INGENIERÍA Y COMPETITIVIDAD
REVISTA CIENTÍFICA Y TECNOLÓGICA

Publicación anticipada

El comité editorial de la revista ***Ingeniería y Competitividad*** informa que este artículo fue aprobado para publicación en el volumen 25 número 1, teniendo en cuenta los requisitos editoriales y los conceptos emitidos por los pares evaluadores. Por lo tanto, se publica anticipadamente para su consulta, descarga y citación provisional, aclarando que esta puede diferir de la versión final, ya que no ha completado las etapas finales del proceso editorial (corrección de estilo, traducción y diagramación) y solo los títulos, datos de autores, palabras clave y resúmenes corresponden a la versión final del artículo.

Como citar:

Chanchí-Golondrino GE, Ospina-Alarcón MA, Muñoz-Sanabria LF. Caracterización de delitos cibernéticos del departamento de Cundinamarca durante el primer semestre de 2021 mediante análisis exploratorio y aprendizaje de máquina. *INGENIERÍA Y COMPETITIVIDAD*, **In press 2023**; e20511760. https://doi.org/10.25100/iyc.v25i1.11760.

Article in press

The editorial committee of the ***Ingeniería y Competitividad*** Journal informs that this article was approved for publication, in volume 25 number 1, considering the editorial requirements and the concepts of the peer reviewers. Therefore, the preliminary version of this article is published for consultation, download and provisional citation purposes, clarifying that this version may differ from the final document, since it has not completed the final stages of the editorial process (proof-editing, translation and layout) and only the titles, authorship, keywords and abstracts will remain unchanged the final version of the article.

How to cite:

Chanchí-Golondrino GE, Ospina-Alarcón MA, Muñoz-Sanabria LF. Characterization of cybercrime in the department of Cundinamarca during the first half of 2021 through exploratory analysis and machine learning. *INGENIERÍA Y COMPETITIVIDAD*, **In press 2023**; e20511760. https://doi.org/10.25100/iyc.v25i1.11760

SYSTEMS ENGINEERING

Characterization of cybercrime in the department of Cundinamarca during the first half of 2021 through exploratory analysis and machine learning.

INGENIERIA DE SISTEMAS

Caracterización de delitos cibernéticos del departamento de Cundinamarca durante el primer semestre de 2021 mediante análisis exploratorio y aprendizaje de máquina

Gabriel Elías Chanchí-Golondrino[1]§, Manuel Alejandro Ospina-Alarcón [1]*, Luis Freddy Muñoz-Sanabria[2]

[1]Universidad de Cartagena, Facultad de Ingeniería, Ingeniería de Sistemas, Cartagena, Colombia

[2]Fundación Universitaria de Popayán, Facultad de Ingeniería, Ingeniería de Sistemas, Popayán, Colombia

gchanchig@unicartagena.edu.co, mospinaa@unicartagena.edu.co, lfreddyms@fup.edu.co

Recibido: 22 de noviembre de 2021 – **Aceptado**: 12 de septiembre de 2022

Abstract

Taking into account the wide diffusion that data analytics has had in different application areas and considering the scarcity of specific datasets associated with cybercrime within open data strategies in Colombia, this article aims to characterize cybercrime in the department of Cundinamarca, through the use of exploratory analysis and machine learning techniques. The present research was developed through 4 methodological phases: data adequacy, exploratory data analysis, application of machine learning models and finally generation of value-added information. For the development of the proposed study, a dataset was formed from the dataset of 35,000 records published by the National Police in the open data portal of Colombia, which addresses high-impact crimes within the department of Cundinamarca and occurred during the first half of 2021. The cybercrime dataset has a total of 1513 records and includes attributes such as: day, quarter, municipality, area, victim, age and crime, so that at the exploratory analysis level, descriptive statistics methods were applied on the different attributes, while at the machine learning level, the association rules and clustering models were applied in order to determine respectively the relationship of the attributes

with the type of crime, and the representative groups formed by relating the age with the type of crime and the municipality with the type of crime. The study developed allowed to demonstrate the usefulness and potential of data analytics techniques in the field of cybersecurity, in order to support decision making by the relevant authorities.

***Keywords**: Cybersecurity, cybercrime, exploratory analysis, machine learning.*

Resumen

Teniendo en cuenta la amplia difusión que ha tenido la analítica de datos en diferentes ámbitos de aplicación y considerando la escasez de *datasets* específicos asociados a los delitos informáticos dentro de las estrategias de datos abiertos en Colombia, este artículo tiene como objetivo realizar la caracterización de los delitos informáticos del departamento de Cundinamarca, mediante el uso de técnicas de análisis exploratorio y machine learning. La presente investigación fue desarrollada mediante 4 fases metodológicas: adecuación de los datos, análisis exploratorio de los datos, aplicación de modelos de machine learning y finalmente generación de información de valor agregado. Para el desarrollo del estudio propuesto, se conformó un conjunto de datos a partir del *dataset* de 35000 registros publicado por la Policía Nacional en el portal de datos abiertos de Colombia, el cual aborda los delitos de alto impacto dentro del departamento de Cundinamarca y ocurridos durante el primer semestre de 2021. El *dataset* de delitos cibernéticos conformado cuenta con un total de 1513 registros e incluye atributos tales como: día, trimestre, municipio, zona, víctima, edad y delito, de tal modo que a nivel del análisis exploratorio se aplicaron métodos de estadística descriptiva sobre los diferentes atributos, mientras que a nivel de machine learning se hizo uso de los modelos de reglas de asociación y clustering con el fin de determinar de manera respectiva la relación de los atributos con el tipo de delito, y los grupos representativos que se forman al relacionar la edad con el tipo de delito y el municipio con el tipo de delito. El estudio desarrollado permitió demostrar la utilidad y potencialidad que tienen las técnicas de analítica de datos en el campo de la ciberseguridad, de cara a apoyar la toma de decisiones por parte de las autoridades pertinentes.

***Palabras clave**: Análisis exploratorio, Aprendizaje de máquinas, ciberseguridad, delito cibernético.*

1. Introducción

En la actualidad, los sistemas de información, la internet y la computación en la nube dan soporte al almacenamiento, gestión y uso de información de carácter personal y organizacional, por lo que se convierten en blanco para quienes desean robar, manipular o afectar a los propietarios de la información [(1–4)].

De acuerdo a lo anterior, los gobiernos y organismos de seguridad reconocen que, dadas las aplicaciones y servicios disponibles en la nube en la actualidad, cada vez existe un mayor riesgo de vulneración a la seguridad a través de los delitos informáticos, ciberterrorismo y las amenazas cibernéticas, por lo que diversas organizaciones a nivel mundial han elevado sus capacidades tecnológicas de seguridad de la información y ciberdefensa para contrarrestar los ciberataques [(5–7)]. En ese mismo sentido, a nivel de Colombia, la Encuesta Nacional de Seguridad Informática de 2022, realizada a 14 sectores diferentes de la economía, reportó como incidentes más representativos a nivel de seguridad informática: los errores humanos con un 38%, los incidentes de phishing con un 32% y los ataques de ingeniería social con un 25% [(8)]. Así mismo, según el reporte de 2020 publicado por la Agencia Federal de Investigación (FBI), los cinco incidentes de ciberseguridad con mayor ocurrencia durante ese mismo año fueron: phishing, no pago-no entrega, extorsión, violación de datos personales y robo de identidad [(9)].

Los delitos cibernéticos, son entendidos como actos ilícitos cometidos a través del uso inadecuado de la tecnología, atentando contra la privacidad de la información de terceras personas, dañando o extrayendo cualquier tipo de datos que se encuentren almacenados en servidores [(10–12)]. Los delincuentes cibernéticos realizan incursiones fraudulentas cada vez más frecuentes y diversas, como el acceso sin autorización a sistemas de información, piratería informática,

fraude financiero, sabotaje informático y pornografía infantil, entre otros. Con el fin de detectar y contrarrestar los delitos cibernéticos, varios países han dispuesto un sistema judicial especializado que permite procesar y castigar dichos delitos [(4)].

En este mismo sentido, con la amplia difusión de la analítica de datos en diferentes ámbitos, es posible su aprovechamiento en la identificación y caracterización de los delitos cibernéticos, de cara a la toma de decisiones por parte de las autoridades pertinentes [(13,14)]. En el contexto colombiano y en el caso específico de la estrategia de datos abiertos, se ha evidenciado la insuficiencia de *datasets* específicos que permitan la identificación y caracterización de este tipo de delitos, por lo que conviene desarrollar estudios que motiven y promuevan el aprovechamiento de la ciencia de datos en el campo de la seguridad.

En este artículo se presenta como contribución el desarrollo de un estudio exploratorio y basado en machine learning para la caracterización de los delitos cibernéticos en el departamento de Cundinamarca durante el primer semestre de 2021. El estudio fue realizado a partir de un subconjunto de datos del dataset de delitos de alto riesgo en el departamento de Cundinamarca publicado por la Policía Nacional en el portal de datos abiertos de Colombia [(15)]. A nivel del estudio exploratorio se realizó un análisis basado en estadística descriptiva sobre los diferentes atributos del conjunto de datos seleccionado. En cuanto al estudio basado en machine learning, se aplicaron modelos basados en reglas de asociación, así como de agrupamiento sobre el conjunto de datos escogido. Para el desarrollo del estudio se hizo uso de la herramienta libre de minería de datos *weka*, la cual permite tanto la aplicación de métodos de estadística descriptiva, como la aplicación de modelos de aprendizaje supervisado y no supervisado [(16–18)]. Del mismo modo se hizo uso de la herramienta libre *GeoDa*, la cual posibilita el análisis espacial de los datos y la aplicación de modelos de clustering sobre los mismos [(19–21)]. Los resultados obtenidos a través del estudio pretender servir de apoyo para la toma de decisiones por parte de las autoridades judiciales, así como servir de base para otros estudios relacionados con delitos no cibernéticos. Esto teniendo en cuenta que el análisis exploratorio realizado y las técnicas de machine learning aplicadas (reglas de inferencia mediante el algoritmo predictive apriori), permiten obtener información de valor agregado a partir del *dataset* estudiado, tal como los días en que se cometen más delitos cibernéticos, el número de delitos por trimestre, el porcentaje de delitos en zonas urbanas y rurales, los municipios en donde se cometieron el mayor número de delitos, los delitos más recurrentes, entre otros. Así mismo, este estudio tiene como ventaja el aprovechamiento de las herramientas de software libre *GeoDa* y *weka*, las cuales al tener licencia GPL tienen como ventaja la posibilidad de ser ejecutadas, modificadas, redistribuidas y redistribuidas con modificaciones sin ninguna restricción. A partir de estas herramientas fueron utilizados las reglas de asociación y clustering, teniendo en cuenta la naturaleza de los datos y la necesidad de conocer la distribución espacial de los datos entre algunos de los atributos del *dataset*.

El resto del artículo está organizado de la siguiente forma: en la sección 2 se presenta las fases metodológicas consideradas para el desarrollo de la presente investigación; en la sección 3 se describen los resultados obtenidos a partir de los análisis exploratorio y basados en machine learning para la caracterización de los delitos cibernéticos ocurridos en el departamento de Cundinamarca durante el primer semestre de 2021; en la sección 4 se presentan las conclusiones y trabajos futuros derivados de la presente investigación.

2. Metodología

Para el desarrollo de la presente investigación, se tuvieron en cuenta 4 fases metodológicas a saber: adecuación del *dataset* y limpieza de datos, análisis exploratorio de los datos, aplicación de modelos de machine learning y generación de información de valor agregado (ver Figura 1).

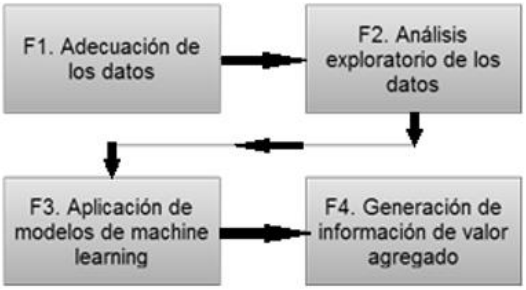

Figura 1. *Metodología considerada. Fuente: elaboración propia.*

En la fase 1 de la metodología, se obtuvo y adecuó el *dataset* de los delitos cibernéticos para el departamento de Cundinamarca, a partir de los *dataset* de 35000 registros con delitos de alto impacto para el mismo departamento, publicado en el portal de datos abiertos de Colombia por la Policía Nacional, del cual se filtraron en específico los correspondientes a la categoría de delitos cibernéticos. Esta adecuación incluyó la adaptación del formato de los datos al formato ARFF requerido por parte de la herramienta de minería de datos empleada, lo cual implicó la inclusión de los atributos en el encabezado del archivo y la separación por comas de los datos asociados a cada atributo dentro de cada registro del *dataset*. A partir del *dataset* conformado en la fase 1, se realizó el análisis exploratorio de los datos, en donde se aplicaron métodos de estadística descriptiva sobre los diferentes atributos del *dataset*, con el fin de determinar aspectos de interés tales como: los días de la semana en los cuales ocurren más delitos, los delitos con mayor frecuencia en el departamento, las edades de las víctimas de los delitos y los municipios en donde se presentan los delitos más frecuentes, entre otros. Lo anterior fue realizado a través del uso de la herramienta de minería de datos y de machine learning *weka*. Así mismo se realizó en esta fase un estudio espacial sobre la distribución de los delitos dentro del departamento, haciendo uso de la herramienta libre *GeoDa*, la cual calcula internamente la distancia entre los municipios haciendo uso de la información geoespacial provista por el *dataset*. Dentro de la fase 3 de la metodología, se aplicaron 2 modelos de machine learning: reglas de inferencia usando el algoritmo Predictive Apriori y clustering mediante el uso del algoritmo KMeans. A través de la aplicación del primer modelo, se obtuvieron un conjunto de reglas de asociación que relacionan los atributos del *dataset*, que pueden ser empleados para la toma de decisiones a partir del *dataset*. Del mismo modo, el segundo modelo permitió determinar los clusters y los centroides alrededor de los cuales se concentran los valores de los atributos del *dataset*. Finalmente, en la fase 4, se consolidan los resultados y se obtienen conclusiones a partir del análisis, que pretenden servir de aporte para la toma de decisiones por parte de las autoridades pertinentes con respecto a los delitos cibernéticos dentro del departamento de Cundinamarca.

3. Resultados y discusión

En esta sección se presentan los resultados tanto del estudio exploratorio, como del análisis basado en machine learning, sobre el *dataset* de delitos cibernéticos para el departamento de Cundinamarca, el cual cuenta con un total de 1513 registros y un total de 7 atributos (día, trimestre, municipio, zona, víctima, edad y delito), los cuales pueden apreciarse de manera más clara en la interfaz de pre-procesamiento de la herramienta de minería de datos *weka* (ver Figura 2), en la cual es posible realizar un análisis descriptivo sobre cada uno de los atributos del *dataset* considerado.

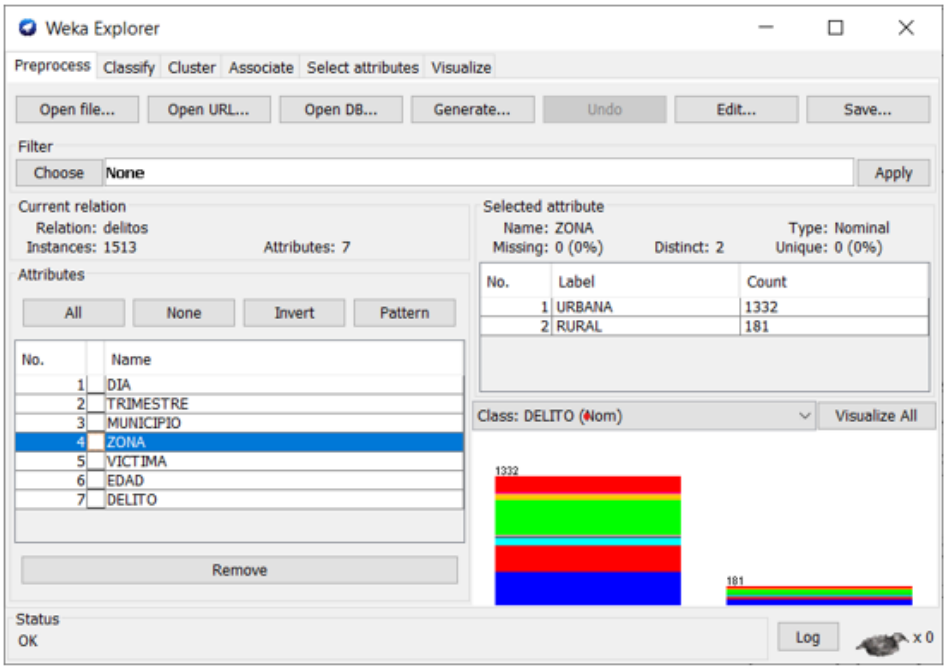

Figura 2. *Interfaz de pre-procesamiento de Weka. Fuente: elaboración Propia.*

A nivel del estudio exploratorio, es importante mencionar que de acuerdo a los análisis realizados sobre los diferentes atributos del *dataset* mediante la herramienta *Weka*, se pudo determinar que los días de la semana en los cuales se cometen con mayor frecuencia los delitos cibernéticos son el miércoles y el viernes, con una frecuencia respectiva de 260 y 256 ocurrencias respectivamente, mientras que los días en los cuales son cometidos con menor frecuencia dichos delitos son el domingo y el sábado con 134 y 177 ocurrencias respectivamente. Lo anterior puede apreciarse de manera más clara en la Figura 3, donde se presentan el número de delitos en cada uno de los días de la semana.

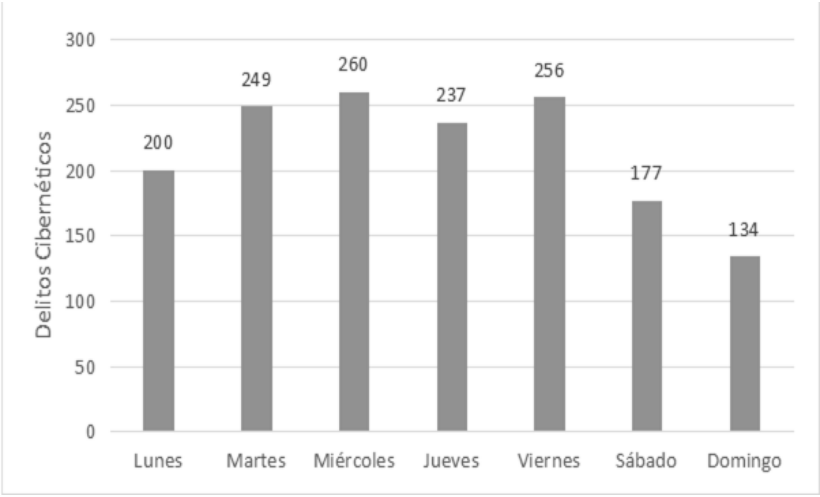

Figura 3. *Delitos cibernéticos por día. Fuente: elaboración propia.*

Continuando con el análisis exploratorio, se pudo determinar que en el primer trimestre del 2021 se cometieron un total de 848 delitos, mientras que en el segundo trimestre de 2021 se cometieron un total de 665 delitos cibernéticos, lo que equivale al 56.05% y al 43.95% de manera respectiva. Del mismo modo, con respecto a las zonas en donde ocurrieron los delitos, un total de 1332 fueron cometidos en zonas urbanas, mientras que 181 ocurrieron en zonas rurales del departamento. En ese orden de ideas, los municipios en donde se cometieron la mayor cantidad de delitos fueron Soacha y Chía, con un total de 366 y 139 ocurrencias de manera respectiva, lo que corresponde al 35.7% y al 13.6% de los delitos en los 10 municipios con mayor número de casos de cibercrimen. En este sentido, en la Figura 4 se presentan los 10 municipios con el mayor número de delitos cibernéticos durante los dos primeros trimestres de 2021, junto con el porcentaje de delitos dentro de los 10 municipios con mayor ocurrencia de casos de cibercrimen.

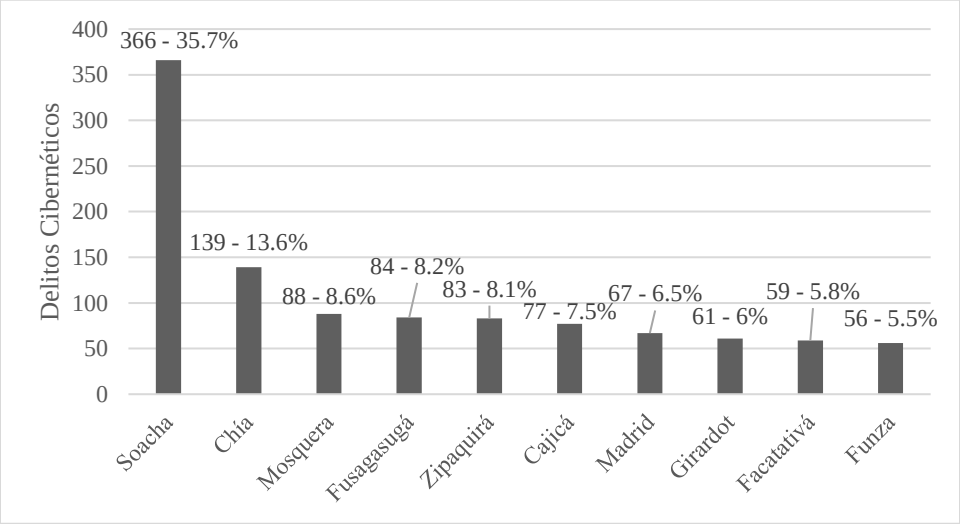

Figura 4*. Delitos cibernéticos por municipio. Fuente: elaboración propia.*

De manera complementaria a los resultados obtenidos anteriormente, mediante el uso de la herramienta libre *GeoDa* es posible visualizar la distribución de los 9 tipos de delitos cibernéticos del *dataset* (1. Violación de datos personales, 2. Acceso abusivo a un sistema informático, 3. Interceptación de datos informáticos, 4. Daño informático, 5. Uso de software malicioso, 6. Hurto por medios informáticos y semejantes, 7. Transferencia no consentida de activos, 8. Obstaculización ilegítima de sistema informático o red de telecomunicación y 9. Suplantación de sitios web para capturar datos personales), dentro del departamento de Cundinamarca, tal como se aprecia en la Figura 5. Cabe mencionar que el *dataset* muestra la distribución de los delitos en los diferentes municipios de Cundinamarca, pero no presenta el mapa con la división geográfica de los mismos.

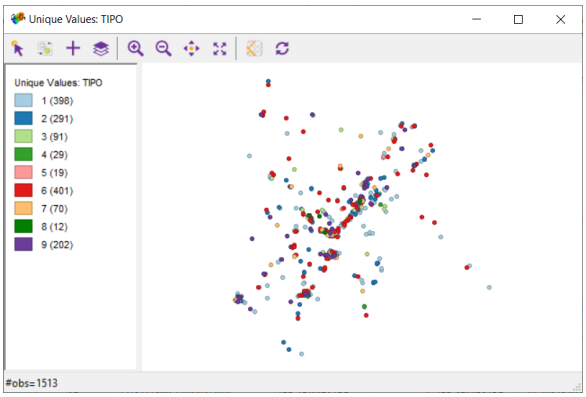

Figura 5. *Distribución de los delitos en Cundimarca. Fuente: elaboración propia.*

En lo referente a las víctimas de los delitos, 842 son mujeres (55.65%), 661 son hombres (43.69%) y 10 no han sido reportadas (0.66%). Del mismo modo con respecto a las edades de las víctimas de los delitos cibernéticos, cabe mencionar que estas oscilan entre los 5 y los 82 años. Así mismo, si se relacionan los 5 grupos etarios (infantes, niños, adolescentes, jóvenes, adultos y adultos mayores) con las víctimas de los delitos cibernéticos, se obtiene que el grupo etario que tiene el mayor número de víctimas es el de los adultos con 968 ocurrencias y el de los jóvenes con 400 ocurrencias. Lo anterior puede apreciarse de manera más clara en la Tabla 1.

A nivel del tipo de los delitos cometidos dentro del departamento de Cundinamarca, es posible mencionar que los delitos cibernéticos más recurrentes se encuentran en las tipologías: hurto por medios informáticos y violación de datos personales, cada uno de los cuales cuenta con 401 y 398 ocurrencias respectivamente. Del mismo modo, el delito que menos se presenta en el departamento de Cundinamarca fue el de obstaculización ilegítima de sistema informático o red de telecomunicación, con un total de 12 ocurrencias. Teniendo en cuenta lo anterior, en la Figura 6 se presenta el número total de incidencias asociadas a cada una de las 9 tipologías o tipos de delitos presentes en el *dataset*.

Tabla 1. *Delitos cibernéticos por grupo etario.*

Grupo Etario	Delitos Cibernéticos
Infantes	0
Niños	2
Adolescentes	12
Jóvenes	400
Adultos	968
Adultos Mayores	86
No reportados	45

Fuente: elaboración propia

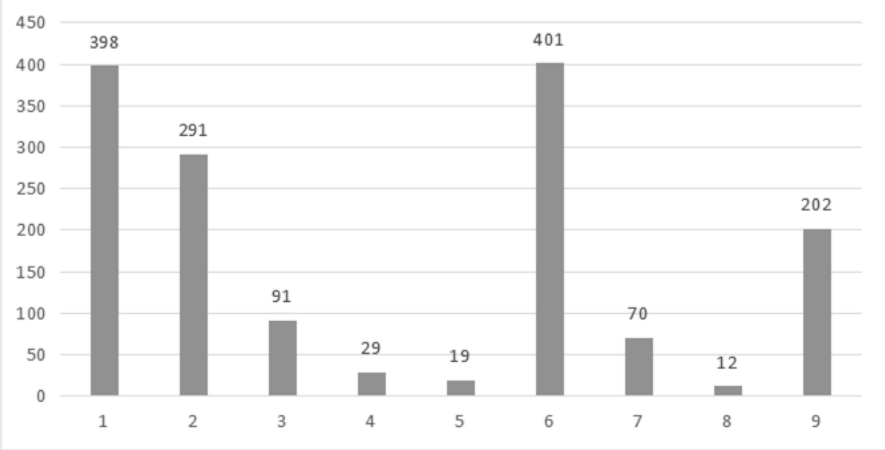

Figura 6. *Ocurrencia de los tipos de delitos cibernéticos. Fuente: elaboración propia.*

Una vez realizado el análisis exploratorio de los datos, se procedió con la aplicación de modelos de *machine learning* y de manera específica, los algoritmos de reglas de asociación (*PredictiveApriori*) y de agrupamiento (*KMeans*), los cuales no requieren el uso de validación cruzada, ni la discriminación de los datos en conjuntos de entrenamiento y prueba, dado que no

son métodos predictivos, como en el caso de los modelos de aprendizaje supervisado. En lo referente al algoritmo de reglas de asociación, se hizo uso de las ventajas provistas por la herramienta *Weka* para obtener un total de 20 reglas de inferencia, de las cuales se seleccionaron las 5 reglas con un porcentaje de acierto o precisión superior al 70%, tal como se presenta en la Figura 7. Así mismo, se configuró en *Weka* el parámetro CAR=False para el modelo de reglas de asociación, el cual permite la obtención de reglas que involucran en el antecedente y consecuente los diferentes atributos que componen el *dataset.*

Para la obtención de las reglas de inferencia en mención, fue necesario asegurarse que todos los atributos fueran categóricos, así como determinar el conjunto de los que arrojan mejores reglas, de tal modo que se tuvo en cuenta los atributos: municipio, víctima, grupo etario (obtenido a partir del atributo edad) y delito (correspondiente a los 9 tipos de delitos identificados en el *dataset*), de tal modo que en la Tabla 2 se presentan las 5 reglas seleccionadas, junto con el porcentaje de acierto. Tal como se aprecia en la Tabla 2, las reglas de asociación obtenidas relacionan el municipio, el grupo etario y el género de la víctima con el tipo de delito.

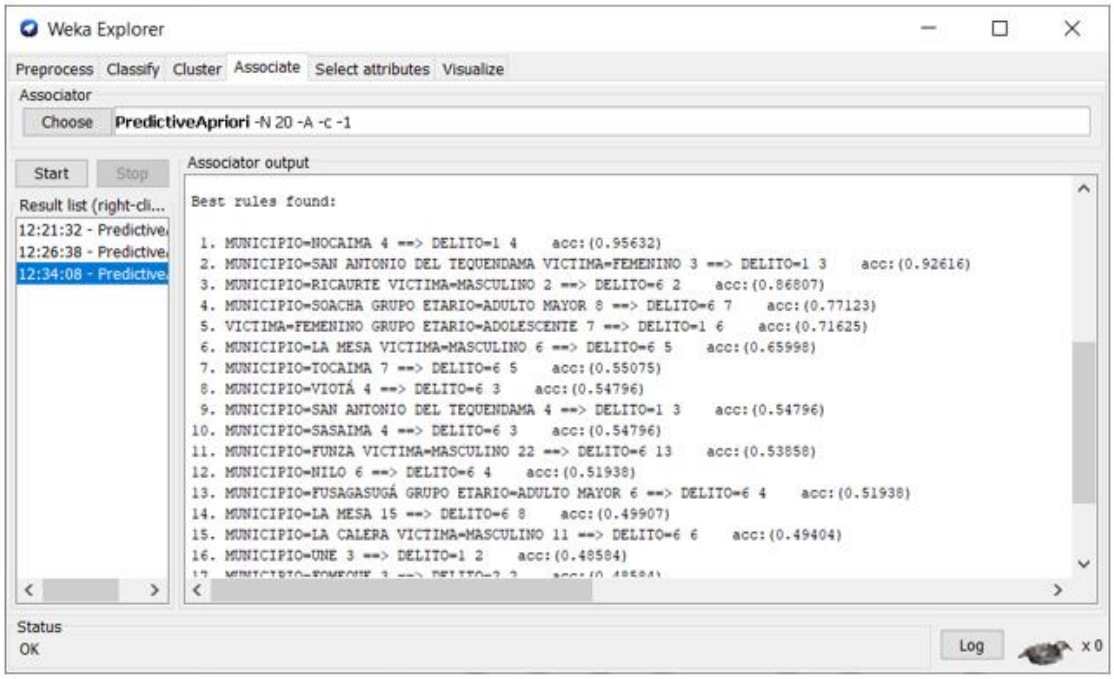

Figura 7. *Aplicación del modelo de reglas de asociación. Fuente: elaboración propia.*

Tabla 2. *Reglas de asociación seleccionadas.*

Regla	Descripción
1	Si MUNICIPIO=”Nocaima” => DELITO=1 (Violación de datos personales) Porcentaje de acierto=95.63%
2	Si MUNICIPIO=”San Antonio del Tequendama” y VÍCTIMA=”Mujer” => DELITO=1 (Violación de datos personales) Porcentaje de acierto=92.61%

3	Si MUNICIPIO=”Ricaurte” y VÍCTIMA=”Hombre” => DELITO=6 (Hurto por medios informáticos) Porcentaje de acierto=86.81%
4	Si MUNICIPIO=”Soacha” y GRUPO ETARIO=”Adulto Mayor” => DELITO=6 (Hurto por medios informáticos) Porcentaje de acierto=77.12%
5	Si VÍCTIMA=”Mujer” y GRUPO ETARIO=”ADOLESCENTE” => DELITO=1 (Violación de datos personales). Porcentaje de acierto=71.63%

Fuente: elaboración propia.

En lo referente a la aplicación de los modelos de agrupamiento, se usaron las funcionalidades de la herramienta *Weka* para la aplicación del algoritmo *KMeans* en la obtención de los *clusters* y centroides de dos modelos (ver Figura 8). El primer modelo tuvo por objetivo relacionar las edades de las víctimas con el tipo de delito cibernético, mientras que el segundo relacionó el municipio con el tipo de delito cibernético.

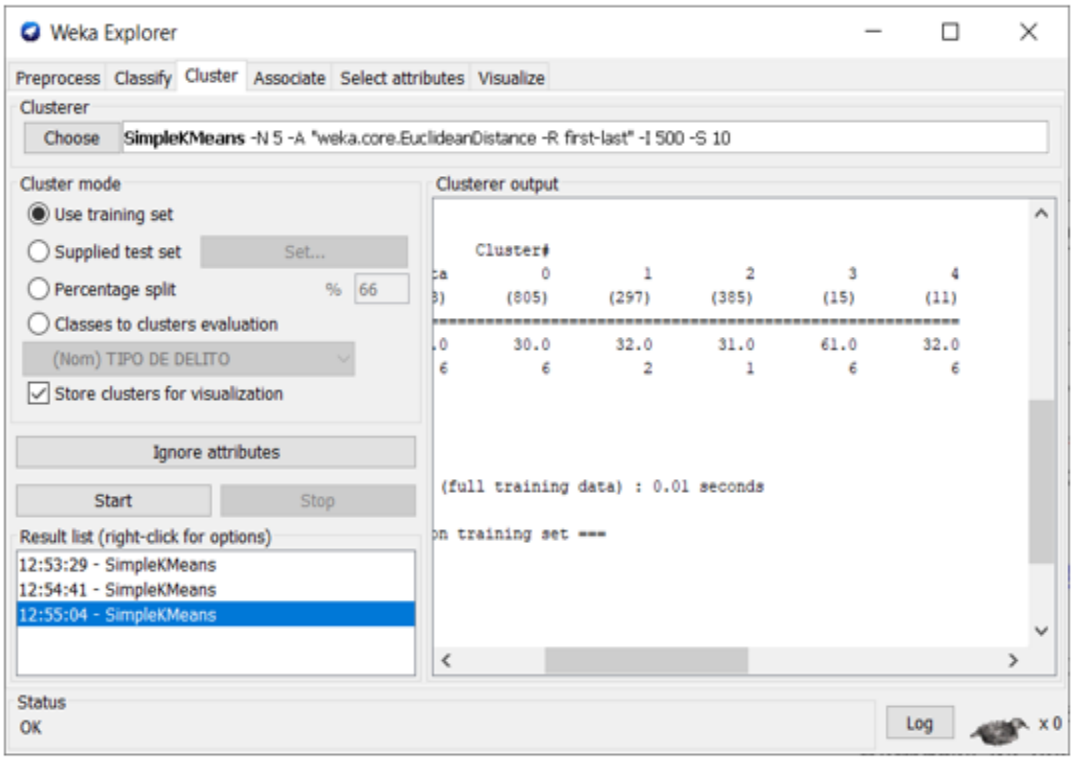

Figura 8. *Aplicación de modelos de clustering. Fuente: elaboración propia*

En cuanto a la aplicación del primer modelo, se obtuvieron un total de cinco *clusters* con sus centroides asociados, los cuales se presentan en la Tabla 3. Se puede apreciar que los *clusters* con mayor porcentaje de instancias son el 0, el 1 y el 2, permitiendo determinar respectivamente como en su mayoría personas de edades cercanas a los 30 años han sido víctimas de los delitos 6 (hurto por medios informáticos), 2 (Acceso abusivo a un sistema informático) y 1 (violación de datos personales).

Tabla 3. *Primer modelo de clustering.*

Cluster	Centroides Asociados
0	Centroide C0 = {Edad=30, Delito=6} Instancias = 805 %Instancias = 53%
1	Centroide C1 = {Edad=32, Delito=2} Instancias = 297 %Instancias = 20%
2	Centroide C2 = {Edad=31, Delito=1} Instancias = 385 %Instancias = 25%
3	Centroide C3 = {Edad=61, Delito=6} Instancias = 15 %Instancias = 1%
4	Centroide C4 = {Edad=32, Delito=6} Instancias = 11 %Instancias = 1%

Fuente: elaboración propia.

En lo referente a la aplicación del segundo modelo de *clustering*, se obtuvieron un total de 5 *clusters* con sus centroides asociados, los cuales son presentados en la Tabla 4. Se puede apreciar que los *clusters* con mayor porcentaje de instancias son el 0, el 1 y el 2, los cuales permiten determinar como el tipo de delito 6 (Hurto por medios informáticos y semejantes) está concentrado en el municipio de "Chia", mientras que los delitos tipo 1 (Violación de datos personales) y 2 (Acceso abusivo a un sistema informático) están concentrados en el municipio de "Soacha".

Tabla 4. *Segundo modelo de clustering.*

Cluster	Centroides Asociados
0	Centroide C0 = {Municipio=Chia, Delito=6} Instancias = 719 %Instancias = 48%
1	Centroide C1 = {Municipio=Soacha, Delito=2} Instancias = 353 %Instancias = 23%
2	Centroide C2 = {Municipio=Soacha, Delito=1} Instancias = 358 %Instancias = 24%
3	Centroide C3 = {Municipio=Funza, Delito=6} Instancias = 32 %Instancias = 2%
4	Centroide C4 = {Municipio=Zipaquirá, Delito=6} Instancias = 51 %Instancias = 3%

Fuente: elaboración propia.

4. Conclusiones

En este artículo se propuso como contribución, el desarrollo de un estudio exploratorio y basado en *machine learning* para el análisis de los delitos cibernéticos dentro del departamento de Cundinamarca, el cual pretende servir de referencia para la toma de decisiones por parte de las autoridades pertinentes, así como la extrapolación del estudio en otras regiones del país.

Para el desarrollo del estudio exploratorio se hizo uso de las ventajas provistas por la herramienta libre *weka*, la cual permite la aplicación de métodos de estadística descriptiva, así como modelos de aprendizaje supervisado y no supervisado. Del mismo modo, se aprovechó las ventajas provistas por la herramienta *GeoDa*, la cual permite el estudio espacial de los datos y la aplicación de modelos de agrupamiento.

A través del uso de los modelos de *machine learning*, en el presente artículo se hizo uso de reglas de asociación para determinar la relación de los atributos categóricos: municipio, víctima y grupo etario, con respecto a los tipos de delitos identificados en el *dataset*. Del mismo modo, la aplicación de los modelos de aprendizaje supervisado o *clustering*, permitieron relacionar

la edad de las víctimas con el tipo de delito, y el municipio con el tipo de delito, de tal forma que se pudo identificar de manera más clara estas relaciones entre atributos.

Tanto para el análisis exploratorio de los datos, como para la aplicación de los modelos de *machine learning*, fue necesario realizar el proceso de limpieza y adecuación de los datos al formato de las herramientas de análisis utilizadas. Así mismo, para el caso de la aplicación de las reglas de asociación, fue necesario convertir en categóricos algunos atributos como la edad, así como determinar aquellos atributos que generaran las mejores reglas a nivel lógico y a nivel de precisión.

Como trabajo futuro derivado de la presente investigación, se pretende realizar un estudio más allá de los delitos cibernéticos, sobre el *dataset* original publicado por la Policía. Del mismo modo, se pretende extrapolar el estudio a otras regiones geográficas del país.

5. Agradecimientos y declaración de financiación

Esta investigación no ha recibido ningún tipo de financiación de ninguna entidad u organización y no existe ningún conflicto de intereses en relación con esta investigación.

6. Referencias

(1). Nicol DM. The Value of Useless Academic Research to the Cyberdefense of Critical Infrastructures. IEEE Secur Priv. 2020 Jan 1;18(1):4–7.
(2). Beuhring A, Salous K. Beyond blacklisting: Cyberdefense in the era of advanced persistent threats. IEEE Secur Priv. 2014 Sep 1;12(5):90–3.
(3). Ospina Díaz MR, Sanabria Rangel PE. Desafíos nacionales frente a la ciberseguridad en el escenario global: un análisis para Colombia. Rev Crim. 2020;62(2):199–217.
(4). Ojeda Pérez J, Rincón Rodríguez F, Arias Flórez M, Daza Martínez L. Delitos informáticos y entorno jurídico vigente en Colombia. Cuad Contab. 2010;11(28):41–66.
(5). Vargas Borbúa R, Reyes Chicango RP, Recalde Herrera L. Ciberdefensa y ciberseguridad, más allá del mundo virtual: modelo ecuatoriano de gobernanza en ciberdefensa. URVIO - Rev Latinoam Estud Segur. 2017 Jun 29;(20):31–45.
(6). Reyna D, Olivera D. Las amenazas cibernéticas. In: 10 Temas de Ciberseguridad. Universidad de Xalapa; 2017. p. 49–72.
(7). Pereira T, Santos H, Mendes I. Challenges and reflections in designing Cyber security curriculum. EDUNINE 2017 - IEEE World Eng Educ Conf Eng Educ - Balanc Gen Spec Form Technol Carriers A Curr Challenge, Proc. 2017 May 2;47–51.
(8). Almanza A. XXII Encuesta Nacional de Seguridad Informática. 2022.
(9). Federal Bureau of Investigation. Internet Crime Report [Internet]. 2020. Available from: https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf
(10). Ortiz-Campos N. Normativa Legal sobre Delitos Informáticos en Ecuador. Rev Científica Hallazgos21 [Internet]. 2019;4(1):100–11. Available from: http://revistas.pucese.edu.ec/hallazgos21/
(11). Acosta MG, Benavides M, García N. Delitos informáticos: Impunidad organizacional y su complejidad en el mundo de los negocios. Rev Venez Gerenc. 2020;25(89):351–68.
(12). Pons V. Internet, la nueva era del delito: ciberdelito, ciberterrorismo, legislación y ciberseguridad. URVIO - Rev Latinoam Estud Segur. 2017;(20):80–93.
(13). Urcuqui C, García M, Osorio JL, Navarro A. Ciberseguridad: Un enfoque desde la

ciencia de datos. Universidad Icesi; 2018. 91 p.

(14). Coyac-Torres JE, Sidorov G, Aguirre-Anaya E. Detección de ciberataques a través del análisis de mensajes de redes sociales : revisión del estado del arte. Res Comput Sci. 2020;149(8):1031–41.

(15). Policia Nacional de Colombia. Dataset de delitos de alto impacto para el departamento de Cundinamarca [Internet]. Available from: https://www.datos.gov.co/dataset/DELITOS-DE-ALTO-IMPACTO-EN-EL-DEPARTAMENTO-DE-CUND/7b35-j7bt/data

(16). Nafie Ali FM, Mohamed Hamed AA. Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents. J Inf Telecommun. 2018;2(3):231–45.

(17). Do Carmo Silva M, Simoes Gomes CF, Alves Lima GB. Utilities Analysis for Latin America and Caribbean Innovation Indicators. IEEE Lat Am Trans. 2018 Nov 1;16(11):2834–40.

(18). Montoya EAQ, Colorado SFJ, Muñoz WYC, Chanchí G. Propuesta de una Arquitectura para Agricultura de Precisión Soportada en IoT. RISTI - Rev Iber Sist e Tecnol Inf [Internet]. 2017 [cited 2020 Aug 1];(24):39–56. Available from: http://www.scielo.mec.pt/scielo.php?pid=S1646-98952017000400005&script=sci_arttext&tlng=es

(19). Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. In: Handbook of applied spatial analysis. Springer; 2010. p. 73–89.

(20). Wu Z, Zhang F, Di D, Wang H. Study of spatial distribution characteristics of river eco-environmental values based on emergy-GeoDa method. Sci Total Environ. 2022;802:149679.

(21). Yang S, Ge M, Li X, Pan C. The spatial distribution of the normal reference values of the activated partial thromboplastin time based on ArcGIS and GeoDA. Int J Biometeorol. 2020;64(5):779–90.