Predicting Dengue Outbreaks with Explainable Machine Learning

Robson Aleixo
Department of Computer Science
University of São Paulo, Brazil
robson.aleixo@usp.br

Fabio Kon
Department of Computer Science
University of São Paulo, Brazil
kon@ime.usp.br

Rudi Rocha
São Paulo School of Business Adm. (FGV)
Institute for Health Policy Studies (IEPS)
rudi.rocha@fgv.br

Marcela Santos Camargo
São Paulo School of Business Adm. (FGV)
Institute for Health Policy Studies (IEPS)
marcela.camargo@fgv.edu.br

Raphael Y. de Camargo
Center of Mathematics, Computing, and Cognition
Federal University of ABC
raphael.camargo@ufabc.edu.br

***Abstract*—Seasonal infectious diseases, such as dengue, have been causing great losses in many countries around the world in terms of deaths, quality of life, and economic burden. In Brazil, this is relevant not only in large cities such as Rio de Janeiro and São Paulo but, according to the Ministry of Health, in another 500 cities throughout the country. Predicting the occurrence of diseases, such as dengue bursts, can be a valuable instrument for public health management as health officials can better prepare and redirect resources to the affected areas. In this paper, we present an explainable machine learning model to forecast the number of dengue occurrences in a large metropolis, Rio de Janeiro. We focus on explainable models, which provide health authorities with the reasons for outbreak predictions, allowing them to plan their actions accordingly. We trained a gradient boosting decision tree algorithm (CatBoost) with data from the National System of Information on Notifiable Diseases (SINAN), weather data, and socio-demographic data from The Brazilian Institute of Geography and Statistics (IBGE).**

***Index Terms*—Dengue Fever, Epidemiology, Public Health, Machine Learning, Explainable AI.**

I. Introduction

The spread of diseases has been a concern for health scientists for a long time due to its severe impact on the lives of citizens. Traditional studies use mathematical models to analyze disease propagation. More recently, researchers started to apply machine learning (ML) techniques to understand disease propagation, which could shed light on the factors that affect this propagation and complement traditional epidemiological models [1].

In this context, we propose the use of machine learning models to predict the spread of seasonal infectious diseases, more specifically, arthropod-borne viruses (arboviruses) such as dengue, chikungunya, and zika, which affect tens of millions of people worldwide every year. According to the CDC[1], there can be up to 400 million people infected with dengue each year in the world, with 100 million sick people and 40 thousand deaths by severe dengue, and it has also been a long-term concern for the Brazilian public health system. A 2020 report from the State of Rio de Janeiro[2], showed severe concern with the epidemic, placing 54% of the municipalities in a state of alert and high-risk. The city of Rio de Janeiro, capital of the state, has 6.32 million inhabitants (3% of the Brazilian population) but concentrated 11.4% of dengue cases in Brazil in 2011 and 2012.

Our objective is to assist public health epidemiological policies by providing predictions of dengue cases (ideally, 1 to 3 months in advance) for each area of the city (i.e., individual districts). These forecasts can be used to guide public health policies within Brazilian cities and states. To achieve this, we developed a machine learning model using health-demographic-environmental information, which includes data on past dengue, zika, and chikungunya cases, temperature, precipitation, health service availability, and demography. We considered data from 160 districts and their neighborhood to capture the spatiotemporal spreading dynamics.

We focus on an explainable model, which allows one to visualize how the ML algorithm used the available features to make predictions. Thus, we avoided using black-box models such as Support Vector Machines and Neural Networks that could hinder the interpretation of the results by public authorities.

To ensure the reproducibility of our results and enable fair comparison among models, we make our code and data available for public use at https://gitlab.com/interscity/health/dengue-prediction.

The major contributions of this paper are (1) an outbreak prediction model using multiple features, including environmental, demographic, epidemiological, and spatial data; (2) the prediction of the expected number of cases for 1, 2, or 3 months in advance with associated explanations for the

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP proc. 14/50937-1, and FAPESP proc. 15/24485-9.

[1] https://www.cdc.gov/dengue/about/index.html

[2] http://www.riocomsaude.rj.gov.br/Publico/MostrarArquivo.aspx?C=NqviPkhBljU%3D

predictions; as well (3) an evaluation of the model using 5 years of data from the Rio de Janeiro metropolis.

II. Related Work

There is extensive literature on dengue, including studies to detect outbreaks, forecast future cases, correlate with climatic and socioeconomic variables, and determine critical regions. A recent survey [2] indicated 966 models created for the analysis of dengue epidemics, including 545 using regression methods, 220 using temporal series, 76 using neural networks, and 50 using decision trees.

Several authors used linear models to predict dengue outbreaks in tropical countries such as Thailand [3], [4], Indonesia [5], Malaysia [6], and Latin America [7]. These works combined multiple predictors, such as climatic and socioeconomic data, and temporal series of dengue cases in the region, showing that they are relevant for model predictions. They use mostly Generalized Linear Models (GLM), although time-series methods [4], [7] and Geographical Weighted Regression (GWR) [6] were also used. Some work also explored the use of neural networks for dengue outbreak predictions [8]. The authors reported varying levels of accuracy, but they are not comparable due to the use of different datasets and performance metrics.

There were also several studies performed in Brazilian cities. Enslen *et al.* [9] did not find a correlation between dengue epidemics and infestation of the Aedes Aegypti mosquito, as measured by a predial infestation index (IIP). Subramanian *et al.* [10] created a stochastic Susceptible-Infected-Recovered (SIR) model, considering specifically the Sorotype Denv1 in the city of Rio de Janeiro, to predict reemergence times of dengue epidemics. Bomfim *et al.* [11] evaluated the impact of urban mobility in dengue dissemination using bus data from the city of Fortaleza. They used an LSTM neural network to detect the moment and intensity of outbreaks and a Susceptible-Exposed-Infected-Recovered (SEIR) model to capture the seasonality in periods of large-scale outbreaks.

Some authors used a classification approach to detect outbreaks, using decision treed and naïve Bayes [12], entropy related techniques [6], [13], and support vector machine (SVM) [14]. These authors used a combination of climatic conditions, temporal series of dengue cases, and social-economical information and obtained different accuracy levels, from 42% to 90%. However, the results are not comparable since they used data from different regions and predicted different time frames.

An important limitation of the existing works is that they provide little or no explanations for the predictions, especially in the case of more complex models, as noted in a recent survey [2]. Linear models provide some interpretability, but they tend to be less accurate in more complex scenarios. We use an explainable boosted decision tree model for outbreak detection, allowing decision-makers to understand how the model uses the available information to make predictions.

A second limitation of previous studies is that it is difficult to compare different methods due to the use of distinct metrics, some of which may be inappropriate, and of different datasets, which are also unavailable or difficult to obtain. We try to reduce these problems by (i) making all data and code publicly available and (ii) evaluating the model using multiple error metrics for regression and classification, in addition to a detailed analysis per district, month of the year, and prediction span.

III. Methods

We developed a model that predicts the number of dengue cases for each district in the city of Rio de Janeiro, Brazil, for periods from one to three months in the future.

A. Data

We obtained the historical data series of monthly dengue cases for each district in the city of Rio de Janeiro, Brazil, from January of 2011 to October of 2020, available on the National Health Notification Information System (SINAN) [15]. We used temperature and precipitation measurement from weather stations from the National Weather Institute (INMET) [16] and evaluated the mean temperature and precipitation per month in the city. Finally, we obtained demographic information from the Brazilian Institute of Geography and Statistics (IBGE) [17], and the number of public health facilities and related information from the National Register of Health Facilities (CNES) [18]. We used the data extracted from these datasets as features for training the machine learning model.

B. Feature Extraction

We selected features encompassing epidemiological, environmental, and demographic data, from which the model could extract valuable information for predictions. The main feature is the past number of dengue cases in districts and their neighbors. For each district we used as input the number of cases over the last three months (`cases m-1`, `cases m-2`, and `cases m-3`). For the neighborhood, we considered the sum of the number of cases among all neighboring districts over the same periods. We thus provide temporal and spatial information on the past number of dengue cases.

We used the precipitation level, mean temperature, and mean air humidity on each month to capture environmental conditions. We also included some additional public health data, which are the LIRAa (an index determined from a survey of Aedes aegypti infestation indicators), and the number of zika and chikungunya cases, since those are transmitted by the same vector as dengue fever. We considered multiple features related to public health facilities, but since they are strongly correlated, we kept only the number of health facilities.

We used a single model for all city districts. To provide information on how a specific district i behaves during dengue outbreaks, we used the `dengue_prevalence` measure, which is the sum of dengue cases in the district in the training set, normalized to the 0 to 1 range. Table I shows the complete list of features and some statistical properties, such as mean, standard deviation, minimum, median, and maximum values.

TABLE I
FEATURES USED AS INPUT TO THE MODEL.

Feature	Description	Mean	Std	Min	Median	Max
cases m-1	number of dengue cases in the last month	7	28	0	1	969
cases m-2	number of dengue cases two months ago	7	28	0	1	969
cases m-3	number of dengue cases three months ago	7	28	0	1	969
dengue_prevalence	Sum of dengue cases in the past (normalized)	0.07	0.11	0	0.04	1
neighbor_cases	Sum of dengue cases in neighboring districts	34	91	0	7	1667
precipitation (mm)	Total precipitation in the last month	95	78	1	65	395
temperature (°C)	Mean temperature in the last month	24	3	20	24	30
air_humidity (%)	Mean air humidity in the last month	73	5	59	74	85
liraa	Aedes aegypti infestation index (for the city)	0.94	0.15	0.6	0.9	1.5
chikungunya	Number of chikungunya cases last month	7	28	0	0	751
zika	Number of zika cases last month	4	23	0	0	528
demographic density	Demographic density	127	98	69	124	678
num_health_unit	Number of health facilities	9	23	0	3	248

C. Regression Methods

We use a boosted tree regression method, CatBoost [19]. We selected this method because it can easily adapt to different types of data, can capture non-linear relationships between features and can be combined with explainability methods. CatBoost does not work with time series directly, but we provide dengue cases for the three previous months and train it to predict the following months. We provide examples from multiple years and districts, which should be enough data for the model to learn the behavior of dengue cases over time, including seasonality.

We compared CatBoost with the Seasonal Autoregressive Integrated Moving Average (SARIMA), using the individual time series for each district. SARIMA can capture seasonality in the data, which is important for dengue and is a standard method for time-series modeling and prediction.

We applied grid search to define Catboost parameters, using 2015 as validation data and 2012 to 2014 as training data. We found the optimal parameters: learning rate of 0.1, plain boosting type, Bernoulli bootstrapping, and lossguide for grow policy. For SARIMA, we applied the model to data from 2012 to 2015 and found the following optimal parameters: p=2, q=0, and d=0 for trend elements, and P=1, Q=0, D=0, and m=12 for seasonal elements. We used the python package `catboost`[3], version 0.24.4, with the default values for parameters not cited. For SARIMA, we used the function SARIMAX from package `stasmodels`[4], version 0.11.0.

D. Training and Testing

We performed individual predictions for each district, starting from each month from 2016 to 2020. We created a separate model for each year, using 5-fold cross-validation, with a single year as the test set and four years as the training set. For instance, for predicting the number of cases in 2017, we used as training data the years 2016, 2018, 2019, and 2020. We perform a 3-month multistep-ahead prediction for each starting month by recursively predicting each month in sequence. We should note that since we used 2015 to optimize the catboost hyperparameters, we excluded it from the cross-validation.

For the SARIMA model, we also created a separate model for each year (2016 to 2020). However, we used the previous four years as training data since SARIMA requires a contiguous time series to extract trend and seasonality features.

E. Evaluation Metrics

The main use of a dengue prediction model is to predict outbreaks before they happen. Since there is no generally accepted quantitative definition for an outbreak, we defined two outbreak levels: severe and mild. We considered an outbreak as severe when the number of cases in a district in a given month is above 99% of all measurements in the training set, and mild when above 95%. Figure 1 shows the distribution of dengue case counts from years 2016 to 2020.

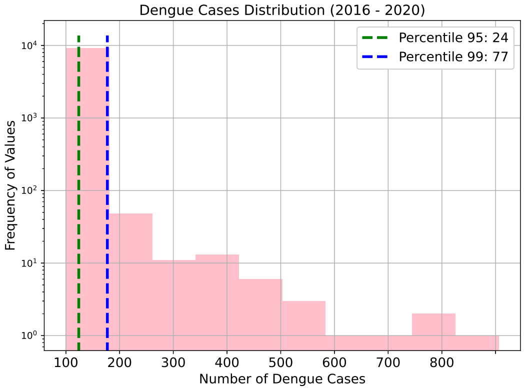


Fig. 1. Histogram with the log-scale frequency of dengue case counts (x-axis) in a single month per district. The dashed vertical lines shows the 95th (green) and 99th (blue) percentiles.

We also evaluated four standard error measurements: the coefficient of determinantion (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). These types of error are normally used in other works and are the standard for regression models. For the R^2, we used the variance of the respective test set.

[3] `https://catboost.ai/`
[4] `https://www.statsmodels.org/`

F. Prediction Explanability

We used the *TreeExplainer* from the `shap`[5] library to understand the relationships between features and model responses and to provide explanations for specific model decisions. This library provides an efficient implementation [20] for the Shapley values [21], which uses a game theoretic approach to compute the importance of variables in tree-based models.

We used the library to generate: (i) a summary of the effects of all the features in all decisions, (ii) examples of the effects of all features in individual decisions, and (iii) the interactions between pairs of features in all decisions.

G. Classification and Regression Models

We could have chosen a classification model instead of a regression one for outbreak prediction. But providing the model with the actual number of cases delivers more information to the model than just the category. Also, in the classification there would be several borderline cases which could introduce noise when training the classification model. Finally, the explanations for the regression model are easier to interpret, since they show how each feature contributed to the predicted number of cases on each instance.

IV. Results

We evaluated the model using data from Rio de Janeiro, Brazil, for periods from one to three months in the future, and present the results for one and three-month predictions. We first show classification and regression metrics for data from all districts and all periods. We then investigate results from individual districts and individual months of the year. Finally, we evaluate model decisions using explainability techniques.

A. Outbreak Prediction

Our model could successfully predict most outbreaks while keeping false positives low for three months in advance (Figures 2C and 2G). It correctly predicted there would be no outbreaks 97% of the time, with only 3% of outbreak misclassificatons (1st line of matrix C). Moreover, when predicting an outbreak as mild, there was a 57% chance of a mild or severe outbreak (320 out of 564), and when predicted as severe, a 86% chance of an outbreak (129 out of 149), as shown in the 2nd and 3rd columns of matrix G. The model also had an adequate recall (rate of detection of future outbreaks) by detecting 76% of all future outbreaks (448 of 587), estimated using the combination of the 2nd and 3rd lines of matrix G. Also, these predictions are three months in advance, giving authorities sufficient time to act before an outbreak occurrence.

Interestingly, for outbreak predictions one month in advance (Figure 2A) the results were similar to three months, which seems counter-intuitive. But as we will see in Section IV-B, the prediction of the number of cases is more reliable for one month, but for the objective of predicting outbreaks, the three-month predictions are sufficiently reliable.

For the SARIMA baseline model (Figure 2, right-hand side matrices), results were less reliable. It also correctly predicted there would be no outbreaks 97% of the time. Moreover, when predicting an outbreak as mild or severe, the chances of an outbreak occurrence were 20% and 25% respectively. But the recall was very low, and the SARIMA model detected only 29% of future outbreaks one month in advance and 14% three months in advance

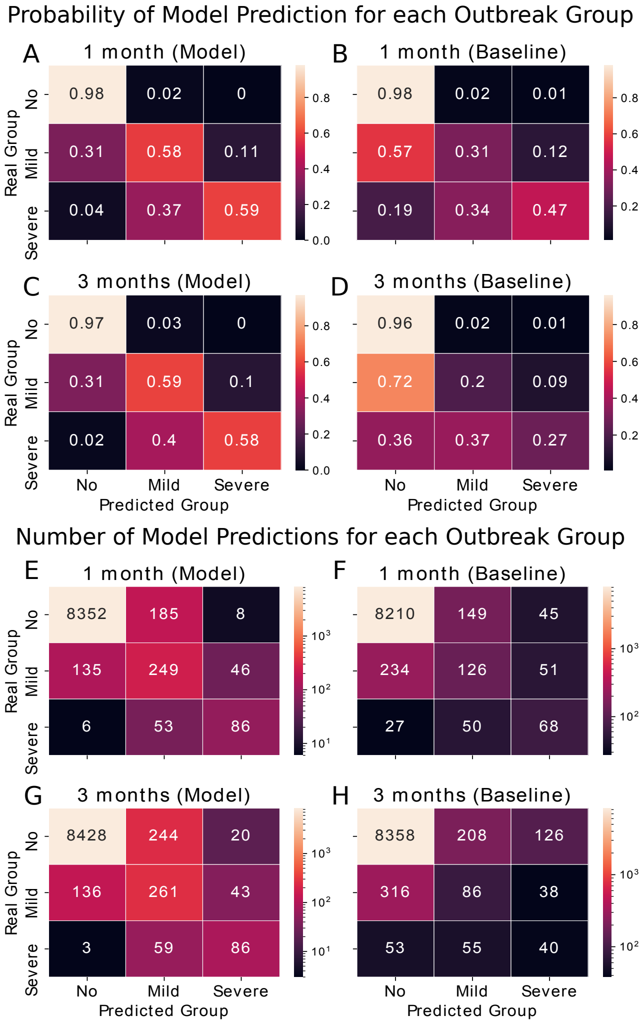


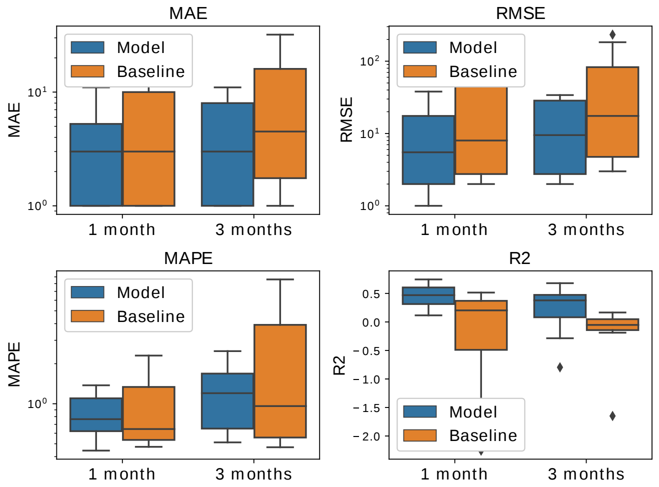
Fig. 2. Confusion matrices for the probability and number of model predictions for each outbreak group for the catboost model (left) and the SARIMA baseline model (right), where columns represent predictions and lines the correct target classification. Severe Mild outbreaks represent samples where the number of cases were very high, in the top 1%, corresponding to 120 or more cases in the district in the month. Outbreaks represent samples in the top 5% (excluding those in group 3), corresponding to 35 or more cases, and No Outbreak the samples with less than 35 cases.

B. Model Regression Performance

We evaluated four standard error measurements: MAE, RMSE, MAPE, and R^2 (Figure 3). The median of MAE for catboost (in blue) was 3.0 for both one-month and three-month predictions, but the 3rd percentile is wider in latter, reaching a value of 8.0, indicating that the error was more prominent for some months. This huge variance between predictions occurs because, for some months, such as in the winter, the number

[5] `https://github.com/slundberg/shap`

of dengue cases is very low and easy to predict, resulting in MAEs close to zero. However, the number of dengue cases is much larger during the summer months, resulting in bigger errors.



Model (1 month)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	1.0	0.44	0.12
25.0	1.0	2.0	0.62	0.31
50.0	3.0	6.0	0.77	0.47
75.0	5.0	18.0	1.1	0.6
100.0	11.0	38.0	1.38	0.75

Baseline (1 month)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	2.0	0.47	-5.13
25.0	1.0	3.0	0.53	-1.15
50.0	3.0	8.0	0.65	0.11
75.0	10.0	45.0	1.34	0.3
100.0	16.0	63.0	2.3	0.51

Model (3 months)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	2.0	0.51	-0.8
25.0	1.0	3.0	0.65	0.08
50.0	3.0	10.0	1.2	0.38
75.0	8.0	28.0	1.69	0.47
100.0	11.0	34.0	2.49	0.68

Baseline (3 months)

Percentile	MAE	RMSE	MAPE	R2
0.0	1.0	3.0	0.47	-86.49
25.0	2.0	5.0	0.56	-8.97
50.0	4.0	18.0	0.96	-0.16
75.0	16.0	83.0	3.92	-0.01
100.0	32.0	233.0	8.61	0.17

Fig. 3. Regression model error metrics (MAE, RMSE, MAPE, and R^2) for one and three months in advance, for the catboost (model) and SARIMA (baseline) models. For each error measurement, we consider the predictions for all districts in a single month.

The RMSE and MAPE errors for the model follow the same pattern from MAE, with the difference that their median for three months was larger than for one month. For the R^2 error, higher values are better, and for one month, the median value was close to 0.47, which means the model predicted almost 50% of the variance in the number of cases. The results were less favorable for three months, with a median of 0.38 and the 25% quartile at 0.08.

When comparing the catboost and SARIMA models using the MAE and RMSE errors, catboost had smaller median values in all scenarios, except for one-month MAE predictions, where the medians had the same value. For three-month predictions, our model reduced the median of MAE from 4.5 to 3.0 and the median of RMSE from 17.5 to 9.5. In both cases, it also reduced the ranges of the 25% and 75% quartiles. In other words, our model reduced both the median and variability of the errors compared to SARIMA. The results were less clear for MAPE, with our model having larger median MAPE errors but smaller errors in the 75% quartile. As we discuss later, MAPE errors are highly influenced by the number of cases in the low season since the number of cases goes into the denominator, and it is not a good error measure for outbreak predictions. Finally, SARIMA performed very poorly when evaluated using the R^2 metric for three months, showing that it cannot explain the variability in the number of dengue cases.

We evaluated the errors for each month of the year to understand how these error measures behave during the typical outbreak season (February through June) and low season (other months) months (Figure 4). We can see that for RMSE the square factor inflates the regression error during outbreak months, resulting in larger RMSE values compared to MAE. During low season, the MAE and RMSE values are very low. For MAPE, the low-season months have a much higher impact, since it considers the ratio of the error with the mean number of cases. R^2 values are difficult to interpret, since R^2 considers the variance of cases of each month in its denominator. If the model overpredicts the number of dengue cases in a given month, the R^2 could be negative, as occurred in February, June and July, which are the months where the outbreak season begins and ends.

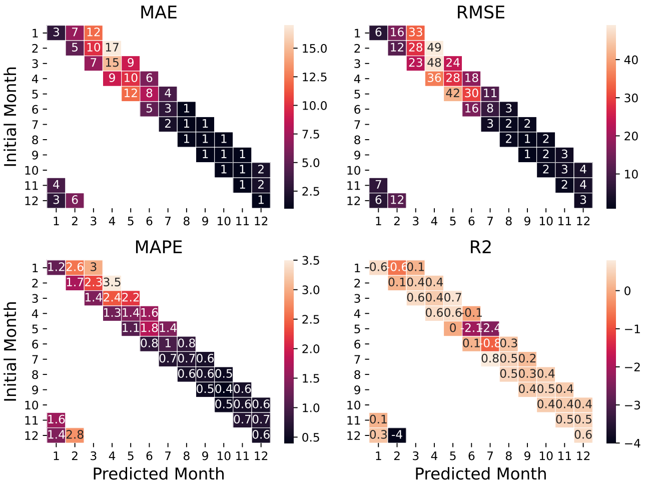


Fig. 4. Regression model error metrics (MAE, RMSE, MAPE, and R^2) for predictions starting on each month of the year. The three values on each matrix line represent the one, two, and three-month predictions.

Although authors typically use the metrics R^2, MAE, and RMSE to assess model performance, they may be challenging to interpret. The main problem is that they are highly dependent on the absolute value of the number of cases. For instance, if our model predicts a severe outbreak with 120 cases, and there are 240 cases, this will result in MAE and RMSE errors much larger than an error where the model predicts 5 cases in a district (no outbreak) while in reality, there were 35 cases (mild outbreak). MAPE is better in this regard since it uses percentual errors and can be helpful if presented monthly, as in Figure 4. While we present the results for each month for all metrics, some authors only present the mean error for all months, which is even worse because the smaller errors in the low season smooth out the larger errors in the high season.

Finally, we show the total number of dengue cases and the predictions from catboost and SARIMA, from 2016 to 2020 (Figure 5). We can see the peaks at the beginning of each year and that, in some years, the peak of cases is small. Our model closely follows the actual number of cases for one-month predictions, while SARIMA results in larger prediction errors. For three months, our model still captures well the behavior of the real number of cases, but it tended to underpredict the number of cases during large outbreaks and overpredict in years with fewer cases. However, even though the model showed a small tendency to produce under and overprediction of dengue outbreaks for the period of three months, the results were accurate enough to predict their occurrence.

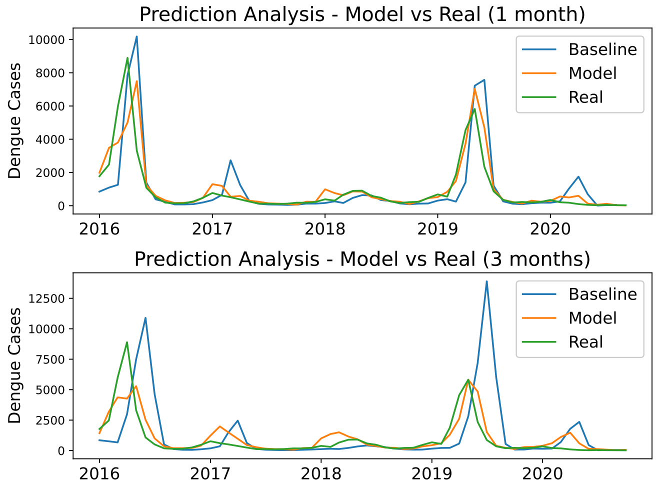


Fig. 5. Sum of dengue cases (green line) in all districts of Rio de Janeiro and of predicted cases using catboost (orange line) and SARIMA (blue line).

C. Evaluation per district

We also analyzed the errors for individual districts in Rio de Janeiro. Figure 6 shows the MAE and F-scores for each district. Districts in darker colors indicate better regression (lower MAE) and classification (higher F-score) results. The main discrepancies between metrics are in districts in the northeastern areas, which had low MAEs but average F-scores, mainly due to the smaller absolute number of cases in these districts, which also lower absolute MAE values.

Large errors in standard metrics do not mean the model is unsuitable for outbreak prediction. Bangu and Realengo districts had high MAE errors, but the model predicted all outbreaks (Figure 7). For 2017 they falsely predicted outbreaks in the three-month prediction, in a year with a minimal number of cases compared to previous years. However, most of the MAE errors for three months were due to the 2016 outbreak, which the model predicted correctly, but underestimated the number of cases. This further indicates why MAE and similar metrics are inappropriate for evaluating dengue prediction models.

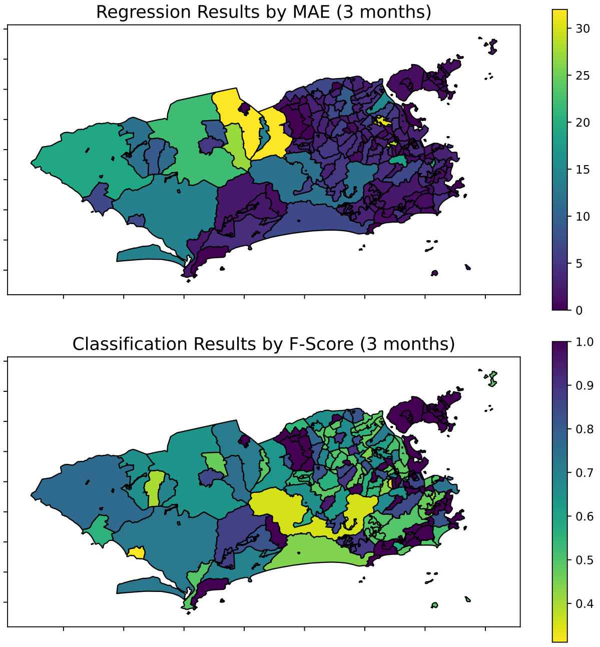
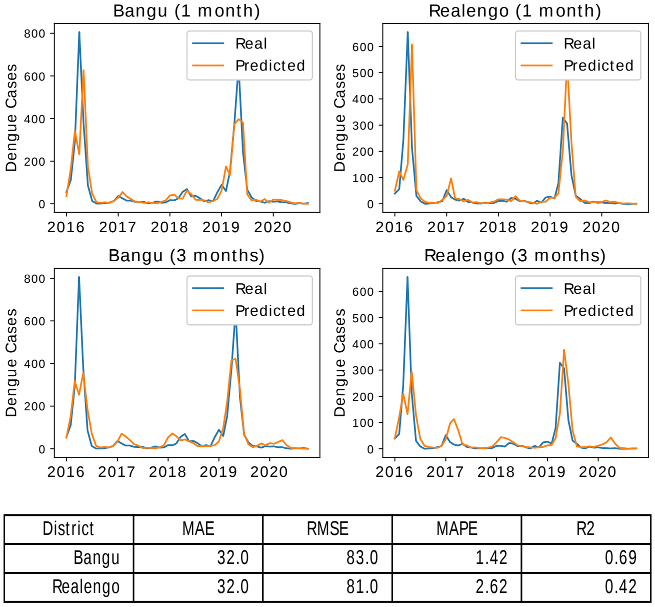

Regression Results by MAE (3 months)
30
25
20
15
10
5
0
Classification Results by F-Score (3 months)
1.0
0.9
0.8
0.7
0.6
0.5
0.4

Fig. 6. Map with the distribution of MAE and F-scores over the city districts for three-month predictions.



District	MAE	RMSE	MAPE	R2
Bangu	32.0	83.0	1.42	0.69
Realengo	32.0	81.0	2.62	0.42

Fig. 7. Number of real and predicted dengue cases per month for Bangu and Realengo districts for one and three-month predictions. The table shows the error metrics for the three-month predictions for the 5 year period on each district.

D. Understanding Model Decisions

We used SHAP values to evaluate the importance of variables on model predictions (Figure 8). We can see that the most relevant feature is the number of cases in the previous month (cases_m-1), followed by the dengue prevalence of the district, precipitation, neighbor_cases, and temperature. These features were expected to be relevant since higher temperatures and water accumulation are the main drivers of the dissemination of *Aedes Egiptys*. The number of cases in the target district and its neighbors indicate dengue's current dissemination levels, while dengue prevalence indicates the

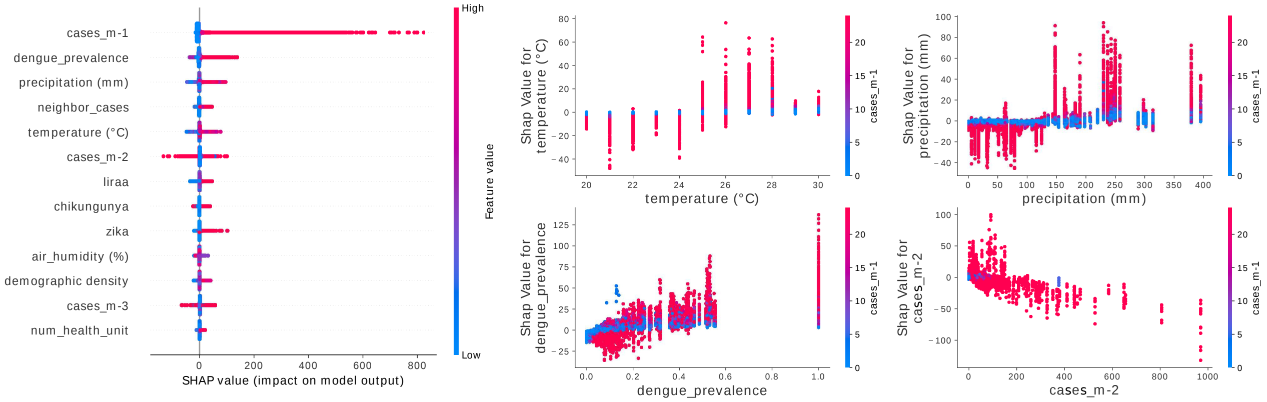


Fig. 8. Feature importance analysis using SHAP values for three-month predictions. (a) Global feature importance for all predictions; (b) interaction between features precipitation, temperature, liraa, and cases m-2 with the feature cases m-1

district's propensity to have larger or smaller outbreaks. Larger values of these variables resulted in predictions of a larger number of cases. Other relevant variables are the number of dengue cases in previous months (cases_m-2 and cases_m-3), the level of *Aedes Egiptys* dissemination (liraa), and the presence of other diseases propagated by the same vector (zika and chikungunya) in the previous month.

We also evaluated the interactions between variables during predictions (Figure 8). We can see that temperature and precipitation values only contribute to the prediction when the number of dengue cases in the previous month (cases_m-1) is above 15. In this case, higher temperature and precipitation levels lead to predictions of more dengue cases and vice-versa. The same applies to the critical neighbor indicator, which is more relevant only for larger values of cases_m-1. Finally, cases_m-2 and cases_m-1 have a more complex interaction where smaller values of cases_m-2 result in predictions of larger outbreaks and vice-versa. The model is probably predicting that the number of cases is increasing (smaller cases_m-2 values) or decreasing (larger cases_m-2 values). In the latter, a very large number of cases may indicate that the outbreak has already reached its peak value.

1) Providing Explainable Predictions: The model can also provide explanations for individual predictions. For instance, Figure 9 shows four examples of predictions: one correct (true positive) and one incorrect (false positive) prediction of a severe outbreak, and one correct (true negative) and one incorrect (false negative) prediction of no outbreak.

In the correctly predicted outbreak, the district had many cases (442) in the last month, increasing from 37 and 93 cases in the earlier months. It is also a district with high dengue_prevalence (1.0). Combined with other factors that positively influenced the predicted number of cases, the model found a precise prediction.

In the false positive case, the number of dengue cases in the previous months followed a similar behavior, with an increasing number of cases from cases_m-3 to cases_m-1. However,

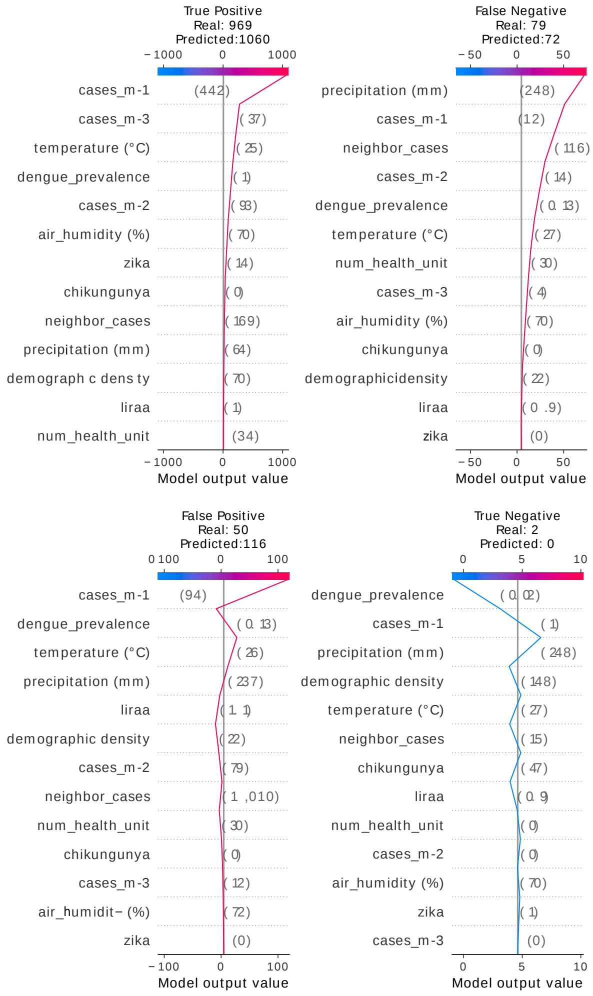


Fig. 9. Local feature importance for individual predictions using SHAP values for three-month predictions.

this number fell from 94 to 50 in the following month, while the model expected a slight increase to 116. However, even with high precipitation levels and a large number of cases in the neighboring districts, the model correctly predicted that there would not be a surge of new cases, in part due to the small dengue prevalence value of 0.13.

The false negative case is interesting because it falls in the frontier between classifying the month as an outbreak or not. Here, the threshold was at 77, and the model predicted 72 cases versus the actual value of 79. The model detected a meaningful increase in cases from 12 in the previous month to 72, mainly due to the significant precipitation levels and dengue cases in the neighboring districts.

Finally, the true negative case was correctly predicted due to the minimal number of cases in previous months and the small value of dengue prevalence. Although there was a large precipitation level, it had a lesser impact than the factors mentioned earlier.

V. Conclusions

In this paper, we described the development of an explainable machine learning model to predict dengue outbreaks within 1 to 3 months in advance and with a fine spatial precision (city districts). We obtained significant improvements with the approach introduced in this paper compared to the widely used SARIMA model. Also, our work indicates that classification metrics seem to be superior to regression metrics, such as R^2, MAE, and RMSE when evaluating a model's ability to predict outbreaks since: (i) regression metrics are highly affected by absolute values in the number of dengue cases; and (ii) the outbreak prediction errors are smoothed out when taking average regression values.

With the explanations provided by our model, public health authorities can interpret individual predictions based on the values of the features used for that prediction. Health professionals can use explanations from SHAP values to justify their actions and provide accountability, which is a significant improvement over black-box results from models such as Support Vector Machines and Neural Networks.

There could be many extensions for this work. Further information, such as vaccination rate and prior exposure to the disease, could also provide additional insights into outbreak behavior. However, the model would need reliable data from multiple years. Also, one could apply the model to other types of outbreaks, such as malaria or zika, as long as there is sufficient data for training the model. Our source code and data are publicly available, and we welcome collaborations from other research groups willing to work with us.

As a next step, enhancing the model presented here with time-series features and other sociodemographic indicators would be interesting. A relevant future work would be to develop an open-source web-based visualization tool capable of displaying the model predictions in an intuitive way to public health professionals and researchers. We will also contact health authorities in cities to discuss improving dengue outbreak prevention using the model predictions.

References

[1] T. L. Wiemken and R. R. Kelley, "Machine learning in epidemiology and health outcomes research," *Annual Review of Public Health*, vol. 41, pp. 21–36, 2020.

[2] P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, "Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes," *IEEE Access*, vol. 6, pp. 53 757–53 795, 2018.

[3] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data," *BMC infectious diseases*, vol. 19, no. 1, pp. 1–16, 2019.

[4] R. Chumpu, N. Khamsemanan, and C. Nattee, "The association between dengue incidences and provincial-level weather variables in thailand from 2001 to 2014," *Plos one*, vol. 14, no. 12, p. e0226945, 2019.

[5] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, and J. Rocklöv, "Prediction of dengue outbreaks based on disease surveillance and meteorological data," *PloS one*, vol. 11, no. 3, p. e0152688, 2016.

[6] G. Zhu, J. Hunter, and Y. Jiang, "Improved prediction of dengue outbreak using the delay permutation entropy," in *2016 IEEE Int. Conf. on Internet of Things (iThings) and Green Computing and Communications (GreenCom) and Cyber, Physical and Social Computing (CPSCom) and Smart Data (SmartData)*, 2016, pp. 828–832.

[7] S. Deb and S. Deb, "An ensemble method for early prediction of dengue outbreak," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2021.

[8] K. Roster and F. A. Rodrigues, "Neural networks for dengue prediction: a systematic review," *arXiv preprint arXiv:2106.12905*, 2021.

[9] A. W. Enslen, A. S. Lima Neto, and M. C. Castro, "Infestation measured by aedes aegypti larval surveys as an indication of future dengue epidemics: an evaluation for brazil," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 114, no. 7, pp. 506–512, 2020.

[10] R. Subramanian, V. Romeo-Aznar, E. Ionides, C. T. Codeço, and M. Pascual, "Predicting re-emergence times of dengue epidemics at low reproductive numbers: Denv1 in rio de janeiro, 1986–1990," *Journal of the Royal Society Interface*, vol. 17, no. 167, p. 20200273, 2020.

[11] R. Bomfim, S. Pei, J. Shaman, T. Yamana, H. A. Makse, J. S. Andrade Jr, A. S. Lima Neto, and V. Furtado, "Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas," *Journal of the Royal Society Interface*, vol. 17, no. 171, p. 20200691, 2020.

[12] A. A. Bakar, Z. Kefli, S. Abdullah, and M. Sahani, "Predictive models for dengue outbreak using multiple rulebase classifiers," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*. IEEE, 2011, pp. 1–6.

[13] C.-C. Chen and H.-C. Chang, "Predicting dengue outbreaks using approximate entropy algorithm and pattern recognition," *Journal of Infection*, vol. 67, no. 1, pp. 65–71, 2013.

[14] S. F. McGough, L. Clemente, J. N. Kutz, and M. Santillana, "A dynamic, ensemble learning approach to forecast dengue fever epidemic years in brazil using weather and population susceptibility cycles," *Journal of the Royal Society Interface*, vol. 18, no. 179, p. 20201006, 2021.

[15] M. R. de Janeiro. (2021) Dengue: dados epidemiológicos. [Online]. Available: http://www.rio.rj.gov.br/web/sms/exibeconteudo?id=2815389

[16] I. N. de Meteorologia. (2021) Dados climáticos. [Online]. Available: https://portal.inmet.gov.br/dadoshistoricos

[17] IBGE. (2021) Dados sociodemográficos. [Online]. Available: https://www.data.rio/datasets/

[18] C. N. de Estabelecimentos de Saúde. (2021) Cadastro nacional de estabelecimentos de saúde. [Online]. Available: http://cnes.datasus.gov.br/pages/downloads/arquivosBaseDados.jsp

[19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6639–6649.

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.