Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

“Trabajo Final de Maestría en Tecnologías de la Información”

“Integración de procesos de explotación de información y tecnología GIS: Aplicación para el hallazgo de patrones de robos y hurtos de la Ciudad de Corrientes”

Autora: Lic. Lorena Elizabeth Flores

Directora: Dra. Sonia Itati Mariño

Co- Director: Lic. Sebastian Martins

Año 2019

“A mis padres Elena y Timoteo,

A mis guías académicos, la profesora Sonia y el profesor Sebastian”

Resumen

Los proyectos de explotación de información son ampliamente desarrollados en diversos dominios de conocimiento, que junto con la aplicación de distintas tecnologías apoyan a la gestión de incidentes. En este contexto, se propone el diseño de un procedimiento para la toma de decisiones integrando tecnología GIS (Geographic Information System) y procesos de explotación de información basados en sistemas inteligentes. La propuesta se valida considerando como caso de estudio el análisis de delitos relacionados con tipos de robo y hurto proveniente del Sistema de Alerta Temprana (SAT) y circunscriptos en la Capital de la Provincia de Corrientes.

Los resultados obtenidos permiten identificar y caracterizar información referente al comportamiento de delitos, zonas de mayor riesgo y personas más propensas a sufrir hechos delictivos, y han generado una descripción de posibles tendencias criminales en donde se observan los tipo de ataque más habituales, rangos de horarios más frecuentes, sustracción de elementos y zonas o barrios con mayor incidencia de delitos, y como el procedimiento lo propone, visualizar estos resultados a través del componente espacial GIS y analizar las capas de los mapas aprovechando el conocimiento producido con el uso de las herramientas de explotación de información.

Palabras Claves: *Sistemas de Información Geográfica, Minería de Datos. Infraestructura de Datos Espaciales, Bases de Datos Espaciales, Información criminal, Patrones delictivos.*

Abstract

Information mining projects are widely developed in various knowledge domains, which, together with the application of different technologies, support the management of incidents. In this context, the design of a decision-making procedure integrating GIS technology (Geographic Information System) and information mining processes based on intelligent systems is proposed. The proposal is validated considering as a case study the analysis of crimes related to types of robbery and theft from the Sistema de Alerta Temprana (SAT) and circumscribed in the Capital of the Province of Corrientes.

The results obtained allow us to identify and characterize information regarding the behavior of crimes, higher risk areas and people more prone to suffer criminal acts, and

have generated a description of possible criminal tendencies where the most habitual types of attacks are observed, schedules ranges more frequent, subtraction of elements and zones or neighborhoods with higher incidence of crime, and as the procedure proposes, visualize these results through the spatial component GIS and analyze the layers of the maps taking advantage of the knowledge produced with the use of exploitation tools of information.

Keywords: *Geographic System Information, Data Mining, Spatial Data Infrastructures, Spatial Data Base, Criminal Information, Criminal Patterns*

Agradecimientos

A Dios por ser mi guía y fortaleza día a día, sin Él no podría haber llegado donde estoy.

A la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste por tantos años de aprendizaje y enseñanza cotidiana.

A mis padres Elena y Timoteo, a mi hermano Leonardo, por su cariño y protección, gracias por acompañarme en cada etapa de mi vida y por su apoyo incondicional.

Un especial agradecimiento a mis profesores, Sonia y Sebastián, por el sustento y acompañamiento diario, su dedicación y predisposición en todo momento.

Índice de contenidos

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1. Introducción

La evolución y el desarrollo de la tecnología han logrado implementar nuevas e innovadoras herramientas para el tratamiento y análisis de la información.

La minería de datos (MD), como tecnología orientada a descubrir patrones significativos en grandes volúmenes de datos se aplica en diferentes dominios de conocimiento. En la actualidad, en el campo de la seguridad permite a los expertos del dominio comprender e interpretar el comportamiento de la delincuencia.

En materia de tratamiento de información, también se incluyen dentro de las tecnologías de análisis el concepto de GIS (Geographic Information System o Sistema de Información Geográfica, SIG su acrónimo en castellano), tal como lo define Chang [1] “*…un sistema informático que permite capturar, almacenar, consultar, analizar y mostrar datos geoespaciales*". Desde su lanzamiento, el GIS ha logrado traspasar las barreras tecnológicas presentando la información espacial a través de mapas georreferenciados. Estos mapas juegan un papel esencial en la vida cotidiana de las personas, su incorporación permite analizar, comunicar y compartir información con el fin de resolver y abordar problemáticas diarias, y asistir a la toma de decisiones más inteligentes.

Desde esta perspectiva, el GIS puede tratarse como una base de datos que contiene información geográfica valiosa vinculada a un determinado territorio de interés [2]. Por ello, siguiendo a [3], resulta viable aplicar MD sobre base de datos georreferenciadas para el hallazgo de patrones y regularidades significativas que resulten en conocimiento de interés sobre el territorio asociado. En el análisis delictivo, la MD aplicada sobre datos espaciales criminales permite la visualización de los hechos delictivos a través de mapas [4], con el objetivo de analizar, clasificar y predecir tendencias del delito de acuerdo a una determinada zona geográfica.

En este Trabajo Final de Maestría (TFM) se propone el diseño y la aplicación práctica de un procedimiento claro y preciso para la toma de decisiones centrado en la integración de tecnología GIS y un método de minería de datos, validado en un contexto geográfico delimitado. Se describen los procesos y herramientas utilizadas para la integración de ambas tecnologías y su aplicación sobre un caso de estudio, el cual

proporciona los elementos necesarios para enfocar una problemática en particular y ser referencia para resolver situaciones similares.

El procedimiento diseñado se validó en la detección y el análisis del comportamiento de hechos delictivos de robos y hurtos que afectan a una ciudad en el período primer semestre del año 2017, aplicando técnicas de minería de datos sobre una base de datos georreferenciada, y apoyados en tecnología GIS e IDE (Infraestructura de Datos Espaciales) para visualizar la información.

El conocimiento obtenido de la integración de la tecnología GIS y algunos algoritmos de minería de datos se reflejó en la construcción del denominado Mapa del Delito o Mapa del Crimen de la ciudad. Ésta es una herramienta cartográfica que permitió mapear y visualizar los patrones de delictualidad obtenidos del análisis realizado, en el cual se identificaron las zonas de mayor riesgo de ocurrencia de hechos delictivos.

Así, la implementación de esta solución tecnológica permitió identificar patrones para detectar y predecir la ocurrencia de estos tipos de delitos, relacionados con ubicaciones geográficas y otras variables de interés, entre las que se mencionan: i) características del hecho: lugar, día y horario de ocurrencia del delito, tipo de elemento sustraído (vehículo, domiciliario, otros), tipo de ataque (forcejeo, arrebato, etc.), tipo de arma u objeto utilizado en la escena y jurisdicción policial interviniente, ii) registro del autor del hecho: características del sospechoso (sexo del delincuente, edad aproximada, etc.), iii) registro del denunciante: sexo y edad de la víctima, entre otras.

Mediante este análisis es posible obtener información para la toma de decisiones, orientada el diseño de diferentes planes de prevención, mejora de la seguridad, alerta de situaciones de riesgo al ciudadano y reducción del impacto de la delincuencia, entre otros posibles impactos positivos.

1.1. Objetivos del Trabajo Final de Maestría

Los objetivos del Trabajo Final de Maestría se exponen a través de un objetivo general y los objetivos específicos que aportaron a su logro.

1.1.1. Objetivo general:

- Integrar tecnologías GIS y métodos comprendidos en la minería de datos para identificar patrones delictivos en un contexto geográfico delimitado.

1.1.2. Objetivos específicos:

- Estudiar, analizar y seleccionar tecnologías de georreferenciación GIS existentes.
- Investigar y profundizar en técnicas y herramientas comprendidas en la minería de datos.
- Elaborar y desarrollar una propuesta integradora de tecnología GIS y métodos comprendidos en la minería de datos, para aplicar procesos de explotación de la información en un dominio específico.
- Verificar la propuesta en un caso de estudio, para producir conocimiento orientado a la detección de hechos delictivos de robos y hurtos en la ciudad de Corrientes.

El primer objetivo específico se expone en la sección 2.1 del trabajo, el segundo en la sección 2.2, el tercero se presenta en la sección 3 y, finalmente el cuarto objetivo se muestra en la sección 5.2.

1.2. Producción científica generada en el trabajo

Durante la elaboración del presente Trabajo Final de Maestría se presentaron resultados preliminares en:

- L. E. Flores y S. I. Mariño, “Propuesta de procedimiento para el análisis delictivo basado en la explotación de la información”. XX Workshop de Investigadores en Ciencias de la Computación, 2018.
- L. E. Flores, S. I. Mariño y S. Martins, "Revisión sistemática de literatura: explotación de información y tecnologías GIS aplicadas para hallar patrones delictivos”, Revista Entorno, no. 67, pp. 30-41, 2019.

- L. E. Flores, S. I. Mariño y S. Martins,"Modelado y simulación de robos y robos basados en redes SOM, TDIDT y Bayesianas. Un caso de estudio”, International Journal of Information Systems and Software Engineering for Big Companies, vol. 6, no. 2, pp. 81-87, 2019.
- L. E. Flores, S. I. Mariño y S. Martins, “Integración de tecnologías de minería de datos y GIS. Generación de conocimiento en torno a la identificación y caracterización de clusters de robos y hurtos en una ciudad Argentina”, III Congreso Internacional de Ciencias de la Computación y Sistemas de Información. CICCSI 2019.

1.3. Estructura del trabajo

El Trabajo Final de Maestría se estructura en siete capítulos. Además, del Capítulo 1, los siguientes son:

Capítulo 2: presenta el estado de la cuestión asociado al TFM. Inicia con la introducción de los sistemas de información geográfica, seguido de la definición de los diferentes conceptos relacionados con minería de datos, explotación de información y los tipos de procesos de explotación de información. Se describen las herramientas open source disponibles para ambas tecnologías y se realiza una comparación entre ellas con el objetivo de seleccionar la herramienta más adecuada para aplicar al procedimiento propuesto. Al final del capítulo, se presenta una sintética comparación de las diferentes metodologías más utilizadas en este tipo de proyecto tecnológico.

Capítulo 3: desarrolla la solución propuesta para la integración de procesos de explotación de información y un sistema de información geográfica.

Capítulo 4: describe la problemática elegida como caso de estudio o validación del procedimiento: el uso de la minería de datos y tecnología GIS en el análisis de la información criminal.

Capítulo 5: detalla el caso de estudio adoptado para verificar la viabilidad de la propuesta. Por ello, se aplica el procedimiento propuesto y las herramientas de minería de datos y GIS seleccionadas.

Capítulo 6: presenta en detalle la etapa del procedimiento propuesto, nombre de etapa, en donde se aplican las técnicas de minería de datos sobre el caso de validación. Al final del capítulo se realiza una discusión de los resultados derivados de la minería de datos.

Capítulo 7: expone la discusión final de los resultados obtenidos. Posteriormente se presentan las conclusiones a partir del desarrollo de este TFM, y los futuros trabajos propuestos comprendidos en esta línea de investigación.

A continuación, el TFM lista las referencias bibliográficas utilizadas para su elaboración.

Como elementos finales se incluyen anexos. El Anexo 1 trata la Revisión Sistemática de la Literatura (RSL) inicialmente elaborada para definir este TFM, referencia la integración de procesos de explotación de información con tecnologías GIS y su aplicación para el hallazgo de patrones delictivos. Los Anexos 2, 3 y 4, presentan la traducción de reglas de pertenencia a los clusters formados con la aplicación de los objetivos de minería de datos Nº 1, 2 y 3 respectivamente definidos para verificar el procedimiento.

Capítulo 2

Estado de la cuestión

2. Estado de la cuestión

El capítulo presenta una síntesis del estado del arte de aquellos temas relacionados con el objetivo de este Trabajo Final de Maestría. En primer lugar, se introduce el concepto de sistema de información geográfica (sección 2.1), se describen los datos geoespaciales y el concepto de sistemas de referencia geográfica, se hace una descripción de la Infraestructura de Datos Espaciales, sus servicios, y las principales tecnologías GIS open source presentes en el mercado. En la siguiente sección, se detallan brevemente los conceptos relacionados con la minería de datos (sección 2.2), los procesos de explotación de información y las principales herramientas para minería de datos open source disponibles en el mercado. Por último, se exponen las características de las principales metodologías para proyectos de minería de datos.

2.1. Sistema de Información Geográfica

En los últimos años, se ha hecho habitual incorporar en la vida cotidiana de las personas, aparatos cuyas tecnologías buscan facilitar las tareas diarias del ser humano a través del uso de mapas como herramientas esenciales para llegar de un lugar al otro en el menor tiempo posible, o simplemente para la búsqueda o ubicación de un punto especifico desconocido, entre otras problemáticas. Esto demuestra precisamente como las potentes herramientas de geolocalización han logrado extender su uso y adaptarse a las necesidades del usuario ayudando a comprender cualquier tipo de información disponible.

El GIS ha revolucionado la sociedad gracias a su amplia variedad y aplicación sobre distintas áreas como la educación, la salud, la seguridad pública, el ambiente, los servicios públicos y las telecomunicaciones, permitiendo una mejor interpretación de la información a través de distintas aplicaciones versátiles, simples y fáciles de utilizar. Desde esta perspectiva, numerosas empresas y comercios incorporan este tipo de tecnología y aprovechan su capacidad para captar información, procesarla, analizarla y difundirla con el objetivo de mejorar sus proyectos y negocios, optimizar recursos y simplificar sus procesos diarios.

Al mismo tiempo, diversas organizaciones gubernamentales nacionales, provinciales y locales han adoptado el uso de GIS como herramienta principal de difusión de datos a la ciudadanía. Esta introducción o innovación se sostiene como una estrategia orientada a

mejorar la calidad de vida a través de la recolección, el proceso, el análisis y la gestión de datos geográficos idóneos a cada entidad, haciendo uso de las mismas para la planificación y el ordenamiento territorial, conformando así una base de datos cartográfica única para cada de ellas.

2.1.1. Introducción a los datos geoespaciales

Uno de los aspectos más importantes de GIS es la georreferenciación. Esta tecnología utiliza datos georreferenciados y los representa a través de diferentes tipos de objetos vectoriales, los cuales pueden ser, un punto, una línea o un polígono. En primer lugar, los puntos representan *“…objetos espaciales que sólo están localizados, no tienen dimensiones, es decir, ni largo (i. e., longitud) ni ancho (anchura). La posición de cada objeto queda fijada a través de las coordenadas de los sistemas de referencia”* [5]. Éstos representan los centros de salud, las comisarías, las estaciones de paradas de colectivos, entre otros objetos de interés.

Por parte, las líneas constituyen *“...una sucesión de puntos y representan objetos espaciales con una dimensión, longitud. Su posición se fija con dos pares de coordenadas”* [5]. Con éste tipo de objeto se visualizan las calles de la ciudad, los recorridos de líneas de colectivos, etc.

Y finalmente, los polígonos que se definen como *“… líneas cerradas y representan objetos espaciales de dos dimensiones: longitud y anchura. La posición de cada objeto se fija con dos o más líneas cuyas coordenadas inicial y final coinciden*” [5]. Éstos se utilizan para representar, por ejemplo, los barrios de la ciudad, las plazas, etc.

De igual forma, es importante indicar el archivo con formato shape como el más utilizado por la tecnología GIS y puede representar cualquiera de las formas mencionadas anteriormente (punto, línea o polígono). El shape se compone de cuatro archivos con distintos tipos de extensiones, el primero de ellos es el archivo shp que representa le geométrica del objeto, el archivo dbf que contiene los atributos descriptivos del objeto espacial, el fichero shx que constituye el índice para el manejo de tablas entre archivos y por último el archivo prj que representa el sistema de coordenadas necesario para ubicar geográficamente el objeto espacial [5]. Para lograr la representación correcta, el archivo shape deber contener el conjunto de archivos mencionados anteriormente.

A continuación, en la Fig. 2.1 se ilustra un ejemplo de un archivo con formato shape de tipo punto, en donde cada punto contiene dos coordenadas únicas (X e Y). Estas coordenadas ubican una posición geográfica específica sobre la superficie de la tierra. Se observa, además, los atributos del objeto espacial (nombre y valor) que representan las características de un punto especifico [6].

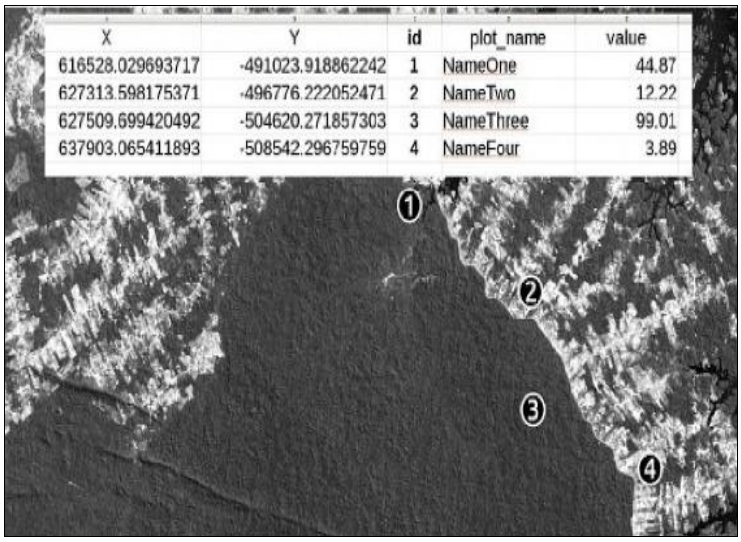


Fig. 2.1. Datos geoespaciales:

Fuente: [6]

Los archivos shape pueden representarse a través de diferentes aplicaciones GIS disponibles en el mercado para realizar distintos tipos de análisis según se requiera.

2.1.2. Sistemas de referencia geográficas

Para definir un objeto espacial es imprescindible mencionar el concepto de sistema de referencia geográfica. Así para lograr una representación correcta y que los objetos espaciales puedan considerarse georreferenciados, éstos deben estar en un marco espacial que los sitúe en un contexto geográfico específico.

El estudio de los diferentes tipos de sistemas de referencia no es sencillo. Sin embargo, se puede mencionar una de las más utilizadas en todo el mundo para aplicaciones topográficas, cartográficas y de navegación, el sistema de referencia WGS84 (World Geodetic System, o en español Sistema Geodésico Mundial), comúnmente utilizado por los GPS (Global Positioning System, o Sistema de Posicionamiento Global, su traducción en castellano), la tecnología más aplicada para determinar la posición de cualquier objeto sobre la Tierra.

En Argentina, el Instituto Geográfico Argentino (IGN) establece el uso de un marco de referencia específico llamado POSGAR (POSiciones Geodésicas ARgentinas) [7]. Este marco se apoya en el sistema de referencia WGS84 y se adopta y aplica en este país para enmarcar cualquier punto geográfico sobre las provincias argentinas [8].

2.1.3. Infraestructura de Datos Espaciales

Las organizaciones públicas o de estado tienden a brindar y compartir información espacial con el fin de proporcionar mayor disponibilidad y acceso a los datos. Este concepto de datos abiertos ha incurrido en un cambio transcendental y se presenta con el nombre de Infraestructura de Datos Espaciales (IDE). Esta infraestructura engloba tanto la unificación como la estandarización de datos geográficos entre los sistemas de información geográfica, logrando de esta forma la interoperabilidad y la colaboración entre ellos, dado que permite consumir y utilizar datos geográficos de otras fuentes para la construcción de nuevas IDE.

De manera forma, se define que una IDE *“…corresponde a un conjunto de tecnologías, políticas y estándares que permiten procesar información georreferenciada o espacial, y facilitan su acceso”* [9]. Es así que este concepto trata de un sistema informático en el cual se integran recursos y se ofrecen servicios geográficos web, y se incluye normas y procedimientos para estandarizar estos datos.

Actualmente en la Argentina existe una entidad que a través de normas y estándares nuclea y regula las diferentes IDE provinciales y locales que conforman lo que se define como Infraestructura de Datos Espaciales de la República Argentina o IDERA. Este organismo pretende lograr la disponibilidad de la información, buscando que los datos geográficos de cada entidad que la conforma puedan ser accedidos de forma gratuita por cualquier otro organismo.

2.1.3.1. Servicios geográficos web OGC

Una de las características principales de una IDE es ofrecer servicios web conocidos como servicios geográficos OGC (Open Geospatial Consortiumu o Consorcio Geoespacial Abierto su traducción en castellano, que referencia a la entidad que lo regula) los cuales permiten trabajar y compartir información espacial entre distintas IDE.

Uno de los servicios web geoespaciales más utilizados y disponibles de un OGC es el Servicio Web de Mapas o WMS (Web Map Service), que ofrece la visualización de mapas y permite la consulta de datos geográficos. Otro servicio más completo, es el Servicio Web de Características o WFS (Web Feature Service) que proporciona los métodos necesarios para la consulta y descarga de datos. Ambos geoservicios se pueden utilizar desde cualquier aplicación GIS disponible.

2.1.4. Software GIS open source

La información geográfica se presenta a través de programas o software GIS disponibles que soportan esta característica espacial, además de otras funciones, es decir, una herramienta que permite analizar, compartir y visualizar información georreferenciada [10]. En la actualidad existen numerosas aplicaciones comerciales y de software libre que admiten este proceso de georreferenciación. Además, permiten trabajar con lo que comúnmente se denominan capas informáticas geográficas, que básicamente contienen los datos espaciales.

En la categoría de software libre u open source, se mencionan algunos como QGIS (Quantum Geographic Information System o Sistema de Información Geográfica Cuántica, su traducción en castellano), gvSIG (Sistema de Información Geográfica de Generalitat Valenciana por su origen en la ciudad de España), OpenEV, GRASS (Geographic Resources Analysis Support System o Sistema de Soporte de Análisis de Recursos Geográficos, su traducción en castellano), entre otros [11], o algunos de los paquetes o software pagos como ArcGIS, Georeferencer, entre los más conocidos.
Entre los principales paquetes GIS de plataformas libres e independientes, se mencionan:

- QGIS: Es una plataforma GIS open source bajo la licencia pública general GNU. Es escalable con complementos y librerías desarrollados en lenguajes Python y C++ [12], posee múltiples características que la hacen atractiva en el mercado y de fácil aprendizaje para los usuarios. Presenta una variedad de funcionalidades y dispone de la descarga de plugins que la complementan para realizar cualquier tipo de análisis. Su desarrollo forma parte de un proyecto oficial de la Fundación Aeroespacial de Código Abierto (OSGeo u Open Source Geosptial Foundation),

que promueve la colaboración de desarrollos de software geomáticos de código abierto [13].

- gvSIG: Se trata de un software GIS open source de escritorio cuyas características principales son la interoperabilidad y su fácil interacción con el usuario común. Trabaja bajo la licencia GNU [14].
- OpenEV: Es una aplicación GIS desarrollado sobre el lenguaje de programación Python y al igual que el resto utiliza la biblioteca GDAL para trabajar y visualizar datos vectoriales y de tipo raster [11].
- GRASS: Se trata de un GIS de software libre que trabaja bajo la licencia GNU. Dispone de potentes herramientas para el procesamiento de distintos tipos de datos para su posterior gestión y análisis [12].

Las librerías Python por otra parte, en entornos GIS, permiten trabajar con datos espaciales. La librería GDAL (Geospatial Data Abstraction Library o Librería de Abstracción de Datos Espaciales, su traducción en castellano) que trabaja con datos en formato ráster, y la OGR que se ocupa de los datos vectoriales [15]. Ambas están disponibles para su descarga y utilización.

Mencionadas las herramientas expuestas anteriormente, se ha realizado un estudio de cada una de ellas considerando algunas de las características principales con las que deben contar las tecnologías de geolocalización. En la Tabla I se presenta un cuadro comparativo, con la finalidad de contribuir en la elección de la herramienta más adecuada para la realización del presente estudio.

Tabla I.

Comparativa de herramientas GIS open source:

Fuente: elaboración propia.

Característica a evaluar / Software GIS	Georreferenciación	Documentación disponible	Servicios OGC	Plugins disponibles	Multiplataforma	Desarrollo de nuevos plugins	Mantenimiento y actualización
QGIS	Este software utiliza un complemento que viene por defecto de fábrica llamado “Georreferenciador GDAL”, y no requiere de ningún tipo de instalación.	La documentación requerida se encuentra disponible a través de la página oficial de QGIS. Dispone de soporte técnico y ayuda a través de un contacto oficial. También ofrece tutoriales de capacitación.	Permite la publicación y descarga de todos los servicios web OGC.	Posee más de 430 plugins disponibles para su descarga.	Se ejecuta bajo múltiples plataformas como: Linux, Unix, Mac OSX, Windows y Android.	Si. Con lenguaje Python y C++	Constantemente la comunidad de desarrollo lo actualiza, última versión lanzado en junio del año 2019
gvSIG	Este software utiliza una extensión de llamada “Teledetección” a partir del cual se puede georreferenciar.	La documentación requerida se encuentra disponible a través de la página oficial de gvSIG y además dispone de soporte técnico y ayuda a través de un contacto oficial.	Permite la descarga de todos los servicios web OGC.	Posee 14 paquetes disponibles con sus correspondientes plugins libres de descarga, con la restricción de que algunos de ellos solo están disponibles para determinas plataformas.	Se ejecuta bajo múltiples plataformas como: Windows, Linux, Mac OSX (solo versión portable)	Si. Con lenguaje Python y Java	La comunidad de desarrollo lo actualiza, última versión lanzado en febrero del año 2018
OpenEV	Carece de complemento para georreferenciar.	Dispone de tutoriales y manuales para su descarga desde el sitio oficial de openEV, y una lista de contactos para consultas.	No permite la publicación y descarga de los servicios web OGC.	Se desconoce con exactitud.	Corre bajo las plataformas de Linux, Windows, Solaris e IRIX	Si. Con lenguaje Python.	No posee constate actualización, última versión lanzada en el año 2004
GRASS	Este software cuenta con 5 módulos para lograr la georreferenciación: i.group, i.targtet, i.points, i.rectify, i.rectify2	La documentación requerida se encuentra disponible a través de la página oficial de GRASS y además de ayuda a través de un contacto oficial.	Permite la descarga de todos los servicios web OGC.	Posee más de 350 plugins disponibles para su descarga.	Se ejecuta bajo múltiples plataformas como: Windows, Linux y Mac OSX	Si. Con lenguaje Python.	La última versión fue lanzado en marzo del año 2019

Al indagar en algunas de las herramientas GIS open source, se mencionaron las principales características de cada una de ellas (Tabla I). A continuación se presentan un resumen de ideas que justifican la elección del software QGIS, como la herramienta más indicada para la realización de este Trabajo Final de Maestría:

- Ofrece conexión a cualquier servicio geográfico OGC y permite la carga de datos brindados por las diferentes Infraestructuras de Datos Espaciales (IDE).
- Es económica y dispone de una variedad de plugins o complementos para realizar cualquier tipo de tarea de forma rápida y sencilla.
- Crea salidas gráficas a través de su diseñador de impresión, el cual lo posiciona como una de las herramientas más viables para generar mapas de calidad.
- Brinda mayor soporte de plataformas disponibles para su instalación.
- Permite manipular distintos tipos de formatos de archivo vectoriales. Esta característica supone una ventaja en la aplicación del procedimiento propuesto en la etapa de conversión de datos explotados, dado que es esencial que la herramienta GIS permita y acepte archivos con distintos tipo de formatos.

2.2. Minería de datos

La inteligencia de negocios o Business Intelligence (BI, por sus siglas en inglés) ofrece un concepto clave desde una perspectiva empresarial. Combina tecnologías y procesos que asisten al análisis de los datos y a la transformación de la información empresarial, con el fin de obtener conocimiento válido que permita la toma de decisiones apropiadas orientadas a optimizar sus recursos y mejorar sus resultados.

La minería de datos es la subdisciplina de los sistemas de información que contribuye a la inteligencia de negocios las herramientas necesarias para explotar la información necesaria y transformarla en conocimiento útil. Para realizar este proceso de búsqueda de patrones se requiere el uso de técnicas de minería de datos (árboles de decisión, redes neuronales artificiales, técnicas bayesianas, etc.) y la aplicación de algoritmos específicos dependiente de la problemática (predicción, clasificación, asociación, agrupamiento, entre otros etc.) [16].

Se presenta una síntesis de los procesos de minería o explotación de información para obtener conocimiento a partir de los datos disponibles [17]:

Descubrimiento de reglas de comportamiento: El proceso se emplea cuando se requiere identificar cuáles son las características o factores que determinan un resultado específico en el dominio del problema. En la Fig. 2.2 se puede observar el proceso a partir de la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) para luego conformar lo que se denomina datos integrados, resultado de la unificación de los mismos. A continuación, se selecciona el atributo clase con base a los datos integrados y se aplica algoritmo de inducción TDIDT (Top Down Induction of Decision Trees) para descubrir el conjunto de reglas que definen el comportamiento de dicha clase [17].

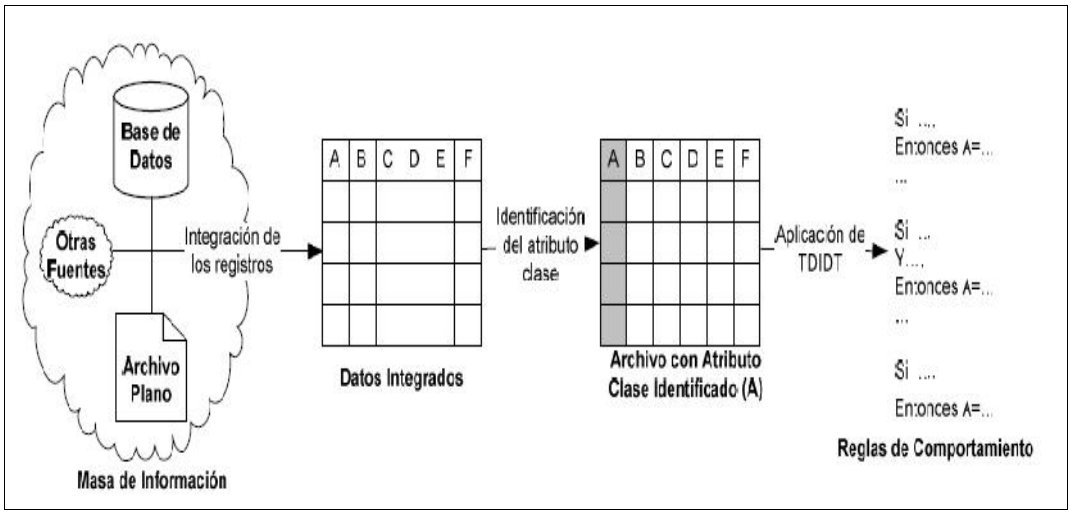


Fig. 2.2. Esquema y subproductos resultantes de aplicar TDIDT
al descubrimiento de reglas de comportamiento:
Fuente [17].

Proceso de descubrimiento de grupos: Este proceso se aplica cuando se requiere obtener una partición de la masa de información del dominio de problema y sobre el cual no se dispone ningún criterio de agrupamiento a priori. En la Fig. 2.3 se puede visualizar el proceso a partir de la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) que se integran entre sí formando una sola fuente de información llamada datos integrados. A continuación, se aplica SOM (Self Organized Maps) y se obtiene una partición del conjunto de registros en distintos

grupos o clusters denominados grupos identificados, y se genera un archivo para cada grupo identificado [17].

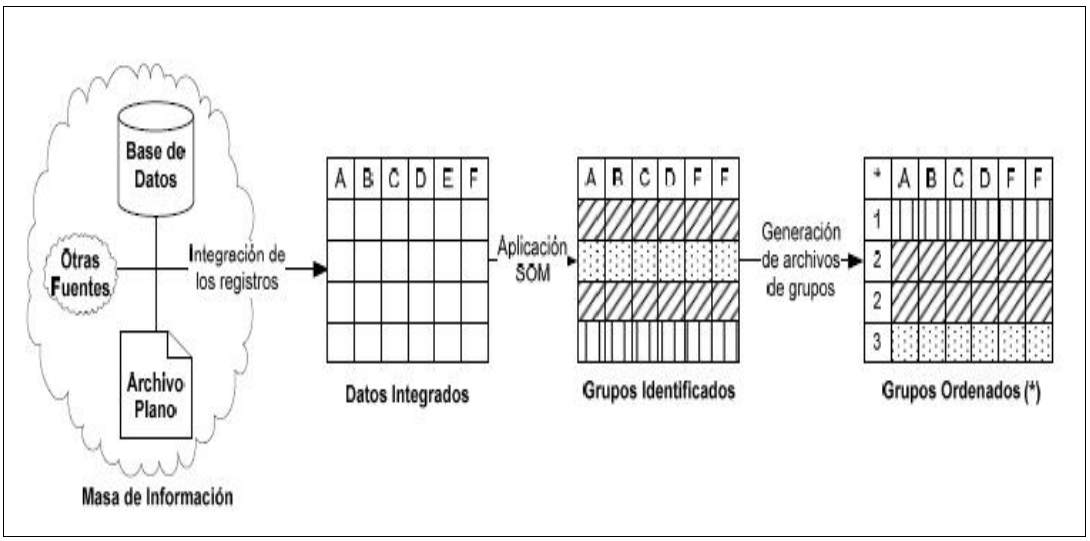


Fig. 2.3. Esquema y subproductos resultantes de aplicar SOM para el descubrimiento de grupos: Fuente: [17]

Proceso de descubrimiento de atributos significativos: Este proceso se aplica para conocer en qué medida los valores de un atributo inciden sobre valor de un atributo clase. Es decir, cuáles de estos factores tienen mayor frecuencia o incidencia.

En la Fig. 2.4 se visualiza el proceso de descubrimiento de atributos significativos a partir de la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) que se integran entre sí formando una sola fuente de información llamada datos integrados. A continuación, se identifica el atributo clase y a éste se le aplica el aprendizaje predictivo Redes Bayesianas, seguido de este proceso se obtiene el árbol de ponderación de interdependencias de atributos el cual representa la frecuencia de cada atributo sobre el atributo clase [17].

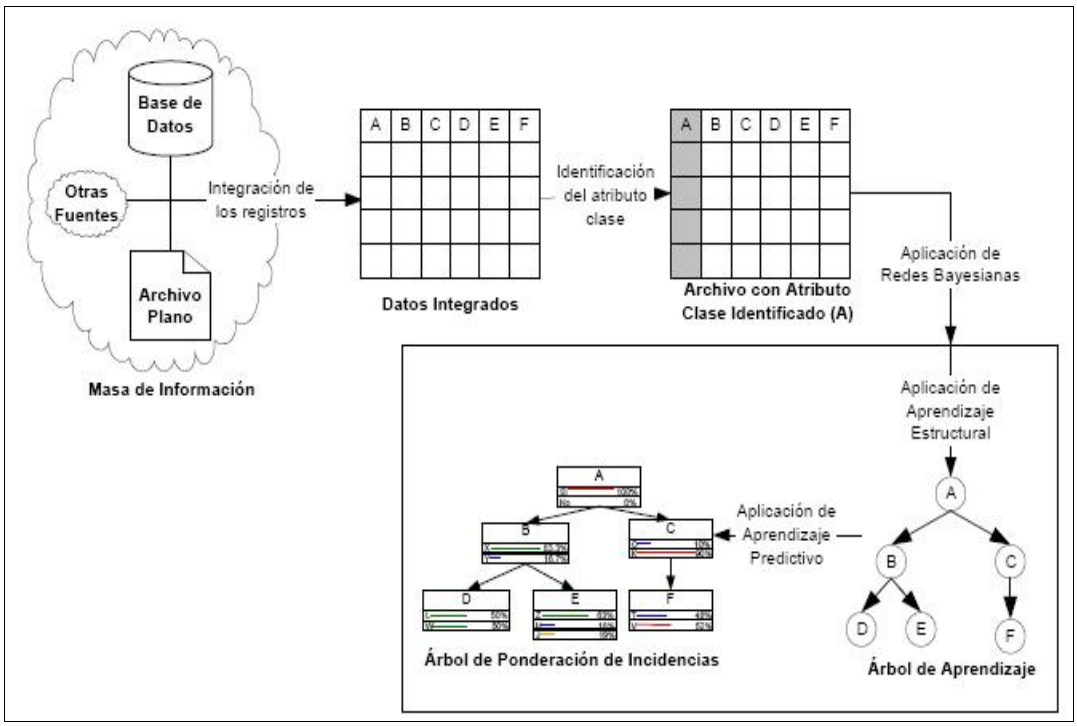


Fig. 2.4. Esquema y productos resultantes para aplicar el descubrimiento de procesos de atributos significativos utilizando redes bayesianas:

Fuente: [17]

Proceso de descubrimiento de reglas de pertenencia a grupos: Este proceso se emplea para identificar cuáles son las características o condiciones de pertenecía de cada una de las clases a una partición desconocida a priori. En la Fig. 2.5 se puede visualizar el proceso de descubrimiento de grupos a partir de la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) que se integran entre sí formando una sola fuente de información llamada datos integrados. A continuación, se aplica SOM y se obtiene una partición del conjunto de registros en distintos grupos o clusters denominados grupos identificados, y a partir del cual se genera un archivo correspondiente a cada grupo identificado. A este conjunto de archivos se lo llama grupos ordenados. El atributo “grupo” de cada grupo ordenado se identifica como el atributo clase de dicho grupo denominado atributo clase identificado (GR). Por último se aplica el algoritmo de inducción TDIDT al atributo clase de cada grupo GR y se obtiene el conjunto de reglas que definen el comportamiento de cada grupo [17].

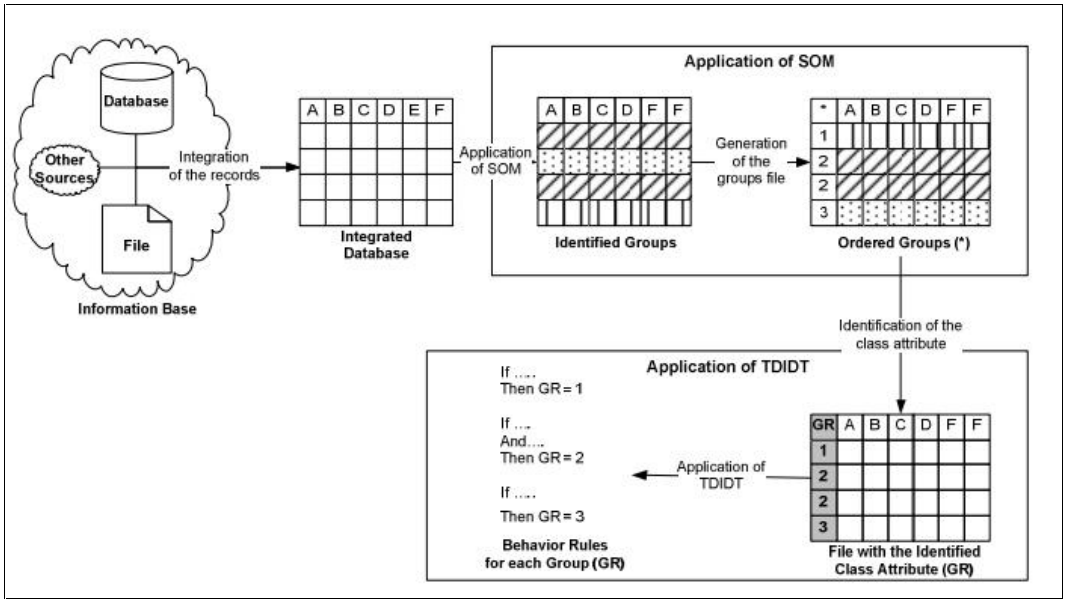


Fig. 2.5. Esquema y productos resultantes de la ejecución del proceso de descubrimiento de la pertenencia a grupos. Reglas que usan SOM y TDIDT:

Fuente: [17]

Proceso de ponderación de reglas de comportamiento o reglas de pertenencia a grupos: Para la aplicación de este proceso en necesario identificar previamente cuando hay grupos y cuando no los hay.

El procedimiento a aplicar cuando hay grupos identificados se observa en la Fig. 2.6, donde se visualiza la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) que se integran entre sí formando una sola fuente de información llamada datos integrados. A continuación, se selecciona el atributo clase para la aplicación del algoritmo de inducción TDIDT con el cual se obtiene el conjunto de reglas que caracterizan a dicho clúster. Por último se aplica Redes Bayesianas al archivo con atributo clase obtenido por la aplicación del algoritmo TDIDT, y a través del cual se obtiene el árbol de aprendizaje que presenta la ponderación de interdependencias de los atributos establecidos como antecedente de las reglas y que tienen mayor incidencia sobre el atributo establecido como consecuente [17].

El procedimiento a aplicar cuando no hay grupos identificados se observa en la Fig. 2.6, en donde se visualiza la identificación de las distintas fuentes de información disponibles (bases de datos, archivos planos, otras fuentes) que se integran entre sí formando una sola fuente de información llamada datos integrados. Con base en los datos integrados se aplican SOM y se obtiene una partición del conjunto de registros en

distintos grupos llamados grupos identificados. Para cada grupo identificado se generará el archivo correspondiente, y a este conjunto de archivos se lo llama grupos ordenados. El atributo “grupo” de cada grupo ordenado se identifica como el atributo clase de dicho grupo, constituyéndose este en un archivo con atributo clase identificado (GR). A continuación, se aplica Redes Bayesianas al archivo con atributo clase obtenido por la aplicación del SOM, y se obtiene el árbol de aprendizaje que determina la ponderación de los atributos que mejor describen la pertenencia al grupo [17].

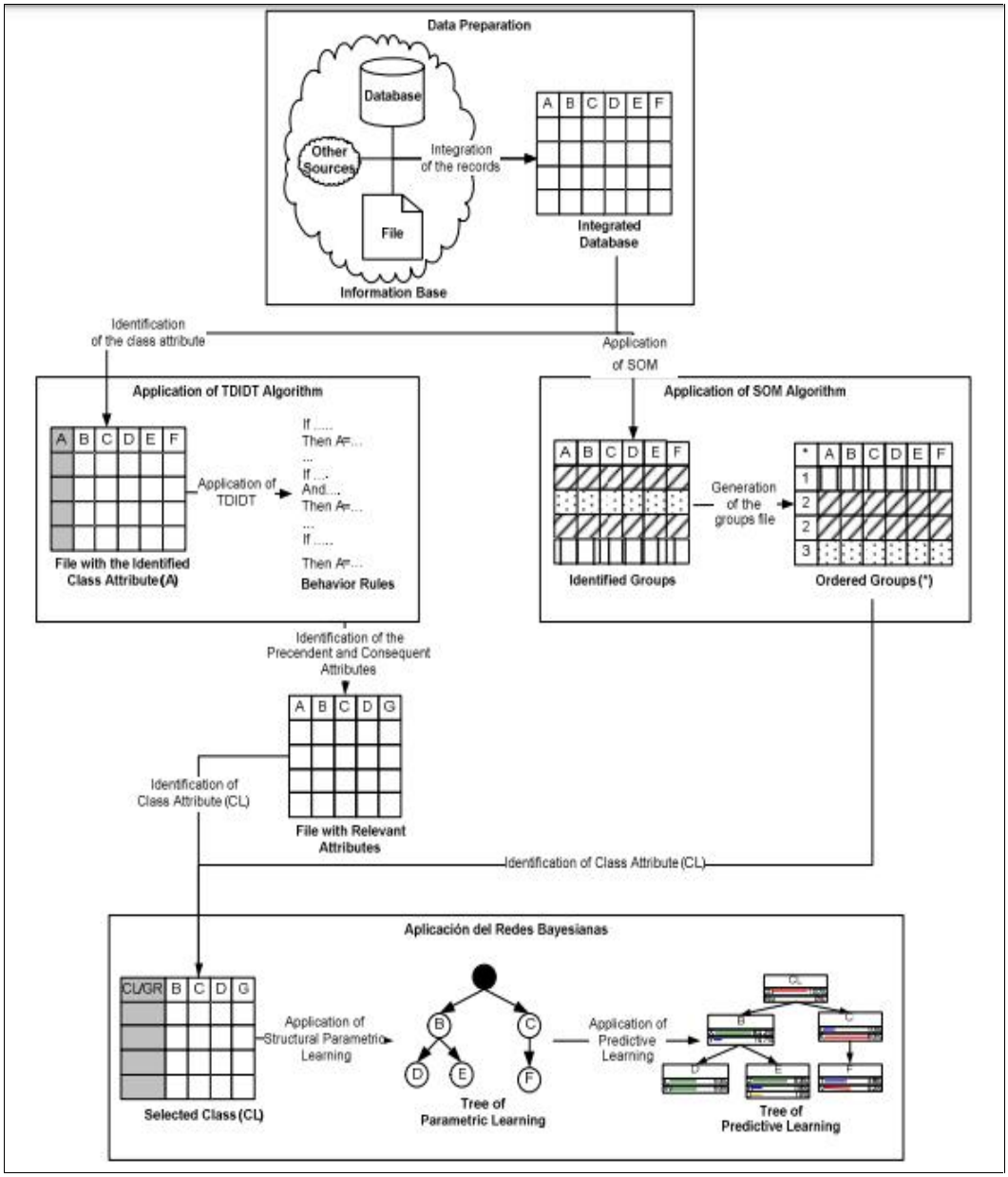


Fig. 2.6. Esquema y productos resultantes del proceso en ejecución de ponderación de comportamiento o reglas de pertenencia a un grupo que utilizan SOM, TDIDT y Redes Neuronales:
Fuente: [17]

2.2.1. Herramientas de minería de datos open source.

Para tener una visión general se presenta a continuación, algunas de las herramientas de minería de datos open source disponibles en el mercado:

La herramienta Tanagra es un software gratuito para fines académicos y de investigación. Propone varios métodos de análisis exploratorio de datos, aprendizaje estadístico, aprendizaje automático, entre otras [18]. Los algoritmos de minería de datos para extraer patrones de conocimiento ocultos en grandes masas de datos pueden clasificarse en dos categorías: aprendizaje supervisado y aprendizaje no supervisado. Tanagra provee una amplia variedad de algoritmos pertenecientes a ambas categorías. Entre algunas características que facilitan su uso se mencionan:

- Software de distribución libre y gratuita.
- Dispone de una interfaz gráfica amigable y fácil de usar.
- Permite la transformación de datos sin requerir de otra herramienta auxiliar.
- Posee una amplia variedad de algoritmos de minería de datos para la aplicación de los diferentes procesos de explotación de información.
- Despliega de forma sencilla los resultados derivados de aplicar los diferentes algoritmos y ofrece salidas de fácil interpretación en diferentes formatos.
- Cuenta con la visualización de gráficas para el análisis de los resultados de aplicar los diferentes algoritmos disponibles.

Otro de los software de minería de datos más conocido y desarrollado en JAVA es el paquete open source Weka (Waikato Environment for Knowledge Analysis o entorno para análisis del conocimiento de la Universidad de Waikato, su traducción en castellano). Se ejecuta bajo la Licencia Pública General de GNU y contiene herramientas para la aplicación de diferentes técnicas de minería de datos como la clasificación, la regresión, el clustering, reglas de asociación, entre otras [19].

ORANGE, es otra herramienta de código abierto para el aprendizaje automático, permite la visualización de datos para principiantes y expertos. Esta herramienta básicamente facilita el análisis de datos interactivos con un gran número de herramientas disponibles [20].

La selección de la herramienta de minería de datos para la aplicación del procedimiento propuesto se fundamentó en una comparativa expuesta en las Tablas II.a y II.b [21] y referenciando al método propuesto en [22]. Del análisis expuesto, se consideró a la herramienta Tanagra como adecuada para su aplicación en este estudio.

Tabla II.a

Comparativa de herramienta de minería de datos open source:

Fuente: [21]

Reporte de Evaluación de Herramientas					
Criterios:					
Evaluación: 1 = Malo, 2 = débil, 3 = Bueno, 4 = Excelente					1 = No, 4 = SI
Herramientas		Peso	Tanagra V.1.4.50	Weka V.3.7.11	Orange V.2.7.8
1. Funcional - Características Técnicas					
Soporte de Metodología / Ciclo de vida	Soporte del proceso	3	2	2	2
Compatibilidad con fuentes de datos	Base de datos	8	--	--	--
	Otras fuentes (word, excel, etc.)	8	3	2	3
Integración	Soporte de distintas técnicas asociadas al proceso de explotación de Información	5	4	4	4
Multilenguaje	Soporta distintas idiomas	2	1	1	1
Técnicas	Variedad de técnicas que provee	18	4	4	4
Reporte y visualización	Permite generar reportes y visualizaciones	12	2	2	2
Multiplataforma	Soporta múltiples plataformas	5	1	4	4
Instalación remota	La administración y mantenimiento son remotos	5	--	--	--
Usuarios Múltiples	Posee perfiles de usuarios	2	1	1	1
Seguridad	Provee seguridad de la información configurada por perfiles	2	1	1	1
Backup	Metodología de backup	2	1	1	1
Amigable	Interfaz de usuario	10	4	2	4
Configuraciones	Permite la configuración del perfil	8			
Documentación	Servicio de soporte y ayuda	5	4	1	3
Conexión	Soporta conexión por: Internet, FTP, ERPs.	2	1	1	1
Soporte de sistemas de mensaje	Soporta compartir información (por mail u otro medio)	3	1	1	1
Total			224	196	234
	Peso del Grupo	40%	89,6	78,4	93,6

Tabla II.b

Comparativa de herramienta de minería de datos open source:

Fuente: [21]

Reporte de Evaluación de Herramientas					
Criterios:					
Evaluación: 1 = Malo, 2 = débil, 3 = Bueno, 4 = Excelente					1 = No, 4 = SI
Herramientas		Peso	Tanagra V.1.4.50	Weka V.3.7.11	Orange V.2.7.8
2. Características del Proveedor					
Características del proveedor	Historia	30	3	3	1
Crecimiento	Perspectiva a futuro	10	2	3	2
Ubicación Geográfica	Oficinas	30	--	--	--
Implementación	Otras implementaciones de la misma herramienta	5	--	--	--
	Contacto con otros clientes	5	--	--	--
Confidencialidad	Confidencialidad de la información	20	--	--	--
Total			110	120	50
	Peso del Grupo	25%	27,5	30	12,5
3. Características del Servicio					
Garantía del producto	Duración y Alcance	30	--	--	--
Mejora	Brinda soporte a versiones previas	20	1	1	1
Licencia	Costo, alcances y soporte postventa	30	--	--	--
Soporte	Tiempo de respuesta y disponibilidad	20	--	--	--
Total			20	20	20
	Peso del Grupo	20%	4	4	4
4. Características Económicas					
Costo del software	Costo de la herramienta	30	--	--	--
Costo del Hardware	Necesidad de mejorar o comprar nuevo hardware compatible con la herramienta	20	--	--	--
Otros costos software	Costos adicionales al producto (backup, web servers, bases de datos, etc.)	20	--	--	--
Licencias	Política de licencia	10	--	--	--
Financiamiento	Existencia	10	--	--	--
Mejoras	Costo promedio de la mejora del producto	10	--	--	--
Total			0	0	0
	Peso del Grupo	-15%	0	0	0
Final					
1. Funcional - Características Técnicas		40%	89,6	78,4	93,6
2. Características del Proveedor		25%	27,5	30	12,5
3. Características del Servicio		20%	4	4	4
4. Características Económicas		-15%	0	0	0
TOTAL			121,1	112,4	110,1

Cabe mencionar además que, en determinados proyectos se puede utilizar un lenguaje de programación con el objetivo de extender las funcionalidades del software de minería de datos escogido. Lenguajes de programación como R [23] o Python [24], son algunos de los más populares. R es utilizado principalmente para la computación estadística [25] y presenta más de 100 paquetes disponibles [26] que, junto con las librerías disponibles de Python [27] posibilitan la extensión del uso de las herramientas clásicas de minería de datos de acuerdo a las necesidades específicas del usuario.

Generalmente, se carece de una herramienta universal que se ajuste y aplique de forma exitosa a cualquier proyecto de explotación de información. No obstante, la elección de las técnicas y herramientas se debe ajustar o adaptar a los objetivos del negocio, la planificación, las características del equipo de trabajo y sus conocimientos.

En el TFM para aplicar el procedimiento propuesto, se optó por Tanagra considerando el estudio comparativo referente a aplicaciones software de minería de datos (Tablas II.a y II.b) en que la indica como la herramienta más efectiva.

2.2.2. Metodologías para proyectos de minería de datos

En [28] se establece que los esfuerzos en el área de la minería de datos se centraron mayoritariamente en indagar las técnicas para la explotación de información y extracción de patrones (tales como árboles de decisión, análisis de conglomerados y reglas de asociación).

Las disciplinas de la informática se caracterizan por proponer diferentes enfoques plasmados en modelos de proceso, metodologías procedimientos y que utilizan herramientas. En este sentido, Ochoa [29], y otros autores señalan la necesidad de un proceso riguroso, preciso y sistematizado que guie el desarrollo de los proyectos [28], [30], [31], es decir, la utilización de modelos de proceso o metodologías.

Se relevaron diversas metodologías para desarrollar procesos de minería de datos. Entre algunas de ellas, se mencionan: KDD (Knowledge Discovery in Databases) [32], SEMMA (Sample, Explore, Modify, Model, and Assess) [33], CRISP-DM (Cross Industry Standard Process for Data Mining) [34], CATALYST [35], MPIMD (Modelo de Proceso de Ingeniería de Minería de Datos) [30], ASD-BI (Adaptive Software Development – Business Intelligence) [36], FMDS (Foundational Methodology for Data Science) [37] y TDSP (Team Data Science Process) [38].

En [39], se realiza un análisis de la evolución de los modelos de procesos y metodologías existentes en la inteligencia de negocios, temática extrapolable al dominio de la minería de datos. Como conclusión de la investigación, se determina que, a partir del 2000, la mayoría de las propuestas estructuraban sus fases (Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación E Implementación) [40] a partir del modelo de proceso CRISP-DM.

En adición, el sitio kdnuggets.com (presidido por Gregory Piatetsky) ha desarrollado entre los años 2002 y 2014 cuatro encuestas respecto al uso de metodologías para el desarrollo de proyectos en minería de datos. En la Fig. 2.7 [41], se ilustran los resultados obtenidos en las últimas dos encuestas. A partir de la misma, CRISP-DM se instala como el estándar de facto [30], [42], confirmando su predominio con más de un 40% de los encuestados. Las siguientes más utilizadas son SEMMA (con el 13% aproximadamente) y KDD (con el 7% aproximadamente).

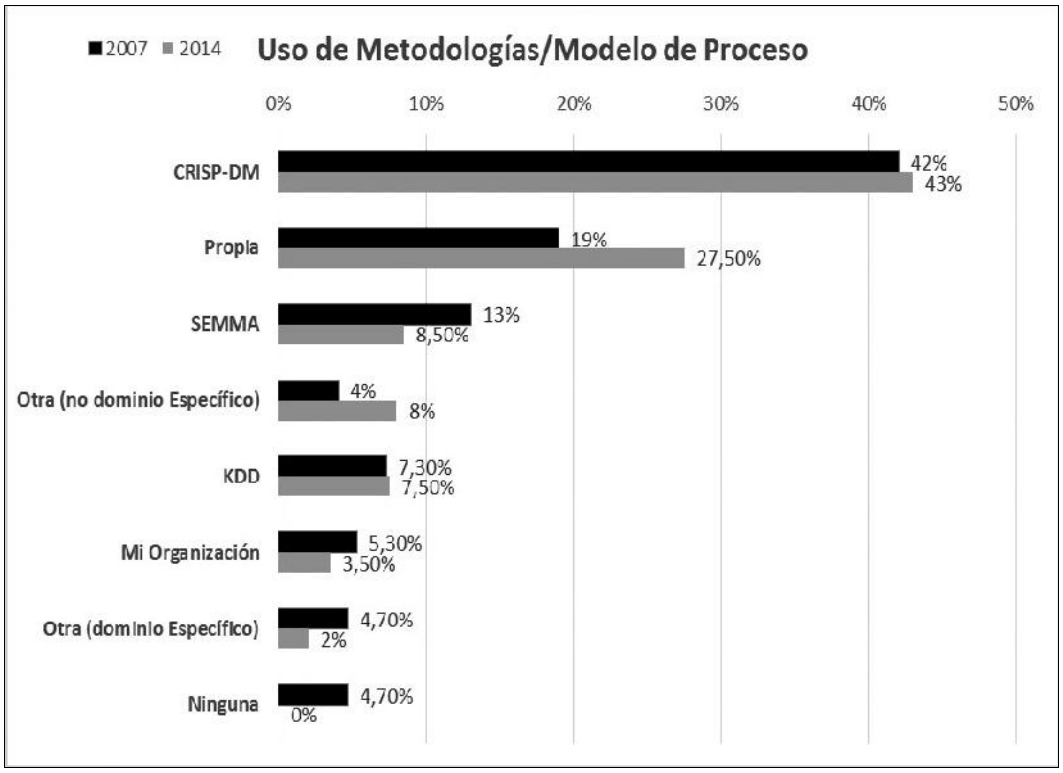


Figura 2.7. Uso de Modelo de Procesos y Metodologías - Resultados Encuestas 2007 y 2014:
Fuente: [41]

Finalmente, CRISP-DM es la única que se acerca al concepto de metodología. Es decir, incluye una guía detallada de cómo se realizan cada una de las tareas involucradas en el proceso, identifica los elementos de entrada y de salida, y las técnicas a aplicar en cada

actividad. Cabe aclarar que se diferencia de las otras propuestas que representan a un Modelo de Proceso, es decir, solo indica la estructura del proyecto y las actividades a desarrollar [41].

Argumentos como los expuestos precedentemente, inclinan a seleccionar CRISP-DM y adaptarla para cumplir los objetivos del presente Trabajo Final de Maestría.

Capítulo 3

Solución propuesta: procedimiento de integración de procesos de explotación de información y tecnología GIS

3. Solución propuesta: procedimiento de integración de procesos de explotación de información y tecnología GIS

En este capítulo se presenta el procedimiento propuesto como solución a la problemática en cuestión. En [43] se expuso el proyecto que guió el presente TFM. Se describen las etapas para integrar un software de escritorio GIS a los procesos de explotación de información. El procedimiento propuesto plantea llevar adelante un proceso de minería de datos sobre la base de datos geoespacial para identificar patrones y comportamiento delimitado por un determinado espacio geográfico, con el fin de que los procesos de análisis que proveen un sistema de información geográfica se potencie con los procesos de explotación de información.

El alcance de este procedimiento está determinado al hallazgo de posibles patrones de comportamiento, el cual utiliza técnicas de minería de datos según la problemática planteada, manipulando una herramienta de geolocalización la cual mostrará en un mapa los resultados obtenidos de aplicar estos algoritmos.

Para la aplicación de los procesos de explotación de información, se seleccionó la metodología CRISP-DM sustentada en el análisis expuesto en la sección 2.3, y como innovación se introduce el uso de GIS en la etapa de evaluación (Fig. 3.1). Los algoritmos seleccionados se determinaron a partir de la propuesta de procesos de explotación de información descripta en [17] y expuesta en la sección 2.2.

Las herramientas se escogieron a partir de experiencias similares reportadas en [3], el análisis del problema de negocio, las características de los datos o formato disponible y cantidad de registros reducida, la disposición de las técnicas de minería de datos requeridas y la experiencia en el uso de las mismas.

En el contexto del presente Trabajo Final de Maestría y en el dominio de validación elegido, no se contemplaron la aplicación de técnicas de procesamiento para grandes datos o entornos Big Data. Sin embargo, se considera de interés ampliar este estudio y se reconoce como un tema a evaluar en futuros trabajos.

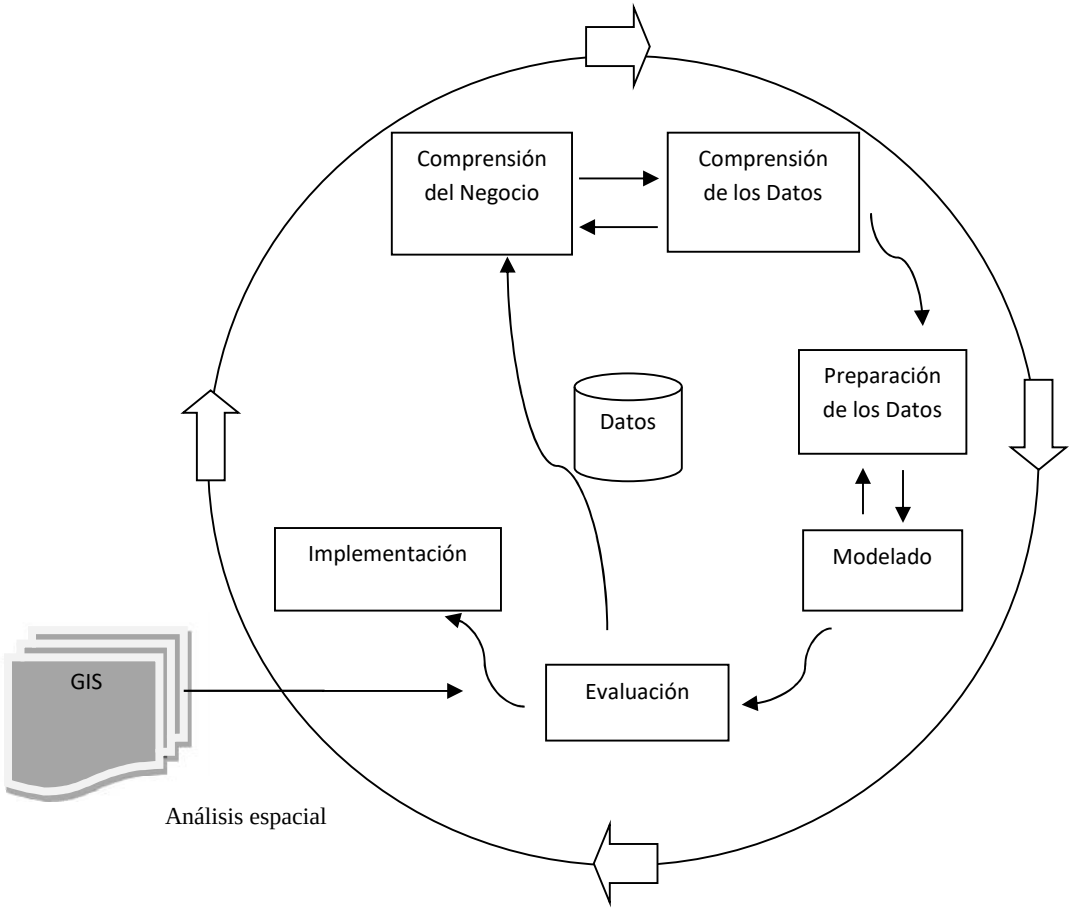


Fig. 3.1. Propuesta de integración de tecnología GIS en la metodología CRISP-DM:
Fuente: elaboración propia

El presente Trabajo Final de Maestría propone un procedimiento para aplicar técnicas de minería de datos sobre datos espaciales, a través de la integración de un sistema de información geográfica y una herramienta de explotación de información. A efectos de verificar este procedimiento, se seleccionó como caso de estudio la información delictiva proveniente de la base de datos del Sistema de Alerta Temprana para la ciudad de Corrientes registrada para el primer semestre del año 2017.

3.1. Requerimientos del procedimiento propuesto

El procedimiento propuesto se integra a través del acceso a los atributos de la capa vectorial geográfica incluidos en el archivo shape, se aplica la conversión de datos para la herramienta de explotación de información, y luego sobre estos datos se emplean las técnicas de minería de datos para identificar y extraer los patrones presentes, cuyo resultado se unifica con los datos espaciales originales utilizando un software GIS.

3.1.1. Requerimientos funcionales

- El sistema permite un análisis más detallado utilizando la herramienta espacial a partir de la información obtenida de aplicar técnicas de minería de datos.
- El procedimiento no realiza la integración de forma automática.

3.1.2. Requerimientos no funcionales

- El procedimiento sólo trabajará con archivos en formato shape y debe contener los cuatro archivos básicos: shp, shx, prj y dbf.
- La integración funciona solo si se mantiene el atributo o campo de unión (campo id del registro) de la capa geográfica en el proceso de explotación de información.

3.2. Etapas del procedimiento de integración propuesto

El procedimiento propuesto como solución a la problemática planteada se puede observar en la Fig. 3.2. Este diagrama de flujo principal visualiza el enfoque dividido en nueve procesos, en donde cada uno de ellos, a su vez, representa otra secuencia de procesos a ejecutar, junto con sus correspondientes entrada y salida de datos.

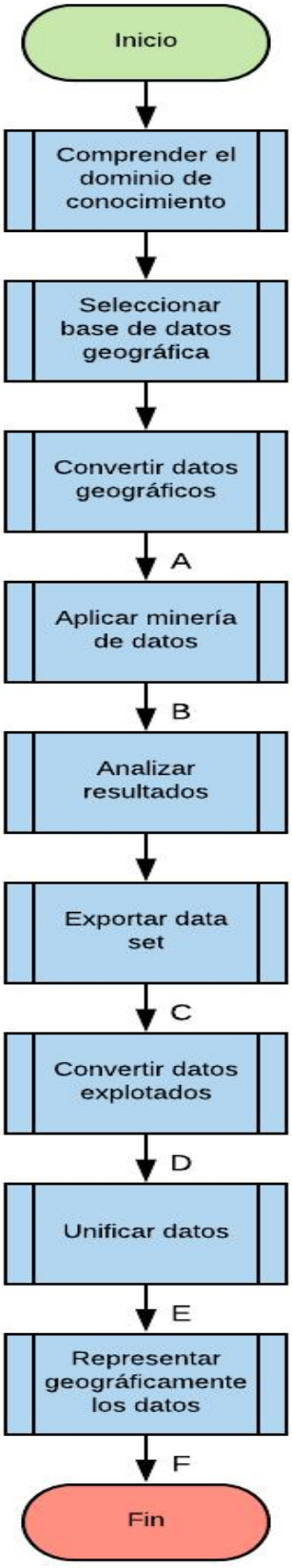


Fig. 3.2. Etapas del procedimiento propuesto: integración GIS y minería de datos:

Fuente: elaboración propia

3.2.1. Proceso de comprensión del dominio del conocimiento

En la Fig. 3.3 se puede observar a la izquierda el proceso de comprensión del dominio del conocimiento del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. El mismo incluye el análisis del campo de estudio del problema que se desea resolver, y la identificación de los objetivos que se requieren alcanzar.

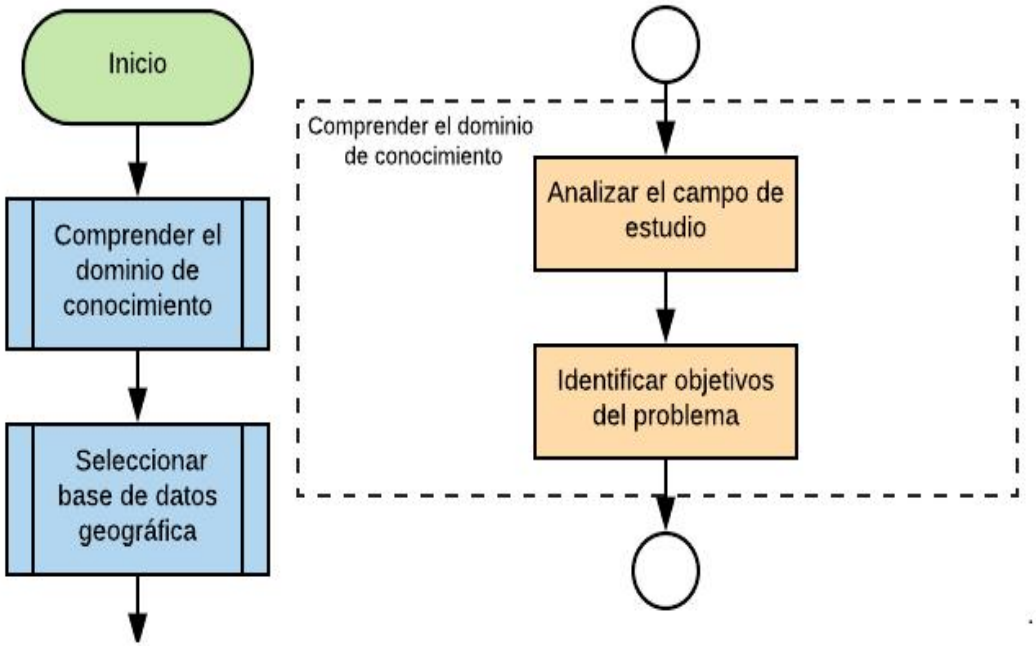


Fig. 3.3. Explotación del proceso de comprensión del negocio:
Fuente: elaboración propia

3.2.2. Proceso de selección de base de datos geográfica

En este proceso de selección de base de datos geográfica se hace uso de la tecnología GIS. A continuación, en la Fig. 3.4 se puede observar a la izquierda este proceso del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada una capa vectorial geográfica, luego se verifica que contenga los cuatro archivos básicos; shp, shx, prj y dbf., y se encuentre referenciada con un sistema de coordenadas adecuado. El siguiente proceso es la representación o visualización de la capa vectorial a través del uso de un software GIS.

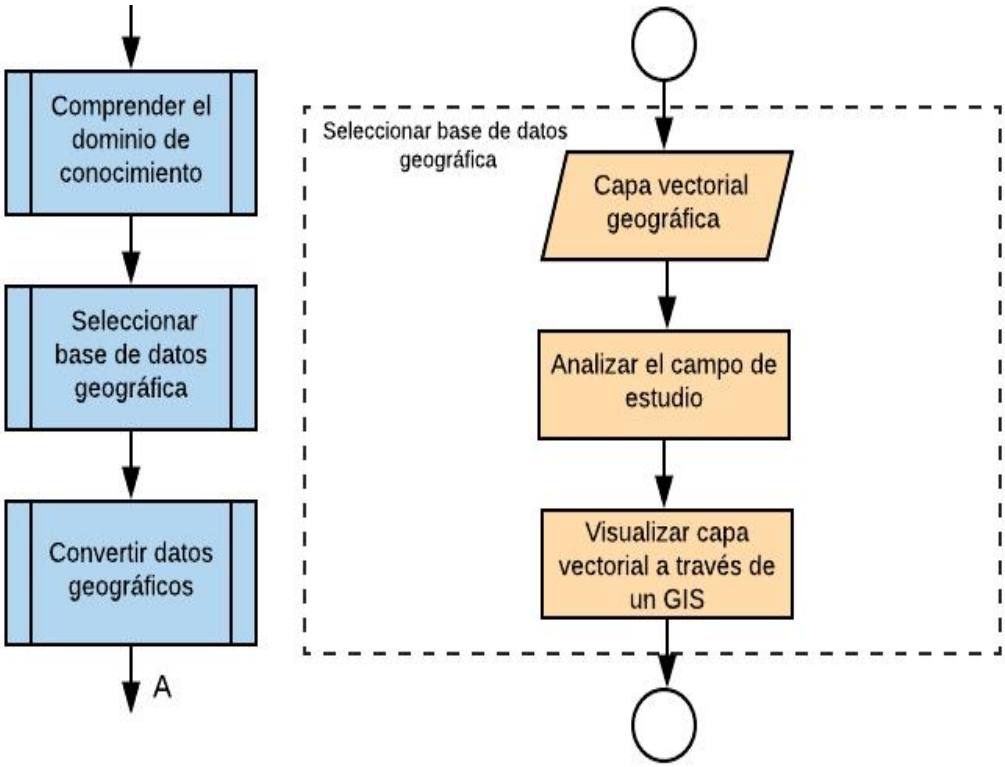


Fig. 3.4. Explotación del proceso de selección de base de datos geográfica:

Fuente: elaboración propia

3.2.3. Proceso de conversión de datos geográficos

En la Fig. 3.5 se puede observar a la izquierda el proceso de conversión de datos geográficos del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada un archivo con los atributos de la capa vectorial (archivo dbf), el siguiente proceso convierte estos datos a un archivo de salida con un formato valido para ser interpretado por una herramienta de explotación de información (A).

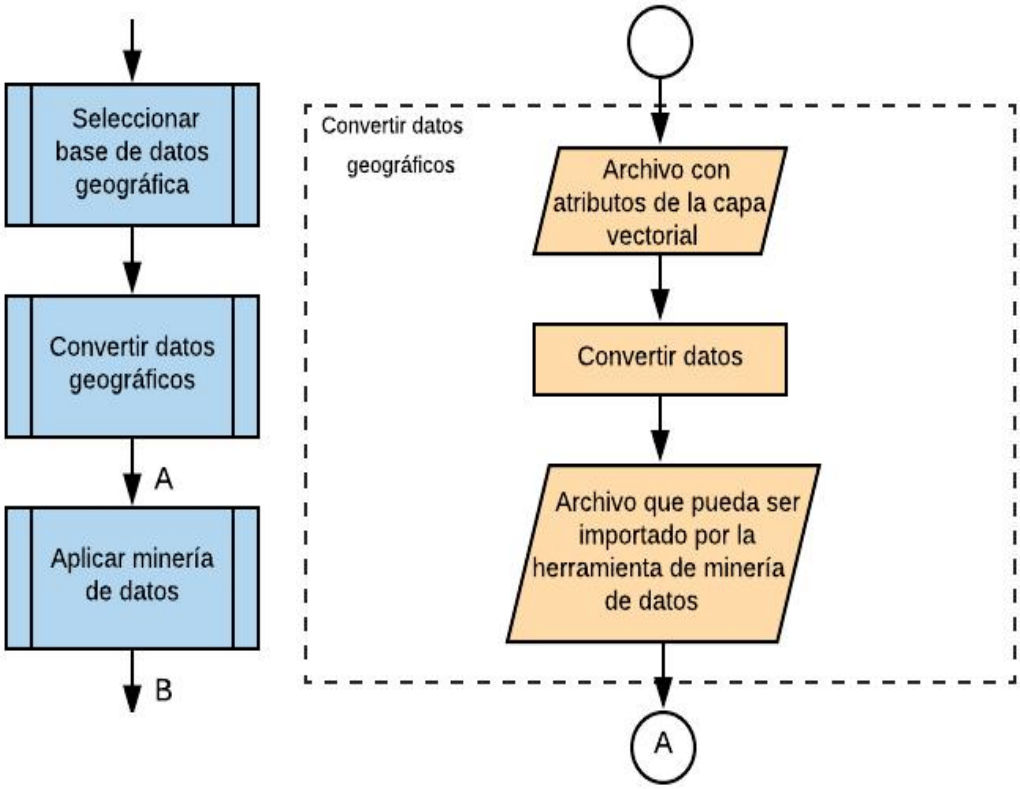


Fig. 3.5. Explotación del proceso de conversión de datos geográficos:
Fuente: elaboración propia

3.2.4. Proceso de aplicación de minería de datos

En este proceso se utiliza una herramienta de explotación de información para aplicar técnicas de minería de datos.

A través de la Fig. 3.6 se puede observar a la izquierda el proceso de aplicación de minería de datos del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada (A) un archivo cuyo formato pueda ser importado por la herramienta de minería de datos. El siguiente proceso representa la acción de aplicar los algoritmos de minería de datos, y cuyos resultados arrojan salidas graficas e informes obtenidos de emplear estos algoritmos (B).

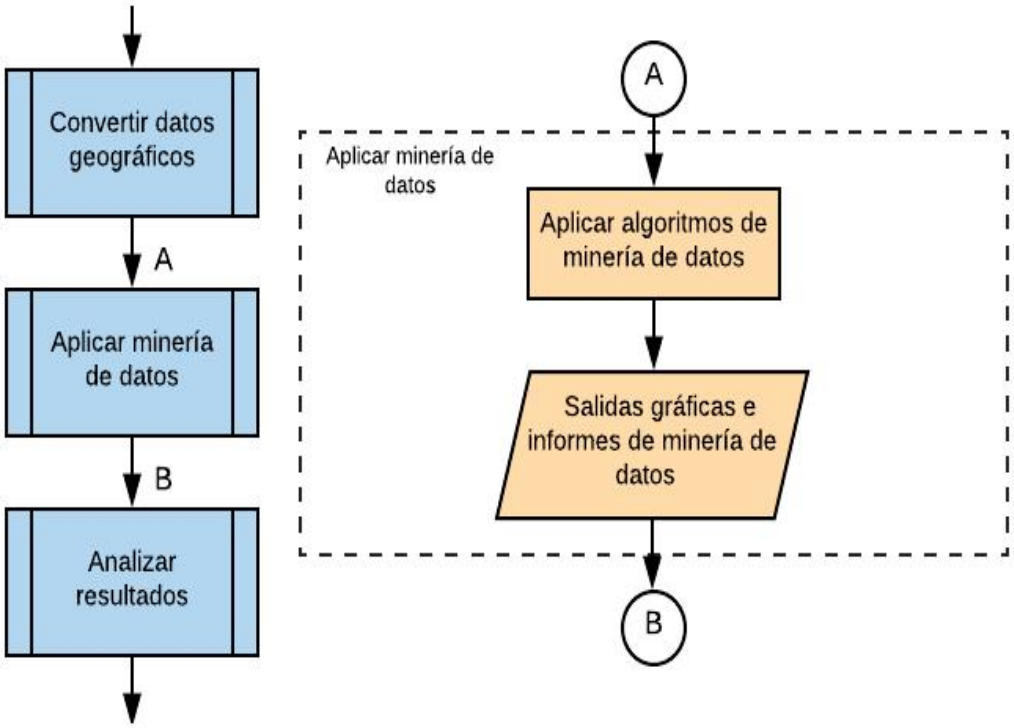


Fig. 3.6. Explotación del proceso de aplicación de algoritmos de minería de datos:
Fuente: elaboración propia

3.2.5. Proceso de análisis de resultados

En la Fig. 3.7 se puede observar a la izquierda el proceso de análisis de resultados del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. El mismo recibe como dato de entrada (B) las salidas graficas e informes obtenidos de emplear los algoritmos de MD, seguido del proceso de lectura de estos resultados, para su posterior análisis y evaluación.

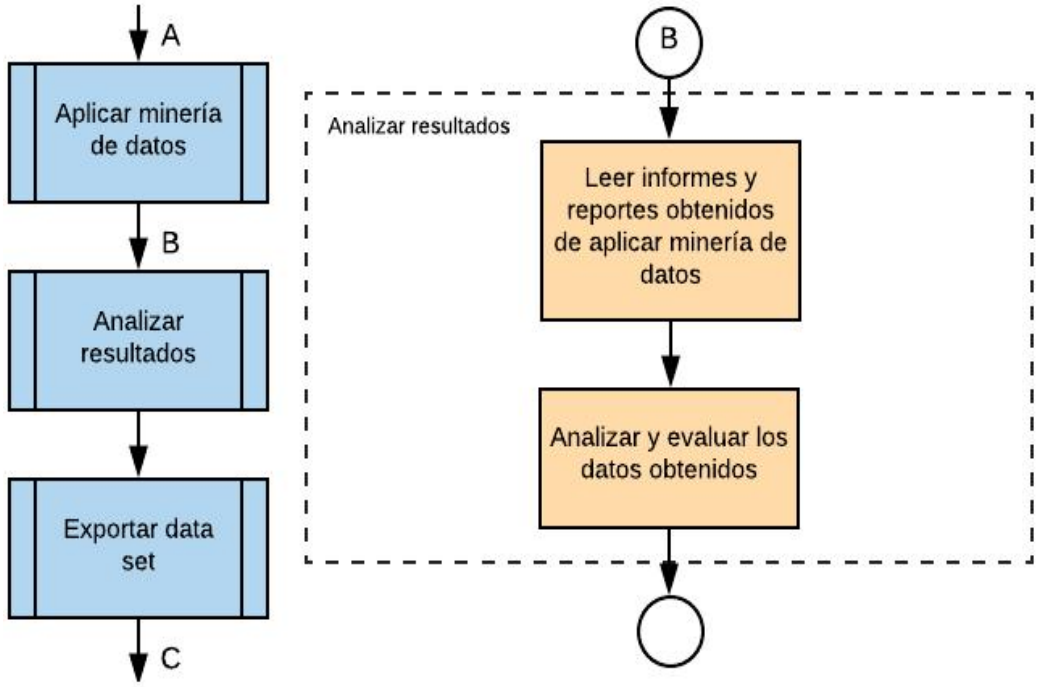


Fig. 3.7. Explotación del proceso de análisis de resultados:
Fuente: elaboración propia

3.2.6. Proceso de exportación de data set

Realizada la aplicación de los algoritmos y el análisis de los resultados que éstos producen, se procede a exportar el conjunto de datos o data set obtenidos por la herramienta de explotación de información.

En la Fig. 3.8 se observa a la izquierda el proceso de exportación de data set del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. La misma incluye la acción de exportar desde la herramienta de MD los datos con los resultados obtenidos de la aplicación de cada algoritmo y cuya salida genera un archivo con datos explotados (C).

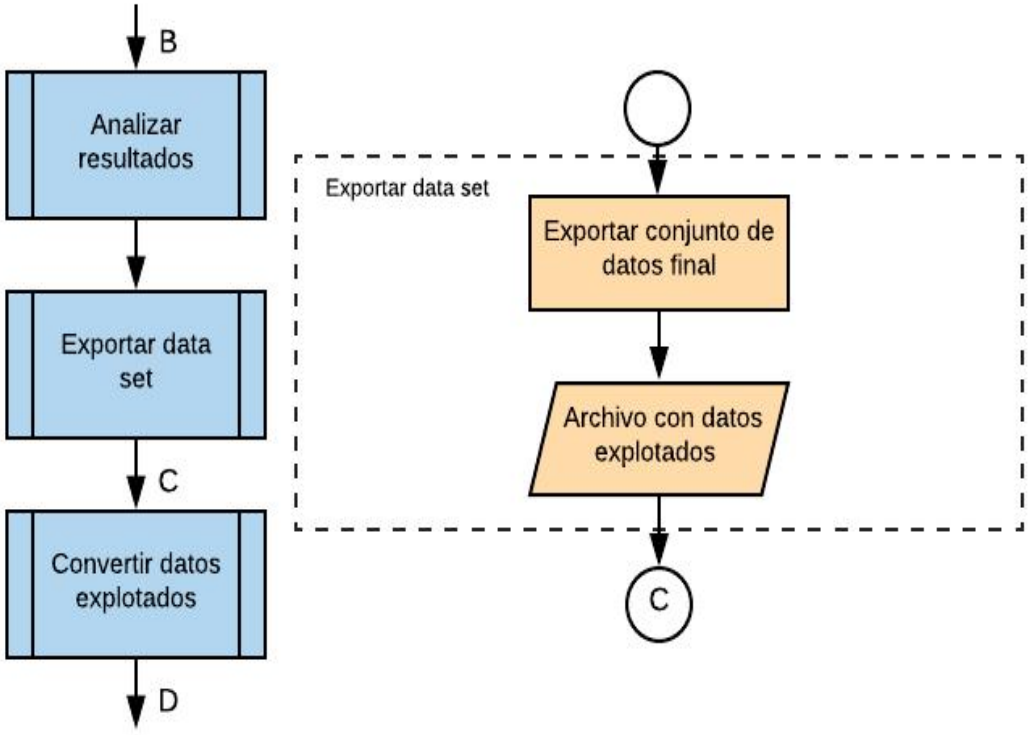


Fig. 3.8. Explotación del proceso de exportación del conjunto de datos:

Fuente: elaboración propia

3.2.7. Proceso de convertir datos explotados

Finalizado el proceso anterior, se requiere nuevamente convertir el archivo exportado por la herramienta de minería de datos, a un archivo cuyo formato sea interpretado por el software GIS.

En la Fig. 3.9 se observa a la izquierda el proceso de conversión de datos explotados del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada (C) un archivo con datos explotados, seguido del proceso de transformar o convertir estos datos a un archivo de salida que pueda ser interpretado e importado por un software GIS (D).

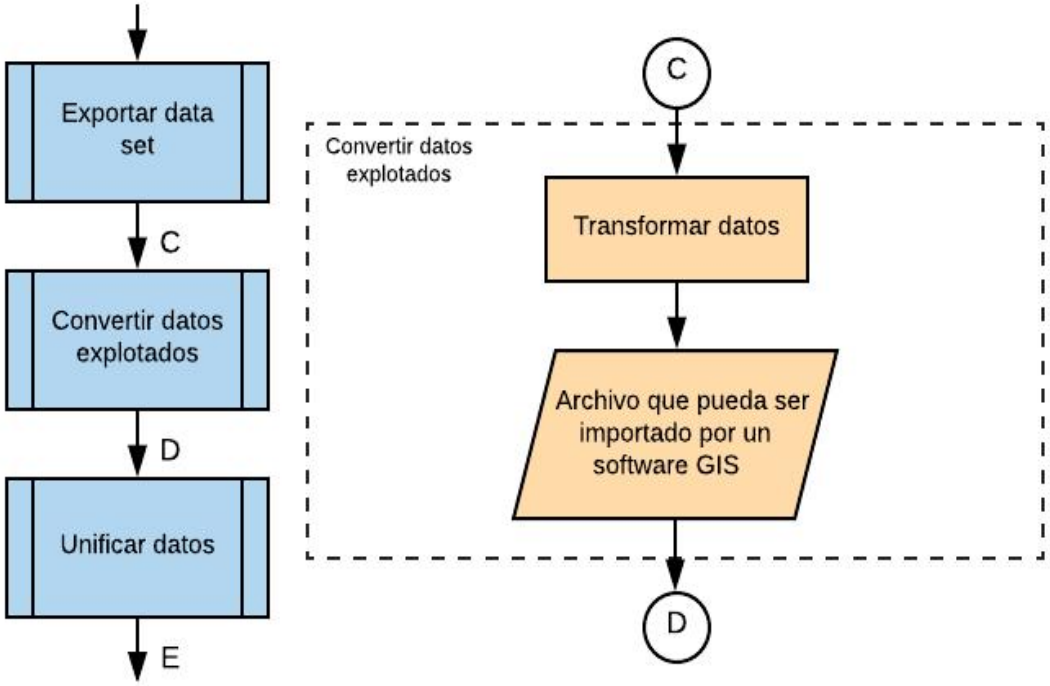


Fig. 3.9. Explotación del proceso de conversión de datos:
Fuente: elaboración propia

3.2.8. Proceso de unificación de datos

Este proceso es uno de los más importantes dado que realiza la unificación de datos entre la herramienta GIS y una herramienta de MD.

En la Fig. 3.10 se observa a la izquierda el proceso de unificación de datos explotados del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada (D) un archivo con datos explotados, en un formato que sea interpretado e importado por un software GIS. Luego, de forma conjunta con la capa vectorial geográfica original se realiza el proceso de unificación de ambos archivos, cuyo archivo de salida representa nueva capa vectorial espacial explotada (E) con la unión de datos, es decir contiene tanto los datos de la herramienta GIS como de la herramienta de MD.

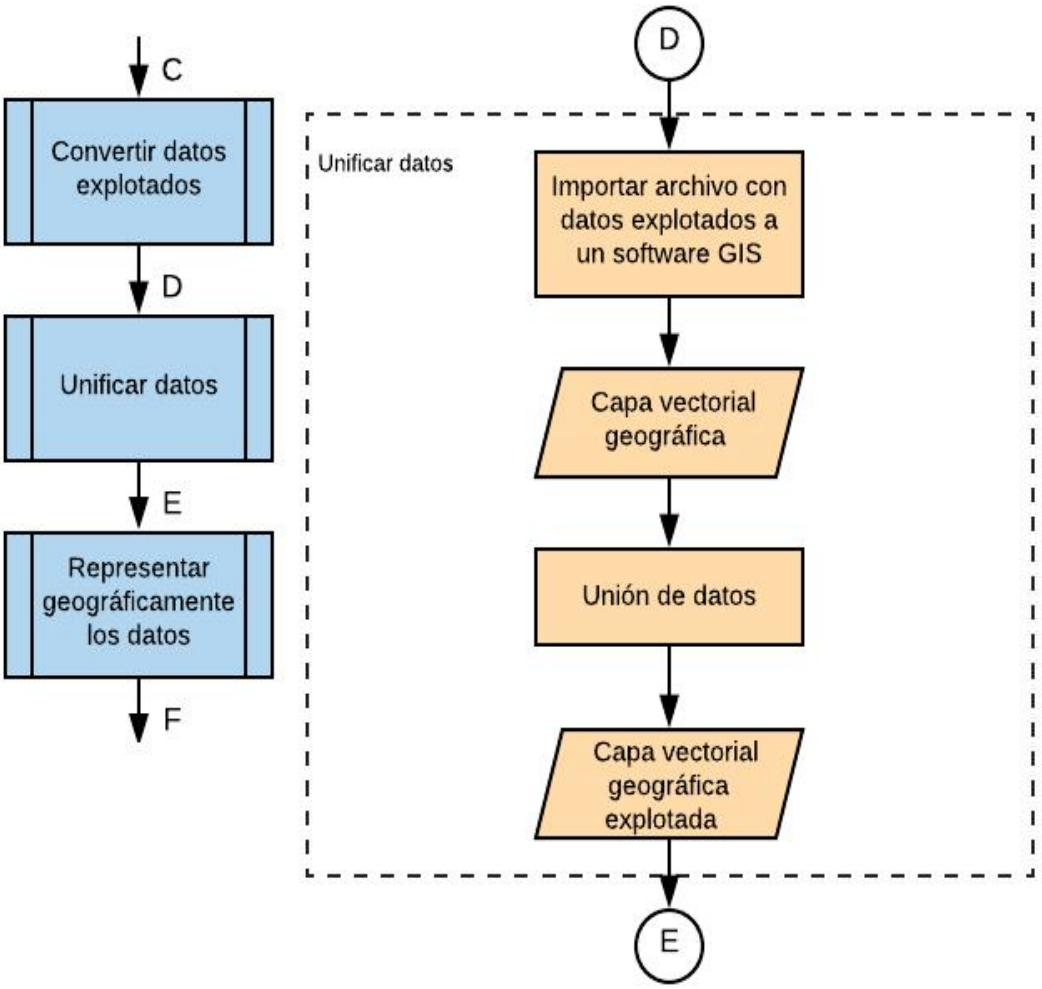


Fig. 3.10. Explotación del proceso de unificación de datos:
Fuente: elaboración propia

3.2.9. Proceso de representación geográficamente los datos

Realizada la unificación de datos, en la Fig. 3.11 se observa a la izquierda el proceso de representación geográfica de datos del diagrama general (Fig. 3.2) que se desea explotar, y a la derecha se visualizan las distintas operaciones de este proceso, que contiene, como primer proceso, al mismo proceso que se quiere explicar. Se recibe como dato de entrada (E) una capa vectorial geoespacial explotada, luego se representa de forma espacial utilizando un software GIS para su interpretación definitiva, y cuya salida genera mapas geográficos que visualizan los resultados de aplicar minería de datos (F).

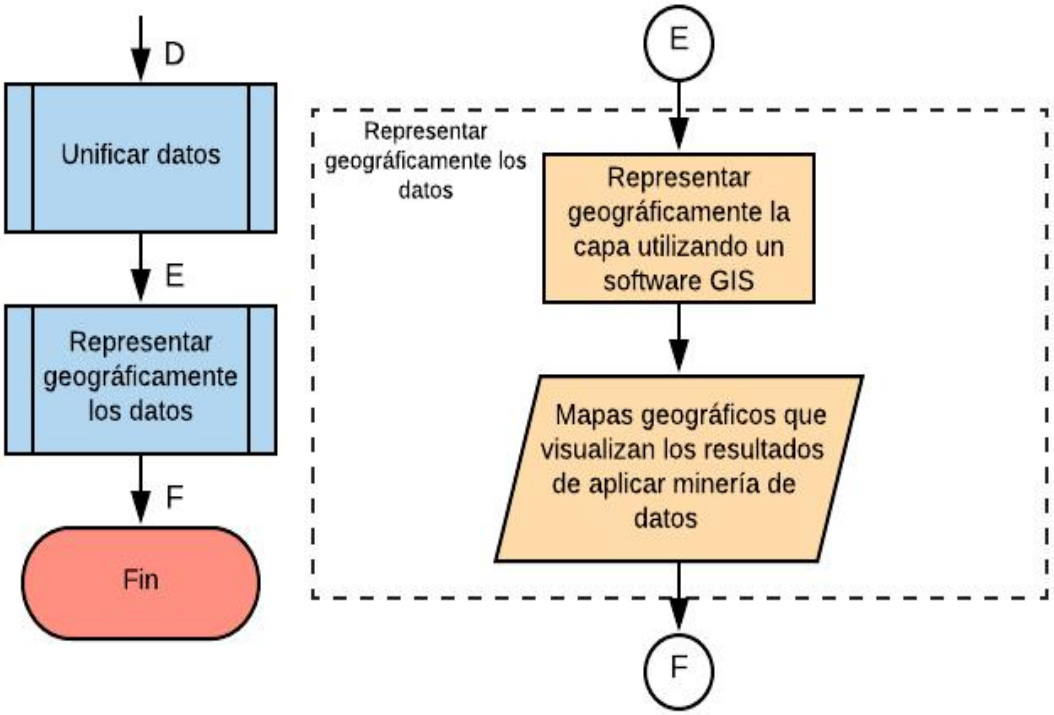


Fig. 3.11. Explotación del proceso de representación de datos geográficos explotados:

Fuente: elaboración propia

Capítulo 4

Delimitación del problema: minería de datos y tecnología GIS para el análisis delictivo.

4. Delimitación del problema: minería de datos y tecnología GIS para el análisis delictivo.

Este capítulo presenta la problemática que se intenta resolver y su delimitación. Se realiza un breve análisis sobre la minería de datos, la tecnología GIS y la integración de ambas para el análisis criminal.

La inseguridad es un problema creciente y complejo en la sociedad actual. La delincuencia en el país y principalmente las situaciones de robos o hurtos [44], se han convertido en una verdadera preocupación que exige el desarrollo e implementación de políticas conducentes a la prevención y detención del delito.

En Argentina, la Dirección Nacional de Estadística Criminal (DNEC) dependiente de la Subsecretaría de Estadística Criminal da cuenta de las estadísticas criminales ocurridas durante el año 2017 con respecto a los delitos contra la propiedad. Según el Sistema Nacional de Información Criminal (SNIC), los hechos más sobresalientes son los siguientes, i) la tasa de robo (que disminuyó un 8% entre 2016 y 2017, pasando de 1.074 a 989 hechos cada 100.000 habitantes, y ii) la tasa de hurto que descendió un 7% entre 2016 a 2017, es decir, paso de 648 a 601 hechos registrados cada 100.000 habitantes. Si bien las tasas de robo y hurto presentan reducciones en los últimos años, estas cifras son las más leves (- 8% y -7% respectivamente) respecto a la disminución de otras tasas criminales, como, por ejemplo, la tasa de víctimas de homicidios dolosos que descendió un 14% entre 2016 a 2017 (-14%) [44].

Siguiendo el análisis anterior, y tal como se observa en la Tabla III, se listan los tres delitos ocurridos con mayor frecuencia por provincia a nivel nacional. Se puede afirmar que los robos representan los hechos delictivos con mayor ocurrencia en Argentina, seguido del delito de hurto y de amenazas.

Tabla III

Lista de los tres delitos más frecuentes por provincia y a nivel nacional. Año 2017:

Fuente: [44]

Provincia	1º más frecuente	2º más frecuente	3º más frecuente
Buenos Aires	Robos	Hurtos	Amenazas
Catamarca	Robos	Amenazas	Lesiones Dolosas
Chaco	Hurtos	Robos	Lesiones Dolosas
Chubut	Robos	Amenazas	Hurtos
CABA	Robos	Hurtos	Lesiones Dolosas
Córdoba	Robos	Otros delitos contra las personas	Hurtos
Corrientes	Robos	Amenazas	Otros delitos en Leyes Especiales
Entre Ríos	Otros delitos en Leyes Especiales	Robos	Hurtos
Formosa	Hurtos	Robos	Lesiones Dolosas
Jujuy	Robos	Hurtos	Otros delitos contra la propiedad
La Pampa	Hurtos	Amenazas	Robos
La Rioja	Lesiones en siniestros viales	Hurtos	Robos
Mendoza	Robos	Hurtos	Amenazas
Misiones	Hurtos	Robos	Amenazas
Neuquén	Robos	Hurtos	Otros delitos contra la propiedad
Rio Negro	Robos	Hurtos	Otros delitos contra la propiedad
Salta	Amenazas	Robos	Hurtos
San Juan	Hurtos	Robos	Lesiones Dolosas
San Luis	Amenazas	Robos	Lesiones Dolosas
Santa Cruz	Robos	Amenazas	Hurtos
Santa Fe	Robos	Amenazas	Lesiones en siniestros viales
Santiago del Estero	Hurtos	Robos	Lesiones Dolosas
Tierra del Fuego	Otros delitos en Leyes Especiales	Amenazas	Hurtos
Tcumán	Amenazas	Robos	Lesiones Dolosas
País	Robos	Hurtos	Amenazas

En el análisis de delitos criminales, el descubrimiento de patrones significativos ha brindado la posibilidad de obtener datos de interés para interpretar y adecuar este conocimiento en la definición de los planes de prevención requeridos. La aplicación de diversas técnicas de minería de datos sobre el campo criminal se ha convertido en una herramienta con un gran potencial que permite diseñar estrategias específicas para esta área, resultando en un proceso automático de extracción de conocimiento útil [45].

En materia de búsqueda de antecedentes, se ha realizado una revisión sistemática de la literatura o RSL (ver Anexo 1) en torno a la aplicación e integración de técnicas y herramientas de minería de datos con tecnología GIS para el hallazgo de patrones delictivos [46], esta RSL fue elaborada en el año 2017 con la finalidad de generar

información que sustente la definición del proyecto de TFM, y cuyos resultados permiten concluir que los registros criminales son fundamentales para el diseño de políticas y planes de prevención del delito. Se propone a futuro replicar la RSL con miras a propiciar estudios longitudinales en torno al uso de tecnología GIS y de minería de datos orientadas a problemáticas gubernamentales, en este caso detección de delitos y hurtos.

En otros estudios [47], se ha determinado como principales técnicas de minería de datos para identificar o prevenir el delito efectivo las técnicas de agrupamiento, de clasificación y de asociación. Del mismo modo, en [48] se ha realizado una revisión de los últimos aplicativos de minería de datos sobre el análisis criminal y se listan los principales usos de minería de la información para identificar y prevenir delitos. Se menciona el sistema Coplink, como una de las implementaciones más exitosas de aplicación de técnica de clustering, la creación de herramientas que juegan un papel importante en el análisis de la criminalidad para la extracción de factores o la creación de perfiles digitales de los delincuentes, y el conocido sistema Crimeless Explorer que combina varias técnicas de minería de datos para mejorar el análisis de vínculos para apoyar en las investigaciones de delitos.

Otros trabajos como [49] aplican la técnica de agrupamiento de minería de datos con GIS para identificar criminales y puntos de acceso mediante el análisis de patrones espaciales. De la misma forma, se hace mención a otra herramienta de minería para combatir el análisis de delitos, una herramienta que combina GIS y la técnica de clustering, la misma se adoptó para la predicción diaria de delitos en India, como se describe en [50].

En Argentina, una de las medidas más utilizadas para combatir la delincuencia es la creación del Sistema de Alerta Temprana (SAT). Esta iniciativa del Ministerio de Justicia y Derechos Humanos es un sistema informático que almacena y concentra datos criminales referentes a los distintos tipos de delitos ocurridos en Argentina, y a partir del cual se realiza un análisis estadístico de la información. Actualmente se desconoce el uso de técnicas o herramientas de minería de datos aplicadas a estos datos.

Por otro parte, en los últimos años, el Instituto Tecnológico de Buenos Aires (ITBA) de forma conjunta con otras Universidades de Buenos Aires, realizan varios estudios enfocados en el campo delictivo. Estos trabajos dan cuenta de la importancia de aplicar

la minería de datos para la exploración y detección de patrones delictivos en Argentina y asistir así en propuestas sobre inteligencia criminal [51], [52].

Otro proyecto de relevancia es la creación del Mapa del Delito de la ciudad Autónoma de Buenos Aires, que representa los hechos delictivos de forma georreferenciada mediante el uso de un mapa interactivo usando GIS. Así, este mapa constituye la base cartográfica de la ciudad de Buenos Aires, y en el cual se observan las zonas con mayor número de ocurrencia de delitos [53].

Del mismo modo, se hace mención a un trabajo comprendido en un proyecto de investigación de la Universidad Nacional de Lanús (UNLa). Este estudio consistió en desarrollar una extensión de software GIS para integrarlo con una herramienta de minería de datos [3]. Actualmente esta aplicación no se encuentra disponible, lo cual dificulta su evaluación y utilización.

De acuerdo con lo expuesto en el análisis anterior (Tabla III), se puede afirmar que en la ciudad seleccionada para validar el procedimiento el robo representa el hecho criminal con mayor porcentaje de ocurrencia en comparativa con otros delitos. Por lo expuesto disponer de herramientas informáticas que automaticen la generación de información podría aportar en la toma de decisiones a las fuerzas de seguridad. La propuesta del procedimiento para combinar tecnologías GIS y MD es innovadora dado que se podría tratar como un instrumento tecnológico para reducir la criminalidad a nivel provincial. Así, se presenta un enfoque que permite adaptar y extender las mencionadas técnicas de minería de datos sobre una base de datos georreferenciada, diseñada y construida con un sistema GIS. Además, resulta de interés visualizar en un mapa de la ciudad las reglas generadas a partir de la aplicación de técnicas de minería de datos utilizando herramientas de georreferenciación. Es decir, integrar esta tecnología permite entender la distribución espacial de determinados hechos en un contexto geográfico específico.

Precisamente, la introducción de estas tecnologías mejoraría los tiempos del proceso de investigación de los posibles delitos y aportaría a la sistematización de las tareas manuales de los agentes policiales, proporcionando una alternativa mediada por las TIC para generar información clara, precisa y confiable asociada a los delitos.

Por lo expuesto, se propone integrar GIS y técnicas de minería de datos como herramientas claves para la detección y predicción de estos hechos delictivos. Para verificar la propuesta, se aplicaron estas tecnologías sobre una base de datos

georreferenciada con información de ocurrencias de delitos de robo y hurto cometidos en la ciudad en el primer semestre del año 2017, con el propósito de realizar un análisis y hallazgo de los patrones más relevantes de estos hechos.

En el contexto de validación, este trabajo contribuye en el ámbito delictivo de la ciudad elegida, ofrece la oportunidad de aprovechar e integrar de manera sencilla distintos aspectos claves en las áreas de conocimiento de minería de datos y de las tecnologías de georreferenciación aún no utilizados en la provincia.

Asimismo, el uso del SIG posibilitó la generación de mapas del delito, éstos ofrecen visualización respecto a la intensidad con que los hechos delictivos se producen, es decir, las zonas calientes en donde se presenta un mayor nivel de delitos.

El Trabajo Final de Maestría permitió profundizar sobre distintos conceptos de la disciplina entre los que se mencionan: explotación de datos basada en sistemas inteligentes, aprendizaje automático, uso de capas geográficas de información, geolocalización de objetos espaciales, entre otros.

Su real implementación producirá información resumida de fácil entendimiento y potencialmente transferible a los decisores de las organizaciones públicas de la ciudad seleccionada en este TFM destinadas a combatir la inseguridad, de modo que puedan incrementar su productividad desde este enfoque de análisis de datos.

Capítulo 5

Validación del procedimiento

5. Validación del procedimiento

En este capítulo se verifica el procedimiento propuesto sobre un caso de estudio orientado a la detección de hechos delictivos de robos y hurtos en la ciudad seleccionada como objeto de estudio en el primer semestre del año 2017. Se presenta el contexto del caso de estudio (sección 5.1.) y la validación del procedimiento aplicada a la misma (sección 5.2).

5.1. Contexto de validación

Desde hace varios años, la Dirección Nacional de Política Criminal (DNPC) lleva adelante la implementación del Sistema Nacional de Estadística Criminal (SNEC) cuya finalidad es centralizar la información sobre el delito en la Argentina. La implementación de este sistema requiere de diferentes fuerzas del país (policiales, federales, nacionales, entre otras) que envían mensualmente los resultados de su actividad sistematizados en dos planillas dirigidas al Sistema Nacional de Información Criminal o Hechos Delictuosos (SNIC) y al Sistema de Alerta Temprana (SAT).

En líneas generales, mientras que en la planilla de hechos delictivos se registran las cantidades por tipo de delito y cantidad de víctimas en algunos delitos en particular, en la planilla del SAT se recaba una mayor cantidad de información específicamente sobre los delitos contra la propiedad, homicidios culposos en hechos de tránsito, homicidios dolosos y suicidios [54].

Desde su creación, el SAT se utiliza por todas las provincias del territorio nacional argentino. Por ello, se considera como la fuente primaria para verificar este procedimiento.

El análisis de la información se centró específicamente sobre los delitos contra la propiedad cometidos en la zona urbana de la ciudad del primer semestre del año 2017, esta información contiene datos de denuncias o tentativas de robo y hurto realizadas por las víctimas en las distintas jurisdicciones policiales ubicadas en la ciudad, que abarca un total de 143 barrios.

5.2. Caso de validación: Comportamiento de delitos de robo y hurto en la ciudad de Corrientes

En esta sección se aplican los procesos definidas del procedimiento (sección 3.2.) para verificar la propuesta sobre un conjunto de datos recibidos del SAT. En la descripción se particulariza el procedimiento mencionando la aplicación de QGIS como software GIS y Tanagra como herramienta de minería de datos.

5.2.1. Comprender el dominio del conocimiento

En los últimos años, los delitos reportados y descubiertos por la policía en esta ciudad evidencian la necesidad de hallar diferentes técnicas y metodologías para lograr su reducción. Por ello, se presenta la urgencia de identificar la forma en que se presentan los delitos de robo y de hurto con el objetivo de hallar patrones de comportamiento de ocurrencias y probables lugares dentro de la ciudad en donde existe la mayor cantidad de hechos.

5.2.2. Seleccionar base de datos geográfica

Para este segundo proceso del procedimiento se seleccionó la base de datos geográfica y se escogió la aplicación de escritorio QGIS como visor de datos GIS, la elección de la misma se justificó en la sección 2.1.4.

Como primer paso se obtuvo la base de datos espacial, y dado que los datos criminales obtenidos del SAT no se encontraban georreferenciados, se realizó un proceso previo utilizando el software QGIS para localizar geográficamente los datos de acuerdo a las ubicaciones y zonas de ocurrencia del hecho, utilizando los atributos calle y altura de la base de datos disponible. Como proceso final se obtuvo una capa espacial formada por los 4 archivos básicos (shp, shx, prj, dbf) y georreferenciada en el Sistema de Referencia de Coordenadas PORGAR 94.

De forma complementaria, se manipuló información espacial suministrada por la Infraestructura de Datos Espaciales de la Municipalidad de la Ciudad de Corrientes o IDEMCC, a través de servicios WFM que contiene la base cartográfica digital de la ciudad de Corrientes referente a elementos básicos como ser; manzanas, barrios, plazas, infraestructura de seguridad (zonas de seguridad urbana municipal, comisarías,

jurisdicciones policiales), sitios con disposición de dinero (bancos, cajeros, centros de pago), sitios de recreación (bares, parrillas, restaurantes), etc.

En la Fig. 5.1 se visualizan las capas base del servicio WMS de la IDEMCC utilizando el software QGIS para representarlas, junto con los diferentes puntos geográficos delictivos cometidos en la ciudad, los cuales se visualizan en color rojo.

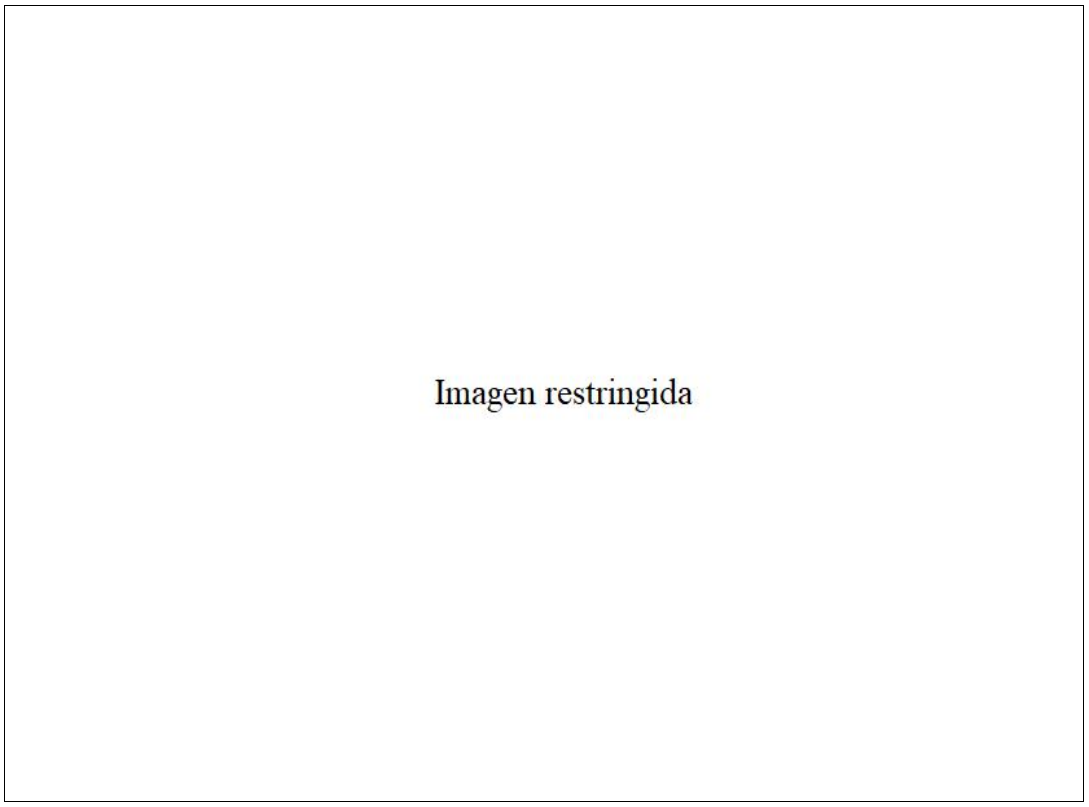


Fig. 5.1. Capas bases WMS de la IDEMCC y capa vectorial de delitos identificados para el contexto de validación: Fuente: elaboración propia

La representación visual a través del mapa de los distintos puntos según los atributos de barrios de la ciudad, jurisdicción policial y tipo de delito (los puntos color naranja indican un robo y los de color azul un hurto) se pueden observar a través de las Figs. 5.2, 5.3 y 5.4.

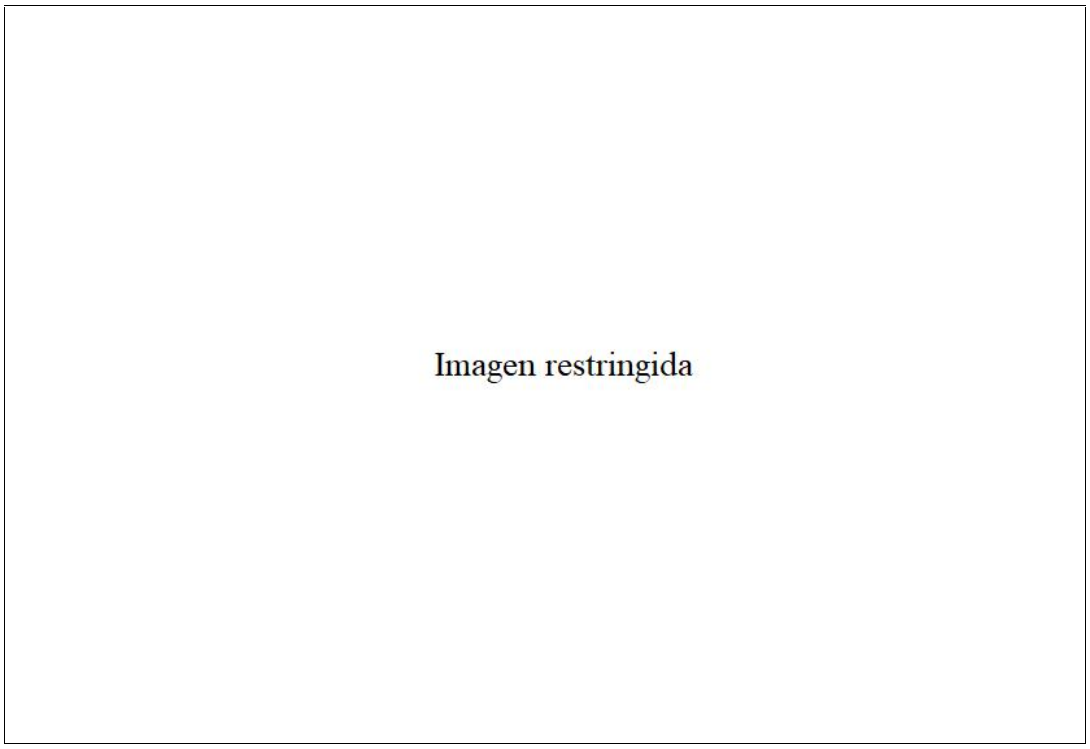


Fig. 5.2. Visualización de puntos delictivos por barrio de la ciudad de Corrientes:
Fuente: elaboración propia

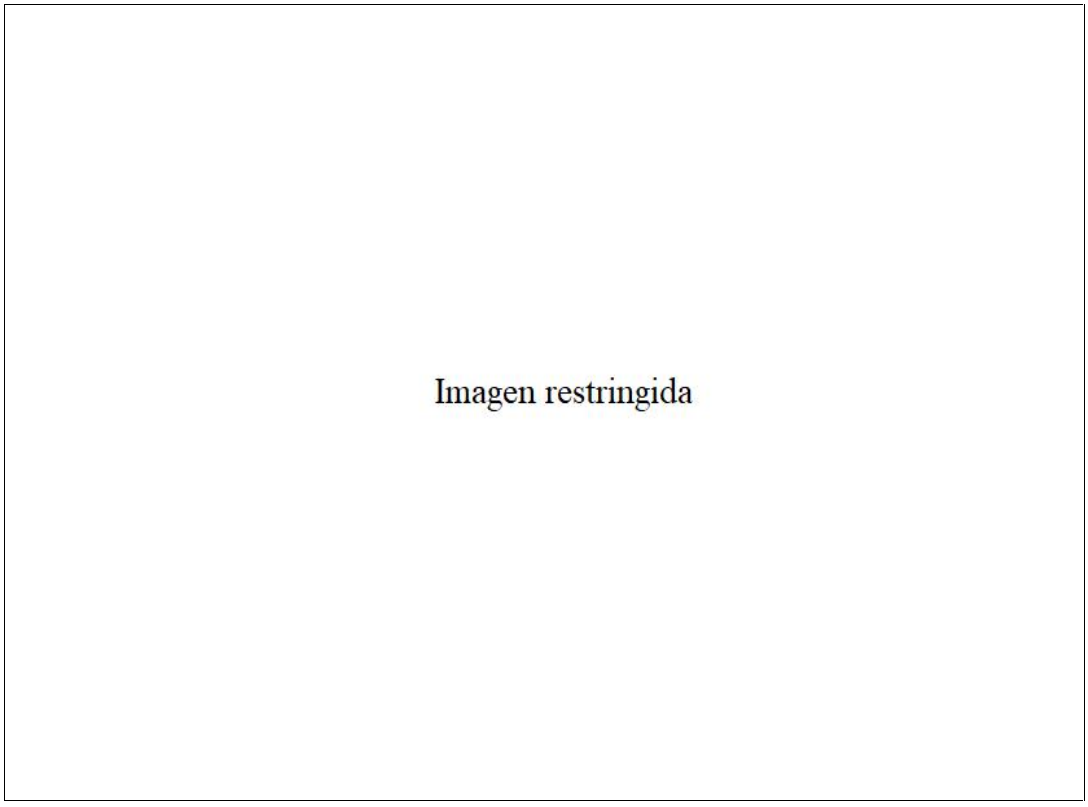


Fig. 5.3. Visualización de puntos delictivos por jurisdicción policial de la ciudad de Corrientes:
Fuente: elaboración propia

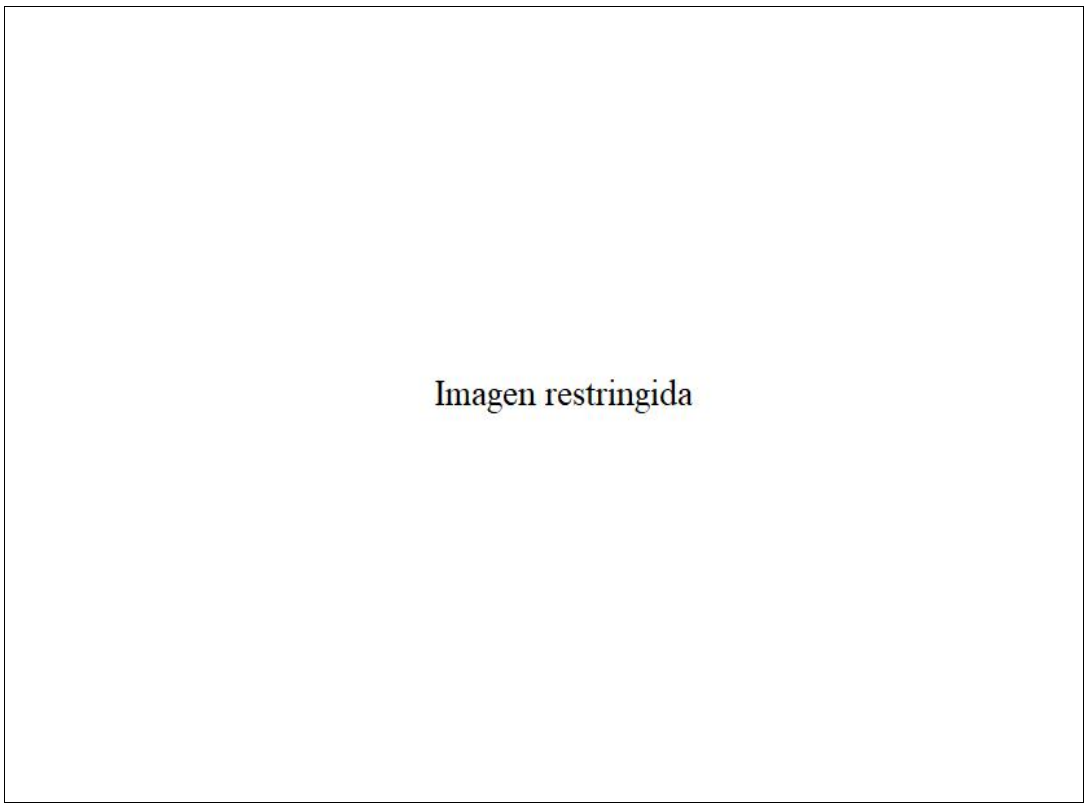


Fig. 5.4. Visualización de puntos delictivos por robo y hurto de la ciudad de Corrientes:
Fuente: elaboración propia

5.2.3. Convertir datos geográficos

En este tercer proceso del procedimiento, se transformó el archivo dbf de la capa vectorial de delitos obtenida en 5.2.2, a un archivo con extensión xls que resulte legible y pueda ser importado por la herramienta de minería de datos.

5.2.4. Aplicar minería de datos

En este cuarto proceso del procedimiento se requiere la aplicación de los algoritmos de minería de datos. Previamente se seleccionó como software la herramienta Tanagra, la elección de la misma se justificó en la sección 2.2.1. Tanagra acepta como archivo de entrada el formato xls, generado en el proceso anterior (sección 5.2.3), y un archivo de salida txt para ejecutar posteriormente la integración de datos con el software GIS.

A continuación, se detallan los objetivos del proyecto de minería de datos definidos en este caso de estudio:

- Como objetivo de minería de datos N° 1, se desea analizar sobre los datos de la base de datos del SAT, patrones de comportamiento relevantes a través de la

agrupación de casos. Y determinar los factores que caracterizan la ocurrencia de delitos de robo y hurto en la ciudad y cuales son aquellos con mayor incidencia dentro de cada grupo.

- Como objetivo de minería de datos N° 2, se desea establecer distintos grupos entre las zonas de mayor porcentaje de ocurrencia de delitos en la ciudad, con el fin de realizar un análisis detallado en dichas regiones.
- Como objetivo de minería de datos N° 3, se desea establecer distintos grupos entre las personas más propensas a sufrir un delito, en orden a comprender con mayor detalle cómo afectó el delito a dichas personas e identificar los factores predominantes en cada grupo.

El análisis de los resultados derivados de ejecutar los tres objetivos de minería de datos descriptos previamente, se detallan en el capítulo 6.

5.2.5. Analizar los resultados

En este quinto proceso del procedimiento, se presenta una síntesis del análisis de los resultados obtenidos al aplicar los diferentes tipos de procesos de explotación de información:

El uso de la minería de datos, y en particular los procesos de explotación de la información seleccionados, han demostrado ser, un medio eficaz para la detección del comportamiento de delitos, al tiempo que ofrecen información útil y efectiva para la toma de decisiones a futuro. Entre la información relevante se menciona: días de mayor porcentaje de ocurrencia de delitos (viernes, sábados y domingos), horarios más susceptibles (horarios de siesta y de madrugada), objetos con mayor número de sustracciones (objetos personales), lugares y tipos de armas más utilizadas durante el delito (vía pública y con arma blanca) y tipo de ataque más sufrido (arrebato).

La aplicación de técnicas de minería de datos para el hallazgo de zonas más peligrosas también ofrece información útil y necesaria. A modo de ejemplo, los barrios con mayor cantidad de delitos son: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6, barrio N° 7 y barrio N° 8.

Además, se identificaron los atributos más significativos resultado del proceso de agrupación con las víctimas que sufrieron los delitos: como el rango de edad y sexo de

las personas con mayor porcentaje a sufrir un robo o hurto (21 a 25 años y de sexo femenino).

La difusión y transferencia de estos resultados podrán tener impacto directo en cuestiones tratadas por organismos de seguridad, dado que el conocimiento producido se podría utilizar en el ámbito policial como una estrategia para apoyar la toma de decisiones.

5.2.6. Exportar data set

A partir de este proceso se exportó el conjunto de datos disponible de la herramienta Tanagra. Este set de datos contiene los resultados de aplicar cada proceso de explotación de información y cuya salida resulto en un archivo de texto (formato txt).

5.2.7. Convertir datos explotados.

En este proceso se realizó la conversión de datos de los resultados obtenidos de aplicar las técnicas de minería de datos obtenidos en la sección 5.2.6 (formato txt), a un archivo interpretable por la herramienta QGIS (archivo con formato xls).

5.2.8. Unificar datos

En este octavo proceso del procedimiento, se unificaron los datos explotados obtenidos en la sección 5.2.7 (archivo con formato xls) con la base de datos de delitos geográfica original.

En la Fig. 5.5, se observa el proceso de unión vectorial entre ambas tablas a través del software QGIS, el cual realiza la unificación de datos por medio de un campo en común (id_registro), seguido de la selección de campos que contienen los resultados de aplicar cada uno de los algoritmos de minería de datos (el campo Cluster_SOM_1 para la formación de clusters, pred_SpvInstance_1 para la caracterización de los grupos y pred_SpvInstance_2 para la ponderación de los atributos). Como resultado final se genera una nueva capa vectorial de delitos con datos explotados (archivos shp, shx, prj y dbf).

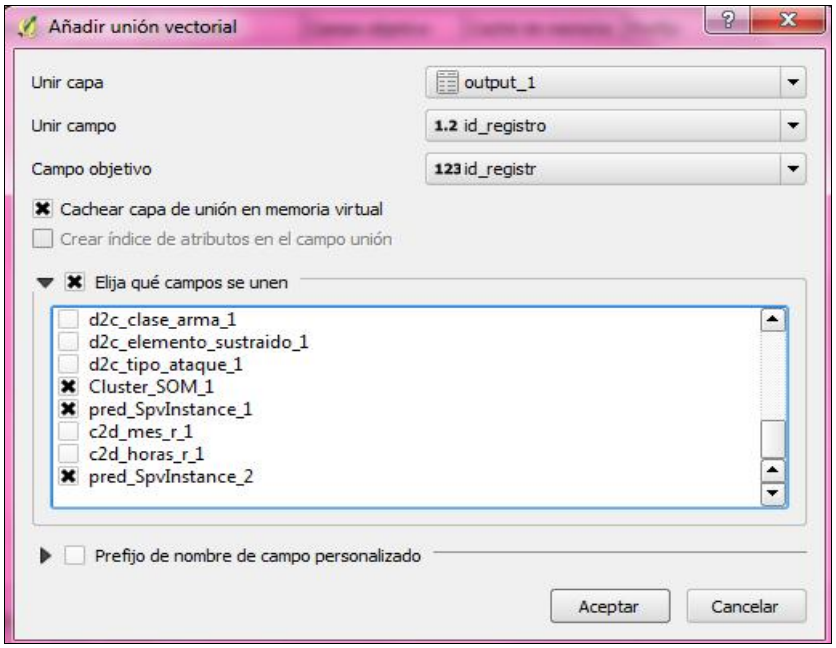


Fig. 5.5. Unión vectorial de capa geográfica de delitos con set de datos del Tanagra:

Fuente: elaboración propia

Se repite este proceso para todos los conjuntos de datos exportados desde la herramienta de MD.

5.2.9. Representar geográficamente los datos

A partir del archivo generado en la sección anterior 5.2.8, se usó la herramienta QGIS para visualizar esta información a través de mapas. Estos mapas son una representación gráfica que reflejan los resultados obtenidos de aplicar minería de datos sobre la base de datos geográfica.

5.2.9.1. Representación geográfica del objetivo de minería de datos Nº 1

Para la representación espacial de los clusters generados en este primer objeto de minería de datos, se utilizó como complemento la generación de un mapa de calor, una herramienta disponible en QGIS capaz de detectar las zonas con mayor concentración de puntos sobre una determinada ubicación geográfica. Estos mapas muestran distintos tonos de color rojo hacia colores más claros según la densidad de los puntos concentrados.

Identificadas las zonas, se analizó la modalidad del clúster c_som_1_1 presente en ellas, con el objetivo de caracterizar los delitos de acuerdo a los resultados derivados de la aplicación de los algoritmos de minería de datos.

En la Fig. 5.6, se aprecia que existe una mayor concentración de puntos sobre las zonas oeste, centro y norte de la ciudad. Los barrios que agrupan estos puntos: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6 y barrio N° 7. La modalidad de delitos para este clúster se caracterizó por el uso de armas mayoritariamente de tipo blancas, los cuales ocurrieron en la vía pública y a través del arrebato. Los días viernes se registraron mayor cantidad de hechos delictivos en el horario de siesta de 12:00 pm a 16:00 pm., en la mayoría de los casos se registraron robos de objetos personales.

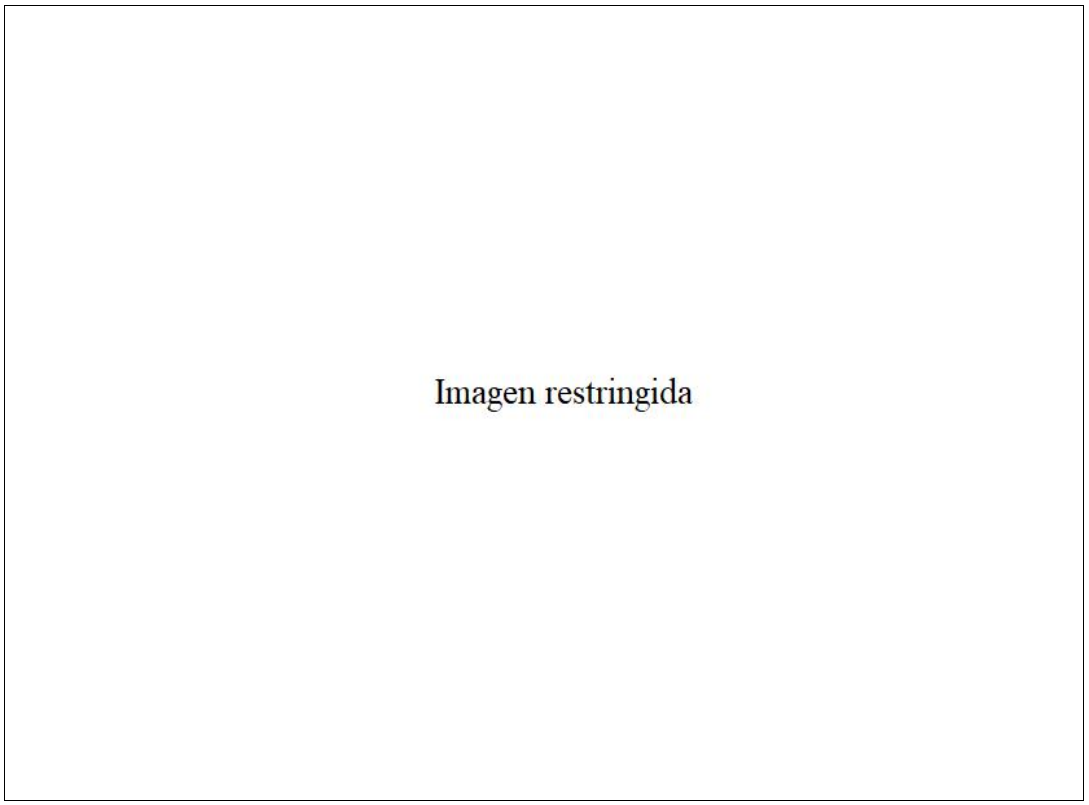


Fig. 5.6. Mapa de calor del clúster c_som_1_1 para caracterizar el comportamiento de delitos

Fuente: elaboración propia

En la Fig. 5.7, se visualiza una mayor concentración de hechos delictivos presentes sobre las zonas norte, oeste y centro de la ciudad ~~de Corrientes~~, correspondiendo al barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6, barrio N° 7, barrio N° 8 y barrio N° 9. Se observa, además, una gran confluencia de delitos sobre la calle N° 1.

Identificadas las zonas, se analizó la modalidad del clúster c_som_1_2 presente en ellas, el cual caracterizó la ocurrencia de delitos por el uso de armas blancas y cuyos robos se produjeron en la vía pública. Se registraron como los días con mayor cantidad de hechos

delictivos los sábados y domingos en el horario de siesta de 12:00 pm a 16:00 pm y de noche de 20:00 pm a 24:00 pm., en la mayoría de los casos se registraron robos de objetos personales en forma de arrebatado o forcejeo.

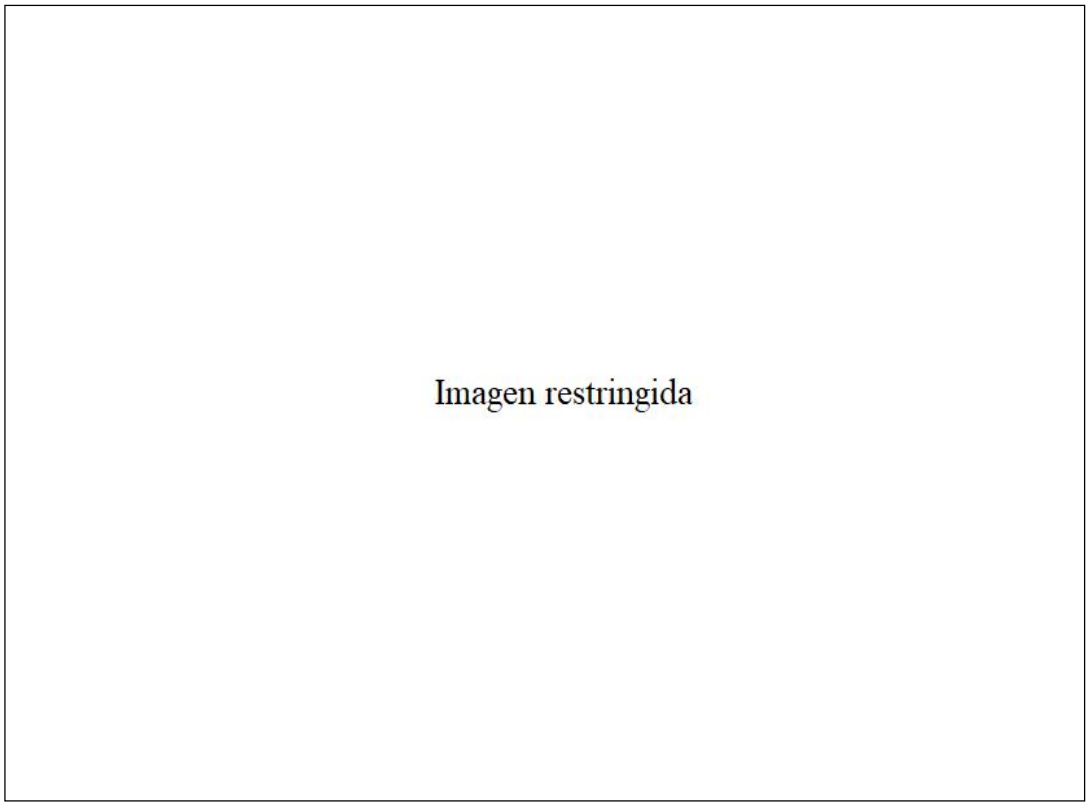


Fig. 5.7. Mapa de calor del clúster c_som_1_2 para caracterizar el comportamiento de delitos:
Fuente: elaboración propia

En la Fig. 5.8, se identifica que existe una mayor concentración de puntos sobre las zonas norte, oeste y centro de la ciudad. Estos puntos se agrupan en los siguientes barrios: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6, barrio N° 7, barrio N° 8, barrio N° 9, barrio N° 10 y barrio N° 11. Se observa, además, una gran confluencia de delitos sobre la calle N° 1 y calle N° 2. Este clúster presenta mayor cantidad de delitos agrupados (128 puntos en total) con respecto al resto de los clusters.

Identificadas las zonas, se precedió a analizar la modalidad del clúster c_som_2_1 presente en ellas, el cual caracterizó la ocurrencia de delitos por robos y hurtos producidos en la vía pública, en un domicilio particular o en el interior de un rodado. En la mayoría de los casos se registraron robos de objetos personales.

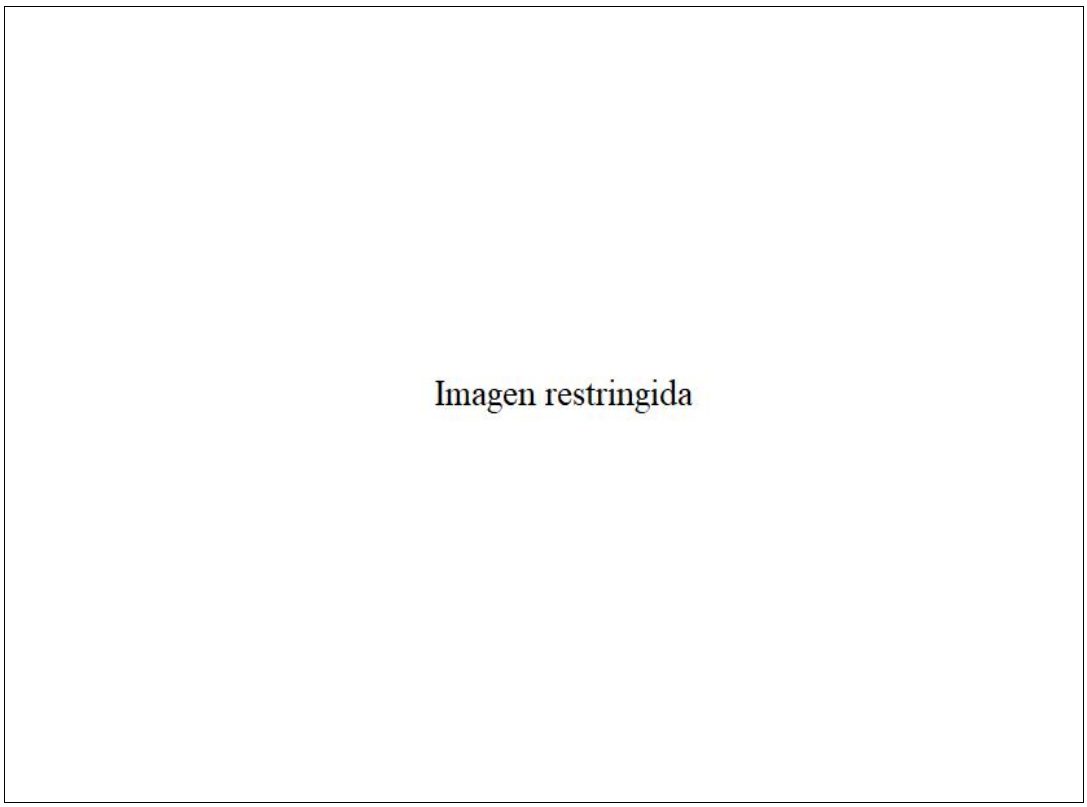


Fig. 5.8. Mapa de calor del clúster c_som_2_1 para caracterizar el comportamiento de delitos:
Fuente: elaboración propia

En la Fig. 5.9, se determina que existe mayor concentración de delitos sobre las zonas norte, oeste y centro de la ciudad. Los barrios que agrupan estos puntos son: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6 y barrio N° 7. Este clúster presenta la menor cantidad de delitos agrupados (61 puntos en total) con respecto al resto de los clusters.

Identificadas las zonas, se estudió la modalidad del clúster c_som_2_2 presente en ellas. Este clúster se caracteriza por el uso en su mayoría de armas blancas y ocurridos en la vía pública. La mayor cantidad de hechos delictivos se registraron los días sábados en el horario de siesta de 12:00 pm a 16:00 pm. Los elementos sustraídos con mayor incidencia son de tipo objeto personal y de tipo motocicleta con un alto porcentaje que evidencia que el robo fue con ataque brutal.

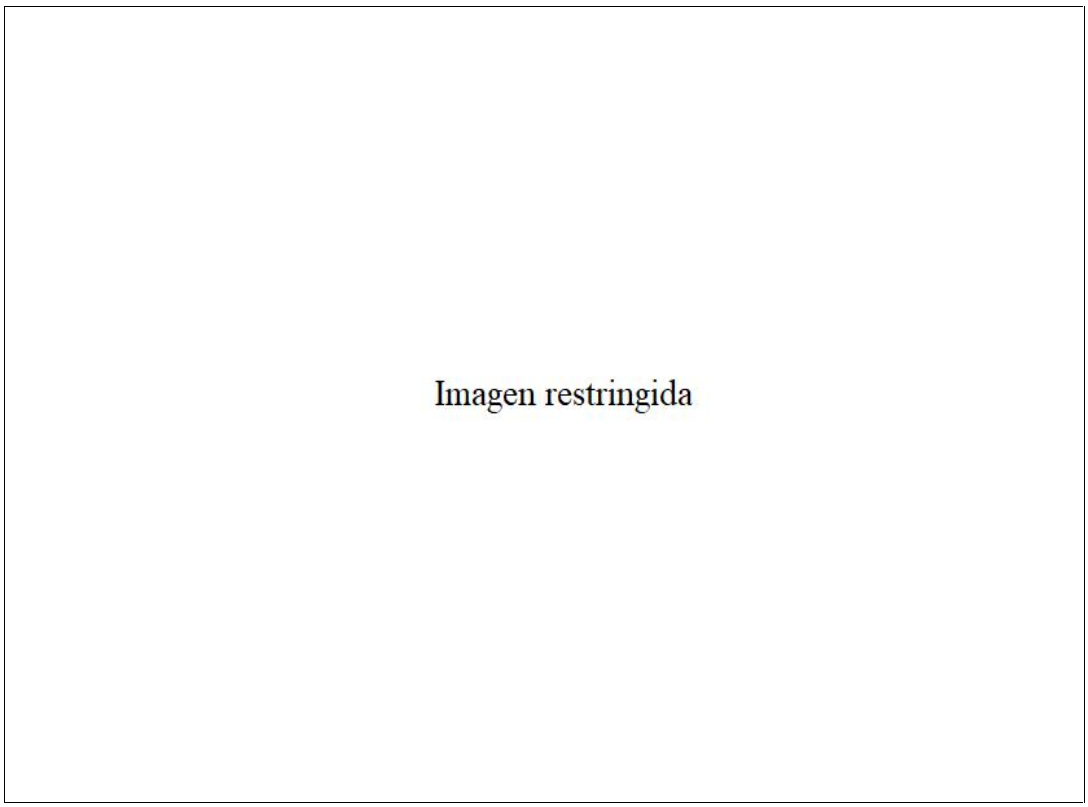


Fig. 5.9. Mapa de calor del clúster c_som_2_2 para caracterizar el comportamiento de delitos:
Fuente: elaboración propia

5.2.9.2. Representación geográfica del objetivo de minería de datos Nº 2

Para analizar geográficamente los clusters generados en este objetivo de minería de datos, se utilizó un mapa de calor general para representar los 4 grupos formados. En la Fig. 5.10 se visualizan cada una de las particiones (puntos de diferentes colores para cada clúster) y la densidad de puntos sobre el mapa. Particularmente, interesa analizar las zonas con alta densidad representadas en el mapa con color rojo.

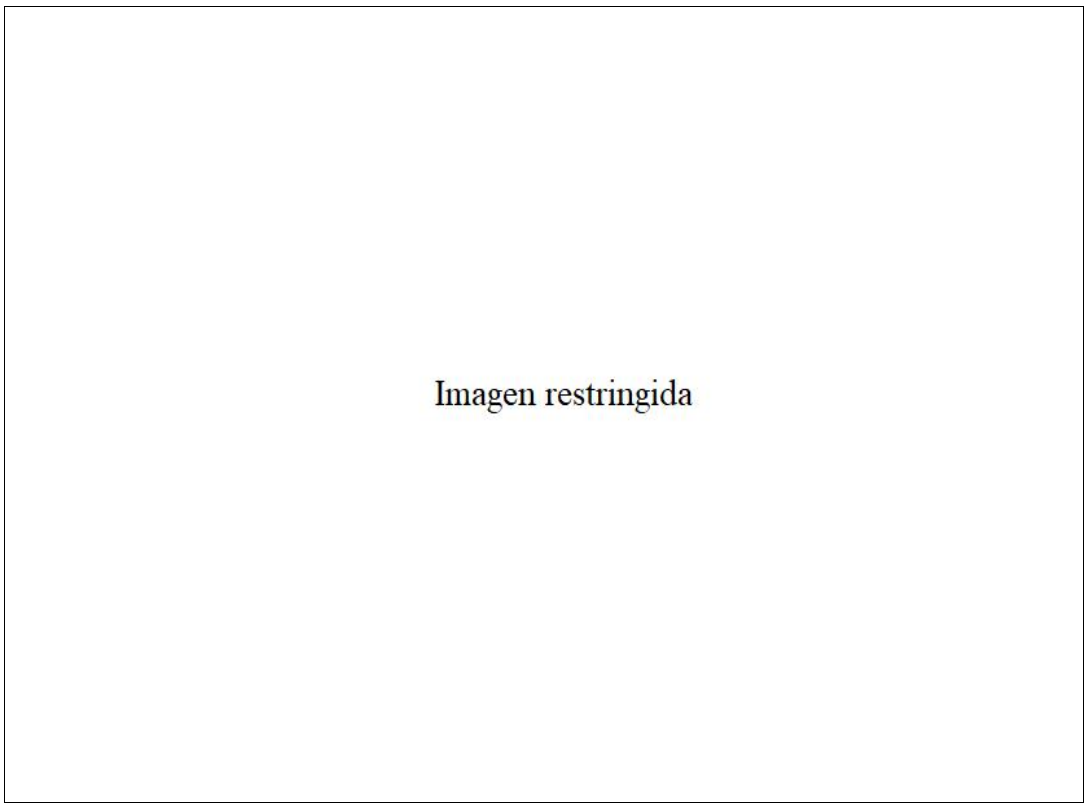


Fig. 5.10. Mapa de calor de los clusters para caracterizar zonas con mayor cantidad de delitos:
Fuente: elaboración propia

En referencia a la ubicación de la actividad delictiva por barrio, el mapa de calor (Fig. 5.11) indica como aquellos con mayor densidad de delitos a: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6, barrio N° 7 y barrio N° 8.

A partir de esta información y utilizando el software QGIS, se creó un mapa del crimen para cada uno de los barrios mencionados previamente, seguido del análisis de los clusters presenten en cada uno de ellos con el fin de describir el tipo del delito de acuerdo a los resultados derivados de aplicación de los algoritmos de minería de datos:

El barrio N° 1 registró un total de 19 delitos (Fig. 5.11), donde cada uno de estos hechos corresponde a un determinado clúster presente en el barrio. Así, el grupo conformado por puntos amarillos corresponden al clúster c_som_1_1, los puntos celestes al clúster c_som_1_2 y los puntos verdes al clúster c_som_2_2.

Se identificó que según las características de los clusters mencionados anteriormente, este barrio se define por la ocurrencia de delitos en su mayoría sobre la calle N° 1 (determinados por los clusters c_som_1_1 y c_som_2_2), en la vía pública y con armas blancas (establecidos por los clusters c_som_1_1 y c_som_2_2).

Los puntos localizados cerca de las paradas de colectivos (icono azul que representa un autobús) también podrían tratarse como un indicador que las personas que sufrieron el hecho podrían haber estado allí situadas, y que el objeto arrebatado haya sido de tipo personal (caracterizados por los clusters c_som_1_1 y c_som_1_2).

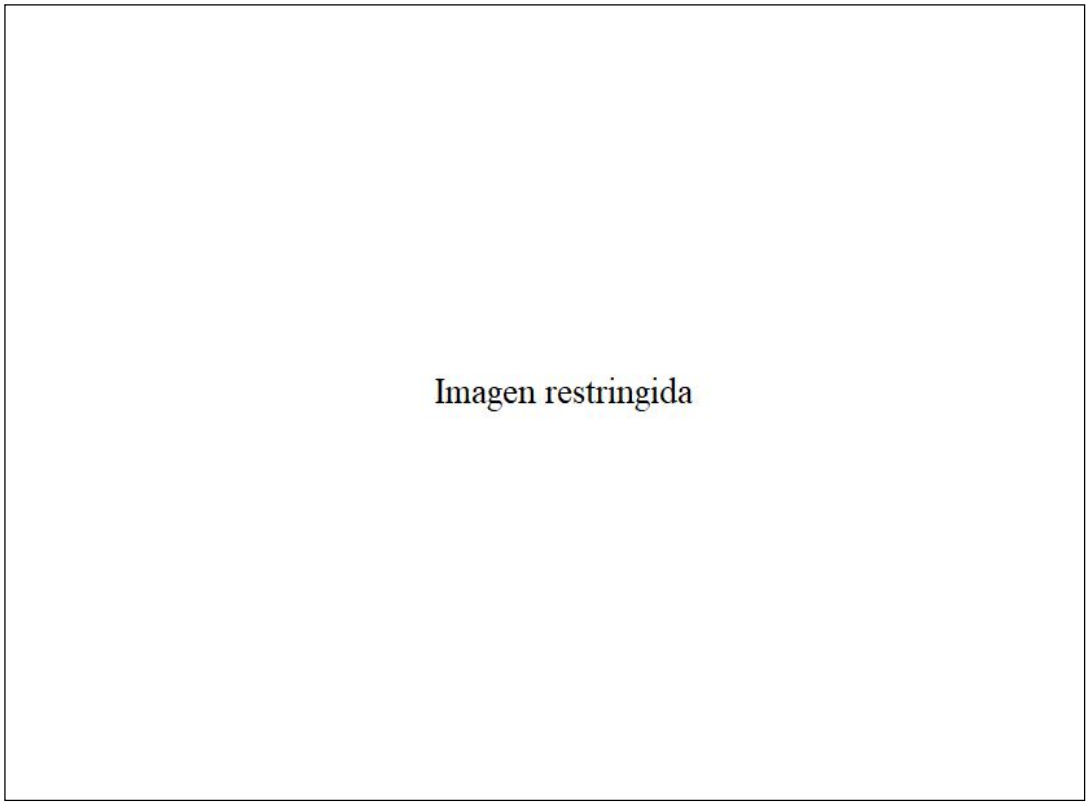


Fig. 5.11. Mapa del crimen del barrio N° 1:
Fuente: elaboración propia

El barrio N° 2 registró un total de 17 delitos (Fig. 5.12), donde cada uno de estos hechos corresponde a un determinado clúster presente. El grupo conformado por puntos rojos corresponde al clúster c_som_2_1 y los puntos verdes al clúster c_som_2_2.

Se identificó que de acuerdo a las características de los clusters mencionados anteriormente, este barrio se define por la ocurrencia de delitos en su mayoría sobre la calle N° 1 y en la vía pública (establecido por el clúster c_som_2_1), con el uso de arma

blanca (caracterizado por los clusters c_som_2_1 y c_som_2_2), y en forma de arrebato (determinado por el clúster c_som_2_1) o ataque brutal (determinado por el clúster c_som_2_2).

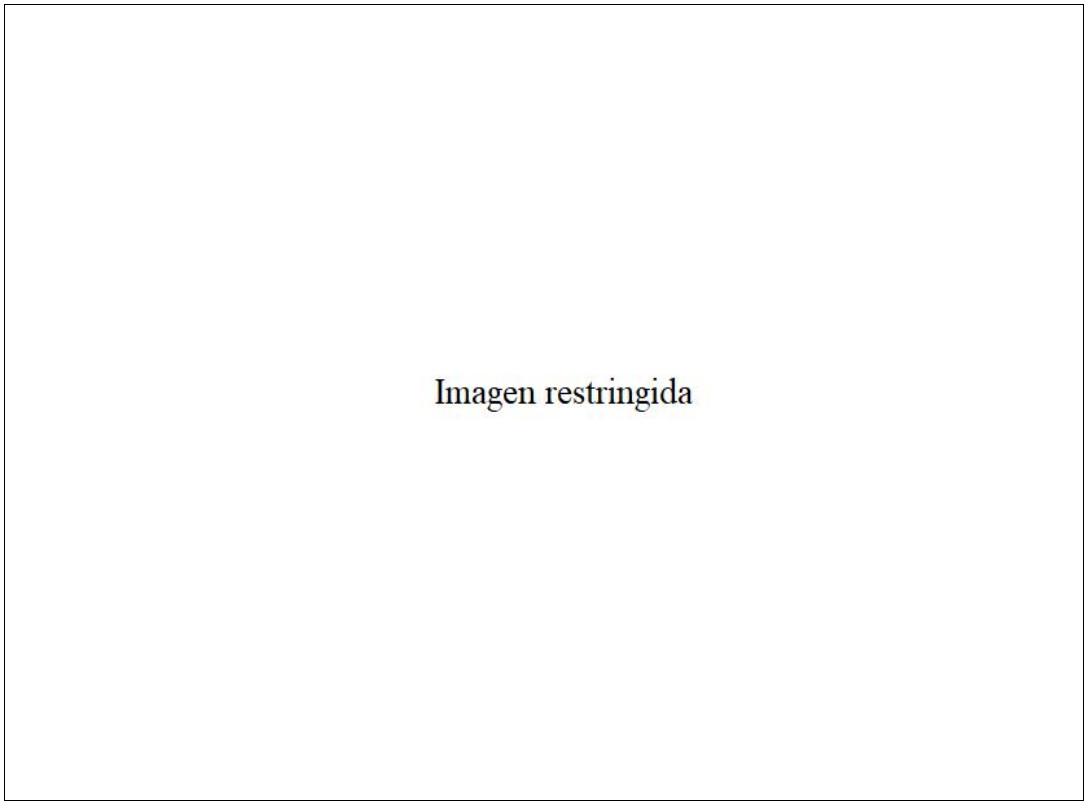


Fig. 5.12. Mapa del crimen del barrio N° 2:
Fuente: elaboración propia

El barrio N° 3 registró un total de 6 delitos, en su mayoría conformado por puntos rojos correspondientes al clúster c_som_2_1 (Fig. 5.13). De acuerdo a las características del clúster mencionado anteriormente, este barrio se define por la ocurrencia de delitos en su mayoría ubicados sobre la calle N°1. El 100% de los casos ocurrieron en la vía pública y con arma blanca y sustracción de algún objeto personal.

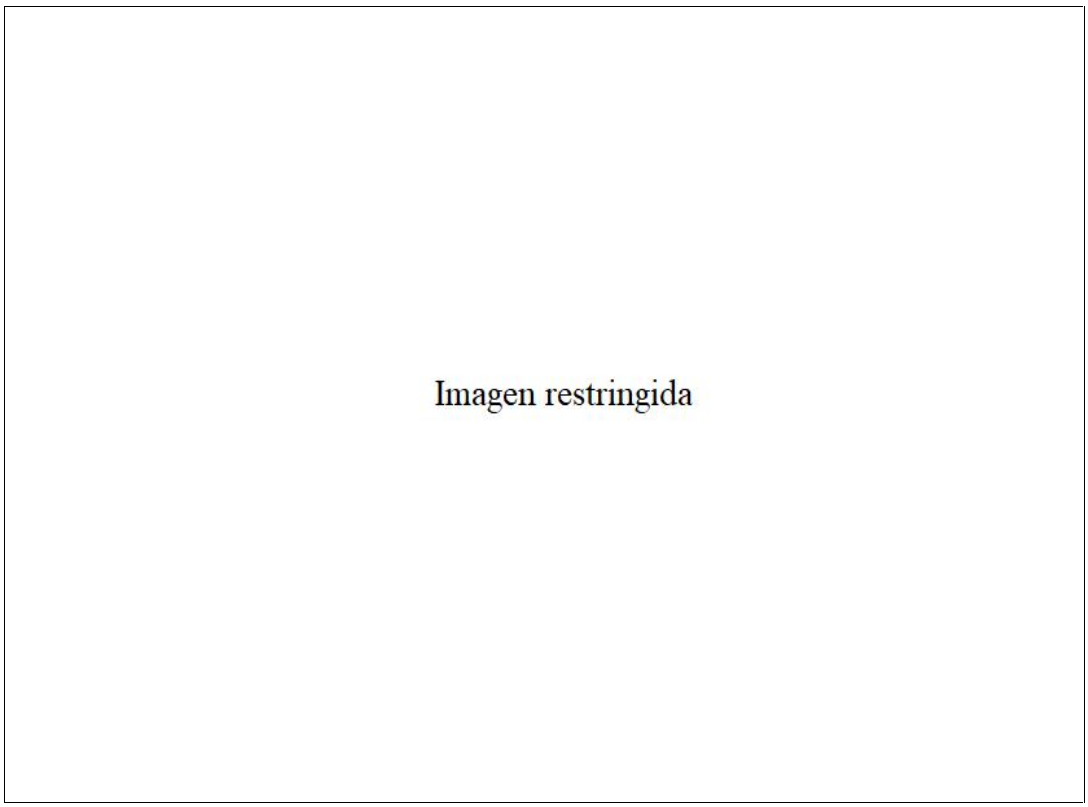


Fig. 5.13. Mapa del crimen del barrio N° 3:

Fuente: elaboración propia

El barrio N° 4 registró un total de 17 delitos (Fig. 5.14), donde cada uno de estos hechos corresponde a un determinado clúster presente en el barrio. El grupo conformado por puntos amarillos corresponde al clúster c_som_1_1 y otro grupo menor conformado por puntos verdes representan al clúster c_som_2_2.

Se identificó que de acuerdo a las características de los clusters mencionados anteriormente, este barrio se define por la ocurrencia de delitos en su mayoría sobre la calle N° 1 y con arma blanca (establecidos por los clusters c_som_1_1 y c_som_2_2). Se observa, además, que los puntos localizados cerca de las paradas de colectivos podrían indicar que allí se ubicaron las personas que sufrieron del hecho. Asimismo, se constató de los delitos ocurrieron en un vehículo o en un domicilio (determinado por el clúster c_som_2_2), o en la vía pública (caracterizado por el clúster c_som_1_1).

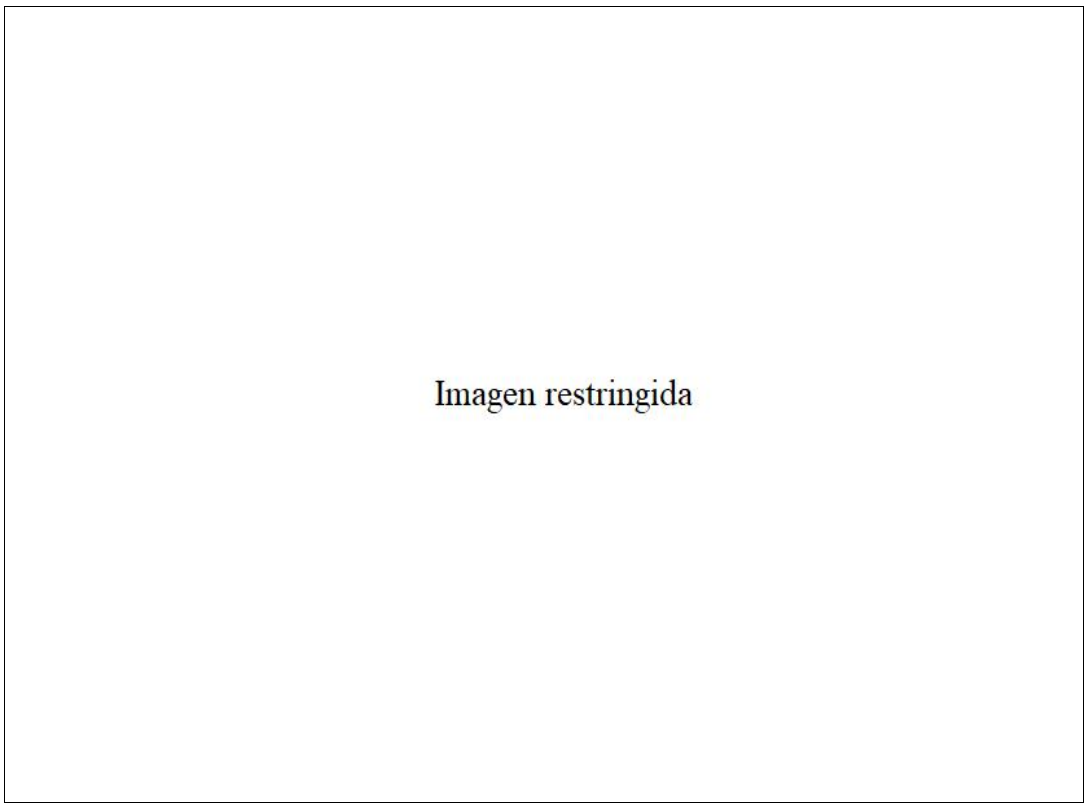


Fig. 5.14. Mapa del crimen del barrio N° 4:

Fuente: elaboración propia

El barrio N° 5 registró un total de 11 hechos delictivos (Fig. 5.15), donde cada uno de estos delitos corresponde a un determinado clúster presente en el barrio. El grupo conformado por puntos rojos corresponde al clúster c_som_2_1, el de los puntos amarillos al clúster c_som_1_1 y el de los puntos verdes al clúster c_som_2_2.

Se determinó de acuerdo a las características de los clusters mencionados anteriormente, que este barrio se define por la presencia de delitos en su mayoría sobre la calle N° 1 (determinado por el clúster c_som_2_1), con arma blanca (establecidos por los clusters c_som_1_1, c_som_2_1 y c_som_2_2) y en forma de arrebato (caracterizados por los clusters c_som_1_1 y c_som_2_1).

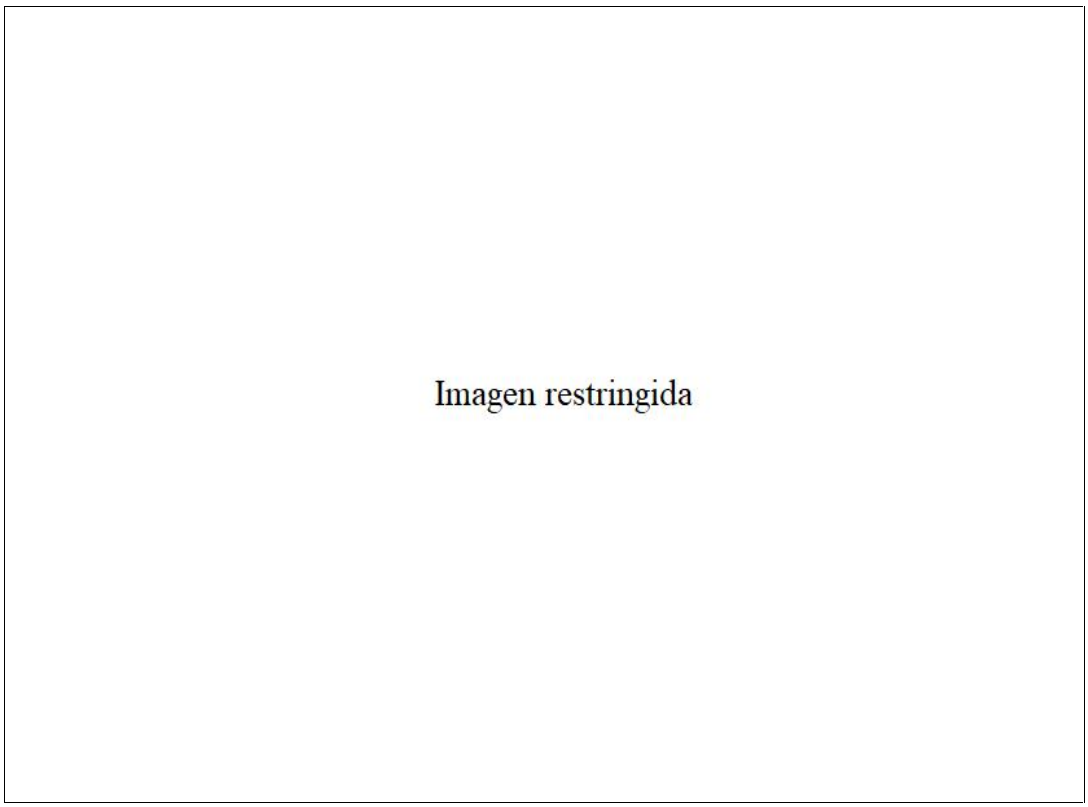


Fig. 5.15. Mapa del crimen del barrio N° 5:

Fuente: elaboración propia

El barrio N° 6 registró la mayor cantidad de delitos ocurridos en la ciudad con un total de 25 puntos (Fig. 5.16), donde cada uno de estos delitos corresponde a un determinado clúster presente en el barrio. El grupo conformado por puntos amarillos corresponden al clúster c_som_1_1 y los puntos verdes al clúster c_som_2_2.

Se apreció según las características de los clusters mencionados anteriormente, este barrio se define por la presencia de delitos en su mayoría sobre la calle N° 1 (establecido por el clúster c_som_1_1) ocurridos en la vía pública y con arma blanca (determinados por los clusters c_som_1_1 y c_som_2_2).

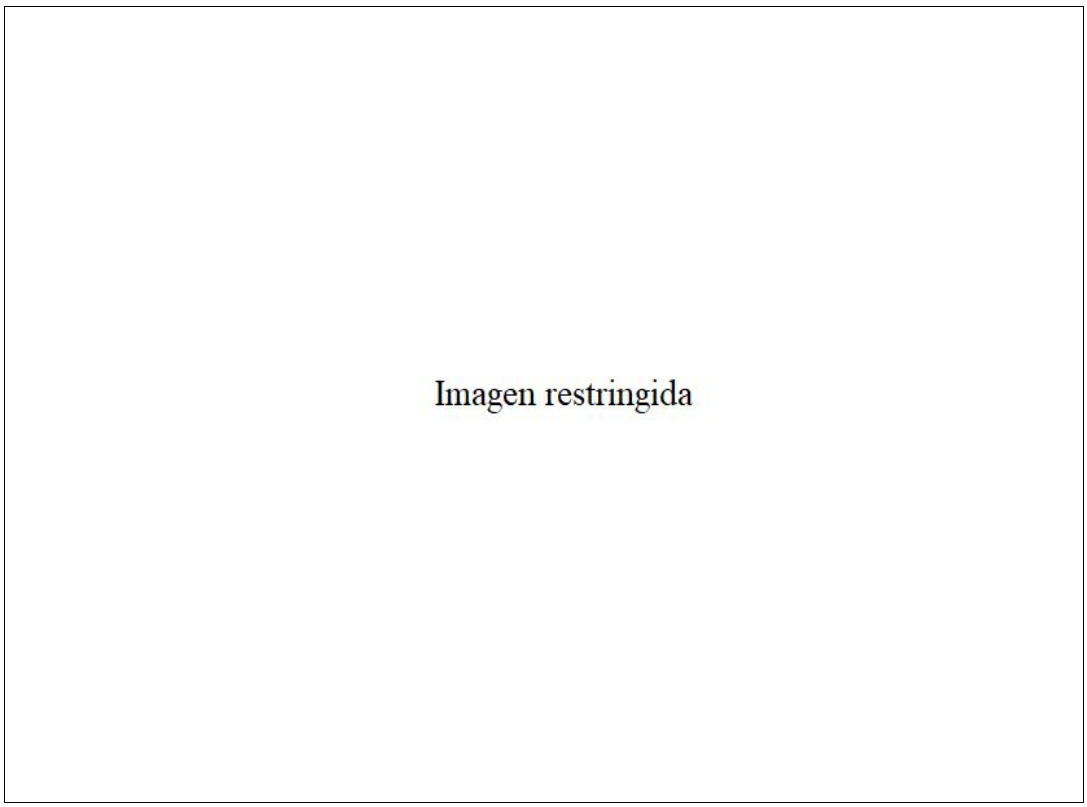


Fig. 5.16. Mapa del crimen del barrio N° 6:

Fuente: elaboración propia

El barrio N° 7 registró un total de 19 puntos (Fig. 5.17), donde cada uno de estos delitos corresponde a un determinado clúster presente. Los puntos amarillos corresponden al clúster c_som_1_1, los puntos celestes al clúster c_som_1_2 y los puntos verdes al clúster c_som_2_2.

Se apreció que de acuerdo a las características de los clusters mencionados anteriormente, este barrio se define por la presencia de delitos en su mayoría sobre la calle N° 1 (característica perteneciente a los clusters c_som_1_1 y c_som_1_2) y donde predominan las paradas de colectivos, que puede dar indicio de la ocurrencia de los robos.

Los casos sucedieron en la vía pública (determinados por los clusters c_som_1_1 y c_som_1_2) con robos de objetos personales (establecidos por los clusters c_som_1_1 y c_som_1_2). En su mayoría se utilizó armas blancas (caracterizados por los clusters c_som_1_1 y c_som_2_2).

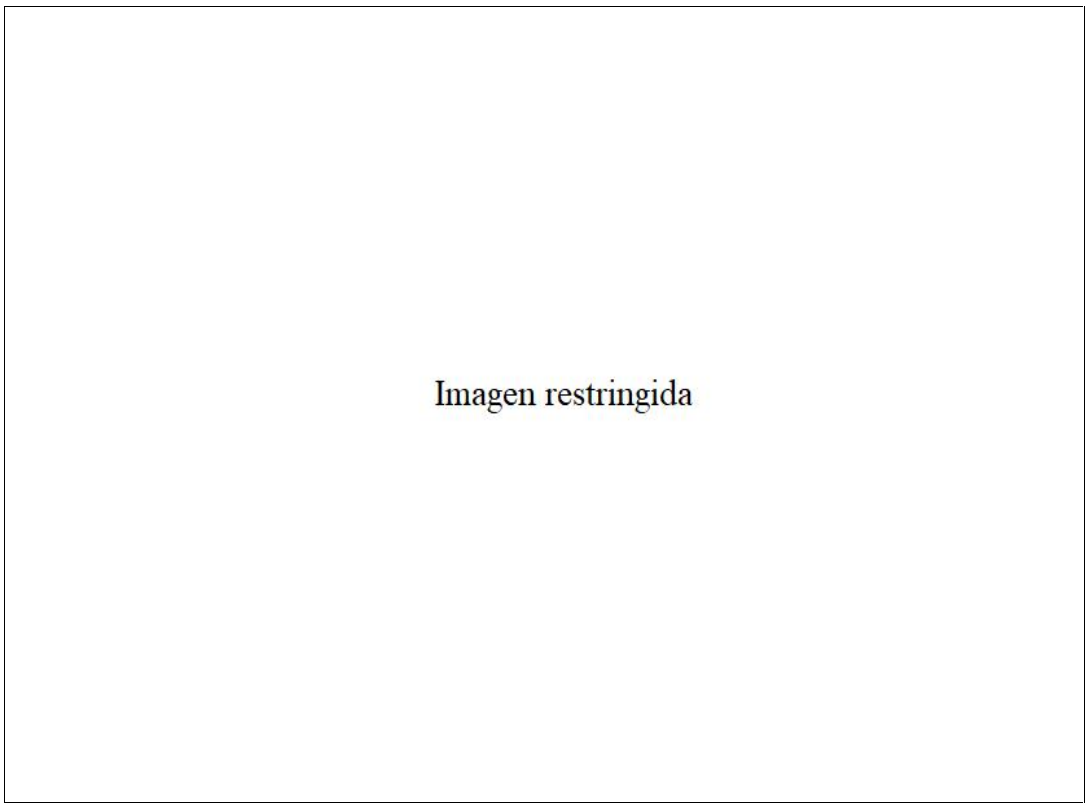


Fig. 5.18. Mapa del crimen del barrio N° 7:

Fuente: elaboración propia

El barrio N° 8 registró un total de 12 delitos (Fig. 5.19), donde cada uno de estos ellos corresponde a un determinado clúster presente. Los grupos están conformados por puntos amarillos correspondientes al clúster c_som_1_1 y aquellos con puntos celestes representan el clúster c_som_1_2.

Se apreció, de acuerdo a las características de los clusters mencionados anteriormente, que este barrio se define por la presencia de delitos en su mayoría sobre la vía pública (determinados por los clusters c_som_1_1 y c_som_1_2) y en forma de arrebato (expresados por el clúster c_som_1_1).

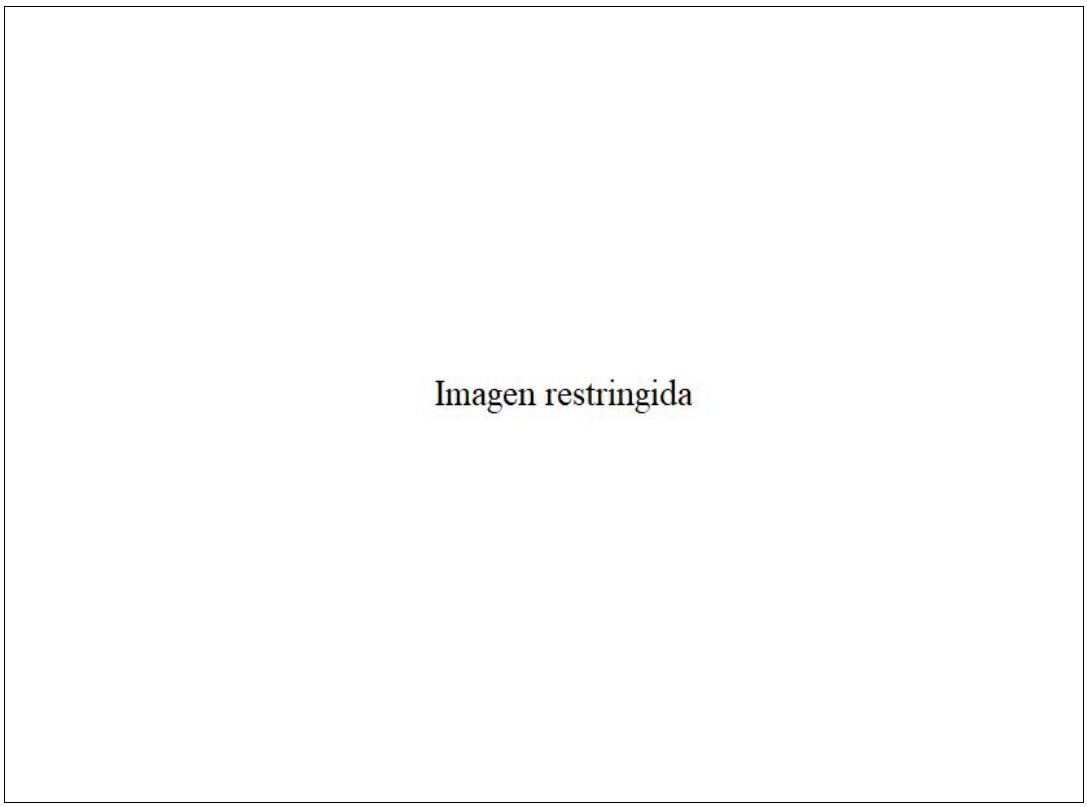


Fig. 5.19. Mapa del crimen del barrio N° 8:

Fuente: elaboración propia

5.2.9.3. Representación geográfica del objetivo de minería de datos Nº 3

Para la representación geográfica de los clusters generados en el tercer objetivo de minería de datos, nuevamente se utilizaron mapas de calor, los cuales muestran distintos tonos de color rojo y colores más claros según la densidad de los puntos concentrados. Estos clusters permiten identificar y caracterizar a las personas más vulnerables o propensas de sufrir algún delito, de acuerdo a los resultados derivados de la aplicación de los algoritmos de minería de datos.

Identificadas las zonas a través de los mapas de calor de los clusters c_som_1_1, c_som_1_2, c_som_2_1 y c_som_2_2 en las Figs. 5.20, 5.21, 5.22 y 5.23 respectivamente, se observa que existe una similitud en cuanto a la ubicación de zonas de mayor concentración de delitos, determinando los siguientes barrios como los más afectados: barrio N° 1, barrio N° 2, barrio N° 3, barrio N° 4, barrio N° 5, barrio N° 6 y barrio N° 7. Se mencionan a continuación los clusters formados y caracterizados por los algoritmos de minería de datos:

El clúster c_som_1_1 agrupó como víctimas más propensas a sufrir un delito a personas de sexo femenino, con un rango de edad entre los 36 a 40 años, y que el 100% de los robos ocurrieron por algún sospechoso de sexo masculino cuya edad ronda entre los 15 y 20 años.

El clúster c_som_1_2 concentró igual número de víctimas del sexo femenino y masculino, con un rango de edad entre los 21 a 25 años, y el 100% de los robos ocurrieron por algún sospechoso de sexo masculino cuya edad ronda entre los 21 y 25 años.

El clúster c_som_2_1 determinó como víctimas en su mayoría personas de sexo femenino, con un rango de edad entre los 21 a 25 años, y que la totalidad de los hurtos ocurrieron por algún sospechoso de sexo masculino cuya edad ronda entre los 15 y 20 años.

Por último, el clúster c_som_2_2 concentró en su mayoría a víctimas de sexo femenino, con un rango de edad entre los 21 a 25 años, y donde el 100% de los robos cometió algún sospechoso de sexo masculino cuya edad ronda entre los 21 y 25 años.

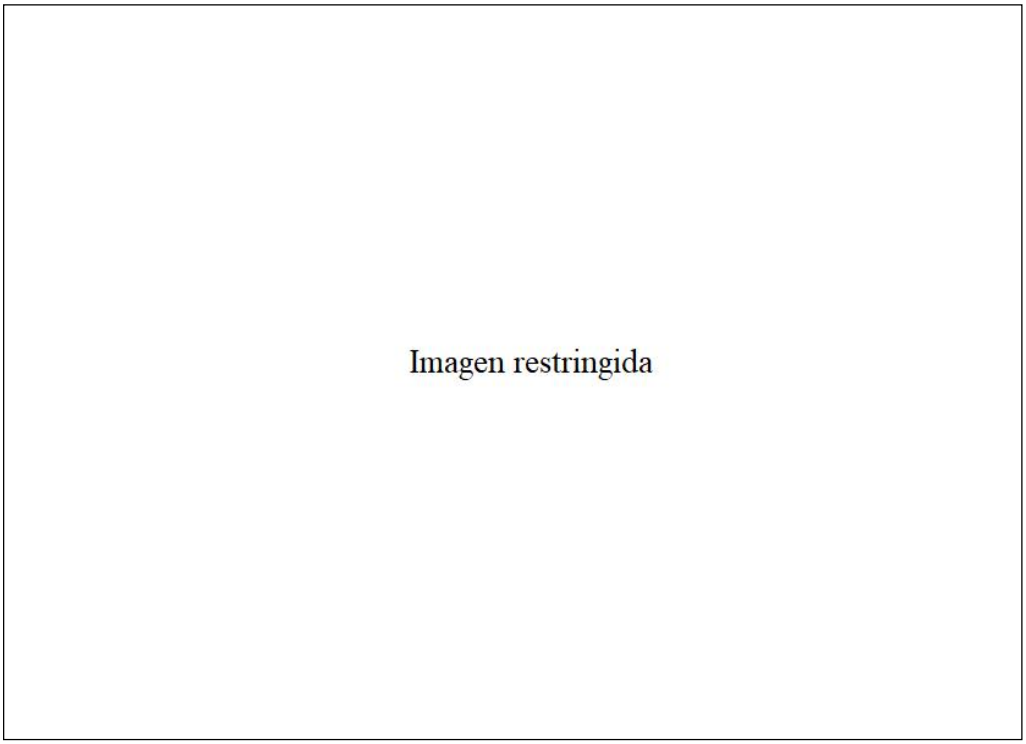


Fig. 5.20. Mapa de calor del clúster c_som_1_1 para caracterización de personas más propensas a sufrir de algún delito;
Fuente: elaboración propia

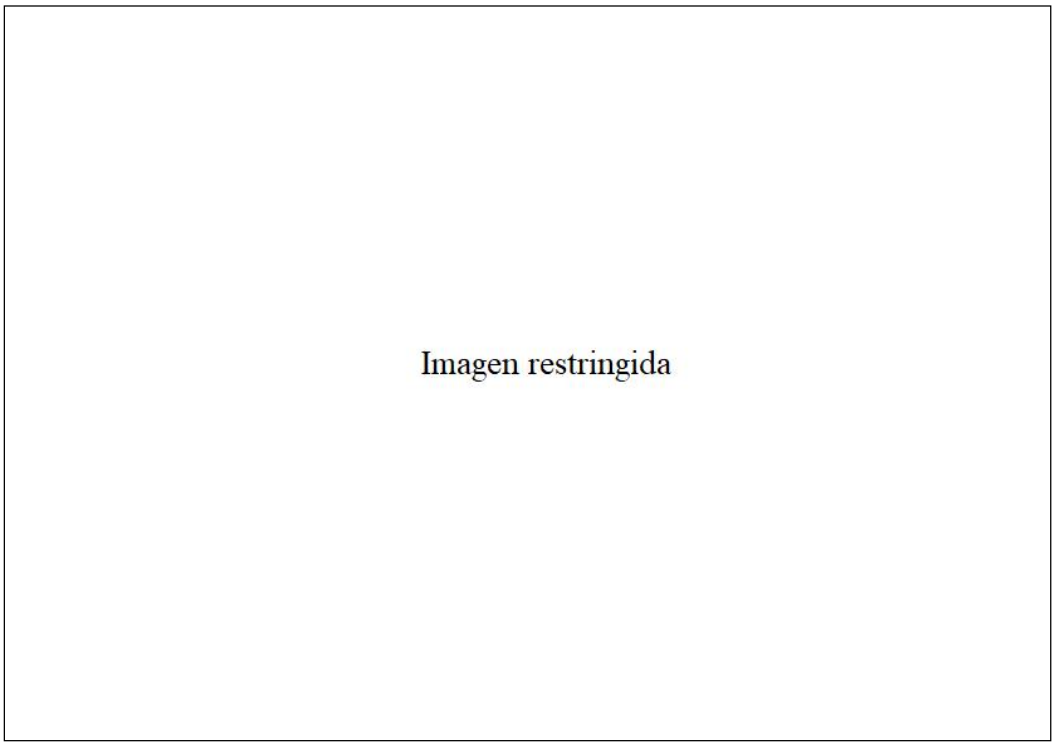


Fig. 5.21. Mapa de calor del clúster c_som_1_2 para caracterización de personas más propensas a sufrir de algún delito:
Fuente: elaboración propia

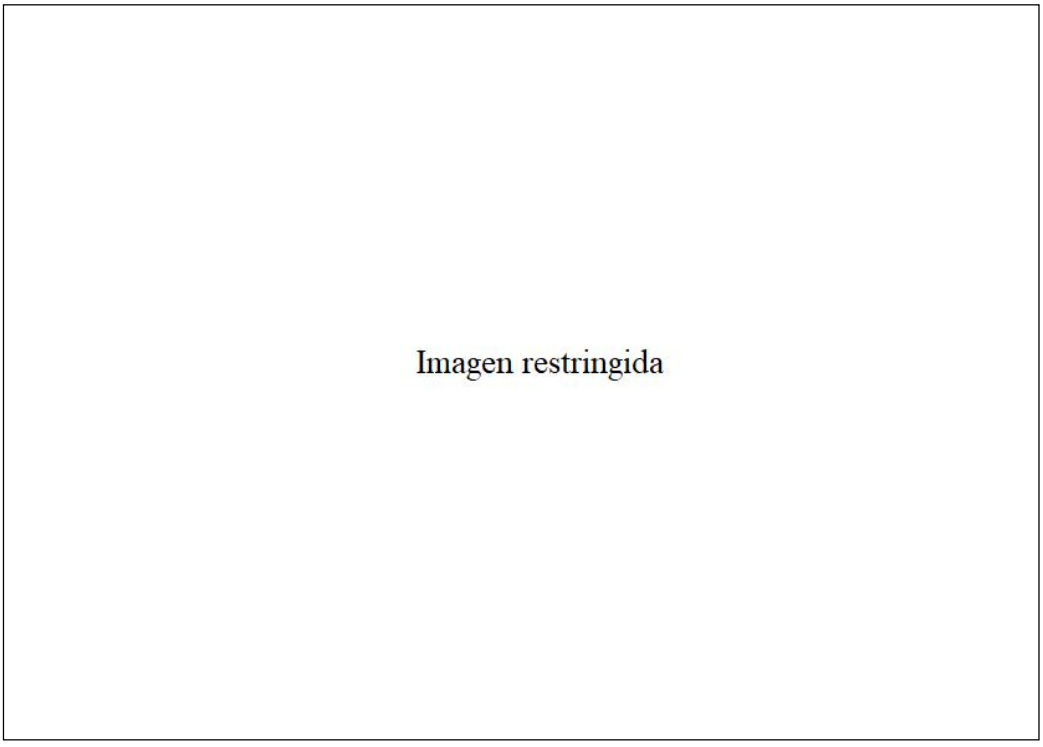


Fig. 5.22. Visualización de clúster c_som_2_1 para caracterización de personas más propensas a sufrir de algún delito:
Fuente: elaboración propia

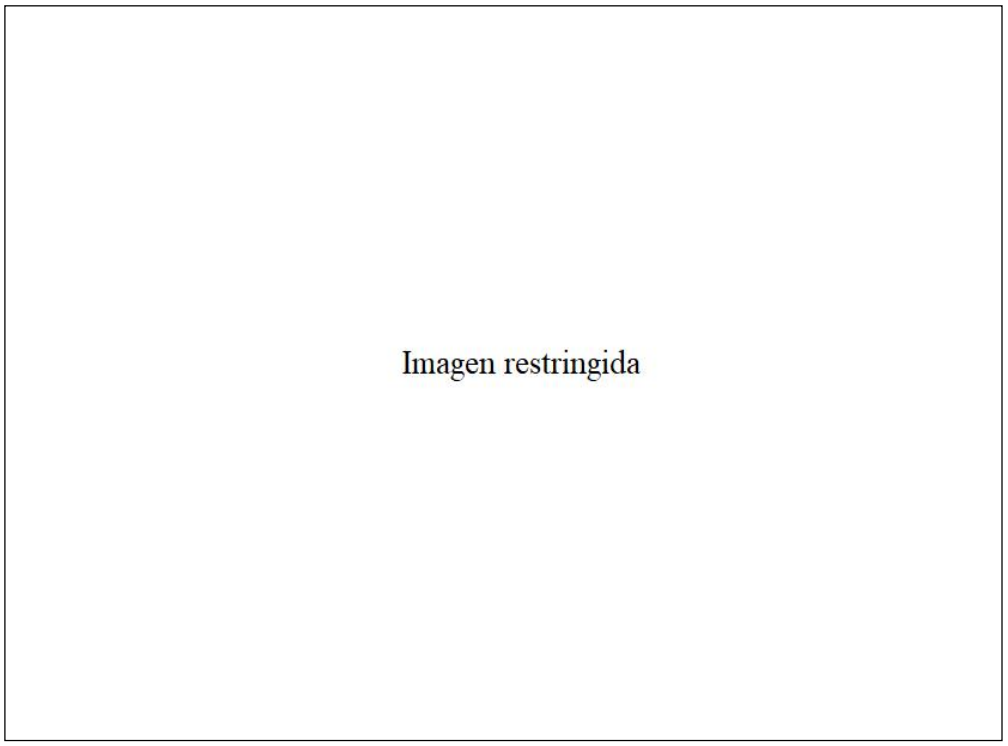


Fig. 5.23. Mapa de calor del clúster c_som_2_2 para caracterización de personas más propensas a sufrir de algún delito:
Fuente: elaboración propia

La Fig. 5.24 muestra el mapa del delito de la ciudad sobre el que se aplicó el procedimiento con fines de validación y que representa la actividad delictiva del territorio elegido como caso de estudio, es decir, el caso de verificación para el procedimiento propuesto en el TFM. Este mapa cartográfico muestra los puntos delictivos resultado del análisis de los procesos de explotación de información sobre los datos espaciales de la información recibida del SAT de robos y hurtos ocurridos durante el primer semestre del año 2017. Se visualizan los clusters con la modalidad del delito obtenido del objeto de minería de datos N°1, junto con el mapa de calor que indican las zonas con mayor concentración de hechos delictivos.

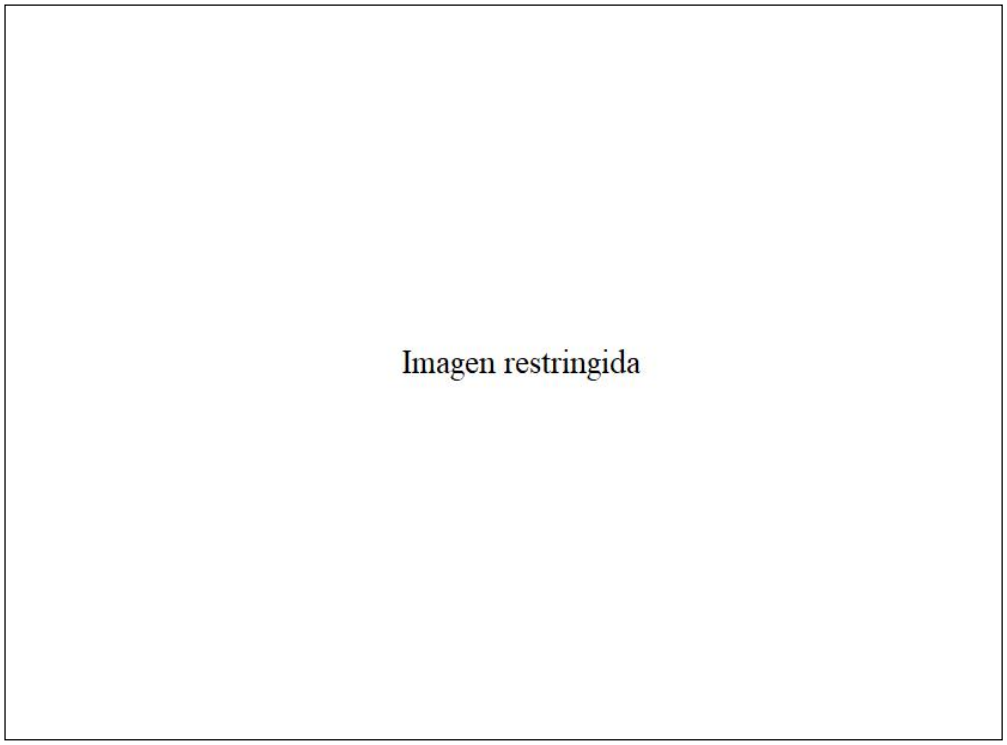


Fig. 5.24. Mapa del delito de ciudad mostrando los puntos delictivos registrados entre enero-junio del año 2017: Fuente: elaboración propia

Estos mapas geográficos representan el producto final de la aplicación del procedimiento propuesto. Es importante mencionar que la detección de patrones delictivos y la geolocalización de los sitios en donde se presentan los incidentes criminales son vitales para el análisis criminal.

Capítulo 6

Aplicación y análisis de minería de datos sobre el caso de validación

6. Aplicación y análisis de minería de datos sobre el caso de validación

En este capítulo se detalla el cuarto proceso del procedimiento propuesto, en el cual se aplican los procesos de explotación de información sobre un caso de estudio orientado a la detección de hechos delictivos de robos y hurtos en la ciudad en el primer semestre del año 2017.

Se muestra la aplicación y resultado de técnicas de minería de datos sobre el caso de validación en la sección 6.1, y se incluyen las discusiones derivadas de la aplicación de la misma en la sección 6.2.

6.1. Aplicación de minería de datos sobre los delitos de robo y hurto en la ciudad de Corrientes

Se utilizó como guía de trabajo para la aplicación de minería de datos la metodología de trabajo CRISP-DM.

6.1.1. Metodología CRISP-DM

A continuación se listan los resultados de aplicar el modelo de proceso al problema, siguiendo las fases de la metodología CRISP-DM.

6.1.1.1. Comprensión del negocio

Se detallan cada una de las tareas de esta primera fase de la metodología CRISP-DM, cuyo propósito fue determinar los objetivos y requisitos del proyecto.

6.1.1.1.1. Determinar los objetivos del negocio

La delincuencia en las ciudades es una constante diaria como consecuencia del número creciente de robos importantes ocurridos, los cuales preocupan a la comunidad y han puesto en alerta al accionar policial. El robo de dinero a comercios y familias de la ciudad, hechos de violencia con arrebatos y robos a mano armada en la vía pública han agravado la inseguridad en la ciudad. En referencia a esta problemática, se plantea la aplicación de la minería de datos.

Previamente se detallan los objetivos del negocio:

Como objetivo de negocio Nº 1 se requiere identificar y diseñar diferentes planes de prevención que permitan mejorar la seguridad, advertir situaciones de riesgo al ciudadano y reducir el impacto de la delincuencia.

Como objetivo de negocio Nº 2 se pretende asegurar el éxito en el accionar policial brindando mayor eficacia y eficiencia ante la detención de zonas con mayor ocurrencia de delitos en la ciudad.

Como objetivo de negocio Nº 3 se busca identificar las situaciones de riesgo para las personas más vulnerables que puedan ser afectadas por la inseguridad.

6.1.1.1.2. Evaluación de la situación

6.1.1.1.2.1. Inventario de recursos

En la Tabla IV, se presenta el inventario de los recursos requeridos para la realización del trabajo:

Tabla IV

Recursos requeridos para la realización del trabajo:

Fuente: elaboración propia

Tipo de recurso	Descripción
Recurso de Hardware	- Una notebook con un procesador doble núcleo y 8 GB de RAM.
Recurso de Software	Software open source Tanagra v1.4.
Recurso de Datos	- Se utilizó una base de datos proporcionada por el SAT, la cual se recibió en formato xls.

6.1.1.1.2.2. Requerimientos y restricciones:

- Requerimientos: Los resultados del proceso deben estar representados de una forma clara, simple y entendible.
- Restricciones: El interés del análisis se encuentra limitado al periodo comprendido entre enero y junio de 2017. La selección de datos del primer semestre del año 2017 se justifica en su relación con la elaboración del plan del Trabajo Final de Maestría.

6.1.1.1.2.3. Gestión del riesgo y plan de contingencia

El impacto del riesgo de realizar este proyecto y sus acciones contingentes se describen en la Tabla V:

Tabla V

Gestión de riesgo y plan de contingencia:

Fuente: elaboración propia

Id. del riesgo	Descripción del riesgo	Impacto	Acciones contingentes
1	Los patrones hallados no logran alcanzar los objetivos propuestos	ALTO	Utilizar distintas técnicas de minería en la fase de modelado. Utilizar distintos parámetros para los modelos obtenidos.
2	Que la baja calidad de los datos detecte patrones que no puedan ser comprendidos en su totalidad	MEDIO	Seleccionar la metodología de trabajo más adecuado para lograr conservar la máxima calidad de los datos. Realizar diferentes representaciones con los resultados de los patrones encontrados.

6.1.1.1.2.4. Terminología del negocio

- Delito: Acción u omisión voluntaria o imprudente penada por la ley [55].
- Robo: Quitar o tomar para sí con violencia o con fuerza lo ajeno [55].
- Hurto: Tomar o retener bienes ajenos contra la voluntad de su dueño, sin intimidación en las personas ni fuerza en las cosas [55].
- Homicidios culposos en hechos de tránsito: Todo hecho que en ocasión o por motivo del tránsito vehicular produzca una muerte en forma involuntaria [54].
- Homicidios dolosos: homicidios dolosos: Comprende todos los homicidios causados en forma intencional por el imputado, ya sea homicidio simple, agravado, en estado de emoción violenta, homicidio preterintencional, homicidio en ocasión de robo y homicidio en riña [54].
- Suicidios: Acción de quitarse voluntariamente la vida [55].
- Criminalidad: Circunstancia que hace que una acción sea criminal [55].

6.1.1.1.2.5. Costes y beneficios

Este proyecto no generó ningún costo adicional dado que los datos pertenecen al propio organismo público que lo provee y las herramientas software utilizadas para el desarrollo del trabajo son de código abierto.

En cuanto a beneficios, dado que el proyecto se encuentra enfocado a cuestiones de seguridad ciudadana, se podrá ofrecer a las áreas encargadas de la misma información relevante para la toma de decisiones y mejorar con ello la calidad de los servicios ofrecidos a la comunidad.

6.1.1.1.3. Determinar los objetivos de minería de datos

A continuación, se detallan los objetivos del proyecto de minería de datos:

Para el objetivo de minería de datos N° 1 se desea analizar según la base de datos del SAT, patrones de comportamiento relevantes a través de la agrupación de casos. Y partir de ello determinar los factores que caracterizan la ocurrencia de delitos de robo y hurto en la ciudad y cuales son aquellos con mayor incidencia dentro de cada grupo. Para lograr dicho objetivo se han establecido las siguientes necesidades:

- Identificar y caracterizar grupos que definan el comportamiento de delitos en base a las características propias del delito (tipo de arma, tipo de ataque, objeto sustraído, lugar del hecho, día y hs del delito, etc.), para comprender con mayor detalle indicadores que definan a dichos grupos.
- A partir del comportamiento definido en el punto anterior, determine cual/es de las características tiene un mayor nivel de incidencia en la ocurrencia de delitos en la ciudad.

Para el objetivo de minería de datos N° 2 y según la base de datos del SAT, se desea establecer distintos grupos entre las zonas de mayor porcentaje de ocurrencia de delitos en la ciudad, con el fin de realizar un análisis detallado en dichas regiones. Para lograr dicho objetivo se han establecido las siguientes necesidades:

- Identificar y caracterizar grupos entre las zonas de mayor ocurrencia de delitos, en orden a comprender con mayor detalle indicadores que definan a dichas zonas.
- Identificar los factores predominantes en cada grupo identificada.

Por último, para el objetivo de minería de datos N° 3 se desea establecer distintos grupos entre las personas más propensas a sufrir un delito, en orden a comprender con mayor detalle cómo afectó el delito a dichas personas. Identificar los factores predominantes en cada grupo identificado.

- Identificar y caracterizar grupos entre las personas más propensas a sufrir un delito.
- Identificar los factores predominantes en cada grupo identificado.

6.1.1.1.4. Realizar el plan del proyecto de minería de datos

La duración total estimada para la ejecución el proyecto fue de 23 semanas. En la Tabla VI se observa la división de las etapas, tiempo estimado, recursos, entrada y salidas utilizados en el trabajo.

Tabla VI

Plan de proyecto:

Fuente: elaboración propia

Etapa	Tarea	Duración estimada (sem)	Entradas	Salidas
Entendimiento de los datos	Recolección de datos	1	Planilla de cálculo	Reporte de recolección inicial de datos
	Describir y explorar datos	1	Reporte de recolección inicial de datos	Reporte y gráficos estadísticos de los datos
	Análisis y verificación de calidad de datos	1	Reporte y gráficos estadísticos de los datos	Reporte de calidad de los datos
Preparación de los datos	Limpiar datos	2	Base de datos	Reporte de limpieza de datos
	Construir datos	1	Base de datos	Reporte de construcción de datos
	Formatear datos	1	Base de datos	Reporte de atributos formateados
	Selección final del conjunto de datos	1	Base de datos	Reporte del conjunto final de datos
Modelado	Seleccionar técnicas de modelado	3	Base de datos y objetivos de minería de datos	Técnicas de modelado
	Seleccionar plan de pruebas	3	Modelos de plan de pruebas	Pruebas del modelo
	Construcción de los modelos	3	Conjunto final de datos	Aplicación de los algoritmos de minería de datos
	Evaluación de modelos	2	Modelos	Evaluación de los resultados de aplicación de modelos
Evaluación	Evaluación de resultados	3	Evaluación de los resultados de aplicación de modelos	Reporte de resultados

6.1.1.2. Comprensión de los datos

En esta segunda fase de la metodología CRISP-DM se realizó la recolección inicial de los datos para establecer un primer contacto con el problema, familiarizarse con los datos y determinar su calidad.

6.1.1.2.1. Recolección de datos iniciales

Para la recolección de los datos se utilizó como fuente una base de datos recibida que contiene los hechos de delitos contra la propiedad (delitos de robo y hurto) ocurridos durante el período del primer semestre del 2017 de la Capital de la Provincia de

Corrientes registrados en el Sistema de Alerta Temprana (SAT), la cual se almacena en una planilla de cálculo.

6.1.1.2.2. Descripción de los datos

Para el desarrollo del trabajo se seleccionaron los delitos de robo y hurto con un total de 366 registros que corresponden a la localidad de la Capital de la Provincia de Corrientes en el periodo de enero hasta junio del año 2017.

Esta información contempla los siguientes datos:

- Registro de hechos: características generales del hecho denunciado (lugar, día, hora, delito denunciado y comisaría interviniente).
- Registro de denunciantes: identidad y características de la víctima o denunciante.
- Registro de autores identificados: edad aproximada, características.
- Registro de elementos robados: información del tipo de objeto, autos, etc.
- Registro de armas utilizada: tipo de arma utilizada durante el delito.

La tabla principal contiene los siguientes datos:

- id_registro: Representa el identificador (único e irrepetible) del registro de cada delito denunciado.
- id_provincia: Representa el identificador del nombre de la provincia en la cual se cometió el hecho delictivo.
- provincia_descrip: Representa el nombre de la provincia en la cual se cometió el hecho.
- id_departamento: Representa el identificador del nombre del departamento de la provincia en el cual se cometió el hecho delictivo.
- departamento_descrip: Representa el nombre del departamento de la provincia en el cual se cometió el hecho.
- id_municipio: Representa el identificador del nombre del municipio en la cual se cometió el hecho delictivo.
- municipio_descrip: Representa el nombre del municipio en la cual se cometió el hecho.

- id_jurisdic_policial: Representa el identificador del nombre de la jurisdicción policial en la cual se cometió el hecho delictivo.
- jurisdic_policial: Representa el nombre de la jurisdicción policial en la cual se cometió el hecho delictivo.
- id_comisaria: Representa el identificador del nombre de la comisaría responsable del hecho.
- comisaria_descrip: Representa el nombre de la comisaría responsable del hecho.
- id_delito: Representa el identificador del nombre del delito cometido.
- delito_descrip: Representa el nombre del delito cometido.
- f_dia: Representa el día del mes en que se cometió el hecho.
- dia_m: Contiene el día de la semana en que se cometió el hecho.
- f_mes: Representa el mes en que se cometió el hecho.
- f_anio: Representa el año en que se cometió el hecho.
- f_hora: Representa el horario en la cual se cometió el hecho.
- barrio_descrip: Contiene el nombre del barrio en la cual se cometió el hecho.
- calle: Representa el nombre de la calle en la cual se cometió el hecho.
- altura: Contiene la numeración de la calle en la cual se cometió el delito.
- id_tipo_lugar: Representa el identificador del tipo de lugar donde se cometió el delito.
- tipo_lugar: Contiene la descripción del tipo de lugar donde se cometió el delito.
- en_ocasion: informa si el delito se cometió en ocasión de algún otro hecho delictivo.
- id_clase_arma: Representa el identificador del tipo de arma con la que se cometió el delito.
- clase_arma: Contiene el tipo de arma con la que se cometió el delito.
- id_elemento_sustraido: Representa el identificador del tipo de elemento sustraído.
- elemento_sustraido: Representa el nombre del tipo de elemento sustraído.
- id_tipo_ataque: Representa el identificador del tipo de ataque.
- tipo_ataque: Representa el nombre del tipo de ataque.

- id_fue_lesionado: Representa el identificador de la descripción, si la persona sufrió de una lesión por parte del delincuente durante el ataque.
- fue_lesionado_descrip: Indica la descripción si la persona sufrió de una lesión por parte del delincuente durante el ataque.
- id_sospechoso_sexo: Representa el identificador del tipo de sexo del atacante.
- sospechoso_sexo_descrip: Representa la descripción del tipo sexo del atacante.
- sospechoso_edad: Representa la edad aproximada del sospechoso.
- id_victima_sexo: Representa el identificador tipo del sexo de la víctima.
- victima_sexo_descrip: Representa la descripción del tipo de sexo de la víctima.
- victima_edad: Contiene la edad de la víctima.
- id_victima_nacionalidad: Representa el identificador de la nacionalidad de la víctima.
- victima_nacionalidad_descrip: Contiene la descripción de la nacionalidad de la víctima.
- id_victima_ocupacion: Representa el identificador de la descripción de la dedicación u ocupación de la víctima.
- victima_ocupacion_descrip: Representa la información de la dedicación u ocupación de la víctima.

Las Tablas VII.a y VII.b describen las variables según su tipo y los valores posibles que puede asumir cada una de ellas.

Tabla VII.a

Descripción de las variables extraídas:

Fuente: elaboración propia

Atributos	Tipo	Valores Posibles	Longitud del campo
id_registro	Entero	[1-366]	Máximo de 5 enteros
id_provincia	Entero	4	Máximo de 5 enteros
provincia_descrip	Varchar	Corrientes	Máximo de 30 caracteres
id_departamento	Entero	1	Máximo de 5 enteros
departamento_descrip	Varchar	Capital	Máximo de 30 caracteres
id_municipio	Entero	1	Máximo de 4 enteros
municipio_descrip	Varchar	Corrientes	Máximo de 30 caracteres
id_jurisdic_policial	Entero	[1-21]	Máximo de 3 enteros
jurisdic_policial	Varchar	[Seccional 1ª- Seccional 21ª]	Máximo de 30 caracteres
id_comisaria	Entero	[1-21]	Máximo de 3 enteros
comisaria_descrip	Varchar	[Comisaría Primera Urbana Capital- Comisaría Vigésimo Primero Urbana Capital]	Máximo de 70 caracteres
id_delito	Entero	[14-15]	Máximo de 3 enteros
delito_descrip	Varchar	[Robo – Hurto]	Máximo de 30 caracteres

Tabla VII.b

Descripción de las variables extraídas:

Fuente: elaboración propia

Atributos	Tipo	Valores Posibles	Longitud del campo
f_dia	Entero	[1-31]	Máximo de 2 enteros
dia_m	Varchar	[Lunes - Domingo]	Máximo de 20 caracteres
f_mes	Varchar	[Enero- Junio]	Máximo de 20 caracteres
f_anio	Entero	2017	Máximo de 4 enteros
f_hora	Numérica	[00:00, 24:00]	Máximo de 5 enteros
barrio_descrip	Varchar	[17 de Agosto - Yapeyu]	Máximo de 50 caracteres
calle	Varchar	[22 de Mayo - Turín]	Máximo de 70 caracteres
altura	Entero	[40 - 6400]	Máximo de 4 enteros
id_tipo_lugar	Entero	[1 - 5]	Máximo de 2 enteros
tipo_lugar	Varchar	[Vía Pública, Domicilio Particular, Comercio, Interior de Rodado, Otro Lugar]	Máximo de 70 caracteres
en_ocasion	Varchar	[Violación, Otro delito, No hubo otro delito]	Máximo de 50 caracteres
id_clase_arma	Entero	[1 - 5]	Máximo de 2 enteros
clase_arma	Varchar	[Arma de Fuego, Arma Blanca, Otra, Ninguna, S/D]	Máximo de 50 caracteres
id_elemento_sustraido	Entero	[1 - 7]	Máximo de 2 enteros
elemento_sustraido	Varchar	[Domiciliario, Vehículo, Motocicleta, Bicicleta, Objeto Personal, Otros, No hubo elemento sustraído]	Máximo de 70 caracteres
id_tipo_ataque	Entero	[1 - 4]	Máximo de 2 enteros
tipo_ataque	Varchar	[Arrebato, Forcejeo, Ataque Brutal, No existió ataque]	Máximo de 50 caracteres
id_fue_lesionado	Entero	[1 - 2]	Máximo de 2 enteros
fue_lesionado_descrip	Varchar	[No fue lesionado - Si fue lesionado]	Máximo de 50 caracteres
id_sospechoso_sexo	Entero	[1 - 3]	Máximo de 2 enteros
sospechoso_sexo_descrip	Varchar	[Masculino, Femenino, S/D]	Máximo de 15 caracteres
sospechoso_edad	Entero	[15 - 56]	Máximo de 3 enteros
id_victima_sexo	Entero	[1 - 2]	Máximo de 2 enteros
victima_sexo_descrip	Varchar	[Masculino, Femenino]	Máximo de 15 caracteres
victima_edad	Entero	[15 - 57]	Máximo de 3 enteros
id_victima_nacionalidad	Entero	[1]	Máximo de 2 enteros
victima_nacionalidad_descrip	Varchar	[Argentina]	Máximo de 15 caracteres
id_victima_ocupacion	Entero	[1 - 3]	Máximo de 2 enteros
victima_ocupacion_descrip	Varchar	[Tiene algún oficio, Tiene alguna profesión, No tiene ni oficio ni profesión]	Máximo de 50 caracteres

6.1.1.2.3. Exploración de los datos

Con la finalidad de explorar los datos, se realizaron consultas y análisis estadísticos simples que revelan propiedades de los datos contenidos en la base de datos delictiva. Se construyeron gráficos de distribución, los cuales se explican a continuación:

6.1.1.2.3.1. Análisis de los atributos

Se analizó la distribución de valores de los siguientes atributos: tipo de delito, día de la semana de ocurrencia del delito, mes de ocurrencia del delito, jurisdicción policial, día de la semana de ocurrencia del hecho, mes de ocurrencia del hecho, rango de horario, tipo de lugar, clase de arma, clase de elemento sustraído, tipo de ataque y sexo de la víctima.

En la Fig. 6.1 se observa la distribución del tipo de delito, en la cual el hecho con mayor porcentaje de delincuencia registra 285 casos de robo y representa el 78% de la cantidad total de delitos, seguida de unos 81 registros de hurto y representa el 22% restante de la totalidad.

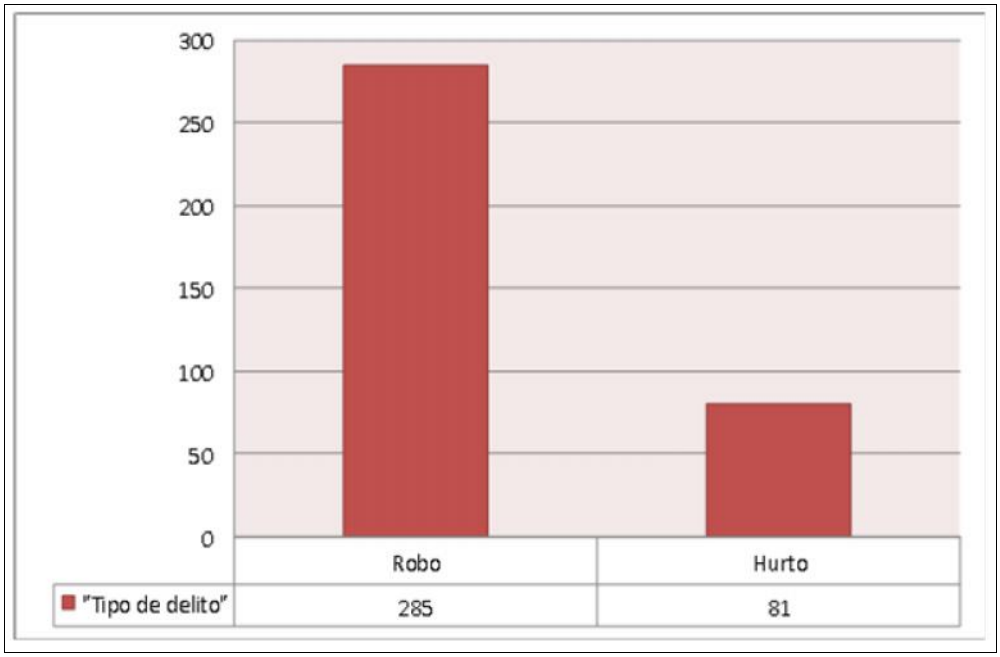


Fig. 6.1. Distribución de delitos por tipo de delito:

Fuente: elaboración propia

Se analiza la alta concentración de ocurrencia de delitos durante el fin de semana (Fig. 6.2). El día de la semana con mayor cantidad de casos se registró el día sábado con 111 casos y el de menor cantidad el día lunes con 32 casos.

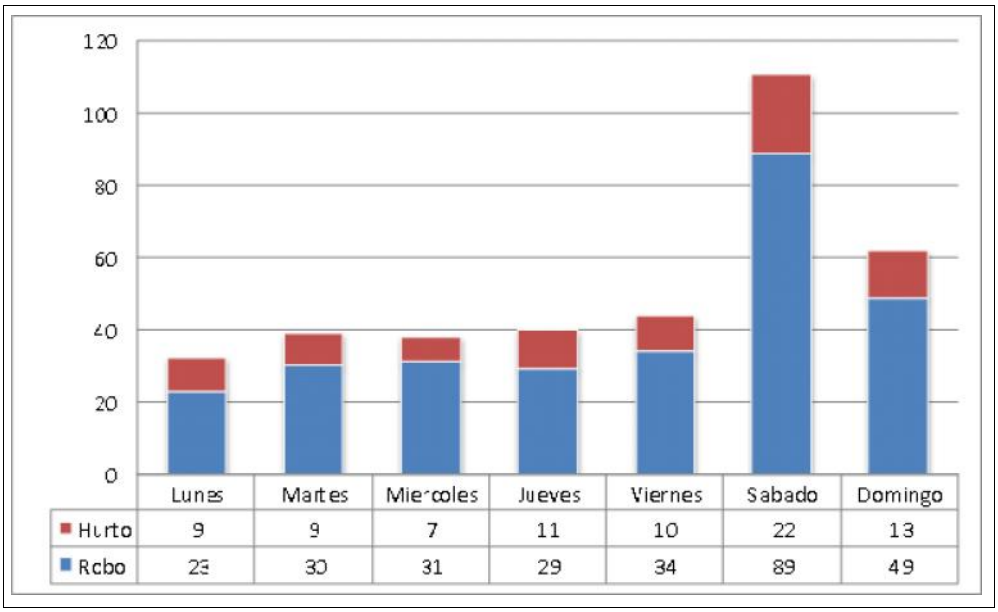


Fig. 6.2. Distribución de delitos por día de la semana:

Fuente: elaboración propia

Como se visualiza en la Fig. 6.3, los meses de abril, marzo y enero del año 2017 presentan la mayor cantidad de casos de delitos registrados. Éstos constituyen el 29%, 22% y 17% respectivamente sobre el total de casos, y con un registro de sólo 20 delitos en el mes de junio el cual equivale al 5% del total de hechos delictivos.

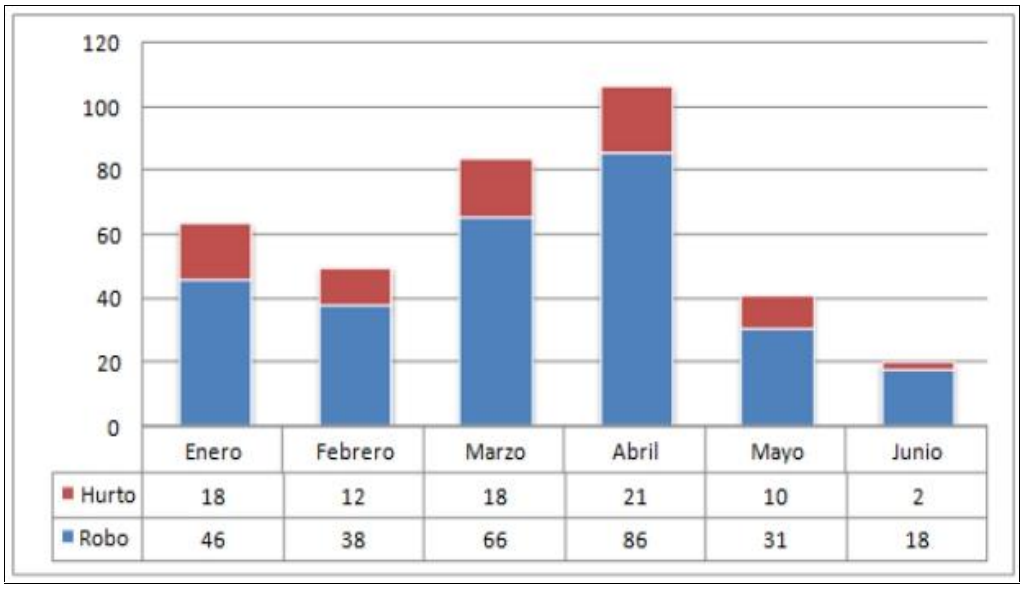


Fig. 6.3. Distribución de delitos por mes:

Fuente: elaboración propia

La Fig. 6.4 muestra la distribución del delito según la jurisdicción policial. Las zonas 7, 2, 9, 19, 16 representan el mayor porcentaje de delitos con un 13%, 11%, 10%, 10% y

8% respectivamente. La menor cantidad de casos se registró en las zonas 6, 17 y 20, representando tan solo el 0.6%, 0.6% y 0.3% del total de hechos.

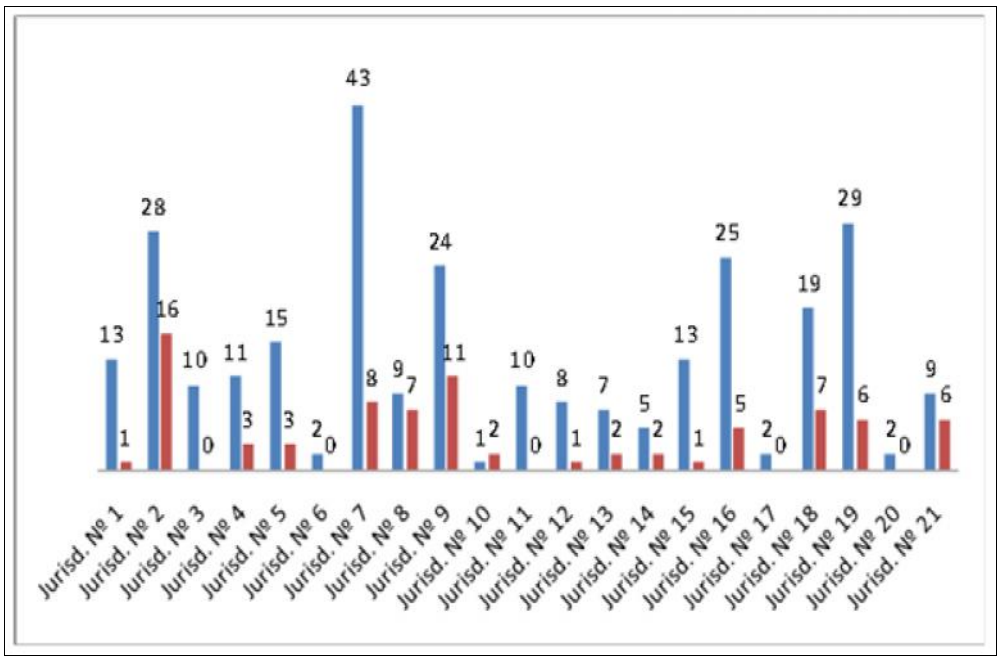


Fig. 6.4. Distribución del delito por jurisdicción policial:

Fuente: elaboración propia

Se muestran los valores de la distribución del delito por rango de horario cometido el hecho (Fig. 6.5). Se observa que las horas más frecuentes de ocurrencia son entre las 12:00hs. y las 16:00hs., con un total de 107 casos, equivalente al 29% del total de hechos.

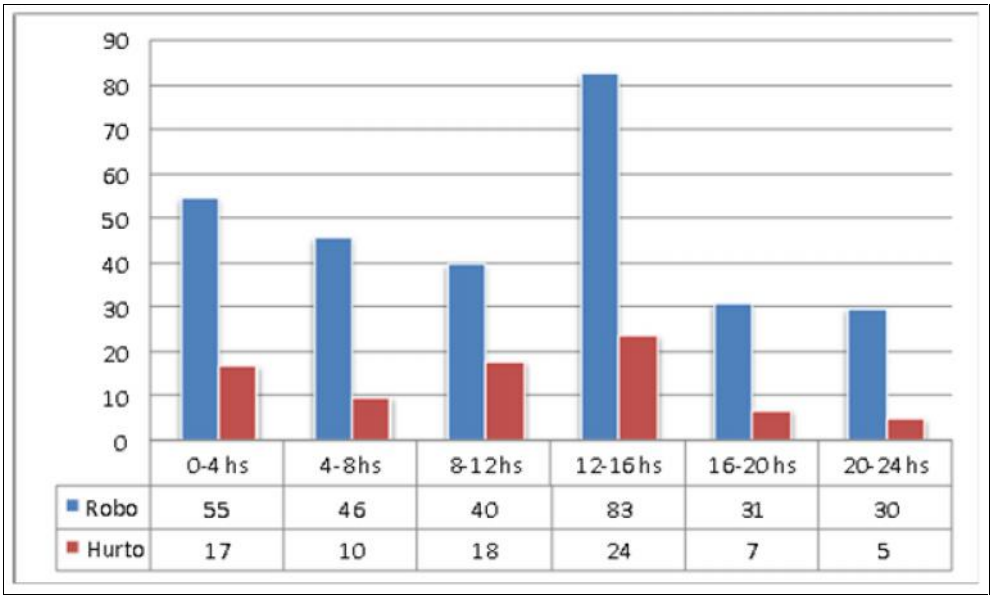


Fig. 6.5. Distribución del delito por rango de horario:

Fuente: elaboración propia

Según la distribución del delito por tipo de lugar ocurrido el delito (Fig. 6.6) se puede observar que la mayor cantidad de casos sucedieron en la vía pública (272 registros), seguido del domicilio particular (46 registros).

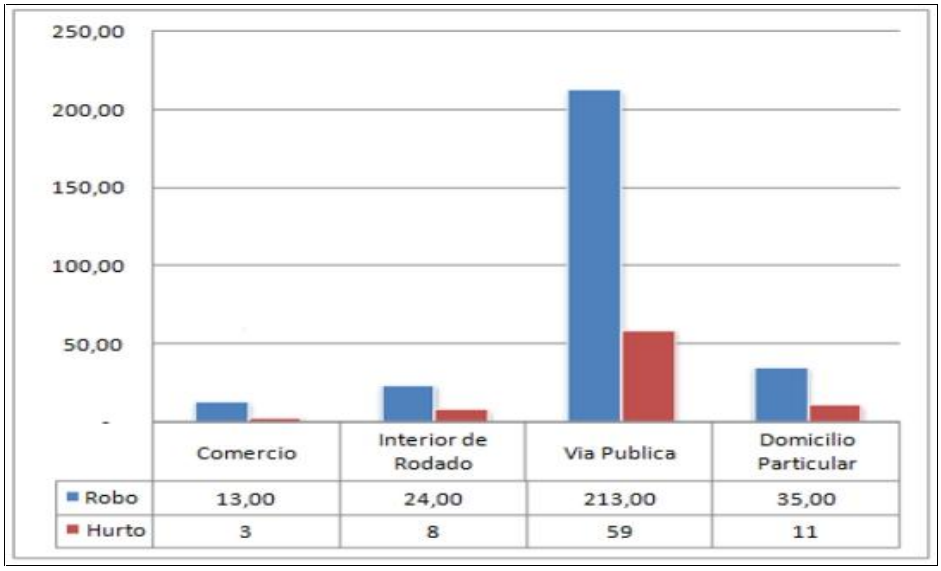


Fig. 6.6. Distribución del delito por tipo de lugar:

Fuente: elaboración propia

En la Fig. 6.7, se visualiza que la mayor cantidad de casos de delitos ocurrieron con arma blanca, existe un total de 211 registros lo cual representa el 58% de la totalidad de casos. Se presentan 81 casos (que corresponden al 22% del total de registros) donde no se registraron ningún tipo de arma, pertenecientes a casos de tipo hurto.

El segundo porcentaje elevado, con un total de 81 registros, representa los casos en donde no se utilizaron armas durante el ataque. Esto se debe o puede deberse a que en su mayoría se trata de delitos de tipo hurto.

Asimismo, en 21 registros (6% del total) se observa que no se informa el tipo de arma.

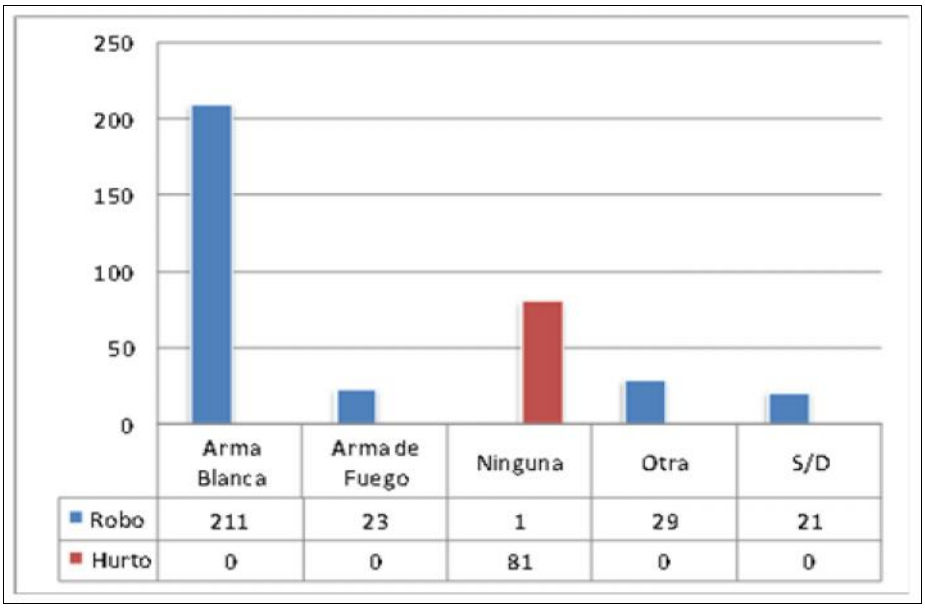


Fig. 6.7. Distribución de delitos por clase de arma:

Fuente: elaboración propia

El análisis de distribución del delito por tipo de elemento sustraído que referencia a objetos personales (Fig. 6.8), representa un total de 224 registros, y equivale a un 61% de la totalidad de casos. Este valor es mayor en comparación con el robo y/o hurto de dinero, que equivale sólo al 0.6% del total con solo 2 registros existentes.

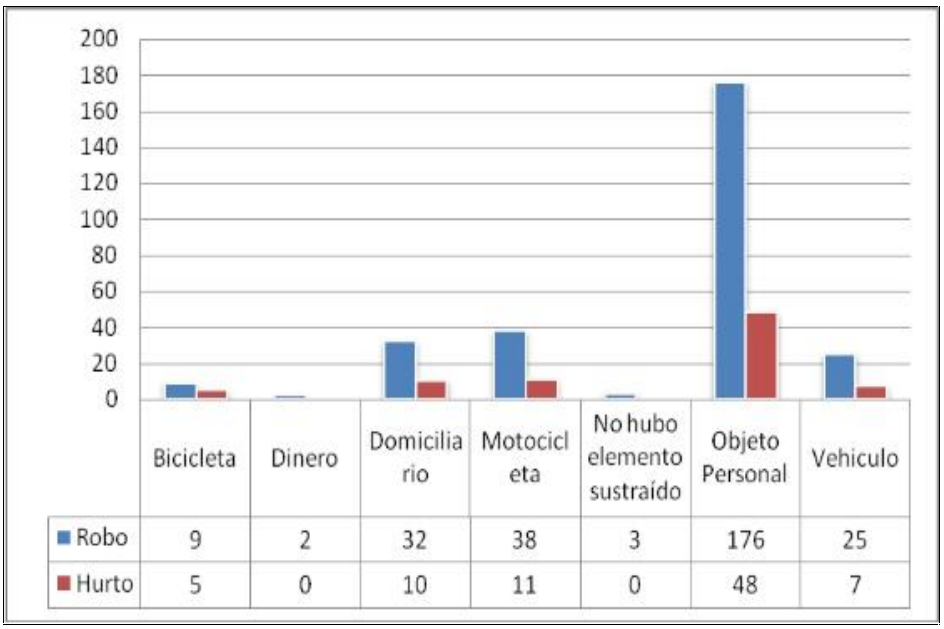


Fig. 6.8. Distribución de delitos por clase de elemento sustraído

Fuente: elaboración propia

La distribución del delito según el tipo de ataque por arrebato presenta la mayor cantidad de casos, con un total de 153 registros (42% del total). A diferencia de otros

campos, existen 111 casos que representan un gran porcentaje del total de registros (30%) en donde no existió ataque. La información permite inferir que en su mayoría se trata de delitos de tipo hurto y no existe ningún tipo de ataque (Fig. 6.9).

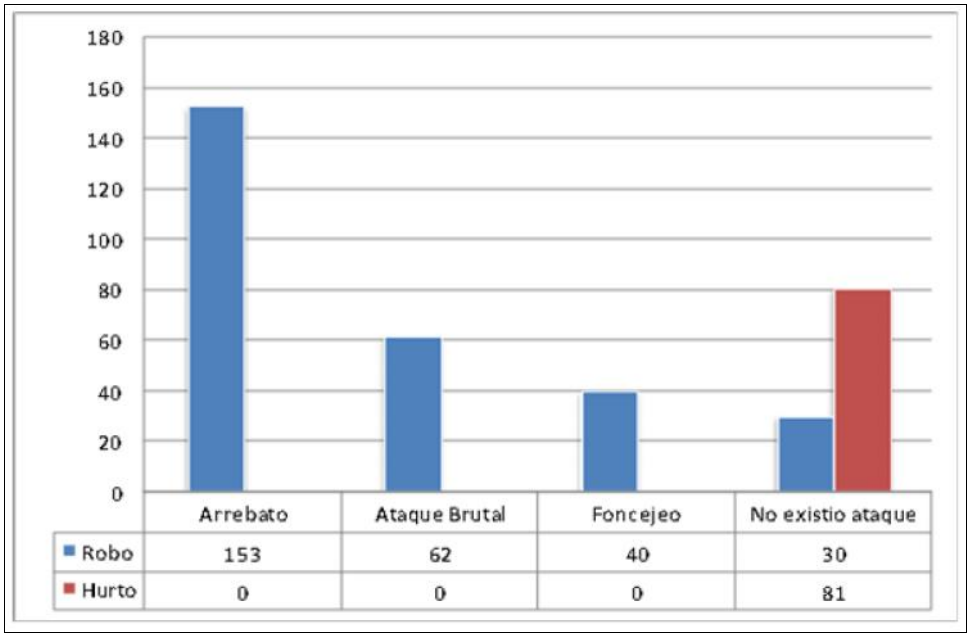


Fig. 6.9. Distribución de delitos por tipo de ataque:

Fuente: elaboración propia

En la Fig. 6.10 se observa un mayor porcentaje (80% de la totalidad) de víctimas que sufrieron un delito del sexo femenino, representadas en 295 registros.

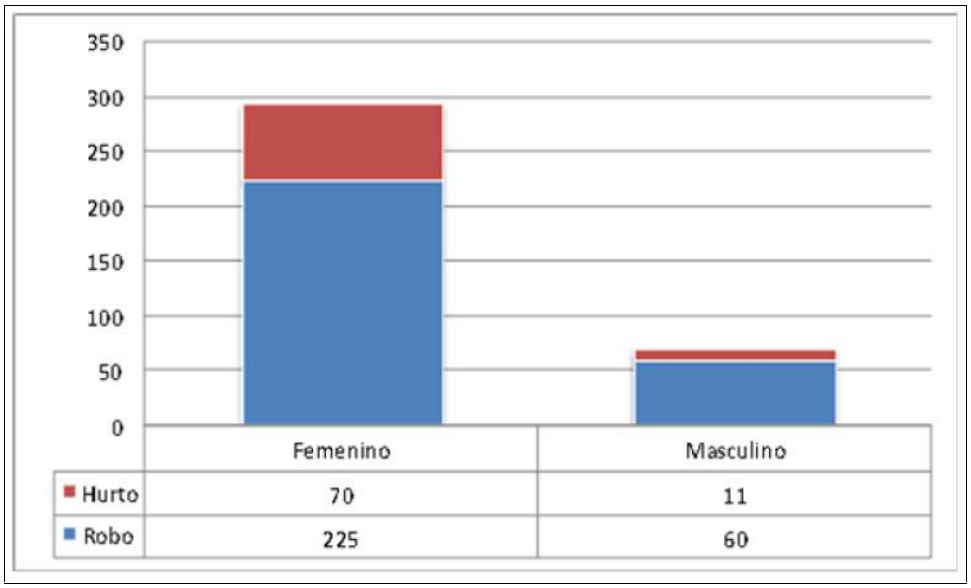


Fig. 6.10. Distribución de delitos por tipo de sexo de la víctima:

Fuente: elaboración propia

Con fines de aportar un detallado análisis, se realizaron gráficos de caja o boxplot con los cuales se determinó el rango de valores de los atributos: jurisdicción policial, rango de horario, tipo de lugar, clase de arma, clase de elemento sustraído y tipo de ataque, los cuales se muestran en las Figs. 6.11, 6.12, 6.13, 6.14, 6.15, y 6.16, respectivamente.

Los valores del gráfico de caja de la distribución del delito por jurisdicción policial son:

- Min (Mínimo valor): 2
- Q1 (Primer Cuartil o la mediana de la mitad menor de los datos): 9
- Q2 (Segundo Cuartil o la mediana de valores): 14
- Q3 (Tercer Cuartil o la mediana de la mitad mayor de los datos): 26
- Max (Máximo valor): 51

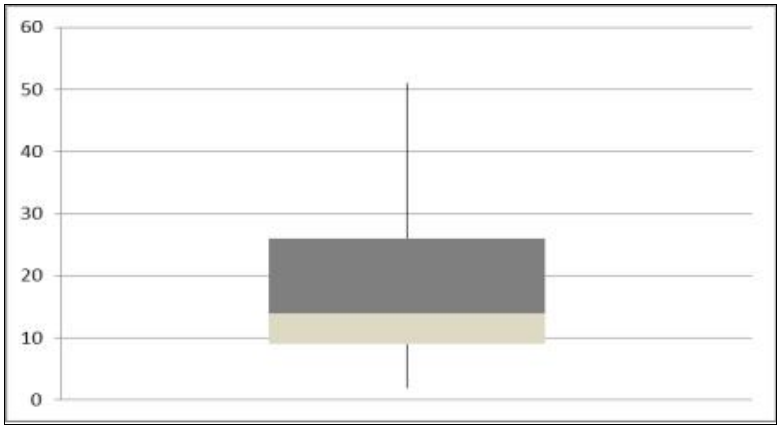


Fig. 6.11. Boxplot de la distribución del delito por jurisdicción policial:

Fuente: elaboración propia

Los valores del gráfico de caja de la distribución del delito por rango de horario son: Min: 35, Q1: 37.25, Q2: 57, Q3: 80.75, Max: 107.

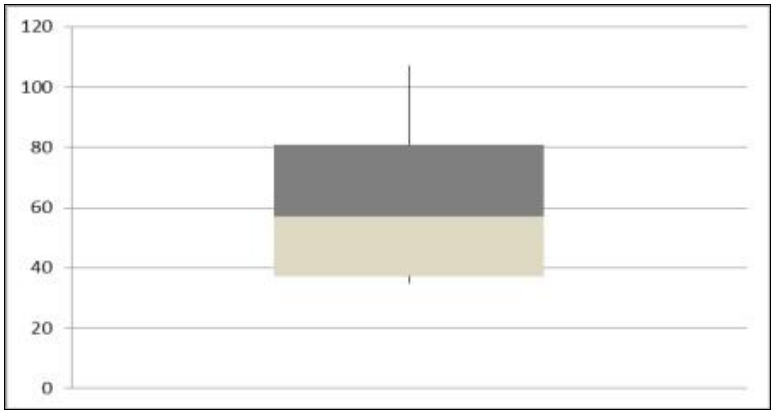


Fig. 6.12. Boxplot de la distribución del delito por rango de horario:

Fuente: elaboración propia

Los valores del gráfico de caja de la distribución del delito por tipo de lugar son: Min: 16, Q1: 28, Q2: 39, Q3:102.5, Max: 272.

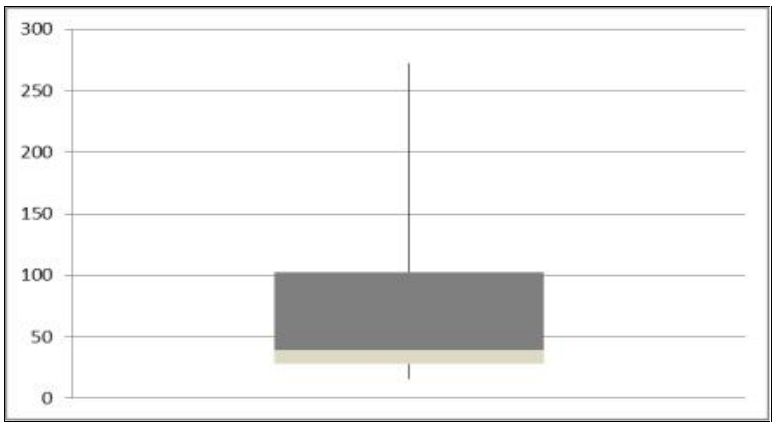


Fig. 6.13. Boxplot de la distribución del delito por tipo de lugar:
Fuente: elaboración propia

Los valores del gráfico de caja de la distribución del delito por clase de arma son: Min: 21, Q1: 22, Q2: 29, Q3: 146.5, Max: 211.

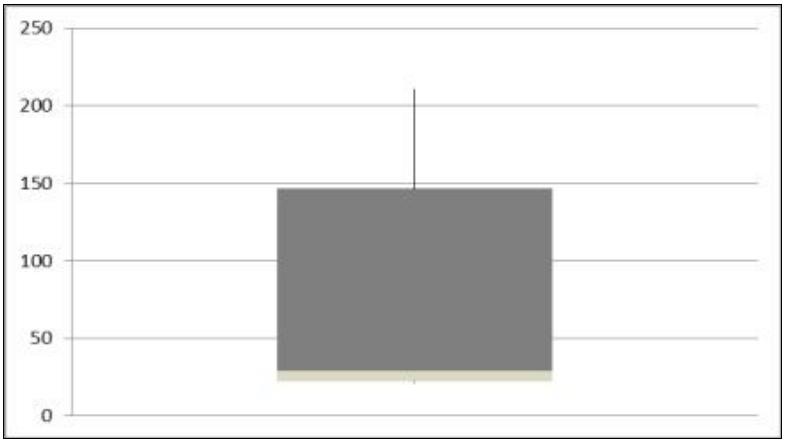


Fig. 6.14. Boxplot de la distribución del delito por clase de arma:
Fuente: elaboración propia

Los valores del gráfico de caja de la distribución del delito por tipo de elemento sustraído son: Min: 2, Q1: 8.5, Q2: 32, Q3: 45.5, Max: 224.

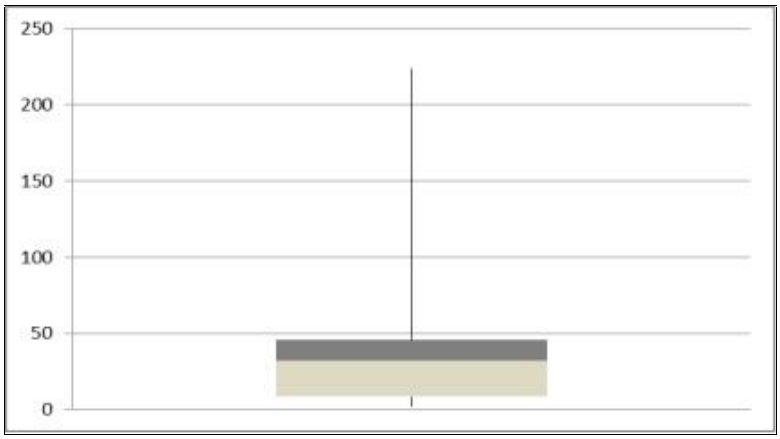


Fig. 6.15. Boxplot de la distribución del delito por tipo de elemento sustraído:

Fuente: elaboración propia

Los valores del gráfico de caja de la distribución del delito por tipo de ataque son: Min: 40, Q1: 45.5, Q2: 86.5, Q3: 142.5, Max: 153.

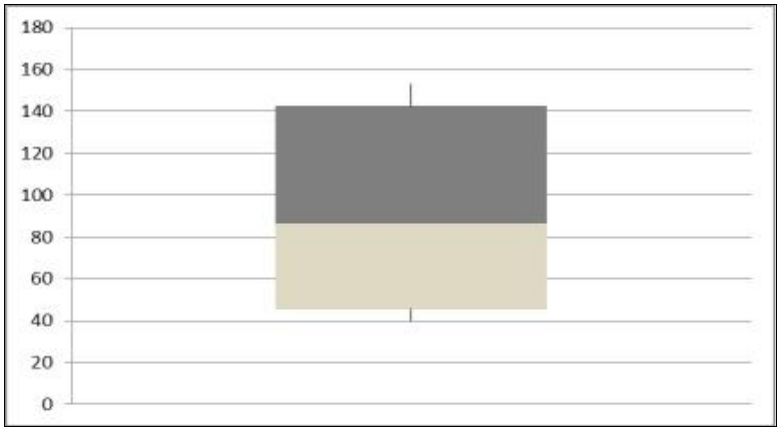


Fig. 6.16. Boxplot de la distribución del delito por tipo de ataque:

Fuente: elaboración propia

6.1.1.2.3. Verificación de calidad de los datos

Finalizada la exploración inicial de los datos se verificó que estos se encuentran completos. Sólo se observaron 22 casos de registros con datos nulos en la variable sospechoso_edad que no representa grandes problemas en cuanto a su calidad.

Se puede afirmar, además, que las variables carecen valores atípicos o datos extremos.

6.1.1.3. Preparación de los datos

En la tercera fase de la metodología CRISP-DM se realizó el proceso para la preparación de datos finales, en donde se detalla la selección, limpieza, construcción y formateo de datos.

6.1.1.3.1. Selección de los datos

El motivo para la inclusión o exclusión de algunos campos, se sustentó en la importancia de los mismos en relación con los objetivos de la minería de datos que se definieron en la fase comprensión del negocio de la metodología CRISP-DM. Por ello, se excluyeron los siguientes atributos, dado que carece mayor relevancia en el proceso de minería de datos expuesto en este TFM:

id_provincia, provincia_descrip, id_departamento, departamento_descrip, id_municipio, municipio_descrip, id_jurisdic_policial, id_comisaria, id_delito, f_dia, f_anio, id_tipo_lugar, id_clase_arma, id_elemento_sustraido, id_tipo_ataque, id_fue_lesionado, id_sospechoso_sexo, id_victima_sexo, id_victima_nacionalidad, id_victima_ocupacion.

Se descartaron todos los atributos de identificación única de la base de datos, y, la descripción de la provincia, del departamento y del municipio, dado que estos campos resultaron irrelevantes para el análisis.

Cabe recordar que el presente estudio se centró en una zona en particular (capital de la Provincia de Corrientes). El campo año también se eliminó dado que el total de datos pertenecen al año 2017. El campo f_dia (día del mes) también se suplantó por el campo dia_m (día de la semana) el cual resultó de la transformación del primero por considerarlo más representativo para el análisis.

En este TFM, se descartaron los campos que no aportaron información dado que mantienen el mismo valor, y otros se transformaron en un nuevo dato con más valor (sección 6.1.1.3.2) según el conocimiento del experto.

Se propone como futuro trabajo incluir algoritmos para determinar las variables relevantes como actividad preliminar a los procesos de explotación de la información.

6.1.1.3.2. Limpieza de los datos

La base de datos contiene toda la información necesaria para cumplir los objetivos de la minería de datos. Sólo se observan 22 casos de registros con datos nulos en la variable sospechoso_edad que no representa grandes problemas en cuanto a su calidad, dada la cantidad mínima de registros y para conservarlos se reemplazó por el valor 0.

6.1.1.3.3. Construcción de los datos

Para lograr un mejor análisis, a partir de la variable f_mes, se creó la variable mes_r, la cual determina una serie de rangos de 2 meses cada uno, para obtener los rangos del mes del delito. Esta variable se representa por los valores de la Tabla VIII:

Tabla VIII

Nuevos estados del atributo mes_r:

Fuente: elaboración propia

Valor	Descripción
1	enero-febrero
2	marzo-abril
3	mayo-junio

Se analizó la variable f_hora y se creó la variable horas_r, la cual determina una serie de rangos de 4 horas cada uno, para obtener los rangos de horarios del delito. Los valores de esta variable se presentan en la Tabla IX:

Tabla IX

Nuevos estados del atributo horas_r:

Fuente: elaboración propia

Valor	Descripción
1	0-4 hs
2	4-8 hs
3	8-12 hs
4	12-16 hs
5	16-20 hs
6	20-24 hs

La variable sospechoso_edad representa la edad aproximada del sospechoso implicado y cuyos valores se encuentran entre los 15 y 35 años. Para un mejor análisis, se modificaron estos valores a una serie de rangos de 5 años de edad y se obtuvieron los rangos de años del sospechoso, la misma fue denominada sospechoso_edad_r. Existen casos nulos que se reemplazaron por el valor 0. Esta variable se representa por los siguientes valores indicados en la Tabla X.

Tabla X

Nuevos estados del atributo sospechoso_edad_r:

Fuente: elaboración propia

Valor	Descripción
0	Valor 0
1	15-20años
2	21-25años
3	26-30 años
4	31-35 años

La variable victima_edad contiene la edad de la víctima que sufrió el ataque y cuyos valores se encuentran entre los 15 y 57 años. Para un mejor análisis, se modificaron estos valores a una serie de rangos de 5 años de edad, se obtuvieron los rangos de años de la víctima, denominada victima_edad_r, la cual se observa en la Tabla XI.

Tabla XI

Nuevos estados del atributo victima_edad_r:

Fuente: elaboración propia

Valor	Descripción
1	15-20 años
2	21-25 años
3	26-30 años
4	31-35 años
5	36-40 años
6	41-45 años
7	46-50 años
8	51-57 años

6.1.1.3.4. Integración de los datos

Finalmente, el conjunto de datos se compuso por los 366 registros originales y 20 atributos, los cuales se sintetizan en las Tablas XII.a y XII.b.

Tabla XII.a

Variables del conjunto de datos final:

Fuente: elaboración propia

Atributos	Tipo	Valores Posibles	Longitud del campo
id_registro	Entero	[1-366]	Máximo de 5 enteros
jurisdic_policial	Varchar	[Seccional 1ª- Seccional 21ª]	Máximo de 30 caracteres
comisaria_descrip	Varchar	[Comisaría Primera Urbana Capital- Comisaría Vigésimo Primero Urbana Capital]	Máximo de 70 caracteres
delito_descrip	Varchar	[Robo - Hurto]	Máximo de 30 caracteres
dia_m	Varchar	[Lunes - Domingo]	Máximo de 20 caracteres
mes_r	Entero	[1- 3]	Máximo de 2 enteros
horas_r	Entero	[1-6]	Máximo de 2 enteros
barrio_descrip	Varchar	[17 de Agosto –Yapeyu]	Máximo de 50 caracteres
calle	Varchar	[22 de Mayo –Turin]	Máximo de 70 caracteres
altura	Entero	[40 – 6400]	Máximo de 4 enteros
tipo_lugar	Varchar	[Vía Pública, Domicilio Particular, Comercio, Interior de Rodado, Otro Lugar]	Máximo de 70 caracteres

Tabla XII.b

Variables del conjunto de datos final:

Fuente: elaboración propia

Atributos	Tipo	Valores Posibles	Longitud del campo
clase_arma	Varchar	[Arma de Fuego, Arma Blanca, Otra, Ninguna, S/D]	Máximo de 50 caracteres
elemento_sustraido	Varchar	[Domiciliario, Vehículo, Motocicleta, Bicicleta, Objeto Personal, Otros, No hubo elemento sustraído]	Máximo de 70 caracteres
tipo_ataque	Varchar	[Arrebato, Forcejeo, Ataque Brutal, No existió ataque]	Máximo de 50 caracteres
fue_lesionado_descrip	Varchar	[No fue lesionado – Si fue lesionado]	Máximo de 50 caracteres
sospechoso_sexo_descrip	Varchar	[Masculino, Femenino, S/D]	Máximo de 15 caracteres
sospechoso_edad_r	Entero	[0–4]	Máximo de 2 enteros
victima_sexo_descrip	Varchar	[Masculino, Femenino]	Máximo de 15 caracteres
victima_edad_r	Entero	[1-8]	Máximo de 2 enteros
victima_ocupacion_descrip	Varchar	[Tiene algún oficio, Tiene alguna profesión, No tiene ni oficio ni profesión]	Máximo de 50 caracteres
victima_ocupacion_descrip	Varchar	[Tiene algún oficio, Tiene alguna profesión, No tiene ni oficio ni profesión]	Máximo de 50 caracteres

6.1.1.3.4. Formatear los datos

Para esta etapa de formateo de datos no se ajustaron los valores de los campos dado que las herramientas software utilizadas permiten manipular datos continuos como discretos. Además, los datos no contienen comas ni caracteres especiales.

6.1.1.4. Modelado

En esta cuarta fase de la metodología CRISP-DM, se realizó la selección de la técnica de modelado para aplicar minería de datos, el diseño de las pruebas a ejecutar sobre el modelo, y la construcción y evaluación de los mismos.

6.1.1.4.1. Selección de técnicas de modelado

Para cada uno de los objetivos de minería de datos planteados se propuso la aplicación de los siguientes procesos de explotación de información que comprende la combinación de las siguientes técnicas de modelado:

- SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos: Para el descubrimiento de reglas de pertenencia a grupos se propone, para el hallazgo de los mismos, la utilización de mapas auto-organizados (SOM) y, una vez identificados los grupos, la utilización de algoritmos de inducción (TDIDT) con el objeto de establecer las reglas de pertenencia a cada uno. En esta técnica se utilizó el algoritmo Kohonen-SOM para el descubrimiento de grupos y el

algoritmo C4.5 para la caracterización o descubrimiento de reglas de cada clúster.

- Redes bayesianas aplicadas a la ponderación de interdependencia entre atributos: Para ponderar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase se propone la utilización de Redes Bayesianas. En esta técnica se utilizó el algoritmo clasificador probabilístico Naive Bayes para la ponderación de los atributos.

Existen múltiples estudios que señalan la variación de la calidad de los resultados obtenidos por los algoritmos según el dominio y las características del proyecto [56]. Este TFM sustenta la elección de los algoritmos utilizados en la propuesta de procesos de explotación de información descripta en [17].

En adición, la elección del algoritmo C4.5 y no de otros algoritmos como Random Forest por ejemplo, se justifica en lo siguiente:

- Interpretabilidad: Dado que comprender la característica del patrón identificado es algo requerido, el algoritmo C4.5 es mejor en este aspecto que Random Forest. Si bien ambos modelos pueden aportar interpretabilidad, C4.5 es una versión más simple que Random Forest. En este contexto, C4.5 permite leer las reglas que explican la decisión, mientras que Random Forest requiere de aplicar técnicas de reducción del bosque a un árbol las cuales son más complejas y se podría perder calidad de respuesta.
- Simplicidad y Generalización: Random Forest es un modelo más complejo, y suele requerir de mayor cantidad de registros para generalizar correctamente y no realizar sobreajuste. Considerando la cantidad de registros, es una decisión esperable utilizar un modelo más simple el cuál pueda explicar el comportamiento requerido.

En resumen, el algoritmo C4.5 posee entre sus características: la simplicidad de su modelo, la capacidad de identificar patrones no lineales, y su interpretabilidad (requisito relevante para el problema planteado).

Sin embargo, se destaca que, si bien no se identificaron estudios del comportamiento de los algoritmos en contextos de datos georreferenciados, el mismo presenta un área de interés para ampliar el análisis del proyecto en futuros trabajos.

Para la construcción de los modelos y su análisis se utilizaron las variables del conjunto de datos final (Tablas XII.a y XII.b.).

6.1.1.4.2. Diseñar las pruebas del modelo

En esta sección se presenta el plan ejecutado para probar la calidad y el contenido del modelo de minería de datos generado a partir de las técnicas seleccionadas. Para evaluar los modelos se utilizaron métricas de la robustez de los mismos [57]:

Se debe analizar cada modelo construido de manera de asegurar que la solución obtenida resuelva eficientemente los objetivos de explotación de información. Esto implica medir y evaluar la calidad de los modelos de la manera más precisa posible, y así garantizar la aplicación y resultados obtenidos de los mismos.

El modelo de descubrimiento de grupos tiene por objetivo la separación representativa de los datos en grupos o clases, sin ningún criterio de agrupamiento a priori, basándose en la similitud de los valores de sus atributos. Todos los datos de un mismo grupo deben tener características comunes, pero a su vez entre los grupos los objetos deben ser diferentes. El factor de calidad del modelo generado se basa en el número de clusters definidos inicialmente. Una forma de evaluar el número correcto de clusters del algoritmo de agrupamiento es utilizar el dendograma o HAC (Hierarchical Aglomerative Clustering o agrupación aglomerada jerárquica). Este método permite la observación de los aglomerados sucesivos y brinda según el análisis de distancias entre los grupos la ayuda necesaria para decidir cuál es el número de clusters más significativo según el conjunto de datos.

Para el modelo de descubrimiento de reglas, el objetivo es obtener un conjunto potencialmente útil de reglas que determinen correctamente una clase. Las métricas utilizadas para este modelo basan su análisis en una serie de fórmulas que determinan la presión de las reglas descubiertas [58]. A su vez, se analiza la matriz de confusión generada por el modelo, la cual permite conocer la clasificación exacta por medio de valores altos en la diagonal principal y valores bajos, en la diagonal secundaria [58].

El modelo de descubrimiento de dependencias significativas consiste en identificar características o factores que tienen mayor incidencia sobre un determinado resultado de un problema, y al igual que con el modelo de descubrimiento de reglas, este modelo también se relaciona con tareas de clasificación de datos y predicción.

Las formulas aplicables a los modelos de descubrimiento de reglas y de dependencias significativas se define por la siguiente expresión [57]:

$$1: EXCT(M)\ \mathrm{X} \quad \frac{NCVCi}{NTC(M)}$$

Dónde:

- EXCT (M) es la exactitud del modelo para clasificar clases o grupos.
- NTC (M) es el número total de casos a utilizar para el modelo.
- NCVCi es el número de casos pertenecientes a la clase Ci (o grupo) correctamente clasificados por el modelo en esa misma clase.

Otra fórmula utilizada en los modelos de descubrimiento de reglas y de dependencias significativas es la tasa de error total, calculada a partir de los valores de la matriz de confusión como se describe a continuación [58]:

$$2 = 1\,\mathrm{Z}(NCVA\,\Gamma\,NCVB) / NTC$$

Dónde:

- NCVA es el número de casos pertenecientes a la clase A (o grupo) correctamente clasificados por el modelo en esa misma clase.
- NCVB es el número de casos pertenecientes a la clase B (o grupo) correctamente clasificados por el modelo en esa misma clase.
- NTC es el número total de casos a utilizar para el modelo.

6.1.1.4.3. Construir el modelo

Se procedió a ejecutar las técnicas de modelado elegidas en 6.1.1.4.1 sobre el conjunto total de dados. En este apartado se describen los ajustes de parámetros de cada técnica de modelado y su ejecución, y puesto que se definieron tres objetivos para la minería de datos, esta sección se dividió en tres partes sub-secciones, una por cada objetivo.

6.1.1.4.3.1. Construcción del modelo para el objetivo de minería de datos Nº 1:

En la Tabla XIII se muestran los valores de los parámetros para las técnicas SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos y, redes bayesianas aplicadas a la ponderación de interdependencia entre atributos:

Tabla XIII

Parámetros seleccionados para el objetivo de minería de datos Nº 1:

Fuente: elaboración propia

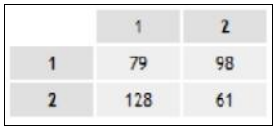
Algoritmo		Justificación
Kohonen SOM		
Row Size	2	Valor por defecto
Col Size	2	Valor por defecto
Distance Normalization	Variance	
Seed Row	Standard	
C4.5		
Min Size of Leaves	5	Valor por defecto
Confidence Level	0.25	Valor por defecto
Naive Bayes		
Use laplacian probestimate	Yes	
Lambda	1	Valor por defecto

A continuación, se presentan los resultados del descubrimiento de reglas de pertenencia a grupos, para identificar y caracterizar grupos que definen el comportamiento de los delitos.

6.1.1.4.3.1.1. Formación de clusters

Para la formación de grupos se definen los atributos de entrada del algoritmo Kohonen SOM: dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque.

Se aplico el algoritmo y se obtuvo cuatro grupos, de los cuales 79 registros forman parte del clúster c_som_1_1, 98 registros conforman el clúster c_som_1_2, 128 forman el clúster c_som_2_1 y 61 del clúster c_som_2_2 (Fig. 6.17).



	1	2
1	79	98
2	128	61

Fig. 6.17. Clusters obtenidos por SOM para identificar el comportamiento del delito:

Fuente: elaboración propia

El algoritmo Kohonen SOM agrupó los clusters basándose en la similitud de los valores de sus atributos. Todos los datos de un mismo grupo deben tener características comunes, pero a su vez, los objetos entre los grupos deben ser diferentes.

6.1.1.4.3.1.2. Descubrimiento del comportamiento del delito

Se establecen a continuación, los atributos para la ejecución del algoritmo C4.5. El atributo clase se define como la variable grupo generada por el algoritmo Kohonen SOM en la sección 6.1.1.4.3.1.1, y los atributos de entrada seleccionados fueron dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque.

Las reglas descubiertas se observan en las Tablas XIV.a, XIV.b y XV.c:

Tabla XIV.a

Reglas generado por algoritmo TDIDT para determinar el comportamiento del delito:

Fuente: elaboración propia

Número de regla	Descripción
1	SI d2c_elemento_sustraido_1 < 2,5000 Y d2c_tipo_lugar_1 >= 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1
2	SI d2c_elemento_sustraido_1 >= 2,5000 Y d2c_tipo_lugar_1 >= 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_2_1
3	SI d2c_tipo_ataque_1 < 1,5000 Y mes_r < 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_2
4	SI horas_r >= 4,5000 Y d2c_tipo_ataque_1 >= 1,5000 Y mes_r < 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_2
5	SI d2c_tipo_lugar_1 < 2,5000 Y d2c_clase_arma_1 < 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_2
6	SI d2c_tipo_lugar_1 >= 2,5000 Y d2c_clase_arma_1 < 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_1
7	SI d2c_elemento_sustraido_1 < 1,5000 Y d2c_clase_arma_1 >= 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_1_1
8	SI d2c_dia_m_1 >= 3,5000 Y horas_r < 2,5000 Y mes_r >= 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1

Tabla XIV.b

Reglas generado por algoritmo TDIDT para determinar el comportamiento del delito:

Fuente: elaboración propia

Número de regla	Descripción
9	SI d2c_clase_arma_1 >= 2,0000 Y d2c_dia_m_1 < 3,5000 Y horas_r < 2,5000 Y mes_r >= 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1
10	SI d2c_tipo_lugar_1 >= 1,5000 Y d2c_elemento_sustraido_1 >= 1,5000 Y d2c_clase_arma_1 >= 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_2
11	SI horas_r >= 4,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_elemento_sustraido_1 >= 1,5000 Y d2c_clase_arma_1 >= 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_2
12	SI mes_r < 1,5000 Y horas_r < 4,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_elemento_sustraido_1 >= 1,5000 Y d2c_clase_arma_1 >= 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_2
13	SI d2c_clase_arma_1 < 2,0000 Y d2c_dia_m_1 < 3,5000 Y horas_r < 2,5000 Y mes_r >= 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_2
14	SI mes_r >= 1,5000 Y horas_r < 4,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_elemento_sustraido_1 >= 1,5000 Y d2c_clase_arma_1 >= 1,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_1
15	SI horas_r < 4,5000 Y d2c_tipo_ataque_1 >= 1,5000 Y mes_r < 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_2_2
16	SI horas_r >= 2,5000 Y mes_r >= 1,5000 Y d2c_dia_m_1 < 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_2

Tabla XIV.c

Reglas generado por algoritmo TDIDT para determinar el comportamiento del delito:

Fuente: elaboración propia

Número de regla	Descripción
17	SI d2c_dia_m_1 >= 4,5000 Y d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1

A través de la clasificación resultado de aplicar el algoritmo C4.5, se obtuvieron 17 reglas que caracterizan a los grupos identificados, los antecedentes de cada clúster se detallan en el Anexo 2. A partir de las reglas generadas, se realizó la interpretación de los clusters:

- Clúster c_som_1_1: Caracterizado por delitos mayoritariamente por medio del forcejeo o arrebato. Incluye casos en la vía pública o en un domicilio particular. Registra mayor actividad delictiva los días miércoles, jueves y viernes, durante los meses de marzo, abril, mayo y junio.
- Clúster c_som_1_2: En principio se trataría de delitos a través del forcejeo o arrebato y con arma blanca. Incluye casos de ocurrencia en la vía pública o en un domicilio particular. Además, se observa la característica que los delitos ocurrieron durante los días sábado, domingo, lunes o martes, en los meses de marzo, abril, mayo y junio.
- Clúster c_som_2_1: Es el que más delitos agrupa. Se determinó que el delito aconteció en algunos casos en forma de ataque brutal y en otros se determinó que no existió ningún tipo de ataque. Predomina el uso de arma fuego y en otros casos, se determinó que no existió ningún elemento de ataque, e incluye lugares del hecho como en la vía pública, en un comercio, en el interior de un rodado, o en un domicilio particular. Este grupo está caracterizado por el robo objetos personales, motocicletas, de tipo domiciliario, vehículos y dinero. No obstante, en otros casos también se observa que no hubo elemento sustraído.
- Clúster c_som_2_2: Este grupo a diferencia del resto, posee mayor descripción de las características de los delitos. Particularmente los delitos ocurrieron con ataque brutal a las víctimas y en otros casos, no existió ningún tipo de ataque. El arma fuego y arma blanca fueron los elementos predominantes en este grupo, así mismo, se determinó que también resultaron casos en donde no hubo elemento

de ataque. Este grupo incluye casos de ocurrencia en vía pública o en algún domicilio particular, en horarios de mañana o siesta. Asimismo, los meses que caracterizan a este grupo son enero y febrero.

Los elementos sustraídos durante el delito que definen a este grupo son de tipo objetos personales, motocicletas, de tipo domiciliario, vehículos y dinero. En otros casos de igual forma se observa que no hubo elemento sustraído.

6.1.1.4.3.1.3. Ponderación de atributos con mayor incidencia en el comportamiento de delitos.

Se definen a continuación, los atributos para la ejecución del algoritmo Naive Bayes, el atributo clase se identifica como la variable de predicción generada por el algoritmo C4.5 en la sección 6.1.1.4.3.1.2, y los atributos de entrada es establecen como dia_m, mes_r, horas_r, delito_descrip, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque.

Los valores de incidencia de los atributos se muestran a continuación en las Figs. 6.18, 6.19, 6.20, 6.21, 6.22, 6.23, 6.24 y 6.25. En cada celda se indica el porcentaje de incidencia (porcentaje más alto) de cada atributo en cada clúster.

	Robo	Hurto	Sum
c_som_1_1	0,9405	0,0595	1,0000
c_som_1_2	1,0000	0,0000	1,0000
c_som_2_1	0,5167	0,4833	1,0000
c_som_2_2	0,7313	0,2687	1,0000
Sum	0,7787	0,2213	1,0000

Fig. 6.18. Ponderación de incidencia para el atributo tipo de delito:

Fuente: elaboración propia

	Sabado	Domingo	Lunes	Martes	Miercoles	Jueves	Viernes	Sum
c_som_1_1	0,0476	0,0357	0,0238	0,0952	0,2619	0,2381	0,2976	1,0000
c_som_1_2	0,5263	0,2421	0,1474	0,0842	0,0000	0,0000	0,0000	1,0000
c_som_2_1	0,2583	0,1917	0,0583	0,1250	0,1083	0,1417	0,1167	1,0000
c_som_2_2	0,3881	0,1940	0,1343	0,1194	0,0448	0,0448	0,0746	1,0000
Sum	0,3033	0,1694	0,0874	0,1066	0,1038	0,1093	0,1202	1,0000

Fig. 6.19. Ponderación de incidencia para el atributo día de la semana:

Fuente: elaboración propia

	Via Publica	Domicilio Particular	Interior de Rodado	Comercio	Sum
c_som_1_1	0,9048	0,0238	0,0119	0,0595	1,0000
c_som_1_2	0,9684	0,0316	0,0000	0,0000	1,0000
c_som_2_1	0,3333	0,3167	0,2583	0,0917	1,0000
c_som_2_2	0,9552	0,0448	0,0000	0,0000	1,0000
Sum	0,7432	0,1257	0,0874	0,0437	1,0000

Fig. 6.20. Ponderación de incidencia para el atributo tipo de lugar:

Fuente: elaboración propia

	Arma Blanca	Ninguna	Arma de Fuego	Otra	S/D	Sum
c_som_1_1	0,7024	0,0714	0,0595	0,1548	0,0119	1,0000
c_som_1_2	0,9263	0,0000	0,0316	0,0421	0,0000	1,0000
c_som_2_1	0,1667	0,4833	0,1167	0,0667	0,1667	1,0000
c_som_2_2	0,6567	0,2687	0,0149	0,0597	0,0000	1,0000
Sum	0,5765	0,2240	0,0628	0,0792	0,0574	1,0000

Fig. 6.21. Ponderación de incidencia para el atributo clase de arma:

Fuente: elaboración propia

	Bicicleta	Objeto Personal	Motocicleta	Domiciliario	Vehiculo	No hubo elemento sustraído	Dinero	Sum
c_som_1_1	0,0952	0,8452	0,0238	0,0119	0,0238	0,0000	0,0000	1,0000
c_som_1_2	0,0632	0,8737	0,0211	0,0105	0,0105	0,0211	0,0000	1,0000
c_som_2_1	0,0000	0,3500	0,0833	0,3083	0,2417	0,0000	0,0167	1,0000
c_som_2_2	0,0000	0,4179	0,5224	0,0448	0,0000	0,0149	0,0000	1,0000
Sum	0,0383	0,6120	0,1339	0,1148	0,0874	0,0082	0,0055	1,0000

Fig. 6.22. Ponderación de incidencia para el atributo tipo de elemento sustraído:

Fuente: elaboración propia

	Forcejeo	Arrebato	No existio ataque	Ataque Brutal	Sum
c_som_1_1	0,1429	0,7976	0,0595	0,0000	1,0000
c_som_1_2	0,2737	0,7263	0,0000	0,0000	1,0000
c_som_2_1	0,0167	0,0417	0,7167	0,2250	1,0000
c_som_2_2	0,0000	0,1791	0,2985	0,5224	1,0000
Sum	0,1093	0,4180	0,3033	0,1694	1,0000

Fig. 6.23. Ponderación de incidencia para el atributo tipo de ataque:

Fuente: elaboración propia

	_1_1,00	_2_2,00	_3_3,00	Sum
c_som_1_1	0,2381	0,6071	0,1548	1,0000
c_som_1_2	0,1053	0,5895	0,3053	1,0000
c_som_2_1	0,1917	0,6167	0,1917	1,0000
c_som_2_2	0,4776	0,4627	0,0597	1,0000
Sum	0,2322	0,5792	0,1885	1,0000

Fig. 6.24. Ponderación de incidencia para el atributo rango del mes:

Fuente: elaboración propia

	_1_1,00	_2_2,00	_3_3,00	_4_4,00	_5_5,00	_6_6,00	Sum
c_som_1_1	0,1905	0,2143	0,1429	0,2857	0,1190	0,0476	1,0000
c_som_1_2	0,1263	0,1263	0,1053	0,2947	0,1684	0,1789	1,0000
c_som_2_1	0,2667	0,1250	0,2000	0,2917	0,0583	0,0583	1,0000
c_som_2_2	0,1791	0,1642	0,1791	0,2985	0,0746	0,1045	1,0000
Sum	0,1967	0,1530	0,1585	0,2923	0,1038	0,0956	1,0000

Fig. 6.25. Ponderación de incidencia para el atributo rango de horas:

Fuente: elaboración propia

A través del estudio de las características del delito con mayor incidencia en la ocurrencia del hecho, se obtuvo el siguiente análisis para cada clúster:

- Clúster c_som_1_1: La mayor cantidad de delitos ocurrieron por robo (94%). En la mayoría de los casos se trata de arrebatos (79,76%) y en la vía pública (90,48%). Existe un alto porcentaje de incidencia (84,52%) de la sustracción de algún objeto personal de la víctima. Se detectó al día viernes, como el día de la semana más frecuente (29,76%), y en cuanto a los valores de incidencia por mes, marzo y abril registraron mayor volumen de delitos (60,71%).
- Clúster c_som_1_2: Este clúster se caracteriza por la ocurrencia de robos en su totalidad por arrebato (79,76%) y de objetos personales. La mayoría de los casos ocurridos sucedieron en vía pública (90,48%) y con arma blanca (92,63%). Los hechos delictivos tienen mayor ocurrencia los días sábados y domingos, en el horario de siesta de 12:00 pm a 16:00 pm y de noche de 20:00 pm a 24:00 pm. Los robos tienen mayor ocurrencia en los meses de marzo y abril según los valores de incidencia.
- Clúster c_som_2_1: Este clúster está determinado por la distribución en cantidades iguales en la ocurrencia de ambos tipos de delitos (robo y hurto). El objeto sustraído con mayor incidencia con respecto al resto fue de tipo objeto personal (35%). De los lugares, el más frecuente fue en la vía pública (33,33%).

- Clúster c_som_2_2: Se analizó que los delitos de robo ocurrieron con mayor frecuencia en los meses de enero y febrero, los días sábados y en el horario de siesta de 12:00 pm a 16:00 pm. Este clúster se caracterizó por presentar mayores casos en la vía pública (95%), utilizando arma blanca (65,67%). Los robos ocurrieron en su totalidad por ataque brutal (52,24%). En cuanto a los elementos sustraídos los más frecuentes fueron motocicletas y objetos personales.

Concluido la ejecución de las técnicas se menciona a modo de resumen que, cuatro de los atributos más significativos de los clusters son tipo de objeto sustraído (objeto personal), lugares y tipos de armas utilizadas durante el delito (vía pública y con arma blanca) y tipo de ataque más sufrido (arrebato).

De los cluster c_som_1_2 y c_som_2_2 se han podido identificar días de mayor porcentaje de ocurrencia de delitos (viernes, sábados y domingos) y horarios más susceptibles (horarios de siesta y de madrugada).

6.1.1.4.3.2. Construcción del modelo para el objetivo de minería de datos Nº 2

Para la ejecución de este modelo, previamente se aplicó un filtro sobre el conjunto final de datos considerando aquellas zonas de mayor ocurrencia de delitos (según el análisis de los datos en la sección 6.1.1.2.3 y la distribución del delito por jurisdicción policial detallado en la Fig. 6.4).

En la Tabla XV se muestran los valores de los parámetros para las técnicas SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos y, redes bayesianas aplicadas a la ponderación de interdependencia entre atributos:

Tabla XV

Parámetros seleccionados para el objetivo de minería de datos Nº 2:

Fuente: elaboración propia

Algoritmo		Justificación
Kohonen SOM		
Row Size	2	Valor por defecto
Col Size	2	Valor por defecto
Distance Normalization	Variance	
Seed Row	Standard	
C4.5		
Min Size of Leaves	5	Valor por defecto
Confidence Level	0.25	Valor por defecto
Naive Bayes		
Use laplacian probestimate	Yes	
Lambda	1	Valor por defecto

A continuación, se presentan los resultados del descubrimiento de reglas de pertenencia a grupos, para identificar y caracterizar grupos que definen el comportamiento de los delitos.

6.1.1.4.3.2.1. Formación de clusters

Para la formación de grupos se definen los atributos de entrada del algoritmo Kohonen SOM: judisdic_policial, delito_descrip, barrio_descrip, calle, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque.

Se aplico el algoritmo y se obtuvo cuatro grupos, de los cuales 72 registros forman parte del clúster c_som_1_1, 45 registros el clúster c_som_1_2, 48 el clúster c_som_2_1 y 30 del clúster c_som_2_2 (Fig. 6.26).

	1	2
1	72	45
2	48	30

Fig. 6.26. Clusters obtenidos por SOM para zonas de mayor ocurrencia de delitos:
Fuente: elaboración propia

El algoritmo Kohonen SOM agrupó los clusters basándose en la similitud de los valores de sus atributos. Todos los datos de un mismo grupo deben tener características comunes, pero a su vez, los objetos entre los grupos deben ser diferentes.

6.1.1.4.3.2.2. Caracterización de grupos de las zonas con mayor ocurrencia de delitos

Se establecen a continuación, los atributos para la ejecución del algoritmo C4.5. El atributo clase se define como la variable grupo generada por el algoritmo Kohonen SOM en la sección 6.1.1.4.3.2.1, y los atributos de entrada seleccionados fueron judisdic_policial, delito_descrip, barrio_descrip, calle, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque.

Las reglas descubiertas se observan en las Tablas XVI:

Tabla XVI.a

Reglas generadas por algoritmo TDIDT para determinar el comportamiento de los grupos de las zonas de mayor ocurrencia de delitos

Fuente: elaboración propia

Número de regla	Descripción
1	SI d2c_tipo_lugar_1 < 2,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_1
2	SI d2c_tipo_lugar_1 >= 2,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_2
3	SI d2c_calle_1 >= 47,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_1
4	SI d2c_barrio_descrip_1 >= 20,5000 Y d2c_calle_1 < 47,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_1
5	SI d2c_elemento_sustraido_1 >= 2,0000 Y d2c_barrio_descrip_1 < 20,5000 Y d2c_calle_1 < 47,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_1
6	SI d2c_elemento_sustraido_1 < 2,0000 Y d2c_barrio_descrip_1 < 20,5000 Y d2c_calle_1 < 47,5000 Y d2c_tipo_lugar_1 < 1,5000 Y d2c_barrio_descrip_1 >= 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_1_1
7	SI d2c_tipo_lugar_1 < 1,5000 Y d2c_barrio_descrip_1 < 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_1_1
8	SI d2c_tipo_lugar_1 >= 1,5000 Y d2c_barrio_descrip_1 < 14,5000 Y d2c_delito_descrip_1 >= 1,5000 ENTONCES c_som_2_2
9	SI d2c_delito_descrip_1 < 1,5000 ENTONCES c_som_1_2

A través de la clasificación resultado de aplicar el algoritmo C4.5, se obtuvieron un total de 9 reglas que caracterizan a los grupos identificados según las zonas de mayor ocurrencia de delitos, los antecedentes de cada clúster se detallan en el Anexo 3. A partir de las reglas generadas, se realizó la interpretación de los clusters:

- Clúster c_som_1_1: Caracterizado por robos mayoritariamente sucedidos en los barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13 y N° 14. Contiene casos en la vía pública y registra objetos personales como el elemento más sustraído por el delincuente.
- Clúster c_som_1_2: Caracterizado por la ocurrencia de delitos en su mayoría fue de tipo hurto.
- Clúster c_som_2_1: Se determinó que el delito aconteció en un domicilio particular y en algunos de los siguientes barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26.
- Clúster c_som_2_2: Particularmente los delitos ocurrieron en el interior de rodado o en un comercio, y en algunos de los siguientes barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13 y N° 14.

6.1.1.4.3.2.3. Ponderación de atributos que definen a los grupos según las zonas de mayor ocurrencia de delitos.

Se definen a continuación, los atributos para la ejecución del algoritmo Naive Bayes, el atributo clase se identifica como la variable de predicción generada por el algoritmo C4.5 en la sección 6.1.1.4.3.2.2 y los atributos de entrada es establecen como judisdic_policial, delito_descrip, barrio_descrip, calle, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque.

Los valores de incidencia de los atributos se muestran a continuación en las Figs. 6.27, 6.28, 6.29, 6.30, 6.31 y 6.32. En cada celda se indica el porcentaje de incidencia (porcentaje más alto) de cada atributo en cada clúster.

	Hurto	Robo	Sum
c_som_1_1	0,0000	1,0000	1,0000
c_som_1_2	1,0000	0,0000	1,0000
c_som_2_1	0,0000	1,0000	1,0000
c_som_2_2	0,0000	1,0000	1,0000
Sum	0,2359	0,7641	1,0000

Fig. 6.27. Ponderación de incidencia para el atributo tipo de delito:

Fuente: elaboración propia

	Seccional 7ª	Seccional 19ª	Seccional 2ª	Seccional 16ª	Seccional 9ª	Sum
c_som_1_1	0,2192	0,0959	0,2466	0,2055	0,2329	1,0000
c_som_1_2	0,1739	0,1304	0,3478	0,1087	0,2391	1,0000
c_som_2_1	0,4043	0,3191	0,1489	0,1277	0,0000	1,0000
c_som_2_2	0,2759	0,2414	0,1034	0,1379	0,2414	1,0000
Sum	0,2615	0,1795	0,2256	0,1538	0,1795	1,0000

Fig. 6.28. Ponderación de incidencia para el atributo tipo de jurisdicción policial:

Fuente: elaboración propia

	Via Publica	Domicilio Particular	Interior de Rodado	Comercio	Sum
c_som_1_1	1,0000	0,0000	0,0000	0,0000	1,0000
c_som_1_2	0,7174	0,1739	0,0652	0,0435	1,0000
c_som_2_1	0,7872	0,2128	0,0000	0,0000	1,0000
c_som_2_2	0,0000	0,4138	0,3448	0,2414	1,0000
Sum	0,7333	0,1538	0,0667	0,0462	1,0000

Fig. 6.29. Ponderación de incidencia para el atributo tipo de lugar:

Fuente: elaboración propia

	Ninguna	Arma de Fuego	Arma Blanca	Otra	S/D	Sum
c_som_1_1	0,0000	0,0137	0,8767	0,0959	0,0137	1,0000
c_som_1_2	1,0000	0,0000	0,0000	0,0000	0,0000	1,0000
c_som_2_1	0,0000	0,0426	0,7660	0,0638	0,1277	1,0000
c_som_2_2	0,0000	0,1724	0,5172	0,1034	0,2069	1,0000
Sum	0,2359	0,0410	0,5897	0,0667	0,0667	1,0000

Fig. 6.30. Ponderación de incidencia para el atributo clase de arma:

Fuente: elaboración propia

	Objeto Personal	Domiciliario	Motocicleta	Bicicleta	Vehiculo	Dinero	No hubo elemento sustraído	Sum
c_som_1_1	0,7808	0,0000	0,1507	0,0411	0,0274	0,0000	0,0000	1,0000
c_som_1_2	0,6304	0,1522	0,0870	0,0652	0,0652	0,0000	0,0000	1,0000
c_som_2_1	0,5957	0,2128	0,1064	0,0426	0,0213	0,0000	0,0213	1,0000
c_som_2_2	0,2414	0,3103	0,1034	0,0000	0,3103	0,0345	0,0000	1,0000
Sum	0,6205	0,1333	0,1179	0,0410	0,0769	0,0051	0,0051	1,0000

Fig. 6.31. Ponderación de incidencia para el atributo tipo de elemento sustraído:

Fuente: elaboración propia

	No existio ataque	Forcejeo	Ataque Brutal	Arrebato	Sum
c_som_1_1	0,0000	0,2192	0,1370	0,6438	1,0000
c_som_1_2	1,0000	0,0000	0,0000	0,0000	1,0000
c_som_2_1	0,1702	0,0638	0,1702	0,5957	1,0000
c_som_2_2	0,2759	0,1379	0,3103	0,2759	1,0000
Sum	0,3179	0,1179	0,1385	0,4256	1,0000

Fig. 6.32. Ponderación de incidencia para el atributo tipo de ataque:

Fuente: elaboración propia

A partir de las zonas de mayor ocurrencia de delitos, se detalla a continuación para cada clúster, el análisis de las características de los hechos criminales que presentan mayor incidencia con respecto a otras:

- Clúster c_som_1_1: Este clúster presenta mayor cantidad de hechos delictivos ocurridos en la jurisdicción policial N° 2 (24,66%), seguido de la jurisdicción N° 9 (23,29%). De los barrios con mayor ocurrencia de delitos se observan los siguientes: N° 1 (15,07%), N° 2 (13,70%), N° 3 (13,70%) y N° 4 (10,96%). En concordancia con lo anterior, los hechos delictivos con mayor porcentaje de delitos fueron sobre las siguientes avenidas: N° 1 (10,96%), N° 2 (06,85%), N° 3 (05,48%) y N° 4 (05,48%). Existe un alto porcentaje de incidencia (78,08%) de que el objeto sustraído fue de tipo objeto personal y una totalidad de ocurrencia de delitos en la vía pública. En cuanto a valores más predominantes

con respecto al tipo de arma y forma de ataque, el arma blanca y el arrebato fueron lo más destacados.

- Clúster c_som_1_2: Este clúster se caracteriza por la ocurrencia de hurtos en su totalidad. El mayor porcentaje de hechos delictivos ocurrieron en la jurisdicción policial N° 2 (34,78%), seguido de la jurisdicción N° 9 (23,91%). Los barrios en donde se detectó mayor actividad delictiva fueron sucedidos en: N° 1 (28,26%) y N° 2 (19,57%). En concordancia con lo anterior se reflejan las calles en donde se detectó mayor porcentaje de delitos: N° 1 (10,87%), N° 2 (13,04%). Existe un alto porcentaje de incidencia de que el objeto sustraído fue de tipo objeto personal (63,04%) y en la vía pública (71,74%).

- Clúster c_som_2_1: Este clúster está caracterizado por delitos de tipo robo. El mayor porcentaje de los hechos delictivos ocurrieron en la jurisdicción policial N° 7 (40,43%), seguido de la jurisdicción N° 19 (31,91%). Los barrios donde se detectó un alto porcentaje de delitos fueron: N° 1 (21,28%), N° 2 (10,64%) y N° 3 (10,64%). Las calles de mayores sucesos de delitos fueron: N° 1 (10,96%) y N° 2 (08,04%). El objeto sustraído con mayor incidencia con respecto al resto es de tipo objeto personal (59,57%). De las modalidades de ataque, el más frecuente por arrebato (59,57%). El objeto sustraído de mayor ocurrencia fue de tipo objeto personal y con un 59,57% de que haya sido de tipo arrebatado, en la vía pública (78,72%) y con arma blanca (76,60%).

- Clúster c_som_2_2: Se analizó que los delitos de robo ocurrieron en su totalidad. El mayor porcentaje de los hechos delictivos ocurrieron en la jurisdicción policial N° 7 (27,59%), seguido de la jurisdicción N° 19 (24,14%) y de la jurisdicción N° 9 (24,14%). Los barrios en donde se detectó mayor porcentaje de delitos fueron: N° 1 (17,24%), N° 2 (13,79%), N° 3 (10,34%), N° 4 (10,34%) y N° 5 (10,34%). Las calles de mayor porcentaje de sucesos delictivos fueron: N° 1 (17,24%). Los delitos presentan un porcentaje alto en donde se utilizó un arma blanca (51,72%) durante el hecho. Existe un mayor porcentaje de incidencia (31,03%) de que el objeto sustraído haya sido un vehículo o de tipo domiciliario, en un domicilio particular (41,38%) y con un 31,03% que haya sido de tipo ataque brutal.

Concluido la ejecución de las técnicas se menciona a modo de resumen que, interesa conocer los barrios con mayor frecuencia de delitos, los cuales según los clúster analizados son: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7 y N° 8. Se observa en la mayoría de los casos delitos ocurridos sobre avenidas principales como la N° 1 por ejemplo.

Solo en el clúster c_som_2_1 se presentan hechos de hurto de objetos personales en su totalidad caracterizado por ocurrir en la vía pública. El clúster c_som_2_2, por otra parte y a diferencia del resto, el robo se produce de algún vehículo o dentro de un domicilio particular.

6.1.1.4.3.3. Construcción del modelo para el objetivo de minería de datos Nº 3:

En la Tabla XVII se muestran los valores de los parámetros para las técnicas SOM y TDIDT aplicados al descubrimiento de reglas de pertenencia a grupos y, redes bayesianas aplicadas a la ponderación de interdependencia entre atributos:

Tabla XVII

Parámetros seleccionados para el objetivo de minería de datos Nº 3:

Fuente: elaboración propia

Algoritmo		Justificación
Kohonen SOM		
Row Size	2	Valor por defecto
Col Size	2	Valor por defecto
Distance Normalization	Variance	
Seed Row	Standard	
C4.5		
Min Size of Leaves	5	Valor por defecto
Confidence Level	0.25	Valor por defecto
Naive Bayes		
Use laplacian probestimate	Yes	
Lambda	1	Valor por defecto

A continuación, se presentan los resultados del descubrimiento de reglas de pertenencia a grupos, para identificar y caracterizar grupos que definen a las personas más propensas a sufrir de un delito.

6.1.1.4.3.3.1. Formación de clusters

Para la formación de grupos se definen los atributos de entrada del algoritmo Kohonen SOM: delito_descrip, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque, sospechoso_edad_r, sospechoso_sexo_descrip, victima_edad_r, victima_sexo_descrip y victima_ocupacion_descrip.

Se aplico el algoritmo y se obtuvo cuatro grupos, de los cuales 61 registros forman parte del clúster c_som_1_1, 78 registros conforman el clúster c_som_1_2, 99 forman parte del clúster c_som_2_1 y 128 del clúster c_som_2_2 (Fig. 6.33).

	1	2
1	61	78
2	99	128

Fig. 6.33. Clusters de registros obtenidos por SOM para identificar grupos entre las personas más propensas a sufrir un delito:

Fuente: elaboración propia

El algoritmo Kohonen SOM agrupó los clusters basándose en la similitud de los valores de sus atributos. Todos los datos de un mismo grupo deben tener características comunes, pero a su vez, los objetos entre los grupos deben ser diferentes.

6.1.1.4.3.3.2 Caracterización de grupos entre las personas más propensas a sufrir un delito:

Se establecen a continuación, los atributos para la ejecución del algoritmo C4.5. El atributo clase se define como la variable grupo generada por el algoritmo Kohonen SOM en la sección 6.1.1.4.3.3.1 y se seleccionan los siguientes atributos de entrada: delito_descrip, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque, sospechoso_edad_r, sospechoso_sexo_descrip, victima_edad_r, victima_sexo_descrip y victima_ocupacion_descrip.

Las reglas descubiertas se observan en las Tablas XVIII.a y XVIII.b:

Tabla XVIII.a

Reglas obtenidas por algoritmo TDIDT para caracterizar grupos entre las personas más propensas a sufrir un delito:

Fuente: elaboración propia

Número de regla	Descripción
1	SI victima_edad >= 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_1_1
2	SI victima_edad < 3,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_2_2
3	SI victima_edad >= 3,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_2

Tabla XVIII.b

Reglas obtenidas por algoritmo TDIDT para caracterizar grupos entre las personas más propensas a sufrir un delito:

Fuente: elaboración propia

Número de regla	Descripción
4	SI d2c_delito_descrip_1 < 1,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_1_2
5	SI d2c_victima_sexo_descrip_1 >= 1,5000 Y d2c_delito_descrip_1 >= 1,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_1_2
6	SI d2c_elemento_sustraido_1 < 4,5000 Y d2c_victima_sexo_descrip_1 < 1,5000 Y d2c_delito_descrip_1 >= 1,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_1
7	SI d2c_elemento_sustraido_1 >= 4,5000 Y d2c_victima_sexo_descrip_1 < 1,5000 Y d2c_delito_descrip_1 >= 1,5000 Y d2c_tipo_lugar_1 >= 1,5000 Y victima_edad < 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_1_2
8	SI victima_edad < 3,5000 Y d2c_tipo_lugar_1 < 1,5000 Y victima_edad < 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_2_2
9	SI sospechoso_edad >= 3,0000 Y victima_edad >= 3,5000 Y d2c_tipo_lugar_1 < 1,5000 Y victima_edad < 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_2_2
10	SI sospechoso_edad < 3,0000 Y victima_edad >= 3,5000 Y d2c_tipo_lugar_1 < 1,5000 Y victima_edad < 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1
11	SI d2c_tipo_lugar_1 < 1,5000 Y victima_edad < 6,5000 Y d2c_tipo_ataque_1 >= 2,5000 ENTONCES c_som_2_1
12	SI victima_edad >= 4,5000 Y d2c_tipo_ataque_1 < 2,5000 ENTONCES c_som_1_1

A través de la clasificación, resultado de aplicar el algoritmo C4.5, se obtuvieron un total de 12 reglas que caracterizan a los grupos identificados, los antecedentes de cada clúster se detallan en el Anexo 4. A partir de las reglas generadas, se realizó la interpretación de los clusters:

- Clúster c_som_1_1: Presenta un rango promedio de edad entre los 30 y 60 años de las personas más propensas a sufrir algún robo, las cuales resistieron un arrebato o forcejeo de sus pertenecías. Este grupo se caracteriza por presentar personas sospechosas en su mayoría con una edad promedio entre los 15 y 25 años.
- Clúster c_som_1_2: Determinó un rango promedio de edad entre los 15 y 45 años de las personas más propensas a sufrir algún robo, de sexo femenino en su mayoría. Las mismas se hallaban en su domicilio particular, en el interior de un rodado, o en algún comercio en el instante del robo y sufrieron ataques de forma brutal en algunos casos y en otros no hubo agresión alguna por parte de los delincuentes.
- Clúster c_som_2_1: Este grupo está caracterizado por un rango promedio de edad entre los 15 y 45 años de personas más propensas a sufrir de algún robo. Las mismas se hallaban en la vía pública en el instante del robo y sufrieron ataques de forma brutal en algunos casos y en otros no hubo agresión alguna por parte de los delincuentes.
- Clúster c_som_2_2: Particularmente este grupo presentó un rango promedio de edad entre los 15 y 35 años de las personas más propensas a sufrir de algún robo. Las mismas se hallaban en la vía pública en el instante del robo y sufrieron ataques en forma de arrebato en algunos casos o forcejeo en otros.

6.1.1.4.3.3.3. Identificación de atributos con mayor incidencia en el comportamiento de grupos entre las personas más propensas a sufrir un delito.

Se definen a continuación, los atributos para la ejecución del algoritmo Naive Bayes, el atributo clase se identifica como la variable de predicción generada por el algoritmo C4.5 en la sección 6.1.1.4.3.3.2, y los atributos de entrada es establecen como delito_descrip, tipo_lugar, clase_arma, elemento_sustraido, tipo_ataque, sospechoso_edad_r, sospechoso_sexo_descrip, victima_edad_r, victima_sexo_descrip y victima_ocupacion_descrip.

Los valores de incidencia de los atributos se muestran en las Figs. 6.34, 6.35, 6.36, 6.37, 6.38, 6.39, 6.40, 6.41, 6.42 y 6.43. En cada celda se indica el porcentaje de incidencia (porcentaje más alto) de cada atributo en cada clúster.

	Robo	Hurto	Sum
c_som_1_1	0,9848	0,0152	1,0000
c_som_1_2	0,8169	0,1831	1,0000
c_som_2_1	0,3366	0,6634	1,0000
c_som_2_2	1,0000	0,0000	1,0000
Sum	0,7787	0,2213	1,0000

Fig. 6.34. Ponderación de incidencia para el atributo tipo de delito:

Fuente: elaboración propia

	Via Publica	Domicilio Particular	Interior de Rodado	Comercio	Sum
c_som_1_1	0,9091	0,0303	0,0152	0,0455	1,0000
c_som_1_2	0,0000	0,5070	0,3944	0,0986	1,0000
c_som_2_1	0,9109	0,0495	0,0198	0,0198	1,0000
c_som_2_2	0,9375	0,0234	0,0078	0,0313	1,0000
Sum	0,7432	0,1257	0,0874	0,0437	1,0000

Fig. 6.35. Ponderación de incidencia para el atributo tipo de lugar:

Fuente: elaboración propia

	Arma Blanca	Ninguna	Arma de Fuego	Otra	S/D	Sum
c_som_1_1	0,8636	0,0152	0,0303	0,0758	0,0152	1,0000
c_som_1_2	0,2676	0,1831	0,2113	0,0563	0,2817	1,0000
c_som_2_1	0,2772	0,6634	0,0099	0,0495	0,0000	1,0000
c_som_2_2	0,8359	0,0078	0,0391	0,1172	0,0000	1,0000
Sum	0,5765	0,2240	0,0628	0,0792	0,0574	1,0000

Fig. 6.36. Ponderación de incidencia para el atributo clase de arma:

Fuente: elaboración propia

	Bicicleta	Objeto Personal	Motocicleta	Domiciliario	Vehiculo	No hubo elemento sustraído	Dinero	Sum
c_som_1_1	0,0000	0,8485	0,1364	0,0000	0,0152	0,0000	0,0000	1,0000
c_som_1_2	0,0000	0,0704	0,0000	0,4930	0,4085	0,0000	0,0282	1,0000
c_som_2_1	0,0495	0,5248	0,3762	0,0495	0,0000	0,0000	0,0000	1,0000
c_som_2_2	0,0703	0,8594	0,0156	0,0156	0,0156	0,0234	0,0000	1,0000
Sum	0,0383	0,6120	0,1339	0,1148	0,0874	0,0082	0,0055	1,0000

Fig. 6.37. Ponderación de incidencia para el atributo tipo de elemento sustraído:

Fuente: elaboración propia

	Forcejeo	Arrebato	No existio ataque	Ataque Brutal	Sum
c_som_1_1	0,2121	0,6970	0,0152	0,0758	1,0000
c_som_1_2	0,0282	0,0423	0,6056	0,3239	1,0000
c_som_2_1	0,0000	0,0000	0,6634	0,3366	1,0000
c_som_2_2	0,1875	0,8125	0,0000	0,0000	1,0000
Sum	0,1093	0,4180	0,3033	0,1694	1,0000

Fig. 6.38. Ponderación de incidencia para el atributo tipo de ataque:

Fuente: elaboración propia

	_1_0,00	_2_1,00	_3_2,00	_4_3,00	_5_4,00	Sum
c_som_1_1	0,0000	0,6364	0,3030	0,0606	0,0000	1,0000
c_som_1_2	0,3239	0,1268	0,3380	0,1690	0,0423	1,0000
c_som_2_1	0,0000	0,2376	0,4752	0,2475	0,0396	1,0000
c_som_2_2	0,0000	0,1875	0,4766	0,1641	0,1719	1,0000
Sum	0,0628	0,2705	0,4180	0,1694	0,0792	1,0000

Fig. 6.39. Ponderación de incidencia para el atributo edad del sospechoso:

Fuente: elaboración propia

	Masculino	S/D	Sum
c_som_1_1	1,0000	0,0000	1,0000
c_som_1_2	0,6761	0,3239	1,0000
c_som_2_1	1,0000	0,0000	1,0000
c_som_2_2	1,0000	0,0000	1,0000
Sum	0,9372	0,0628	1,0000

Fig. 6.40. Ponderación de incidencia para el atributo sexo del sospechoso:

Fuente: elaboración propia

	_1_1,00	_2_2,00	_3_3,00	_4_4,00	_5_5,00	_6_6,00	_7_7,00	_8_8,00	Sum
c_som_1_1	0,0000	0,0000	0,0000	0,1970	0,3485	0,1970	0,1061	0,1515	1,0000
c_som_1_2	0,0986	0,3944	0,1972	0,3099	0,0000	0,0000	0,0000	0,0000	1,0000
c_som_2_1	0,1287	0,4752	0,2673	0,0792	0,0198	0,0297	0,0000	0,0000	1,0000
c_som_2_2	0,1719	0,5078	0,1484	0,1719	0,0000	0,0000	0,0000	0,0000	1,0000
Sum	0,1148	0,3852	0,1639	0,1776	0,0683	0,0437	0,0191	0,0273	1,0000

Fig. 6.41. Ponderación de incidencia para el atributo edad de la víctima:

Fuente: elaboración propia

	Femenino	Masculino	Sum
c_som_1_1	0,8636	0,1364	1,0000
c_som_1_2	0,4648	0,5352	1,0000
c_som_2_1	0,8911	0,1089	1,0000
c_som_2_2	0,8984	0,1016	1,0000
Sum	0,8060	0,1940	1,0000

Fig. 6.42. Ponderación de incidencia para el atributo sexo de la víctima:

Fuente: elaboración propia

	No tiene ni oficio ni profesion	Tiene algun oficio	Tiene alguna profesion	Sum
c_som_1_1	0,3788	0,4848	0,1364	1,0000
c_som_1_2	0,5211	0,4789	0,0000	1,0000
c_som_2_1	0,5248	0,4158	0,0594	1,0000
c_som_2_2	0,4766	0,4844	0,0391	1,0000
Sum	0,4809	0,4645	0,0546	1,0000

Fig. 6.43. Ponderación de incidencia para el atributo ocupación de la víctima:

Fuente: elaboración propia

El estudio de las características del delito con mayor incidencia en la ocurrencia del hecho, generó el siguiente análisis para cada clúster:

- Clúster c_som_1_1: El mayor porcentaje (86,36%) de los hechos delictivos lo sufrieron personas de sexo femenino, en un rango de edad de 36 a 40 años (34,85%). No se registran casos de víctimas con edad de 15 a 30 años. La totalidad de las víctimas sufrieron el delito por algún sospechoso de sexo masculino cuya edad ronda entre los 15 y 20 años (33,38%), en la vía pública y en un 69,70% correspondió al tipo arrebato.

- Clúster c_som_1_2: Este clúster se caracteriza por contener una proporción igual en el número de mujeres y hombres que sufrieron algún delito, en un rango de edad de 21 a 25 años (39,44%), el 50,70% de incidencia trata el delito ocurrido en un domicilio particular y en un 60,56% de casos no existió ataque. No se registran casos de víctimas con edad a partir de los 36 años. El 67,61% de los robos cometió algún sospechoso de sexo masculino cuya edad ronda entre los 21 y 25 años.
- Clúster c_som_2_1: Este clúster determinó que existe una mayor cantidad de casos (66,34%) en donde el delito fue de tipo hurto. El mayor porcentaje (89,11%) de los hechos delictivos, sufrieron personas de sexo femenino y en un rango de edad de 21 a 25 años (47,52%). No se registran casos de víctimas con edad de 46 años en adelante. La totalidad de los delitos cometió algún sospechoso de sexo masculino cuya edad ronda entre los 15 y 20 años, y en la vía pública y sin utilizar algún tipo de arma durante el ataque (66,34%).
- Clúster c_som_2_2: Se analizó que el mayor porcentaje (89,84%) de los hechos delictivos lo sufrieron personas de sexo femenino, en un rango de edad de 21 a 25 años (50,78%). No se registran casos de víctimas con edad de 36 años en adelante. La totalidad de los robos los cometieron sospechosos de sexo masculino cuya edad ronda entre los 21 y 25 años (50,78%), en la vía pública y con un tipo de arma blanca (83,59%).

Concluida la ejecución de las técnicas se menciona a modo de resumen que, se identificaron los atributos más significativos resultado del proceso de agrupación con las víctimas que sufrieron los delitos: como el rango de edad y sexo de las personas con mayor porcentaje a sufrir un robo o hurto (21 a 25 años y de sexo femenino).

6.1.1.4.4. Evaluar el modelo

Se detalla a continuación la evaluación del modelo de descubrimiento de grupos y del descubrimiento de reglas, utilizando las métricas propuestas en la etapa de diseño de pruebas (sección 6.1.1.4.2) y aplicado a cada uno de los objetivos de minería de datos.

En el objetivo de minería de datos Nº 1 el algoritmo SOM agrupó los clusters según características similares. El problema de aplicar este algoritmo es decidir el número correcto de clúster y dado que no existe un criterio óptimo para la elección de un

número de grupos se utilizó como parámetro el valor por defecto propuesto por el algoritmo SOM (2 filas y 2 columnas, 4 grupos en total).

La valoración se expone considerando la problemática en cuestión, y la visualización proporcionada por el dendograma como una herramienta de agrupación jerárquica. El dendograma despliega la formación de clusters a través de un árbol jerárquico de agrupaciones, este método permite examinar los aglomerados sucesivos y brinda el número de clusters más significativo a aplicar según el análisis de distancias entre los grupos.

La gráfica del dendograma presenta de manera simple la distancia (basada en la similitud) entre los registros. A partir de dicha medición de distancia, se determinan los puntos de agrupación, uniendo aquellos pares de individuos que se encuentran más cercanos de manera iterativa hasta integrar todos a un mismo grupo. Es decir, básicamente mide la distancia entre los puntos en cada corte y une a los individuos a partir de su cercanía. Por último, se define la cantidad de clusters que optimiza (maximiza) la diferencia entre clusters.

La Fig. 6.44, representa el dendograma a través del cual se observa la formación de los clusters para este primer objetivo de minería de datos.

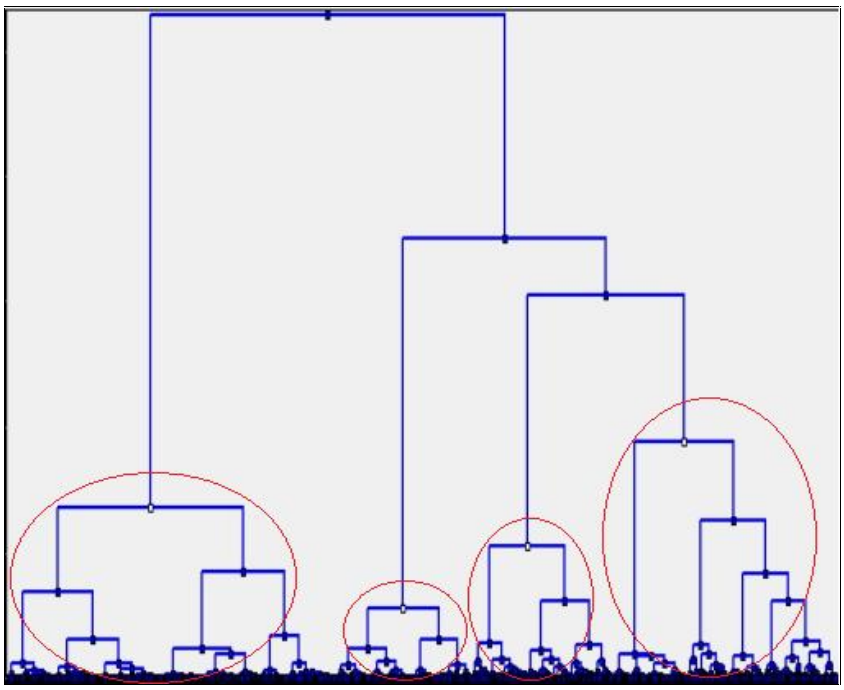


Fig. 6.44. Dendograma correspondiente al algoritmo SOM para identificar el comportamiento del delito: Fuente: elaboración propia

El análisis del dendograma se muestra en la Fig. 6.45, en donde a partir del conjunto de datos (366 registros) se observan los resultados de la formación de 10 clusters y la selección del número óptimo de grupos a elegir. En este caso la partición más significativa es la formación de cuatro grupos (fila marcada en color verde) según la cercanía entre los objetos del mismo clúster (0.37) y la brecha de distancia entre los grupos (0.29), por lo que se asume correcto la selección del número de clusters aplicados en la técnica SOM. Si bien se observa que la distancia máxima (0,44) entre grupos se logra con la formación de 2 clusters, la cercanía entre sus miembros es menor que el propuesto como óptimo (4 grupos), por lo que se entiende que el número que maximiza las diferencias entre clusters debe ser relativo a la variación dentro de los mismos.

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,1652	0,4444
3	0,2749	0,1154
4	0,3702	0,2914
5	0,4291	0,1310
6	0,4715	0,0260
7	0,5107	0,0504
8	0,5437	0,0528
9	0,5700	0,0015
10	0,5961	0,0360

Fig. 6.45. Selección del número óptimo de clusters para identificar el comportamiento del delito:

Fuente: elaboración propia

A continuación, se evaluó a partir de la matriz de confusión y la tabla de predicción de valores, la exactitud del modelo y la tasa de error total calculada a partir del modelo de descubrimiento de reglas de comportamiento.

Los valores localizados en la diagonal principal de la matriz de confusión son las clasificaciones correctas y en el cual se observa el mayor número de registros, es decir, se indica una clasificación eficaz (Fig. 6.46). Mientras que los valores de la diagonal secundaria representan los errores entre las clases, como ejemplo se observa en la primera fila de la matriz en donde el modelo clasifica por error 3 registros en la clase c_som_1_2 cuando deberían pertenecer al grupo c_som_1_1.

	c_som_1_1	c_som_1_2	c_som_2_1	c_som_2_2	Sum
c_som_1_1	73	3	0	3	79
c_som_1_2	7	90	0	1	98
c_som_2_1	2	0	118	8	128
c_som_2_2	2	2	2	55	61
Sum	84	95	120	67	366

Fig. 6.46. Matriz de confusión:

Fuente: elaboración propia

La tabla de predicción de valores (Fig. 6.47) indica la efectividad del algoritmo para predecir una clase en particular. Se observa que la eficacia del método para determinar las clases es aceptable en todos los casos dado que hay un 92,41% en el clúster c_som_1_1, un 91,84% en el clúster c_som_1_2, un 92,19% en el clúster c_som_2_1 y un 90,16% en el clúster c_som_2_2.

Value	Recall	1-Precision
c_som_1_1	0,9241	0,1310
c_som_1_2	0,9184	0,0526
c_som_2_1	0,9219	0,0167
c_som_2_2	0,9016	0,1791

Fig. 6.47. Tabla de predicción de valores:

Fuente: elaboración propia

A continuación, se calculan las métricas de tasa de exactitud y error total del modelo, se aplican para el modelo de descubrimiento de reglas y de dependencias significativas:

Tasa de exactitud = 91,20%

Tasa de error = 08,20%

En el objetivo de minería de datos Nº 2, el algoritmo SOM agrupó los clusters según características similares. Para la elección de un número de grupos se utilizó como parámetro el valor por defecto propuesto por el algoritmo SOM (2 filas y 2 columnas, 4 grupos en total).

La Fig. 6.48, representa el dendograma a través del cual se observa la formación de los clusters para este segundo objetivo de minería de datos.

En ésta gráfica se observa la formación de un solo clúster, donde éste a su vez forma otros 2 grupos, a la izquierda del mismo se agrupa el primer clúster, y luego a su derecha se identifican otros 2 grupos, que, a su vez, a su izquierda forma el segundo clúster y luego a su derecha los otros 2 grupos restantes. Identificando 4 clúster en total (círculos de color rojo).

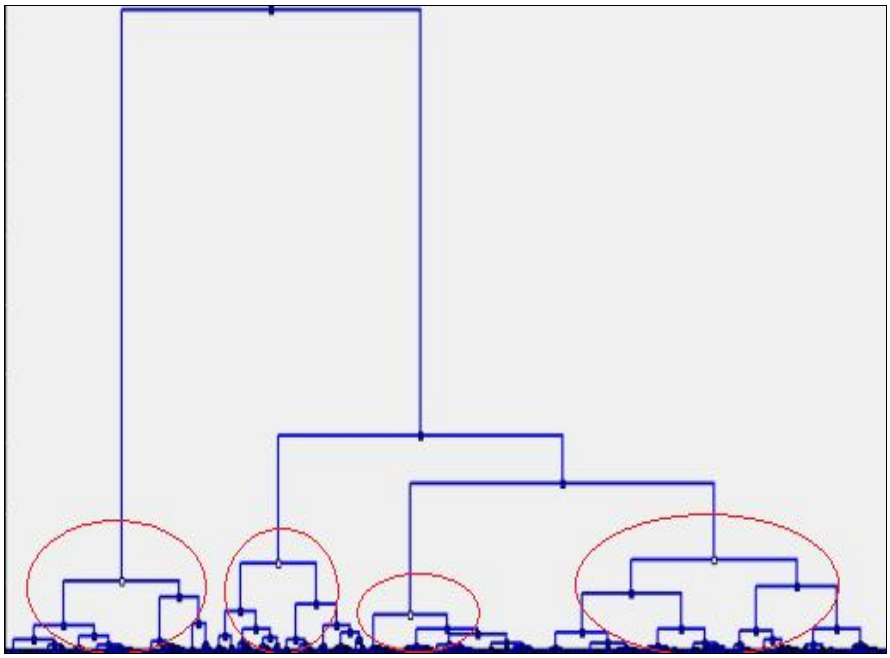


Fig. 6.48. Dendograma del algoritmo SOM para identificar grupos entre las zonas de mayor ocurrencia de delitos: Fuente: elaboración propia

El análisis del dendograma se observa en la Fig. 6.49, a partir del conjunto de datos (366 registros) se muestran los resultados de la formación de 10 clusters y la selección del número óptimo de grupos a elegir. En este caso la partición más significativa es la formación de cuatro grupos (fila marcada en color verde) según la cercanía entre los objetos del mismo clúster (0.46) y la brecha de distancia entre los grupos (0.27), por lo que se asume correcto la selección del número de clúster aplicados en la técnica SOM. Si bien se observa que la distancia máxima (1,53) entre grupos se logra con la formación de 2 clusters, la cercanía entre sus miembros es menor que el propuesto como óptimo (4 grupos), por lo que el número que maximiza las diferencias entre clusters debe ser relativo a la variación dentro de los mismos.

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,2887	1,5319
3	0,3858	0,1733
4	0,4613	0,2741
5	0,5026	0,0104
6	0,5425	0,0665
7	0,5742	0,0139
8	0,6040	0,0263
9	0,6306	0,0156
10	0,6553	0,0235

Fig. 6.49. Selección del número óptimo de clusters para identificar grupos entre las zonas de mayor ocurrencia de delitos:

Fuente: elaboración propia

Se evaluó a partir de la matriz de confusión y la tabla de predicción de valores, la exactitud del modelo y la tasa de error total, calculada a partir del modelo de descubrimiento de reglas de comportamiento.

Los valores de la diagonal principal de la matriz de confusión representan las clasificaciones correctas y corresponde al mayor número de registros, esto indica una clasificación eficaz (Fig. 6.50). Mientras que los valores de la diagonal secundaria representan los errores entre las clases, como ejemplo se observa en la primera fila de la matriz en donde el modelo determina erróneamente 5 registros en la clase c_som_1_2 cuando deberían pertenecer al grupo c_som_1_1.

	c_som_1_1	c_som_1_2	c_som_2_1	c_som_2_2	Sum
c_som_1_1	67	0	0	5	72
c_som_1_2	0	45	0	0	45
c_som_2_1	4	0	43	1	48
c_som_2_2	2	1	4	23	30
Sum	73	46	47	29	195

Fig. 6.50. Matriz de confusión:

Fuente: elaboración propia

La tabla de predicción de valores (Fig. 6.51) indica cuan efectivo es el algoritmo para predecir una clase en particular. Se observa que la eficacia del método para determinar las clases es aceptable en todos los casos dado que hay un 93,06% de eficacia en el clúster c_som_1_1, un 100% en el clúster c_som_1_2, un 89,58% en el clúster c_som_2_1 y un 76,67% en el clúster c_som_2_2.

Value	Recall	1-Precision
c_som_1_1	0,9306	0,0822
c_som_1_2	1,0000	0,0217
c_som_2_1	0,8958	0,0851
c_som_2_2	0,7667	0,2069

Fig. 6.51. Tabla de predicción de valores:
Fuente: elaboración propia

Las métricas de tasa de exactitud y error total del modelo, se aplican para el modelo de descubrimiento de reglas y de dependencias significativas:

Tasa de exactitud= 91,28%

Tasa de error = 08,72%

En el objetivo de minería de datos Nº 3 el algoritmo SOM agrupó los clusters según características similares. Para la elección de un número de grupos se utilizó como parámetro el valor por defecto propuesto por el algoritmo SOM (2 filas y 2 columnas, 4 grupos en total).

La Fig. 6.52, representa el dendograma a través del cual se observa la formación de los clusters para este tercer objetivo de minería de datos.

En ésta gráfica se observa la formación de un solo grupo, donde éste a su vez forma otros 2. A la izquierda del mismo se agrupa el primer clúster, y luego a su derecha se identifican otros 2 grupos, a la izquierda de este se forma el segundo clúster y luego a su derecha se forman otras 2 agrupaciones, el cual a su izquierda se agrupa el tercer clúster, y luego a su derecha el cuarto clúster restante. Identificando 4 clusters al total (círculos de color rojo).

Fig. 6.52. Dendograma del algoritmo SOM para identificar grupos entre las personas más propensas a sufrir un delito:
Fuente: elaboración propia

La Fig. 6.53 presenta el análisis del dendograma, a partir del conjunto de datos (366 registros). Se observan los resultados de la formación de 10 clusters y la selección del número óptimo de grupos a elegir, en este caso la partición más significativa es la formación de cuatro grupos (fila marcada en color verde) según las cercanías entre los objetos del mismo clúster (0.37) y la brecha de distancia entre los grupos (0.29), por lo que se asume correcto la selección del número de clusters aplicados en la técnica SOM. Si bien se observa que la distancia máxima (0,44) entre grupos se logra con la formación de 2 clusters, la cercanía entre sus miembros es menor que el propuesto como óptimo (4 grupos), por lo que se asume que el número que maximiza las diferencias entre clusters debe ser relativo a la variación dentro de los mismos.

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,1652	0,4444
3	0,2749	0,1154
4	0,3702	0,2914
5	0,4291	0,1310
6	0,4715	0,0260
7	0,5107	0,0504
8	0,5437	0,0528
9	0,5700	0,0015
10	0,5961	0,0360

Fig. 6.53. Selección del número óptimo de clusters para identificar grupos entre las personas más propensas a sufrir un delito:
Fuente: elaboración propia

A continuación, se evaluó a partir de la matriz de confusión y la tabla de predicción de valores, la exactitud del modelo y la tasa de error total considerando el modelo de descubrimiento de reglas de comportamiento.

Los valores localizados en la diagonal principal de la matriz de confusión representan las clasificaciones correctas y se observa el mayor número de registros, esto indica una clasificación eficaz (Fig. 6.54). Mientras que los valores de la diagonal secundaria representan los errores entre las clases, como ejemplo se observa en la segunda fila de la matriz en donde el modelo clasifica por error 3 registros en la clase c_som_1_1 cuando deberían pertenecer al grupo c_som_1_2.

	c_som_1_1	c_som_1_2	c_som_2_1	c_som_2_2	Sum
c_som_1_1	61	0	0	0	61
c_som_1_2	3	71	2	2	78
c_som_2_1	0	0	99	0	99
c_som_2_2	2	0	0	126	128
Sum	66	71	101	128	366

Fig. 6.54. Matriz de confusión:
Fuente: elaboración propia

La tabla de predicción de valores de la Fig. 6.55 indica la efectividad del algoritmo para predecir una clase en particular. Se observa que la eficacia del método para determinar las clases es aceptable en todos los casos dado que hay un 100% de coincidencia en el clúster c_som_1_1, un 91,03% en el clúster c_som_1_2, un 100% en el clúster c_som_2_1 y un 98,44% en el clúster c_som_2_2.

Value	Recall	1-Precision
c_som_1_1	1,0000	0,0758
c_som_1_2	0,9103	0,0000
c_som_2_1	1,0000	0,0198
c_som_2_2	0,9844	0,0156

Fig. 6.55. Tabla de predicción de valores:
Fuente: elaboración propia

Además, se calculan las métricas de tasa de exactitud y error total del modelo para el modelo de descubrimiento de reglas y de dependencias significativas:

Tasa de exactitud= 97,46%

Tasa de error = 02,46%

Los modelos obtenidos resultado de los objetivos de minería de datos han demostrado lo siguiente:

- Una selección de agrupamientos adecuada y un alto porcentaje de precisión dado que la tasa de exactitud de todos los modelos con los parámetros especificados para la ejecución de los mismos a partir del conjunto de datos seleccionados arrojan un porcentaje mayor al 90%. Este porcentaje indica que no es necesario reiterar las etapas de aplicación de los modelos propuestos.
- La tasa de error en todos los casos tiene un promedio que no supera el 10% lo cual indica una alta capacidad predictiva en cuanto a los casos que son de interés. Entre estos casos se menciona: el comportamiento de los delitos ocurridos en la ciudad, las zonas con mayor porcentaje de incidencia de ocurrencia conjuntamente con la caracterización de las personas más propensas a sufrir de alguno de estos hechos.

6.1.1.5. Evaluación

En esta última etapa de la metodología CRISP-DM, se expone la evaluación de los resultados obtenidos considerando los objetivos del negocio:

- Evaluación para el modelo del objetivo de negocio Nº 1. Este modelo es calificado como aceptable dado que permite diseñar planes de prevención de riesgo a partir de las predicciones realizadas en cuanto a los días de la semana con mayor porcentaje de delitos, ubicación y tipo de arma más utilizada, y el tipo de objeto con mayor porcentaje de robo.
- Evaluación para el modelo del objetivo de negocio Nº 2. Este segundo modelo también es aceptable, ya que se pueden asegurar los barrios y las calles de mayor riesgo de ocurrencia de delitos y a través de esta información aumentar la eficacia del accionar de las fuerzas policiales logrando la ubicación de los mismos en las zonas más peligrosos de la ciudad.

- Evaluación para el modelo del objetivo de negocio Nº 3. Para este objetivo se logró un modelo viable, ofrece información de las personas con determinada edad y sexo que podrán ser más vulnerables ante estos hechos de robo y hurto, y quienes son las personas con cierta edad y determinado sexo que pueden ser posibles sospechosos.

6.2. Discusión de los resultados obtenidos de aplicar las técnicas de minería de datos sobre el caso de validación.

En el análisis de los resultados obtenidos en cuanto al objetivo de minería de datos Nº 1 se observa que el delito se caracteriza por variables como: la modalidad, el empleo de armas, los objetos sustraídos, lugar de ocurrencia, el rango horario, los días de la semana y los meses con más delitos ocurridos.

Con respecto a la modalidad del delito estudiada en cada uno de los clusters, el arrebato es la constante en los cuatro grupos, el mayor número se registra en la zona de la vía pública. Predominan como los objetos sustraídos aquellos de tipo personal. En relación del tipo de arma utilizada, prevalecen las armas blancas.

La información producida indica la preferencia de los delincuentes en actuar los días viernes, sábados y domingos. En cuanto al rango horario, la tendencia de los hechos delictivos se registró entre las 12 p.m. y 16 p.m.

Respecto a los meses en los cuales se presentaron más delitos, los clusters coinciden en detectar a marzo y abril. Si bien este resultado no ofrece mayor información, como líneas futuras se podrían aplicar estas técnicas de minería de datos para los años siguientes, y a partir de los resultados conocer la influencia de delito por época o por estaciones del año.

Del análisis de zonas con mayor ocurrencia de delitos estudiadas a partir del objetivo de minería de datos Nº 2, los resultados de los modelos desarrollados establecen los barrios y calles con mayor número de delitos, entre los que se mencionan a los barrios N° 1 y N° 2. La calle en donde se ubican la mayor cantidad de hechos delictivos es la calle N° 1. Los valores que predominan en ambos grupos son: el delito ocurrido en la vía pública, siendo los objetos personales las mayores sustracciones registradas y prevaleciendo el uso de armas blancas. Otro grupo comprende a los barrios N° 3, N° 4, N° 5, N° 6 y N° 7. Las calles en donde se ubicaron la mayor cantidad de hechos son:

calles N° 2, N° 3 y N° 4. En este sector se localizan gran cantidad de paradas de colectivos, lo cual conlleva a una alta afluencia de personas y se podría considerar un motivo por el cual los delitos ocurran en su mayoría sobre la vía pública.

También se menciona al barrio N° 8 el cual registró el número más elevado de delitos, en particular en la calle N° 5. Los valores que más predominan en este barrio son hechos ocurridos sobre la vía pública, siendo los objetos personales sustraídos en mayor número.

El objetivo de minería de datos Nº 3, presenta otro análisis para conocer y caracterizar el robo y hurto a personas más propensas o con mayor posibilidad de ser víctima. La ponderación de atributos utilizando Redes Bayesianas permitió comprobar que, de acuerdo a los valores de incidencia de los atributos edad y sexo de la víctima, se determinó un porcentaje muy alto de mujeres entre los 21 a 25 años.

Capítulo 7

Discusión final de los resultados, conclusiones y trabajos futuros

7. Discusión final de los resultados, conclusiones y trabajos futuros

Este capítulo presenta la discusión final de los resultados derivados de la elaboración del presente Trabajo Final de Maestría (sección 7.1), seguido de las conclusiones y futuras líneas de investigación relacionadas con temas de interés identificados durante el desarrollo del trabajo (sección 7.2).

7.1. Discusión final de los resultados

El Trabajo Final de Maestría, tuvo como objetivo principal elaborar un procedimiento, a través del cual se integraron métodos y herramientas de las tecnologías GIS y de la minería de datos, con la finalidad de aportar nuevas formas de análisis de los datos y generación de información sobre un dominio particular.

Específicamente, con fines de producir información de apoyo a toma de decisiones centradas en tecnologías de minería de datos se aplicaron las técnicas SOM, TDIDT y Naive Bayes, optando respectivamente por: el algoritmo Kohonen-SOM para el descubrimiento de grupos, C4.5 para la caracterización o descubrimiento de reglas de cada clúster y Redes bayesianas para la ponderación de atributos significativos.

Construido el procedimiento, se procedió a su validación. Se optó como dominio de aplicación la caracterización de robos y hurtos en una ciudad a partir de los datos registrados para el primer semestre del año 2017 en el SAT.

Estos procesos de explotación de información aplicados sobre una base de datos georreferenciada posibilitaron el hallazgo de patrones significativos con el fin de descubrir información valiosa sobre un determinado territorio asociado. En este contexto, el uso las herramientas de minería de datos enfatizando en las técnicas y algoritmos mencionados para la resolución de la problemática planteada y siguiendo el proceso descripto, permitió:

- Evaluar la relación entre múltiples variables (modelando el proceder de un experto) para determinar el patrón que explica el comportamiento de la problemática planteada.
- Facilitar la implementación continua de procesos analíticos, permitiendo ajustar el conocimiento a los cambios del dominio.

- Ampliar este tipo de análisis, reduciendo los tiempos y costos de aplicación, complementando el proceder de los expertos en el dominio.

A partir del caso de estudio basado en los hechos delictivos, se propusieron tres objetivos de minería de datos para hallar patrones que indiquen las características de los delitos, zonas de mayor riesgo y víctimas más propensas a sufrirlos. A continuación, se discuten los principales hallazgos de este estudio:

- En referencia al primer objetivo de minería de datos: la caracterización del comportamiento de delitos utilizando variables temporales (días y horarios) ofrece información relevante como: días de mayor porcentaje de ocurrencia de delitos (viernes, sábados y domingos), horarios más susceptibles (horarios de siesta y de madrugada), objetos con mayor número de sustracciones (objetos personales), lugares y tipos de armas más utilizadas durante el delito (vía pública y con arma blanca) y tipo de ataque más sufrido (arrebato).
- Respecto al segundo objetivo de minería de datos: la caracterización de zonas más peligrosas utilizando variables de ubicación (barrios y calles), ofrece información útil y necesaria. Los barrios con mayor cantidad de delitos son: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7 y N° 8.
- En relación al tercer objetivo de minería de datos: la identificaron de los atributos más significativos resultado del proceso de agrupación con las víctimas que sufrieron los delitos ofrece información útil como el rango de edad y sexo de las personas con mayor porcentaje a sufrir un robo o hurto (21 a 25 años y de sexo femenino).

Los mapas delictivos creados como uno de los aportes de la aplicación del procedimiento propuesto, corresponden a la relación entre el delito ocurrido en un espacio geográfico y tiempo determinado. Cada mapa de calor muestra cantidades asociadas a los delitos. Es decir, para los objetivos de minería de datos 1, 2 y 3, los mapas como representaciones gráficas no brindaron grandes resultados dado que muestran zonas similares. Sin embargo, es posible obtener mayor información descriptiva que caracteriza a cada delito, al aplicar las técnicas de minería de datos al conjunto de registros relevados, es decir, estas técnicas se requieren para hallar y caracterizar a los patrones delictivos.

Los resultados obtenidos del proceso relacionado al segundo objetivo de minería de datos permiten concluir que el análisis resulta más significativo para este caso, dado que, incluía variables relativas a ubicaciones, las cuales lograron mejorar el hallazgo de barrios y calles con mayor ocurrencia de robos y hurtos.

7.2. Conclusiones y trabajos futuros

La elaboración del presente Trabajo Final de Maestría, desde una perspectiva personal y disciplinar, permitió profundizar los conocimientos teóricos y metodológicos y propiciar un entorno de experiencia práctica en referencia a la construcción de procedimientos, la tecnología GIS y la minería de datos.

Particularmente, se propuso una solución genérica aplicada a distintos casos de estudio o situaciones problemáticas de interés. Éste procedimiento se sustentó en el estudio, selección y análisis de algunas tecnologías de georreferenciación GIS/IDE existentes y en métodos, técnicas y herramientas comprendidas en la minería de datos, como estrategia de explotación de la información.

Con la finalidad de verificar el procedimiento descripto en el Capitulo 4, se seleccionó como caso de estudio la explotación de información delimitada a la detección de hechos delictivos de robos y hurtos en el primer semestre del año 2017. Una real aplicación en un dominio que produce conocimiento en torno a la seguridad de los habitantes de la ciudad seleccionada a efectos de validación.

La propuesta se verificó mediante la definición de tres objetivos de minería de datos. Así, los modelos de explotación de información representativos de estos objetivos de MD y los resultados obtenidos produjeron información relevante orientada a identificar y caracterizar el comportamiento de los delitos, las zonas de mayor riesgo y las personas más expuestas. Las características evaluadas generaron una descripción de posibles tendencias delictivas en donde predominan: los días de mayor porcentaje de ocurrencia de delitos (viernes, sábados y domingos), horarios más susceptibles (siesta), objetos con mayor número de sustracciones (objetos personales), lugares y tipos de armas más utilizadas durante el delito (vía pública y con arma blanca) y tipo de ataque más sufrido (arrebato).

La utilización de las técnicas de minería de datos para el hallazgo de las zonas más peligrosas también ofrece información útil y relevante para conocer los barrios con

mayor cantidad de delitos (N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7 y N° 8). Además, se identificaron los atributos más significativos, resultado del proceso de agrupación de las víctimas que sufrieron los delitos: y surgieron como datos relevantes el rango de edad y sexo de las personas con mayor porcentaje a sufrir un robo o hurto (21 a 25 años y sexo femenino).

Los modelos desarrollados permitieron aplicar los algoritmos de minería de datos SOM, TDIDT y Naive Bayes. Estas técnicas de agrupamiento, caracterización y ponderación de atributos resultaron eficientes sobre el análisis de los objetos espaciales seleccionados.

Por lo expuesto en párrafos precedentes, en el presente Trabajo Final de Maestría se abordaron e integraron aspectos teóricos, metodológicos y prácticos. Como línea de trabajo futuro se propone el diseño y desarrollo de un mapa interactivo que se podría destinar a un ámbito de toma de decisiones de alcance provincial, regional o nacional con capas geográficas competentes al dominio. Es decir, se pretende maximizar el objetivo del presente TFM de modo que los datos se puedan consultar y descargar disponiéndolos para cualquier otro organismo público, privado o comunidad que los requiera. Por ello, se espera que la información producida aporte a las fuerzas de seguridad en la definición de políticas de prevención que asistan a la seguridad social del ciudadano, a la sistematización de las tareas manuales de los agentes policiales y la generación de información de calidad que apoye la toma de decisiones.

En adición, como futura líneas de trabajo se propone ampliar el estudio y el análisis de las técnicas y herramientas a utilizar en contextos de grandes datos, además de incluir algoritmos para determinar las variables relevantes como actividad preliminar a la aplicación de las técnicas de minería de datos.

Referencias

[1] K. T. Chang, "Introduction," in *Introduction to Geographic Information Systems*, 8th ed. New York: McGraw-Hill, pp. 1-10, 2015.

[2] R. García-Martínez, P. Britos, P. Pesado, R. Bertone, F. Pollo-Cattaneo, D. Rodríguez, P. Pyteland and J. Vanrell, "Towards an Information Mining Engineering," in *Software Engineering, Methods, Modeling and Teaching*, Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6, pp. 83-99, 2011.

[3] R. García-Martínez, H. Merlino, D. Rodríguez, S. Martins, E. Baldizzoni, E. Diez, H. Amatriain, F. Ribeiro, A. Segura, P. Santamaría, F. Mieres and D. Aguirre, "Explotación de información geográfica basada en integración de ambientes de trabajo," in *XX Congreso Argentino de Ciencias de la Computación*, Argentina, 2014.

[4] A. Peña Suarez, "Modelo para la Caracterización del Delito en la Ciudad de Bogotá, Aplicando Técnicas de Minería de Datos Espaciales," Tesis de Maestría en Ciencias de la Información y de las Comunicaciones. Universidad Distrital Francisco José de Caldas. Bogotá, Colombia, 2017.

[5] Y. Cabrero Ortega and A. García Pérez, "Introducción al QGIS," in *Análisis estadístico con datos espaciales con QGIS y R*, Madrid: Editorial UNED, pp. 1-16, 2015.

[6] M. Wegmann, B. Leutner and S. Dech, "Spatial Data and Software," in *Remote Sensing and GIS for Ecologists*, 1st ed., Pelagic Publishing, 2016.

[7] Instituto Geográfico Nacional, (2019, Junio 30). "Geodésica, Introducción". [En línea]. Disponible en: http://www.ign.gob.ar/NuestrasActividades/Geodesia/Introduccion

[8] Instituto Geográfico Nacional de Argentina, (2019, Julio 30). "Introducción. POSGAR 94. Posiciones Geodésicas Argentinas". [En línea]. Disponible en: http://www.ign.gob.ar/NuestrasActividades/Geodesia/Posgar94

[9] M. C. Mata Montes, "Innovando en carreras técnicas: Las IDE's como recurso educativo," in *Innovación educativa en las enseñanzas técnicas*, vol. 2, España: Ediciones de la Universidad de Castilla La Mancha, pp. 1179-1182, 2015.

[10] A. Pérez Navarro, A. Botella Plana, A. Olivella González, C. Olmedillas Hernández and J. Rodríguez Lloret, "Software GIS," in *Introducción a los sistemas de información geográfica y geotelemática*, 1ra ed. Barcelona: Editorial UOC, pp. 259-270, 2011.

[11] D. McInerney and P. Kempeneers, "Introduction," in *Open Source Geospatial Tools: Applications in Earth Observation*, 1ra ed., Barcelona: Springer, pp. 11-20, 2014.

[12] M. J. Lopez Garcia, P. Carmona, J. Salom and J. M. Albertos, "Tecnologías de la información geográfica: Perspectivas multidisciplinares en la sociedad del conocimiento", in *XVIII Congreso Nacional de Tecnología de Información Geográfica*, Departamento de Geográfica, Universidad de Valencia, 2018.

[13] N. Baghdadi, C. Mallet and M. Zribi, "Introduction to QGIS," in *QGIS and G*eneric Tools, vol.1. Wiley, pp.1-16, 2018.

[14] Asociación gvSIG, (2019, Junio 30). "Conoce gvSIG Desktop, el Sistema de Información Geográfica libre". [En línea]. Disponible en: http://www.gvsig.com/es/productos/gvsig-desktop

[15] E. Westra, "Python Libraries for Geospatial Development," in *Python Geospatial Development*, 3rd. ed., pp. 47-58, 2016.
[16] M. Anoopkumar and A. M. J. Md. Zubair Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," in *International Conference on Data Mining and Advanced Computing*, India, 2016.
[17] R. García-Martínez, P. Britos and D. Rodríguez, "Information mining processes based on intelligent systems" in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Berlin, Heidelberg, pp. 402-410, 2013.
[18] R. Rakotomalala, (2019, Junio 30). "TANAGRA project". [En línea]. Disponible en: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[19] S. K. Strydom and M. Strydom, "Classification Techniques and Data Mining Tools Used in Medical Bioinformatics," in *Big Data Governance and Perspectives in Knowledge Management*, IGI Global, pp. 105-116, 2018.
[20] J. Demšar, B and Zupan B, "Orange: Data Mining Fruitful and Fun - A Historical Perspective," in *Informática* no. 37, pp. 55–60, 2013.
[21] S. Martins, "Modelo de proceso para proyectos de explotación de información," Tesis doctoral en ciencias informáticas, Universidad Nacional de la Plata. Sin publicar.
[22] P. Britos, H. Merlino, E. Fernández, M. Ochoa, E. Diez, and R. García Martínez, "Tool Selection Methodology in Data Mining," in *Proceedings V Ibero-American Symposium on Software Engineering*, pp. 85-90, 2006.
[23] A. Cirillo, "Why to Choose R for Your Data Mining and Where to Start," in *R Data Mining*, Packt Publishing Ltd., pp. 1-37, 2017.
[24] P. J. Deitel and H. Deitel, "Introducción a las computatoras y Python 1," in Python for Programmers, Prentice Hall, pp. 5-29, 2019.
[25] Y. Zhao and Y. Cen, Data Mining Applications with R, 1st ed. Academic Press, 2013.
[26] V. Porcu, "Getting Started," in *Python for Data Mining Quick Syntax Reference*, Apress, pp. 1-12, 2018.
[27] N. Greeneltch, Python Data Mining Quick Start Guide. Packt Publishing Ltd, 2019.
[28] J. M. Moine, "Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo," Tesis de Maestría en Ingeniería de Software, Facultad de Informática, Universidad Nacional de la Plata, Buenos Aires, Argentina, 2013.
[29] A. Ochoa, E. Fernández, P. Britos and R. García-Martínez. Metodologías de Ingeniería Informática. Editorial Nueva Librería. ISBN 978-987-1104-54-3. 2008.
[30] O. Marbán, G. Mariscal and J. Segovia, "A data mining & knowledge discovery process model" In *Data mining and knowledge discovery in real life applications*. Intech Open. 2009.
[31] S. Martins, P. Pesado and R. García-Martínez, "Information Mining Projects Management", in *Proceedings 28th International Conference on Software Engineering & Knowledge Engineering*, pp. 504-509, 2016.
[32] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery, in databases" in *AI magazine*, vol. 17, no. 3, 1996.

[33] S.Hawamdeh and H. Chang, "Data Analytics for Deriving Knowledge From User Feedback," in Analytics and Knowledge Management, CRC Press, 2018.
[34] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide. 2000.
[35] D. Pyle, "Methodology," in *Business modeling and data mining*. Morgan Kaufmann, pp. 533-540, 2003.
[36] M. Alnoukari, "ASD-BI: A Business Intelligence Modeling and Integration Framework based on Agile Methodologies", (Doctoral dissertation, Arab Academy for Banking and Financial Sciences), 2010.
[37] J. B. Rollins. Foundational methodology for data science. White Paper. IBM Analytics. 2015.
[38] Microsoft Azure, (2019, Junio 30). "Team Data Science Process". [En línea]. Disponible en: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/
[39] G. Mariscal, O. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," in The Knowledge Engineering Review, vol. 25, no 2, pp. 137-166, 2010.
[40] M. Tiwari, R. Dixit and A. Kesharwani, "*Data Mining Algorithms,"* in Data Mining Principles, Process Model and Applications, Educreation Publishing, pp. 18-25, 2017.
[41] S. Martins, "Modelo de Proceso para Proyectos de Ingeniería de Explotación de Información", in *XXI Congreso Iberoamericano en Ingeniería de Software*, 2018.
[42] G. Piatetsky, (2019, Junio 30). "KDnuggets". [En línea]. Disponible en: https://www.kdnuggets.com
[43] L. E. Flores, S. I. Mariño and S. Martins, "Propuesta de procedimiento para el análisis delictivo basado en la explotación de la información" in *XX Workshop de Investigadores en Ciencias de la Computación*, 2018.
[44] Ministerio de Seguridad de la Nación Argentina, (2019, Julio 25). "Estadísticas Criminales en la República Argentina – Año 2017 Informe". [En línea]. Disponible en: https://estadisticascriminales.minseg.gob.ar/reports/Informe%20SNIC%202017.pdf
[45] M. Colleen, "Process Models for Data Mining and Analysis," in *Data Mining and Predictive Analysis*, 2nd ed., Butterworth-Heinemann, pp. 45-65, 2015.
[46] L. E. Flores and S. I. Mariño, "Revisión sistemática de literatura: explotación de información y tecnologías GIS aplicadas para hallar patrones delictivos" in *Revista Entorno*, de Utec Universidad Tecnológica de El Salvador, Editorial no. 67, Argentina, pp. 30-41, 2019.
[47] M. Hosseinkhani, S. Koochakzaei, S. Keikhaee, and Y. H. Amin, "*Detecting suspicion information on the web using crime Data Mining techniques*," in International Journal of Advanced Computer Science and Information Technology (IJACSIT), vol. 3, no. 1, pp 32–41, 2014.
[48] H. Hassani , X. Huang , E. S. Silva and M. Ghodsi, "*A review of data mining applications in crime,*" in Statistical Analysis and Data Mining, 1rt ed., vol. 9, no. 3. New York, USA, pp 139-154, 2016.
[49] M. Sukanya, T. Kalaikumaran, and S. Karthik, "*Criminals and crime hotspot detection using data mining algorithms: clustering and classification,"* in

International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 10, pp. 225–227, 2012.

[50] M. Vijayakumar, S. Karthick, and N. Prakash, *"The dayto-day crime forecasting analysis of using spatial-temporal clustering simulation,"* in Statistical Analysis and Data Mining, vol. 4, no. 1, pp 1–6, 2013.

[51] F. Valenga, E. Fernández, H. Merlino, D. Rodríguez, C. Procopio, P. Britos and R. García-Martínez, "Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina," in *VII Jornadas de Ingeniería del Software e Ingeniería del Conocimiento*. Argentina, pp. 258-270, 2008.

[52] P. Britos, E. Fernández, H. Merlino, F. Pollo-Cataneo, D. Rodríguez, C. Procopio, C. Rancan and R. García Martínez, "Explotación de información aplicada a inteligencia criminal en Argentina", in *XIII Congreso Argentino de Ciencias de la Computación*. Argentina, 2008.

[53] Gobierno de la Ciudad de Buenos Aires, (2019, Julio 27). "Buenos Aires Ciudad. Mapa del delito". [En línea]. Disponible en: https://mapa.seguridadciudad.gob.ar/

[54] Ministerio de Seguridad de la Nación Argentina, Subsecretaria de Política de Seguridad e Intervención Territorial, Dirección Nacional de Gestión de la Información Criminal (2019, Julio 20) "Sistema Nacional de Información Criminal. Manual de Instrucciones (SNIT - SAT)" [En línea]. Disponible en: https://policia.chubut.gov.ar/media/documentos/Manual%20Actualizado%20SNIC-SAT.pdf

[55] Real Academia Española (2019, Julio 30) "Diccionario de la lengua española" [En línea]. Disponible en: https://dle.rae.es/

[56] G. Ciciliani, S. Martins and H. Merlino, "Análisis preliminar del rendimiento de algoritmos para el proceso de descubrimiento de reglas de pertenencia a grupos," in XXIV *Congreso Argentino de Ciencias de la Computación*, 2018.

[57] R. García Martínez, P. Britos, S. Martins and E. Baldizzoni, "Métricas para Proyectos de Explotación," in *Ingeniería de Proyectos de Explotación de Información*, 1ra. ed., Editorial Nueva Librería, Buenos Aires, Argentina, 2015.

[58] D. Basso, "Propuesta de Métricas para Proyectos de Explotación de Información," in *Revista Latinoamericana de Ingeniería de Software*, vol. 2, no. 4, pp. 157-218, 2014.

[59] B. Kitchenham, (2019, Julio 30). "Procedures for Performing Systematic Reviews". [En línea]. Disponible en: http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf

[60] P. Britos, "Procesos de explotación de información basados en sistemas inteligentes," Tesis Doctoral en Ciencias Informáticas, Facultad de Informática, Universidad Nacional de la Plata, Buenos Aires, Argentina, 2008.

[61] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery in Databases: An Overview," in AI Magazine, vol. 13, no 5, pp. 57-70, 1992.

[62] C. Pérez López and D. S. González, "Minería de datos: Conceptos, Técnicas y Sistemas," in *Minería de datos: técnicas y herramientas*, Thomson Ediciones, España: Editorial Paraninfo, pp. 1-10, 2007.

Anexos

Anexo 1. Revisión sistemática de la literatura: integración de procesos de explotación de información con tecnologías GIS y su aplicación para el hallazgo de patrones delictivos.

En materia de búsqueda de antecedentes, se realizó una revisión sistemática de la literatura o RSL en torno a la aplicación e integración de técnicas y herramientas de minería de datos con tecnología GIS para el hallazgo de patrones delictivos, esta RSL fue elaborada en el año 2017 con la finalidad de generar información que sustente la definición del proyecto de TFM.

Esta Revisión Sistemática de la Literatura (RSL) es una estrategia relevante como método de investigación. En este Anexo se presenta una RSL referente a la integración de procesos de explotación de información con tecnologías GIS y su aplicación para el hallazgo de patrones delictivos. En su elaboración se adoptó / aplicó el método propuesto por Kitchenham [59] y se definen los conceptos de explotación de información, minería de datos y Knowledge Discovery in Databases (KDD).

La explotación de información, constituye la sub-disciplina de la Informática que aporta a la Inteligencia de Negocio, las herramientas para la transformación de información en conocimiento. Un proceso de explotación de información se define como un grupo de tareas relacionadas lógicamente [60] que, a partir de una masa de información con un grado de valor para la organización, se ejecutan para lograr piezas de conocimiento sobre el funcionamiento de algún aspecto de esta, con un grado de valor mayor que la información original [61].

Para lograr este objetivo se utiliza las técnicas de minería de datos. Se define la minería de datos (data mining) [62] como el proceso automático mediante el cual se extrae conocimiento comprensible y útil que previamente era desconocido en las bases de datos, y en diversos formatos.

La minería de datos es un elemento fundamental de un proceso más amplio que tiene como objetivo el descubrimiento de conocimiento en grandes bases de datos [64], en inglés Knowledge Discovery in Databases.

Un proceso de explotación de información utiliza las técnicas de minería de datos para el descubrimiento de conocimiento en grandes bases de datos. Así, en esta RSL se

indagan conceptos aplicados al descubrimiento o hallazgo de patrones en base de datos delictivos o de crímenes.

A continuación, se presentan las etapas del método propuesto en [61] y seguidas en esta indagación:

La planificación de la RSL:

El principal objetivo de este documento es presentar una RSL, que pretende a través de esta técnica estudiar e identificar reportes de trabajos y experiencias relacionados con la integración de los procesos de explotación de información con los sistemas de información geográfica, y su utilización en el hallazgo de patrones delictivos. Ésta información aporta al conocimiento en torno al estado actual y síntesis de la literatura existente.

Atendiendo a este objetivo se plantean las preguntas de investigación propuestas en la Tabla I.

Tabla I

Preguntas de Investigación:

Fuente: elaboración propia

Id. Pregunta	Descripción
PI- 1	¿Qué tipo de propuesta proponen actualmente para integrar métodos de minería de datos con herramientas GIS?
PI- 2	¿Cómo se ha validado la propuesta de integración?
PI- 3	¿Existe algún artefacto software que automatice la propuesta?
PI- 4	¿Cuáles son los métodos de minería de datos actualmente aplicados al hallazgo de patrones delictivos?
PI- 5	¿Cuáles son los beneficios y limitaciones de aplicar minería de datos en el análisis delictual?
PI- 6	¿Qué herramientas o tecnologías GIS se aplican actualmente para el hallazgo de patrones delictivos y cuáles son los principales conceptos que están siendo investigados?
PI- 7	¿Cuál es la confiabilidad y el rendimiento de aplicar GIS en el análisis delictivo?

Para formar la cadena de búsqueda se considera una serie de palabras claves y sus respectivas palabras relacionadas, como se presentan en la Tabla II:

Tabla II

Definición de la cadena de búsqueda:

Fuente: elaboración propia

Palabras Claves	Palabras relacionadas
data mining	Process information exploitation, Knowledge Discovery in Databases
geographic informations system	GIS
crime analysis	criminal analysis, criminal patterns

Utilizando operadores lógicos y uniendo con AND las palabras claves y con OR las palabras relacionadas se obtuvo la siguiente cadena de búsqueda: (data mining OR Process Information Exploitation OR Knowledge Discovery in Databases) AND (geographic information system OR GIS) AND (crime analysis OR criminal analysis OR criminal patterns).

La ejecución de la búsqueda de los estudios primarios se realizó en los siguientes motores de búsqueda Web, se definió como período el 01/01/1990 al 01/10/2017 (Tabla III):

Tabla III

Búsqueda de los estudios primarios en los repositorios:

Fuente: elaboración propia

Repositorios	url
IEEE Digital Library	https://ieeexplore.ieee.org/Xplore/home.jsp
Science Direct	https://www.sciencedirect.com/
ACM Digital Library	www.acm.org/
Google Scholar	https://scholar.google.com/
SEDICI	sedici.unlp.edu.ar/

Se establecieron los siguientes criterios de inclusión para la RSL:

- Los artículos relacionados con el tema de interés a través del análisis del título, el resumen y las palabras claves. Además, se analizó como se trataban las palabras clave en el contenido total de cada artículo para decidir si tenía que ser seleccionado en el contexto de la revisión sistemática como estudio relevante (candidato potencial a convertirse en estudio primario).
- Los artículos publicados en congresos, workshops y revistas científicas.
- Los libros y tesis doctorales o de maestría.

Se consideraron los siguientes criterios de exclusión para la RSL:

- Los artículos publicados en revistas no indexadas.
- Los documentos que sean publicaciones de tutoriales.

Ejecución de la RSL:

Buscar en bases de datos científicas y extraer contenidos y datos relevantes (iterando el proceso en varias etapas):

- El criterio de la ejecución de la revisión sistemática se sustentó en el modelo iterativo e incremental.
- Es iterativo porque la ejecución (búsqueda, extracción de información y visualización de resultados) de la revisión sistemática se aplicó primero completamente en una fuente de búsqueda, y luego sobre las otras.
- Es incremental porque el documento (que es el producto) de la revisión sistemática crece y evoluciona en cada iteración hasta convertirse en el definitivo (el cual contiene los resultados).

La obtención de los estudios primarios se realizó de acuerdo al siguiente procedimiento para cada fuente de búsqueda:

- Realizar la búsqueda según la cadena de búsqueda aplicada al título (artículos encontrados), teniendo en cuenta las facilidades que proporciona cada fuente, para filtrar artículos.
- Los artículos encontrados o resultado obtenido se seleccionaron del proceso de inclusión o exclusión a partir del análisis del título y el abstract (artículos restantes).
- Conformar los estudios primarios, a partir de la lectura del texto completo (estudios primarios).
- Distribuir los artículos encontrados en cada fuente al aplicar los tres primeros pasos del procedimiento de búsqueda detallados anteriormente.

En la Tabla IV se presentan los estudios primarios, resultado de ejecutar la revisión sistemática.

Tabla IV

Distribución de los artículos encontrados por fuente:

Fuente: elaboración propia

Fuente de búsqueda	Artículos encontrados	Artículos restantes	Estudios primarios	Porcentaje por fuente
IEEE Digital Library	1.184	47	6	32%
Science Direct	363	32	1	5%
ACM Digital Library	1.089	108	3	16%
Google Scholar	965	90	4	21%
SEDICI	563	54	5	26%
TOTAL	4.164	331	19	100%

Para la obtención de estudios primarios se han excluido los artículos aplicando los criterios de inclusión y exclusión y aquellos repetidos. Es decir, artículos previamente localizados en las otras fuentes de búsqueda consultadas. Identificados los estudios primarios, se extrajo información de cada fuente y se ordenaron cronológicamente tal como se observa en las Tablas V.a, V.b y V.c.

Tabla V.a.

Extracción de información por fuente:

Fuente: elaboración propia

Fuente	Título	Autores	Congreso, workshops, revistas, libros o tesis doctorales.	Año	ID- PI
IEEE Digital Library	Association Rules Mining with GIS: An Application to Taiwan Census 2000	Chin-Jui Chang, Shiahn-Wern Shyue	2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery	2009	PI- 1, PI- 2, PI- 6
IEEE Digital Library	Research of GIS-based Spatial Data Mining Model	Wang Jinlin, Chen Xi , Zhou Kefa ,Zhang Haibo, Wang Wei	2009 Second International Workshop on Knowledge Discovery and Data Mining	2009	PI- 1, PI- 2, PI- 6
IEEE Digital Library	Analysis of Crime Factors Correlation Based on Data Mining Technology	Zhang Ying	2016 International Conference on Robots & Intelligent System (ICRIS)	2016	PI- 4, PI- 5
IEEE Digital Library	Data Mining and Predictive Analytics in Public Safety and Security	Colleen McCue	Magazine IT Professional	2006	PI- 4, PI- 5
IEEE Digital Library	Crime Data Mining: A General Framework and Some Examples	H. Chen, W. Chung, JJ Xu	Computer	2004	PI- 4, PI- 5
Sciene Direct	Geographic Knowledge Discovery and Data Mining	Robert Laurini	Geographic Knowledge Infrastructure Applications to Territorial Intelligence and Smart Cities	2017	PI- 1, PI- 2
ACM Digital Library	Criminal network analysis and visualization	Jennifer Xu, Hsinchun Chen	Magazine Communication of the ACM	2005	PI- 4, PI- 5

Tabla V.b.

Extracción de información por fuente:

Fuente: elaboración propia

Fuente	Título	Autores	Congreso, workshops, revistas, libros o tesis doctorales.	Año	ID- PI
ACM Digital Library	Crime Pattern Detection Using Data Mining	Shyam Varan Nath	IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology	2006	PI- 4, PI- 5
ACM Digital Library	Mining for offender group detection and story of a police operation	Fatih Ozgul, Julian Bondy and Hakan Aksoy	Proceedings of the sixth Australasian conference on Data mining and analytics	2007	PI- 4, PI- 5
Google Scholar	Crime Modeling and Mapping using Geospatial Technologies	Michael Leitner	Editorial Springer	2013	PI- 6, PI- 7
Google Scholar	Geographic Data Mining and Knowledge Discovery	Harvey j. Miller, Jiawei Han	Editorial Chapman & Hall – Second Edition	2009	PI- 1, PI- 2
Google Scholar	Application of GIS in Crime Analysis: A Gateway to Safe City	Gupta, R, Rajitha, K., Basu, S., Mittal, S.K.	14th Annual International Conference and Exhibition on Geospatial Information Technology and Applications	2012	PI- 6,PI- 7
Google Scholar	Crime Analysis with Crime Mapping	Rachel Boba Santos	SAGE Publications	2016	PI-6, PI- 7
SEDICI	Aplicación de minería de datos para la exploración y detección de patrones delictivos en Argentina	Valenga F., Perversi I., Fernández E., Merlino H., Rodríguez D. ,Britos P. y García-Martínez R	XIII Congreso Argentino de Ciencias de la Computación	2007	PI- 4, PI- 5
SEDICI	Explotación de información aplicada a inteligencia criminal en Argentina	Britos, P., Fernández, E., Merlino, H., Pollo-Cataneo, F., Rodríguez, D., Procopio, C., Rancan, C., García-Martínez	XIII Congreso Argentino de Ciencias de la Computación.	2008	PI- 4, PI- 5
SEDICI	Explotación de información geográfica basada en integración de ambientes de trabajo	García-Martínez, R., Merlino, H., Rodríguez, D., Martins, S., Baldizzoni, E., Diez, E., Amatriain, H., Ribeiro, F., Segura, A., Santamaría, P., Mieres, F., Aguirre, D	XX Congreso Argentino de Ciencias de la Computación	2014	PI- 1, PI- 2, PI- 3, PI- 4, PI- 5, PI- 6, PI- 7
SEDICI	Identificación y detección de patrones delictivos basada en minería de datos	Perversi, I., Valenga, F., Fernández, E., Britos P., García-Martínez, R	IX Workshop de Investigadores en Ciencias de la Computación	2007	PI- 4, PI- 5

Tabla V.c.

Extracción de información por fuente:

Fuente: elaboración propia

SEDICI	Identificación de patrones característicos de la población carcelera mediante minería de datos	Gutiérrez Rüegg, P., Merlino, H., Rancan, C., Procopio, C., Rodríguez, D., Britos, P., García-Martínez, R.

Reporte de resultados:

A continuación, se presentan los resultados obtenidos organizados por pregunta de investigación.

¿Qué tipo de propuesta se proponen actualmente para integrar métodos de minería de datos con herramientas GIS? (PI-1)

Se analizaron los 19 estudios primarios para hallar y estudiar las propuestas de integración de métodos de minería de datos con tecnologías o herramientas GIS. En la Tabla VI se observa la distribución de estudios primarios considerando la propuesta de cada uno. En la categoría **conocimiento**, se han incluido aquellos artículos que muestran evidencia empírica de la propuesta para integrar la minería de datos con GIS. Evidencia recogida a través de experimentos o investigación en acción.

Existe un artículo en donde se aplica una extensión de un software GIS para integrar con minería de datos. El cual no está disponible para la utilización en el público en general.

Bajo la categoría **otras**, se consideraron artículos que proponen métodos no formales: casos de estudios o que no especifican como se realizó la integración.

Y por último la categoría de **no existe propuesta**, hace referencia a aquellos artículos que no proponen ningún método de integración.

Tabla VI

Distribución de tipo de propuesta:

Fuente: elaboración propia

Conocimiento	Extensión de software	Otras	No existe propuesta
2	1	2	14
11%	5%	11%	73%

¿Cómo se ha validado la propuesta de integración? (PI-2)

Como se observa en la Tabla VII la mayoría de los artículos solo se limitan a mostrar un ejemplo o casos de uso para ilustrar la viabilidad de la propuesta. Lo que demuestra que es sumamente necesaria la validación de las propuestas para recoger resultados empíricos sobre la efectividad de uso de las mismas.

Tabla VII

Métodos de validación de las propuestas:

Fuente: elaboración propia

Caso de estudio	Ejemplo	Solo propuesta	No existe propuesta
3	1	1	14
17 %	5 %	5 %	73%

¿Existe algún artefacto software que automatice la propuesta? (PI-3)

Analizados los estudios primarios, se observó que existe un artículo que aplica una extensión de un software GIS para integrar con técnicas de minería de datos. El mismo se desarrolló y formó parte de un proyecto de investigación de la Universidad Nacional de Lanús, sin embargo, actualmente, esta aplicación no está disponible para su uso o descarga al público en general.

¿Cuáles son los métodos de minería de datos actualmente aplicados al hallazgo de patrones delictivos? (PI-4)

Los artículos primarios ofrecieron información referente al uso predominante de los métodos de minería de datos: clúster, clasificación y asociación (Fig. 1).

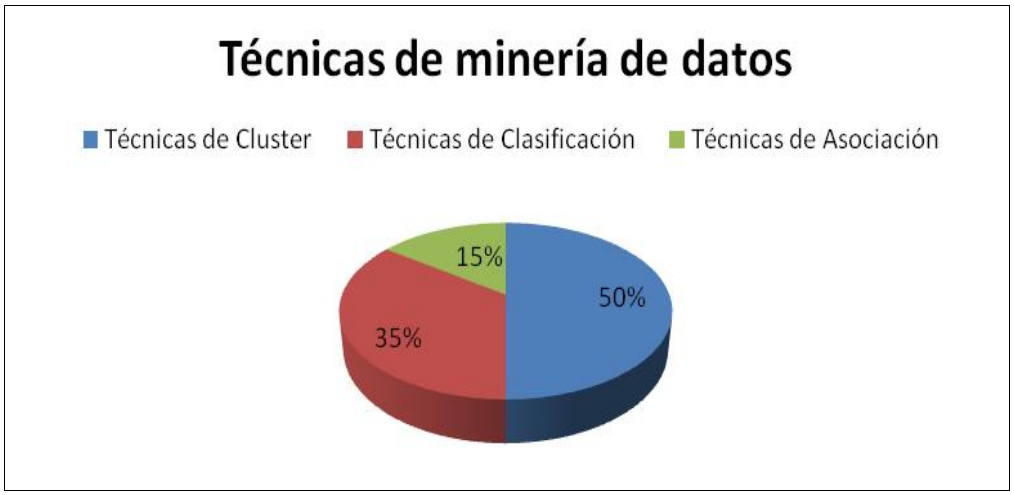


Fig. 1. Distribución por técnicas de minería de datos aplicadas al hallazgo de patrones delictivos:

Fuente: elaboración propia

Se identificó que las técnicas de aprendizaje no supervisado para detectar la agrupación de clases son las más utilizadas, seguida de las técnicas aprendizaje supervisado para la clasificación y finalmente las técnicas de asociación.

¿Cuáles son los beneficios y limitaciones de aplicar minería de datos en el análisis delictual? (PI-5)

Los estudios primarios brindaron información respecto a los principales beneficios y las limitaciones de aplicar minería de datos al análisis delictual:

- Los beneficios del análisis predictivo aplicado a la seguridad y la protección son la identificación temprana y caracterización de una posible amenaza presenta.
- Mediante la aplicación de estas técnicas la prevención es casi siempre menos costosa que la recuperación. Especialmente si se mide en términos humanos.
- Las técnicas de agrupamiento permiten identificar sospechosos que conducen crímenes de manera similar o distinguir entre grupos pertenecientes a diferentes pandillas.
- Los investigadores pueden aplicar minería de reglas de asociación para descubrir perfiles de los intrusos en una red, con el fin de aportar en la detección de una futura red potencial de ataques.
- Las técnicas de clasificación encuentran propiedades comunes entre diferentes entidades delictivas y las organiza en clases predefinidos.
- Algunas de las limitaciones de usar estas técnicas incluyen que el análisis del patrón del crimen solo puede ayudar al detective, no sustituirlos. Además, la minería de datos es sensible a la calidad de los datos de entrada que pueden ser inexactos, la falta de información también es una problemática y puede ser propenso a errores dado el ingreso erróneo de datos.

¿Qué herramientas o tecnologías GIS se aplican actualmente para el hallazgo de patrones delictivos y cuáles son los principales conceptos que están siendo investigados? (PI-6)

Se observó de los 19 estudios primarios hallados las tecnologías GIS actualmente aplicadas en el hallazgo de patrones delictivos. En la Tabla VIII se muestra el número de trabajos que cumple este criterio por herramienta.

Tabla VIII

Herramientas GIS utilizadas en el análisis delictual:

Fuente: elaboración propia

ArcGIS Desktop	gvSIG Desktop	Otros	No aplica ninguna herramienta
2	3	1	13
11%	16%	5%	68%

Se indagó en torno al uso del software bajo licencia ArcGIS Desktop comercializado por la empresa ESRI. Este GIS permitir llevar a cabo multitud de tareas relacionadas con estadísticas espaciales y determinados procesos centrados en el análisis del crimen.

Por otro lado, se observa el uso del software gvSIG, paquete de software de uso libre para el manejo de información geográfica con precisión cartográfica.

En la categoría **otros**, se incluyen aquellos artículos que mencionan el uso de tecnologías GIS, sin especificar el nombre de la herramienta.

La categoría **no aplica ninguna herramienta**, representa a los artículos que no aplican ninguna herramienta GIS.

Así, según los estudios primarios los principales conceptos en investigación relacionados con la aplicación de GIS en el análisis de delitos son:

- Generación de mapas del delito: ofrecen una imagen rápida, concreta y fácilmente interpretable de la intensidad con que los hechos delictivos se producen.
- Identificación de las zonas calientes: se trata de un área geográfica que presenta un nivel de delitos o desorden más elevado que el promedio. Son agrupaciones y conglomerados de delitos que pueden existir a diferentes escalas.
- Determinación de los puntos calientes, es decir las ubicaciones donde principalmente ocurren los delitos.

¿Cuál es la confiabilidad y el rendimiento de aplicar GIS en el análisis delictivo? (PI-7)

Los estudios primarios también brindaron información sobre los principales beneficios y las limitaciones de aplicar tecnologías GIS en el análisis delictual:

- Los GIS permiten que los personales de las fuerzas de seguridad puedan: planificar de forma efectiva las respuestas ante emergencias, determinar las prioridades en las medidas de mitigación, analizar los eventos pasados y predecir los eventos futuros.
- Los GIS en el análisis de delitos facilita el mapeo, la visualización y análisis de puntos calientes del crimen.
- Los GIS ayudan a los oficiales del crimen a determinar sitios potenciales de crimen al examinar complejos criterios aparentemente no relacionados y desplegarlos en un mapa.

Conclusiones en torno a la RSL:

En este documento se expuso el procedimiento que permitió recopilar y analizar la literatura existente sobre la integración de procesos de explotación de información con tecnologías GIS y su aplicación para el hallazgo de patrones delictivos en el periodo comprendido entre el 1 de enero de 1990 y 1 de octubre de 2017.

Los 19 estudios primarios permitieron responder las preguntas de investigación y elaborar un estado de la cuestión preliminar, guiando el presente TFM.

Los análisis de los resultados obtenidos permiten afirmar que:

- La aplicación de las tecnologías GIS es un tema de suma actualidad para el análisis de patrones delictivos. El análisis de artículos ha permitido observar su aplicación y utilidad.
- Las utilizaciones de GIS junto con las técnicas de minería de datos han demostrado tener un potencial sumamente alto para el análisis de delitos en cualquier región. Las características que aportan estas tecnologías en este tipo de análisis que reflejan claridad y conocimiento clave para entender la problemática.

- Existe una mayor tendencia a las propuestas basadas en casos de uso, para lograr entender su aplicación y verificarla.
- También existe una amplia cantidad de artículos que basan la demostración utilizando software GIS bajo una licencia paga.
- Se observa una clara necesidad de verificar las propuestas a través de métodos empíricos.

Es fundamental disponer de algún procedimiento claro y preciso que pueda ayudar a integrar ambas tecnologías dado su amplia utilidad en el dominio de conocimiento.

Anexo 2. Traducción de reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 1.

En este Anexo se incluye la interpretación de las reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 1, correspondiente a las Tablas XIV.a, XIV.b y XV.c.

Las reglas de pertenencia al clúster c_som_1_1 se describen como:

- Si el elemento sustraído fue una bicicleta u objeto personal, y el tipo de lugar fue en el interior de un rodado o comercio, y si el ataque fue de tipo forcejeo o arrebato.
- Si el elemento sustraído fue una bicicleta y si la clase de arma fue de tipo arma de fuego u otra o ninguna, y si el ataque fue de tipo ataque brutal o no existió ataque alguno.
- Si el día de la semana fue el día martes, y el rango de horario en el que se cometió el delito fue de 0-4 am o de 4-8 am, y el mes fue entre marzo y junio, y el tipo de lugar fue en la vía pública o en un domicilio particular y si el ataque fue de tipo forcejeo u arrebato.
- Si la clase de arma fue de tipo arma de fuego u otra o ninguna, y si día de semana fue sábado, domingo, lunes o martes, y si el rango de horario en el que se cometió el delito fue de 0-4 am o de 4-8 am, y el mes fue entre marzo y junio, y el tipo de lugar fue en la vía pública o en un domicilio particular y si el ataque fue de tipo forcejeo u arrebato.
- Si el día de la semana fue el día miércoles, jueves o viernes, y si el tipo de lugar fue en la vía pública o en un domicilio particular y si el ataque fue de tipo forcejeo u arrebato.

Las reglas de pertenencia al clúster c_som_1_2 se describen como:

- Si el ataque fue de tipo forcejeo, y si ocurrió en el mes de enero o febrero, y si el día fue un sábado, domingo, lunes o martes, y si el tipo de lugar fue en la vía pública o en un domicilio particular.

- Si el rango de horario en el que se cometió el delito fue de 16-20 pm o de 20-24 pm, y si el ataque fue de tipo arrebato, y si el mes fue entre enero y febrero, y si el día fue un sábado, domingo, lunes o martes, y si el tipo de lugar fue en la vía pública o en un domicilio particular.
- Si la clase de arma fue de tipo arma blanca, y si el día fue un martes, y si el rango de horario en el que se cometió el delito fue de 0-4 am o de 4-8 am, y si el mes fue entre marzo y junio, y si el tipo de lugar fue en la vía pública o en un domicilio particular, y si el ataque fue de tipo arrebato o forcejeo.
- Si el rango de horario en el que se cometió el delito fue de 8 am o de 24 pm, y si el mes fue entre marzo y junio, y si día de semana fue sábado, domingo, lunes o martes, y si el tipo de lugar fue en la vía pública o en un domicilio particular, y si el ataque fue de tipo arrebato o forcejeo.

Las reglas de pertenencia al clúster c_som_2_1 se describen como:

- Si el elemento sustraído fue una motocicleta, de tipo domiciliario, vehículo, dinero o si no hubo elemento sustraído, y si el tipo de lugar fue en un comercio o interior de un rodado, y si el ataque fue de tipo forcejo o arrebato.
- Si el tipo de lugar fue en un comercio o en el interior de un rodado, y si la clase de arma fue de tipo arma blanca, y si el ataque fue de tipo ataque brutal o no existió ataque.
- Si el tipo de lugar fue en un comercio, interior de un rodado, o domicilio particular, si el elemento sustraído fue un objeto personal, una motocicleta, de tipo domiciliario, vehículo, dinero o si no hubo elemento sustraído, y si la clase de arma fue de tipo arma de fuego, u otra o ninguna, y si el ataque fue de tipo ataque brutal o no existió ataque.
- Si el mes fue entre marzo y junio, y si el rango de horario en el que se cometió el delito fue de 0-4 am, o 4-8 am, o 8-12 pm o 16-20 pm, y si el tipo de lugar fue en la vía pública, y si el elemento sustraído fue un objeto personal, una motocicleta, de tipo domiciliario, vehículo, dinero o si no hubo elemento sustraído, y si la clase de arma fue de tipo arma de fuego, u otra o ninguna, y si el ataque fue de tipo ataque brutal o no existió ataque.

Las reglas de pertenencia al clúster c_som_2_2 se describen como:

- Si el tipo de lugar fue en la vía pública o domicilio particular, y si la clase de arma fue de tipo arma blanca, y si el ataque fue de tipo ataque brutal o no existió ataque.
- Si el rango de horario en el que se cometió el delito fue de 16-20 pm, o 20-24 pm, y si el tipo de lugar fue en la vía pública, y si el elemento sustraído fue un objeto personal, una motocicleta, de tipo domiciliario, vehículo, dinero o si no hubo elemento sustraído, y si el ataque fue de tipo ataque brutal o no existió ataque.
- Si el mes en el que ocurrió el delito fue en enero o febrero, y rango de horario en el que se cometió el delito fue de 0-4 am, o 4-8 am, 8-12 am, 12-16 pm, y si el tipo de lugar fue en la vía pública, y si el elemento sustraído fue un objeto personal, una motocicleta, de tipo domiciliario, vehículo, dinero o si no hubo elemento sustraído, y si la clase de arma utilizada fue arma de fuego, otra o ninguna y si el ataque fue de tipo ataque brutal o no existió ataque.
- Si el rango de horario en el que se cometió el delito fue de 0-4 am, o 4-8 am, 8-12 am, 12-16 pm, y si el ataque fue de tipo arrebato, ataque brutal o no existió ataque, y si el mes en el que ocurrió el delito fue en enero o febrero, y si ocurrió el día sábado, domingo, lunes o martes, y si el tipo de lugar fue en la vía pública o domicilio particular, y si el ataque fue de tipo forcejeo o arrebato.

Anexo 3. Traducción de reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 2.

En este Anexo se expone la interpretación de las reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 2, correspondiente a las Tablas XVI.a y XVI.b.

Las reglas de pertenencia al clúster c_som_1_1 se describen como:

- Si el elemento sustraído de tipo objeto personal; si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26.
- ; y si el lugar fue en la vía pública; y si el delito fue de tipo robo.
- Si el tipo de lugar donde se cometió el delito fue en la vía pública; y si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17; y si el delito fue de tipo robo.

Las reglas de pertenencia al clúster c_som_1_2 se describen como:

- Si el delito fue de tipo hurto.

Las reglas de pertenencia al clúster c_som_2_1 se describen como:

- Si el tipo de lugar donde se cometió el delito fue en un domicilio particular; y si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25, N° 26, N° 27, N° 28 y N° 29; y si el delito fue de tipo robo.
- Si el tipo de lugar donde se cometió el delito fue en un domicilio particular; y si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20 y N° 21; y si el delito ocurrió en algunas de las siguientes calles: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13 y N° 14; y si el delito fue de tipo robo.

- Si los barrios en donde se cometió el delito fueron algunas de las siguientes: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26; y si las calles fueron N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21 y N° 22; y si el delito fue de tipo robo.
- Si el elemento sustraído de tipo domiciliario, una motocicleta, una bicicleta, un vehículo, dinero o si no hubo elemento sustraído; si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7 y N° 8; y si las calles fueron: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26; y si el lugar fue en la vía pública; y si el delito fue de tipo robo.

Las reglas de pertenencia al clúster c_som_2_2 se describen como:

- Si el tipo de lugar donde se cometió el delito fue en el interior de rodado o en un comercio; y si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26; y si el delito fue de tipo robo.
- Si el tipo de lugar donde se cometió el delito fue en un domicilio particular, o interior de rodado o en un comercio; y si el delito ocurrió en algunos de los siguientes barrios: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17 y N° 18; y si el delito fue de tipo robo.

Anexo 4. Traducción de reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 3.

En este Anexo se expone la interpretación de las reglas de pertenencia a los clusters formados en el objetivo de minería de datos Nº 3, correspondiente a las Tablas XVIII.a y XVIII.b.

Las reglas de pertenencia al clúster c_som_1_1 se describen como:

- Si la edad de la víctima fue entre los 46 y 57 años; y si el tipo de ataque fue de tipo ataque brutal o si no existió ataque alguno.
- Si la edad del sospechoso fue entre los 15 y 25 años; y si la edad de la víctima fue entre los 31 y 35 años; y si el lugar en donde ocurrió el delito fue en la vía pública; y si el tipo de ataque fue de tipo arrebato o forcejeo.
- Si la edad de la víctima fue entre los 36 y 57 años; y si el tipo de ataque fue de tipo arrebato o forcejeo.

Las reglas de pertenencia al clúster c_som_1_2 se describen como:

- Si la edad de la víctima fue entre los 31 y 35 años; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si el tipo de ataque fue de tipo arrebato o forcejeo.
- Si el delito fue de tipo robo; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si la edad de la víctima fue entre los 15 y 45 años; y si el tipo de ataque fue de tipo ataque brutal o si no existió ataque alguno.
- Si el delito ocurrido lo sufrió alguna persona de sexo masculino; si el delito fue de tipo hurto; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si la edad de la víctima fue entre los 15 y 45 años.
- Si el elemento sustraído fue un vehículo, dinero o si no hubo elemento sustraído; si el delito ocurrido lo sufrió alguna persona de sexo femenino; si el delito fue de tipo hurto; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si la edad de la víctima fue

entre los 15 y 45 años; y si el tipo de ataque fue de tipo arrebato, ataque brutal o si no existió ataque alguno.

Las reglas de pertenencia al clúster c_som_2_1 se describen como:

- Si el elemento sustraído fue una bicicleta, algún objeto personal, una motocicleta o de tipo domicilio; si el delito ocurrido lo sufrió alguna persona de sexo femenino; si el delito fue de tipo hurto; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si la edad de la víctima fue entre los 15 y 45 años; y si el tipo de ataque fue de tipo ataque brutal o si no existió ataque alguno.
- Si la edad de la víctima fue entre los 15 y 45 años; y si el lugar en donde ocurrió el delito fue en la vía pública; si el tipo de ataque fue de tipo ataque brutal o no existió ataque alguno.

Las reglas de pertenencia al clúster c_som_2_1 se describen como:

- Si la edad de la víctima fue entre los 15 y 35 años; y si el lugar en donde ocurrió el delito fue en un domicilio particular, en el interior de un rodado, o en algún comercio; y si el tipo de ataque fue de tipo arrebato o forcejeo.
- Si la edad de la víctima fue entre los 15 y 35 años; y si el lugar en donde ocurrió el delito fue en la vía pública; y si el tipo de ataque fue de tipo arrebato o forcejeo.
- Si la edad del sospechoso fue entre los 26 y 35 años; y si la edad de la víctima fue entre los 31 y 35 años; y si el lugar en donde ocurrió el delito fue en la vía pública; y si el tipo de ataque fue de tipo arrebato o forcejeo.