridad:

1) Subir y verificar los ficheros episodios.sql y codificacion.sql (la aplicación detecta cuál es cada uno y comprueba el formato antes de añadir nada; solo amplía, nunca sobrescribe). 2) Ver el resultado de los modelos sobre la serie ampliada. 3) Reentrenar de verdad el modelo elegido —o todos— sobre el periodo indicado, actualizando antes Google Trends (tarda varios minutos). 4) Regenerar las predicciones guardadas que la aplicación muestra por defecto. 5) Actualizar Google Trends (re-descarga la ventana completa para que los índices no se queden congelados).

Además, «Restaurar de fábrica» devuelve tanto los datos como los modelos a su estado original, por si se ha cargado algo por error.

Predicción de afluencia a urgencias (agregada y por áreas asistenciales) y de ingresos

URGENCIAS - HOSPITAL PUERTA DE HIERRO

INICIO PROCESADO DATASET URGENCIAS ÁREA INGRESOS INFORME CARGAR DATOS

Cargar datos nuevos y ampliar la base de datos

En este apartado se pueden subir nuevos ficheros proporcionados por el hospital (episodios.sql y codificacion.sql). Antes de añadir nada, la aplicación **comprueba que tienen el formato correcto**, muestra qué días se añadirían y pide confirmación. Solo se permite **ampliar** (añadir días posteriores a los que ya hay); nunca se sobrescribe lo existente.

La base actual llega hasta **2026-06-30** (1642 días). Si te equivocas, abajo tienes el botón de **restaurar de fábrica**.

1 · Subir y verificar

FICHERO 1

EPISODIOS.SQL

EL REGISTRO DE **EPISODIOS**: UNO POR VISITA A URGENCIAS, CON SUS FECHAS DE LLEGADA Y ALTA, EDAD, SERVICIO...

FICHERO 2

CODIFICACION.SQL

LOS **CÓDIGOS CLÍNICOS** (CIE-10) ASIGNADOS A CADA EPISODIO: DIAGNÓSTICOS, ANTECEDENTES Y PROCEDIMIENTOS.

VERIFICAR FORMATO

Haz clic en una casilla (o arrastra el fichero encima) para elegirlo. **El orden no importa**: si los cruzas, la aplicación detecta cuál es cada uno por su contenido; las casillas son solo una guía visual.

2 · Ver el resultado sobre los datos nuevos

Abre la pestaña de **urgencias** y **muestra** el error de los modelos sobre la serie ampliada, calculándolo en vivo si hiciera falta. Solo **enseña** el resultado; no modifica la caché guardada.

VER RESULTADO

Tarda unos segundos. Al terminar, en la pestaña de **urgencias** puede cambiarse el modelo, las fechas o el horizonte.

3 · Reentrenar los modelos

Reentrena **de verdad** el modelo elegido (o todos) sobre el periodo de datos que indiques, y **actualiza Google Trends** antes de empezar. Al terminar regenera las predicciones guardadas con los pesos nuevos.

Es un **reentreno real**: puede tardar **varios minutos** (el transformer y la fiabilidad son los lentos) y requiere internet para los trends. No cierres la ventana mientras corre. Con **Restaurar de fábrica** vuelves a los checkpoints originales.

MODELO

Todos los modelos

ENTRENAR DESDE

01/01/2022

ENTRENAR HASTA

30/06/2026

REENTRENAR

4 · Regenerar los resultados guardados

Vuelve a ejecutar los **4 modelos** (no se entrenan, solo predican) sobre toda la serie a +1..+7 días y **guarda** el resultado en la caché que la aplicación muestra por defecto. Úsalo cuando hayas cambiado datos y quieras que las predicciones guardadas queden al día.

Tarda **~1-2 minutos**. El resultado es determinista: con los mismos datos y modelos sale siempre lo mismo.

REGENERAR PREDICCIONES GUARDADAS

5 · Actualizar Google Trends

Los modelos de afluencia usan 6 índices de **Google Trends** (gripe, fiebre, urgencias, covid, gastroenteritis, bronquiolitis) como señal adelantada. Este botón **los re-descarga** de la ventana completa (2022 → hoy) y regenera las predicciones afectadas, para que nunca se queden congelados en un valor viejo.

Requiere **conexión a internet** y tarda **~1-2 minutos** (descarga + regeneración). Google Trends es una fuente con límite de peticiones: si falla, reinténtalo en unos minutos.

ACTUALIZAR GOOGLE TRENDS AHORA

Restaurar de fábrica

Devuelve la base de datos a su estado original, **descartando todos los días que hayas añadido**. Úsalo si te has equivocado al cargar.

RESTAURAR DE FÁBRICA

Informe Experimental

Predicción de llegadas diarias a Urgencias

Informe experimental: modelos del estado del arte, análisis de la señal supervisora y validación prospectiva

GHEODE • Universidad de Alcalá • Septiembre 2026

Resumen

Se documentan, uno a uno, los experimentos realizados para predecir el número diario de llegadas a Urgencias (cohorte de Madrid, 778 472 episodios de 2022–2025 para entrenar y validar, más 98 727 episodios del primer semestre de 2026 reservados como *test prospectivo*; horizontes +1 a +7 días). Cada experimento se presenta con su *motivación*, *método*, *semillas* lanzadas, *resultados* y *conclusiones*. El informe se organiza en tres partes: la comparación de **modelos del estado del arte** (Parte I), el **análisis de variables** que explica de dónde proviene la capacidad predictiva (Parte II) y el **paso al despliegue** (Parte III): validación con pesos congelados sobre un semestre real posterior al entrenamiento, predicción con *intervalos de fiabilidad* calibrables y desagregación por área asistencial e ingresos. La conclusión transversal es que la señal útil se reduce a la autocorrelación de la serie más un factor de *estado epidémico*, captable de forma equivalente por búsquedas de Google Trends o por el *case-mix* interno del hospital; el resto de exógenas no aporta valor desplegable y el error topa en ~5,3 % de MAPE, un techo que el test prospectivo confirma (5,2–5,5 % en 2026-H1 para cuatro modelos distintos) y que coincide con el suelo de ruido del propio dato. El error de cada día no es predecible como cantidad puntual ($r = 0,16$), pero sí *acotable*: una banda P10–P90 cubre el 77 % de los días en validación y el 71 % en el semestre prospectivo, y una perilla α traslada la tasa de infraestimación elegida de validación a test con un desvío de ± 3 puntos.

1 Datos y protocolo común

Datos. 778 472 episodios de Urgencias (2022–2025) agregados a una serie diaria contigua de 1 461 días (media ≈ 533 llegadas/día; estacionalidad semanal y anual). Hay tres familias de predictores: (i) *autorregresivos* (historial de la propia serie), (ii) *calendario* (día de la semana, mes, festivos), y dos fuentes de *contexto*: (iii) *variables de la casa* o *case-mix* (edad, estancia, nº de códigos y diagnósticos, 24 capítulos CIE...) y (iv) *exógenas externas* (clima ERA5, calidad del aire y polen, vigilancia gripal ERVISS, Google Trends; 85 columnas).

Datos prospectivos (Parte III). Tras congelar los modelos se recibieron 98 727 episodios del primer semestre de 2026 (181 días), primero enero y después el semestre completo, que extienden la serie a 1 642 días (2022-01-01 $\rightarrow$ 2026-06-30). Ese semestre no interviene en ningún ajuste ni selección: es un test prospectivo genuino.

Protocolo A (Partes I y II; idéntico en todos sus experimentos). *Backtest de origen rodante*: en cada día-origen se reajusta el modelo con todo el histórico disponible y se predicen los 7 días siguientes. Partición temporal por la fecha del primer día pronosticado: entrenamiento ($\leq 2024-12-31$), validación (2025-H1) y **test (2025-H2)**, que no se usa hasta el final. Métricas MAE, RMSE y MAPE por horizonte.

Protocolo B (Parte III, despliegue). A diferencia del backtest, en despliegue los pesos están *congelados*: se entrena una vez con datos anteriores a 2025-07-01, se selecciona por validación en 2025-H2 (el antiguo test) y se mide en **2026-H1** sin reajuste alguno. Es el régimen real de un sistema en producción y una prueba más dura que el backtest. Las cifras de la Parte III se recalculan con

el pipeline final (la caché que sirve la aplicación: 184 días de validación y 181 de test) y pueden diferir en $\leq 0,3$ puntos de las obtenidas en el momento de la selección.

Multi-semilla y banda de ruido. El transformer es estocástico; su varianza de semilla ($\pm 0,3$ puntos de MAPE) supera a muchas de las diferencias estudiadas. Por eso todo se reporta como media $\pm$ desviación sobre varias semillas, y se define una *banda de ruido* = error estándar de la media ($sd/\sqrt{n}$): solo se da por real un efecto que la supere.

Anti-fuga. Escalado y winsorización calculados *solo con train*; los predictores autorregresivos y de case-mix entran únicamente con retardos del pasado; el calendario futuro (determinista) entra como covariable del decoder; las banderas binarias no se estandarizan.

Calibración de la ventana de entrada. Antes de los experimentos se fijó la longitud de la ventana del transformer probando 28/32/35/42 días (3 semillas). Gana **28 días** (4 semanas exactas): múltiplo entero del ciclo semanal; ventanas mayores diluyen.

Ventana	28 (4×7)	32	35 (5×7)	42 (6×7)
MAPE +7d	5,65	6,06	5,87	5,85

2 Parte I – Modelos del estado del arte

Esta parte establece *cuánto se puede predecir* con las principales familias de modelos de la literatura (GLM, árboles/boosting, descomposición, deep learning), todas bajo el protocolo común. La pregunta de *qué variables* importan se aborda en la Parte II.

2.1 Experimento 1 – SOTA clásico/ML con todas las variables de la casa

Por qué. La revisión del estado del arte prioriza GLM, gradient boosting y random forest, y sugiere que el dato interno del hospital debería bastar. Queremos su rendimiento usando *toda* la información de la casa.

Cómo. Tres modelos –GLM de Poisson (enlace log, familia natural para conteos), Random Forest y XGBoost (objetivo `count:poisson`)– con features: calendario + autorregresivo + las 34 variables de case-mix (valor del día y media móvil de 7 días). Backtest rodante a 7 días.

Semillas. Poisson es *determinista* (1 corrida, exacta). RF y XGBoost dependen de `random_state`; se ejecutan con **10 semillas** y se reporta media $\pm$ sd.

Resultados.

Modelo	semillas	MAE	RMSE	MAPE +1d	MAPE +7d
XGBoost + casa	10	29,07 $\pm$ 0,07	37,50	4,81 $\pm$ 0,03	5,89 $\pm$ 0,06
Poisson GLM + casa	1	31,40	39,39	5,18	6,41
Random Forest + casa	10	32,02 $\pm$ 0,08	40,74	5,39 $\pm$ 0,02	6,48 $\pm$ 0,03

Conclusión. XGBoost es el mejor de los tres y, a +1 día (4,81 %), iguala al mejor modelo global; además es *extremadamente estable* entre semillas ($\pm 0,03$, frente a $\pm 0,2$ – $0,3$ del transformer), porque el promedio de $\sim 1\,000$ reajustes del backtest diluye la aleatoriedad de los árboles. Todos se degradan a +7 días: al predecir de forma *directa* (sin realimentar sus salidas) pierden fuelle en horizonte largo.

2.2 Experimento 2 – Modelos de referencia y comparación de familias

Por qué. Situar a los modelos del SOTA frente al transformer encoder-decoder (producción), Prophet, un baseline ingenuo estacional y el *modelo propio del hospital* (el incumbente a batir).

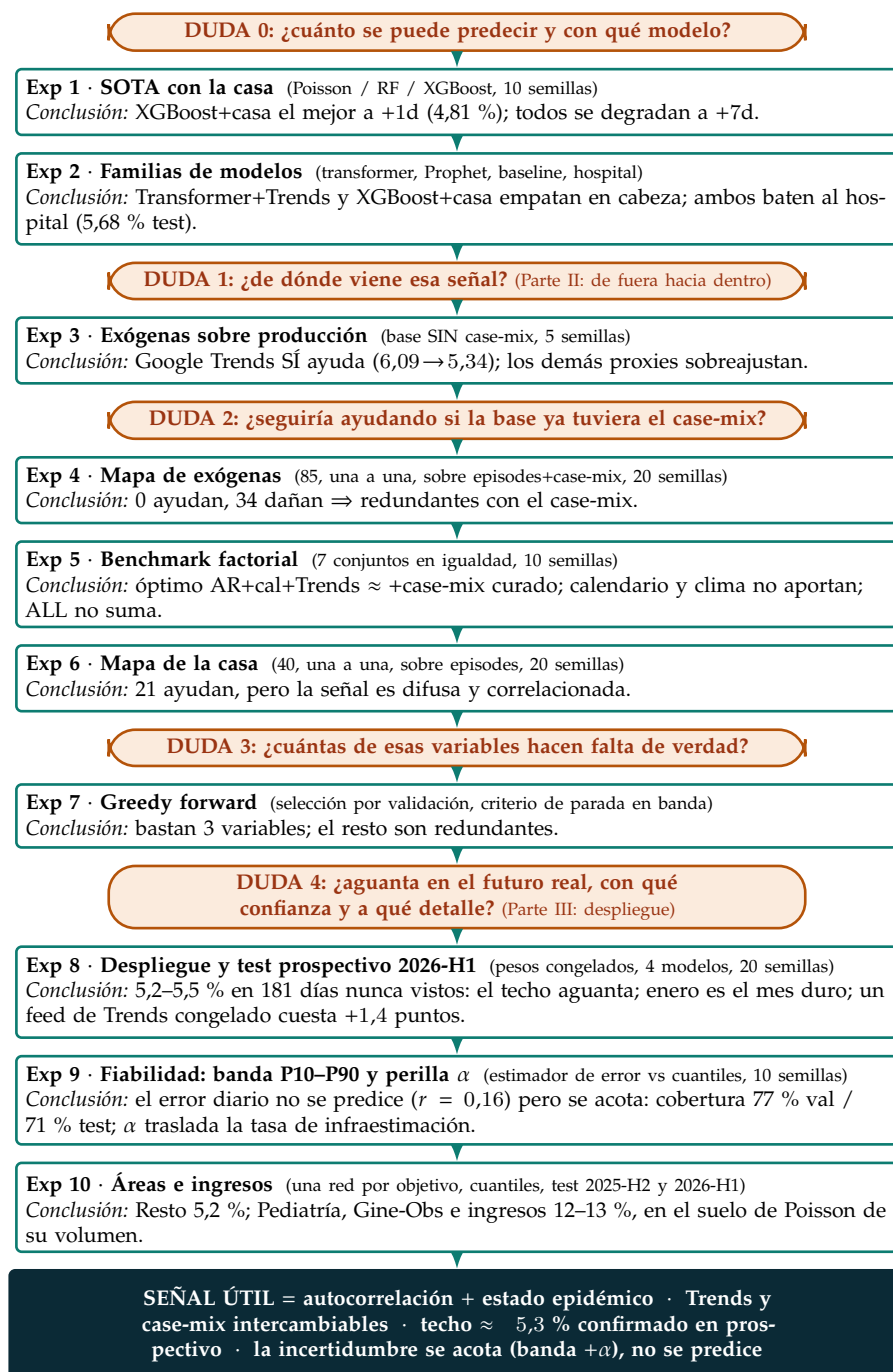


Figura 1: **Mapa lógico del estudio.** Cada **duda** (naranja) motiva un **experimento** (verde) cuya conclusión abre la siguiente: primero *cuánto* se puede predecir (Parte I), luego *de dónde* sale la capacidad predictiva (Parte II) y por último si *aguanta en despliegue*, con qué confianza y a qué detalle (Parte III).

Cómo. Mismo backtest. El transformer es autorregresivo con calendario futuro conocido; Prophet modela tendencia + estacionalidad + festivos. El modelo del hospital se evalúa con sus predicciones archivadas.

Semillas. Transformer: 10–20 (media $\pm$ sd). Prophet y baseline: deterministas. Hospital: dato observado.

Resultados (consolidado; las filas marcadas † se derivan en la Parte II).

Modelo	MAPE +1d	MAPE +7d
Transformer + Trends [†]	4,79 $\pm$ 0,21	5,34 $\pm$ 0,31
XGBoost + casa	4,81 $\pm$ 0,03	5,89 $\pm$ 0,06
Transformer + case-mix curado [†]	4,90 $\pm$ 0,12	5,35 $\pm$ 0,20
Prophet	$\approx$ 5,0	$\approx$ 5,3
Poisson GLM + casa	5,18	6,41
Random Forest + casa	5,39 $\pm$ 0,02	6,48 $\pm$ 0,03
Transformer producción	5,33 $\pm$ 0,37	5,89 $\pm$ 0,34
Modelo del hospital	—	$\approx$ 5,7 (test)

Conclusión. Dos rutas empatan en cabeza: *Transformer+Trends* (mejor a +7d, 5,34 %) y *XGBoost+casa* (mejor a +1d, 4,81 %). Ambas baten al modelo del hospital, que en el mismo periodo de test (2025-H2) hace 5,68 % (MAE 30,6); su $\sim$ 9–10 % de otras épocas no es comparable y no debe usarse para inflar el margen. El mejor modelo, su RMSE incluido, es el *Transformer+Trends*: a +1d MAE 26,0 $\pm$ 1,3, RMSE 33,7 $\pm$ 1,6, MAPE 4,8 %.

3 Parte II – Análisis de variables: ¿de dónde viene la señal?

Establecido el rendimiento, dissectamos *qué* predictores aportan. El hilo lógico va de fuera hacia dentro: primero si las exógenas externas sirven (Exp. 3–4), luego qué combinación es óptima (Exp. 5), y por último qué variables internas son la señal y cuántas hacen falta (Exp. 6–7).

3.1 Experimento 3 – ¿Mejoran las exógenas al modelo de producción?

Por qué. La literatura post-COVID sugiere que proxies (gripe, búsquedas) adaptan los modelos al *concept drift*. Lo probamos sobre la base de producción (episodos + clima + calendario), que *no* lleva case-mix.

Cómo. Ablación: se añade cada bloque exógeno entero (Trends, gripe ERVISS, calendario extendido...) y se reentrena el transformer completo.

Semillas. 5 (media $\pm$ sd). Durante este experimento se detectaron y corrigieron dos artefactos: banderas binarias estandarizadas que inyectaban picos $z \approx 10$, y extrapolación de test fuera del soporte de train (mitigada con winsorización).

Resultados. El bloque de Google Trends baja el MAPE medio de 5,61 % a 5,08 % (a +7d: 6,09 $\rightarrow$ 5,34). El resto de bloques empeoran la red por sobreajuste.

Conclusión. Sobre una base sin case-mix, *Google Trends sí ayuda* (aporta el estado epidémico ausente); los demás proxies sobreajustan. Esto deja una pregunta clave para el Exp. 4: ¿seguiría ayudando si la base ya tuviera el case-mix del hospital?

3.2 Experimento 4 – Mapa de exógenas: efecto individual de las 85

Por qué. Medir, con escrutinio fuerte, el efecto de *cada* exógena externa sobre una base que ya incluye el case-mix.

Cómo. Se añade cada una de las 85 exógenas, de una en una, sobre episodios + case-mix, con sus retardos anti-fuga (gripe +14d, Trends +7d, aire +1d). Influencia = MAPE(con) – MAPE(base).

Semillas. 20 (banda $\pm 0,06$). Base: MAPE +1d 5,25 % / +7d 5,64 %.

Resultados. De las 85: **0 ayudan** por encima del ruido y **34 dañan** (overfitting medible). La mejor (trend_fiebre, +0,04) es estadísticamente cero.

Conclusión. Ninguna exógena externa aporta cuando el modelo ya tiene el case-mix interno. La razón es redundancia: el case-mix ya contiene la señal epidémica que los proxies aportarían. Esto reconcilia con el Exp. 3: Trends solo ayuda en ausencia de case-mix.

3.3 Experimento 5 – Benchmark factorial de conjuntos de variables

Por qué. Los experimentos previos usaban bases distintas; aquí se comparan conjuntos de variables en igualdad de condiciones (mismo transformer, mismo protocolo) para decidir la combinación óptima.

Cómo. Siete recetas de entrada, variando solo el bloque añadido sobre la autorregresión (AR) y el calendario.

Semillas. 10 (banda $\pm 0,08$ a +1d, $\pm 0,10$ a +7d).

Resultados.

Conjunto	MAPE +1d	MAPE +7d
AR (solo episodios)	5,33 $\pm$ 0,37	5,89 $\pm$ 0,34
AR + calendario	5,33 $\pm$ 0,37	5,89 $\pm$ 0,34
AR + cal + clima	5,40 $\pm$ 0,49	6,21 $\pm$ 0,56
AR + cal + case-mix curado	4,90 $\pm$ 0,12	5,35 $\pm$ 0,20
AR + cal + case-mix (34)	5,24 $\pm$ 0,26	5,64 $\pm$ 0,32
AR + cal + Trends	4,79 $\pm$ 0,21	5,34 $\pm$ 0,31
ALL (todo junto)	5,02 $\pm$ 0,17	5,35 $\pm$ 0,26

Conclusión. (1) Óptimo: AR+cal+Trends, empatado con case-mix curado (5,34 vs 5,35). (2) El calendario explícito no aporta sobre AR (la ventana de 28d ya lleva el ciclo semanal). (3) El clima estorba. (4) Volcar las 34 de casa (5,64) es peor que 6 curadas (5,35): hay que curar. (5) Juntarlo todo (ALL) no suma: Trends y case-mix son redundantes.

Tres formas de medir una variable (y qué concluye cada una)

Los Exp. 6 y 7 usan dos métodos complementarios; un tercero (*leave-one-out*) es su espejo. El esquema aclara qué pregunta responde cada uno y por qué hacen falta los dos: el mapa individual ve *quién tiene señal*, pero solo el greedy (o el LOO) ve *quién tiene señal que los demás no aportan ya*.

3.4 Experimento 6 – Mapa de la casa: señal individual interna

Por qué. Identificar qué variables del propio hospital llevan la señal.

Cómo. Se añade cada una de las 40 variables internas, de una en una, sobre episodios.

Semillas. 20 (banda $\pm 0,08$). Base: +1d 5,18 / +7d 5,89.

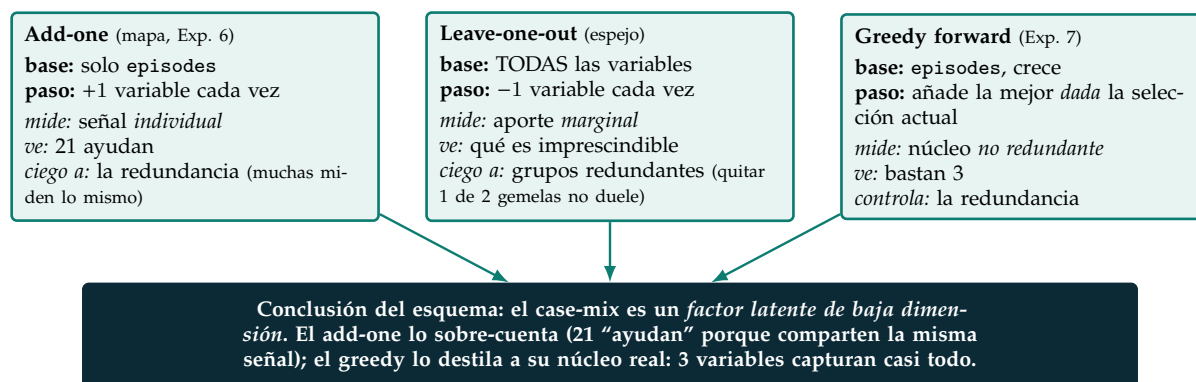


Figura 2: **Métodos de análisis de variables.** Add-one y leave-one-out son duales (construir vs. destruir); ambos son *ciegos a la redundancia* entre variables correlacionadas. El greedy la controla por construcción, y por eso da el conjunto mínimo. Misma conclusión por dos caminos = robusta.

Resultados. 21 variables ayudan individualmente, 6 neutras, 13 dañan. Lideran edad_num (+0,53 a +7d) y n_diagnostico (+0,50), proxies de carga/gravedad del día; siguen capítulos CIE de gran volumen (obstetricia, oído, traumatismos, infecciosas). El valor se concentra a +7d: a +1d casi todas son neutras o negativas.

Conclusión. El case-mix lleva señal *real pero difusa*: muchas variables puntúan parecido porque están correlacionadas (un único factor de “carga reciente del servicio”). El ranking individual no distingue señal de redundancia; eso lo resuelve el Exp. 7.

3.5 Experimento 7 – Greedy forward: núcleo mínimo no redundante

Por qué. Hallar el conjunto *mínimo* de variables de casa que captura la señal, eliminando la redundancia que el mapa individual no ve.

Cómo. Selección hacia delante: partiendo de episodios, en cada paso se añade la variable que más baja el MAPE de *validación por encima* de las ya elegidas; se para cuando la mejora no supera la banda de ruido. La selección por validación evita ajustar al test.

Semillas. 3 por candidato; criterio de parada en banda $\pm 0,06$.

Resultados.

Paso	Variable añadida	val +7d	test +7d
0	(solo episodios)	5,89	6,03
1	cie_chap_N (genitourinario)	5,73	5,50
2	cie_chap_H60_95 (oído)	5,64	5,55
3	edad_num (edad media)	5,55	5,72
4	— (parada: mejora < banda)	—	—

Conclusión. Bastan 3 variables para capturar casi toda la señal del case-mix: de las 21 que ayudaban sueltas, el resto son redundantes. El case-mix es, por tanto, un factor latente de baja dimensión.

4 Parte III – Despliegue: validación prospectiva y fiabilidad

Las Partes I y II responden *cuánto* se puede predecir y *con qué*. Queda la pregunta que decide si el sistema sirve: ¿aguanta con pesos congelados sobre un futuro que no ha visto, sabe decir cuánto fiarse de cada cifra, y baja al detalle que el hospital necesita (áreas, camas)? Esta parte lo prueba sobre el primer semestre de 2026, recibido *después* de fijar los modelos.

4.1 Experimento 8 – Despliegue con pesos congelados y validación prospectiva (2026-H1)

Por qué. El backtest rodante reajusta el modelo en cada origen; un sistema en producción no puede. Hay que medir la degradación real de unos pesos fijos sobre datos posteriores al entrenamiento y comprobar que las conclusiones de la Parte II (qué variables) sobreviven fuera de muestra.

Cómo. Se fija el conjunto de variables que la Parte II señaló como útil: autocorrelación (ventana de 28 días) + *case-mix* útil (7 variables, unión del greedy y del curado: edad media, estancia, n.º de diagnósticos y capítulos CIE H60–95, P, N y OTHER) + calendario (4) + Google Trends (6 índices, retardo 7 d, winsorizados con train). Con *esa misma* información se entrenan tres modelos puntuales: el transformer de producción (autorregresivo, ventana de 28 d), *XGB-flat* (la misma ventana de 28 d aplanada como vector) y *XGB-sum* (retardos + media móvil de 7 d), más el P50 del modelo de cuantiles del Exp. 9. Nuevo split: train < 2025-07-01 (antiguo train + val), validación = 2025-H2 (antiguo test) y **test prospectivo = 2026-H1** (181 días), recibido en dos envíos tras congelar los pesos: primero enero (19 días útiles) y después el semestre completo. Nada de 2026 interviene en el ajuste ni en la selección.

Semillas. 20 por modelo; se despliega la mejor por MAPE de validación (media $\pm$ sd a +1/+7 d: transformer 4,99 $\pm$ 0,19 / 5,30 $\pm$ 0,14; XGB-flat 4,82 $\pm$ 0,04 / 5,65 $\pm$ 0,05; XGB-sum 5,13 $\pm$ 0,04 / 5,69 $\pm$ 0,05). Se probó además sin calendario (20 semillas más por modelo); se retuvo la configuración con calendario.

Resultados. MAPE por horizonte del modelo desplegado (mejor semilla), en validación y en el semestre prospectivo:

	Modelo	+1	+2	+3	+4	+5	+6	+7	media	MAE +1
val 25-H2	Transformer	5,06	4,91	4,81	4,91	4,94	4,99	5,06	4,95	27,2
	XGB-flat	4,99	5,19	5,30	5,48	5,48	5,47	5,59	5,36	26,7
	XGB-sum	5,25	5,34	5,34	5,46	5,57	5,52	5,54	5,43	28,2
	Fiabilidad (P50)	4,82	4,75	4,71	4,85	4,89	4,92	4,97	4,84	25,8
test 26-H1	Transformer	5,36	5,23	5,34	5,38	5,49	5,57	5,64	5,43	28,6
	XGB-flat	5,40	5,46	5,42	5,53	5,49	5,48	5,45	5,46	29,2
	XGB-sum	4,86	5,11	5,22	5,30	5,26	5,27	5,28	5,19	26,1
	Fiabilidad (P50)	5,23	5,28	5,41	5,50	5,54	5,73	5,74	5,49	28,0

1. *Validación.* Los dos modelos autorregresivos (transformer y P50) lideran (4,95 y 4,84 de media) y son *planos* en horizonte: al realimentar sus salidas con el calendario futuro apenas se degradan de +1 a +7 d. Los dos XGBoost, que predicen cada horizonte de forma directa, pierden 0,3–0,6 puntos entre +1 y +7 d, como en el Exp. 1.
2. *Test prospectivo.* Los cuatro modelos caen en la franja **5,2–5,5 %** (medias 5,19–5,49) sobre 181 días que ningún modelo había visto: **el techo de la Parte II se confirma fuera de muestra**. El orden se baraja (XGB-sum es nominalmente el mejor, 4,86 a +1 d) pero la dispersión entre modelos (0,3 puntos) es del tamaño de la banda de semilla: no hay un ganador único. La degradación val $\rightarrow$ test es de +0,5 puntos para transformer y P50 y de $\pm$ 0,2 para los XGBoost.
3. *Horizonte.* El error crece con la antelación de forma moderada (transformer: MAE 28,6 $\rightarrow$ 29,7 y MAPE 5,36 $\rightarrow$ 5,64 de +1 a +7 d), con la convergencia clásica de los pronósticos sucesivos hacia el valor real a medida que se acerca el origen (figura del anexo).

4. *Enero, el mes duro.* Los 19 primeros días de 2026 dieron 7,4–8,0 % (P50 7,40; transformer 7,53; XGB-sum 7,82; XGB-flat 7,98; MAE 39–42): rebotes posfestivos y ola gripal. Sobre el semestre completo el efecto se diluye hasta la franja de 5,2–5,5 %. Es el error representativo de un mes de invierno, no el del modelo.

Incidente y lección: el feed de Trends. En enero Google devolvió 429 y los índices se persistieron desde el último valor conocido (marcado como tal). Al integrar el semestre completo, la carga incremental siguió arrastrando ese último valor: entre 185 y 248 días *planos* en 2026-H1 para los seis índices, en silencio. Los modelos de afluencia usan Trends (los del Exp. 10 no). Con el feed congelado el MAPE del P50 en 2026-H1 era **6,69 %**; tras re-descargar la ventana completa 2022 → 2026-06 (obligatoriamente entera: Trends normaliza 0–100 al máximo del rango pedido; el histórico 2022–2025 no cambió, ratio $\approx 1,0$) bajó a **5,25 %** (transformer: MAE 32,6 → 28,7; XGB-sum 29,2 → 26,4). El feed rancio costó +1,4 puntos, más que toda la ganancia que Trends aporta (Exp. 3: –0,75 a +7 d). Lección: una variable exógena *viva* es un pasivo operativo tanto como un activo de modelado; el cargador descarga ahora la ventana completa en cada carga y avisa si Google no responde (nunca más persistencia silenciosa).

Conclusión. (1) El techo $\approx 5,3$ % *aguant*a con pesos congelados sobre un semestre real nunca visto (5,2–5,5 % para cuatro modelos). (2) No hay ganador único: las diferencias entre familias son del tamaño del ruido de semilla, lo que hace razonable un *ensemble*. (3) Enero concentra el error (7,4–8,0 %). (4) Las variables externas en vivo exigen monitorización: un feed congelado degrada más de lo que aporta.

4.2 Experimento 9 – Fiabilidad: del punto al intervalo y la perilla α

Por qué. Si el error topa en el ruido de etiqueta, el siguiente paso útil no es arañar décimas de MAPE sino *cuantificar la incertidumbre de cada día*. Dos preguntas: ¿se puede predecir cuánto va a fallar el modelo mañana? y, si no, ¿se puede acotar?

Cómo. Dos enfoques sobre la misma entrada del Exp. 8 (episodes + case-mix útil + Trends + calendario, 28 d). **(A) Estimador de error:** un segundo transformer (misma receta) entrenado para predecir el error absoluto a +1 d del primario. **(B) Cuantiles:** un encoder no autorregresivo (misos bloques de producción) con una cabeza monótona $P_{10} \leq P_{50} \leq P_{90}$ entrenada con pérdida *pinball* sobre el *residuo* del transformer puntual; la banda es $\hat{y} + [r_{10}, r_{90}]$ y la fiabilidad del día es su anchura. Dos refinamientos operativos: reentrenó excluyendo dos viernes imposibles de 2023 (174 y 248 llegadas, extracciones parciales: imputados con la mediana estacional en la entrada y enmascarados en la pérdida) y la **perilla α** : pronóstico = $P_{50} + \alpha (P_{90} - P_{50})$, un margen que se adapta a la incertidumbre de cada día en vez de sumar una cantidad fija.

Semillas. 10 en (A) y 10 en (B), selección por MAE de $\hat{e}$ y por pinball de validación respectivamente; 3 en el reentrenó sin outliers.

Resultados. **(A) falla:** el MAE del error estimado es $15,3 \pm 0,2$ y su correlación con el error real $r = 0,16 \pm 0,04$ en las 10 semillas; $\hat{e}$ se aplana en 15–25 sea cual sea el día (figura del anexo). El error de un día concreto *no es predecible* a partir de las entradas: es, literalmente, el ruido. **(B) funciona:** en validación la banda $[P_{10}, P_{90}]$ cubre el $74,7 \pm 3,6$ % de los días (10 semillas; nominal 80 %) con anchura ≈ 73 llegadas, y su P50 es además el mejor pronóstico puntual (MAPE 4,68 de media entre semillas; 4,58 la desplegada). Sin outliers el efecto en test es nulo ($5,25 \rightarrow 5,23$). Cobertura por horizonte de la banda desplegada:

Cobertura (%)	+1	+2	+3	+4	+5	+6	+7	media
Validación 2025-H2	77,2	77,2	76,6	76,1	76,6	77,2	76,6	76,8
Test 2026-H1	72,9	71,3	70,2	72,4	72,9	69,1	65,7	70,6

La banda, aprendida sobre residuos de 2022–2025, resulta algo estrecha para un semestre más volátil (anchura ≈ 75 en ambos periodos): *sub-cubre* 6 puntos en prospectivo y la sub-cobertura crece con el horizonte. En los 19 primeros días de enero, con el feed de Trends congelado, la cobertura fue del 42 %. La perilla α traslada la tasa de infraestimación (días en que el real supera al pronóstico) de validación a test:

α	$\approx$ cuantil	corto val	corto test	margen medio	MAPE val	MAPE test
0,00	P50	57,1 %	55,8 %	+0	4,82	5,23
0,25	$\sim$ P60	46,7 %	46,4 %	+9	4,81	5,24
0,50	$\sim$ P70	32,1 %	32,0 %	+19	5,28	5,66
0,75	$\sim$ P80	19,0 %	22,1 %	+28	6,19	6,52
1,00	P90	12,5 %	16,0 %	+38	7,50	7,62
1,25	—	9,8 %	8,3 %	+47	8,95	8,96
1,50	—	5,4 %	5,0 %	+57	10,50	10,48
2,00	—	2,2 %	2,8 %	+76	13,84	13,70

El MAPE sube *a propósito* con α (se predice alto adrede); la cifra que importa es el porcentaje de días corto, que se traslada de val a test con un desvío máximo de 3,5 puntos ($\alpha = 1$: 12,5 vs 16,0 %; nominal 10 %). La elección de α es una decisión de costes (regla del vendedor de periódicos, $q^* = C_u/(C_u + C_o)$): si quedarse corto cuesta tres veces más que pasarse, $q^* = 0,75$ y $\alpha \approx 0,75$.

Conclusión. El error de un día no se predice (estimador de error $r = 0,16$), pero se acota: la banda P10–P90 está razonablemente calibrada en validación (77 %) y sub-cubre en prospectivo (71 %), más cuanto mayor es el horizonte, así que conviene recalibrarla periódicamente (p. ej. cada semestre, o con un ajuste conformal). La perilla α es la interfaz operativa correcta: convierte la banda en una tasa de infraestimación elegida, y esa tasa se traslada al futuro con ± 3 puntos.

4.3 Experimento 10 – Desagregación: áreas asistenciales e ingresos

Por qué. El hospital pidió la previsión por área (Pediatría, Ginecología-Obstetricia y Resto) y la de ingresos diarios para planificar camas. Son series de menor volumen: hay que saber cuánto se puede exigirles.

Cómo. Misma arquitectura de producción, *una red por objetivo* (norma del equipo: nada de modelos auxiliares). Entrada: 28 d de [objetivo + temperatura + precipitación + calendario] y el calendario futuro; a propósito *sin* case-mix ni Trends, para disponer de modelos sin dependencia de Google ni del retardo de codificación. Áreas por código de servicio (Pediatría: 8 servicios; Gine-Obs: 3; Resto: 11); ingresos = episodios con destino hospitalización. Cabeza de cuantiles por objetivo como en el Exp. 9. Split del Protocolo A (test 2025-H2) y, después, extensión al semestre prospectivo 2026-H1 con el envío nuevo (área por patrón del nombre de servicio; puerta de verificación: reproduce diciembre de 2025 con diferencia máxima 0,01; las tres áreas suman el total diario en 179 de 181 días).

Semillas. 1 por objetivo en los modelos puntuales (selección por época en validación); 3 por objetivo en los cuantiles (por pinball).

Resultados. A +1 d, con el *naïve* estacional como referencia y la cobertura de la banda a +1 d:

Objetivo	media/día	test 2025-H2			test 2026-H1		cobertura 80 %	
		MAE	MAPE	naïve MAE	MAE	MAPE	25-H2	26-H1
Resto (general)	395	20,4	5,2 %	29,3	22,1	5,6 %	79,2	78,5
Pediatría	97	12,4	13,6 %	14,3	14,7	16,6 %	80,9	70,7
Gine-Obstetricia	47	6,1	13,0 %	8,4	6,2	13,2 %	77,0	75,7
Ingresos (camas)	52	6,6	12,7 %	8,6	6,4	12,9 %	83,7	77,3

Todos baten al *naïve* (un 25–30 % menos de MAE) y el error crece poco con el horizonte (Resto: MAE 20,4 → 23,3 de +1 a +7 d; ingresos 6,6 → 6,9; Gine-Obs plano). Resto se comporta como el total (5,2–5,6 %). Los tres objetivos pequeños rondan el 13 %, y *ahí no hay nada que arañar*: para un conteo de Poisson de media λ el mejor MAE posible es $\approx 0,8\sqrt{\lambda}$, es decir, un suelo *relativo* de $\approx 0,8/\sqrt{\lambda}$: 11,7 % para Gine-Obs ($\lambda \approx 47$), 11,1 % para ingresos (≈ 52), 8,1 % para Pediatría (≈ 97), 4,0 % para Resto y 3,5 % para el total (≈ 533). Gine-Obs e ingresos están *sobre* su suelo; Pediatría queda 5 puntos por encima del suyo (sobredispersión: brotes epidémicos infantiles, la serie donde un proxy epidémico más ayudaría; en 2026-H1 se degrada a 16,6 %); Resto y el total dejan una brecha de $\approx 1,3$ puntos que explica el ruido de etiqueta ($\sim 17/\text{día} \approx 3,2$ % del total; combinando ambos suelos en cuadratura sale $\approx 4,7$ %, justo el nivel alcanzado). Es una estimación orientativa (aproximación normal, independencia), pero cuadra.

Conclusión. La desagregación funciona hasta donde el volumen lo permite: Resto como el total (≈ 5 %); Pediatría, Gine-Obs e ingresos en el 12–13 % que dicta su suelo de Poisson, con bandas bien calibradas en 2025-H2 (77–84 %) y algo estrechas en prospectivo (71–79 %). Los ingresos por área (3–5/día) no admiten pronóstico diario; para camas la cifra útil es el total. Aviso: la tasa de ingreso subió ~ 8 % relativo en 2025 (*label shift*), motivo para recalibrar con cada envío.

5 Discusión y conclusiones

Dónde está la señal. Por orden: (1) la *autocorrelación* del recuento (el grueso; el calendario explícito no añade nada porque la ventana de 28 días ya codifica el ciclo semanal); (2) el *estado epidémico* (única ganancia aditiva real, $\sim -0,55$ puntos); y poco más.

Qué se pisa (redundancia), con su evidencia.

- *Trends* $\leftrightarrow$ *case-mix*: juntas no suman (ALL = 5,35 = cada una sola; Exp. 5). Dos ventanas a la misma ola.
- *AR* $\leftrightarrow$ *día de la semana*: $AR = AR + \text{cal}$ (Exp. 5).
- *case-mix entre sí*: bastan 3 de 21 (Exp. 7); las 34 juntas empeoran respecto a 6 curadas (Exp. 5).
- *exógenas externas*: nulas sobre *case-mix* (Exp. 4, 20 semillas).

El techo, ahora cuantificado y confirmado. Varias configuraciones independientes convergen a $\approx 5,3$ % de MAPE, y el test prospectivo lo ratifica: cuatro modelos distintos, con pesos congelados, dan 5,2–5,5 % en 181 días nunca vistos (Exp. 8). No es casualidad: el suelo de Poisson del volumen diario ($\approx 3,5$ %) y el ruido de etiqueta (~ 17 episodios/día de desacuerdo entre dos fuentes de conteo, $\approx 3,2$ %) combinados en cuadratura dan $\approx 4,7$ % (Exp. 10), justo el nivel que alcanza el mejor modelo a +1 d. Ningún regresor diario lo cruza; mejorar más exige reducir el ruido de etiqueta, no añadir variables.

La incertidumbre se acota, no se predice. Intentar predecir el error de mañana fracasa ($r = 0,16$, Exp. 9): ese error *es* el ruido. Lo que sí funciona es emitir cuantiles: la banda P10–P90 cubre el 77 % en validación y el 71 % en prospectivo, y la perilla α = margen elegido traslada la tasa de infraestimación al futuro con ± 3 puntos. Es la interfaz que el gestor necesita: decide cuánto quiere quedarse corto, no cuánto MAPE quiere.

Recomendación (lo desplegado). Los cuatro modelos del Exp. 8 con la banda y la perilla α del

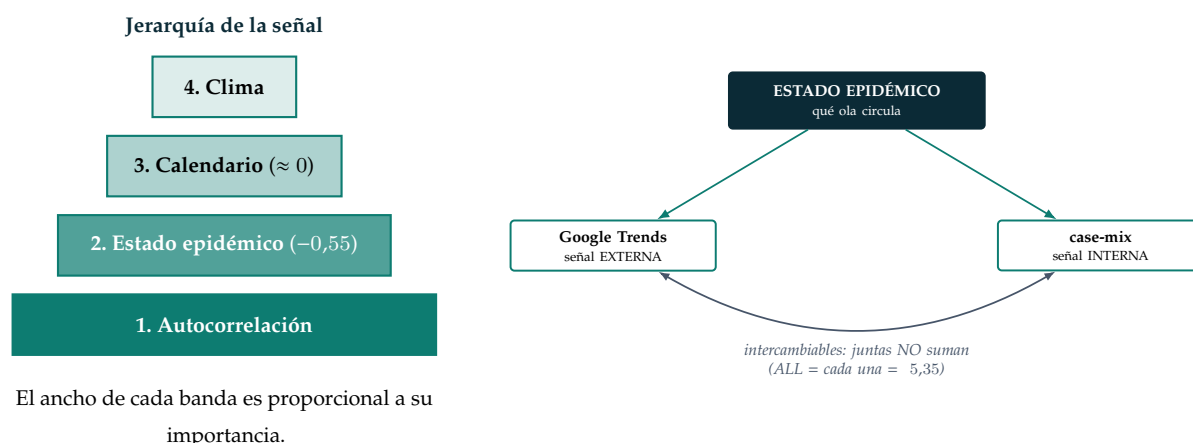


Figura 3: **Izquierda:** jerarquía de la señal (ancho $\propto$ importancia). La autocorrelación es la base; el estado epidémico es la única ganancia aditiva; calendario y clima no aportan. **Derecha:** por qué Trends y case-mix son redundantes — son dos medidas del mismo factor latente, así que añadir una cuando ya está la otra no suma.

Exp. 9 y los modelos por área e ingresos del Exp. 10 están integrados en una aplicación web autónoma (ejecuta los modelos sin reentrenar, carga envíos nuevos, reentrena bajo demanda y emite el parte de previsión a 7 días con banda y α). Tres consejos operativos: (1) promediar transformer y XGB-sum (no hay ganador único y sus errores se descorrelacionan por familia); (2) monitorizar el feed de Trends y re-descargar la ventana completa en cada carga (un feed congelado cuesta más de lo que aporta); (3) recalibrar la banda y el α con cada semestre recibido, porque la cobertura se erosiona con el drift.

Limitaciones. Un solo centro/región; suelo de ruido intrínseco; posible drift de la era COVID; las fuentes externas pueden cambiar, fallar o quedarse rancias (incidente del Exp. 8); el case-mix interno tiene un retardo de codificación que condiciona su uso en tiempo real; la banda de fiabilidad sub-cubre 6–9 puntos fuera de muestra y crece con el horizonte; los objetivos de bajo volumen (Gine-Obs, ingresos) están ya en su suelo de Poisson y Pediatría se degrada en prospectivo (16,6 %); la tasa de ingreso muestra *label shift*; el test prospectivo cubre un solo semestre (invierno y primavera), no un año completo; y no se han aplicado contrastes formales (Diebold–Mariano) a las diferencias entre modelos, que aquí se juzgan frente a la banda de semilla.

Anexo A. Figuras

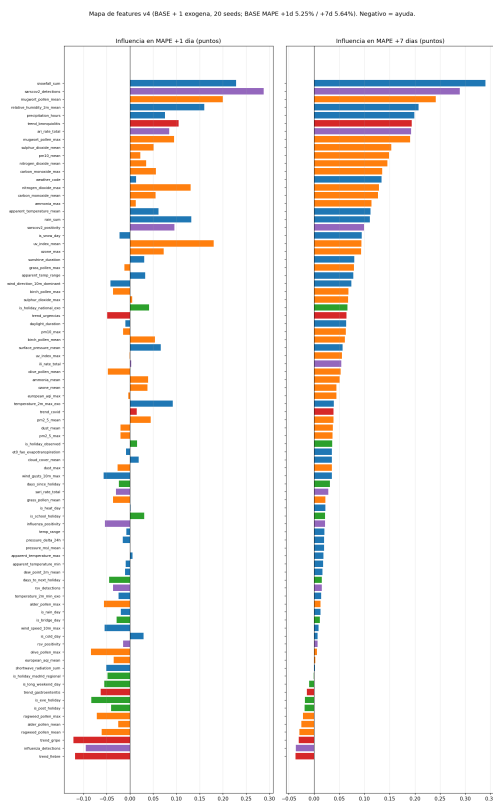


Figura 4: Influencia individual de las 85 variables exógenas sobre la base episodes+case-mix (20 semillas). Ninguna supera la banda de ruido.

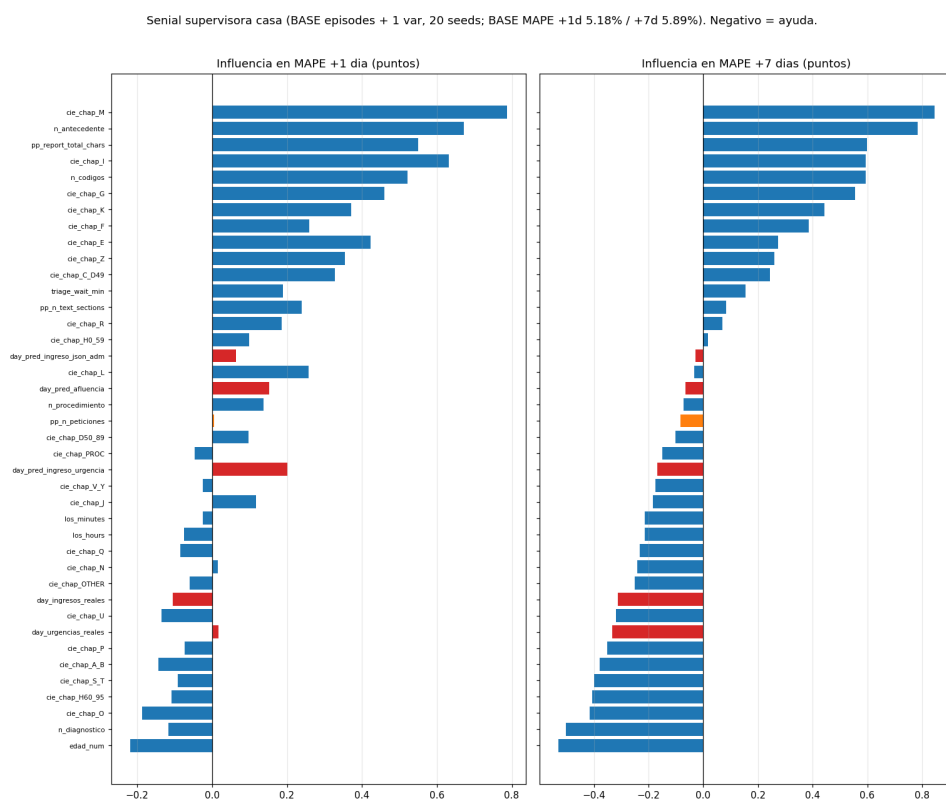


Figura 5: Influencia individual de las 40 variables de la casa sobre episodios (20 semillas). Señal real pero difusa y redundante.

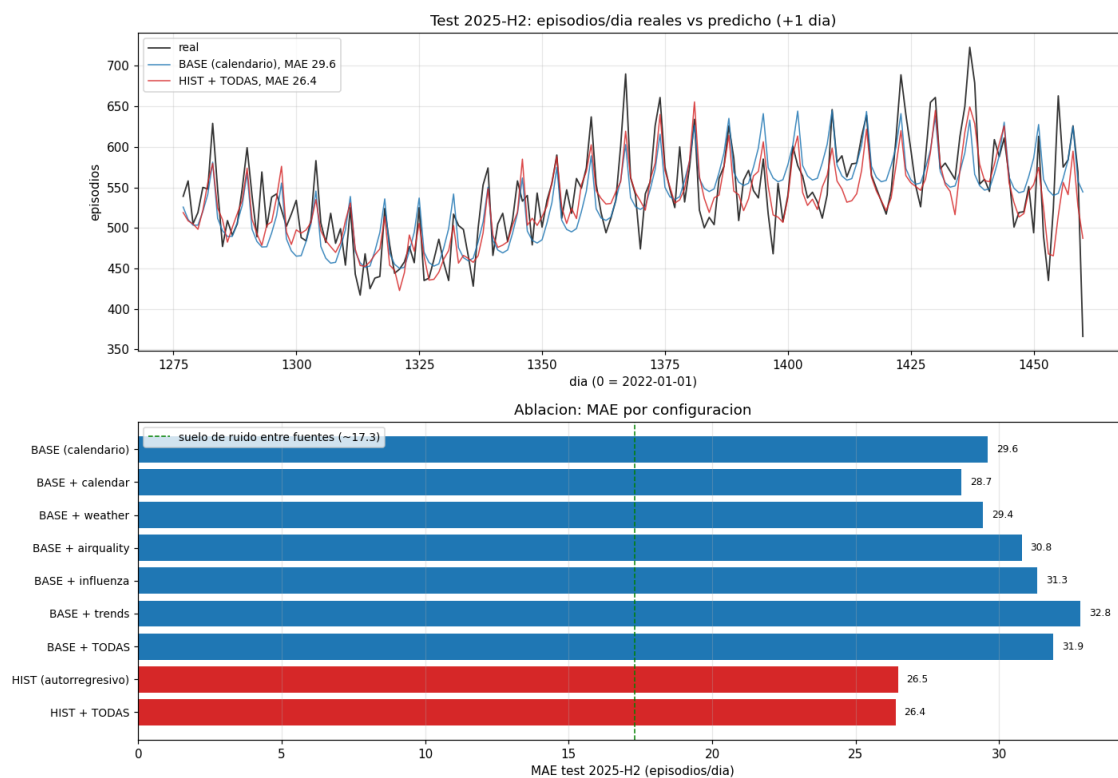


Figura 6: Estudio de ablación de exógenas (exploratorio): serie real vs predicho y ranking de MAE por configuración.

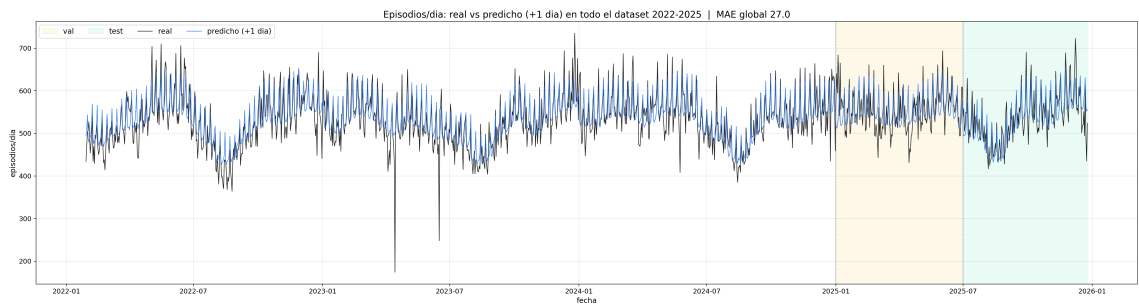


Figura 7: Transformer: pronóstico a 7 días sobre el test (rodante).

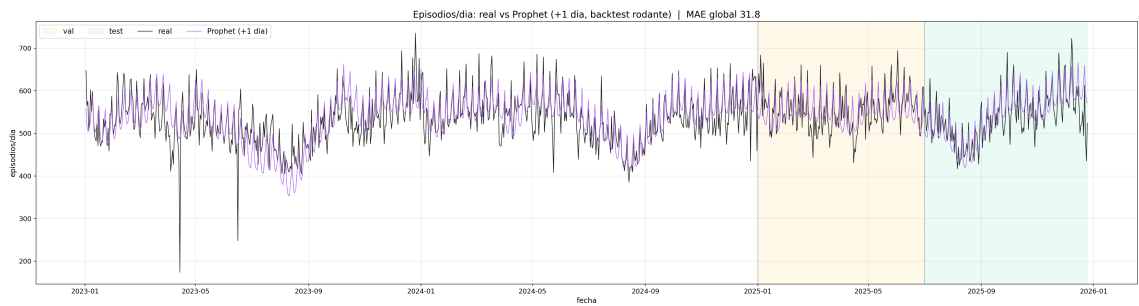


Figura 8: Prophet: pronóstico y componentes sobre la serie diaria.

USADAS por los modelos

Autocorrelacion

▼ episodios (ventana de 28 días)

el grueso de la señal

Case-mix util (7, union greedy + curado)

▼ edad_num

▼ los_hours

▼ n_diagnostico

▼ cie_chap_H60_95

▼ cie_chap_P

▼ cie_chap_OTHER

▼ cie_chap_N

Calendario (4)

▼ day_of_week, month

▼ is_weekend, is_holiday_any

Google Trends (6)

▼ gripe, fiebre, urgencias

▼ covid, gastroenteritis, bronquiolitis

DESCARTADAS

Case-mix restante (~27 de 34)

✗ n_codigos, cie_chap_M/I/G/K...

redundante (greedy/curado)

Clima

✗ ERA5: temperatura, precipitacion

no aporta / estorba

✗ weather: humedad, viento, presion

nulo

Aire y polen (32)

✗ PM2.5, NO2, ozono, polenes

nulo

Vigilancia gripal ERVISS

✗ SARI/ARI, positividad gripe/RSV

redundante con Trends

Administrativas / cuasi-id

✗ hosp_ref, centro_ref, resp_*, fuente

drift / cuasi-id

Fuga o retrospectivas

✗ day_* (pred/reales hospital)

LEAKY + cobertura parcial

✗ LOS, triage_wait, motivo_alta

solo al cerrar el episodio

Config final: episodios (AR) + case-mix util (7) + calendario (4) + Google Trends (6). El resto no supera la banda de ruido, es redundante, leaky o no disponible en tiempo real.

Figura 9: Configuración final desplegada (Exp. 8): variables usadas (izquierda) y descartadas con su razón (derecha).

MAPE por horizonte - config casemix+cal+trends

VALIDACION (2025-H2)

Modelo	+1d	+2d	+3d	+4d	+5d	+6d	+7d
Transformer	4.73%	4.71%	4.64%	4.78%	4.80%	4.84%	5.18%
XGB-flat	4.76%	4.99%	5.14%	5.30%	5.34%	5.35%	5.63%
XGB-sum	5.07%	5.19%	5.30%	5.39%	5.42%	5.51%	5.59%

TRAIN

Modelo	+1d	+2d	+3d	+4d	+5d	+6d	+7d
Transformer	5.12%	5.24%	5.36%	5.49%	5.56%	5.55%	5.59%
XGB-flat	4.03%	3.92%	3.85%	3.81%	3.89%	4.01%	4.25%
XGB-sum	3.88%	3.73%	3.61%	3.58%	3.62%	3.75%	4.03%

Figura 10: Exp. 8. MAPE por horizonte de los tres modelos puntuales en validación (arriba) y entrenamiento (abajo), en el momento de la selección.

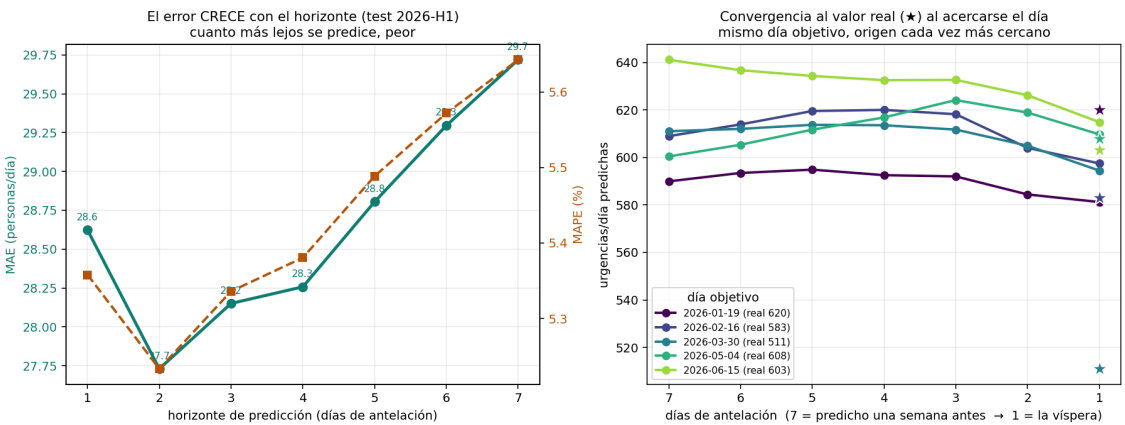


Figura 11: Exp. 8. Izquierda: MAE y MAPE del transformer desplegado por horizonte en el test prospectivo 2026-H1 (28,6 → 29,7 de +1 a +7 d). Derecha: convergencia de los pronósticos sucesivos de cinco días objetivo hacia su valor real (★) a medida que se acerca el origen.

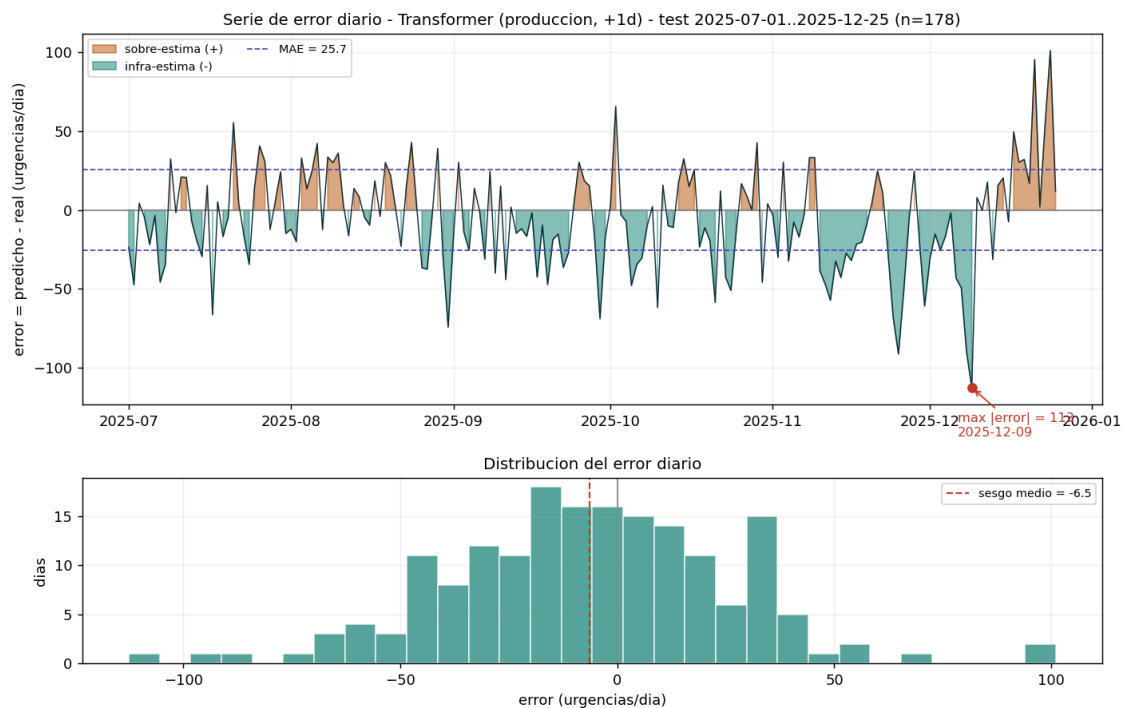


Figura 12: Exp. 8. Serie de error diario del transformer desplegado a +1 d sobre 2025-H2 (MAE 25,7; máximo 113 el 2025-12-09).

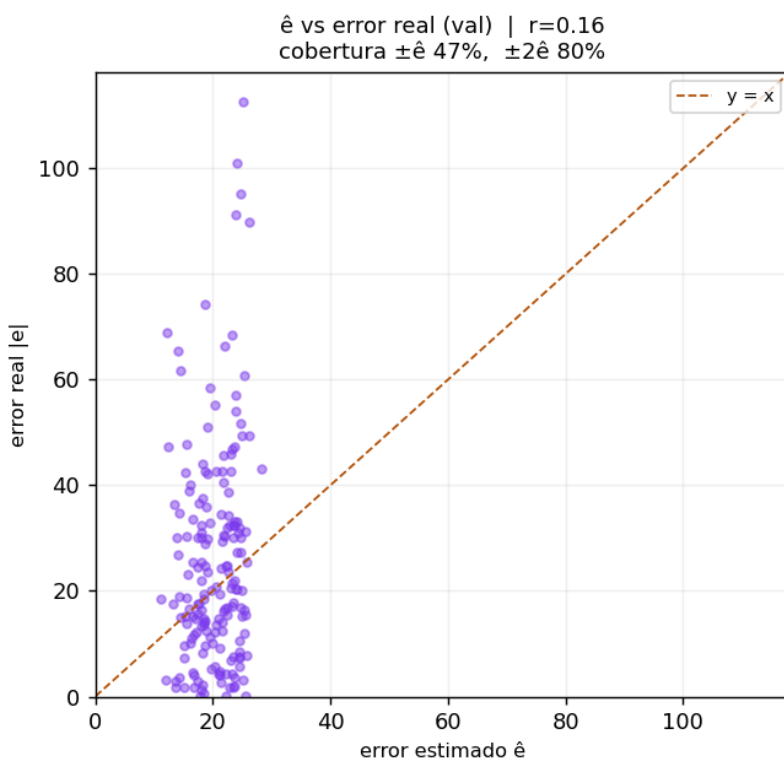


Figura 13: Exp. 9 (A). Error estimado $\hat{e}$ frente a error real $|e|$ en validación: $r = 0,16$; $\hat{e}$ se aplana en 15–25 sea cual sea el día. El error diario no es predecible.

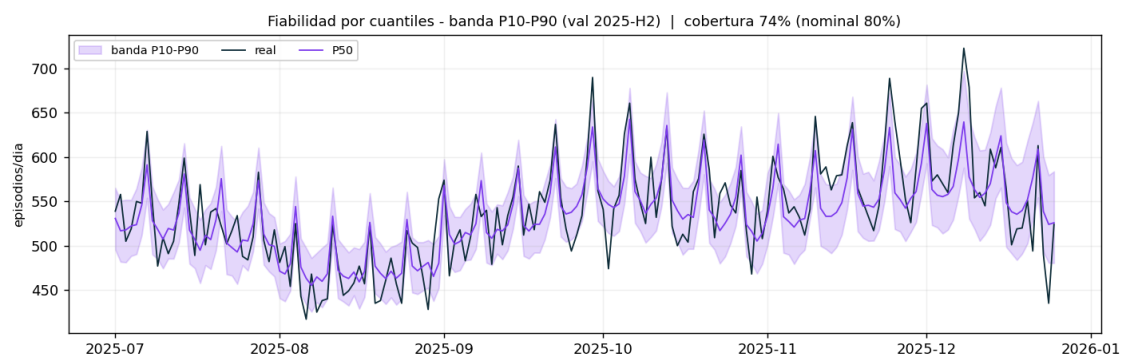


Figura 14: Exp. 9 (B). Banda P10-P90 del modelo de cuantiles sobre la validación 2025-H2 (cobertura 74 %, nominal 80 %), con el real y el P50.

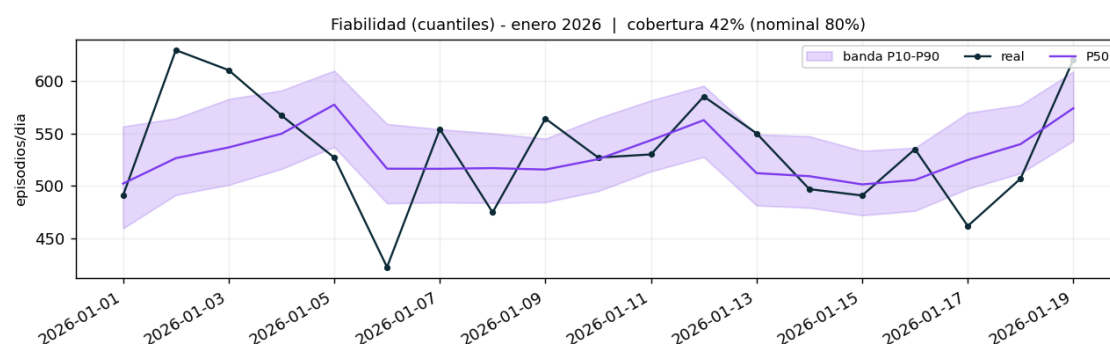


Figura 15: Exp. 8/9. Los 19 primeros días de enero de 2026 con el feed de Trends persistido: cobertura del 42 % y MAPE 7,4 %; los rebotes posfestivos quedan fuera de banda.

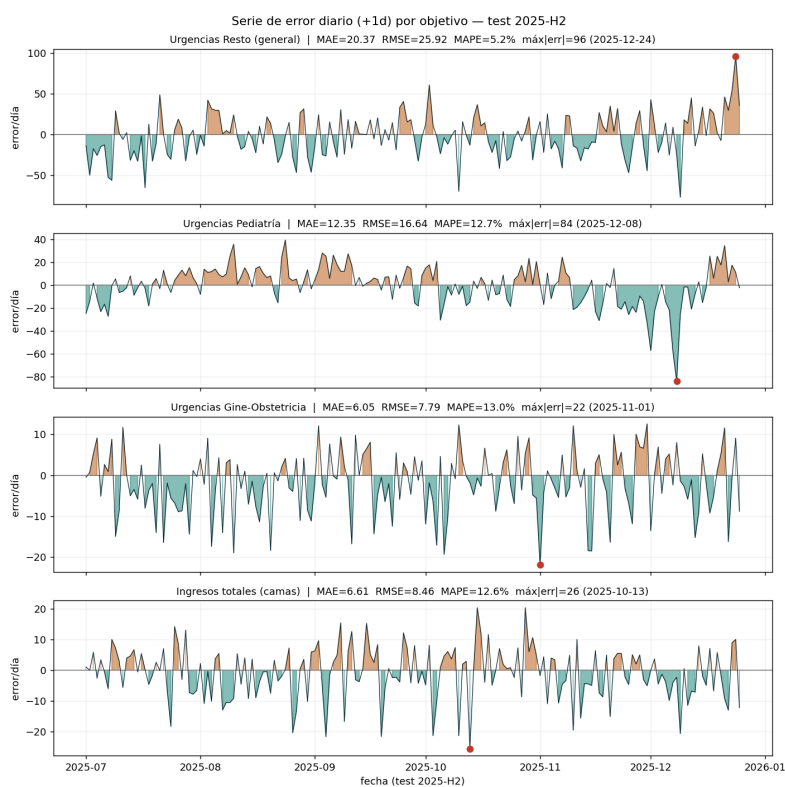


Figura 16: Exp. 10. Serie de error diario (+1 d) por objetivo en el test 2025-H2, con el máximo marcado.

Anexo B. Influencia de las 85 variables exógenas (20 semillas)

Base (episodes+case-mix): MAPE +1d 5.25 %, +7d 5.64 %. Ganancia = MAPE(base) - MAPE(con); positivo = ayuda.

#	variable	bloque	gan.+7d	gan.+1d
1	trend_fiebre	trends	+0.04	+0.12
2	influenza_detections	influenza	+0.04	+0.09
3	trend_gripe	trends	+0.03	+0.12
4	ragweed_pollen_mean	airquality	+0.03	+0.06
5	alder_pollen_mean	airquality	+0.03	+0.02
6	ragweed_pollen_max	airquality	+0.02	+0.07
7	is_post_holiday	calendar	+0.02	+0.04
8	is_eve_holiday	calendar	+0.02	+0.08
9	trend_gastroenteritis	trends	+0.01	+0.06
10	is_long_weekend_day	calendar	+0.01	+0.06
11	is_holiday_madrid_regional	calendar	+0.00	+0.05
12	shortwave_radiation_sum	weather	-0.00	+0.05
13	european_aqi_mean	airquality	-0.00	+0.03
14	olive_pollen_max	airquality	-0.01	+0.08
15	rsv_positivity	influenza	-0.01	+0.01
16	is_cold_day	weather	-0.01	-0.03
17	wind_speed_10m_max	weather	-0.01	+0.05
18	is_bridge_day	calendar	-0.01	+0.03
19	is_rain_day	weather	-0.01	+0.02
20	alder_pollen_max	airquality	-0.01	+0.06
21	temperature_2m_min_exo	weather	-0.01	+0.02
22	rsv_detections	influenza	-0.02	+0.04
23	days_to_next_holiday	calendar	-0.02	+0.04
24	dew_point_2m_mean	weather	-0.02	+0.01
25	apparent_temperature_min	weather	-0.02	+0.01
26	apparent_temperature_max	weather	-0.02	-0.01
27	pressure_msl_mean	weather	-0.02	-0.00
28	pressure_delta_24h	weather	-0.02	+0.01
29	temp_range	weather	-0.02	+0.01
30	influenza_positivity	influenza	-0.02	+0.05
31	is_school_holiday	calendar	-0.02	-0.03
32	is_heat_day	weather	-0.02	-0.00
33	grass_pollen_mean	airquality	-0.02	+0.04
34	sari_rate_total	influenza	-0.03	+0.03
35	days_since_holiday	calendar	-0.03	+0.02
36	wind_gusts_10m_max	weather	-0.04	+0.06
37	dust_max	airquality	-0.04	+0.03
38	cloud_cover_mean	weather	-0.04	-0.02
39	et0_fao_evapotranspiration	weather	-0.04	+0.01
40	is_holiday_observed	calendar	-0.04	-0.02
41	pm2_5_max	airquality	-0.04	+0.02
42	dust_mean	airquality	-0.04	+0.02
43	pm2_5_mean	airquality	-0.04	-0.04
44	trend_covid	trends	-0.04	-0.02
45	temperature_2m_max_exo	weather	-0.04	-0.09
46	european_aqi_max	airquality	-0.04	+0.00
47	ozone_mean	airquality	-0.04	-0.04
48	ammonia_mean	airquality	-0.05	-0.04
49	olive_pollen_mean	airquality	-0.05	+0.05

#	variable	bloque	gan.+7d	gan.+1d
50	ili_rate_total	influenza	-0.05	-0.00
51	uv_index_max	airquality	-0.06	+0.00
52	surface_pressure_mean	weather	-0.06	-0.07
53	birch_pollen_mean	airquality	-0.06	-0.05
54	pm10_max	airquality	-0.06	+0.01
55	daylight_duration	weather	-0.06	+0.01
56	trend_urgencias	trends	-0.06	+0.05
57	is_holiday_national_exo	calendar	-0.07	-0.04
58	sulphur_dioxide_max	airquality	-0.07	-0.00
59	birch_pollen_max	airquality	-0.07	+0.04
60	wind_direction_10m_dominant	weather	-0.07	+0.04
61	apparent_temp_range	weather	-0.08	-0.03
62	grass_pollen_max	airquality	-0.08	+0.01
63	sunshine_duration	weather	-0.08	-0.03
64	ozone_max	airquality	-0.09	-0.07
65	uv_index_mean	airquality	-0.09	-0.18
66	is_snow_day	weather	-0.09	+0.02
67	sarscov2_positivity	influenza	-0.10	-0.10
68	rain_sum	weather	-0.11	-0.13
69	apparent_temperature_mean	weather	-0.11	-0.06
70	ammonia_max	airquality	-0.11	-0.01
71	carbon_monoxide_mean	airquality	-0.13	-0.06
72	nitrogen_dioxide_max	airquality	-0.13	-0.13
73	weather_code	weather	-0.13	-0.01
74	carbon_monoxide_max	airquality	-0.14	-0.06
75	nitrogen_dioxide_mean	airquality	-0.15	-0.04
76	pm10_mean	airquality	-0.15	-0.02
77	sulphur_dioxide_mean	airquality	-0.15	-0.05
78	mugwort_pollen_max	airquality	-0.19	-0.09
79	ari_rate_total	influenza	-0.19	-0.08
80	trend_bronquiolitis	trends	-0.19	-0.10
81	precipitation_hours	weather	-0.20	-0.08
82	relative_humidity_2m_mean	weather	-0.21	-0.16
83	mugwort_pollen_mean	airquality	-0.24	-0.20
84	sarscov2_detections	influenza	-0.29	-0.29
85	snowfall_sum	weather	-0.34	-0.23

Anexo C. Influencia de las 40 variables de la casa (20 semillas)

Base (solo episodios): MAPE +1d 5.18 %, +7d 5.89 %.

#	variable	tipo	gan.+7d	gan.+1d
1	edad_num	casemix	+0.53	+0.22
2	n_diagnostico	casemix	+0.50	+0.12
3	cie_chap_0	casemix	+0.41	+0.19
4	cie_chap_H60_95	casemix	+0.41	+0.11
5	cie_chap_S_T	casemix	+0.40	+0.09
6	cie_chap_A_B	casemix	+0.38	+0.14
7	cie_chap_P	casemix	+0.35	+0.07
8	day_urgencias_reales	redund	+0.33	-0.02
9	cie_chap_U	casemix	+0.32	+0.14
10	day_ingresos_reales	redund	+0.31	+0.11
11	cie_chap_OTHER	casemix	+0.25	+0.06
12	cie_chap_N	casemix	+0.24	-0.02
13	cie_chap_Q	casemix	+0.23	+0.08
14	los_hours	casemix	+0.21	+0.08
15	los_minutes	casemix	+0.21	+0.03
16	cie_chap_J	casemix	+0.18	-0.12
17	cie_chap_V_Y	casemix	+0.17	+0.02
18	day_pred_ingreso_urgencia	redund	+0.17	-0.20
19	cie_chap_PROC	casemix	+0.15	+0.05
20	cie_chap_D50_89	casemix	+0.10	-0.10
21	pp_n_peticiones	lowcov	+0.08	-0.00
22	n_procedimiento	casemix	+0.07	-0.14
23	day_pred_afluencia	redund	+0.07	-0.15
24	cie_chap_L	casemix	+0.03	-0.26
25	day_pred_ingreso_json_adm	redund	+0.03	-0.06
26	cie_chap_H0_59	casemix	-0.02	-0.10
27	cie_chap_R	casemix	-0.07	-0.19
28	pp_n_text_sections	casemix	-0.08	-0.24
29	triage_wait_min	casemix	-0.16	-0.19
30	cie_chap_C_D49	casemix	-0.24	-0.33
31	cie_chap_Z	casemix	-0.26	-0.35
32	cie_chap_E	casemix	-0.27	-0.42
33	cie_chap_F	casemix	-0.39	-0.26
34	cie_chap_K	casemix	-0.44	-0.37
35	cie_chap_G	casemix	-0.56	-0.46
36	n_codigos	casemix	-0.59	-0.52
37	cie_chap_I	casemix	-0.59	-0.63
38	pp_report_total_chars	casemix	-0.60	-0.55
39	n_antecedente	casemix	-0.79	-0.67
40	cie_chap_M	casemix	-0.85	-0.79

Anexo D. Benchmark factorial: cada semilla (10)

MAPE +7d por semilla

seed	AR	+cal	+clima	+cmx6	+cmx34	+Trends	ALL
0	5.50	5.50	6.01	5.62	5.53	5.23	5.63
1	5.92	5.92	6.09	5.05	6.22	5.00	5.08
2	6.25	6.25	6.41	5.35	5.80	5.00	5.23
3	5.36	5.36	5.89	5.69	5.35	5.63	5.73
4	5.61	5.61	5.73	5.24	5.36	5.13	5.34
5	5.60	5.60	5.93	5.56	5.40	5.01	5.65
6	6.32	6.32	6.88	5.35	5.41	5.81	5.06
7	5.98	5.98	5.73	5.22	5.66	5.76	5.39
8	6.36	6.36	7.58	5.15	5.43	5.21	5.51
9	6.00	6.00	5.86	5.28	6.20	5.58	4.91
media	5.89	5.89	6.21	5.35	5.64	5.34	5.35
sd	0.34	0.34	0.56	0.20	0.32	0.31	0.26

MAPE +1d por semilla

seed	AR	+cal	+clima	+cmx6	+cmx34	+Trends	ALL
0	4.77	4.77	5.01	4.76	5.11	4.58	5.21
1	5.80	5.80	5.28	4.78	5.74	4.63	4.74
2	5.62	5.62	5.60	4.90	5.34	4.70	4.95
3	4.97	4.97	5.08	5.19	5.18	4.84	5.05
4	5.18	5.18	4.89	4.82	4.98	4.55	4.93
5	4.89	4.89	5.11	4.87	4.87	4.59	5.29
6	5.63	5.63	6.13	4.91	5.01	5.09	4.95
7	5.06	5.06	5.17	4.95	5.39	5.14	5.18
8	5.79	5.79	6.48	4.97	5.20	4.79	5.11
9	5.57	5.57	5.31	4.87	5.60	5.03	4.79
media	5.33	5.33	5.40	4.90	5.24	4.79	5.02
sd	0.37	0.37	0.49	0.12	0.26	0.21	0.17

Anexo E. SOTA (RF y XGBoost): cada semilla (10)

XGBoost + casa

seed	MAE	RMSE	MAPE +1d	MAPE +7d
0	29.12	37.57	4.77	5.93
1	29.01	37.37	4.75	5.90
2	28.97	37.47	4.79	5.83
3	28.99	37.49	4.82	5.77
4	29.22	37.61	4.81	5.97
5	29.17	37.55	4.83	5.95
6	29.03	37.53	4.83	5.82
7	29.06	37.40	4.81	5.90
8	29.07	37.42	4.83	5.88
9	29.09	37.58	4.83	5.89

Random Forest + casa

seed	MAE	RMSE	MAPE +1d	MAPE +7d
0	32.05	40.73	5.41	6.47
1	32.04	40.79	5.41	6.52
2	31.87	40.72	5.37	6.43
3	31.93	40.65	5.38	6.46
4	32.13	40.89	5.39	6.51
5	32.01	40.67	5.38	6.45
6	32.14	40.84	5.42	6.52
7	32.06	40.85	5.42	6.47
8	32.06	40.70	5.39	6.50
9	31.92	40.58	5.39	6.44

Anexo F. Greedy forward: traza de selección

paso	variable añadida	val +7d	test +7d
0	(solo episodes)	5.891	6.025
1	cie_chap_N	5.726	5.495
2	cie_chap_H60_95	5.636	5.545
3	edad_num	5.546	5.717

Anexo G. Despliegue por horizonte (Exp. 8)

Recalculado desde la caché final de la aplicación (predictions.csv). MAPE en %, MAE en llegadas/día.

MAPE

split	modelo	+1	+2	+3	+4	+5	+6	+7	media	Δ_{7-1}
val 2025-H2	Transformer	5.06	4.91	4.81	4.91	4.94	4.99	5.06	4.95	+0.01
val 2025-H2	XGB-flat	4.99	5.19	5.30	5.48	5.48	5.47	5.59	5.36	+0.59
val 2025-H2	XGB-sum	5.25	5.34	5.34	5.46	5.57	5.52	5.54	5.43	+0.29
val 2025-H2	Fiabilidad (P50)	4.82	4.75	4.71	4.85	4.89	4.92	4.97	4.84	+0.16
test 2026-H1	Transformer	5.36	5.23	5.34	5.38	5.49	5.57	5.64	5.43	+0.28
test 2026-H1	XGB-flat	5.40	5.46	5.42	5.53	5.49	5.48	5.45	5.46	+0.05
test 2026-H1	XGB-sum	4.86	5.11	5.22	5.30	5.26	5.27	5.28	5.19	+0.41
test 2026-H1	Fiabilidad (P50)	5.23	5.28	5.41	5.50	5.54	5.73	5.74	5.49	+0.51

MAE

split	modelo	+1	+2	+3	+4	+5	+6	+7	media	Δ_{7-1}
val 2025-H2	Transformer	27.20	26.22	25.64	26.00	26.14	26.45	26.84	26.36	-0.36
val 2025-H2	XGB-flat	26.66	27.72	28.35	29.23	29.15	29.06	29.76	28.56	+3.09
val 2025-H2	XGB-sum	28.22	28.69	28.72	29.28	29.87	29.68	29.85	29.19	+1.63
val 2025-H2	Fiabilidad (P50)	25.85	25.37	25.10	25.74	25.74	25.95	26.36	25.73	+0.51
test 2026-H1	Transformer	28.62	27.73	28.15	28.26	28.80	29.30	29.72	28.65	+1.10
test 2026-H1	XGB-flat	29.22	29.51	29.30	29.89	29.66	29.59	29.46	29.52	+0.24
test 2026-H1	XGB-sum	26.09	27.48	28.04	28.44	28.18	28.21	28.29	27.82	+2.20
test 2026-H1	Fiabilidad (P50)	27.96	28.00	28.62	29.00	29.17	30.19	30.39	29.05	+2.43

Anexo H. Fiabilidad: cobertura y α (Exp. 9)

Banda desplegada (mejor semilla, sin outliers). Cobertura nominal 80 %. “Corto” = días en que el real supera al pronóstico (+1 d).

split	cob.+1	cob.+2	cob.+3	cob.+4	cob.+5	cob.+6	cob.+7	media	anchura	n
val 2025-H2	77.2	77.2	76.6	76.1	76.6	77.2	76.6	76.8	75.1	184
test 2026-H1	72.9	71.3	70.2	72.4	72.9	69.1	65.7	70.6	75.2	181

α	corto val	corto test	margen val	margen test	MAPE val	MAPE test
0.00	57.1 %	55.8 %	+0.0	+0.0	4.82	5.23
0.25	46.7 %	46.4 %	+9.6	+9.3	4.81	5.24
0.50	32.1 %	32.0 %	+19.2	+18.6	5.28	5.66
0.75	19.0 %	22.1 %	+28.8	+27.9	6.19	6.52
1.00	12.5 %	16.0 %	+38.5	+37.1	7.50	7.62
1.25	9.8 %	8.3 %	+48.1	+46.4	8.95	8.96
1.50	5.4 %	5.0 %	+57.7	+55.7	10.50	10.48
2.00	2.2 %	2.8 %	+76.9	+74.3	13.84	13.70

Anexo I. Fiabilidad por semilla (Exp. 9)

(B) Cuantiles: 10 semillas (validación 2025-H2; ★ = desplegada)

seed	pinball	cobertura 80	anchura	MAPE P50
0	0.6926	70.2	69.1	4.72
1	0.6964	78.1	78.7	4.77
2 ★	0.6732	74.2	72.7	4.58
3	0.6956	74.2	73.7	4.64
4	0.6975	80.9	80.4	4.71
5	0.6797	71.9	70.8	4.66
6	0.6822	77.5	74.8	4.61
7	0.6835	71.3	68.6	4.67
8	0.6970	78.7	75.6	4.73
9	0.6989	70.2	70.4	4.72

(A) Estimador de error: 10 semillas (MAE y correlación de $\hat{e}$ con el error real, validación)

seed	MAE($\hat{e}$)	RMSE	r
0	15.16	21.04	0.120
1	15.31	20.85	0.166
2	15.73	21.39	0.094
3	15.21	21.17	0.196
4	15.14	21.13	0.157
5	15.35	20.59	0.234
6	15.31	21.06	0.159
7	15.25	20.86	0.173
8	15.20	20.94	0.201
9	15.22	20.75	0.118

Anexo J. Despliegue por semilla (Exp. 8)

MAPE de validación a +1d / +7d por semilla; ★ = semilla desplegada (mejor a +1d).

seed	Transformer		XGB-flat		XGB-sum	
	+1d	+7d	+1d	+7d	+1d	+7d
0	4.73★	5.27	4.77	5.59	5.10	5.73
1	4.93	5.23	4.86	5.64	5.14	5.76
2	4.90	5.25	4.82	5.65	5.15	5.69
3	5.07	5.45	4.79	5.67	5.09	5.63
4	4.86	5.14	4.82	5.69	5.20	5.77
5	5.34	5.36	4.78	5.68	5.18	5.67
6	5.02	5.18	4.84	5.69	5.07	5.71
7	4.93	5.45	4.82	5.69	5.18	5.66
8	5.29	5.40	4.86	5.68	5.13	5.64
9	4.98	5.42	4.91	5.73	5.17	5.75
10	4.81	5.16	4.83	5.62	5.20	5.62
11	4.84	4.94	4.79	5.65	5.15	5.66
12	4.88	5.19	4.82	5.65	5.07★	5.59
13	5.10	5.43	4.77	5.56	5.10	5.72
14	4.77	5.32	4.78	5.62	5.15	5.76
15	4.79	5.19	4.81	5.58	5.08	5.72
16	4.96	5.24	4.81	5.65	5.11	5.74
17	5.24	5.35	4.80	5.55	5.08	5.67
18	5.37	5.62	4.89	5.78	5.09	5.73
19	4.95	5.36	4.76★	5.63	5.10	5.61
media ± sd	4.99±0.19	5.30±0.14	4.82±0.04	5.65±0.05	5.13±0.04	5.69±0.05

Anexo K. Áreas e ingresos (Exp. 10)

Recalculado desde `predictions_multitarget.csv`. MAE en llegadas/día por horizonte (+1... + 7); MAPE a +1 d; cobertura de la banda P10–P90 (%) a +1 d y media de +1... + 7.

val 2025-H2

objetivo	media/día	MAE +1	MAE +2	MAE +3	MAE +4	MAE +5	MAE +6	MAE +7	MAPE	cob. +1	cob. med
Resto	395	20.4	21.6	21.8	22.3	22.5	22.5	23.3	5.2 %	79.2	72.7
Pediatría	97	12.4	11.1	10.8	10.8	10.9	11.3	11.6	13.6 %	80.9	74.2
Gine-Obs	47	6.1	6.0	6.0	6.0	6.0	6.0	6.1	13.0 %	77.0	75.1
Ingresos	52	6.6	6.5	6.5	6.6	6.7	6.7	6.8	12.7 %	83.7	81.5

test 2026-H1

objetivo	media/día	MAE +1	MAE +2	MAE +3	MAE +4	MAE +5	MAE +6	MAE +7	MAPE	cob. +1	cob. med
Resto	400	22.1	22.3	22.9	23.0	23.1	23.4	23.2	5.6 %	78.5	73.6
Pediatría	99	14.7	13.6	13.2	13.1	13.0	12.8	12.9	16.6 %	70.7	74.3
Gine-Obs	47	6.2	6.0	5.8	5.7	5.7	5.8	5.9	13.2 %	75.7	78.2
Ingresos	55	6.4	6.3	6.2	6.2	6.2	6.2	6.3	12.9 %	77.3	78.5