Digital Object Identifier: 10.24054/rcta.v2i44.3021

Analysis and implementation of clustering in dengue cases using an unsupervised learning algorithm

Análisis e implementación de clustering en casos de dengue mediante algoritmo de aprendizaje no supervisado.

MSc. Miguel Alberto Rincón Pinzón[1], MSc. Carlos Alberto Mejía Rodríguez [1]
MSc. Erney Alberto Ramirez Camargo[1], Esp. Lina Marcela Arévalo Vergel [1]

[1]***Universidad Popular del Cesar,*** *Systems Engineering, GIDEATIC Research Group, Aguachica, César, Colombia.*

Correspondence: *calbertomejia@unicesar.edu.co*

Received: *january 21, 2024.* ***Accepted:*** *june 11, 2024.* ***Published:*** *july 24, 2024.*

How to cite: *M. A. Rincón Pinzón, C. A. Mejía Rodríguez, E. A. Ramírez Camargo, and L. M. Arévalo Vergel, "Analysis and implementation of clustering in dengue cases using unsupervised learning algorithm", RCTA, vol. 2, no. 44, pp. 104–111, Jul. 2024. Retrieved from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3021*

Abstract: This study focuses on the application of unsupervised learning algorithms, specifically clustering techniques to analyze the incidence of dengue in San Juan, Puerto Rico, and Iquitos, Peru. The main objective is to test the effectiveness of these algorithms in identifying hidden patterns in the dataset, composed of environmental, climatic and dengue case information. The research allowed us to verify the importance of selecting the appropriate clustering technique, evidenced by the variable performance of the methods used. The results reveal the usefulness of unsupervised learning to understand the spread of dengue, highlighting the need to carefully consider the choice of algorithm for future epidemiological and environmental analyses.

Key words: Unsupervised Learning, Clustering, Dengue, Segmentation.

Resumen: Este estudio se enfoca en la aplicación de algoritmos de aprendizaje no supervisado, específicamente técnicas de clustering para analizar la incidencia del dengue en San Juan, Puerto Rico, e Iquitos, Perú. El objetivo principal es probar la eficacia de estos algoritmos en la identificación de patrones ocultos en el conjunto de datos, compuesto por información ambiental, climática y casos de dengue. La investigación permitió comprobar la importancia de seleccionar la técnica de clusterización adecuada, evidenciada por el rendimiento variable de los métodos utilizados. Los resultados revelan la utilidad del aprendizaje no supervisado para comprender la propagación del dengue, resaltando la necesidad de considerar cuidadosamente la elección del algoritmo para análisis epidemiológicos y ambientales futuros.

Palabras clave: Aprendizaje No Supervisado, Clustering, Dengue, Segmentación.

1. INTRODUCTION

Today's technology has redefined the basic concept of data. Once limited to being text and numbers in spreadsheets or relational databases, today data is a dynamic asset, created and massively consumed by owners of digital devices [1]. This paradigm shift not only redefines our perception of data, but also drives the evolution of fields such as machine learning. This discipline, which according to [2], is a branch of artificial intelligence that focuses on designing systems to learn from data through training. These systems can improve with experience, generating predictive models based on previous learning. In this discipline, there are two main types of algorithms: supervised and unsupervised, chosen based on the desired output.

Supervised algorithms are useful in situations where there are fewer labeled instances to learn and a large amount of unlabeled data. In these cases, supervised algorithms, which include classification and regression problems, are presented as an optimal solution [3].

Unsupervised algorithms are distinguished by lacking an objective function, which implies learning without a predefined goal. This approach focuses on uncovering patterns and associations in the data. These algorithms have a variety of practical applications, from optimizing shelf locations to identifying correlated mechanical failures, to name a few examples [4].

Clustering analysis and dimensionality reduction are prominent examples of the application of unsupervised learning. Another example can be the management of MP3 files without metadata, grouping emerges as the optimal solution to automatically organize similar songs into identified categories [5].

The process of clustering, an unsupervised learning method, is used to divide inputs into groups, which were not known in advance. This process involves the formation of groups based on the similarities between the instances [6]. In this context, this article focuses on the application of unsupervised learning, with emphasis on clustering techniques, to analyze and understand patterns related to dengue incidence in two different locations: San Juan, Puerto Rico, and Iquitos, Peru.

Some relevant unsupervised learning algorithms that involve clustering are k-Means, Hierarchical Cluster Analysis (HCA), and Expectation Maximization [7].

The K-Means algorithm is used to cluster and reveal patterns in unlabeled data, the main objective is to separate groups with similar characteristics and assign them to clusters. K-Means, a common tool for this purpose, seeks to identify K centroids representative of the center of clusters and assign labels to training data. This is especially true when working with unlabeled data, and the number of groups, K, represents the goal of the clustering [8]. An example of clustering with k-means can be seen in Figure 1.

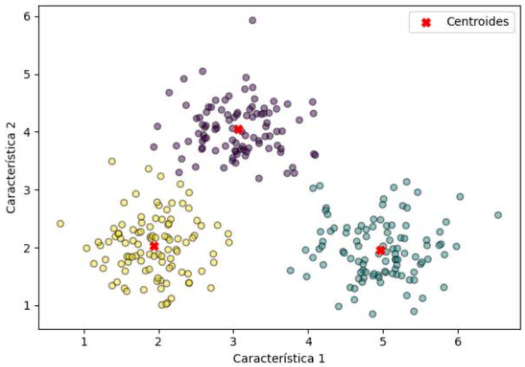

Fig. 1. *Clustering with k-means.*
Source: *Authors' own creation.*

The K-Means algorithm has several practical applications. In the business environment, it is used for market segmentation, identifying customer groups with defined characteristics. This allows for segment-specific treatment. In addition, in the classification of books, movies or other documents, as well as in the detection of fraud and criminal activities, the K-Means algorithm analyzes data to identify patterns and similarities, making it easier to understand the behavior of customers or users [9].

Hierarchical clustering is another clustering strategy that employs graph theory and unsupervised machine learning techniques to organize related or related elements in a hierarchical manner. This approach makes it possible to recognize the underlying hierarchical structure in the dataset or universe in question [10] . Hierarchical clustering starts by creating as many clusters as there are instances in the dataset, each initially containing only one instance. It then iteratively identifies the two clusters with the minimum distance between them, such as the Euclidean distance, and merges them into a new cluster. This process is repeated until only a single cluster remains. The result is a dendrogram that visualizes the hierarchical organization of the instances [11]. An example is presented in Figure 2.

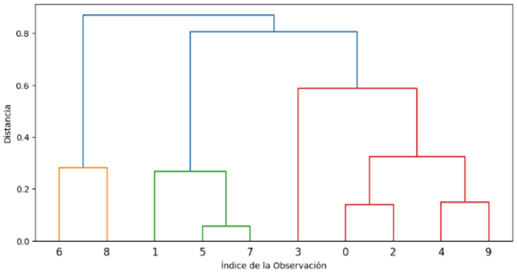

Fig. 2. *Dedogram Hierarchical clustering*
Source: *Authors' own creation.*

Relevant aspects of machine learning are feature selection and dimensionality reduction to improve model performance on feature-rich datasets. These practices not only optimize the effectiveness of supervised algorithms but are also critical in the context of unsupervised algorithms, where the identification and choice of relevant features and the efficient management of dimensionality contribute significantly to the quality of the results [12].

Dengue affects populations in tropical and subtropical regions, and its incidence is influenced by environmental, climatic, and geographical factors [13]. Understanding the relationship between these variables and the spread of dengue is essential to developing effective prevention and control strategies [14]. In this study, unsupervised learning was applied to a dataset encompassing environmental, climatic, and dengue case information in two different locations for the purpose of clustering. The main purpose of this approach is twofold: first, to identify patterns and relationships between variables that can provide valuable information on the spread of dengue; and second, reducing the dimensionality of the data to improve the accuracy of weekly case predictions.

Commonly, the development of a machine learning project goes through several stages, including data collection and cleansing, feature engineering, as well as model training and testing. The result of this process is a model capable of predicting a dependent variable or identifying patterns in the data [15].

According to [16], to harness or obtain knowledge from large volumes of information, specific methodologies have been designed that offer a structured route to obtain, debug and effectively apply the knowledge acquired. Crucial examples are the KDD (Knowledge Discovery in Databases) and CRISP-DM (Cross-Industry Standard Process for Data Mining) model. KDD establishes key stages, such as selection, preparation, pattern finding, and model evaluation. CRISP-DM, derived from KDD, adapts to the overall needs, organizing the process from understanding the business to implementing results. These methodologies provide a step-by-step guide to successfully carrying out data mining projects.

The results and findings of this study point to the importance of the right choice of clustering technique, as different methods offer clustering depending on the nature of the data. Likewise, the need to consider the scale of the variables used in the analysis was evidenced, since certain techniques may be sensitive to differences in the scale, which could require a prior normalization of the data. Taken together, the study contributes to the understanding of the spread of dengue in two critical regions and highlights the relevance of unsupervised learning techniques in the exploration of epidemiological and environmental data.

2. METHODOLOGY

In engineering, data-driven methods, such as data mining and machine learning, are increasingly being employed to develop solutions and optimize processes [17].

Data mining project managers benefit from using a standard process model, such as CRISP-DM, as they reduce costs and times, facilitate knowledge transfer, and promote the reuse of best practices [18].

The CRISP-DM methodology, according to [19], provides an open-access framework based on the KDD process for data mining projects. It consists of six key phases: business understanding, data understanding, data preparation, modeling, assessment, and implementation, as shown in Figure 3.

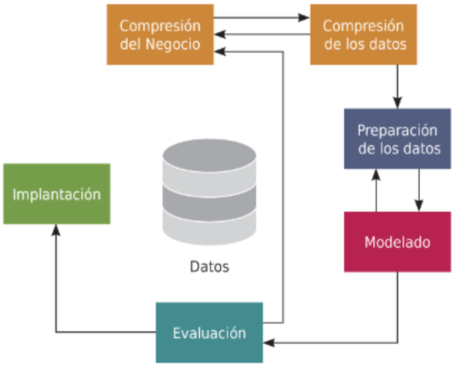

Fig. 3. *CRISP-DM Flow* ***Source:*** *Castillo Romero, J. A. (2019). Big data. IFCT128PO. IC Editorial.*

For the analysis of dengue cases, the CRISP-DM methodology will be applied, this structured framework will guide the clustering process, facilitating the understanding of patterns in epidemiological information.

3. RESULTS

The structuring of the work follows the stages and actions established by the CRISP-DM methodology. The results obtained in each phase are detailed below.

3.1 Understanding the business.

When a project is carried out following the CRISP-DM methodology, it starts by discussing the project with stakeholders to precisely define their requirements and expectations. After clarifying these aspects, you can begin to analyze the data to assess the possibility of meeting those goals [20].

The purpose of this phase is to convert the required objectives into measurable technical objectives, to gather existing knowledge about physical and process-related interactions, to develop a data mining idea and a technical concept for data acquisition [21].

As for the particular project advanced, in the business understanding phase, the study focuses on participation in a Machine Learning (ML) competition of the DrivenData platform. The competition focuses on predicting "total_cases" in the test dataset, which spans five years for San Juan and Iquitos, with three years respectively. However, the objective of this research is to perform clustering to reduce dimensionality and discover patterns that can facilitate the subsequent prediction of cases.

In this paper, we propose to carry out an analysis of the dataset using an Unsupervised Learning approach. In a first phase, a comprehensive evaluation of the data set will be carried out, verifying the existence of significant differences in the variables between the cities of Iquitos and San Juan that facilitate an adequate clustering. To achieve this, the KMEANS method will be employed.

Subsequently, the analysis will be segmented for each city to verify the presence of heterogeneous segments. It is also proposed to evaluate the Hierarchical by Agglomeration methods.

3.2 Understanding the data.

The variables (columns) present in the training database are identified. From Python coding the dataset is loaded and summarized, it is observed that the dataset consists of 23 variables and 1456 records. Below is a list of the variables:

City & Date Indicators:
"city", "week_start_date".

Daily Weather Station Measurements:
"station_max_temp_c", "station_min_temp_c", "station_avg_temp_c", "station_precip_mm", "station_diur_temp_rng_c".

Satellite precipitation measurements:
"precipitation_amt_mm."

Climate Forecast System Measurements:
"reanalysis_sat_precip_amt_mm",
"reanalysis_dew_point_temp_k",
"reanalysis_air_temp_k",
"reanalysis_relative_humidity_percent",
"reanalysis_specific_humidity_g_per_kg",
"reanalysis_precip_amt_kg_per_m2",
"reanalysis_max_air_temp_k",
"reanalysis_min_air_temp_k",
"reanalysis_avg_temp_k", "reanalysis_tdtr_k"

Satellite vegetation:
"ndvi_se", "ndvi_sw", "ndvi_ne", "ndvi_nw".

Different queries are made in terms of statistical summaries, null counts, among others that allow us to understand the variables. A graphical result of the outlier exploration is presented in Figure 4.

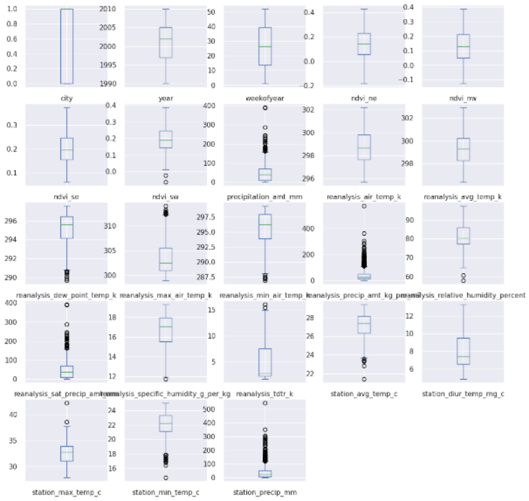

Fig. 4. *Exploring Outliers*
Source: *Authors' own creation.*

3.3 Data preparation

A widely employed modeling procedure in the creation of machine learning applications is the cross-industry standard process for data mining (CRISP-DM) [22]. Following the incremental path of the process corresponds to the preparation of the data, in this phase data optimization is done.

Null values are eliminated, and outliers are adjusted by the correction process of upper and lower outliers.

3.4 Modeling

3.4.1. Clustering KMEANS Method

The main purpose of K-Means is to partition a specific dataset into K clusters (where K is a hyperparameter) and provide the centroid for each data sample [23].

The KMEANS method is applied, a summary of the main techniques and hyperparameters used is presented in Table 1.

***Table 1:** For KMEANS Adjustment* Hyperparameters

Hyperparameters	Settings/Settings
n_clusters	3
Init	'k-means++'
max_iter	300
n_init	10

***Source:** Authors' own creation.*

In terms of techniques, dimensionality reduction is applied with PCA (Principal Component Analysis), PCA is used with n_components set to 2 to reduce the dimensionality of the data to two dimensions, allowing its visualization in a two-dimensional plane. The original dataset (df_kmeans) and centroids are transformed using the trained PCA model.

The graph is generated for analysis of results, the graph obtained is visualized in figure 5.

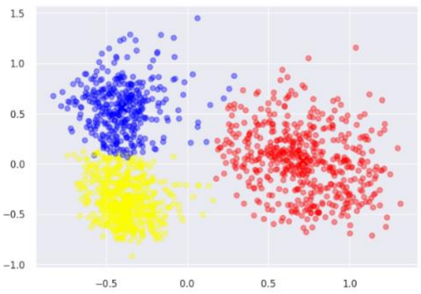

***Fig. 5.** Result of Clustering by KMEANS Method.*
***Source:** Authors' own creation.*

3.4.2. Agglomerative Hierarchical Method

The Hierarchical Method is a grouping technique that organizes data into a hierarchical structure, represented as a tree or dendrogram.

To create a dendrogram with respect to dengue data, the "linkage" function is used to calculate the distances between the point pairs and determine how the clusters should be merged. In this case, the 'ward' binding method is used, which minimizes intracluster variance when merging clusters. The result represents the clustering hierarchy and allows the cluster structure to be identified at different levels of similarity. The resulting figure shows the relationship between the points and how they are grouped into clusters, see Figure 6.

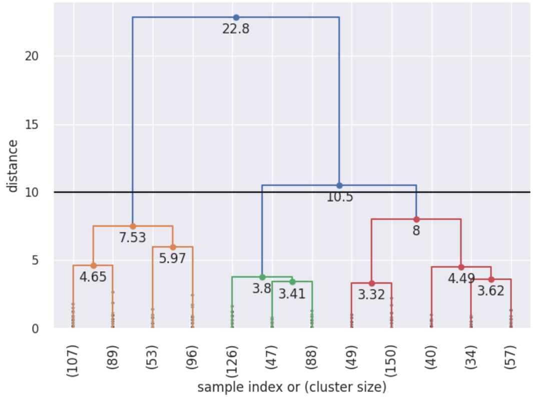

***Fig. 6**. Hierarchical Clustering Dendrogram.*
***Source:** Authors' own creation.*

3.5 Evaluation

In the evaluation phase, the results are verified in terms of the achievement of the defined objectives. Therefore, it is necessary to interpret the results and define additional actions. In addition, the process should be reviewed in general [24].

When analyzing the dendrogram, it is observed that approximately 5 clearly differentiated groups are identified, marking this number in the position where the vertical distance is maximum (10). This result suggests the presence of distinct structures in the dengue data, allowing a first approach to the identification of significant groups.

3.6 Implementation

This last phase is responsible for the implementation of the results of the data mining project to ensure their availability and meet the needs of the end users. However, CRISP-DM does not detail the specification of the implementation requirements.

For this reason, professionals adapt the reference process to measure the compliance of the implemented solution and its final use. It is important to note that projects based on unsupervised models may not need deployment, as their goal is to discover features and interpret them in the context of a specific problem [25].

Since this is an unsupervised machine learning project, no additional deployment or deployment is required. The results are obtained through the application of the selected algorithms, highlighting that the Agglomerative Hierarchical Method showed a superior performance in the identification of groups.

4. CONCLUSIONS

The efficacy and usefulness of unsupervised learning, in particular, clustering techniques, in the exploration of epidemiological and environmental data related to the spread of dengue in San Juan, Puerto Rico, and Iquitos, Peru, is demonstrated. The results underscore the importance of understanding patterns and relationships hidden in the data, which can provide critical insights for dengue prevention and control in these regions.

A fundamental aspect that has been highlighted throughout this study is the appropriate choice of clustering technique. In the development of the analysis, it was observed that for this particular case the hierarchical method worked quite well, producing clusters of similar size and adapting appropriately to the data, while the use of density-based methods showed poor performance in the identification of groups. These results highlight the need to carefully select the most appropriate clustering technique according to the nature of the data and the objectives of the analysis.

In addition, the importance of considering the scale of the variables used in the analysis has been demonstrated, as some techniques are sensitive to differences in the scale of the variables. Pre-normalization of data is an essential step to ensure consistent and reliable results in the clustering process.

The results of this study provide valuable information for those interested in this field or professionals working with this data, using supervised algorithms in future analyses.

It is recommended to continue exploring and applying unsupervised learning techniques in epidemiological and environmental research, to continue uncovering hidden patterns and causal relationships that help address public health challenges more efficiently and effectively.

REFERENCES

[1] J. Quddus, *Machine Learning with Apache Spark Quick Start Guide: Uncover Patterns, Derive Actionable Insights, and Learn from Big Data Using MLlib*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5626693

[2] J. Bell, *Machine Learning: Hands-On for Developers and Technical Professionals*. Somerset, UNITED STATES: John Wiley & Sons, Incorporated, 2014. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1818248

[3] U. N. Dulhare, K. Ahmad, and K. A. Bin Ahmad, *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6268187

[4] R. Gopalakrishnan and A. Venkateswarlu, *Machine Learning for Mobile: Practical Guide to Building Intelligent Mobile Applications Powered by Machine Learning*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5628277

[5] R. Karim, *Machine Learning with Scala Quick Start Guide: Leverage Popular Machine Learning Algorithms and Techniques and Implement Them in Scala*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5764277

[6] M. A. Jabbar, *Machine Learning Methods for Signal, Image and Speech Processing*. Aalborg, DENMARK: River Publishers, 2021. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=29002971

[7] O. Campesato, *Artificial Intelligence, Machine Learning, and Deep Learning*.

Bloomfield, UNITED STATES: Mercury Learning & Information, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6032875

[8] W.-M. Lee, *Python Machine Learning*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5747364

[9] Z. Nagy, *Artificial Intelligence and Machine Learning Fundamentals: Develop Real-World Applications Powered by the Latest AI Advances*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5620491

[10] E. Jurczenko, *Machine Learning for Asset Management: New Developments and Financial Applications*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6268186

[11] G. Kyriakides and K. G. Margaritis, *Hands-On Ensemble Learning with Python: Build Highly Optimized Ensemble Machine Learning Models Using Scikit-Learn and Keras*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5837325

[12] K. Ramasubramanian and J. Moolayil, *Applied Supervised Learning with R: Use Machine Learning Libraries of R to Build Models That Solve Business Problems and Predict Future Trends*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5784240

[13] H. Gutierrez-Barbosa, S. Medina-Moreno, J. C. Zapata, and J. V Chua, "Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country," *Trop Med Infect Dis*, Vol. 5, No. 4, 2020, DOI: 10.3390/tropicalmed5040156.

[14] R. Gangula, L. Thirupathi, R. Parupati, K. Sreeveda, and S. Gattoju, "Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns," *Mater Today Proc*, vol. 80, pp. 3458–3463, 2023, doi: https://doi.org/10.1016/j.matpr.2021.07.270.

[15] S. Mehrotra and A. Grade, *Apache Spark Quick Start Guide: Quickly Learn the Art of Writing Efficient Big Data Applications with Apache Spark*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5675596

[16] S. Maldonado, *Analytics and Big Data: Data Science applied to the business world*. RIL editors, 2022. [Online]. Available: https://elibro.net/es/lc/biblioupc/titulos/225562

[17] E. Alogogianni and M. Virvou, "Addressing the issue of undeclared work – Part I: Applying associative classification per the CRISP-DM methodology," *Intelligent decision technologies*, vol. 15, no. 4, pp. 721–747, 2021.

[18] W. Y. Ayele, "Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset," *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 6, 2020, doi: https://doi.org/10.14569/IJACSA.2020.0110603.

[19] J. A. Castillo Romero, *Big data. IFCT128PO*. IC Editorial, 2019. [Online]. Available: https://elibro.net/es/lc/biblioupc/titulos/124254

[20] A. So, T. V. Joseph, R. Thas John, A. Worsley, and S. Asare, *The Data Science Workshop: A New, Interactive Approach to Learning Data Science*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6033288

[21] H. Wiemer, L. Drowatzky, and S. Ihlenfeldt, "Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model," *Applied Sciences*, Vol. 9, No. 12, 2019, doi: https://doi.org/10.3390/app9122407.

[22] M. T. H. Suhendar and Y. Widyani, "Machine Learning Application Development Guidelines Using CRISP-DM and Scrum Concept," in *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)*, 2023, pp.

168–173. doi: 10.1109/ICoDSE59534.2023.10291438.

[23] A. Cheng, "Evaluating Fintech industry's risks: A preliminary analysis based on CRISP-DM framework," *Financ Res Lett*, vol. 55, p. 103966, 2023, doi: https://doi.org/10.1016/j.frl.2023.103966.

[24] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: https://doi.org/10.1016/j.procs.2021.01.199.

[25] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data Knowl Eng*, vol. 139, p. 102013, 2022, doi: https://doi.org/10.1016/j.datak.2022.102013.