IA LATAM
II Congreso Académico

SALUD Y BIENESTAR

Anticipar brotes de dengue en Argentina con aprendizaje automático

Clima, casos recientes y validación sin maquillaje

Javier Balda · Juan Caracoix · Facundo Casas

Universidad Católica Argentina · Argentina



5 — 6 NOVIEMBRE 2026 · ONLINE Y GRATUITO · congreso2.comunidadialatam.org

AGENDA

Lo que veremos hoy

01

Contexto

El dengue en Argentina y por qué predecirlo

02

Enfoque

Datos, variables y modelos

03

Resultados

Qué tan bien anticipamos y dónde fallamos

04

Aprendizajes

Límites, hallazgos y próximos pasos

01

Contexto y problema

El dengue en Argentina: de la calma al brote récord de 2024

582 mil

casos de dengue confirmados en 2024, casi cuatro veces los de 2023

Fuente: SNVS, Ministerio de Salud de Argentina · 582.265 casos en 2024 y 146.876 en 2023

819.793 casos confirmados entre 2018 y 2025.

El 71 % de la serie ocurrió en un solo año: 2024.

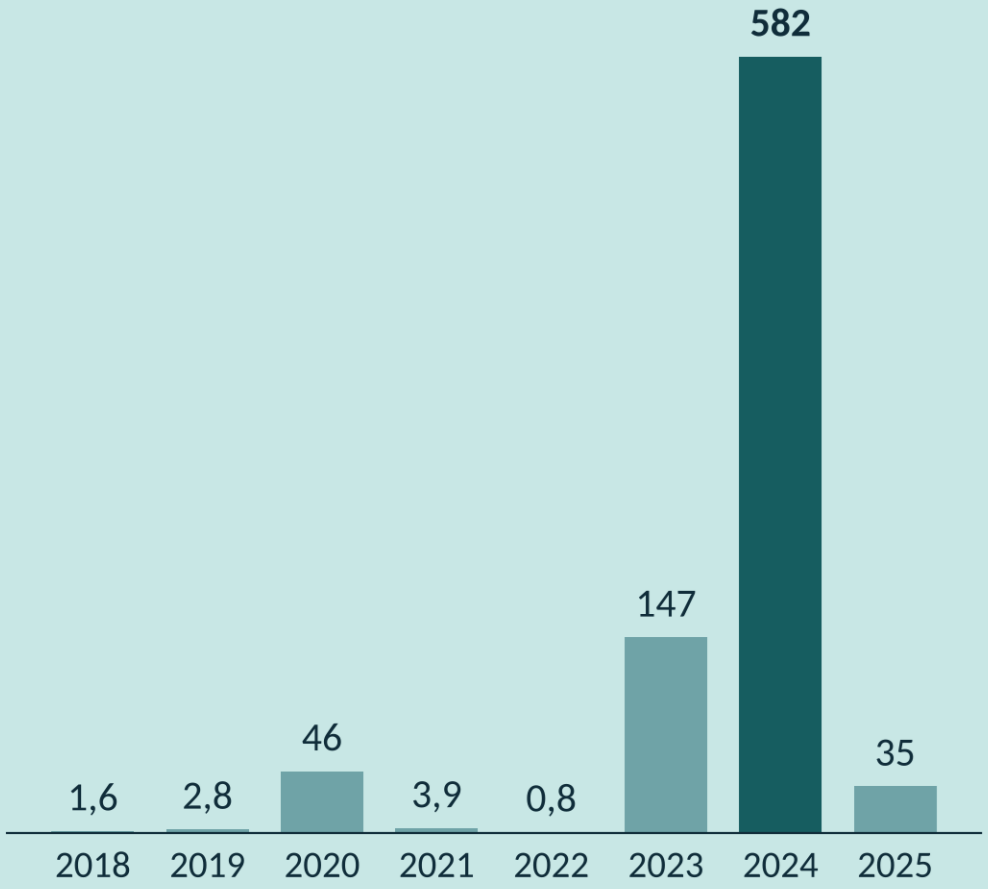
01 / CONTEXTO

De años tranquilos a un brote sin precedentes

- El dengue pasó de miles de casos a cientos de miles en dos años.
- 2024 es nuestro año de prueba: el brote más exigente para un modelo.
- 2025 volvió a 35 mil casos: la serie es muy irregular.

Casos confirmados de dengue por año

Miles de casos · Argentina, 2018–2025



Fuente: SNVS, Ministerio de Salud de Argentina

Pasar de reaccionar a anticipar

El punto de partida

- Casos por departamento y semana, publicados por el SNVS
- Series cortas: datos departamentales desde 2018
- Subregistro, sobre todo en zonas rurales

Lo que propusimos

- Pronóstico de la semana siguiente por departamento
- Casos recientes y clima rezagado como señales
- Validación temporal contra un baseline simple

02

Enfoque

Datos, variables y modelos: las decisiones de diseño

Cuatro fuentes públicas, un dataset maestro

01

SNVS

819.793 casos por departamento y semana, 2018–2025

02

NASA POWER

16 variables climáticas diarias en 24 provincias

03

Censo 2022

Población y vivienda por departamento (INDEC)

04

RENABAP

Barrios populares y acceso a agua y cloacas

Antes de modelar, hubo que ordenar los datos

- Dos esquemas de columnas en el SNVS (2018–2022 y 2023–2025), unificados automáticamente.
- 350 filas con departamento desconocido, excluidas.
- Nombres con errores de codificación y tildes, normalizados.
- Clima unido por provincia y semana: cobertura del 87 % del dataset maestro.
- Dataset maestro: 17.017 filas y 577 departamentos; 192 con al menos 30 observaciones para modelar.

Dos tipos de señales para cada departamento

Memoria epidemiológica

- Casos de las últimas 1 a 6 semanas
- Medias móviles de 4 y 8 semanas
- Máximo y variabilidad del último mes

Clima con rezago

- Temperatura, humedad, lluvia y punto de rocío
- Rezagos de 2, 4 y 6 semanas
- Base biológica: el ciclo del mosquito y del virus toma de 1 a 6 semanas

CÓDIGO

Elegimos variables sin espiar el futuro

```
# Selección solo con años < 2020: sin fuga de datos
df_fs = df_feat[df_feat['anio'] < initial_train_end]

# Tres criterios, promediados y normalizados
scores['corr']          = X.corrwith(y).abs()
scores['mutual_info']   = mutual_info_regression(X, y)
scores['rf_importance'] = rf_sel.feature_importances_

# 84 candidatas  ->  30 variables seleccionadas
```

Tres modelos frente al mismo examen



Baseline

Media móvil de 4 semanas: la vara mínima que hay que superar.



LightGBM

Gradient boosting con 30 variables: el candidato a uso operativo.



EpiGNN

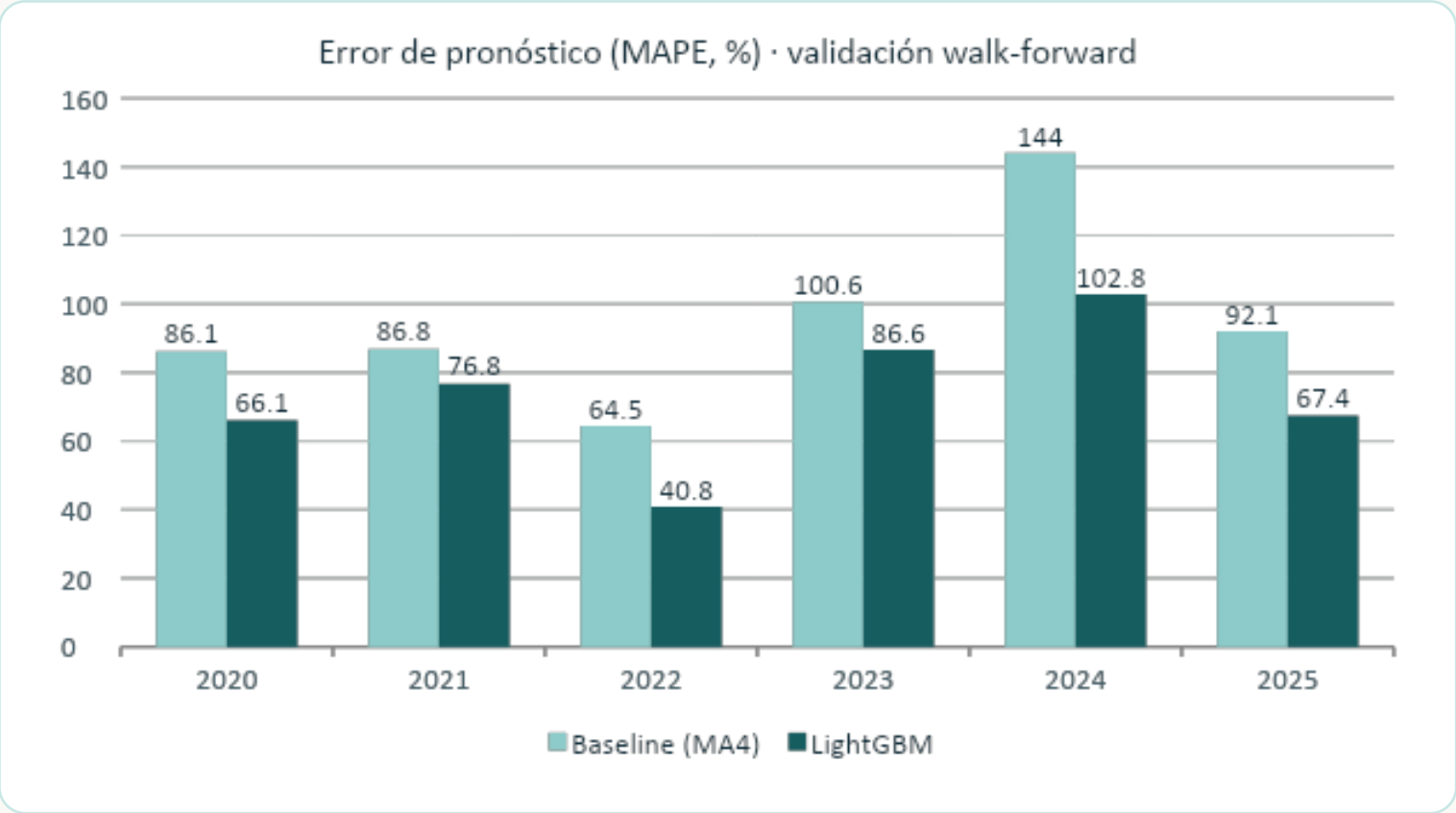
Red neuronal de grafos sobre 447 departamentos: extensión exploratoria.

03

Resultados

Qué tan bien anticipamos y dónde fallamos

LightGBM supera al baseline los seis años



LECTURA

El modelo reduce el error entre 10 y 41 puntos porcentuales cada año.

La mayor mejora es en 2024, el brote más fuerte.

En 2020 y 2024, los IC 95 % por bootstrap no se superponen.

EpiGNN gana en el papel, pero no es comparable

Modelo	MAE	MAPE (≥5 casos)	RMSE
Baseline (MA4)	67,4	144,0 %	160,9
EpiGNN (exploratorio)	14,8	83,4 %	99,6
LightGBM (recomendado)	**50,3**	**101,1 %**	**149,5**

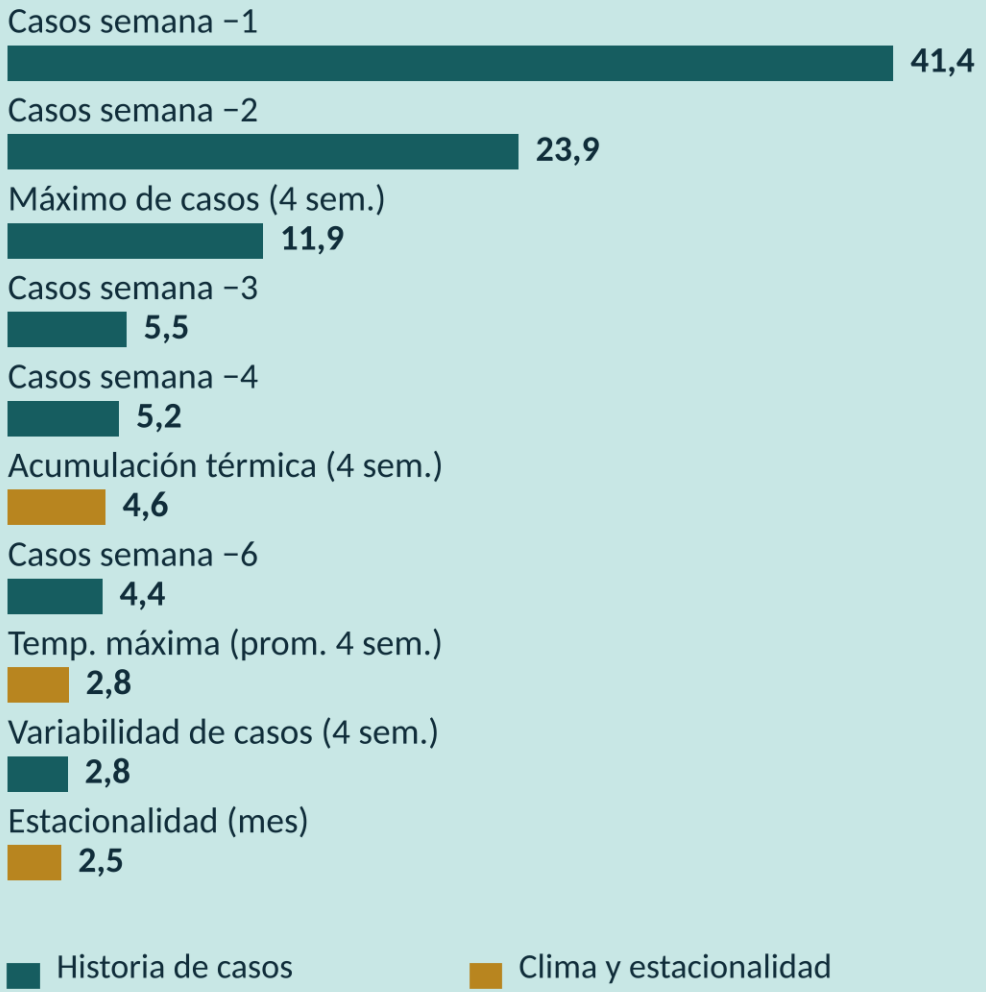
Test 2024. El MAE y el RMSE de EpiGNN se calculan sobre 447 nodos × 52 semanas con ceros imputados; los de LightGBM, sobre 3.732 observaciones reales.

Pesa más la semana anterior que el clima

- Los casos de la semana anterior son la señal dominante (SHAP).
- La primera variable climática es la acumulación térmica de 4 semanas, en el 6.º lugar.
- El clima aporta, pero en segundo plano.

Qué pesa más en la predicción

Importancia media |SHAP| · LightGBM · test 2024



8 de 167

departamentos alcanzaron el objetivo de MAPE ≤ 30 % en 2024

Fuente: test 2024 con LightGBM · MAPE sobre observaciones con ≥5 casos

Seis de los ocho son comunas de la Ciudad de Buenos Aires.

En el conjunto de 2024, el MAPE del modelo es 101 %.

La mejora es robusta, pero lejos del objetivo

- Con umbrales de 1, 5, 10 y 20 casos, LightGBM supera al baseline en todos.
- Con casos ≥ 20, el MAPE baja a 57 %, frente a 92 % del baseline.
- Sobre 2025 (844 observaciones): MAE de 25,0 y MAPE de 68 %.
- Los residuos tienen sesgo de +13 casos y no son normales: tiende a subestimar los picos.

04

Aprendizajes y límites

Lo que funcionó, lo que no y lo que sigue

Lo que el modelo todavía no sabe

- El subregistro del SNVS agrega ruido que el modelo no puede corregir.
- Sin serotipos, sin movilidad entre departamentos y sin control vectorial.
- Series cortas: datos departamentales disponibles recién desde 2018.
- Cada departamento se predice por separado; EpiGNN, que modela el contagio entre vecinos, aún es inestable.

Tres ideas para llevarse



Hallazgo

Los casos recientes pesan más que el clima; LightGBM supera al baseline cada año.



Impacto

Una alerta semanal que apoye a la vigilancia, no que la reemplace.



Próximo paso

Sumar movilidad, serotipos y control vectorial, y reentrenar cada año.

IA LATAM

II Congreso Académico

¡Gracias!

¿Preguntas?

Javier Balda · Juan Caracoix · Facundo Casas

Universidad Católica Argentina

Consultoría de datos · Laboratorio III



AUSPICIADORES

Gracias por apoyar este espacio de conocimiento abierto para Latinoamérica



OpenChapitas



TTPSEC