Final - Análisis Predictivo

PREDICCIÓN DE INCIDENCIA ANUAL DE DENGUE Y ZIKA

Geronimo Fretes

CASO DE NEGOCIO

Problema

- El dengue genera picos de demanda sanitaria y presión operativa territorial.

Decisión que habilita

- Priorización de prevención/recursos por departamento antes del pico.

Usuario final

- Gestión sanitaria / planificación territorial.

Salida del modelo

- Predicción de incidencia anual por departamento.

DATASET

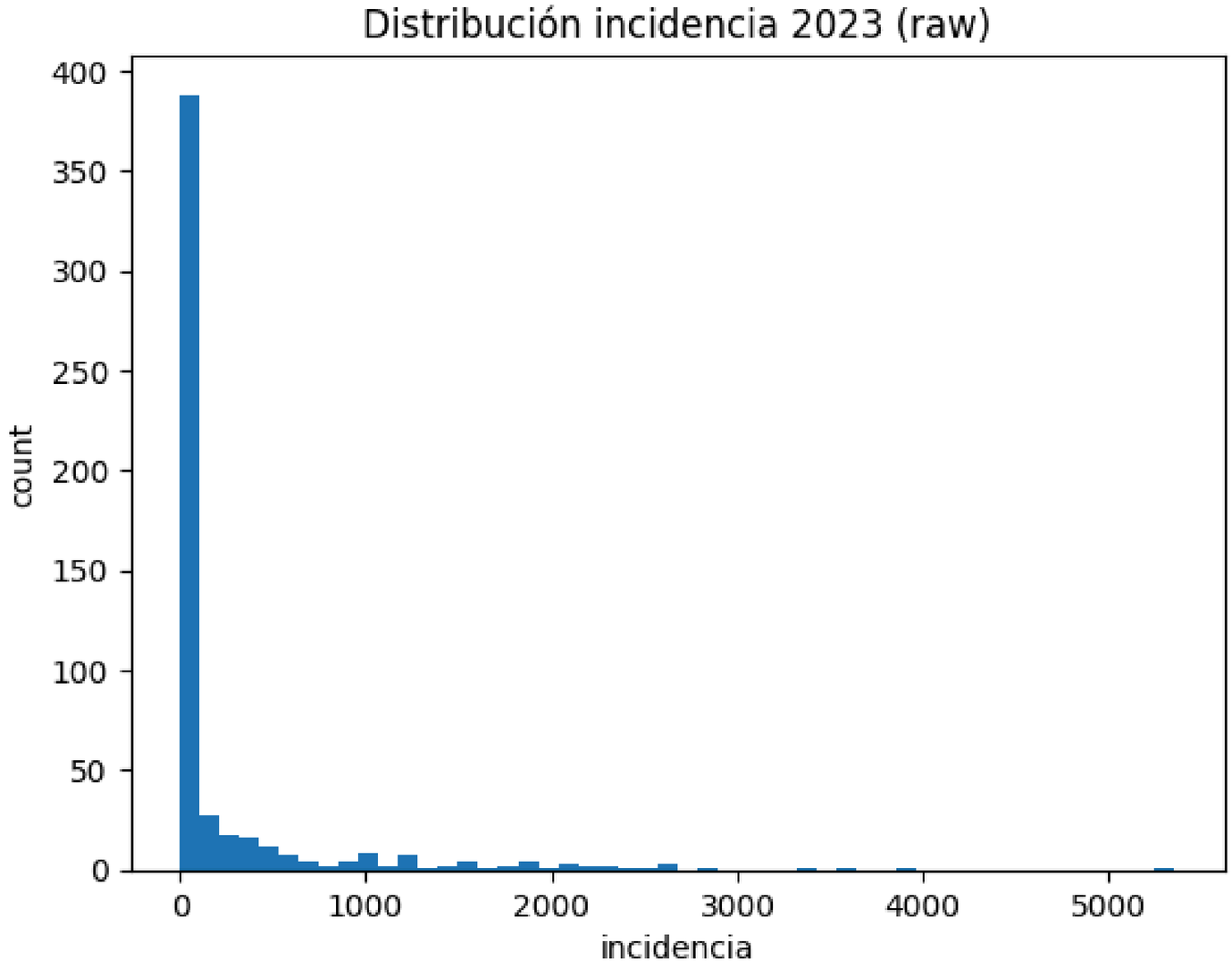
Unidad observacional: Departamento (n = 527)

Granularidad: anual, por año objetivo

Construcción: dataset integrado con variables socio-habitacionales, vulnerabilidad y ambiente + histórico dengue.

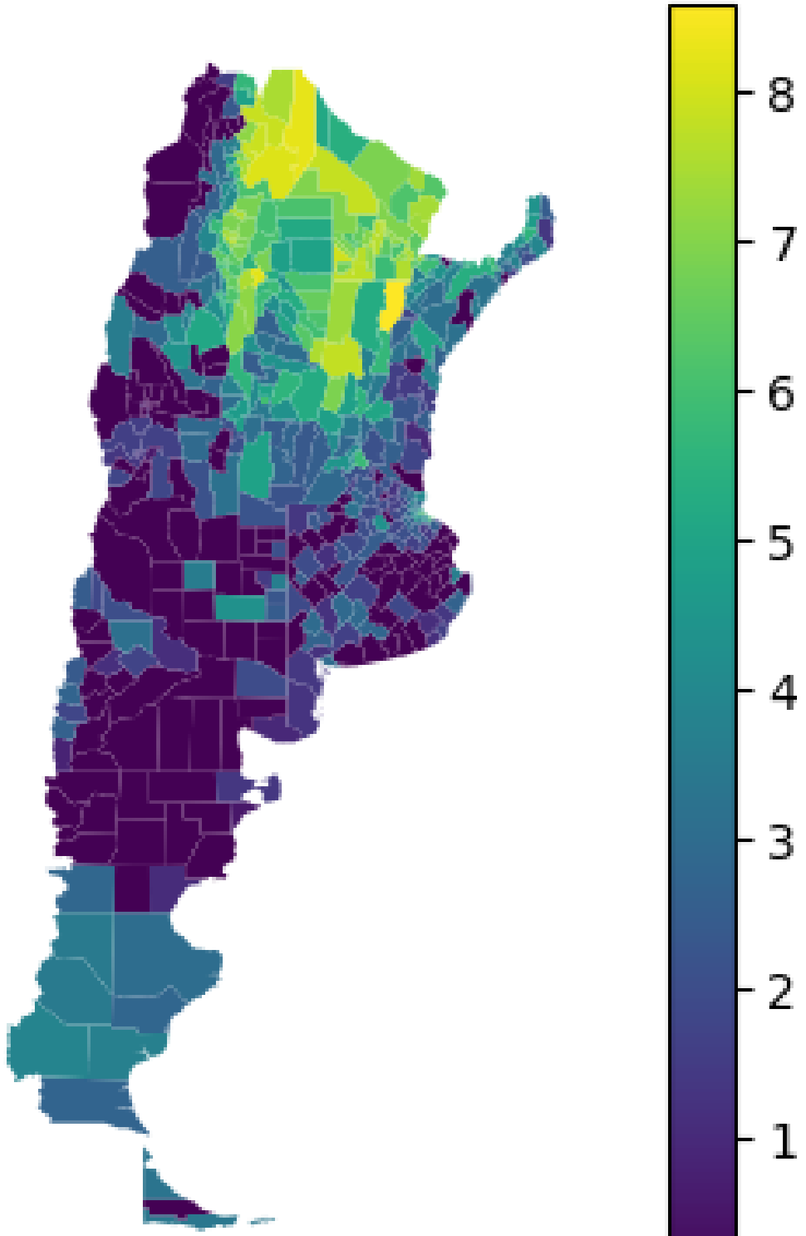
Dimensión: 223 features numéricas

ANÁLISIS EXPLROATORIO



ANÁLISIS EXPLROATORIO

ANÁLISIS EXPLROATORIO



DEFINICIÓN DEL OBJETIVO Y MÉTRICA

- **Target**

  y=log(1+incidencia anual)

  para estabilizar escala y outliers.

- **Métrica principal**

  RMSE en espacio log1p

BASELINE

Predicción: $\hat{y}2024 = y2023$

Intuición: si el dengue fuera estable, el último año sería un buen predictor.

Resultados del baseline

- RMSE = 2.9
- MAE = 2.4
- R^2 = -0.8

PARTICIÓN Y EVALUACIÓN

Evaluación interna

- CV con StratifiedKFold por bins del target.
- Cantidad de particiones = 5
- CV multi-semilla
- Score = promedio de RMSE de las semillas.

Evaluación externa

- Entrenar con datos del 2023 y obtener predicciones para el 2024
- Evaluar contra incidencia real del 2024

SELECCIÓN DE MODELOS

Búsqueda de configuraciones

Optuna optimiza simultáneamente:

hiperparámetros de CatBoost

top_k features (prefiltradas por orden “ortogonal”)

MODELO FINAL

Algoritmo: CatBoostRegressor

Features: top_k = 20

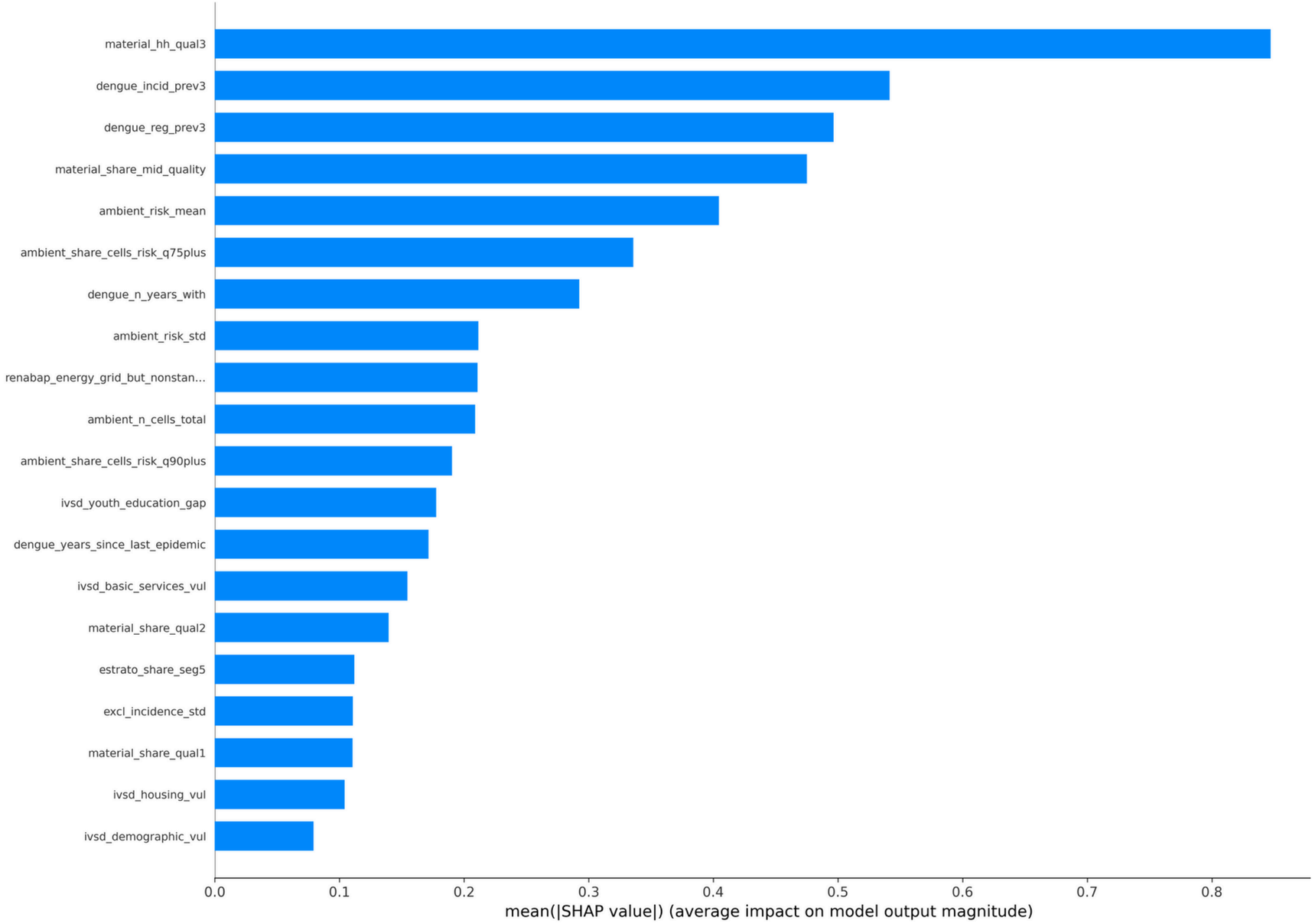
- learning_rate =0.02
- depth = 9
- l2_leaf_reg = 0.2
- bagging_temperature = 3.5
- border_count = 103
- random_strength = 2
- min_data_in_leaf = 25
- grow_policy = Lossguide
- iterations = 1332

ESTABILIDAD DEL MODELO FINAL

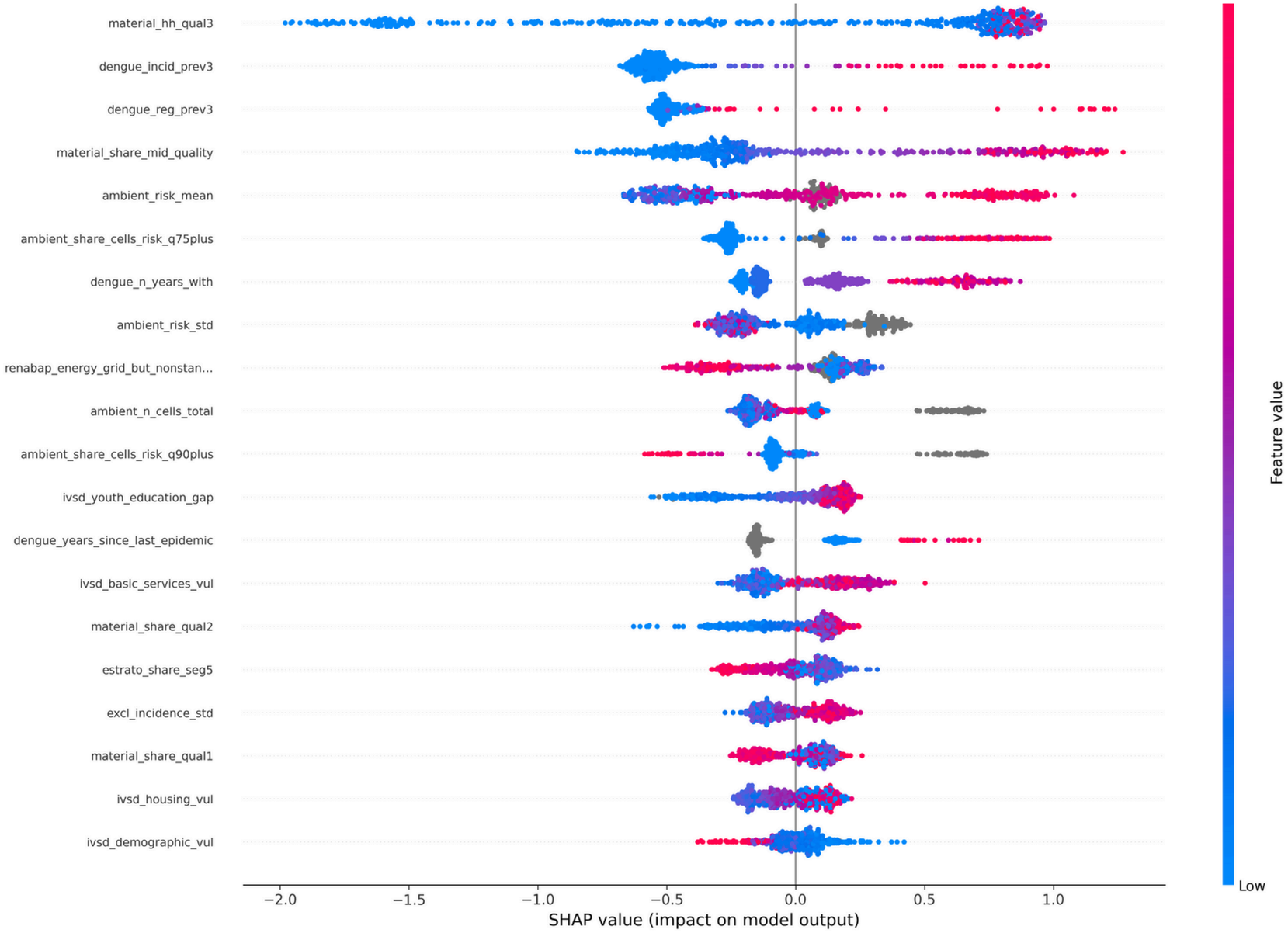
Estabilidad bajo particiones alternativas (20 repeticiones)

- RMSE promedio: 1.2486
- Desvío: 0.0103
- IC 95%: [1.2441, 1.2531]

INTERPRETABILIDAD



INTERPRETABILIDAD



EVALUACIÓN EXTERNA VS BASELINE

Predicciones 2024

- RMSE = 3.1
- R^2 = -1

Baseline

- RMSE: 2.9
- R^2: -0.8

LIMITACIONES Y MEJORAS

Limitaciones

- Generalización temporal no garantizada
- Features estáticas mayormente del 2022
- Cambio de régimen en 2024

Mejoras concretas

- Entrenar multi-año (2018–2023) con validación temporal
- Sumar drivers temporales como el clima (lluvia/temp)