Dengue Risk Analysis in Latin America (2020–2024)

Socio-environmental Determinants of Dengue: A Predictive Modelling Perspective

Giuseppe Leonardi
Data Science – UNICT
Data Analysis for Public Health

Project Goals

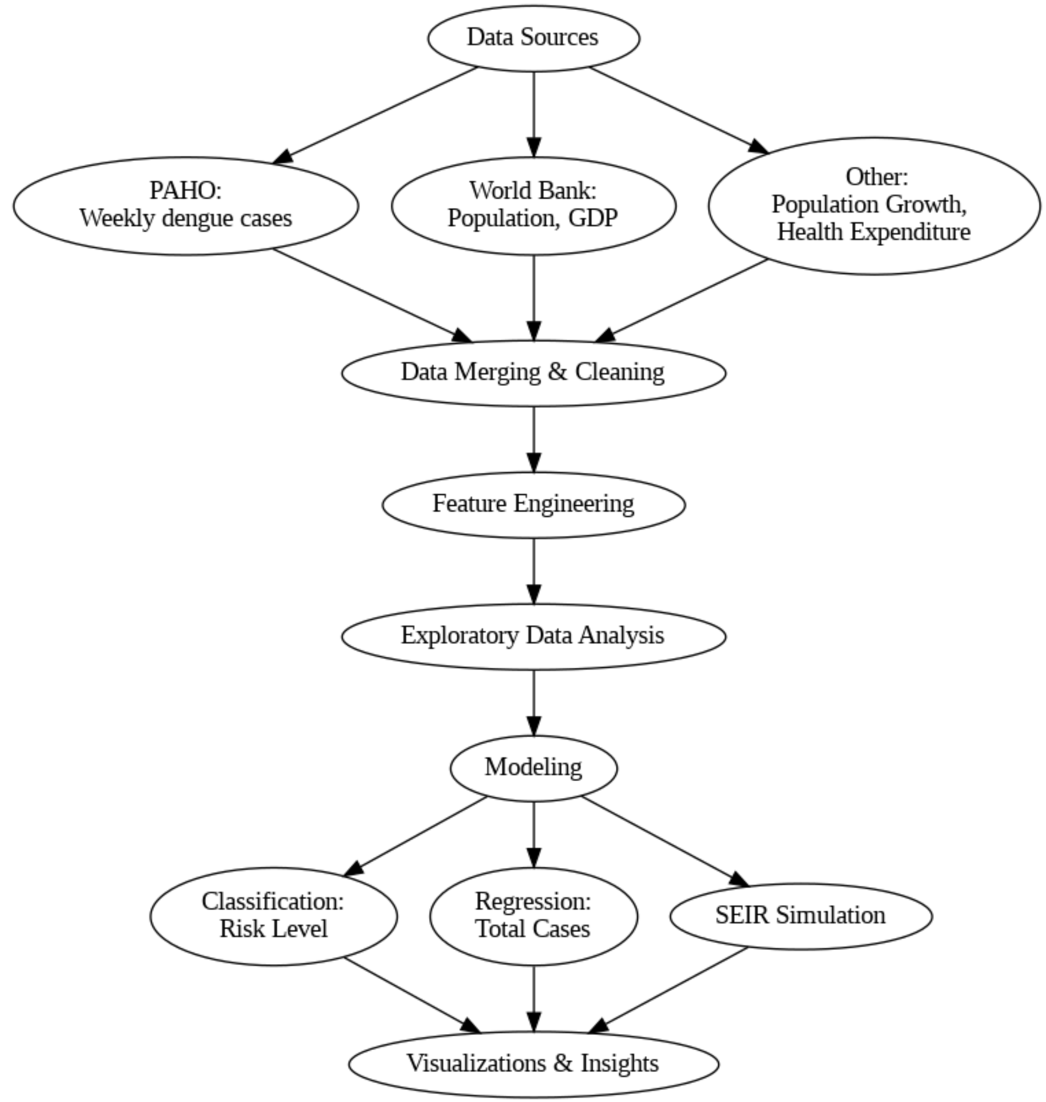
- Understand the evolution of Dengue in Latin America (2020-2024)
- Identify key socio-environmental determinants of Dengue
- Build predictive models to classify and estimate Dengue risk and incidence

Problems

- Difficulty in obtaining data in the same format (each country collects data differently)
- No environmental data found (30-year historical averages)
- Excessive fragmentation in data collection and many datasets are behind paywalls

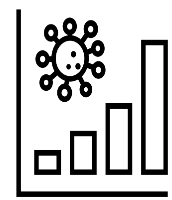
Pipeline

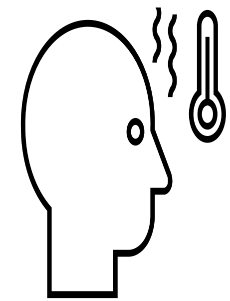
- Epidemiological data (PAHO): dengue cases, severe cases, deaths;
- Socioeconomic & environmental data: World Bank & others;
- Derived indicators: incidence rate, CFR, mortality rate, risk level;
- 10 countries × 5 years = 50 observations



Data Preparation & Feature Engineering

- Populations manually added from official sources (Dati Macro, Statista, Worldometer);
- Calculated epidemiological metrics:
  - $incidence_{100k} = \left(\frac{total\_cases}{population}\right) * 100000$
  - $mortality\_rate_{100k} = \left(\frac{deaths}{population}\right) * 100000$
  - $cfr_{100k} = \left(\frac{deaths}{total\_cases}\right) * 100000$





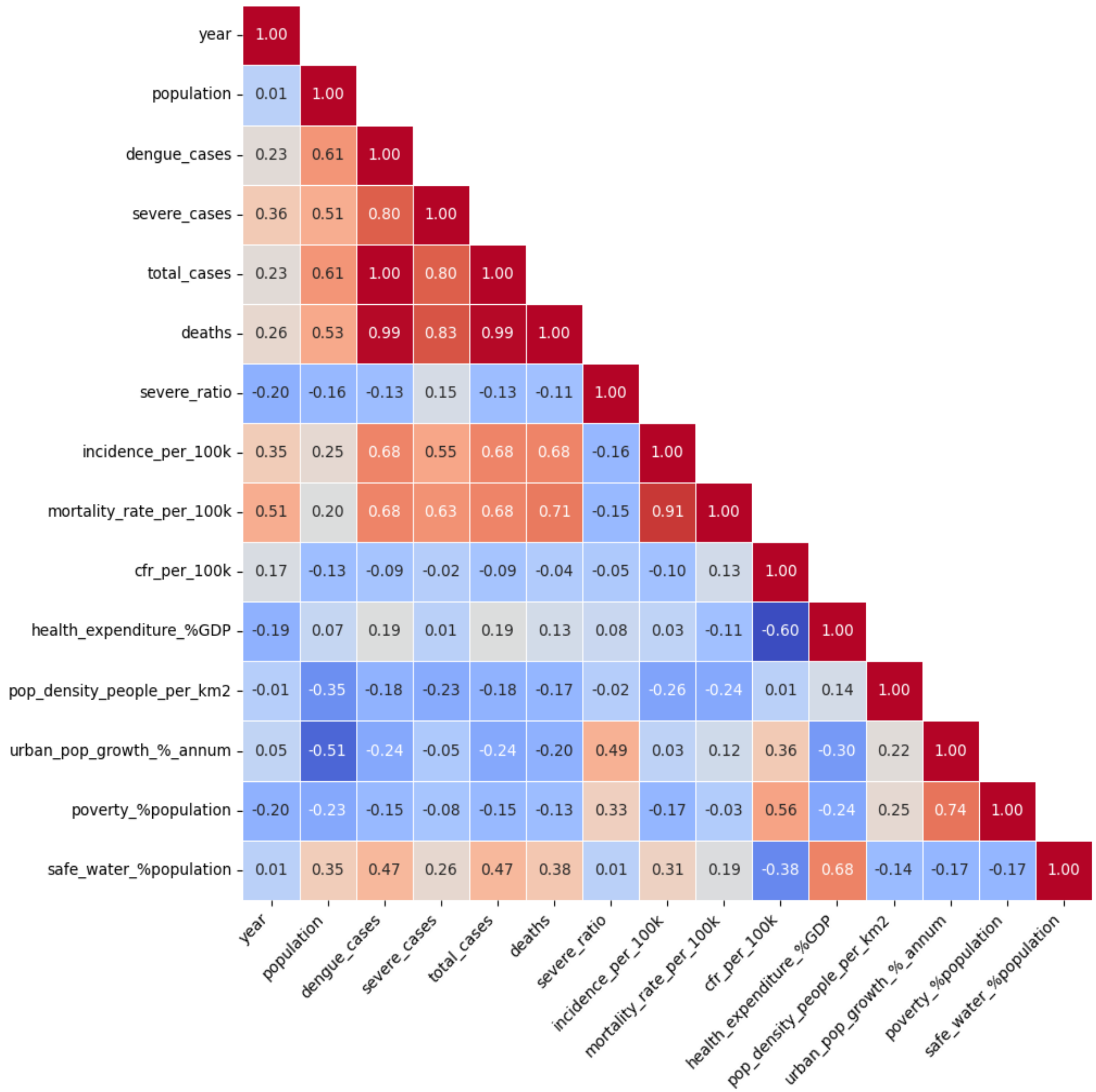
- Created risk level with quantile-based classification on incidence
- Socioeconomic indicators cleaned and merged:
  - Health expenditure, urban growth, density, poverty, safe water
  - Missing data estimated using 10-year historical averages
- Merged all tables by country and year, with standardization and renaming of features

Variables in final dataset:

- country, year, region, population, dengue_cases, severe_cases, total_cases, deaths, severe_ratio
- incidence_per_100k, mortality_rate_per_100k, cfr_per_100k, risk_level
- health_exp_%GDP, pop_density_people_per_km2, urban_pop_growth_%_annum, poverty_%population, safe_water_%population

Exploratory Analysis & Correlations

- Strong correlation: dengue_cases ≈ deaths ≈ incidence
- Negative correlation: health expenditure vs CFR
- Poverty & urban growth show weak positive links to dengue burden
- Safe drinking water shows correlation with dengue indicators



Spearman Correlation

- Very strong positive correlations among dengue_cases, total_cases, deaths, and incidence: All indicators describe a shared disease burden (ρ > 0.95).
- Negative correlations in health_expenditure_%GDP vs. cfr_per_100k (ρ = –0.65) & population_density vs. dengue_cases (ρ = –0.49)

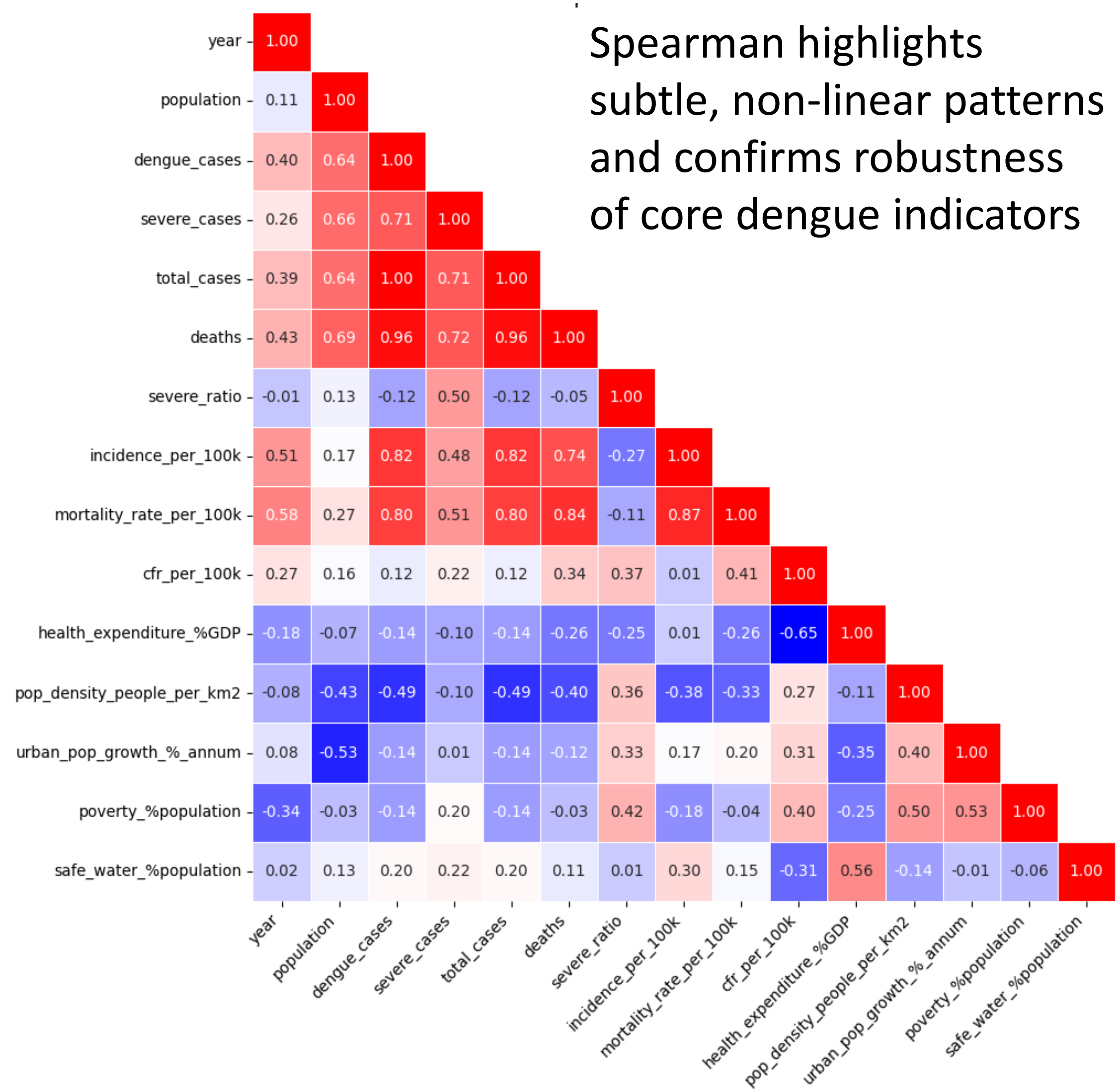
- Suggests protective roles of health spending and urban infrastructure.

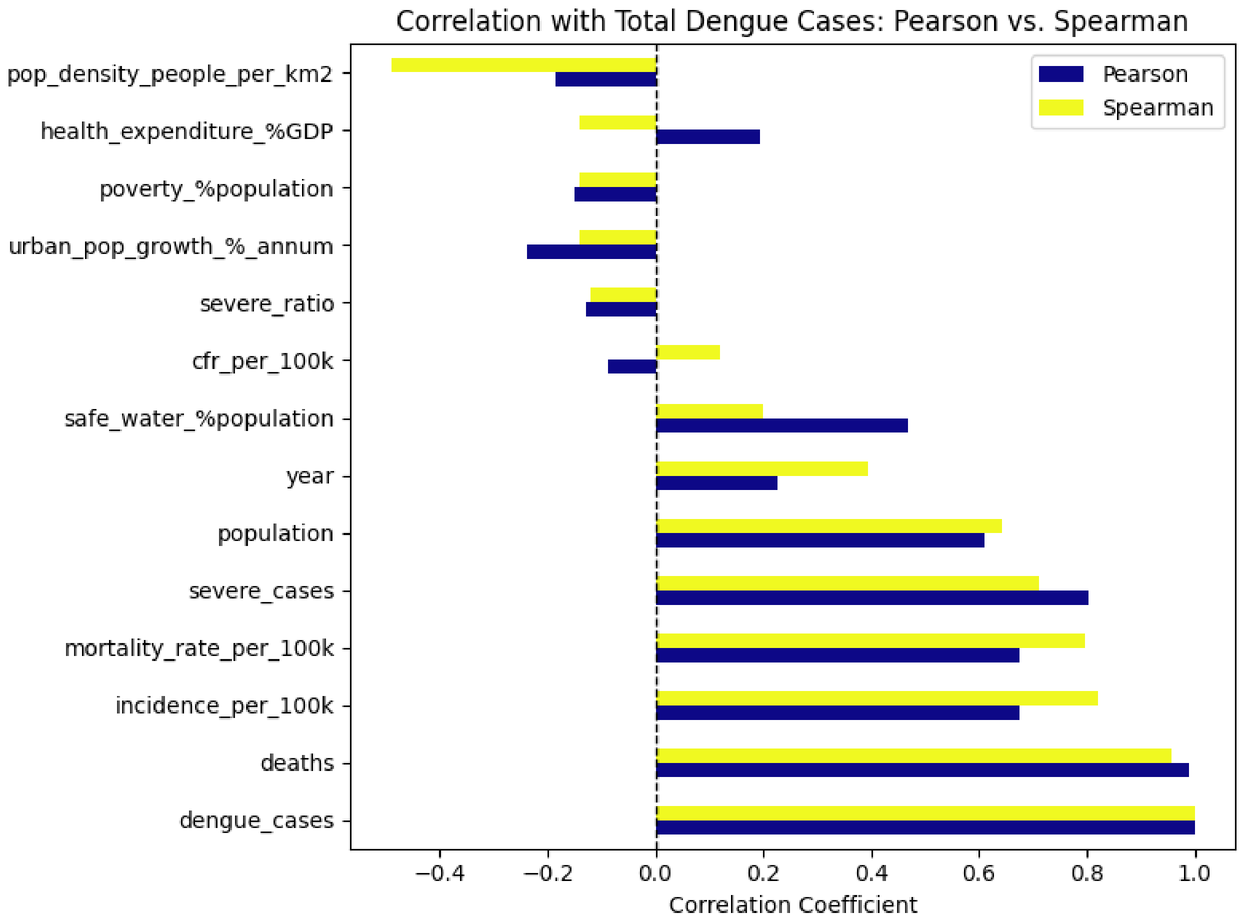
- severe_ratio increases with urban_growth (ρ = 0.33) and poverty (ρ = 0.42)

- Highlights the social determinants of disease severity.

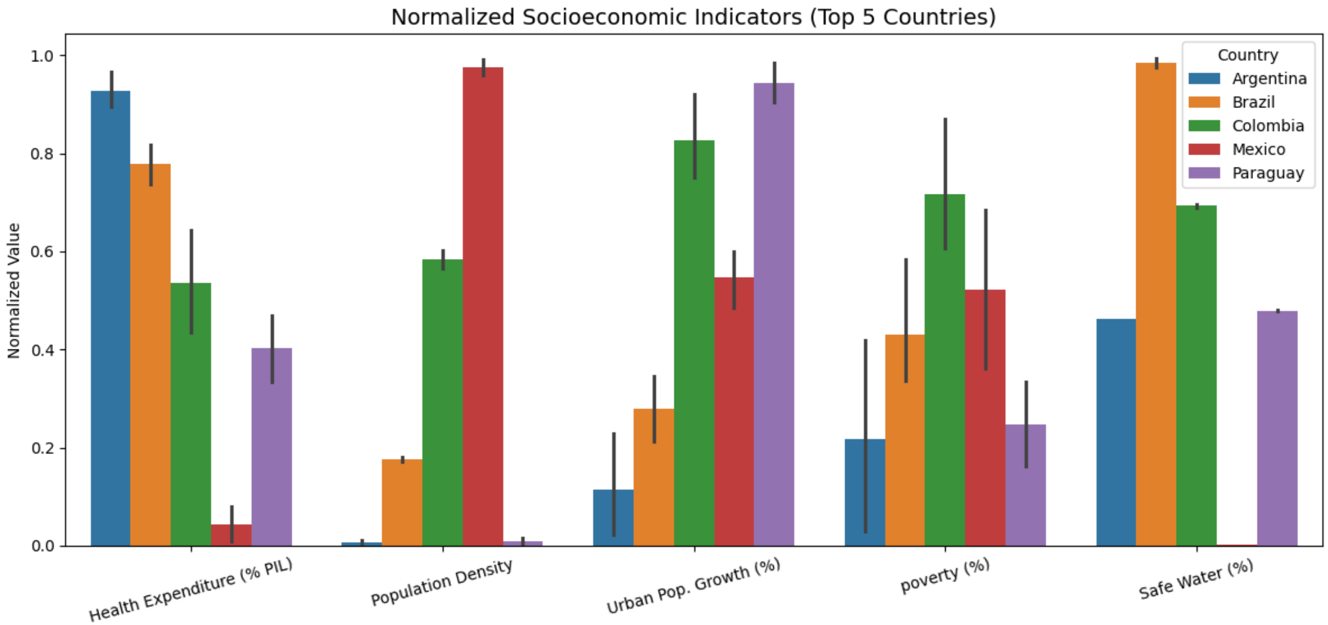
- Access to safe water correlates with better surveillance:→ ρ = 0.30 with incidence, ρ = 0.56 with health_expenditure

Spearman highlights subtle, non-linear patterns and confirms robustness of core dengue indicators



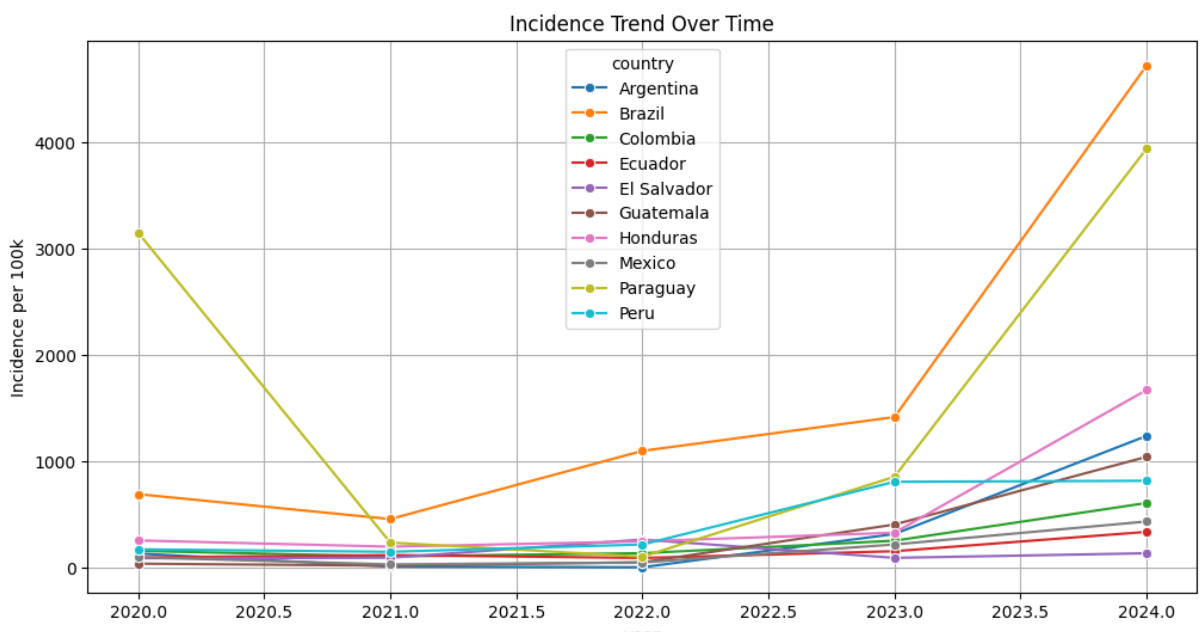


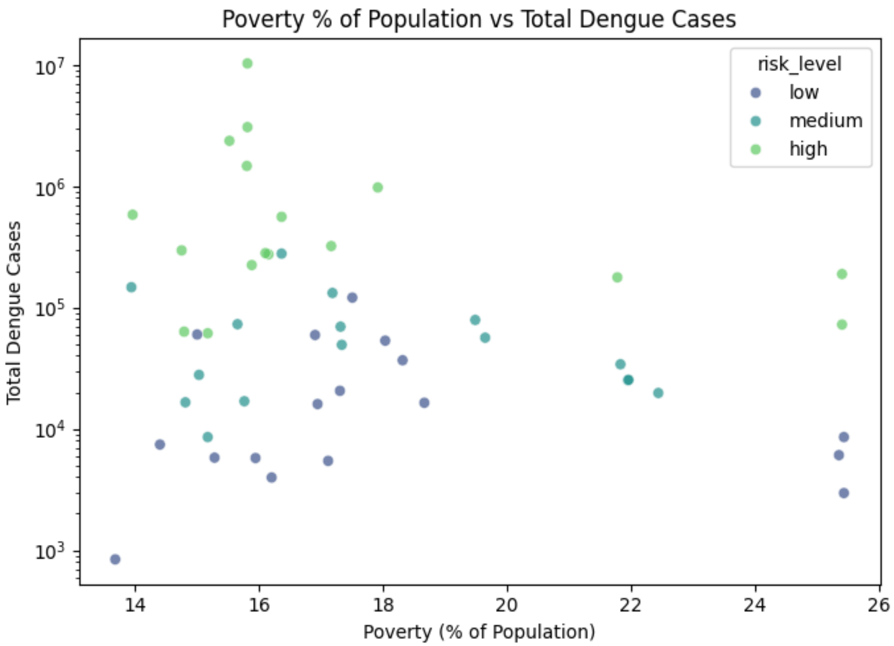
The bar chart shows strong positive correlations between total_cases and key variables using both Pearson and Spearman methods. Pearson generally reports higher correlations, reflecting linear relationships, while Spearman's lower values suggest some non-linear patterns or outliers.

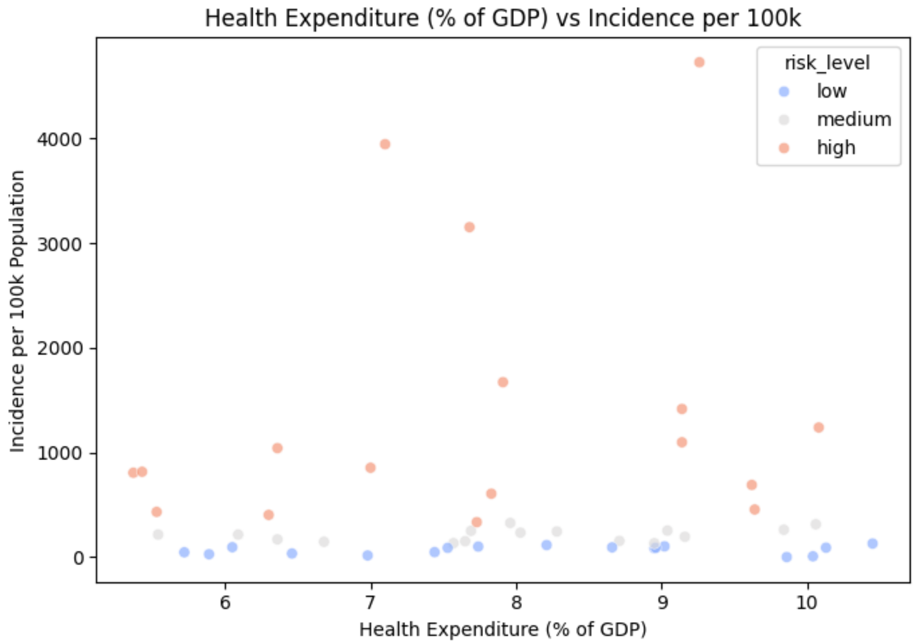


This chart compares normalized indicators across the top 5 acted countries. Patterns vary e.g., Paraguay shows high poverty and urban growth, while Brazil has moderate values across most metrics.

The time series reveals diverging trends between countries. Some show persistent increases over the years, while others have more erratic or declining patterns.







This panel explores how structural conditions relate to dengue burden across countries:

- Poverty shows a possible link with higher case numbers, especially in high-risk zones.
- Health spending exhibits a weak inverse trend with incidence, suggesting limited protective effects.
- Access to safe water appears loosely connected to lower mortality, hinting at infrastructure's role.

Though none of the trends are strongly linear, these patterns suggest that socio-economic vulnerabilities may contribute to worse dengue outcomes.

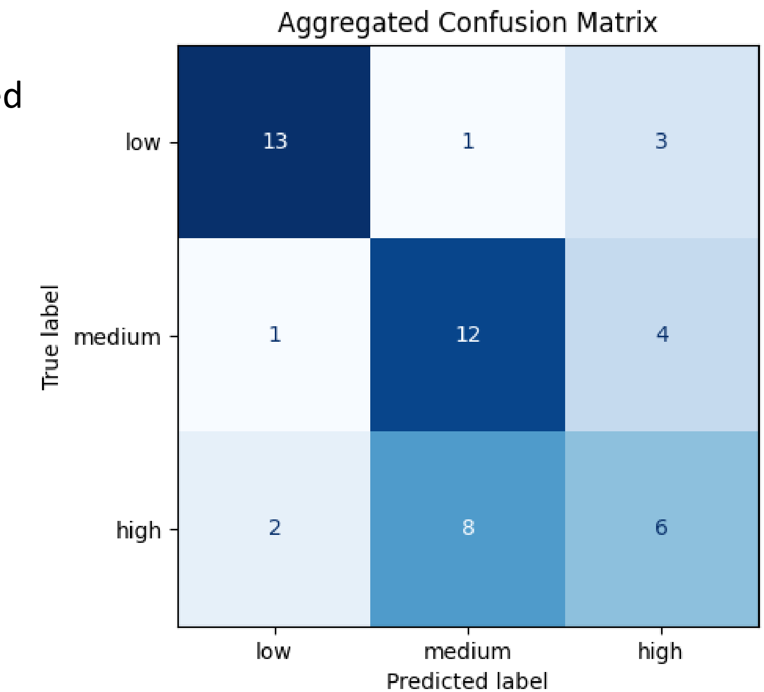
Classification on Dengue Risk

• *Objective:*
Predict categorical dengue risk levels (**low**, **medium**, **high**) based on socio-environmental and epidemiological indicators.

• *Models evaluated:*
<u>Logistic Regression </u>(baseline, interpretable)
<u>Random Forest </u>(nonlinear, robust to noise)
<u>XGBoost</u> (optimized boosting, high performance)

• Results overview
- Best performance: *Logistic Regression*
Accuracy ≈ 64%, balanced across all classes
- XGBoost performed better in identifying low-risk areas, but struggled with high-risk minority class
- All models trained on standardized, engineered dataset (18 features × 50 samples)

• Insights
- Risk levels can be predicted with moderate accuracy using structural variables alone
- Classification is sensitive to class imbalance (few “high” risk entries)
- Suggests real-world dengue risk has strong deterministic patterns, despite low sample size



Classification Report (aggregated):

	precision	recall	f1-score	support
low	0.81	0.76	0.79	17
medium	0.57	0.71	0.63	17
high	0.46	0.38	0.41	16
accuracy			0.62	50
macro avg	0.62	0.62	0.61	50
weighted avg	0.62	0.62	0.62	50

Regression on Dengue cases

- Objective:

Estimate the actual number of dengue cases (total_cases) based on structural, demographic predictors.

- Models evaluated:

- Linear Regression (fast, interpretable)
- Random Forest (interactions and nonlinearity)
- XGBoost (optimized for predictive accuracy)

- Best model:

XGBoost Regressor

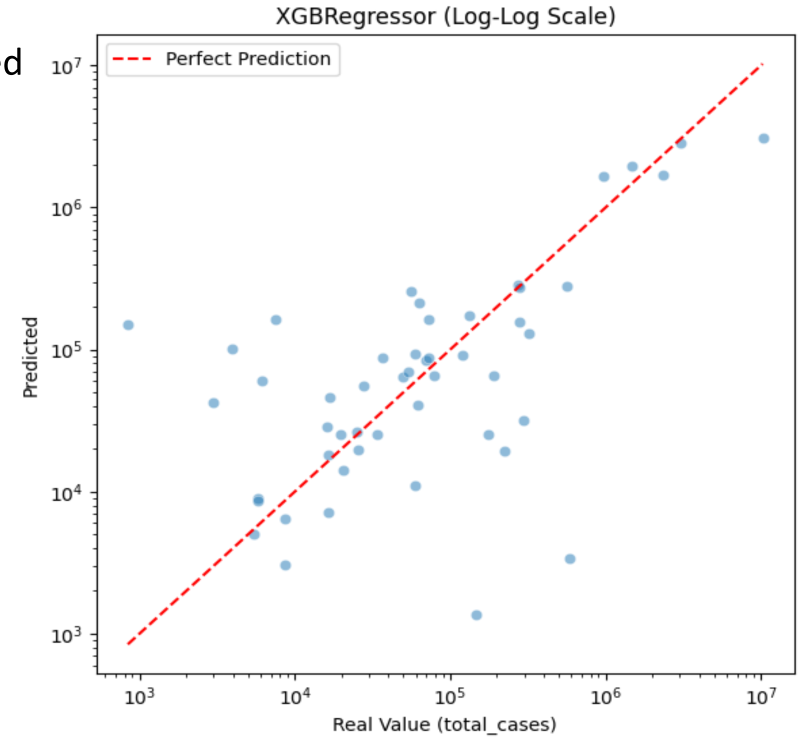
$R^2 \approx 0.52$, robust across folds

- Models performed better at estimating medium/low case loads, with some underestimation on extreme peaks

- Residual analysis shows performance drop in very high incidence scenarios

- Insights:

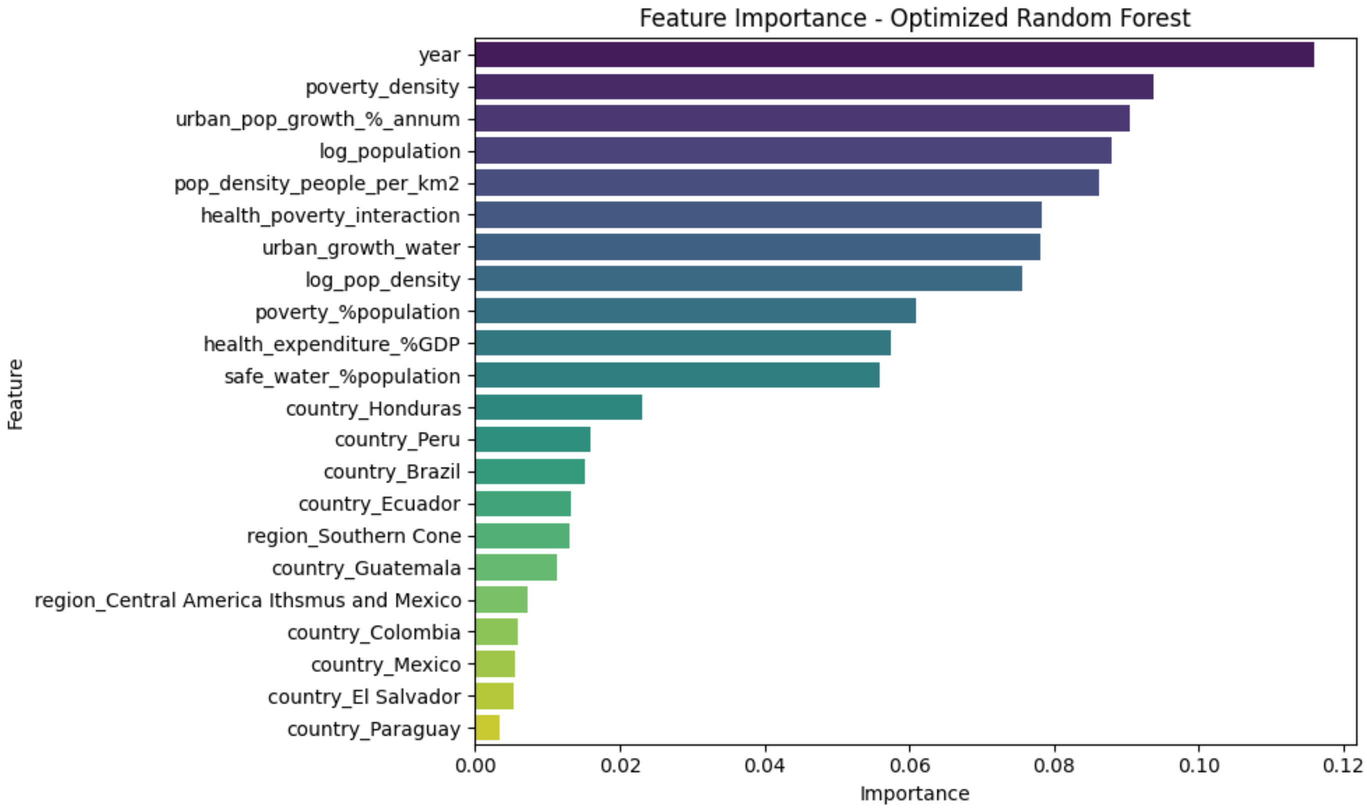
- Despite the small dataset, regression models capture meaningful trends
- Applied log-transformation to total_cases to reduce skewness and improve learning stability
- Log-scaling is essential when dealing with skewed public health data




```
=== XGBoost (Regression) ===
=== Logarithmic scale metrics ===
MAE (log): 1.08
RMSE (log): 1.60
R² (log): 0.068

=== Real scale ===
MAE: 254042.02
RMSE : 1038093.41
R²: 0.529
```

Feature Importance



- Top predictors (both tasks):
  - Urban population growth
  - Population density
  - Poverty density (poverty × density)
  - log_population

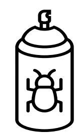
- Structural and socioeconomic factors dominate

SEIR Epidemiological Model

This SEIR extension incorporates both human and mosquito populations to capture transmission loops and biological delays.

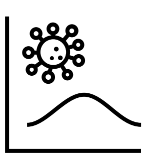
Model Structure:
Includes both human and mosquito with these compartments:
- **Humans:** S → E → I → R
- **Mosquitoes:** Sv → Ev → Iv

Key Parameters:
- $\boldsymbol{\beta_{hv}}$: vector-to-human transmission
- $\boldsymbol{\beta_{vh}}$ human-to-vector transmission
- $\boldsymbol{\sigma}$, $\boldsymbol{\sigma_{v}}$: incubation rates (humans and vectors)
- $\boldsymbol{\gamma}$: human recovery rate
- $\boldsymbol{\mu_{v}}$: mosquito mortality

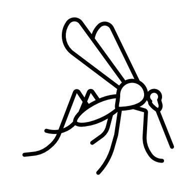


Simulation Goals:
Explore epidemic behavior in country-specific scenarios, assess impact of interventions, understand role of incubation delays and transmission loops.

Insights:
- Vector biology critically shapes outbreak size and timing
- Small changes in $\boldsymbol{\beta}$ or $\boldsymbol{\mu_{v}}$ alter transmission cycles significantly;
- Control strategies targeting vectors (e.g., insecticides) are highly effective;



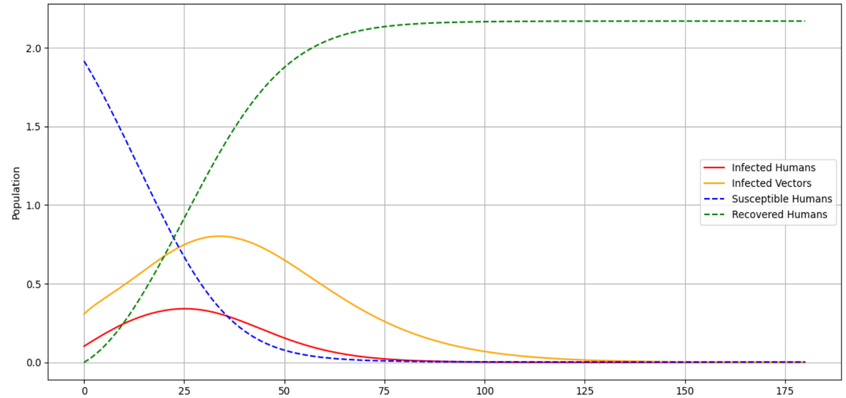
- Transmission efficiency and vector lifespan shape the epidemic.
- Targeting mosquitoes can suppress outbreaks (even in the absence of vaccines).



Brazil 2024: SEIR Simulation Scenarios

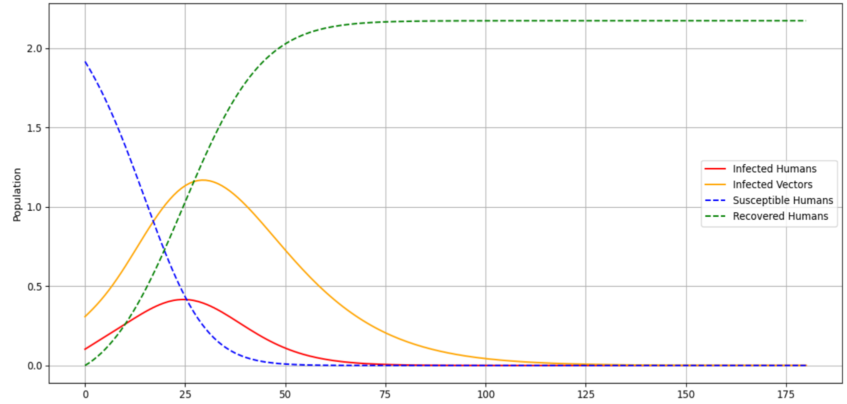
Scenario 1 – Baseline Dynamics

- $\beta_{hv} = 0.3, \beta_{vh} = 0.2, \sigma = \frac{1}{5}, \gamma = \frac{1}{7}, \sigma_v = \frac{1}{10}, \mu_v = \frac{1}{14}$
- Gradual rise and fall of infections in both populations
- Peak delay between humans and vectors



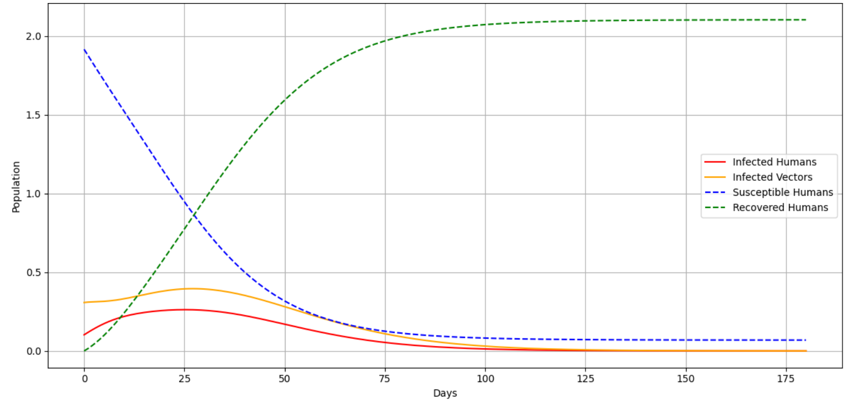
Scenario 2 – Increased Human-to-Vector Transmission

- $\beta_{vh} = 0.4$
- Faster outbreak, sharper epidemic peak
- Highlights impact of small β changes



Scenario 3 – Vector Control Intervention

- $\mu_v = \frac{1}{7}$ (shorter mosquito lifespan)
- Flattened curve, fewer total cases
- Demonstrates effect of entomological strategies



Conclusions

This project explored dengue dynamics in Latin America using both data-driven and mechanistic approaches. The integration of statistical modeling, machine learning, and epidemic theory provided a multi-layered view of risk prediction and disease behavior.

- Key takeaways:
  - Dengue burden is significantly shaped by structural and socioeconomic factors such as population density, poverty, and urban growth
  - Classification and regression models showed promising results despite limited data, and can support targeted surveillance
  - The SEIR model provided insight into potential outbreak trajectories and the role of transmission dynamics

- Limitations & Future work:
  - Lack of climate and entomological (vector) data limits biological resolution
  - Future extensions may include temporal modeling, weather integration, and higher-granularity data

Predictive analytics and epidemiological modeling, even when based on structural data alone, can powerfully inform public health strategies in resource-limited contexts

Thanks for your attention!

GIUSEPPE LEONARDI