ISSN: 1692-7257 - Volume 1 – Number 43 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i43.2822

Application of machine learning and CRISP-DM methodology for accurate severity classification of dengue

Aplicación de machine learning y metodología CRISP-DM para la clasificación precisa de severidad en casos de dengue

MSc. Carlos Alberto Mejía Rodríguez [1], MSc. Miguel Alberto Rincón Pinzón [1]
MSc. Luís Palmera Quintero [1], Esp. Lina Marcela Arévalo Vergel [1]

[1] ***Universidad Popular del Cesar,*** *Ingeniería de sistemas, Grupo de Investigación GIDEATIC, Aguachica, César, Colombia.*

Correspondence: *calbertomejia@unicesar.edu.co*

Received: *October 15, 2023.* ***Accepted:*** *December 17, 2023.* ***Published:*** *March 16, 2024.*

How to cite: *C. A. Mejía Rodríquez, M. A. Rincón Pinzón, L. M. Palmera Quintero y L. M. Arévalo Vergel, "Aplicación del aprendizaje automático y la metodología CRISP-DM para la clasificación precisa de la gravedad del dengue", RCTA, vol. 1, no. 43, pp. 78-85, marzo de 2024.*
Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/2822

Abstract: The project focuses on accurately classifying the severity of Dengue cases in Casanare, Colombia, using Machine Learning (ML) and the CRISP-DM methodology. The target variable is "final classification," which categorizes cases into Dengue without warning signs and Dengue with warning signs. Several models and techniques were tested, with 'RandomForest' standing out as the most effective due to its high performance, achieving an accuracy of 100%. This improvement in classification will enable early and accurate identification of case severity, which, in turn, can enhance medical care and intervention strategies. The "Dengue Cases in Casanare by hospital service, person type, symptoms, and hospital status" database was used to support the analysis. The ML model is developed with the primary goal of reducing Dengue's complications and impact in the region.

Keywords: CRISP-DM, Data Science, Dengue, Machine Learning.

Resumen: El proyecto se centra en clasificar con precisión la severidad de los casos de Dengue en Casanare, Colombia, utilizando Machine Learning (ML) y la metodología CRISP-DM. La variable objetivo es “clasificación final”, que categoriza los casos en dengue sin signos de alarma y con signos de alarma. Se probaron varios modelos y técnicas, destacando 'RandomForest' como el más efectivo debido a su alto rendimiento, alcanzando una precisión del 100%. La mejora en la clasificación permitirá una identificación temprana y precisa de la gravedad de los casos, lo que, a su vez, puede mejorar la atención médica y las estrategias de intervención. Se utilizo la base de datos “Casos de Dengue en Casanare por servicio hospitalario, relación tipo de persona, síntomas y estado hospitalario” para respaldar el análisis. El modelo de ML se desarrolla con el objetivo principal de reducir las complicaciones y el impacto del Dengue en la región.

Palabras clave: Ciencia de Datos, CRISP-DM, Dengue, Machine Learning.

1. INTRODUCTION

Data capture today is more massive than ever; however, its application should not be limited to simple storage. As [1] highlights, "To the extent that our organizations recognize the great value that data has, we will witness many more implementations." This recognition has given rise to a revolutionary concept known as Big Data. For [2] the term Big Data, often translated as massive data, emerged at the beginning of the 21st century, especially in scientific fields such as astronomy and genetics, driven by the explosion in data availability. Notable examples include the Sloan Digital Sky Survey project, which generated more data in months than in the entire history of astronomy thus far, and the human genome project, which produced enormous amounts of genetic data. However, in recent years, the massification of data has spread to all areas with the increase in devices connected to the Internet, the rise of social networks and the Internet of Things (IoT). Furthermore, much of this data is openly accessible, allowing for global exploitation. But despite the abundance of data, its true value lies in its analysis and interpretation.

Within the scope of Big Data, it is essential to understand the concepts of data, information and knowledge. According to [3], data is the most elementary and crude unit that can be analyzed. Information is the result of operations performed on the data, losing certain details, but gaining generality. Finally, knowledge is an abstraction that guides the transformation of data into information, considering context and balancing applicability and understanding of nuances for accurate interpretation. A general way to understand the relationship between these elements is to recognize their adaptability to different users, who, due to their individual perspectives, may place them at different levels of abstraction. For one user, what constitutes information may be considered data by another. So, depending on the approach and the level of aggregation, the entity that is considered data, information or knowledge varies.

To extract knowledge from large amounts of information, it is necessary to carry out a data mining process. According to [4], data mining focuses on taking advantage of large amounts of information. However, to transcend the mere data analysis stage, specific methodologies have been designed with the aim of improving and optimizing the knowledge extraction process. These methodologies offer a structured framework to obtain, refine and apply acquired knowledge effectively in organizational settings. The KDD (Knowledge Discovery in Databases) model and the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology are fundamental approaches in the field of data mining. KDD establishes key stages for successful projects, including selection, preparation, pattern search, evaluation and model refinement. CRISP-DM, on the other hand, is based on KDD, but is specifically tailored to industrial needs, organizing the process into phases, processes and activities, from business understanding and data preparation to modeling, evaluation and implementation of results.

In the modeling stages, Machine Learning (ML) methods come into play and, according to [5], these algorithms can learn from the input data to improve performance on specific tasks through training processes. This interdisciplinary field combines statistics and computer science and is divided into supervised learning (with input examples and known targets) and unsupervised learning (without additional information). Problems like classification and regression are supervised, while feature extraction and clustering are examples of unsupervised learning.

Now, [6] highlight that, in the health sector, both diagnosis and medical decision-making involve reasoning under uncertainty. Doctors evaluate the information provided by the patient, the patient's medical history, and the patient's experience to determine the likelihood that the patient may have a particular condition. Therefore, it is crucial to make an accurate estimate of the associated risks to make appropriate medical decisions.

Regarding the application of machine learning in epidemiological research, [7] emphasize the critical importance that epidemic control has had throughout history. In response to this challenge, they propose the use of ML techniques as a novel tool to develop optimal control strategies that address multiple types of interventions. This historical approach and the application of advanced technology highlight the urgent need to effectively address epidemic control in contemporary society.

Dengue as a viral disease can raise control and prevention challenges. According to [8], dengue represents a critical challenge in Colombia, a hyperendemic country for this disease, and understanding epidemiological trends is essential for effective health policies. According to [9], dengue encompasses a wide variety of symptoms, from mild to severe. Early detection and appropriate

management are vital to reduce mortality. To achieve this, ML models such as Logistic Regression and Decision Trees, among others, can be used. These models are trained with data from dengue cases, including symptoms and test results, and combine or complement the strengths of different models, thus improving the accuracy of predictions.

Following these theories, a study was carried out to classify the severity of dengue in patients, considering their symptoms and using a database compiled from Colombian Open Data. The study was guided by the CRISP-DM methodology, which involves data preparation, variable definition, and model evaluation. The Machine Learning models used were Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest. In the evaluation process, the most effective model in diagnosing the severity of dengue is selected. This initiative contributes significantly to the management and early treatment of the disease, thus reducing its impact on public health in the region.

2. METHODOLOGY

The research is carried out following the CRISP-DM methodology, which according to [10], establishes the necessary stages for carrying out data mining projects. It is a free access guide based on the KDD process. The CRISP-DM methodology is made up of six main phases: understanding the business, understanding the data, data preparation, modeling, evaluation and implementation. A diagram of the phase flow is presented in Figure 1.

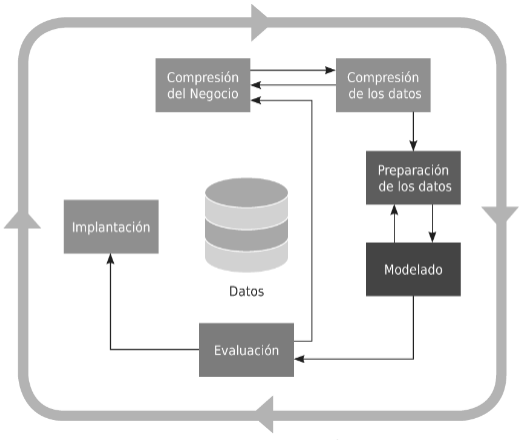

Fig.1. *CRISP-DM flow*
Source: Castillo Romero, JA (2019). Big data. IFCT128PO. IC Editorial.

The study is based on the data set "Dengue Cases in Casanare by hospital service," downloaded from the Colombian Open Data portal. This dataset consists of 54 columns and 2010 rows, each representing a case of dengue treated in hospital centers. The data was collected between October 2022 and February 2023. The target variable, called "clasfina" or "Final Classification," is used to determine the severity of dengue in patients and is divided into two categories: "Dengue without warning signs " and "Dengue with warning signs." This variable is essential, since it is what the models are intended to be able to learn to predict. The goal is to develop the best Machine Learning model to classify the severity of the disease in new patients.

3. RESULTS

The work is organized following the phases and activities of the CRISP-DM methodology. The results of each phase are described below.

3.1 Understanding the business

3.1.1. Determine the objectives of the Organization

According to [11] dengue is a viral infection transmitted by mosquitoes, common in tropical and subtropical areas. In many cases, infected people have no symptoms, but when symptoms do occur, they include high fever, headaches, nausea, and skin rashes. Most recover within one or two weeks, but occasionally the disease becomes severe and requires hospitalization, which can even be fatal. Although there are medications to relieve dengue symptoms, there is currently no specific treatment.

The information on the entity providing the data is presented in table 1.

Table 1: *Entity Information*

Area/dependency	Municipal Health and Environment Secretariat
Name/Entity	ESE Health Yopal
Department	Casanare
Municipality	Yopal
Order	Territorial
Sector	Health and Social Protection

3.1.2. Determination of project objectives

Develop a prediction model for the final diagnosis of the severity of dengue in patients treated in hospitals based on their historical care data. It is established that the minimum effectiveness of the prediction model must be 99%.

3.1.3. Plan tasks.

In this phase, it is recommended to adopt agile development approaches, Scrum is especially recommended, since the main objective is to develop ML models, which are related to the development of computer algorithms. And as indicated [12] Scrum is a widely recognized framework in the software development industry. This involves forming teams, organizing work into meaningful general tasks, and breaking them down into detailed activities.

3.2 Understanding data

The data set corresponds to the list of Dengue cases that occurred in the municipality of Yopal, department of Casanare in Colombia, disaggregated by type of person, symptoms and hospital services, each row is a case attended. The details of the dataset are presented in table 2.

Table 2: *Data Information*

Language	Spanish
Geographic coverage	Municipal – Yopal
Broadcast date	2022-10-24
Last update	2023-02-17
Rows	2010
Columns	54
Regulatory URL	Go.

3.2.1. Data exploration

The ideal at this point would be to present a detailed description of each variable or column in the data set. However, for reasons of practicality in the report, only the detail of the target variable is provided in Table 3.

Table 3. Target variable detail

Name Attribute	Final
Description	Classification of patients according to symptoms and follow-up protocol
Type of data	Categorical/ String
Distribution of values	Destinations: 3 Nulls: 0 Values: Dengue with alarm symptoms: 1020 (50.7%). Dengue without warning symptoms: 988 (49.3%)

A statistical summary of the numerical variables is relevant in this section. A concrete example of this type of summary is found in Table 4, where descriptive statistics of the variable "age_" are detailed.

Table 4: *Statistical description of the "age" variable*

count	2010
pissn	19.736816
std	18.016333
min	1
25%	7
fifty%	13
75%	27
Max	101

The results in Table 4 indicate the presence of outliers, this can also be seen in the data distribution in Figure 2.

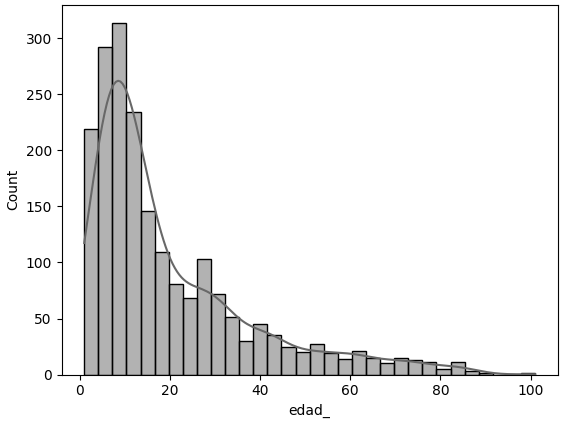

Fig.2. *Distribution of the age of the patients.*

The summary of the variable indicates that cases are significantly concentrated in patients under 13 years of age, which suggests a marked trend of the disease in children. These results support the statement made by ("Treatments for Dengue: A Global Dengue Alliance to Address Unmet Needs," 2023), which maintains that dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), two severe variants of the disease, have a higher incidence in children under 15 years of age.

This demographic analysis, which highlights children's vulnerability to severe dengue, has applications beyond medical research. In the context of machine learning, it could be used as an indicator to establish clusters in unlabeled data.

3.3 Data preparation

According to [13] "Data, in the real world, is usually incomplete and has inconsistencies. This is why it is necessary to prepare them before running analytics models."

A frequent problem in Big data projects are null and outlier values. Regarding the latter [14], they emphasize that the presence of outliers in the data

can have a significant impact on the evaluation of data models. ML, influencing the perception of model performance and the interpretation of important features.

For [15] it is crucial to understand that, in environments with a high number of dimensions, some algorithms do not work effectively. Dimensionality reduction addresses this challenge by converting a data set with multiple dimensions (variables or attributes) into data with lower dimensions, while ensuring that relevant information is preserved in a concise manner. This technique plays a vital role in simplifying complex data sets and is widely applied in the field of machine learning. It should be noted that there are distinctions between feature selection and dimensionality reduction, although both approaches are primarily intended to improve both efficiency and accuracy in data analysis.

Preparing the data implies ensuring its quality, and to achieve this, it is essential to carry out data cleaning, which according to [16] "In the data cleaning process (in English, data cleaning or data scrubbing) activities are carried out to detect, eliminate or correct corrupted or inappropriate instances in the data sets.

To carry out these activities, it is essential to review and establish business rules in collaboration with the various actors involved, such as service providers and users. Through this collaboration, valid data ranges or formats can be defined. In some cases, documentation may be the only means necessary to establish these rules, or it may complement the process.

In this phase of data preparation, measures are taken to improve the quality and usefulness of the data. Variables that do not provide significant information are identified and eliminated, such as those in which more than 95% of their values are identical. Missing values are replaced using the mode for categorical variables and the mean for numerical variables. Additionally, the entry date is transformed into the month number to consider possible seasonal patterns. A patient's days with symptoms in care are calculated by subtracting the date of symptom onset from the date of admission.

Continuing with the dimension reduction, a Backward Elimination algorithm is used, also known as Backward Elimination, a technique that, according to [17], has as its main purpose the construction of a high-quality multiple regression model that have as few attributes as possible, but without sacrificing the predictive capacity of the model.

At the end of this phase, the data set is defined as follows: 1994 rows and 12 columns. Table 3 presents the final variables that will be used in the modeling.

Table 5: *Variables classified for modeling.*

Column	***Dwriting***	***Guy***
week	Epidemiological week according to the current calendar	Number
sex_	Patient sex	Text
Lifecycle	Patient's life stage	Text
dolretroo	Semiological findings Retro ocular pain	Text
malgias	Myalgia semiological findings	Text
arthralgia	Semiological findings Arthralgias	Text
abdominal_pain	Semiological findings Abdominal pain	Text
threw up	Semiological findings vomit	Text
pac_hos_	The patient is hospitalized	Text
conduct	Patient behavior	Text
synt_days	Days with patient symptoms	Number
final	Final classification	Text

3.4 Modeling

According to [18], Machine Learning modeling uses labeled historical data to train models that can make accurate predictions on new unlabeled data. It starts with data collection, followed by division into training and test sets. The training set is used to teach the model to identify patterns and relationships between features and results. The evaluation is performed on the test set to measure the accuracy of the model. If necessary, parameters are adjusted or different algorithms are tested. Once deemed effective, the model is used to make predictions on new data.

Following this direction, libraries implemented in Python are used that provide various machine learning techniques to process data and generate models. The code corresponding to these activities is compiled in Jupyter notebooks available on the*GitHub repository*. Experimentation will include testing various algorithms and optimizing hyperparameters. The results of these evaluations will be presented in summary in the following sections.

The algorithms used are LogisticRegression, KNeighborsClassifier and RandomForestClassifier.

Logistic Regression, for [19], is a valuable analytical approach to solve classification problems, such as determining whether a new sample best relates to a specific category. The most common logistic regression model addresses a binary outcome; something that can take two values.

This approach can be applied to the context of classifying the severity of dengue in patients treated in health centers, allowing differentiation between cases of Dengue with alarm symptoms and cases of Dengue without alarm symptoms.

KNeighborsClassifier, according to [20] is a classifier that is based on the evidence provided by learning instances close to the pattern being classified. The parameters of the method are adjusted so as to maximize the evidential likelihood, which is an extension of the likelihood function designed to handle uncertain data. This classifier has been shown to outperform other methods in partially supervised learning situations.

The training set was used to train the KNeighborsClassifier model. This model fits the training data and learns to identify patterns that relate patient characteristics to the severity of Dengue.

RandomForestClassifier, according to [21] is used to try to improve accuracy compared to linear regression, since random forest can better approximate the relationship between targets and features.

Below, the outstanding results of each technique are presented along with their respective parameter combinations.

A first step is to use "Label Encoding" to convert categorical variables into numerical format, thus facilitating their processing by machine learning algorithms. The target variable is coded as 0 for Dengue with alarm symptoms and 1 for Dengue without alarm symptoms.

3.4.1. LogisticRegression

Parameters:C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver= 'lbfgs', tol=0.0001, verbose=0, warm_start=False.
Performance obtained:0.9038076152304609

Figure 3 presents the confusion matrix that summarizes the results of the model using the LogisticRegression technique in terms of successes and failures in the classification.

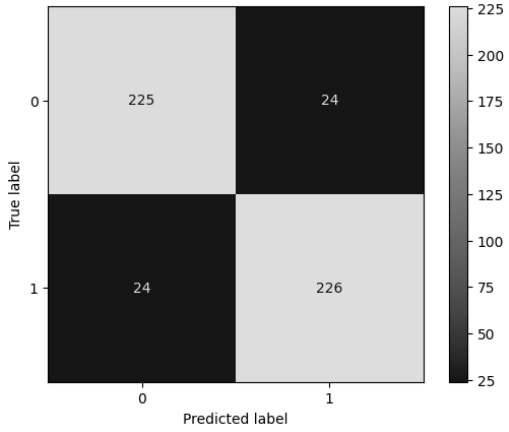

Fig.3. *LogisticRegression model confusion matrix*

3.4.2. KNeighborsClassifier

Parameters:algorithm='kd_tree', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=10, p=2, weights='uniform
Performance obtained:0.7414829659318637

Figure 4 shows the confusion matrix that summarizes the performance of the model using the KNeighborsClassifier technique.

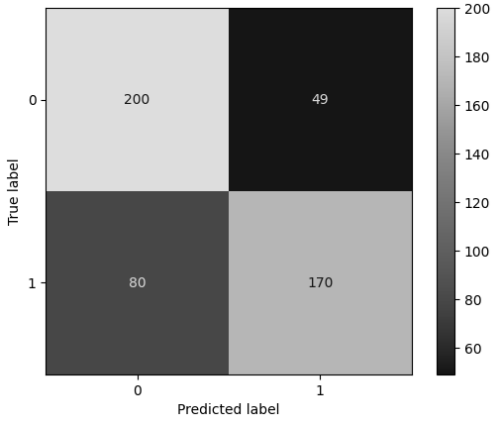

Fig.4. *KNeighborsClassifier model confusion matrix*

3.4.3. RandomForestClassifier

Parameters: bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf= 0.0, n_estimators=100, n_jobs=None, oob_score=False, random_state=42, verbose=0, warm_start=False
Performance obtained: 1.0

Figure 5 shows the confusion matrix that summarizes the performance of the model using the KNeighborsClassifier technique.

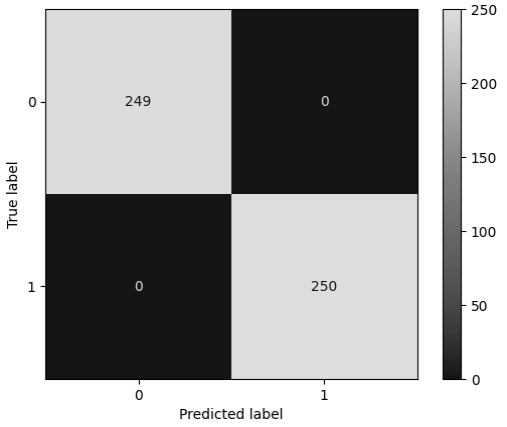

Fig.5. *RandomForestClassifier model confusion matrix*

During the model testing process, several models were employed, and specific configurations were applied to improve their performance. Key hyperparameters were tuned for each model to find optimal combinations.

3.5 Evaluation

After experimenting with the different models and testing different settings in the hyperparameters, it is the RandomForestClassifier technique that best adapts to the characteristics of the data and with which 100% efficiency is obtained in classifying the complexity of dengue.

3.6 Implementation

It would be ideal to evaluate the performance of the model with new data from the supplying entity, but these are not frequently updated in the portal where they are openly available. Another implementation alternative is to integrate the model with the data registration platform to automatically generate classifications of the severity of dengue. This would provide the professional with the option of using these predictions as a support tool when making the final diagnosis, whether dengue with alarm symptoms or without alarm symptoms.

4. CONCLUSIONS

The CRISP-DM methodology was applied to achieve an efficient classification of dengue severity in patients from Casanare, Colombia, using machine learning techniques such as LogisticRegression, KNeighborsClassifier and RandomForestClassifier.

The results show that the RandomForestClassifier model achieved a 100% correct classification rate in the severity of dengue patients, which could be essential for the early treatment of the disease and have a significant impact on public health in the region.

For future research, the integration of these models into the data registration platform is suggested to generate automatic predictions, providing health professionals with a support tool in diagnoses.

The effectiveness of data mining and machine learning as predictive tools for the management of dengue is proven, with important implications for medical care and disease prevention in the region.

REFERENCES

[1] Medina L., E. H. (2023). Big Data: Los Datos como Generadores de Valor. Universidad Peruana de Ciencias Aplicadas.

[2] Casas R., J., Nin G., J., & Julbe L., F. (2019). Big data: análisis de datos en entornos masivos. Editorial UOC.

[3] López M., J. J. y Zarza, G. (2017). La ingeniería del big data: cómo trabajar con datos. Editorial UOC. Barcelona, España.

[4] Maldonado, S. (2022). Analytics y Big Data: ciencia de los Datos aplicada al mundo de los negocios. RIL editores.

[5] Suarez L, A. A., Vazquez S., C. R., & Huffel, S. Van. (2018). Machine learning approaches for ambulatory electrocardiography signal processing.

[6] Rios Insua, D., & Gomez-Ullate Oteiza, D. (2019). Big data: conceptos, tecnologias y

aplicaciones. Editorial CSIC Consejo Superior de Investigaciones Cientificas.

[7] Arnst, M., Louppe, G., Van Hulle, R., Gillet, L., Bureau, F., & Denoël, V. (2022). A hybrid stochastic model and its Bayesian identification for infectious disease screening in a university campus with application to massive COVID-19 screening at the University of Liège. Mathematical Biosciences, 347. https://doi.org/10.1016/j.mbs.2022.108805

[8] Gutierrez-Barbosa, H., Medina-Moreno, S., Zapata, J. C., & Chua, J. V. (2020). Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country. Tropical Medicine and Infectious Disease, 5(4). https://doi.org/10.3390/tropicalmed5040156

[9] Gangula, R., Thirupathi, L., Parupati, R., Sreeveda, K., & Gattoju, S. (2023). Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns. Materials Today: Proceedings, 80, 3458–3463. https://doi.org/https://doi.org/10.1016/j.matpr.2021.07.270

[10] Castillo Romero, J. A. (2019). Big data. IFCT128PO. IC Editorial.

[11] Organización Mundial de La Salud. (2023). Dengue y dengue grave. WHO.

[12] Kadenic, M. D., Koumaditis, K., & Junker-Jensen, L. (2023). Mastering scrum with a focus on team maturity and key components of scrum. Information and Software Technology, 153, 107079. https://doi.org/https://doi.org/10.1016/j.infsof.2022.107079

[13] Treatments for dengue: a Global Dengue Alliance to address unmet needs. (2023). The Lancet Global Health. https://doi.org/https://doi.org/10.1016/S2214-109X(23)00362-5

[14] Nariya, M. K., Mills, C. E., Sorger, P. K., & Sokolov, A. (2023). Paired evaluation of machine-learning models characterizes effects of confounders and outliers. Patterns, 4(8), 100791. https://doi.org/https://doi.org/10.1016/j.patter.2023.100791

[15] Menoyo R., D., Garcia L., E., & Garcia C., A. (2021). Fundamentos de la ciencia de datos. Editorial Universidad de Alcala.

[16] Minguillon, J., Casas, J., & Minguillon, J. (2017). Mineria de datos: modelos y algoritmos. Editorial UOC.

[17] Kotu, V., & Deshpande, B. (2019). Chapter 14 - Feature Selection. In V. Kotu & B. Deshpande (Eds.), Data Science (Second Edition) (pp. 467–490). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00014-9

[18] Caballero, R., & Martin, E. (2022). Las bases de big data y de la inteligencia artificial. Los libros de la Catarata.

[19] Edgar, T. W., & Manz, D. O. (2017). Chapter 4 - Exploratory Study. In T. W. Edgar & D. O. Manz (Eds.), Research Methods for Cyber Security (pp. 95–130). Syngress. https://doi.org/https://doi.org/10.1016/B978-0-12-805349-2.00004-2

[20] Denœux, T., Kanjanatarakul, O., & Sriboonchitta, S. (2019). A new evidential K-nearest neighbor rule based on contextual discounting with partially supervised learning. International Journal of Approximate Reasoning, 113, 287–302. https://doi.org/https://doi.org/10.1016/j.ijar.2019.07.009

[21] Malik, A., Javeri, Y. T., Shah, M., & Mangrulkar, R. (2022). Chapter 11 - Impact analysis of COVID-19 news headlines on global economy. In R. C. Poonia, B. Agarwal, S. Kumar, M. S. Khan, G. Marques, & J. Nayak (Eds.), Cyber-Physical Systems (pp. 189–206). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-824557-6.00001-7