ISSN: 1692-7257 - Volumen 2 – Número 44 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i44.3021

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

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

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,*** *Ingeniería de sistemas, Grupo de Investigación GIDEATIC, Aguachica, César, Colombia.*

Correspondencia: *calbertomejia@unicesar.edu.co*

Recibido: *21 enero 2024.* ***Aceptado:*** *11 junio 2024.* ***Publicado:*** *24 julio 2024.*

Cómo citar: *M. A. Rincón Pinzón, C. A. Mejía Rodríguez, E. A. Ramírez Camargo, y L. M. Arévalo Vergel, «Análisis e implementación de clustering en casos de dengue mediante algoritmo de aprendizaje no supervisado», RCTA, vol. 2, n.º 44, pp. 104–111, jul. 2024. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3021*

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.

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 evaluate the effectiveness of these algorithms in identifying hidden patterns in the data set, composed of environmental, climatic information and dengue cases. 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 for understanding the spread of dengue, highlighting the need to carefully consider the choice of algorithm for future epidemiological and environmental analyses.

Keywords: Unsupervised Learning, Clustering, Dengue, Segmentation.

1. INTRODUCCIÓN

La tecnología actual ha redefinido el concepto básico de datos. Antes limitados a ser texto y números en hojas de cálculo o bases de datos relacionales, hoy los datos son activos dinámicos, creados y consumidos masivamente por poseedores de dispositivos digitales [1]. Este cambio de paradigma no solo redefine nuestra percepción de los datos, sino que también impulsa la evolución de campos como el aprendizaje automático. Esta disciplina, que según [2], es una rama de la inteligencia artificial que se centra en diseñar sistemas para aprender de los datos mediante entrenamiento. Estos sistemas pueden mejorar con la experiencia, generando modelos predictivos basados en el aprendizaje previo. En esta disciplina, existen dos tipos principales de algoritmos: supervisado y no supervisado, elegidos en función de la salida deseada.

Los algoritmos supervisados resultan útiles en situaciones donde hay un número menor de instancias etiquetadas para aprender y una gran cantidad de datos no etiquetados. En estos casos, los algoritmos supervisados, que incluyen problemas de clasificación y regresión, se presentan como una solución óptima [3].

Los algoritmos no supervisados se distinguen por carecer de una función objetivo, lo que implica un aprendizaje sin una meta predefinida. Este enfoque se centra en descubrir patrones y asociaciones en los datos. Estos algoritmos tienen diversas aplicaciones prácticas, desde optimizar ubicaciones en estanterías hasta identificar fallos mecánicos correlacionados, por citar algunos ejemplos [4].

El análisis de clustering y la reducción de dimensionalidad son ejemplos destacados de la aplicación de aprendizaje no supervisado. Otro ejemplo puede ser la gestión de archivos MP3 sin metadatos, el agrupamiento emerge como la solución óptima para organizar automáticamente canciones similares en categorías identificadas [5].

El proceso de clustering, un método de aprendizaje no supervisado se emplea para dividir las entradas en grupos, los cuales no eran conocidos de antemano. Este proceso implica la formación de grupos según las similitudes entre las instancias [6]. En este contexto, el presente artículo se centra en la aplicación del aprendizaje no supervisado, con énfasis en técnicas de clustering, para analizar y comprender patrones relacionados con la incidencia del dengue en dos ubicaciones distintas: San Juan, Puerto Rico, e Iquitos, Perú.

Algunos algoritmos relevantes de aprendizaje no supervisado que involucran agrupamiento son: k-Means, Análisis Jerárquico de Clústeres (HCA), y Expectation Maximization [7].

El algoritmo K-Means se emplea para realizar clustering y revelar patrones en datos no etiquetados, el objetivo principal es separar grupos con características similares y asignarlos a clústeres. K-Means, es una herramienta común para este propósito, busca identificar K centroides representativos del centro de los clústeres y asignar etiquetas a los datos de entrenamiento. Esto se aplica especialmente cuando se trabaja con datos no etiquetados, y el número de grupos, K, representa el objetivo del agrupamiento [8]. Ejemplo de clustering con k-means se puede apreciar en la Figura 1.

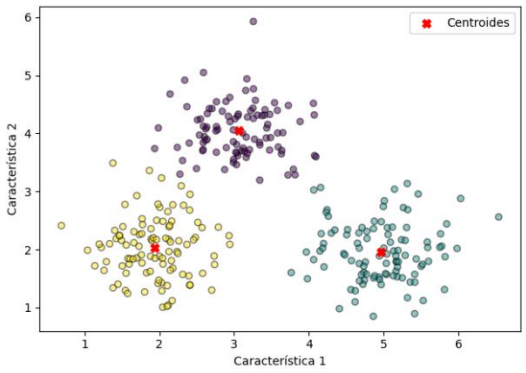

Fig. 1. *Clustering con k-means.*
Fuente: *elaboración propia.*

El algoritmo K-Means tiene diversas aplicaciones prácticas. En el ámbito empresarial, se utiliza para la segmentación de mercados, identificando grupos de clientes con características definidas. Esto permite un tratamiento específico para cada segmento. Además, en la clasificación de libros, películas u otros documentos, así como en la detección de fraudes y actividades criminales, el algoritmo K-Means analiza datos para identificar patrones y similitudes, facilitando la comprensión del comportamiento de clientes o usuarios [9].

El clustering jerárquico es otra estrategia de agrupamiento que emplea técnicas de teoría de grafos y aprendizaje automático no supervisado para organizar elementos afines o relacionados de forma jerárquica. Este enfoque permite reconocer la estructura jerárquica subyacente en el conjunto de datos o universo en cuestión [10]. El clustering

jerárquico comienza creando tantos clústeres como instancias en el conjunto de datos, cada uno inicialmente conteniendo solo una instancia. Luego, de manera iterativa, identifica los dos clústeres con la distancia mínima entre ellos, como la distancia euclidiana, y los fusiona en un nuevo clúster. Este proceso se repite hasta que solo queda un único clúster. El resultado es un dendrograma que visualiza la organización jerárquica de las instancias [11]. Se presenta un ejemplo en la Figura 2.

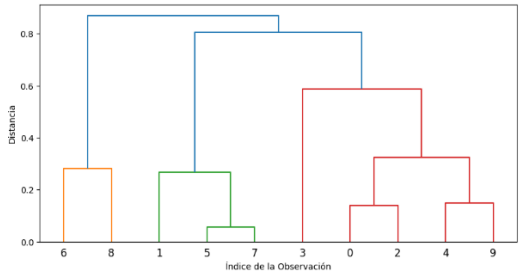

Fig. 2. *Dedograma clustering jerárquico*
Fuente: *elaboración propia.*

Aspectos relevantes del aprendizaje automático son la selección de características y la reducción de dimensionalidad para mejorar el rendimiento de los modelos en conjuntos de datos con numerosas características. Estas prácticas no solo optimizan la eficacia de los algoritmos supervisados, sino que también son fundamentales en el contexto de los algoritmos no supervisados, donde la identificación y elección de características pertinentes y la gestión eficiente de la dimensionalidad contribuyen significativamente a la calidad de los resultados [12].

El dengue afecta a poblaciones en regiones tropicales y subtropicales, y su incidencia está influenciada por factores ambientales, climáticos y geográficos [13]. Comprender la relación entre estas variables y la propagación del dengue es esencial para desarrollar estrategias efectivas de prevención y control [14]. En este estudio, se aplicó el aprendizaje no supervisado a un conjunto de datos que abarca información ambiental, climática y casos de dengue en dos ubicaciones diferentes con el propósito de realizar clustering. La finalidad principal de este enfoque es doble: en primer lugar, identificar patrones y relaciones entre las variables que puedan ofrecer información valiosa sobre la propagación del dengue; y, en segundo lugar, reducir la dimensionalidad de los datos para mejorar la precisión de las predicciones de casos semanales.

Comúnmente, el desarrollo de un proyecto de aprendizaje automático atraviesa diversas etapas, que incluyen la recolección y limpieza de datos, la ingeniería de características, así como el entrenamiento y las pruebas del modelo. El resultado final de este proceso es un modelo capaz de predecir una variable dependiente o identificar patrones en los datos [15].

Según [16], para aprovechar u obtener conocimiento de grandes volúmenes de información se han diseñado metodologías específicas que ofrecen una ruta estructurada para obtener, depurar y aplicar eficazmente el conocimiento adquirido. Ejemplos cruciales son el modelo KDD (Knowledge Discovery in Databases) y CRISP-DM (Cross-Industry Standard Process for Data Mining). KDD establece etapas clave, como selección, preparación, búsqueda de patrones y evaluación de modelos. CRISP-DM, derivado de KDD, se adapta a las necesidades generales, organizando el proceso desde la comprensión del negocio hasta la implementación de resultados. Estas metodologías proporcionan una guía paso a paso para llevar a cabo exitosamente proyectos de minería de datos.

Los resultados y hallazgos de este estudio señalan la importancia de la elección adecuada de la técnica de clusterización, ya que los diferentes métodos ofrecen agrupamientos en función de la naturaleza de los datos. Asimismo, se evidenció la necesidad de considerar la escala de las variables utilizadas en el análisis, ya que ciertas técnicas pueden ser sensibles a las diferencias en la escala, lo que podría requerir una normalización previa de los datos. En conjunto, el estudio contribuye a la comprensión de la propagación del dengue en dos regiones críticas y destaca la relevancia de las técnicas de aprendizaje no supervisado en la exploración de datos epidemiológicos y ambientales.

2. METODOLOGÍA

En ingeniería, los métodos basados en datos, como la minería de datos y el aprendizaje automático, están siendo cada vez más empleados para desarrollar soluciones y optimizar procesos [17] .

Los administradores de proyectos de minería de datos se benefician al usar un modelo estándar de procesos, como CRISP-DM, ya que reducen costos y tiempos, facilitan la transferencia de conocimientos y promueven la reutilización de las mejores prácticas [18].

La metodología CRISP-DM, según [19], proporciona un marco de acceso libre basado en el proceso KDD para proyectos de minería de datos. Consta de seis fases clave: comprensión del

negocio, comprensión de los datos, preparación de datos, modelado, evaluación e implementación, como se muestra en la figura 3.

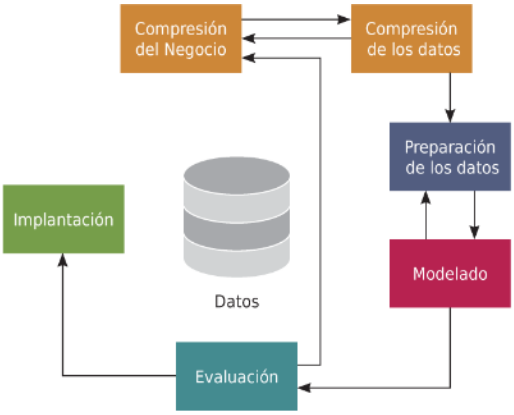

Fig. 3. *Flujo CRISP-DM*
Fuente: *Castillo Romero, J. A. (2019). Big data. IFCT128PO. IC Editorial.*

Para el análisis de los casos de dengue, se aplicará la metodología CRISP-DM, este marco estructurado guiará el proceso de clustering, facilitando la comprensión de patrones en la información epidemiológica.

3. RESULTADOS

La estructuración del trabajo sigue las etapas y acciones establecidas por la metodología CRISP-DM. A continuación, se detallan los resultados obtenidos en cada fase.

3.1 Comprensión del negocio

Cuando se lleva a cabo un proyecto siguiendo la metodología CRISP-DM, se inicia discutiendo el proyecto con los interesados para definir de manera precisa sus requisitos y expectativas. Después de aclarar estos aspectos, se puede comenzar a analizar los datos para evaluar la posibilidad de cumplir con esos objetivos [20].

El propósito de esta fase es convertir los objetivos requeridos en objetivos técnicos mensurables, recopilar conocimientos existentes sobre las interacciones físicas y relacionadas con el proceso, desarrollar una idea de minería de datos y un concepto técnico para la adquisición de datos [21].

En cuanto al proyecto particular adelantado, en la fase de comprensión del negocio, el estudio se centra en la participación en una competencia de Machine Learning (ML) de la plataforma DrivenData. La competencia se centra en predecir "total_cases" en el conjunto de datos de prueba, que abarca cinco años para San Juan e Iquitos, con tres años respectivamente. Aunque el objetivo de esta investigación es realizar clustering para reducir la dimensionalidad y descubrir patrones que puedan facilitar la predicción posterior de los casos.

En este trabajo se plantea llevar a cabo un análisis del conjunto de datos mediante un enfoque de Aprendizaje No Supervisado. En una primera fase, se realizará una evaluación integral del conjunto de datos, verificando la existencia de diferencias significativas en las variables entre las ciudades de Iquitos y San Juan que faciliten una adecuada clusterización. Para lograr esto, se empleará el método KMEANS.

Posteriormente, se segmentará el análisis para cada ciudad con el propósito de verificar la presencia de segmentos heterogéneos. Se propone además la evaluación de los métodos Jerárquico por Aglomeración.

3.2 Comprensión de los datos

Se identifican las variables (columnas) presentes en la base de datos de entrenamiento. A partir de codificación en Python se cargar y resumir el conjunto de datos, se observa que el dataset consta de 23 variables y 1456 registros. A continuación, se presenta un listado de las variables:

Indicadores de ciudad y fecha:
“city”, “week_start_date”.

Mediciones diarias de la estación meteorológica:
“station_max_temp_c”, “station_min_temp_c”, “station_avg_temp_c”, “station_precip_mm”, “station_diur_temp_rng_c”.

Mediciones de precipitación por satélite:
“precipitation_amt_mm”.

Mediciones del sistema de pronóstico climático:
“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”

Vegetación satelital:

"ndvi_se", "ndvi_sw", "ndvi_ne", "ndvi_nw".

Se realizan diferentes consultas en términos de resúmenes estadísticos, conteo de nulos, entre otros que permitan comprender las variables. Un resultado gráfico de la exploración de valores atípicos se presenta en la figura 4.

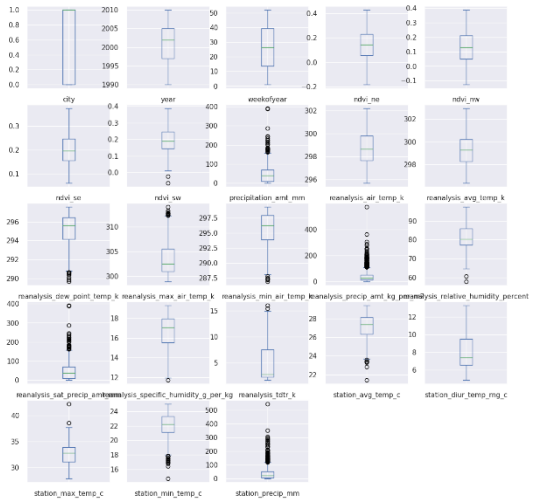

Fig. 4. *Exploración de valores atípicos*
Fuente: *elaboración propia.*

3.3 Preparación de los datos

Un procedimiento de modelado ampliamente empleado en la creación de aplicaciones de aprendizaje automático es el proceso estándar intersectorial para la minería de datos (CRISP-DM) [22]. Siguiendo la ruta incremental del proceso corresponde la preparación de los datos, en esta fase se hace optimización de datos.

Se realiza eliminación de valores nulo y ajuste de valores atípicos por proceso de corrección de outlier superiores e inferiores.

3.4 Modelado

3.4.1. Clustering Método KMEANS

La finalidad principal de K-Means consiste en particionar un conjunto de datos específico en K clústeres (siendo K un hiperparámetro) y ofrecer el centroide para cada muestra de datos [23].

Se aplica el método KMEANS, un resumen de las principales técnicas e hiperparámetros utilizados se presenta en la Tabla 1.

Tabla 1: *Para KMEANS ajuste de* hiperparámetros

Hiperparámetros	Valores/Configuración
n_clusters	3
init	'k-means++'
max_iter	300
n_init	10

Fuente: *elaboración propia.*

En cuanto a las técnicas, se aplica reducción de dimensionalidad con PCA (Análisis de Componentes Principales), se emplea PCA con n_components establecido en 2 para reducir la dimensionalidad de los datos a dos dimensiones, permitiendo su visualización en un plano bidimensional. El conjunto de datos original (df_kmeans) y los centroides se transforman usando el modelo PCA entrenado.

Se genera el grafico para análisis de resultados, la gráfica obtenida se visualiza en la figura 5.

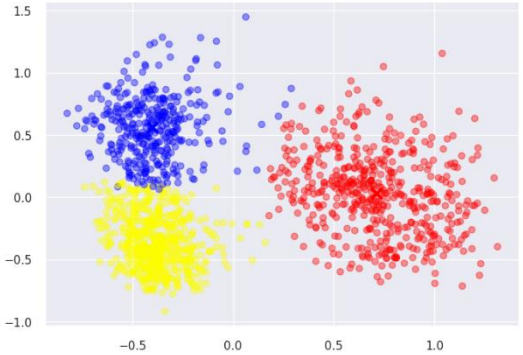

Fig. 5. *Resultado de Clustering por Método KMEANS.*
Fuente: *elaboración propia.*

3.4.2. Método Jerárquico Aglomerativo

El Método Jerárquico es una técnica de agrupamiento que organiza los datos en una estructura jerárquica, representada como un árbol o dendrograma.

Para crear un dendrograma con respecto a los datos del dengue se utiliza la función "linkage" se para calcular las distancias entre los pares de puntos y determinar cómo se deben fusionar los clústeres. En este caso, se utiliza el método de enlace 'ward', que minimiza la varianza intraclúster al fusionar los clústeres. El resultado representa la jerarquía de agrupamiento y permitiendo identificar la estructura de clústeres a diferentes niveles de similitud. La figura resultante muestra la relación entre los puntos y cómo se agrupan en clústeres, ver figura 6.

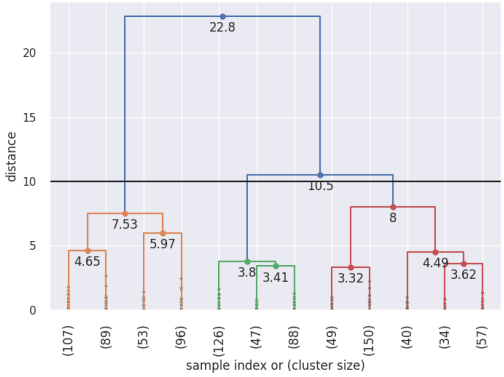

Fig. 6. *Hierarchical Clustering Dendrogram.*
Fuente: *elaboración propia.*

3.5 Evaluación

En la fase de evaluación, los resultados se verifican en términos de alcance de los objetivos definidos. Por lo tanto, es necesario interpretar los resultados y definir acciones adicionales. Además, el proceso debe ser revisado en general [24].

Al analizar el dendrograma, se observa la identificación de aproximadamente 5 grupos claramente diferenciados, marcando este número en la posición donde la distancia vertical es máxima (10). Este resultado sugiere la presencia de estructuras distintas en los datos del dengue, permitiendo una primera aproximación a la identificación de grupos significativos.

3.6 Implementación

Esta última fase se encarga de la implementación de los resultados del proyecto de minería de datos para asegurar su disponibilidad y cumplir con las necesidades de los usuarios finales. Sin embargo, CRISP-DM no detalla las especificaciones de los requisitos de implementación. Por esta razón, los profesionales adaptan el proceso de referencia, para obtener medir el cumplimiento de la solución implementada y su uso final. Es importante señalar que los proyectos basados en modelos no supervisados pueden no necesitar despliegue, ya que su objetivo es descubrir características e interpretarlas en el contexto de un problema específico [25].

Dado que se trata de un proyecto de aprendizaje automático no supervisado, no se requiere despliegue ni implementación adicional. Los resultados se obtienen mediante la aplicación de los algoritmos seleccionados, destacando que el Método Jerárquico Aglomerativo mostró un rendimiento superior en la identificación de grupos.

4. CONCLUSIONES

Se demuestra la eficacia y la utilidad del aprendizaje no supervisado, en particular, de las técnicas de clustering, en la exploración de datos epidemiológicos y ambientales relacionados con la propagación del dengue en San Juan, Puerto Rico, e Iquitos, Perú. Los resultados obtenidos subrayan la importancia de entender los patrones y las relaciones ocultas en los datos, lo que puede proporcionar conocimientos críticos para la prevención y el control del dengue en estas regiones.

Un aspecto fundamental que se ha resaltado a lo largo de este estudio es la elección adecuada de la técnica de clusterización. En el desarrollo del análisis, se observó que para este caso particular el método jerárquico funcionó bastante bien, produciendo clústeres de tamaño similar y adaptándose adecuadamente a los datos, mientras que el uso de métodos basados en densidad mostró un rendimiento deficiente en la identificación de grupos. Estos resultados destacan la necesidad de seleccionar con cuidado la técnica de clusterización más adecuada según la naturaleza de los datos y los objetivos del análisis.

Además, se ha demostrado la importancia de considerar la escala de las variables utilizadas en el análisis, ya que algunas técnicas son sensibles a las diferencias en la escala de las variables. La normalización previa de los datos se revela como un paso esencial para garantizar resultados coherentes y confiables en el proceso de clusterización.

Los resultados de este estudio ofrecen información valiosos para los interesados en este campo o profesionales que trabajen con estos datos, empleando algoritmos supervisados en futuros análisis.

Se recomienda seguir explorando y aplicando técnicas de aprendizaje no supervisado en la investigación epidemiológica y ambiental, con el fin de continuar descubriendo patrones ocultos y relaciones de causalidad que ayuden a abordar desafíos de salud pública de manera más eficiente y efectiva.

REFERENCIAS

[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 y Big Data: ciencia de los Datos aplicada al mundo de los negocios*. RIL editores, 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.