

Traditional Machine Learning based on Atmospheric Conditions for Prediction of Dengue Presence

Brenda Sofía Sánchez López¹, Daniela Candiotti Nolberto¹,
José Antonio Taquía Gutiérrez², Yvan García López²

¹ Universidad de Lima,
Peru

² Universidad de Lima,
Instituto de Investigación Científica,
Peru

{20163586, 20162914}@aloe.ulima.edu.pe,
{jtaquia, ygarcia}@ulima.edu.pe

Abstract. The dengue virus has become an increasingly critical problem for humanity due to its extensive spread. This is transmitted through a vector that sprouts in certain climatic conditions (tropical and subtropical climates). The transmission of the disease can be associated with certain climatic variables that reinforce the outbreak. Data were collected on dengue cases by epidemiological week registered in Loreto-Peru from January 1, 2016, to January 31, 2022. Likewise, data on meteorological variables (maximum and minimum temperature; dry and humid bulb temperature; wind speed and total precipitation in the area). In this study, four Machine learning modeling techniques were considered: Support Vector Machine (SVM), Decision Tree, Random Forest and AdaBoost; and the parameters defined to evaluate the models are: Accuracy, Precision, Recall and F-1. As a result, optimal AUC values were obtained in a range from 0.818 to 0.996 for the SVM, Random Forest and AdaBoost algorithms, likewise, in all weather stations the ROC curve showed good performance for all models, except for the Decision Tree algorithm. As a conclusion for this study, we propose the optimal model to associate dengue cases with climatic conditions is SVM.

Keywords. Dengue outbreak, machine learning, SVM, classification, meteorology.

1 Introduction

Infectious diseases are the leading cause of death in the world. About 13 million people die annually from these diseases. The dengue virus has one of the highest infection rates per year. It is estimated that at least half of the world's population is at risk and estimates that there are between 100 and 400 million infections per year.

The problem is that there is still no specific treatment for dengue, and early detection of severe dengue-associated disease and access to appropriate hospital care only reduces mortality rates of the severe type to less than 1% [32].

Dengue transmission occurs by the bite of the female *Aedes aegypti* vector and by larval transmission of the mosquito, this species is present in all tropical regions, which are characterized by high temperatures and moderate to high relative humidity [20].

In the second case, the mosquito lays its eggs in natural and human-created (artificial) water containers, transmitting the virus orally. In Peru, dengue disease has manifested itself at different times, but it was not a serious problem in the past, to the point that it disappeared in 1956.

However, there was a reintroduction of the dengue vector, detected in 1984 in Loreto, which then spread to the San Martín region and the central jungle, and up to 2011 it was identified in 269 districts and 18 regions (almost one third of the country) [4].

Recently, in 2019 and 2020, the cumulative incidence of dengue per 100,000 population increased by 12.73% with 15,338 and 18,818 dengue cases in 2019 and 2020, respectively [18].

Likewise, for the year 2020, the official Health Minister indicates that Peru ranks third in the Americas with the highest mortality rate for dengue with data from cases reported in the Peruvian surveillance system out of a total of 48,858 dengue cases, including 72 deaths [18].

During 2019, there was evidence of an increase in dengue cases due to the low budget on the part of the state for fumigation of infected areas. Then, there was a decrease in cases coinciding with the appearance of COVID-19, which caused a quarantine declaration that prevented the correct capture of dengue cases.

One of the modern solutions being employed in medicine is the use of Machine learning (ML) tools. This branch of artificial intelligence (AI) uses algorithms, expressed as sets of mathematical methods which explain the relationships between variables.

1.1 Objectives

The present study focused on answering the question of whether it is possible to find an optimal ML model to associate dengue cases with climatic conditions in the city of Iquitos, located in Loreto. This is deployed in the following specific objectives:

- Compare the performance of different ML models using meteorological and categorical variables.
- Identify an optimal model for the classification of dengue cases in the city of Iquitos.
- Compare results obtained with previous research to complement the analysis and avoid possible biases.

The hypothesis is that there is an ML model that can effectively classify dengue cases in the city of

Iquitos and provide valid and reliable results based on model validation indicators.

2 Literature Review

After learning about the usefulness of this discipline, many authors who have applied ML models in the medical field investigate the prediction of dengue spread in San Juan (Puerto Rico) and Iquitos (Peru), using climatic variables with three different models (Interpolation, Gradient Boosting Regression [GBR] and Random Forest [RF]) to predict the number of dengue cases reported each week, with GBR being the model that presented the best performance for this type of study, with an MAE of 24.11 and 7.36 for San Juan and Iquitos, respectively [1].

Likewise, they investigated dengue in Malaysia using ML models (CART, Artificial Neural Network [ANN], Support Vector Machine [SVM] and Naive Bayes [NB]), these models focus on climate data from tropical locations where SVM showed the best prediction performance (Accuracy = 70%, Sensitivity = 14%, Specificity = 95%, Precision = 56%); however, the test sample increased to 63.54% compared to 14.4% for the unbalanced data; likewise, in this study, a binary variable was created for the prediction of dengue outbreak based on weekly incidence data in Selangor [24].

Adding to the topic, dengue in Indonesia was investigated with ML models reaching the conclusion that Random Forest with 10-fold cross-validation is the most accurate algorithm (58%) for predicting dengue in critical phase [28]. Also, they used meteorological data from 5 provinces and ML models (Support Vector Regression [SVR], reductive linear regression model, GBM, NBM, LASSO and GAM) to predict the occurrence of dengue in China, the most accurate was SVR model [12].

And they used ML to investigate infectious diseases where dengue is also included, for this they used 7 supervised models (SVM, Decision Tree, RF, NB, ANN, Bootstrap Aggregating and AdaBoost) and unsupervised learning methods (PCA and K-main) [31]. Similarly, in a paper they applied 4 ML models (general additive modeling, seasonal autoregressive integrated moving average or SARIMA, random forest, and gradient

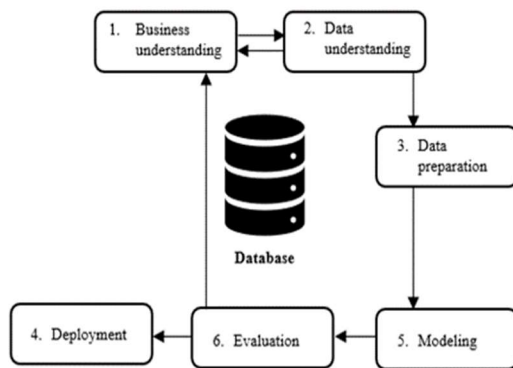


Fig. 1. CRISP-DM model

augmentation) to predict dengue in Manila, for this they used meteorological data, reported dengue case data and population statistics [5].

Finally, a study compared regression and time series statistical models with ML algorithms to predict dengue cases and outbreaks 4 to 12 weeks in advance, using meteorological data and dengue case reports, resulting in an error rate of 21% and 33% lower in ML models than in regression and time series [3].

We also found sources that explain the classification of dengue types using ML models. SVM technique was applied to identify whether machine learning identifies and classifies dengue with patient symptom data.

The result with cross-validation of 10 shows a performance of SMV algorithm with sensitivity of 0.4723, specificity of 0.9759, accuracy of 0.9042 and hazard ratio of 0.8343 [11].

Following the events that occurred in the year 2020 with the emergence of the pandemic caused by COVID-19, many researches have focused their efforts on presenting solutions. Firstly, an ML model for predicting COVID-19 based on symptoms was exposed; this model, consisted of 8 binary features in the form of a simple questionnaire and presented 0.9 AUC with 95% CI: 0.892-0.905 [33].

An unsupervised model for community detection using a COVID-19 dataset was exposed, with a Principal Component Analysis (PCA) algorithm and the K-Means method to efficiently cluster countries according to similarities in the different COVID-19 cases counted [7]. The most

frequent symptoms of COVID-19 affected patients were studied to predict in advance the key features of a patient, the models they used are: Convolutional Neural Network, Neural Network with Cross-Validation and Random Forest.

Authors Vaishya et al. collected more recent information on AI for COVID-19 from previous studies to identify its possible application for this disease, during the work they identified 7 possible applications for COVID-19 [29].

Similarly, another example of unsupervised learning whose objective was to test the effectiveness of predictive analytics to determine an area infected by an epidemic disease, with a backpropagation method to analyze a large set of diverse data categorized into physical network (population density and hotspot), geographic (climate and geodemography), clinical studies (clinical case classification and vaccination follow-up) and social media (geographic mapping) [13].

Another point evaluated above is the preference of Machine Learning over Big Data. In another source they explain that Big Data reduces its quality when incomplete data is presented, or disease peculiarities appear by region.

For this, they propose a joint Decision Tree and Map Reduce model for structured and unstructured data to help predict sub-diseases that may arise because of a main disease, obtaining an accuracy of 94.8% in the model [30].

Regarding the national level, the use of artificial intelligence to optimize the diagnostic process of tuberculosis was presented in the eRx project, which consists of an application that reads X-ray images based on artificial intelligence methods using convolutional neural networks, which detects pulmonary anomalies and preliminary clinical evidence of the disease [9].

3 Methods

The present research employed the CRISP-DM methodology which consists of 6 phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The ML models to be built are of the supervised type and include data on dengue cases and meteorological conditions, and a target

variable called "Class" was generated as a dependent variable.

Then, based on the most representative studies previously cited, the supervised models with optimal parameters and showing the best results were selected.

3.1 Algorithms

3.1.1 Support Vector Machine (SVM)

It is a supervised learning technique used to analyze data used in regression and classification. In addition to linear classifications, SVM efficiently performs nonlinear classifications by making use of the "kernel trick", which implicitly maps the input data to a higher dimensionality space by drawing margins between the classes.

The distances drawn between the two are maximized, thus reducing the classification error [16].

The parameters defined for this model are Cost (C) of 1.00, Regression loss epsilon (ϵ) of 0.10 and Kernel function of Polynomial type (1):

$$(gx*y+c)^d. \quad (1)$$

With the values of $g = 0.06$, $c = 0.5$ and $d = 2.5$.

3.1.2 Decision Tree (DT)

It is a supervised learning algorithm widely used for classification problems. It works by splitting the data into two or more related sets, first measuring the entropy of each point; then relying on the variables with maximum gain or minimum entropy to split sets, and finally repeating these two steps over and over again [26].

The parameters used for the construction of the model are minimum number of instances in leaves of 15, maximum node division of 5 and maximum depth of 150.

3.1.3 Random Forest (RF)

It is a supervised learning method used for classification and regression. This model is built with several fused decision trees, in order to obtain a more accurate prediction [15].

The parameters needed to build the model are number of trees of 50, number of attributes to be considered in each split of 2 and maximum node split of 50.

3.1.4 AdaBoost (AB)

It is a statistical classification meta-algorithm that trains weak trainees to adapt to successive errors, so that as the training is replicated on new trainees, they improve their performance.

The parameters established for model building are number of estimators of 15, learning rate of 0.1, fixed seed for random generator of 10, SAMME.R ranking algorithm (updates the weight of the base estimator with probability estimates) and square-type regression loss function.

3.2 Performance Metrics

The present study compared the models detailed above, seeking the best result for the classification of the target variable.

It is noteworthy that, neural techniques were not employed due to the limited data set used to test the models.

Next, to evaluate the performance of the models, the area under the ROC curve (AUC) will be calculated to ensure the correct fit.

In addition, the model validation indicators will be found: Accuracy (2), Precision (3), Recall (4) and F1 (5):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \quad (2)$$

$$\text{Precision (P)} = \frac{TP}{TP + FP}. \quad (3)$$

$$\text{Recall (R)} = \frac{TP}{TP + FN}. \quad (4)$$

$$F1 = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (5)$$

*TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

4 Data Collection

For the acquisition of the data set of dengue cases, the virtual health situation room, CDC Peru, was used.

This page shows reported cases including confirmed, probable and suspected cases by Epidemiological Week (SE), ordered demographically by department and sex, and also

Table 1. Recollected data sample

Features	2017	2018	2019
T Max (°C)	31.97	29.31	32.58
T Dry Bulb (°C) [7h]	24.57	23.69	25.15
T Dry Bulb (°C) [13h]	30.03	28.46	31.50
T Wet Bulb (°C) [13h]	28.09	26.66	27.63
Precipitation (mm)	19.90	09.20	43.30
Dengue cases	54	14	49

Table 2. Results of validation indicators (San Roque)

Model	AUC	CA	F-1	P	R
DT	0.718	0.934	0.902	0.873	0.934
SVM	0.952	0.975	0.965	0.956	0.975
RF	0.987	0.934	0.904	0.876	0.934
AB	0.967	0.992	0.992	0.992	0.992

Table 3. Results of validation indicators (Amazonas)

Model	AUC	CA	F-1	P	R
DT	0.417	0.977	0.965	0.954	0.977
SVM	0.895	0.995	0.993	0.991	0.995
RF	0.821	0.986	0.982	0.981	0.986
AB	0.818	0.995	0.993	0.991	0.995

Table 4. Results of validation indicators (Puerto Almenara)

Model	AUC	CA	F-1	P	R
DT	0.589	0.934	0.902	0.872	0.934
SVM	0.994	0.977	0.970	0.970	0.977
RF	0.996	0.949	0.931	0.917	0.949
AB	0.970	0.996	0.996	0.996	0.996

includes the cumulative incidences per 100,000 inhabitants in the area studied.

The records collected correspond to the districts of the city of Iquitos: Iquitos, Belen, Punchana and San Juan Bautista, ranging from SE 1 of 2010 to SE 18 of 2022.

From the dengue records, a categorical variable "Class" was generated that classifies

cases into four groups (A, B, C and D), taking the maximum and minimum value of cases and dividing them by ranges.

Meteorological data were collected from the National Meteorology and Hydrology Service of Peru (SENAMHI) website, where the stations of San Roque, Amazonas, Puerto Almenara and Moralillo were selected, and the independent numerical variables corresponding to maximum and minimum temperature in degrees Celsius [°C]; dry and wet bulb temperature in degrees Celsius [°C] at set hours 07, 13 and 19h.

Were extracted, total precipitation in the area [mm] and wind speed [m/s], ranging from SE 1 of 2016 to SE 4 of 2022 [25].

Data cleaning was performed, with more than 1900 records, detecting missing values for meteorological data on some days of each year, for which the missing data were filled in by applying statistical tools, as appropriate; except for the variable "total precipitation in the area", where the highest value of the week was chosen, since it has the greatest impact on the tropical behavior of the climate.

Finally, there were 243 records for San Roque, 213 for Amazonas, 257 for Puerto Almenara and 246 for Moralillo.

5 Results and Discussion

5.1 Numerical Results

For San Roque station, the Decision Tree model presents an AUC with a lower performance than the others, but with respect to the rest of the parameters, its results are optimal. Meanwhile, for the rest of the models, good results are shown, evidencing a good model fit.

In the case of the Amazonas station, the best performance is presented by the SVM algorithm in all the indicators obtained, while the Decision Tree model shows the worst performance in AUC.

With respect to the two remaining models, they present an AUC that shows an optimal fit, above 0.8.

Continuing with the Puerto Almenara station, the best performance is shown by the AdaBoost algorithm; it should be noted that, although there are higher values, as in the case of Random

Forest, they could be considered over-adjusted because they are very close to 1.

Finally, at Moralillo station, all the results obtained are optimal, including those of the Decision Tree model, which could be considered the least over-fitted, but if the Recall is also analyzed, the Random Forest algorithm shows a more adequate value of 0.874, which counterbalances with the possible over-fitting shown in AUC.

5.2 Graphical Results

The ROC curve (figure 2) shows that the fit is correct for all algorithms being above 0.5 and below 1, only with respect to the Decision Tree algorithm a drop-in performance can be seen.

The results of the Moralillo station (figure 3) show an optimal classification in all algorithms, being the dataset that shows the best relationship with the meteorological variables.

5.3 Proposed Improvements

The tables previously shown show the performance of the ML implemented on the climatic data of the four Iquitos stations. It is also observed that the best accuracy is provided by the SVM algorithm with a result ranging from 0.873 to 0.970.

From the data obtained, it is observed that the SVM algorithm classified optimally in most of the stations keeping similarity with a research indicating that this algorithm is the best predictor in terms of precision and accuracy of dengue outbreaks without overfitting in Malaysia [24].

Random Forest is a regression model that better captures nonlinear dynamics as is the case of the atmospheric variables used in this research because there are seasons with higher rainfall and temperature that condition the appearance of dengue-carrying mosquitoes [3].

To obtain an optimal classification, the Decision Tree, SVM, Random Forest and AdaBoost models were compared with a 5-fold cross-validation, all of them having a good Recall, but SMV was identified as optimal due to its accuracy and AUC values.

In this research, there were certain limitations with the meteorological data since information was

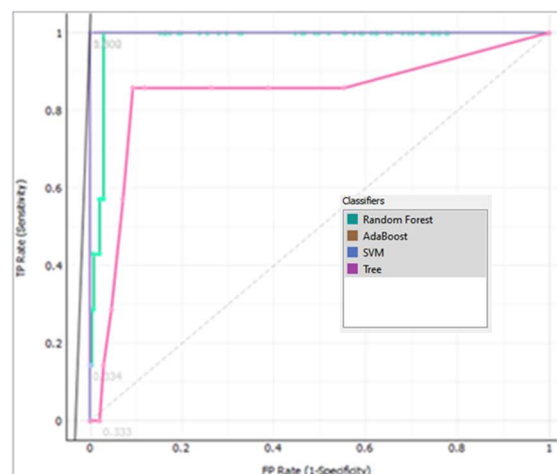


Fig. 2. ROC curve (San Roque station)

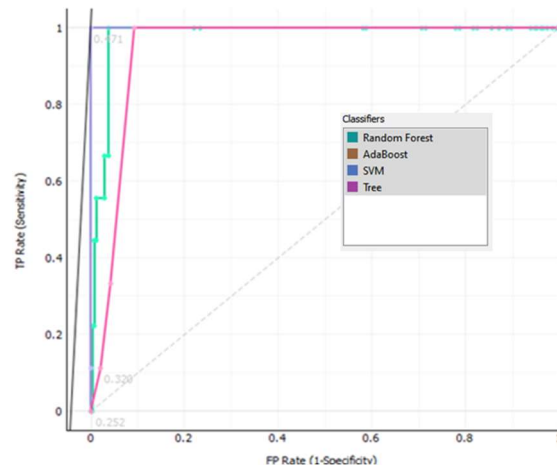


Fig. 3. ROC curve (Moralillo station)

not found for all months of the year, which made obtaining results complicated.

Part of the data used in Orange for this research was completed with statistical tools according to each variable. A similar data manipulation was performed in the research of the authors Anuranjan et al., who cleaned the data by interpolation [1].

Likewise, the results obtained are similar in methodology, data search procedures, ML models and obtaining results to the main reference papers used.

These provided information to complement the analysis of the results and confirm the hypothesis proposed.

Table 5. Validation result of 10 folds - San Roque

Model	AUC	CA	F-1	P	R
DT	0.796	0.942	0.919	0.909	0.942
SVM	0.956	0.971	0.96	0.953	0.971
RF	0.985	0.934	0.906	0.88	0.934
AB	1.000	1.000	1.000	1.000	1.000

Table 6. Validation result of 10 folds - Moraillo

Model	AUC	CA	F-1	P	R
DT	0.969	0.927	0.905	0.896	0.927
SVM	0.994	0.955	0.949	0.95	0.955
RF	0.977	0.882	0.837	0.801	0.884
AB	0.991	0.996	0.996	0.996	0.996

Table 7. Comparison of validations

Method	AUC	CA	F-1	P	R
80/20 train/test	0.712	0.939	0.91	0.881	0.94
5 folds cv(*)	0.718	0.934	0.90	0.873	0.93
10 folds cv	0.796	0.942	0.92	0.909	0.94

(*) cross validation

Finally, it is considered that in order to obtain more accurate results in this research, more support tools such as the use of R or Python programming, search methods and data manipulation are required.

5.4 Validation

Validation allows the performance of the model to be tested, to corroborate whether the results shown correctly quantify the relationships between variables and that when new data are added there is no need to adjust the model or over-fitting occurs.

The cross-validation method separates the data, first performing training and then validation. Orange Data Mining software was used for the modeling.

¹ www.ulima.edu.pe/en/scientific-research

The clean data was tested by cross-validation with 5 folds, dividing the data into five equal parts. In the same way happens for 10 folds. The complete results for 10 folds are shown below.

As can be seen, in this validation the model can learn better as in the case of AdaBoost, so it is correct to select this number of folds.

Table 7 shows the comparison of validations using a sample of the San Roque station with the DT model.

6 Conclusion

We were able to compare the performance of four ML models relating meteorological variables to dengue cases, obtaining optimal results in most of the performance indicators.

The SVM was identified as the optimal model for the classification of dengue cases in the city of Iquitos, due to its good performance.

The results obtained were compared with other research that applied similar techniques, since the research is similar in methodology, data search, ML models and obtaining results to the main reference papers used.

These provided information to complement the analysis of the results and confirm the hypothesis proposed.

Acknowledgments

This paper has received the support from the Universidad de Lima – Scientific Research Institute¹.

References

1. Singh, C., Anuranjan. (2019). Predicting dengue spread in San Juan and Iquitos using machine learning. DOI: 10.13140/RG.2.2.24207.74406.
2. Fatima, A., Manimeglai, D. (2012). Predictive analysis for the Arbovirus-Dengue using SVM classification. International Journal of

- engineering & Technology. Vol. 2, No. 3, pp 521–527.
3. **Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., Markuzon, N. (2020).** Weekly dengue forecasts in Iquitos. *Plos Neglected tropical diseases*, pp. 1–26. DOI: 10.1371/journal.pntd.0008710.
 4. **Cabezas, C., Fiestas, V., García-Mendoza, M., Palomino, M., Mamani, E. Donaires, F. (2015).** Dengue en el Perú: A un cuarto de siglo de su reemergencia. *Revista Peruana de Medicina Experimental y Salud Publica*, Vol. 32, Vol. 1, pp. 146–156.
 5. **Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., Watanabe, K. (2018).** Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infectious Diseases*, Vol. 18, No.183, pp. 1–15. DOI: 10.1186/s12879-018-3066-0.
 6. **Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. (2022).** Sala virtual de situación de salud, CDC Perú. <https://www.dge.gob.pe/portalnuevo/>.
 7. **Chaudhary, L., Singh, B. (2021).** Community detection using unsupervised machine learning techniques on COVID-19 dataset. *Social Network Analysis and Mining*, Vol. 11, No. 28, pp 1–9. DOI: 10.1007/s13278-021-00734-2.
 8. **Chen, M., Hao, Y., Hwang, K., Wang, L., Wang, L. (2017).** Disease prediction by machine learning over big data from healthcare communities. *Journals & Magazines, IEEE Access*, Vol.5, pp 869–879. DOI: 10.1109/access.2017.2694446.
 9. **Curioso, W. H., Brunette, M. J. (2020).** Inteligencia artificial e innovación para optimizar el proceso de diagnóstico de la tuberculosis. *Revista Peruana de Medicina Experimental y Salud Pública*, Vol. 37, No. 3. pp. 554–558. DOI:10.17843/rpmesp.2020.373.5585.
 10. **Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. (2020).** Dirección general de epidemiología, casos de dengue en el Perú hasta la SE 13 – 2020, Diapositiva de PowerPoint.
 11. **Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., Ma, W. (2017).** Developing a dengue forecast model using machine learning: A case study in China. *PLOS Neglected tropical diseases*, Vol. 11, No. 10. DOI: 10.1371/journal.pntd.0005973.
 12. **Ibrahim, N., Akhir, N. S. M., Hassan, F. H. (2017).** Predictive analysis effectiveness in determining the epidemic disease infected area. *AIP Conference Proceedings*. Vol. 1891, No. 1. DOI: 10.1063/1.5005397.
 13. **Lalmuanawna, S., Hussain, J., Chhakchhuak, L. (2020).** Applications of machine learning and artificial intelligence for Covid 19 (SARS- CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, Vol. 139, pp. 110059. DOI: 10.1016/j.chaos.2020.110059.
 14. **Li, X., Yang, X., Wen, W. (2019).** Diagnosis of methylmalonic acidemia using machine learning methods. *Proceedings of the 4th International Conference on Machine Learning Technologies*, pp 7–14. DOI: 10.1145/3340997.3341000.
 15. **Maesh, B. (2020).** Machine learning algorithms: A review. *International Journal of Science and Research*, Vol. 9, pp 381–386. DOI: 10.21275/ART20203995.
 16. **Marques, G., Pires, I. M., García, N. M. (2020).** Diabetes disease through machine learning: A comparative study. *4th International Conference on Computer Science and Artificial Intelligence*, pp 74–79. DOI: 10.1145/3445815.3445828.
 17. **Márquez-Benítez, Y., Monroy-Cortés, K. J., Martínez-Montenegro, E. G., Peña-García, V. H., Monroy-Díaz, Á. L. (2019).** Influencia de la temperatura ambiental en el mosquito *Aedes SPP* y la transmisión del virus del dengue. *CesMEDICINA*, pp. 44–47. DOI: 10.21615/cesmedicina.33.1.5.
 18. **Ministerio de salud. (2020).** Incremento de trasmisión de dengue, con ocurrencia de brotes y elevada letalidad en el país. www.dge.gob.pe/epipublic/uploads/alertas/alertas_202028.

19. **Ministerio de salud. (2022).** Situación del dengue en el Perú. www.dge.gob.pe/portalnuevo/vigilanciaepidemiologica/subsistema-de-vigilancia/dengue/situacion-deldengue-en-el-peru, Accessed on May.
20. **Ochoa, O. R., Casanova, M. M., Díaz, D. M. (2015).** Análisis sobre el dengue, su agente transmisor y estrategias de prevención y control. *Revista Archivo Médico de Camagüey*, Vol. 19, No. 2.
21. **Pu, X., Deng D., Chu C., Zhou T., Liu J. (2021).** High-dimensional hepatopath data analysis by machine learning for predicting HBV-related fibrosis. *Scientific Report*, Vol. 11, DOI: 10.1038/s41598-021-84556-4.
22. **Rasool, A., Tao, R., Kashif, K., Khan, W., Agbedanu, P., Choudhry, N. (2020).** Statistic solution for machine learning to analyze heart disease data: Proceedings of the 12th International Conference on Machine Learning and Computing, ICMLC'20, pp. 134–139. DOI: 10.1145/3383972.3384061.
23. **Sajana, T., Navya, M., Gayathri, Y., Reshma, N. (2018).** Classification of Dengue using Machine Learning Techniques. *International Journal of Engineering and Technology*, Vol. 7, No. 2–3, pp. 212–218, DOI: 10.14419/ijet.v7i2.32.15570.
24. **Salim, N., Wah, Y., Reeves, C., Smith, M., Yaacob, W., Mudin, R., Dapari, R., Sapri, N., Haque, U. (2021).** Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. *Scientific reports*, Vol. 11, No. 939. DOI: 10.1038/s41598-020-79193-2.
25. **Servicio Nacional de Meteorología e Hidrología del Perú. (2022).** Mapa de estaciones. www.senamhi.gob.pe/mapas/mapa-estacionesapadepesta1.php.
26. **Garg, A., Sharma, B., Khan, R. (2021).** Heart disease prediction using machine learning techniques. 1st International Conference on Computational Research and Data Analytics. Vol. 1022. DOI: doi:10.1088/1757-899X/1022/1/012046.
27. **Sidey-Gibbons, J., Sidey-Gibbons, C. (2019).** Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, Vol. 19, No. 1. DOI: 10.1186/s12874-019-0681-4.
28. **Silitonga, P., Dewi, B., Bustamam, A., Shaori Al-Ash, H. (2021).** Evaluation of dengue model performances developed using artificial neural network and random forest classifiers. *Procedia Computer Science*, pp. 135–143. DOI: 10.1016/j.procs.2020.12.018.
29. **Vaishya, R., Javaid, M., Haleem-Khan, I., Haleem, A. (2020).** Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Vol. 14, No. 4, pp. 337–339. DOI: 10.1016/j.dsx.2020.04.012.
30. **Vinitha, S., Sweetlin, S., Vinusha, H., Sajini, S. (2018).** Disease prediction using machine learning over big data. *SSRN Electronic Journal*, Vol. 8, pp. 556–559. DOI: 10.2139/ssrn.3458775.
31. **Wong, Z. S. Y., Zhou, J., Zhang, Q. (2018).** Artificial Intelligence for infectious disease big data analytics. *Infect Dis Health*, Vol. 24, No. 1, pp. 44–48. DOI: 10.1016/j.idh.2018.10.002.
32. **World Health Organization Dengue and Severe Dengue (2021).** www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue
33. **Zoabi, Y., Deri-Rozov, S., Shomron, N. (2021).** Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, Vol. 4, No. 1, pp. 1–5. DOI: 10.1038/s41746-020-00372-6.

*Article received on 22/10/2022; accepted on 20/03/2023.
Corresponding author is José Antonio Taquíá Gutiérrez.*