VICTEC. Revista Académica y Científica
https://doi.org/10.61395/victec.v4i7.109
pp.95 - 104
VICTEC Vicente Tecnológico
ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

Análisis de Vulnerabilidades de Ciberseguridad Mediante Técnicas de Ciencia de Datos

Analysis of Cybersecurity Vulnerabilities Using Data Science Techniques

Autores:
Suanny Tigselema-Egre [1]
Ricardo Villarroel-Molina [2]
Javier Guaña-Moya [3]
Wilson Iván Sánchez Paredes [4]

RESUMEN

El análisis de ciberseguridad utilizando ciencia de datos y aprendizaje automático desempeña un papel crucial en la era digital actual, donde la seguridad de la información se ha vuelto imperativa para las organizaciones. Este estudio se enfoca en la intersección entre ciberseguridad y ciencia de datos, utilizando métodos de aprendizaje automático y análisis de texto para comprender y fortalecer las defensas contra amenazas. Se destaca la importancia del TF-IDF (Frecuencia de Términos-Inversa de Documentos) como una herramienta para evaluar la relevancia de los términos en documentos y su aplicación en la clasificación de vulnerabilidades. El clasificador Multinomial Naive Bayes se presenta como una herramienta eficiente en la clasificación de texto, calculando probabilidades de pertenencia a clases específicas basadas en la frecuencia de términos. Se detallan las fórmulas esenciales utilizadas en este algoritmo, como la probabilidad condicional y la distribución multinomial. La metodología KDD (Knowledge Discovery in Databases) guía el proceso, desde la recopilación de datos en plataformas como Kaggle hasta la selección, limpieza y transformación de datos. El uso de `TfidfVectorizer` facilita la discretización de datos de texto, y el método `GridSearchCV` optimiza los hiperparámetros del modelo, alcanzando una

Recibido: 05/09/2023 **Aceptado:** 30/01/2023 **Publicado:** 17/02/2024

[1] Universidad Técnica Estatal de Quevedo, Email: suanny.tigselema2016@uteq.edu.ec ORCID: https://orcid.org/0009-0000-8887-7330
[2] Universidad Técnica Estatal de Quevedo, Email: rvillarroelm@uteq.edu.ec ORCID: https://orcid.org/0000-0002-6171-9815
[3] Instituto Superior Tecnológico Japón, Email: eguana@itsjapon.edu.ec ORCID: https://orcid.org/0000-0003-4296-0299
[4] Universidad Técnica de Ambato, Email: wilsons92@live.com ORCID: https://orcid.org/0009-0009-2379-4548

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

exactitud del 97.36%. Finalmente, La matriz de confusión revela un buen rendimiento general, aunque se identifican áreas de mejora, especialmente en la clase 'High'.

Palabras claves: ciberseguridad; naive bayes; vulnerabilidades.

ABSTRACT

Cybersecurity analysis using data science and machine learning plays a crucial role in today's digital era, where information security has become imperative for organizations. This study focuses on the intersection between cybersecurity and data science, using machine learning methods and text analytics to understand and strengthen defenses against threats. The importance of TF-IDF (Term-Inverse Document Frequency) is highlighted as a tool to evaluate the relevance of terms in documents and its application in vulnerability classification. The Multinomial Naive Bayes classifier is presented as an efficient tool in text classification, calculating probabilities of belonging to specific classes based on the frequency of terms. The essential formulas used in this algorithm, such as conditional probability and multinomial distribution, are detailed. The KDD (Knowledge Discovery in Databases) methodology guides the process, from data collection on platforms like Kaggle to data selection, cleaning and transformation. The use of `TfidfVectorizer` facilitates the discretization of text data, and the `GridSearchCV` method optimizes the model's hyperparameters, achieving an accuracy of 97.36%. Finally, the confusion matrix reveals good overall performance, although areas for improvement are identified, especially in the 'High' class.

Keywords: cybersecurity; naïve bayes; vulnerabilities.

INTRODUCCIÓN

La seguridad de la información se ha vuelto crucial para las organizaciones en un mundo cada vez más interconectado y digitalizado [1]. Los objetivos primordiales de la ciberseguridad abarcan desde la prevención y detección de intrusiones hasta la identificación de amenazas como el malware y prevención del fraude. La ciencia de datos aplicada en el dominio de la ciberseguridad emplea el aprendizaje de máquina para prevenir, identificar y remediar posibles amenazas a la ciberseguridad [2]. La intersección entre la ciberseguridad y la ciencia de datos ofrece una oportunidad única para comprender el panorama actual de amenazas, anticipar y fortalecer las defensas contra futuros ataques. A través de la combinación de métodos estadísticos, aprendizaje automático y análisis de texto, este análisis busca proporcionar una visión integral de las vulnerabilidades, desde su

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

descubrimiento hasta su resolución. La capacidad de prever posibles puntos de vulnerabilidad, comprender la gravedad de las amenazas y optimizar los tiempos de respuesta son objetivos clave que se abordarán en este estudio, con el objetivo de fortalecer la postura de seguridad cibernética en entornos cada vez más dinámicos y desafiantes.

REVISIÓN DE LITERATURA

El TF (Frecuencia de Términos) y el IDF (Frecuencia de Documentos Inversa) son conceptos utilizados en el procesamiento de lenguaje natural y la recuperación de información [3]. Para entender estos términos independientemente se tiene que: El TF es una medida que indica la frecuencia con la que un término específico aparece en un documento. Se calcula dividiendo el número de veces que aparece un término en un documento por el número total de términos en ese documento. El TF es útil para determinar la importancia de un término dentro de un documento específico. Y que el IDF es una medida que indica la importancia de un término en un conjunto de documentos. Se calcula dividiendo el número total de documentos en el conjunto por el número de documentos que contienen el término específico. El IDF es útil para determinar la relevancia de un término en todo el conjunto de documentos.

TF-IDF es un algoritmo que combina TF e IDF para asignar pesos a los términos en un documento, destacando su importancia relativa en un conjunto de documentos. Esta técnica, ampliamente utilizada en recuperación de información y clasificación de documentos, pesa las palabras en función de su relevancia. La fórmula básica consiste en multiplicar la frecuencia de término (TF) por la frecuencia inversa de documento (IDF).

Multinomial Naive Bayes

El clasificador Multinomial Naive Bayes (MultinomialNB) es un algoritmo utilizado en el procesamiento de lenguaje natural y la clasificación de documentos. Se basa en el teorema de Bayes y se utiliza principalmente en tareas de clasificación de texto. El algoritmo calcula la probabilidad de cada etiqueta para una muestra dada y luego selecciona la etiqueta con la probabilidad más alta como resultado. El MultinomialNB asume que cada característica clasificada no está relacionada con ninguna otra característica, lo que significa que la presencia o ausencia de una característica no afecta la presencia o ausencia de otra característica. Este clasificador es ampliamente utilizado debido a su eficiencia computacional y facilidad de implementación [4].

Fórmulas utilizadas en el clasificador MultinomialNB [5];

Probabilidad condicional

- $P(c|d) = P(c) * P(d|c) / P(d)$

Donde:

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

- P(c|d) es la probabilidad de que el documento d pertenezca a la clase c.
- P(c) es la probabilidad a priori de la clase c.
- P(d|c) es la probabilidad condicional de que el documento d pertenezca a la clase c.
- P(d) es la probabilidad marginal del documento d.

Calcula la probabilidad de que un documento pertenezca a una clase específica dada la presencia de ciertas características en el documento. Es fundamental para la clasificación de texto utilizando el algoritmo MultinomialNB.

Distribución multinomial:

- $P(d|c) = \Pi |V| \; t=1 \; P(wt|c)xt$

Donde:

- P(d|c) es la probabilidad de que el documento d pertenezca a la clase c.
- |V| es el tamaño del vocabulario utilizado para entrenar el modelo.
- P(wt|c) es la probabilidad de que el término wt aparezca en un documento de la clase c.
- xt es la frecuencia del término wt en el documento d.

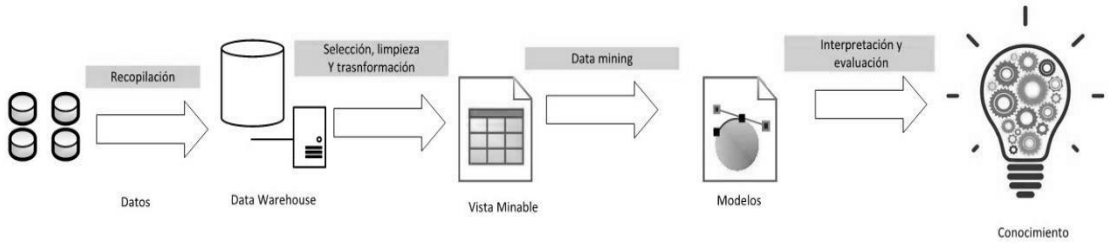
Se utiliza para calcular la probabilidad condicional de un documento dada una clase, considerando la frecuencia de aparición de cada término en el documento y en la clase. Es fundamental para modelar la distribución de palabras en un documento y su relación con una clase específica en el contexto de la clasificación de texto con MultinomialNB.

METODOLOGÍA

Como metodología se eligió KDD (Knowledge Discovery in Databases), basada en un bien definido proceso KDD de múltiples pasos (Fig. 1), para el descubrimiento de conocimiento en grandes coleccione de datos [6], [7].

Figura 1

Etapas KDD



Recopilación

La recopilación de datos se llevó a cabo mediante una exhaustiva búsqueda en plataformas de datos abiertos, siendo Kaggle la principal fuente utilizada bajo la temática "cybersecurity". El conjunto de datos seleccionado para el análisis es "Cybersecurity Risk (2022 CISA Vulnerability)", el cual ofrece una visión de

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

las vulnerabilidades reportada por Cybersecurity and Infrastructure Security Agency (CISA) durante el año 2022 [8]. Este conjunto de datos ha sido elegido debido a su amplitud, calidad y representatividad en cuanto a la información disponible sobre vulnerabilidades cibernéticas. Además, se ha validado la autenticidad y confiabilidad de los datos mediante revisiones detalladas de la fuente. A continuación, se describe las columnas presentes en el dataset:

Tabla 1

Descripción de las columnas del dataset.

Columna	Descripción	Tipo
Vendor_project	Nombre del proyecto del proveedor asociado a la vulnerabilidad	String
Product	El nombre del producto asociado a la	String
Vulnerability_name	El nombre de la vulnerabilidad.	String
Date_added	La fecha en que la vulnerabilidad se añadió al	Date
Short_description	Breve descripción de la vulnerabilidad.	String
Required_action	La acción requerida para remediar la	String
Due_date	La fecha en la que debe completarse la acción	Date
Notes	Notas adicionales sobre la vulnerabilidad.	String
Grp	El grupo asociado a la vulnerabilidad.	String
Pub_date	Fecha de publicación de la vulnerabilidad.	Date
Cvss	La puntuación de Common Vulnerability Scoring System asociada a la vulnerabilidad	String
Cwe	Enumeración de la debilidad común asociada a la vulnerabilidad	String
Vector	El vector asociado a la vulnerabilidad.	String
Complexity	La complejidad asociada a la vulnerabilidad.	String
Severity	El nivel de gravedad asociado a la	String

Selección, limpieza y transformación

El procesamiento de datos desempeña un papel crucial en la identificación de vulnerabilidades de ciberseguridad. La normalización, limpieza y transformación adecuada de datos permiten una interpretación más precisa de las amenazas, se tratan los valores faltantes o también conocidos como 'na', en este caso se los elimina, adicionalmente se aplica TfidfVectorizer para discretizar los valores de la variable 'short_description', esta técnica se utiliza para convertir datos de texto en vectores numéricos, asignando ponderaciones a las palabras en función de su frecuencia [7].

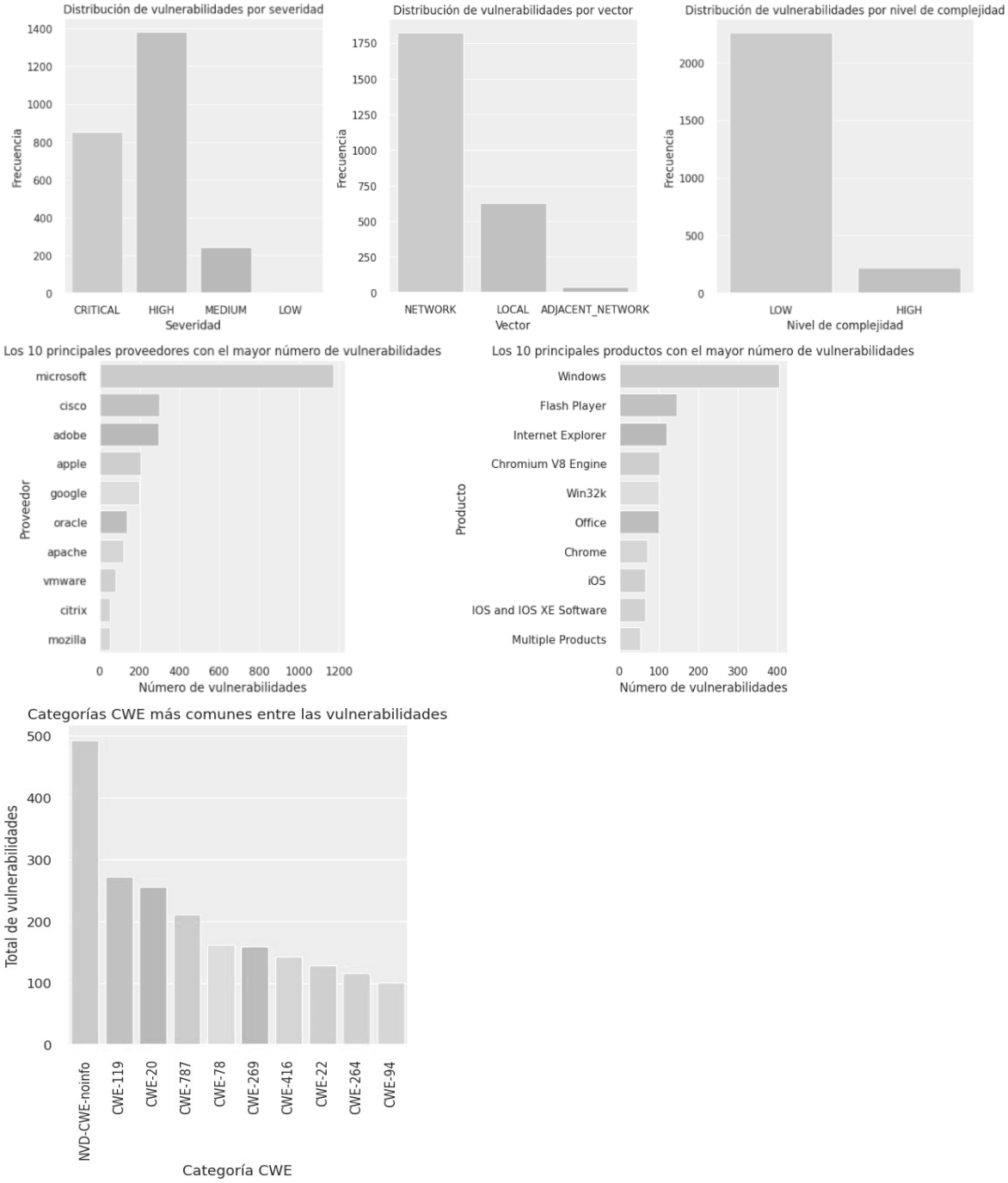
Además, el uso de algoritmos de aprendizaje automático facilita la detección temprana de patrones anómalos, contribuyendo así a la prevención proactiva de posibles ataques [6].

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

Minería de datos

Figura 2

Análisis descriptivo del dataset.



ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

Las gráficas en torno a la severidad, demuestran que existen mayor cantidad de vulnerabilidades con una severidad alta seguidos por vulnerabilidades de severidad critica, esto puede ser causado por que la Cybersecurity and Infrastructure Security Agency (CISA) se centra en las amenazas cibernéticas de mayor relevancia.

El medio de ataque que prevalece en los datos de las vulnerabilidades es el vector de redes las cuales pueden llevarse a cabo desde internet y el de menor ocurrencia es el de redes adyacentes las cuales únicamente pueden ser llevadas a cabo desde la misma red física o lógica.

La complejidad asociada a la vulnerabilidad nos dice que tantos conocimientos o recursos se necesitan para poder atacar a la vulnerabilidad y los datos demuestran que la mayor cantidad de vulnerabilidades tienen una complejidad baja.

Los proveedores de productos o servicios informáticos pueden ser objeto de ataques ya sea por fines económicos o de relevancia geopolítica, en este sentido podemos evidenciar que Microsoft, Cisco, Adobe, Apple y Google son los proveedores con mayor número de vulnerabilidades siendo todos estos con base en Estados Unidos. En este sentido los productos con mayores vulnerabilidades reportadas por la CISA son el sistema operativo Windows de Microsoft, el software de Flash Player de Adobe y las tecnologías de navegadores de internet de Internet Explorer y el motor de JavaScript y Web Asamblea V8.

Las categorías de las vulnerabilidades CWE es utilizado para ayudar a los desarrolladores de productos informáticos a evitar los errores más comunes y categorizarlos siendo la segunda categoría más frecuente en las vulnerabilidades la de un manejo deficiente de la restricción inadecuada de operaciones dentro de los límites de un búfer de memoria (CWE-119) [9], [10].

RESULTADOS

La elección de este enfoque metodológico se fundamenta en su capacidad para procesar descripciones breves de vulnerabilidades, permitiendo una clasificación precisa y eficiente. A continuación, se detallarán los resultados obtenidos, destacando la relevancia y aplicabilidad de las técnicas de preprocesamiento y clasificación de la severidad de las vulnerabilidades y comprensión de las vulnerabilidades reportadas en el ámbito de la ciberseguridad. Para el ajuste de los hiperparámetros del modelo se utilizó el método GridSearchCV, encuentra la combinación óptima de hiperparámetros de tal manera que maximice la exactitud del modelo.

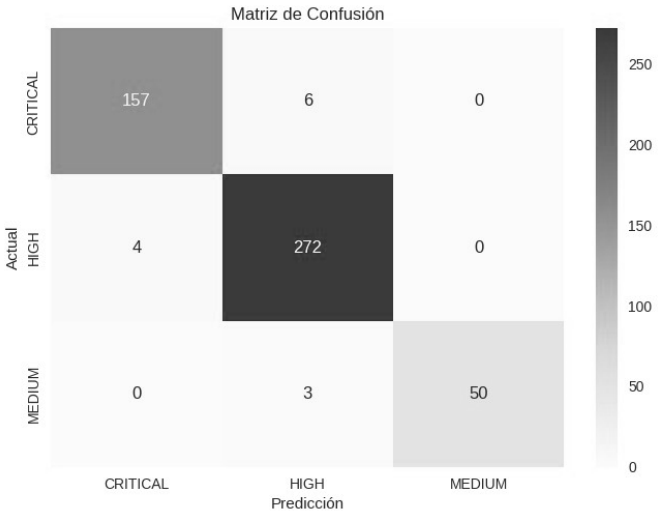
El modelo obtenido proporciona una exactitud de 97,36% lo que sugiere que tiene un buen rendimiento general. La matriz de confusión (Fig. 3) proporciona detalles sobre cómo el modelo se desempeña en cada clase. Se observa que

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

el modelo tiene dificultades con la clase 'High' ya que da falsos positivos con las clases 'Medium' y 'Critical', sin embargo, funciona bien para las clases 'Medium' y 'Critical'.

Figura 3

Matriz de confusión resultado del test con el modelo construido basado en el algoritmo Multinomial Naive Bayes.



DISCUSIÓN

El tema de la influencia de los estereotipos de género en la construcción de la identidad del niño es de gran importancia y ha sido abordado en diversas investigaciones y estudios, como los presentados en la bibliografía proporcionada. Las entrevistas realizadas a educadores, padres y expertos en el tema son fundamentales para complementar el análisis de estos estudios y obtener una perspectiva más amplia y detallada de la problemática.

En virtud de lo expuesto, es importante destacar que los estudios en la bibliografía se enfocan en diferentes aspectos de la influencia de los estereotipos de género en la construcción de la identidad del niño, como la influencia de los cuentos clásicos en la perpetuación de estereotipos de género (Asensi, 2019), la influencia de los medios de comunicación y la televisión en la formación de la identidad de género (Avalos, 2009), la influencia de los estereotipos de género en el ambiente escolar sobre el desarrollo psicosocial de los niños y niñas (Balseca & Nataly, 2020), la construcción de la identidad personal, sexual y de género durante el juego en la infancia (Caballero, 2021), y la influencia de los roles y estereotipos de género en la publicidad infantil (Del Rosario & Vargas, 2018), entre otros.

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

En primer lugar, las entrevistas realizadas a educadores, padres y expertos en el tema permiten obtener una visión más amplia de cómo los estereotipos de género se reflejan en diferentes ámbitos de la vida del niño, y cómo afectan a su identidad de género y a su desarrollo en general. Además, estas entrevistas pueden proporcionar información valiosa sobre las estrategias y herramientas que se utilizan para abordar y modificar estos estereotipos en diferentes contextos, como el aula, el hogar y la sociedad en general.

Por ejemplo, las entrevistas pueden permitir conocer cómo los educadores están trabajando para fomentar la igualdad de género en el aula y cómo están abordando los estereotipos de género en su práctica educativa. Igualmente pueden permitir conocer la opinión de los padres sobre la influencia de los estereotipos de género en la identidad de género de sus hijos, así como las estrategias que utilizan para fomentar la igualdad de género en el hogar.

En segundo lugar, las entrevistas a expertos en el tema pueden proporcionar información valiosa sobre las tendencias actuales en la investigación sobre la influencia de los estereotipos de género en la construcción de la identidad del niño, así como sobre las herramientas y estrategias más efectivas para abordar y modificar estos estereotipos.

En síntesis, las entrevistas realizadas a educadores, padres y expertos en el tema son fundamentales para obtener una visión más amplia y detallada de la problemática de los estereotipos de género en la construcción de la identidad del niño, así como para conocer las estrategias y herramientas más efectivas para abordar y modificar estos estereotipos en diferentes ámbitos de la vida del niño.

CONCLUSIONES

La aplicación del algoritmo Multinomial Naive Bayes permitió una evaluación detallada del desempeño del modelo en diferentes clases, proporcionando una visión integral de su capacidad predictiva. La alta exactitud obtenida, del 97.36%, sugiere que el modelo es eficaz en la clasificación de las vulnerabilidades en las categorías CRITICAL, HIGH y MEDIUM. Sin embargo, la matriz de confusión reveló la presencia de algunos falsos positivos y falsos negativos, indicando áreas donde el modelo puede ser mejorado.

REFERENCIAS BIBLIOGRÁFICAS

[1]. I. V. Peña, «Gestión de Riesgos en Ciberseguridad» p. 3, 2023.

[2]. I. M. d. Diego y A. Ferández Isabel, Ciencia de datos para la

ISSN 2737-6214 - Vol 5 - Núm 8 (Enero - Junio 2024)

ciberseguridad, Madrid: RA-MA Editorial, 2020.

[3]. G. A. Dalaorao, A. M. Sison y R. P. Medina, «Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy» EEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA), 2019.

[4]. S. Xu, Y. Li y W. Zheng, «Bayesian Multinomial Naïve Bayes Classifier to Text Classification» Lecture Notes in Electrical Engineering, 2017.

[5]. E. Anguiano-Hernández, «Naive Bayes Multinomial para clasificación de texto usando un esquema de pesado por clases,» 2009. [En línea].

[6]. I. H. Witten, E. Frank, M. A. Hall y C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems), New Zealand: Morgan Kaufmann, 2016.

[7]. J. L. Dias, M. K. Sott, C. C. Ferrão, J. C. Furtado, and J. A. R. Moraes, «Data mining and knowledge discovery in databases for urban solid waste management: A scientific literature review,» Waste Management & Research, vol. 39, no. 11, pp. 1331-1340, 2021

[8]. A. R. Bruce, «Cyber security during international conflict,» Tech. Rep., 2022.

[9]. R. Esparza Tortosa, «Análisis y corrección de vulnerabilidades de un producto software con SonarQube,» 2023.

[10]. M. E. de Vega Martín, «Metodología de benchmark de herramientas SAST,» 2023.