RESEARCH NOTE / RELEASE 1.2 / 27 OCTOBER 2026

NiñoSalud AI

Testing climate-informed surveillance with public data from northern Peru

Cristian Daniel Medina Gomez
Independent undergraduate-level research and learning project. Background in Economics and Management at Bocconi. Developed with AI support; no institutional endorsement.

Abstract

NiñoSalud is a reproducible, exploratory prototype for evaluating the usefulness and limits of public-health data and machine learning. The main benchmark compares four-week-ahead predictions of reported acute diarrheal disease (EDA) episodes in Piura's eight provinces. On 992 matched retrospective targets in 2023–2025, persistence achieved a mean absolute error (MAE) of 45.088 episodes, compared with 52.763 for a Random Forest without climate and 50.294 with climate. A separate regional dengue audit showed that date-interval summaries depend materially on inclusion thresholds. These results support transparent source checking and baseline comparison, not operational early-warning claims. The dashboard displays saved analyses; it is neither a clinical product nor a forecast of future epidemics.

1. Motivation and questions

The project connects an undergraduate quantitative and management background with an emerging interest in data science and health sciences. Peru provides a concrete setting in which epidemiological records, weather observations and territorial exposure can be studied together. El Niño motivates the preparedness question, but no effect is attributed to El Niño solely from contemporaneous trends. The general need to assess model inputs, validation and uncertainty is consistent with WHO/TDR and US CDC guidance [S9–S10].

Two questions organize the work: (i) does adding climate improve a fixed predictive algorithm, and does that algorithm beat a simple baseline? (ii) how stable are regional surveillance indicators when the rules used to interpret recorded dates change?

2. Data and analytical units

Component	Unit / scope	Role
Primary EDA + weather [S1]	3,344 province-week rows, 2018–2025; eight Piura provinces	Forecast benchmark
CDC age-group summaries [S2]	Annual Piura totals; both age groups	Independent export reconciliation
Regional dengue [S5]	Separate cohorts 2023, 2025 and partial 2026	Recorded-date audit
Territorial snapshots [S6–S8]	District risk, population and registered facilities	Context only; not model inputs

Counts refer to reported episodes or records, not necessarily unique people or independently confirmed infections. Province and department are distinct geographic levels. Source snapshots and definitions are retained rather than silently harmonized.

3. Primary EDA experiment

For origin week t, the target is the count in t+4, not a four-week cumulative total. Inputs include current and prior EDA counts, a trailing mean, the previous year's target-week count, seasonal terms and province indicators. The climate variant adds observed temperature, humidity, rainfall, short trailing summaries and source-reference anomalies. No measured weather after t is used.

Both forests use 140 trees, maximum depth 12, minimum leaf size 5 and seed 42. Training uses 1,408 retained rows in 2019–2022. The common evaluation sample contains 352 rows in 2023, 352 in 2024 and 288 in 2025. Same-year lag/target availability excludes boundary weeks. Missing humidity in source weeks 19–22 of 2025 also excludes dependent origins. This conservative complete-case rule is shared by every method.

The evaluation period has already informed exploratory development. It is not an untouched final holdout. Source reporting vintages and weather availability were not reconstructed as they would have existed in real time. Performance therefore describes a retrospective experiment, not an operational backtest.

Table 1. Mean absolute error on the same 992 targets

Method	MAE
Persistence: last observed week	45.088
Trailing four-week mean	46.673
Random Forest + observed climate	50.294
Random Forest without climate	52.763
Same target week in the previous year	71.954

Climate reduced the forest's error by 2.469 episodes, approximately 4.7%, relative to the otherwise matched no-climate forest. The climate forest nevertheless had 5.206 more MAE than persistence. Equal weighting of province-week rows does not remove the influence of large-count provinces on absolute errors. The repository also reports year- and province-specific results.

4. Uncertainty and supporting benchmark

A paired bootstrap resampled 24 province-year sequences 5,000 times (seed 20261027). The approximate 95% percentile interval for climate-forest MAE minus persistence MAE is −0.474 to +13.349. Resampling blocks retains within-sequence dependence, but provinces and years may remain correlated. The interval is descriptive, not a formal equivalence test or a prediction interval for future outcomes.

An auxiliary department-wide dengue benchmark uses different units and years: training 2018–2022, 88 evaluation rows in 2023–2024. Persistence MAE is 805.9 versus 911.5 for its climate forest. It remains a separate supporting experiment, not additional evidence pooled with the primary EDA test.

5. Reconciliation before modelling

For Piura in 2024, the historical national EDA file [S4] contains 15,451 episodes in the age-5+ field, compared with 57,363 in a newer CDC export [S2]. The historical file's later weeks contain a prolonged zero sequence. The discrepancy is documented rather than interpreted as a sudden disease decline. The raw file is retained unchanged outside the public release. Annual totals are not redistributed across weeks to create artificial observations.

The primary model instead uses the CDC province-week table [S1]. Its annual provincial sums reconcile with the supplied age-group summaries for 2020–2025, including 90,579 total episodes in 2024 and 82,623 in 2025. Agreement between exports is a consistency check, not independent verification of the underlying surveillance system.

6. Regional dengue: definitions change the question

The GORE Piura CSV [S5] contains 47,082 records before geographic filtering, with non-contiguous cohorts: 2023, 2025 and 2026. The 2023 snapshot is partial, 2024 is absent and 2026 is partial. The supplied dictionary defines FECHA_NOTIFICACION as the recorded attendance date at a health centre. It cannot automatically be treated as verified first contact or central reporting time. The source does not provide a per-row diagnostic-status field for matching to national confirmed-only counts.

For the supplied 2025 cohort, nonnegative differences between attendance and symptom-onset dates were summarized under several maximum-inclusion thresholds. A same-calendar-year restriction was also evaluated. Numerical plausibility does not establish validity of the dates.

Table 2. Median recorded interval by inclusion cutoff

Maximum included interval	Piura: days	Huancabamba: days
30 days	3	5
60 days	3	10
90 days	3	15
180 days	3	18
365 days	3	19

At the 0–90-day cutoff, 5,468 of 5,507 Piura records and 249 of 275 Huancabamba records are retained. The between-province contrast is visible, but its magnitude is cutoff-sensitive. It should not be presented as proof of delayed treatment, unequal access or a causal effect of climate.

Descriptive seasonal-reference plots use only pre-evaluation observations, but include pandemic years in the reference. Exceeding a historical percentile is not a confirmed outbreak. No independent outbreak labels, validated detection sensitivity or verified early-warning lead time are available.

7. Reproducibility and verification boundary

The public bundle contains aggregate data, analytical scripts, model outputs and tests. The core EDA input was reconstructed from the supplied source workbook and matched all 3,344 rows. The matched forests, bootstrap, sensitivity tables and auxiliary dengue results were recomputed during release preparation. Patient-level regional source records are not distributed; rerunning that component requires separately acquiring the identified snapshot.

The static interface displays saved calculations. It does not train AI in the browser or connect to live patient systems. Automated browser checks load the exact bundled HTML and JavaScript in Chromium; local URL navigation is restricted in the preparation environment, so hosted GitHub Pages and final account workflows still require an owner-side check. Legacy map geometry and some contextual views are archived extracts rather than a complete raw-geospatial rebuild.

8. What this project does and does not establish

The work demonstrates a traceable exploratory workflow and an interpretable negative result: climate features helped the tested forest relative to its no-climate variant, but did not establish superiority over persistence. Missingness, revised surveillance, coarse aggregation, correlated samples and repeated inspection of evaluation years constrain generalization. Date-interval sensitivity and territorial exposure cannot establish clinical or causal effects. No 2027 forecast, patient advice or operational alert is offered.

A future study should freeze the protocol before testing new time periods or regions, investigate data availability at forecast time, and obtain public-health feedback. These are proposed extensions, not completed validations. Generative-AI assistance in implementation and writing is documented; the project remains independent undergraduate-level work.

References and source inventory

[S1] CDC Perú. Observatorio de Clima y Salud. Supplied workbook: clima_departamento.xlsx.
[S2] CDC Perú. Sala Situacional EDA. Exported age-group indicators, Piura.
[S3–S4] MINSA. Dengue, EDA and zoonosis surveillance catalogues. Historical supplied CSVs.
[S5] GORE Piura. Casos de Dengue en la Región Piura. CSV and accompanying data dictionary.
[S6–S8] CENEPRED risk scenario; INEI district population projections; SUSALUD RENIPRESS. Source-specific terms apply.
[S9] WHO/TDR (2017). Operational Guide: EWARS for Dengue Outbreaks. ISBN 9789241512053.
[S10] US CDC CFA (2026). How to assess a modeling framework and its results.

Full source URLs, snapshot hashes, eligibility rules and exact rerun commands: docs/DATA_SOURCES.md, docs/MODEL_CARD.md and docs/REPRODUCIBILITY.md in the repository. References do not imply source-provider endorsement.