Computational and Applied Mathematics (2022) 41:393
https://doi.org/10.1007/s40314-022-02101-z

Machine learning algorithms for dengue risk assessment: a case study for São Luís do Maranhão

Fernanda Paula Rocha[1] · **Mateus Giesbrecht**[1]

Received: 2 August 2022 / Revised: 14 October 2022 / Accepted: 15 October 2022 /
Published online: 15 November 2022

Abstract
This study aims to assess dengue fever risk using Machine Learning techniques, such as logistic regressions, linear discriminant analyses, Naive Bayes, decision tree, and random forest classifiers. This kind of approach to epidemiological problems has been developed to detect risks for diseases occurrence and allows to create public policies based on mathematical models to prevent public health problems. In this study, the models were trained with data from the municipality of São Luís do Maranhão, state of Maranhão, Brazil. The majority of related works analyze states, countries, or continental levels, with greater availability of data. To apply the approach to such a small region, some oversampling techniques were used. The number of cases per neighborhood from 2014 to and 2020 and climatic, territorial, and environmental data was used as input variables to estimate the probability of dengue occurrence in the municipality. Due to the unbalanced database, we used the SMOTE, ADASYN, and DBSMOTE oversampling techniques. The DBSMOTE-trained Random Forest classifier achieved the best results with a 75.1% AUC, 75.43% sensitivity and a 60.53% specificity.

Keywords Dengue · Classification · Machine learning · Random forest · Logistic regression · Naive bayes

Mathematics Subject Classification 92

1 Introduction

Machine learning techniques have been applied to deal with many different diseases, from diseases known for a long time, to more recent diseases, such as COVID-19 (Chumachenko et al. 2022). In recent years, modeling the epidemiological dynamics of dengue has grown

Communicated by Rafael Villanueva.

✉ Fernanda Paula Rocha
fernanda.rocha507@gmail.com

Mateus Giesbrecht
mateus@fee.unicamp.br

[1] Department of Electronics and Biomedical Engineering, Campinas State University, Campinas, São Paulo, Brazil

as a method to prevent and control future outbreaks, especially in endemic areas. Disease models based on Machine learning techniques were recently introduced to assist decision-making processes (Batista et al. 2021). Models start from information containing biological, climatic, geographical, and other data. Then, classifier algorithms operate these data to obtain a possible optimal model, which is tested on unused data to perform predictions.

This article analyzes dengue behavior in São Luís do Maranhão from 2014 to 2020 using historical case data obtained from the Municipal Department of Health of São Luís, climate data from the National Institute of Meteorology, and neighborhood locations and vegetation indices from Google Earth Engine servers. Based on basic knowledge on the dynamics of the disease, a database containing variables to train classification algorithms, obtains the occurrence probability of dengue cases in certain locations, and analyzes the factors indicating an increase in cases, was created.

This study compares five classic classification algorithms in the machine learning literature, i.e., Logistic Regression (LR) (James et al. 2013), Linear Discriminant Analysis (LDA) (Xanthopoulos et al. 2013), Naive Bayes (NB) (Wongkar and Angdresey 2019), Decision Tree (DT) (Myles et al. 2004), and Random Florest (Cutler et al. 2007). The main objective is to analyze factors influencing the increase of dengue cases in São Luís/MA via quantitative engineering methods applied to epidemiology. It also aims to explore how supervised classification algorithms, applied to monthly epidemiological data distributed by neighborhood, perform to predict the future probability of cases occurrence.

1.1 Dengue fever in São Luís do Maranhão

Dengue has become a worldwide public health problem and is currently the insect-borne arbovirus with the greatest mortality rate. According to the World Health Organization (WHO), dengue incidence has increased eightfold in the last two decades, from 505,430 cases in 2000 to more than 2.4 million in 2010 and 5.2 million in 2019 (Organization 2022).

The study (Silva et al. 2016), analyzed the relation between temperature, rainfall, and dengue occurrence in São Luís/MA, observing increases in cases after rain periods. In addition to climate, behavioral and structural factors support the disease spread, despite government surveillance to control it. The Brazilian Ministry of Health has implemented strategies to struggle against dengue, such as the indices of building infestation—measured via the sampling index survey (LIA)—and the larval index rapid assay for Aedes aegypti (LIRAa)—obtained by collecting larvae in homes and wastelands—to identify predominant breeding sites. Another measure is spraying insecticide in those sites.

According to Ministry of Health data, the state of Maranhão reported 80,150 cases of the disease between 2001 and 2013 and 52,432 cases from 2014 to 2020. The state capital reported 6,379 cases from 2014 to 2020.

Data in Fig. 1 indicate that most cases in 2016 occurred from January to July. According to Brasil (2021), dengue shows a seasonal behavior, occurring mainly between October and May, as seen from 2015 to 2016.

2 Machine learning applications in epidemiology

This section shows some applications of machine learning algorithms analyzing and predicting the behavior of communicable diseases, such as, dengue, Zika virus, yellow fever, Ebola, and Marburg disease.

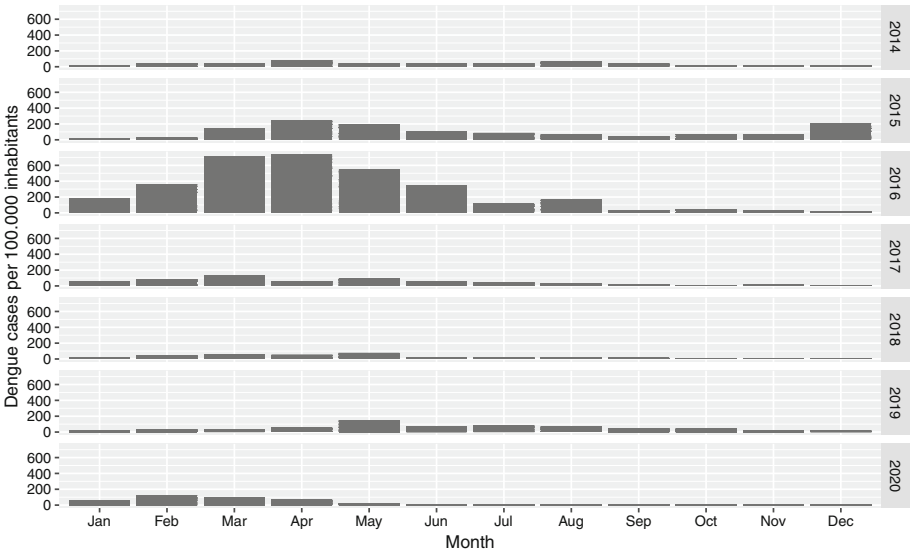


Fig. 1 Population data of positive cases obtained from the Municipal Health Department from 2014 to 2020

The dengue virus has rapidly expanded in recent years. It initially only inhabited tropical climates, having since moved to subtropical and temperate climates. Of the many factors affecting dengue virus transmission, temperature is the most frequently investigated factor as it affects the survival of the female vector. Temperature analysis can determine when and where viral transmission will persist or cease.

Aedes Aegypti and *Aedes Albopictus* are the two vectors which can transmit the virus. The difference between species may lie in their intestinal salivary glands (Lambrechts et al. 2010). A study (Brady et al. 2014) defined the thermal limits of *Aedes Aegypti* and the persistence of Aedes albopictus via a parameterized model considering the interaction between viral incubation lengths - which depend on temperature and survival of the adult vector. That study related temperature and viral transmission potential via vector capacity equations previously used in Gething et al. (2011). It employed a generalized regression model with data obtained from laboratory and field studies to analyze this relation. It formulated and adjusted the incubation period of the Aedes aegypti virus to a constant temperature, subsequently assessing differences in viral behavior during this period in both vector species.

Results showed that *Aedes Aegypti* is better suited to warmer regions, whereas *Aedes Albopictus* better inhabits regions with less pronounced temperatures, in which case it shows a 42 times higher adequacy index. Still, *Aedes Albopictus*, can adapt to regions best suited for *Aedes Aegypti* The obtained maps can restrict the extent to which transmission occurs and its endemicity.

The analysis of the transmission of dengue virus in Kraemer et al. (2015) relates it to spatial data, assuming that not every individual has the same chance of contacting any other individual. Generally, hypotheses assume a constant contact likelihood. For this, the authors collected temperature, water availability, and vegetation cover data, and daily case information from a hospital in Punjab from 2011 to 2014. They entered the data into a regression tree model to perform their analyses (Finkenstädt and Grenfell 2000).

Their results show that regional variability directly affects estimates of the basic reproduction number (R_0) by measuring viral transmissibility in the studied region and elsewhere.

Those results also show the significant differences in the heterogenous mixture of urban and rural configurations as human behavior directly influenced this distribution, and that the interaction between mixture parameters and infection strength has potentially large implications to optimize targeted interventions.

Another study (MacCormack-Gelles et al. 2020) sought to understand whether the larval index rapid assay for Aedes aegypti (LIRAa) helps to reduce mosquito-borne diseases. It considered information from several years to analyze whether the index is a good predictor of the risk of dengue transmission. The study was conducted in Fortaleza, capital of the state of Ceará, Brazil. The authors obtained data on dengue cases reported between 2012 and 2015 from the Brazilian Notifiable Diseases Information System and on local precipitation from Climate Hazards Group InfraRed Precipitation Station. As model outputs, the authors calculated dengue incidence rates per neighborhood in intervals of one, two, and four weeks after LIRAa, using two logistic models to analyze their data. They concluded that the LIRAa schedule is inadequate and that its sampling fails to prioritize certain densely populated areas.

In 2015, Zika Virus (ZIKV) infections were confirmed in several Brazilian states; eight months later, the World Health Organization (WHO) declared a ZIKV epidemic in the Americas. Immunologically inapt populations and a large number of Aedes aegypti mosquitoes in favorable environments are two reasons for this spread.

The study (Rocklöv et al. 2016) aimed to analyze the entry and expansion of the ZIKV in Europe, especially with the arrival of summer and the entry of travelers from the Americas. The study developed mathematical models highly dependent on average temperatures and daytime temperature variation. The models analyzed the monthly flows of air passengers arriving in European cities from American cities affected by ZIKV. They also considered monthly viral (R_0) estimates in areas in which transmitting mosquitoes inhabit in Europe and those in which humans reside in possible transmission areas. The authors developed three models, selecting that which best fits the transmission dynamics in the Americas. They then adapted the models to Europe using European climate data to estimate (R_0), comparing it with the (R_0) of the same region, given by government agencies, to validate the estimated R0. They found that predictions and observations agreed with each other.

The authors found that the periods in which travelers arrived in Europe coincide with the peak of estimated vector capacity and that areas with climates more prone to the virus suffer from the possible autochthonous ZIKV transmission, which can lead to outbreaks. Thus, the authors suggest greater attention to authorities in the warmer periods of the year. They also highlight that their study assesses the transmission potential of ZIKV in Europe rather than the probability of a ZIKV epidemic.

The study (Bogoch et al. 2016) aimed to assess the spread of ZIKV in Africa and the Asian Pacific. These regions have a large influx of travelers from the Americas, focus of cases. The study aimed to evaluate the risks of importing the virus via travelers and its ecological adequacy in new regions. In addition to analyzing the monthly flows of passengers from the Americas, models of intra-mosquito dynamics of dengue virus also assessed climate adequacy for ZIKV virus transmission and national health investments. The literature lacks a specific ZIKV model due to data scarcity, but evidence suggests that the dengue and Zika viruses share many common epidemiological features. Then the model was combined with global temperature data and then travelers were mapped.

The study concluded that countries that invest more in health decrease the risk of an epidemic, even with the large flow of travelers and with populations living in areas with higher risks of autochthonous transmission. Countries with a moderately high volume of travelers arriving from the Americas, but low investment in health, may be the most likely to transmit these viruses. Finally, analyses emphasize how local populations could benefit from

greater health investments. A similar approach was used in Rocklöv et al. (2016), but in that work the authors ignored the financial analyses of health investments.

Another study (Messina et al. 2016) shows which regions of the world have the adequate environmental conditions for Zika virus transmission. The authors used a species distribution model and boosted regression trees, including places which reported the disease in humans and places which did not, and environmental and socioeconomic variables. The study generated a database from information on ZIKV occurrence sites in the literature, and data on Brazil came directly from its national Ministry of Health.

The authors averaged of 300 boosted regression trees, predicting that tropical and subtropical zones show a greater chance of ZIKV transmission and that most of the Americas have adequate transmission environments. The models showed that annual accumulated precipitation particularly influences the risk of ZIKV. The study used cross-validation to evaluate the model.

Another virus that has been studied using Machine learning techniques is the Ebola. The study (Pigott et al. 2014) assessed the zoonotic transmission of the Ebola virus in Africa due to its high lethality rate. The adopted procedure resembles that in Pigott et al. (2015) to treat the Marburg Virus, as we will discussed later. To develop their research, the authors used BRT classification models to define the appropriate areas for an Ebola outbreak. To analyze model results, they used data on the location of cases in humans and three bat and primate species. They also used environmental variables collected from satellite information.

Although unprotected direct contact with infected individuals and cadavers can spread the disease, the study focused on the viral transmission between animals and humans (called zoonotic transmission) to predict a future outbreak based on the data cited above. The authors adopted this approach because the health system at these sites is unable to control the spread of the disease if the number of infected increases, raising the risk of a large outbreak.

The study classified countries into two sets, those which reported Ebola cases and those which failed to do so. They used the area under the curve (AUC) to measure the performance of the classification model (Prati et al. 2008). Their results predicted that 22.2 million people live in areas suitable for the zoonotic transmission of the Ebola virus, approximately 97% of which inhabit rural areas. Some 15.2 million people live in those areas, and the Democratic Republic of the Congo, Guinea, and Uganda are the countries with the highest number of reported cases. Moreover, seven million people live in areas with no recorded cases, with high numbers in Cameroon, the Central African Republic, and Nigeria. The classifier also found a relation between the Ebola virus and the environmental variables included in the model. As a conclusion, zoonotic transmission is more likely to occur in regions dominated by tropical forest.

From 2014 to 2016, West Africa suffered a geographically extensive outbreak of the Ebola virus. Research raised the hypothesis that the interaction between increased urbanization and human mobility contributed to the outbreak. To enable future response planning, works such as Kraemer et al. (2017) seek to understand what causes the spread of the Ebola virus and at what time the outbreak occurs. In total, three mathematical models aimed to analyze the human movement described in Simini et al. (2012) in which data were obtained from censuses and cell phones. It found that analyzing human mobility via mobile phones can considerably explain the dynamics of Ebola virus transmission at the studied site.

The authors of Kraemer et al. (2017) focused on the spread of yellow fever in Angola and the Democratic Republic of the Congo from 2015 to 2016. Due to the limited stock of vaccines in these countries, research needed to analyze how the virus spread during the outbreak to prioritize districts with higher contamination potential. They needed tools for this and, initially, a standard logistic model to infer infection risks. Moreover, the authors

found correlations between high population density and early contamination in a district via Pearson's correlation. The authors used the Cox Model to analyze the most populated districts, finding that transmission risks increased in these districts.

The study concluded that ecological and demographic factors significantly contributed to the continuous spread of yellow fever. Estimates provide regions which should be prioritized for population vaccination. Although these analyses always have limitations, such as the future inclusion of vaccine supply and delivery, public policies require such research.

A study (Pigott et al. 2015) on a new virus which emerged in 1967—named Marburg Virus—assessed the symptoms of the disease, finding high fever, hemorrhage, and organ failure. The Democratic Republic of the Congo suffered an outbreak in 1998, the source of which was bat colonies in local gold mines, and another in 2004, in the province of Uige, which caused interpersonal contamination. The authors found that non-human primates are susceptible to the disease and that human contamination had occurred in laboratories.

The authors set up their database by analyzing references which reported the disease and its contamination sites. They aimed to define the areas in which the zoonotic transmission of the Marburg Virus could occur, identify the number of people in at-risk locations, better understand the transmission of the virus, and raise awareness about the risk of outbreaks, which may arise from a delay in finding initial cases. For this, the study employed boosted regression trees, better detailed in Elith et al. (2008). The model builds a set of trees based on binary decisions, given a database, environmental variables, and areas with similar environments. The authors used four information components, the database in which the virus was transmitted from animals to humans, infections reported only in animals, environmental variables, and data from places in which the disease was unreported.

Then, the authors used driven regression trees to define environmentally appropriate areas for zoonotic transmission. The model requires present and past information for its predictions. Thus, the authors generated a dataset by randomly sampling 10,000 locations across Africa. In total, they used 500 submodels to compare factors influencing the sites with new viral occurrences. After training the model in the dataset, the authors created a prediction map indicating the most likely locations for the virus.

The study proposed two model variations as a result of the small amount of available data. Thus, the first model found geological characteristics, distances from infection sites, Karst formations—a terrain in which rock corrosion generates caves, relevant information since bats (an infection source) usually reside in these sites—and finally, vegetation indices. The second model considered a larger territorial area and some environmental factors.

Both models showed similar predictions of high zoonotic transmission in countries which reported cases of Marburg Virus. The second model predicted that 27 countries were among those who reported the disease or not and 105 million people lived in these areas. The first model predicted 19 countries, with 75 million people in risk areas, and that these people were also at-risk areas according to the second model.

Both models showed similar predictions of high zoonotic transmission in countries which reported cases of Marburg Virus. The first model predicted 19 countries, with 75 million people in risk areas. The second model predicted 27 countries, including those who reported the disease or not, and 105 million people living in those areas. Analysis found that temperature and vegetation determined the spatial distribution of the virus and that geological characteristics influence the risk level of an area. Finally, the maps obtained in the research can be used for clinical recommendations to diagnose cases with characteristics of this disease.

In this paper, the epidemiological dynamic of dengue fever in São Luiz do Maranhão will be assessed using machine learning techniques. The main contributions to the current state-of-the-art, besides the analysis of a new region itself, are the analyses per neighborhood within

the same city, allowing an adequate planning of resources application by the municipality government, the application of oversampling techniques, necessary due to the data imbalance and the comparison of different machine learning algorithms. The methods applied in this work are discussed in the following section.

3 Methodology

3.1 Data collection and processing

Over the years, a huge amount of data is generated from financial transactions, internet browsing, environmental monitoring, among others. Thus, data present different formats, e.g., graphs, time series, images, audios, and videos, sometimes making impossible the direct application of data on machine learning algorithms. In such cases, preprocessing techniques are suggested to make data more suitable for such algorithms. Some of these techniques were applied in this study to improve its applicability and results.

The data were obtained through the request protocol 00001.000235/2021-24 on the website (União CGU 2021). The data availability is guaranteed by law 12.527/2011 and is subject to the elaboration of a project adressed to the Municipal Health Secretariat of São Luís do Maranhão describing the purpose of the use of the requested information. Separate worksheets were sent by year with information on neighborhoods, the number of dengue cases in each, notification period, and deaths. Initially, the data received were in more than one data set, which led to the need to integrate them, generating a single dataset. In this group, 219 neighborhoods in the municipality of São Luís which had some confirmed case of the disease from 2014 to 2020 were catalogued. However, when we conducted a confirmation study of these sites via Google Maps, we noticed duplicate locations, nominal and rational data combination, and condominiums with duplicate locations. In these cases, these attributes were manually eliminated, totalizing 123 neighborhoods to be included in our model.

In addition to the time series of dengue cases at São Luís, climate variables were included in the model since those are directly related to the number of cases, as shown by Xavier et al. (2021). The monthly data collected from the National Institute of Meteorology were average temperature, precipitation, and humidity at the municipality of São Luís, obtained at Instituto Nacional de Meteorologia do Brasil - INMET (2021) in worksheets separated by years. This data are not split by neighborhood. Then, those climate variables were applied to all neighborhoods due to the small territorial extension of the assessed municipality.

Environmental variables such as the enhanced vegetation index (EVI) and the normalized difference vegetation index (NDVI) were also analyzed in this study, and collected via Google Earth Engine (GEE), available at GEE (2021). Both quantify the green vegetation of a location, but EVI was included in the model since it can correct data for atmospheric influences and canopy background signals, shows greater sensitivity in densely vegetated areas, and is used in several epidemiological studies, such as Pigott et al. (2015). These variables were obtained by the Landsat 8 satellite. Operating since 2013, it belongs to the Landsat series which began in 1972 and remains active until the date of this study. The series was developed by a project of the National Aeronautics and Space Administration and the United States Geological Survey to observe the natural resources of the Earth.

In addition to environmental variables, neighborhood locations were obtained by centralizing a point in each location, which generated a time series with EVI and NDVI monthly values and the latitude and longitude of each point.

3.2 Data unbalance processing

Classifiers often face unbalance problems with real data sets since they tend to classify majority classes. In our dataset, 78% of its cases were associated with sites with unconfirmed dengue cases, whereas 22% of dataset cases were related to sites with confirmed instances of the disease over the analyzed period.

A possible solution for some of these cases is to balance the current dataset by including new data. However, this approach is unfeasible in most practical applications, such as the one studied in this paper. Thus, some techniques to artificially balance the dataset are used, such as adding artificially created data to minority classes, which incurs the risk of these data failing to represent actual situations. In the present study, that would mean assigning confirmed cases to neighborhoods without notifications, inducing an inadequate data model. Another possible problem is the increased possibility of overfitting, in which the model is overadjusted, i.e., the artificial repetitions of minority class samples make the model fit the training data set so well that it is rendered ineffective for a new dataset. Data can also be excluded from the majority class, though this can delete important model data or cause underfitting, in which the model fails to fit the training dataset. To overcome these problems, some oversampling techniques reported in literature are discussed in the following section. Those techniques are applied to the training and validation datasets, after they are separated from the test dataset

3.2.1 Oversampling techniques

Some oversampling techniques consist of synthetically creating new minority class observations to match category proportion, ensuring the inclusion of all information at higher computational costs. In this case, random minority class data are replicated, which may lead to overfitting since the process duplicates existing data. To try to correct for overfitting, a more sophisticated technique was proposed, known as SMOTE, which generates synthetic minority class data from neighbors, thus avoiding duplicating data (Chawla et al. 2002). The method consists in calculating which are the closest neighbors and their characteristics, from which new data will be created.

Some variations of SMOTE are used in this work, such as the density-based synthetic minority over-sampling technique (DBSMOTE) introduced in Bunkhumpornpat et al. (2011). The method is based on cluster density, which groups a point set based on a distance measure (a Euclidean one, for example), creating minority class clusters which are then used to enlarge the minority class. According to the results in Bunkhumpornpat et al. (2011), this is, in practice, a more efficient method than SMOTE.

A variation of SMOTE is the adaptive synthetic sampling approach for imbalanced learning (ADASYN), elaborated by He et al. (2008). The method initially behaves as SMOTE, though it creates synthetic data based on data density, generating more synthetic data in regions with low minority data density and ignoring (or creating few minority data) in high-density spaces.

In Rana et al. (2022), two classifier models are generated with four oversampling techniques, included SMOTE and ADASYN, on an unbalanced dengue data set. The models, trained using synthetic samples generated by the ADASYN technique, provided higher accuracy by creating synthetic samples based on the density of the data. The DBSMOTE technique related to classification models used in epidemiological problems is uncommon, so we chose to apply it in this work, as performed in Kotb and Ming (2021). In that paper, a comparison between different oversampling techniques of the SMOTE family was performed on

an unbalanced dataset from the insurance industry. The results confirm that the techniques improve the accuracy of the models.

3.3 Principal component analysis

Principal component analysis (PCA) can be applied to data to reduce the size of the variable space and obtain a set of orthogonal (uncorrelated) axes that capture much of the original data variability. Though many variables may be correlated, this fails to contribute to discrimination, such as individuals' income and educational attainment. In our case, a high number of variables can lead to the curse of dimensionality. This term describes what happens when a region in space is divided into regular cells. Cell numbers will exponentially grow with the size of the space, exponentially raising the number of samples to ensure that all cells are filled, thus implying higher computational costs to solve the problem.

Simplifying data without losing important information helps processing by reducing computational time and algorithm complexity. Moreover, PCA enables data transformation to eliminate redundancies and preserve important information.

PCA projects data into a new space to reduce variable correlation. The first projection axis is the one showing the greatest variation, the second, the second largest variation perpendicular to the first, and so on.

Considering that the dataset has p observations with n attributes, and each observation is represented by a column vector $X_1, X_2 \ldots, X_p$, a $n \times p$ matrix X can be defined as:

$$X = [X_1, X_2, \ldots, X_p] = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}. \tag{1}$$

Then a matrix $A = X^T X$ containing variances and covariances can be obtained as follows

$$A = \begin{bmatrix} \text{Var}\,(X_1) & \text{Cov}\,(X_1X_2) & \text{Cov}\,(X_1X_3) & \cdots & \text{Cov}\left(X_1X_p\right) \\ \text{Cov}\,(X_2X_1) & \text{Var}\,(X_2) & \text{Cov}\,(X_2X_3) & \cdots & \text{Cov}\left(X_2X_p\right) \\ \text{Cov}\,(X_3X_1) & \text{Cov}\,(X_3X_2) & \text{Var}\,(X_3) & \cdots & \text{Cov}\left(X_3X_p\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}\left(X_pX_1\right) & \text{Cov}\left(X_pX_2\right) & \text{Cov}\left(X_pX_3\right) & \cdots & \text{Var}\left(X_p\right) \end{bmatrix} \tag{2}$$

whose size depends on its attributes. If $\text{Cov}\left(X_i, X_j\right), i, j = 1, \ldots, p$ and $i \neq j$ are non-null, there will be a correlation between the variables. If the original variables are all uncorrelated, the main components will be the original variables themselves.

Original variable correlation indicates redundant dimensions. Thus, the new uncorrelated variables are combinations of the original variables, reducing database size.

After obtaining covariance matrix A, its eigenvalues and eigenvectors are calculated by singular value decomposition of X, so that:

$$A = Q\Sigma Q^{-1} \tag{3}$$

in which $Q_{n\times n}$ contains the eigenvectors of matrix A and $\Sigma_{n\times p}$ the diagonal eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$. Then, eigenvectors are ordered according to the eigenvalues and the original data is multiplied by the main eigenvectors, obtaining the main components.

In Ramachandran et al. (2022), PCA is used to select the most relevant attributes for the models, with neural networks being used to evaluate the performance of models for dengue diagnosis predictions. The results in that reference were improved using the PCA and, for this reason, we chose this technique as the feature selection method in this study.

3.4 *K*-fold technique

One of the steps in predictive learning is the random division of the original data set into a training set and a testing set. The training set is used to train the model and the test set is used to evaluate how well the model is learning with new input data.

Separating data into only two disjoint parts can bring divergent results depending on the information contained in each set, and K-fold cross-validation approach minimizes these problems. The technique basically involves partitioning the training data and readjusting the competing models for each subsample to obtain additional information about the model fit.

The method consists of dividing the data into K equal parts, fitting the model using $K-1$ parts, and the remaining part is for validation. This process is repeated K times (at each time a different partition will be the validation), then the results are combined to obtain the average of the errors obtained.

3.5 Classifiers

For some problems, model response variables are non-quantitative and often defined as categorical. Classifiers are algorithms which can predict qualitative responses.

By predicting a qualitative response to an observation, we can understand how to classify that observation by assigning it to a category or class. Many classification methods are based on calculating the probability that a sample belongs to each qualitative variable category or class and assigning the class with the highest probability to the sample.

In this study, the following five classification techniques are used: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF). A brief description of each is presented in sequel.

3.5.1 Multiple logistic regression

The variable which answers our research hypothesis is qualitative, i.e., Yes for the sites which reported dengue cases and No, otherwise. Binary logistic regression is widely used for this type of problem. Our case contains multiple predictors, i.e., temperature, vegetation index, humidity, precipitation, altitude, and longitude. We should emphasize that the logistic regression model output is a real value indicating the probability of an event occurring in a given location. This is possible thanks to the separation of classes established for a probability value indicating that a class is positive, i.e., dengue cases either occurred or did not. This probability value is empirically defined from the collected data.

A probability value was established as standard for all classifiers used in this study. Its choice was indicated by 120 simulations with different predictor, training set, and test combinations. The linear regression models are given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d = \beta^T x \tag{4}$$

in which, $X^T = [X_1, X_2, \ldots, X_d]$ is the set of regressors and $\beta = [\beta_1, \beta_2, \ldots, \beta_d]$, the parameter vector. If Y is considered a categorical variable, regression models can classify

data. So, the probability of dengue occurrence is defined by

$$p(y = 1 \mid x) \text{ e } p(y = 0 \mid x), \tag{5}$$

indicating the probability of Y belonging to a particular category.

But since probabilities vary from 0 to 1, a logarithmic transformation of event occurrence probabilities is necessary. Mathematically,

$$\text{odds} = \frac{P}{1-P} \Longrightarrow \ln\left(\frac{P}{1-P}\right), \tag{6}$$

and that of non-occurrence by

$$p(y = 1 \mid x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}, \tag{7}$$

and that of non-occurrence by

$$p(y = 0 \mid x) = 1 - p(y = 1x) = \frac{1}{1 + e^{\beta^T x}}. \tag{8}$$

To estimate model parameters, the training set $T = (x_i, y_i)_{i=1}^{N}$ is used and estimation used the maximum likelihood method $\hat{\beta} = \text{argmax}_{\beta}\, p(y \mid X; \beta)$.

A literature review is performed in Hoyos et al. (2021), indicating the most used dengue models based on machine learning techniques. In the results, logistic models are the most widely used, with 59.1% being used in modeling for dengue diagnosis. For this reason, this technique is tested in this paper.

3.5.2 Linear discriminant analysis

Linear discriminant analysis (LDA) is an alternative and less direct approach to estimating the probabilities of belonging to a given class. LDA is very similar to PCA, but it not only finds component axes that maximize data variance, but also assesses the axes maximizing multiple class separation.

Comparing it with logistic regression, LDA performs well when classes are well separated—despite having some deficiencies, such as its incapacity to efficiently separate nonlinearly separable classes and to find a smaller space in which to project. LDA assumes that the data shows Gaussian distributions, and each class has identical covariance matrices.

We wish to classify a random variable X into a class k, in which $k \geq 2$, with density

$$f_k(X) = Pr(X = x \mid Y = k). \tag{9}$$

Discriminant analysis aims to divide the data space into k regions representing the classes to assign x to class k, if in the k region.

It assumes that $X = (X_1, X_2, \ldots, X_p)$ is extracted from a multivariate Gaussian distribution in which $\mathbf{X} \sim N(\mu, \Sigma)$. The occurrence probabilities $\pi_1, \ldots, \pi_k$ of each obtained class and the allocation of x to each class k will occur if $k = \arg\max \pi_i f_i(\mathbf{x})$, following Bayes' theorem.

This study used LDA given the limitations of logistic regression (such as well-separated class instability) and its capacity to address each point as a linear method for multiclass classification problems.

3.5.3 Naive Bayes

The Naive Bayes classifier is derived from Bayesian decision theory, which ranks objects according to the highest possible likelihood. The highest conditional probability indicates the chance of objects belonging to *Yes* or *No* classes. If the conditional probability is known for each class, the error obtained is as small as possible (optimal classifier). However, in most cases, joint probability distribution is unknown and estimating it is quite complicated. To solve this problem, variables are assumed to be independent. This implies that the joint probability distribution is equal to the product of marginal distributions. Mathematically, given M classes $\omega_1, \omega_2, \ldots, \omega_M$ and a random variable x, the conditional probability $p(\omega_i \mid x)$ of the random variable x belonging to each class i is indicated by

$$P(\omega_i \mid x) = \frac{P(x \mid \omega_i)P(\omega_i)}{P(x)}. \tag{10}$$

Bayesian decision theory ranks objects according to their highest likelihood, i.e., the chance of an object belonging to classes one, two or three, for example, is defined by its highest probability. Considering that each attribute is independent of each other,

$$P(\omega_i \mid x) = \prod_{i=1}^{n} P(\omega_i \mid x) \quad \text{and} \quad P(x) = \prod_{i=1}^{n} P(x). \tag{11}$$

The algorithm rates the most likely class as

$$\hat{y} = \underset{i \in \{1,\ldots,M\}}{\operatorname{argmax}} \; p\left(\omega_i \mid x\right). \tag{12}$$

Despite the limitation of the attribute independence hypothesis, the Naive Bayes classifier is robust and performs well for many real data. One of its advantages is that all required probabilities can be calculated from the training data in a single step. i.e., just looking at the training set it is possible to estimate all probabilities, which is relatively efficient if compared to other methods.

In Ozer et al. (2021), a prediction model based on machine learning algorithms is proposed. The model uses medical records related to dengue (chikungunya and zika) arbovirus data to help physicians evaluate whether or not a patient with a suspected arboviral virus should be hospitalized. Among the models, LDA achieved a high of 96.43%. Due to its success in a classification problem, this technique was applied in this paper.

3.5.4 Decision tree

Tree-based methods divide variable spaces into simple regions, in which training observation averages or modes in the region to which they belong are typically used to predict a given observation. The rules for dividing space can be summarized in a decision tree. These can be used in classification and regression problems.

In a tree structure, leaf nodes (response) contain the predictions, and internal nodes, attribute tests. Each possible attribute value has a branch to another subtree. The root node is usually the most important attribute since it is at the top and indicates the most relevant attribute to split the tree. After the tree is built, classification is done by "navigating" the tree until a leaf node is reached. A tree with combinations of numerical and categorical variables can be built, which makes data normalization unnecessary.

In this work, the Gini index was the metric used to find branches, choose nodes, and analyze how purely separated classes were in a tree.

A decision tree was used because it employs simple understanding and interpretation, since we know how variables relate to each other and can thus understand how decisions are made. Another advantage of the method is that it does not need data normalization, since it only compares attributes and branches rather than calculating distances to assess similarity. The algorithm can be simultaneously used with numerical and categorical data and is robust to outliers, since it only cares about how many elements are in each class. It can also be used in large volumes of data. Another advantage is that it is a nonparametric method.

In Sarma et al. (2020), a new approach is proposed by for dengue prediction on a dataset with patients medical history, based on Decision Tree and Random Forest. The models were evaluated using the ROC curve, demonstrating greater predictive ability of the Decision Tree model, with 79% accuracy.

3.5.5 Random forest

Random forests produce multiple trees to produce a single consensus prediction, improving predictive accuracy.

In random forests, each partition randomly selects X predictors out of an n total, in which $X \approx \sqrt{n}$, according to Gareth et al. (2013).

Data is sampled to produce several trees which then relate observations to attributes. Attribute sampling generates trees which are not dominated by a highly discriminating attribute, as occurs in decision trees.

In some cases, an attribute completely separates two classes, tending to be at the top of the tree as its root node. Thus, the tree is always constructed from it. When attributes are sampled, that one with high discriminatory power may not be at the top of the tree. Thus, other attributes may be used for this purpose, allowing forests with greater variability.

Random Forests are a very important method but the number of required trees must be defined. This number of trees ends up being a hyperparameter of the model.

In Sarma et al. (2020), the Random Forest model is used compared to the Decision Tree model. The model achieved 70% of accuracy on the training set and 68% of accuracy on the test set. Although the Random forest was not the most successful model in Sarma et al. (2020), we decided to test the model in this paper to compare its results with other methods.

3.6 Model validation

The choice of the best model involves reducing variance and bias. Variance must be reduced so that the model presents a proper performance for unseen data and errors must be as small as possible, reducing the bias. Unfortunately, there is no standard scientific method to achieve these objectives, which will thus depend on the choice of evaluative metrics and the problem in question.

Training errors are a poor estimate of test set errors, since large test sets can present small training errors, indicating overfitting. Overfitting occurs when a model fits previously observed datasets very well, but proves ineffective in predicting new results. Generally, complex models show low biases and high variances (overfitting).

In this study, K-fold cross-validation was used to reduce the relation between model bias and variance. In this technique, K is the number of equal subdivisions, or folds, of the training set. Different data are used to train and validate each subdivision. i.e., results will differ according to each K iteration. Training errors are calculated by applying the statistical method in the training data, whereas validation errors are the average of errors resulting from

the prediction of a new observation (which was not part of the training data). For example, if the training set is divided into 10 equal folds and the process starts with $K - 1$ parts to train the model, the tenth part is intended for predictions, checking the error, until all possible subdivision variations are completed. This cross-validation reduces variance and can select the best model. In this work, applications of the method for $K = 3$, $K = 5$, and $K = 10$ were simulated.

Many machine learning algorithms include one or more hyper parameters, which enable the algorithm to adapt its behavior to a specific dataset. Then, optimization must be used to find out a set of hyperparameters which perform better for a given dataset. In cross-validation, all labeled data is used, but there may be variations in the results of each classification, and the average of all classifications may reduce the variance of the entire process. Validation not only determines model accuracy but also choose attributes and models. This evinces that K-fold cross-validation is used to select both model hyperparameters and configured models. After validation, the entire training set is used to fit the classification model to be applied to the test set.

Since actual data sets are usually unbalanced, stratified cross-validation is one way to validate algorithms so each fold contains proportional class distributions. Stratified cross-validation keeps the proportion of the original dataset at each fold, which is extremely important. If the database shows 20% of positive classes and 80% otherwise, this validation distributes this proportionality across each fold. If stratified validation is ignored and the set has more negative cases, folds may contain only negative cases, producing poor results in model tests with folds containing only positive cases, since the model learned only from negative classes.

Even after the oversampling techniques mentioned in Sect. 3.2, stratified cross-validation is still useful to guarantee that each fold contains a balanced dataset.

3.7 Model validation metrics

Even if a single algorithm is chosen, parameter variations produce different models. For example, to evaluate decision tree models, the Gini index was used in this study to assess the purity of the separation of each tree. However, when using another criterion, such as entropy, other trees are generated. Thus, varying parameters for each classifier will obtain different results.

This study used metrics adequate to the problem of dengue prediction in São Luís/MA based on its occurrence probability. Sensitivity is obtained by dividing TP by the sum of TP and false negatives (FN). This metric is used to evaluate the ability of the model to successfully detect results classified as positive. It especially assesses how well models classify sites with dengue cases, even in the presence of false positives. Specificity is obtained by dividing true negatives (TN) by the sum of FP and TN. This metric evaluates the number of places classified as where dengue does not occur in relation to the total number of places where dengue indeed does not occur. A high specificity indicates that the majority of true negative classifications are correct, providing more confidence for not intervening in that places to reduce the dengue contamination risk.

Another metric used in this work is the receiver operating characteristic curve. This is a popular diagnostic tool for classifiers in balanced and unbalanced binary prediction problems, since it is unbiased to majority or minority classes (Weiss 2013). To evaluate this metric, the performance of each classifier is represented by a curve in a space defined by the number of

false positives (horizontal axis) and true positives (vertical axis). Each point from this curve is obtained by evaluating the classifier for a given threshold.

4 Results

4.1 Exploratory analysis

This subsection will address an exploratory analysis of our dataset to quantitatively summarize the main data characteristics via summary statistics. This stage is incapable of definitive conclusions without inferential statistics. Thus, we will describe the relevant aspects and outline hypotheses. Table 1 briefly describes the used summarized statistics.

Mode estimates the frequency of each attribute in our dataset. The Cases attribute shows a 0.67 average number of dengue cases in the entire dataset, with a 0 median and a 312-maximum count. Average rainfall during the studied period was 177.8 mm. However, as values show a greatly heterogeneous distribution, we deemed reasonable working with the median, 106.5 mm, as large values affect it less and data distribution is asymmetric.

Temperature showed a 26.75 °C median in this period, a valid value since the municipality shows a tropical climate throughout the year. Moreover, the most frequent values found were 0, i.e., the National Institute of Meteorology was unable to evaluate the temperature in certain periods. Thus, we assigned the value 0 for each missing value, treating precipitation in the same way.

EVI show a 0.28 average and a 0.95 maximum value. Since values vary between -1 and 1 (in which the closer to 1 a value is, the denser the vegetation), these values indicate densely vegetated areas.

Humidity was the attribute with the most missing data, and we obtained some metrics with null results. Although it is a climatic variable which can influence cases, it may perform poorly in the results due to its missing data.

To assess variables and their level of importance for subsequent models, we applied PCA.

4.2 PCA processing

To extract the most important variables from the dataset responsible for the best model performance, we applied PCA. By analyzing the results in Fig. 2, we obtained six main principal components (PC), each of which explained the percentage of total dataset variation. PC1 explains 28.67% of the total variance, whereas PC2, 19.28%. Together, they explained almost half of the set variation. PC6 explains only 4% of variation, indicating a low variance which PCA treated as noise. PC6 is related to humidity data due to its scarce information as the National Institute of Meteorology dataset excluded it in most months. Thus, the modeling process will not take this variable into account.

4.3 Predictive model performance

This subsection will show model results for both the original dataset and the oversampled ones, as described in previous sections. In summary, we applied five classifiers to four different sets, resulting in an analysis comparing 20 different results.

Table 1 Dataset descriptive statistics

Measures	Cases	Latitude	Longitude	Precipitation	Temperature	EVI	Moisture
Min.	0	− 44.26	− 44.38	0	0	− 0.20	0
1° quartile	0	− 2.59	− 44.28	7.6	26.29	0.19	0
Median	0	− 2.55	− 44.25	106.5	26.75	0.25	0
Mean	0.67	− 2.90	− 43.59	177.8	25.29	0.28	19.44
3° quartile	0	− 2.53	− 44.22	327.5	27.42	0.33	0
Max.	312	− 2.47	− 2.51	818.2	28.45	0.95	88.67
Mode	0	− 2.54	− 44.24	0	0	0.22	0
Standard deviation	5.23	3.74	5.16	197.78	6.39	0.14	34.84
Variance	27.41	14.02	27.73	39124.42	40.90	0.02	1214.19

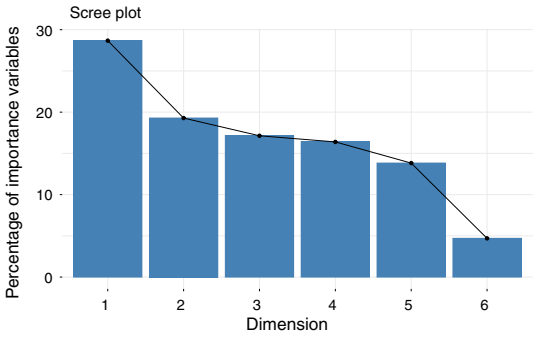
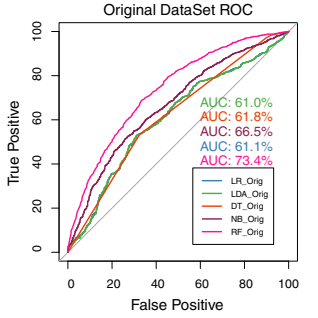


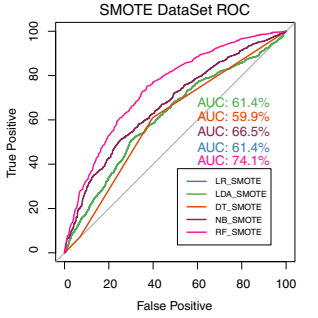
Fig. 2 Percentage of variations explained by each main component (CP)

The models used stratified K-fold cross-validation, in which the number of samples of each class in each fold is proportional to that of the original training set. We randomly tested the strategy for $K = 3$, 5, and 10, finding that 10 groups resulted in better performance.

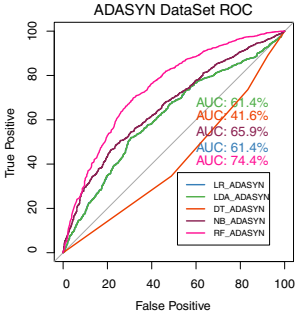
We used area under the curve (AUC), sensitivity (true positive rate), and specificity (true negative rate) to evaluate test dataset models. We selected the final model based on its receiver operating characteristic (ROC), considering sensitivity and specificity with different classification thresholds. AUC simplifies ROC by a single value, relating all its thresholds and calculating its area, remaining invariant to scales, since it processes classification accuracy



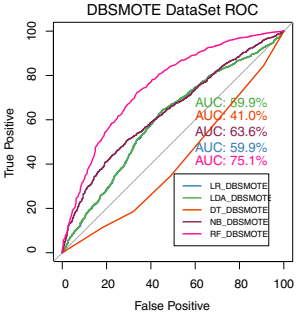
(a) AUC of the Original dataset



(b) AUC of the SMOTE dataset



(c) AUC of the ADASYN dataset



(d) AUC of the DBSMOTE dataset

Fig. 3 Comparison of the ROC curve for the different models and sets

rather than absolute values. Figure 3 shows our graphical results. RF was the only model to perform comparatively well, with AUC scores above 70%, regardless of the test set. This method obtained higher values in balanced datasets. The largest 75.1% AUC stems from the DBSMOTE-balanced model. Naive Bayes showed the second-best performance in all test sets, though with a relatively high difference, compared to RF. LR and LDA performed similarly in all test sets. The worst performing algorithm, DT, obtained its best AUC (61.8%) with the unbalanced set—the worst results among balanced-trained models—and the lowest AUC (41%) with the SMOTE set, worse than the random result.

The RF model with DBSMOTE achieved the better AUC. This happens since the other models classified a greater number of false positives, unlike RF, which classified a greater number of false negatives and a higher number of True Positives, absent in the other models. By observing the sensitivity, we can quantify the true positives, i.e. the percentage of places where dengue occurs and that are classified as where it indeed occurs. Therefore, classifiers with high sensitivity make it possible to implement control policies through sanitary surveillance teams and insecticide application. Implying a reduction in spending on hospitalization and medication for the treatment of patients. The specificity quantifies the true negatives, i.e., the places where dengue does not occur and that are correctly classified, and a high specificity is important to avoid financial investment in places not prone to outbreaks. Table 2 shows the sensitivity results for the training set for different probability thresholds. The values in bold are the highest performing.

Performing the isolated sensitivity analysis, to only minimize the locations that the model classifies as not having possible outbreaks and that have the potential for them to occur, that is, to reduce the False Negative results, the best models with sensitivity higher than 90% are obtained at thresholds 0.2, 0.3 and 0.4, trained on the SMOTE, ADASYN, DBSMOTE sets. The results using the original data are poor, as expected, due to the unbalance between classes.

Specificity is also used as a tool to evaluate models. It is calculated as the number of correct negative predictions divided by the total negative numbers, measuring how well the model can detect negative events. The numerical results on the training set are in Table 3.

Analyzing only the specificity, aiming to reduce the false positives, minimizing the costs of investments in dengue control in places classified as possible outbreaks, it was obtained the best models with specificity greater than 90% in the thresholds 0.6 to 0.9, trained on the sets Original, SMOTE, ADASYN and DBSMOTE. Note that, from threshold 0.3 on, the specificity is higher than 90% in the Original set, but that can be justified by the predominance of negative values in the set, with the model being trained on the unbalanced data set.

In summary, if the AUC and specificity criteria are considered, the RF trained with data balanced by DBSMOTE is the adequate classifier to predict the occurrence of dengue in a neighborhood given the climatic and EVI conditions, with greater specificities for thresholds above 0.7. On the other hand, if sensitivity is considered, the other classifiers also achieved values better than 90% for thresholds smaller than 0.3.

In Tables 4 and 5, the sensitivities and the specificities for the test set are shown for the different classifiers, trained with the different balancing techniques. Surprisingly, the sensitivity is low for the RF, while the other classifiers achieve better results for thresholds smaller than 0.3. As a conclusion, if the choice of the classifier was based on AUC, the result is a model with poor results on the test set. For this reason, it is fundamental to establish the metrics correctly to choose the classifiers for a given problem.

In Fig. 4, the DBSMOTE-trained RF classifier showed that Aedes aegypti adapts better to warmer regions with some rain volume. Especially in those with precipitations greater than 200 mm, the probability of dengue occurrence exceeds 80%—though low volumes still

Table 2 Model sensitivity in the training set

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.2				
DT	64.36	**100**	**100**	99.07
NB	78.96	99.89	**100**	99.91
LDA	65.04	**100**	**100**	99.95
LR	66.03	**100**	**100**	99.95
RF	**100**	**100**	**100**	**100**
Threshold 0.3				
DT	11.13	96.57	92.43	60
NB	51.18	98.14	99.82	99.48
LDA	11.01	97.64	**100**	96.86
LR	11.01	97.62	**100**	96.81
RF	**100**	99.96	99.98	99.95
Threshold 0.4				
DT	9	61.14	92.43	60.06
NB	28.65	93.13	98.97	97.45
LDA	3	89.73	97.30	65.68
LR	14	90.97	97.30	66.74
RF	99.75	99.83	**99.84**	99.74
Threshold 0.5				
DT	9	61.14	65.86	60
NB	4	88.14	92.63	94.03
LDA	0	36.74	48.85	31.70
LR	0	36.86	48.97	31.82
RF	98.33	99.38	**99.40**	99.23
Threshold 0.6				
DT	0	0	65.86	60
NB	1	79.08	85.47	88.99
LDA	0	7	9	6
LR	0	7	9	6
RF	88.24	97.88	**97.98**	96.52
Threshold 0.7				
DT	0	0	0	7
NB	0	58.54	70.85	80
LDA	0	0	0	0
LR	0	0	0	0
RF	56.25	90.82	**92.24**	87.02
Threshold 0.8				
DT	0	0	6	7
NB	0	27.02	35.72	**62.15**
LDA	0	0	0	0
LR	0	0	0	0
RF	0	0	0	0

Table 2 continued

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.9				
DT	0	0	6	7.90
NB	0	1	1	23.18
LDA	0	0	0	0
LR	0	0	0	0
RF	4	59.61	**66.11**	58.62

Table 3 Model specificity in the training set

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.2				
DT	54.99	0	0	0
NB	42.69	2	0	2
LDA	53.50	0	0	0
LR	50.83	0	0	0
RF	**91.26**	87.80	87.64	89.71
Threshold 0.3				
DT	97.16	7	26.88	77.19
NB	68.88	9	2	6
LDA	91.61	3	0	6
LR	91.65	3	0	6
RF	**97.67**	96.47	96.65	97.63
Threshold 0.4				
DT	98	61.85	26.88	77.14
NB	88	23.07	7	13.19
LDA	97.34	16.13	3	49.92
LR	98.52	14.71	3	48.85
RF	99.29	99.06	99.22	**99.48**
Threshold 0.5				
DT	98	61.85	64.16	77.19
NB	98.73	33.01	22.25	24.51
LDA	**99.99**	78.58	65.16	82.89
LR	99.94	78.44	65.07	82.86
RF	99.84	99.82	99.95	99.95
Threshold 0.6				
DT	**100**	**100**	65.86	77.19
NB	99.75	46.29	34.36	35.37
LDA	99.96	94.55	93.07	98.49
LR	99.98	94.55	93	95.54
RF	**100**	99.96	99.98	**100**

Table 3 continued

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.7				
DT	**100**	**100**	**100**	**100**
NB	99.72	67.22	51.93	47.39
LDA	**100**	99.53	99.48	99.51
LR	**100**	99.66	99.32	99.53
RF	**100**	99.98	**100**	**100**
Threshold 0.8				
DT	**100**	**100**	**100**	**100**
NB	99.98	89.89	81.98	66.58
LDA	**100**	9.99	**100**	**100**
LR	**100**	9.99	**100**	**100**
RF	**100**	**100**	**100**	**100**
Threshold 0.9				
DT	**100**	**100**	**100**	**100**
NB	**100**	99.96	99.85	93.47
LDA	**100**	**100**	**100**	**100**
LR	**100**	**100**	**100**	**100**
RF	**100**	**100**	**100**	**100**

show some occurrences. These places show an average temperature of 27 °C. The EVI scale ranges from − 1 and 1, in which negative values correspond to places with significant water accumulation, whereas values between 0 and 1 refer to local vegetation. Our results shows that areas with denser vegetation present higher case occurrence probabilities.

5 Discussion

In Brady et al. (2014), the thermal limits of Aedes Aegypti are analyzed, relating temperature and the potential for virus transmission. The results revealed that the Aedes Aegypti adapts better to warmer regions, as predicted by the Random Forest machine learning model developed in this study. Its predictive efficiency obtained an AUC of 75.1% with a threshold of 0.2. Its sensitivity and specificity on the training set reached 100% and 89.1%, respectively. In the test set, the model reached sensitivity and specificity values of 75.43% and 60.53%, respectively.

There is also a strong influence of temperature in the studies performed by Kraemer et al. (2015), and vegetation cover. In this study, the results related to EVI indicate that the sites with the highest probability of dengue occurrence present a dense vegetation. This index is important because some locations have mosquito-producing containers filled with rainwater, and the shading caused by the vegetation prevents water evaporation, facilitating mosquito reproduction. The precipitation has an influence on mosquito reproduction, and its combination with high temperatures and places with water accumulation generates favorable environments for the increase of cases.

Table 4 Model sensitivity in the test set

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.2				
DT	67	**100**	**100**	98.98
NB	80.78	99.71	99.85	99.56
LDA	68.93	**100**	**100**	99.85
LR	70	**100**	**100**	99.71
RF	70.52	76.88	78.18	75.43
Threshold 0.3				
DT	9	95.92	83.38	35.69
NB	53.61	97.25	99.56	98.55
LDA	10.69	98.12	**100**	96.67
LR	10.54	98.12	**100**	96.53
RF	50.34	65.17	64.74	61.27
Threshold 0.4				
DT	7	60.26	83.38	35.69
NB	28.46	92.34	**98.55**	96.53
LDA	2	91.62	98.12	66.62
LR	1	93.21	98.12	67.34
RF	41	51.59	51.16	50.14
Threshold 0.5				
DT	7	60.26	48.84	35.69
NB	3	88.29	**93.79**	91.62
LDA	0	38.15	54.19	29.33
LR	0	38.72	54.19	29.48
RF	27.31	39.16	39.01	36.27
Threshold 0.6				
DT	0	0	48.84	35.69
NB	1	78.18	**87.57**	85.98
LDA	0	6	8	4
LR	0	6	8	4
RF	17.48	28.61	29.19	25.72
Threshold 0.7				
DT	0	0	0	0
NB	0	55.06	72.54	**74.25**
LDA	0	0	0	0
LR	0	0	0	0
RF	8	18.93	18.06	15.46

Table 4 continued

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.8				
DT	0	0	0	0
NB	0	22.68	36.12	**49.42**
LDA	0	0	0	0
LR	0	0	0	0
RF	2	9	9	8
Threshold 0.9				
DT	0	0	0	0
NB	0	0	0	**13.72**
LDA	0	0	0	0
LR	0	0	0	0
RF	0	3	2	2.45

Table 5 Model specificity in the test set

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.2				
DT	53.64	0	0	0
NB	41.75	1	0	1
LDA	52.72	0	0	0
LR	50.31	0	0	0
RF	**63.27**	56.96	55.17	60.53
Threshold 0.3				
DT	**97.25**	7	26.38	76.94
NB	68.26	9	1	5
LDA	90.44	3	0	6
LR	90.52	3	0	6
RF	78.40	72.21	72.46	75.78
Threshold 0.4				
DT	98.33	60.91	26.38	76.94
NB	88	20.98	6	12.59
LDA	97.17	16.04	3	50.02
LR	**98.58**	14.37	3	48.61
RF	87.91	81.89	82.55	85
Threshold 0.5				
DT	98.33	60.91	65.43	76.94
NB	98.42	32.28	20.86	23.18
LDA	98.91	77.73	63.98	82.05
LR	**99.95**	77.56	63.98	81.96
RF	93.81	89.90	89.32	91.35

Table 5 continued

	Original (%)	SMOTE (%)	ADASYN (%)	DBSMOTE (%)
Threshold 0.6				
DT	**100**	**100**	65.43	76.94
NB	94.37	45.37	33.49	33.53
LDA	**100**	94.26	92.60	95.38
LR	**100**	94.26	92.56	95.47
RF	97.17	94.26	94.26	95.30
Threshold 0.7				
DT	**100**	**100**	**100**	**100**
NB	99.79	66.42	51.56	46.82
LDA	**100**	99.66	99.58	99.58
LR	**100**	99.75	99.37	99.62
RF	98.50	96.92	97.17	97.79
Threshold 0.8				
DT	**100**	**100**	**100**	**100**
NB	99.95	89.94	80.93	66.35
LDA	**100**	**100**	**100**	**100**
LR	**100**	**100**	**100**	**100**
RF	99.62	98.58	98.63	98.87
Threshold 0.9				
DT	**100**	**100**	**100**	**100**
NB	**100**	**100**	99.91	92.89
LDA	**100**	**100**	**100**	**100**
LR	**100**	**100**	**100**	**100**
RF	99.91	99.70	99.87	99.75

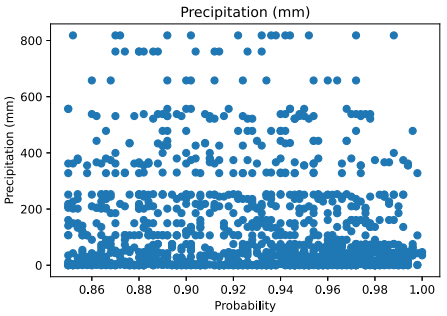
In Kraemer et al. (2015), transmission analysis was done using spatial data with assumptions about individuals and their influence on dengue transmission. The analysis occurred using a boosted regression tree model, with the same variables used in this study. In addition to human behavior influencing virus transmission, climate and environmental variables also contribute to an increase in transmission, as identified by the best model presented here.

When observing the results obtained through PCA, we have that the first principal component has large positive associations with the climatic variables (temperature and humidity). The second principal component has large negative association with the vegetation cover variable. The third component has positive association with the temperature variable.

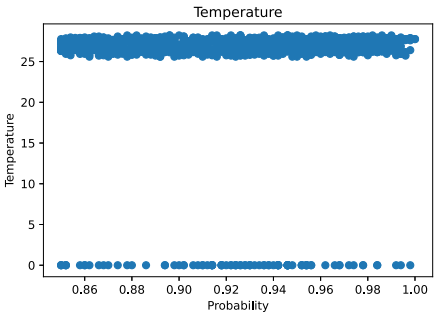
Although the conclusions are in agreement with other studies, machine learning methods allow us to quantify the factors and give numerical answers and predictions for the phenomenon.

6 Conclusion

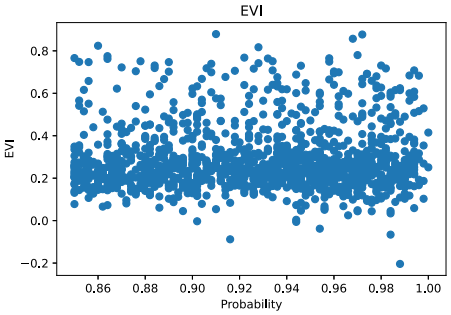
This study applied machine learning to predict the probability of dengue occurring in a given location and the main ambient factors related to the disease occurrence. Using government



(a) Precipitation in relation to the class with probability greater than or equal to 80%



(b) Temperature in relation to the class with probability greater than or equal to 80%



(c) EVI in relation to the class with probability greater than or equal to 80%

Fig. 4 Results of precipitation, temperature, and EVI if the probability of interest class is greater than or equal to 80%

data, we could assess which of five machine learning classifier algorithms developed the ideal predictive model. Our choice included assessments other than AUC results since this score alone fails to necessarily guarantee the best classifier. Since this study aimed to predict the probability of the occurrence of positive dengue cases, we expect that the chosen model will optimally classify positive cases, even if it includes false positives, as our priority was minimizing prevention costs.

Estimates enable the creation of effective public control policies, concentrating insecticide spraying in places of higher expected dengue incidence and optimizing the use of public financial resources. Moreover, the technique allows the definition of policies able to define the possible spatial distribution of vectors, geographically recognize areas of interest, and concentrate the building infestation index to these locations.

Funding This work was supported by the Brazilian agency CAPES (Grant no. 88887.486268/2020-00).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Instituto Nacional de Meteorologia do Brasil - INMET (2021) Normas Climatológicas. https://portal.inmet.gov.br/. Accessed 01 Jan 2021

Batista E, Araújo W, Lira R, Batista L (2021) Predicting dengue cases through Machine Learning and Deep Learning: a systematic review. Res Soc Dev 10:e33101119347. https://rsdjournal.org/index.php/rsd/article/view/19347

Bogoch I, Brady O, Kraemer M, German M, Creatore M, Brent S, Watts A, Hay S, Kulkarni M, Brownstein J, Khan K (2016) Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. Lancet Infect Dis 16:1237–1245

Brady O, Golding N, Pigott D, Kraemer M, Messina J, Reiner R Jr, Scott T, Smith D, Gething P, Hay S (2014) Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus transmission. Parasites Vect 7:1–17

Brasil S (2021) Plano de contingência nacional para epidemias de Dengue. https://bvsms.saude.gov.br. Accessed 2 Feb 2021

Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2011) DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique. Appl Intell 36:664–684

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) SMOTE: synthetic minority over-sampling technique. AI Access Foundation

Chumachenko D, Meniailov I, Bazilevych K, Chumachenko T, Yakovlev S (2022) Investigation of statistical machine learning models for COVID-19 epidemic process simulation: random forest, k-nearest neighbors, gradient boosting. Computation 10. https://www.mdpi.com/2079-3197/10/6/86

Cutler D, Edwards T Jr, Beard K, Cutler A, Hess K, Gibson J, Lawler J (2007) Random forests for classification in ecology. Ecology 88:2783–2792

Elith J, Leathwick J, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Finkenstädt B, Grenfell B (2000) Time series modelling of childhood diseases: a dynamical systems approach. J R Stat Soc Ser C (Appl Stat) 49:187–205. https://doi.org/10.1111/1467-9876.00187

Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with applications in R. Springer, New York

GEE G (2021) Earth engine data catalog. https://developers.google.com/earth-engine/datasets/catalog/landsat. Accessed 04 Mar 2021

Gething P, Van Boeckel T, Smith D, Guerra C, Patil A, Snow R, Hay S (2011) Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. Parasites Vect 4:92

He H, Bai Y, Garcia E, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pp 1322–1328

Hoyos W, Aguilar J, Toro M (2021) Dengue models based on machine learning techniques: a systematic literature review. Artif Intell Med 119:102157

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications In R, pp 59–126

Kotb M, Ming R (2021) Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2021.0120970

Kraemer M, Perkins T, Cummings D, Zakar R, Hay S, Smith D, Reiner R (2015) Big city, small world: density, contact rates, and transmission of dengue across Pakistan. J R Soc Interface 12:20150468. https://doi.org/10.1098/rsif.2015.0468

Kraemer M, Golding N, Bisanzio D, Bhatt S, Pigott D, Faria N, Pybus O, Smith D, Tatem A, Hay S (2017) Others predicting the geographic spread of the 2014–2016 west Africa Ebola virus disease outbreak. Am J Trop Med Hyg 95:47–47

Kraemer M, Faria N, Reiner R, Golding N, Nikolay B, Stasse S, Johansson M, Salje H, Faye O, Wint G, Niedrig M, Shearer F, Hill S, Thompson R, Bisanzio D, Taveira N, Nax H, Pradelski B, Nsoesie E, Murphy N, Bogoch I, Khan K, Brownstein J, Tatem A, De Oliveira T, Smith D, Sall A, Pybus O, Hay S, Cauchemez S (2017) Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. Lancet Infect Dis 17:330–338

Lambrechts L, Scott T, Gubler D (2010) Consequences of the expanding global distribution of Aedes albopictus for dengue virus transmission. PLoS Negl Trop Dis 4:e646

MacCormack-Gelles B, Lima Neto A, Sousa G, Do Nascimento O, Castro M (2020) Evaluation of the usefulness of Aedes aegypti rapid larval surveys to anticipate seasonal dengue transmission between 2012–2015 in Fortaleza, Brazil. Acta Trop 205:105391

Messina J, Kraemer M, Brady O, Pigott D, Shearer F, Weiss D, Golding N, Ruktanonchai C, Gething P, Cohn E, Brownstein J, Khan K, Tatem A, Jaenisch T, Murray C, Marinho F, Scott T, Hay S (2016) Mapping global environmental suitability for Zika virus. ELife 5:e15272. https://doi.org/10.7554/eLife.15272

Myles A, Feudale R, Liu Y, Woody N, Brown S (2004) An introduction to decision tree modeling. J Chemom 18:275–285

Organization W (2022) Dengue and severe dengue. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue. Accessed 12 Apr 2022

Ozer I, Cetin O, Gorur K, Temurtas F (2021) Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset. Neural Comput Appl 33:14975–14989

Pigott D, Golding N, Mylne A, Huang Z, Henry A, Weiss D, Brady O, Kraemer M, Smith D, Moyes C, Bhatt S, Gething P, Horby P, Bogoch I, Brownstein J, Mekaru S, Tatem A, Khan K, Hay S (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. ELife 3:e04395

Pigott D, Golding N, Mylne A, Huang Z, Weiss D, Brady O, Kraemer M, Hay S (2015) Mapping the zoonotic niche of Marburg virus disease in Africa. Trans R Soc Trop Med Hyg 109:366–378

Prati R, Batista G, Monard M (2008) Curvas ROC para avaliação de classificadores. Rev IEEE Am Latina 6:215–222

Ramachandran L, Rathnayaka R, Wickramaarachchi W (2022) Finding the best feature selection method for dengue diagnosis predictions

Rana S, Boruah A, Biswas S, Chakraborty M, Purkayastha B (2022) Dengue fever prediction using machine learning analytics. In: 2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON), vol 1, pp 126–130

Rocklöv J, Quam M, Sudre B, German M, Kraemer M, Brady O, Bogoch I, Liu-Helmersson J, Wilder-Smith A, Semenza J, Ong M, Aaslav K, Khan K (2016) Assessing seasonal risks for the introduction and mosquito-borne spread of zika virus in Europe. EBioMedicine 9:250–256

Sarma D, Hossain S, Mittra T, Bhuiya M, Saha I, Chakma R (2020) Dengue prediction using machine learning algorithms. In: 2020 IEEE 8th R10 humanitarian technology conference (R10-HTC), pp 1–6

Silva F, Santos A, Corrêa R, Caldas A (2016) Temporal relationship between rainfall, temperature and occurrence of dengue cases in São Luís, Maranhão, Brazil. Ciencia Saude Coletiva 21:641–646

Simini F, González M, Maritan A, Barabási A (2012) A universal model for mobility and migration patterns. Nature 484:96–100. https://doi.org/10.1038/nature10856

União CGU C (2021) Plataforma Integrada de Ouvidoria e Acesso á informação—Fala. BR. https://falabr.cgu.gov.br/. Accessed 2 Feb 2021

Weiss G (2013) Foundations of imbalanced learning. Imbalanc Learn:13–41

Wongkar M, Angdresey A (2019) Sentiment analysis using naive bayes algorithm of the data crawler: Twitter. In: 2019 Fourth international conference on informatics and computing (ICIC), pp 1–5

Xanthopoulos P, Pardalos P, Trafalis T (2013) Linear discriminant analysis. Robust Data Min:27–33

Xavier L, Honório N, Pessanha J, Peiter P (2021) Analysis of climate factors and dengue incidence in the metropolitan region of Rio de Janeiro, Brazil. PLoS One 16:1–15