ESCUELA POLITÉCNICA NACIONAL

DEPARTAMENTO DE INGENIERÍA EN SISTEMAS

APLICACIÓN DE TÉCNICAS DE MINERÍA DESCRIPTIVA DE DATOS PARA EL DESCUBRIMIENTO Y VISUALIZACIÓN DE PATRONES EN DELITOS DE ROBO EN ECUADOR

MAGISTER EN SISTEMAS DE LA

INFORMACIÓN CON MENCIÓN EN INTELIGENCIA DE

NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

Cisneros Morillo Jaime Gabriel

DIRECTOR: María Gabriela Pérez Hernández

Quito, Junio 2024

AVAL

Como director del trabajo de titulación “Aplicación de técnicas de minería descriptiva de datos para el descubrimiento y visualización de patrones en delitos de robo en ecuador” desarrollado por Jaime Gabriel Cisneros Morillo, estudiante de la Maestría de Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la defensa oral.

María Gabriela Pérez Hernández, PhD.

DIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, Jaime Gabriel Cisneros Morillo, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración dejo constancia de que la Escuela Politécnica Nacional podrá hacer uso del presente trabajo según los términos estipulados en la Ley, Reglamentos y Normas vigentes.

Jaime Gabriel Cisneros Morillo

DEDICATORIA

El presente trabajo está dedicado a mi padre Jaime Cisneros, un hombre de lucha y trabajo, quien con amor y disciplina me crio y enseño toda su sabiduría. Desde pequeño me transmitió su amor por la lectura y la ciencia demostrando que todo se puede conseguir con constancia y trabajo duro.

También está dedicado a mis dos madres Jenny y Gloria, quienes durante toda la vida me han cuidado, apoyado y consentido con su amor y cariño característico.

Gracias a estas 3 personas el presente trabajo y cualquier logro futuro es y serán posibles.

AGRADECIMIENTO

Gracias a mi madre Jenny, por su apoyo durante todo el proceso académico.

Mi más sincero agradecimiento al Msc. Lenin Falconí, sin el cual el presente trabajo no se hubiese concretado. Muchas gracias por su ayuda, tiempo y consejo durante todo el desarrollo del presente trabajo.

Gracias a la Dr. María Pérez, por su guía, tiempo y paciencia durante el desarrollo de esta tesis.

ÍNDICE DE CONTENIDO

RESUMEN

PALABRAS CLAVE: Minería de datos, robo, reglas de asociación, patrones de robo

El presente trabajo aborda el problema creciente de la delincuencia, específicamente los delitos de robo en Ecuador, presentando un enfoque que utiliza técnicas de Minería descriptiva de Datos para gestionar la información de delitos de robo proporcionada por la Fiscalía General del Estado. Con estas técnicas se busca identificar patrones y visualizar información útil de los datos, con el objetivo de identificar noticias del delito de robo relacionadas. El estudio se enfoca en el uso de algoritmos de *clustering*, para identificar zonas geográficas relevantes en cada provincia, seguidos de la extracción de reglas de asociación para encontrar patrones de delitos de robo. Se concluye que la metodología CRISP-DM es efectiva, así como el uso de *K-means* para identificar zonas geográficas significativas en cada provincia y *FP-growth* para la minería de reglas de asociación. Además, se determina el soporte mínimo óptimo para la extracción de reglas de asociación, que deriven en patrones de robo, que después relacionarán noticias del delito. Los resultados obtenidos permiten identificar patrones delictivos y noticias relacionadas, contribuyendo a una mejor comprensión y gestión de la delincuencia en el país.

ABSTRACT

KEYWORDS: Data mining, theft, association rules, theft patterns

This paper addresses the growing problem of crime, specifically theft crimes in Ecuador, by presenting an approach that uses descriptive data mining techniques to manage information on theft crimes provided by the Fiscalía General del Estado. These techniques seek to identify patterns and visualize useful information from the data, with the objective of identifying related theft crime news. The study focuses on the use of clustering algorithms to identify relevant geographic zones in each province, followed by the extraction of association rules to find theft crime patterns. It is concluded that the CRISP-DM methodology is effective, as well as the use of K-means to identify significant geographical zones in each province and FP-growth for association rule mining. In addition, the optimal minimum support for the extraction of association rules, which derive theft patterns, which will then relate crime news, is determined. The results obtained allow the identification of crime patterns and related news, contributing to a better understanding and management of crime in the country.

1. INTRODUCCIÓN

La delincuencia es un problema complejo que afecta a todos los países del mundo. En Ecuador, el problema de la delincuencia ha ido en aumento en los últimos años. Una muestra de esta lacra social se presenta en la Figura 1, problema que se ha convertido en uno de los principales desafíos que afronta el país. Según los registros de la Fiscalía General del Estado (FGE), en 2022 se presentaron 262.341 denuncias de robo [1]. El incremento del número de denuncias investigadas genera congestión en los sistemas de justicia. Pues, el Fiscal, a través de diferentes diligencias y pericias, trata de juntar evidencias para descubrir responsables y justificar la formulación de cargos, en un proceso que en la actualidad, se realiza de manera manual[2].

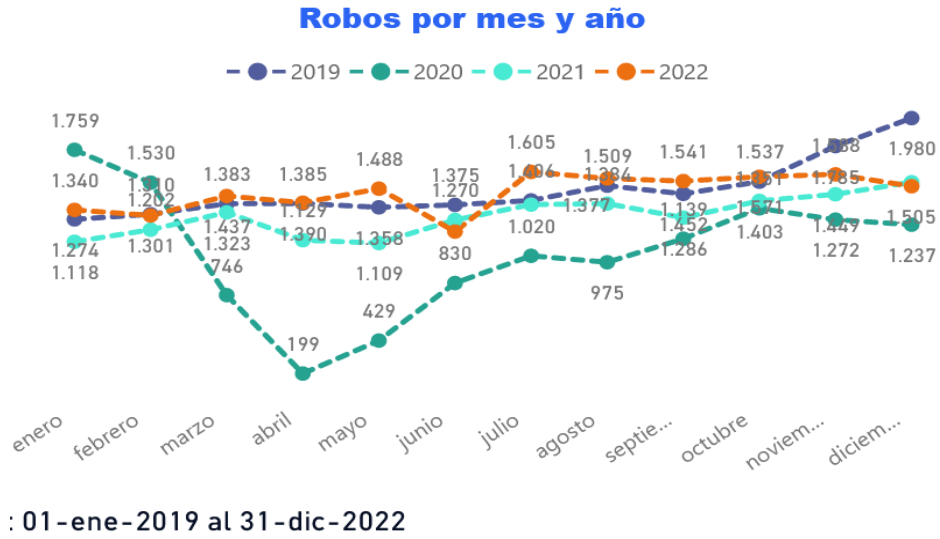


Figura 1. Denuncias de robo en Ecuador por mes, en los años 2019 – 2022 [1]

En este contexto, las técnicas descriptivas de Minería de Datos contribuyen al descubrimiento de patrones que pueden asistir en la automatización y procesamiento de grandes volúmenes de información, para generar procesos más agiles y oportunos en la gestión procesal penal de la FGE.

Para comprender el marco del presente trabajo, en los siguientes apartados se definirá brevemente conceptos teóricos penales fundamentales, del estudio.

1.1. Conceptos principales de esta área

Delito: Es la infracción penal sancionada con pena privativa de libertad mayor a 30 días [3]

Noticia de delito: Es un término empleado por la FGE para denominar todas las formas en las que se conoce una infracción penal o delito[1]

Robo: de acuerdo al art 189 del código integral penal (CIP), la persona que mediante amenazas o violencias sustraiga o se apodere de cosa mueble ajena, sea que la violencia tenga lugar antes del acto para facilitarlo, en el momento de cometerlo o después de cometido para procurar impunidad, será sancionada con pena privativa de libertad de cinco a siete años. Cuando el robo se produce únicamente con fuerza en las cosas, será sancionada con pena privativa de libertad de tres a cinco años. Si se ejecuta utilizando sustancias que afecten la capacidad volitiva, cognitiva y motriz, con el fin de someter a la víctima, de dejarla en estado de somnolencia, inconciencia o indefensión o para obligarla a ejecutar actos que con conciencia y voluntad no los habría ejecutado, será sancionada con pena privativa de libertad de cinco a siete años. Si a consecuencia del robo se ocasionan lesiones de las previstas en el numeral 5 del artículo 152 se sancionará con pena privativa de libertad de siete a diez años. Si el delito se comete sobre bienes públicos, se impondrá la pena máxima, dependiendo de las circunstancias de la infracción, aumentadas en un tercio. Si a consecuencia del robo se ocasiona la muerte, la pena privativa de libertad será de veintidós a veintiséis años. La o el servidor policial o militar que robe material bélico, como armas, municiones, explosivos o equipos de uso policial o militar, será sancionado con pena privativa de libertad de cinco a siete años [3].

1.2. Tipos de robo

Dentro del delito de robo la FGE distingue seis tipos, tal como se muestra en la Figura 2.

Tipo de robo por periodo (año / mes)

TIPO DE ROBO	2019	2020	2021	2022	**Total**
Robo a personas	31.002	20.126	25.440	31.485	**108.053**
Robo de motos	8.020	6.666	9.178	14.567	**38.431**
Robo a domicilio	11.099	7.369	8.198	8.386	**35.052**
Robo de bienes, accesorios y autopartes	9.686	6.214	8.000	8.354	**32.254**
Robo de carros	5.653	4.596	6.911	11.372	**28.532**
Robo de unidades económicas	5.731	4.078	4.857	5.353	**20.019**
Total	**71.191**	**49.049**	**62.584**	**79.517**	**262.341**

Figura 2. Clasificación de delitos de robo de acuerdo a la FGE y su conteo por año a nivel nacional [1]

A continuación, se describe cada uno de estos tipos de robo:

Robo a personas: Evento que se caracteriza cuando una persona o grupo de personas mediante amenazas o violencia sobre la o las víctimas, sustraiga o se apodere de un bien mueble propio o del que sea custodio, que porte en el momento del hecho, sea en un lugar público o privado.

Robo a domicilio: Evento que se caracteriza cuando una persona o grupo de personas ingrese a un domicilio ajeno mediante amenazas, violentando o haciendo uso de la fuerza, con el fin de sustraer o apoderarse de un bien u objeto que se encuentre en el domicilio o sea parte del bien inmueble, excepto vehículos a motor [4]

Robo de carros: Evento que se caracteriza cuando una persona o grupo de personas mediante amenazas, violencia o uso de la fuerza, sustraiga totalmente un carro propio o en custodio, sea en un lugar público o privado, independientemente que posterior al evento sea recuperado total o parcialmente el carro. Considerando como carros: camión, automóviles, cabezales, tanqueros, tráileres, buses, camionetas, retroexcavadoras, tractores, equipos camineros; excepto motocicletas, cuadrones, y vehículos no terrestres [4]

Robo de motos: Evento que se caracteriza cuando una persona o grupo de personas mediante amenazas, violencia o uso de la fuerza, sustraiga totalmente una motocicleta, sea en un lugar público o privado. Considerando como motocicleta: motos, pasolas, cuadrones, tricimotos [4].

Unidades económicas: Persona natural o jurídica, que, bajo una sola dirección o control, combina actividades y recursos con la finalidad de producir y/o vender bienes y servicios, independientemente de poseer o no Registro Único de Contribuyente (RUC) o Régimen Impositivo Simplificado (RISE) [4].

Robo a unidades económicas: Evento que se caracteriza cuando una persona o grupo de personas concurran a una unidad económica y, mediante amenazas, violencia o uso de la fuerza, sustraen o se apoderen de bienes, dinero u objetos propios de esta actividad [4].

Robo de bienes, accesorios y autopartes de vehículos: Evento que se caracteriza cuando una persona o grupo de personas mediante violencia o uso de la fuerza sobre los vehículos (carros y motos), sustraigan o se apodere de uno o varios accesorios, autopartes del vehículo o bienes que estén al interior del vehículo, sea en un lugar público o privado. Cuando la víctima se encuentre presente y también le hayan sustraído un bien que no sea accesorio o autoparte del vehículo, en este caso primará el indicador de robo a personas [4].

Dado la gran cantidad de registros de delito de robo, esta tesis propone gestionar la información de delitos de robos proporcionada por la FGE con el objetivo de visualizar y extraer información útil, utilizar técnicas de minería descriptiva de datos para el descubrimiento y visualización de patrones que permitan descubrir patrones relevantes del delito de robo y así proporcionar un aporte a la carga procesal de la FGE.

1.3. Objetivo General

Aplicar técnicas de minería descriptiva de datos para el descubrimiento de patrones en delitos de robo.

1.4. Objetivos Específicos

- Revisar el estado del arte de técnicas de minería descriptiva de datos aplicadas en el reconocimiento de patrones delictivos, mediante la exploración de artículos científicos relevantes como revisiones sistemáticas de literatura realizadas en este ámbito o de existir, propuestas similares.
- Implementar un *dashboard* para el análisis de los datos proporcionados por la FGE.

- Revisar y determinar los modelos/técnicas de minería descriptiva de datos apropiados para descubrir noticias del delito relacionadas en los datos generados por la FGE
- Evaluar el desempeño de los modelos seleccionados para el descubrimiento de patrones en delitos de robo.

1.5. Marco Teórico

1.5.1. Minería de datos

La minería de datos es una rama interdisciplinar de las ciencias computacionales que procesa grandes cantidades de datos y tiene como objetivo, obtener información con una estructura comprensible para su posterior uso. Dentro de este análisis convergen herramientas de: Inteligencia artificial, Aprendizaje de máquina, Estadística e inteligencia de negocios [5].

En el análisis de información de crímenes, la minería de datos se considera una poderosa herramienta que permite a las organizaciones de investigación y seguridad, descubrir patrones significativos dentro de grandes cantidades de datos [6].

Como se presenta en la Figura 3, dentro de la minería de datos se puede identificar cinco técnicas principales que son: agrupación (*clusstering*), clasificación, minería de reglas de asociación, análisis de regresión y detección de datos anómalos [7].

Figura 3 Principales técnicas dentro de la minería de datos [7]

Estas técnicas se pueden enfocar para resolver distintas tareas de minería de datos, de acuerdo con [8] las tareas de minería de datos se clasifican en dos tipos: Predictivas y Descriptivas Figura 4.

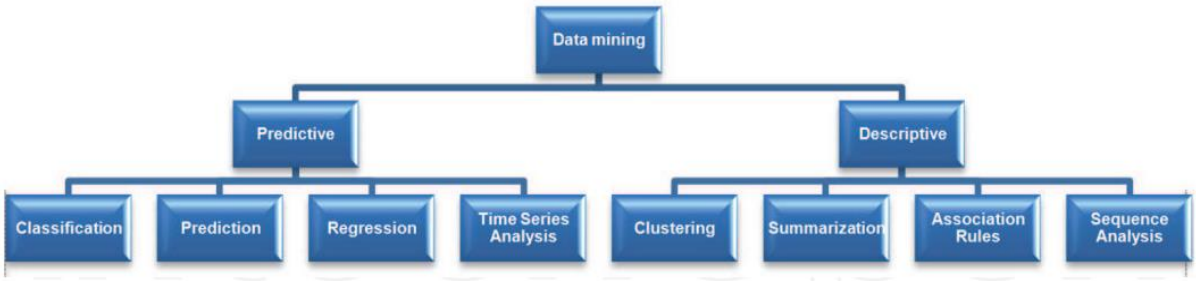


Figura 4.Técnicas de minería de datos de acuerdo a su enfoque [9]

1.5.2. Minería descriptiva de datos

El enfoque descriptivo de la minería de datos identifica patrones y relaciones en los datos. Un modelo descriptivo explora propiedades de los datos examinados y generalmente es utilizada para identificar correlaciones, frecuencias y tabulaciones cruzadas dentro de la base de datos [9].

Se define a las técnicas de minería descriptiva de datos como métodos que nos permiten encontrar subgrupos y descubrir patrones útiles. De acuerdo con [9] y tal como se presenta en la Figura 4, la minería descriptiva de datos abarca las siguientes técnicas: Summarization (resumen), clustering (agrupación), association rules (minería de reglas de asociación) sequence Analysis (análisis de secuencias).

A continuación, conceptualizaremos cada una de las técnicas mencionadas haciendo énfasis en las utilizadas en el presente trabajo.

1.5.3. Sumarización (Summarization)

Se define a la técnica de *summarization* como el proceso de obtener una versión concisa e informativa de los datos originales, donde los términos “conciso” e “informativo” dependen del contexto en el que se está utilizando dicha técnica. La sumarización hace que los datos sean comprensibles para análisis posteriores, es una técnica ampliamente usada en el procesamiento de grandes cantidades de datos[10]. En resumen, la técnica de sumarización produce resúmenes de los datos con alta calidad.

Dentro del amplio campo de contextos de la *sumarización* se puede identificar dos enfoques principales se acuerdo al tipo de datos, estos son: Datos estructurados y datos

no estructurados, en la Figura 5 se muestra los distintos métodos de aplicación para cada enfoque.

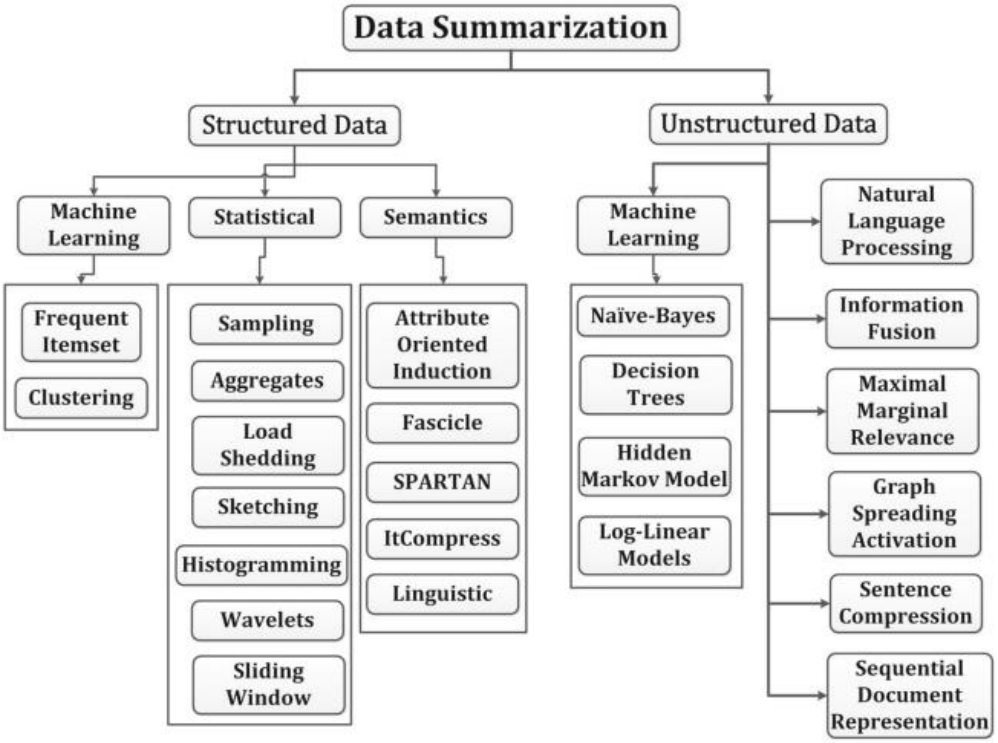


Figura 5 Clasificación de métodos de *sumarización* de acuerdo con el tipo de datos [10]

Ya que los datos para el desarrollo del presente proyecto son estructurados, se revisa brevemente los principales métodos de la técnica de *sumarización* en este enfoque. A continuación, se presentan los métodos que se utilizan en el presente trabajo.

Métodos estadísticos

Utilizan herramientas estadísticas para obtener información útil y resumida de los datos, como se muestra en la Figura 5 y de acuerdo con [10], existen 7 métodos estadísticos principales, de los cuales profundizaremos en:

- **Agregados (*Aggregates*):** Consiste en la compilación de datos mediante operaciones matemáticas como sumas, medidas de tendencia central, conteos, mínimos y máximos. Este método es utilizado para condensar grandes volúmenes de datos en métricas clave que resumen la información esencial del conjunto de datos. Los agregados proporcionan una visión general rápida y fácil de interpretar, facilitando la comprensión de las tendencias y patrones generales[11].

- ***Histogramming*:** Esta técnica involucra la creación de histogramas para representar la distribución de un conjunto de datos. Un histograma divide los datos en intervalos o bins y cuenta la frecuencia de los datos en cada bin. Este método proporciona una visualización clara de la distribución de los datos, identificando patrones como la centralización, dispersión y presencia de *outliers* [12].

Aprendizaje de máquina

Comprende en el uso métodos de aprendizaje de maquina con el objetivo de resumir grandes cantidades de información de manera efectiva[13]. Dentro de esta clasificación tenemos

- **Conjuntos de elemento frecuentes (*Frecuent Itemset*):** Es una técnica de *sumarización* basada en conjuntos de elementos frecuentes, que permite identificar y resumir patrones y asociaciones comunes en grandes conjuntos de datos transaccionales. Utilizando algoritmos como *Apriori* y *FP-Growth*, se pueden extraer estos conjuntos frecuentes y utilizarlos para crear resúmenes informativos que capturen la esencia de los datos originales [11].

- **Clustering:** Es uno de los algoritmos más importantes dentro del aprendizaje de máquina, cuyo objetivo es encontrar grupos naturales o intrínsecos dentro de una base de datos no etiquetada, en el contexto del resumen de datos, este método obtiene un nuevo conjunto de datos más pequeño con información esencial del conjunto original [14]. En la siguiente sección se revisará más a fondo esta técnica con el enfoque de descubrimiento de patrones.

1.5.4. Clustering

El clustering o agrupación de datos, consiste en dividir a los datos que comparten características comunes en conjuntos que han sido determinados bajo un criterio de importancia o usabilidad, se lo puede definir como un método de clasificación no supervisado que agrupa datos de tal manera que, los datos pertenecientes a un grupo o cluster son relativamente idénticos entre sí, y muy diferentes de los que pertenecen a otro grupo [15].

En la minería descriptiva de datos el clustering se utiliza para segmentar los datos en grupos significativos basados en la distancia que existe entre cada individuo, con el objetivo de descubrir conjuntos de patrones.

Junto con el desarrollo de la inteligencia artificial, metodologías computacionales y mayor acceso a bases de datos masivos, se han desarrollado diferentes métodos de implementación de clustering con el objetivo de manejar distintos tipos de datos y mejorar su eficiencia. Como se muestra en la figura Los principales son: Clustering jerárquico, basado en densidad, basado en partición y clustering basado en malla [7].

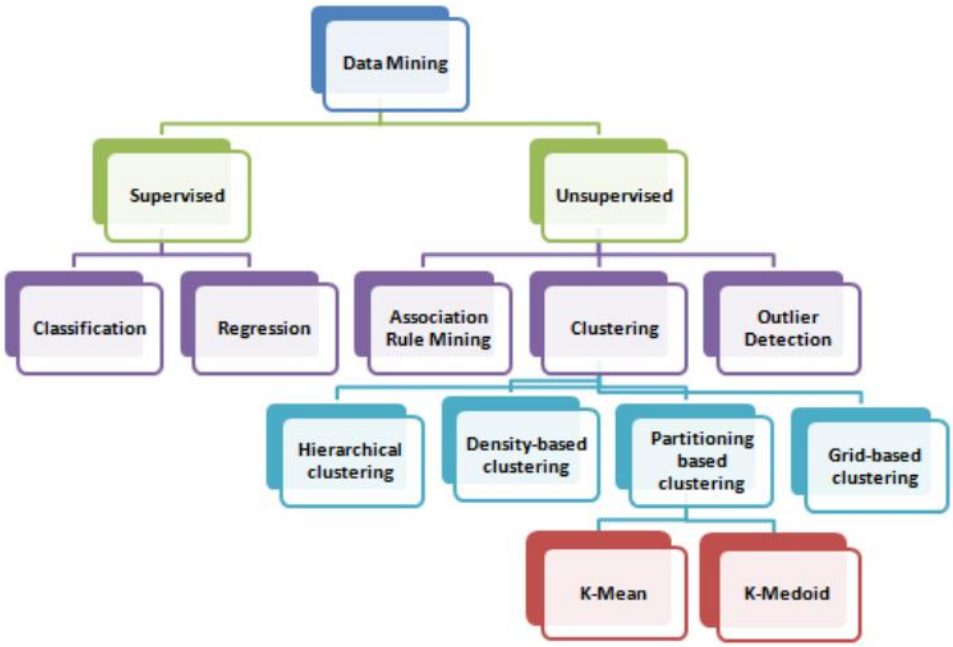


Figura 6 Clasificación de método de clustering

A continuación, revisaremos cada uno de los métodos mostrados en la figura 6.

Clustering jerárquico

Esta técnica construye clústeres basándose en similitudes, proporciona una visualización jerárquica de los conjuntos formados y permite identificar el nivel de profundidad de los mismos. Consiste en la creación de grupos jerárquicos mediante dendogramas. Un dendograma es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud [16].

Clustering basado en densidad

Es una técnica de *clustering* se basada en la densidad de los datos representados como puntos en el espacio, e identifica los clústeres tomando las regiones con mayor densidad, estos son divididos por regiones con menor densidad de datos. El valor de densidad de un punto es determinado, evaluando la cantidad de vecinos que posee en un radio determinado Esta técnica no se ve afectada por la existencia de datos atípicos ni la

existencia de *clusters* no convexos, además de ser muy efectiva en el análisis de bases de datos multidimensionales [15]

De acuerdo con [7], dentro de los algoritmos de clustering basados en densidad más comunes tenemos:

- DBSCAN (Density Based Spatial Clustering of Applications with Noise) agrupación espacial de aplicaciones con ruido.
- OPTICS (Ordering Points To Identify the Clustering Structure) Identificación del cluster mediante ordenación de puntos.
- DENCLUE (DENsitybased CLUstEring).

Clustering basado en partición

Esta técnica construye *clusters* basándose en un proceso iterativo, este proceso clasifica *n* datos en *k* clusters mediante una función de criterio predefinida, esta función asigna el registro *i* al cluster *j* mediante optimización de resultados [17].

El *clustering* particional puede ser dividido en dos subclases: partición fuerte; donde cada objeto pertenece solo a un *cluster* y particional suave donde cada objeto puede pertenecer a más de un *cluster* [18], los algoritmos de *K-Means* y *K-Medoid* son las técnicas de *clustering* particional más utilizadas [7] las mimas que revisaremos a continuación.

K-Means

Es uno de los algoritmos más populares en el análisis y clasificación de patrones, este divide los datos en k clusters, asignando a cada uno un cluster con el centroide más cercano al mismo.

al final se evalúa el error cuadrado y se repite la operación con nuevos centroides. La cantidad de clusters, los puntos centroides iniciales y el método para medir la distancia son meta parámetros del algoritmo. Si bien es uno de los algoritmos más utilizados por su escalabilidad y complejidad lineal, presenta problemas ocasionados por datos atípicos o clusters no convexos [14].

Debido a su alta popularidad el algoritmo de K-Means he evolucionado con varias adaptaciones para aumentar su eficiencia y escalabilidad para distintas situaciones y tipos de datos, de acuerdo con [7] dentro de las principales variantes de este algoritmo tenemos:

- Fyzzy C Mean

- K- Means++
- MinMax K-Means
- K*Means
- K-Means con clasificación

K-Medoid

También denominado como partición entrono a medoides (PAM) por sus siglas en inglés.

Se denomina medoide de un conglomerado al objeto cuya diferencia con el resto de miembros del conglomerado es mínima [19].

Este algoritmo define un conglomerado de datos mediante su medoide indicando la estructura más céntrica dentro de la agrupación, el algoritmo es más eficiente ante la presencia de datos atípicos y en general fue desarrollado para mejorar las falencias de la técnica de K-Means, el algoritmo define de manera heurística o aleatoria medoides iniciales para representar a una cantidad previamente definida de conglomerados, identifica los puntos con mayor similitud con el medoide en base a la distancia con el mismo y por último se redefine el medoide que representa al conglomerado, este procedimiento se repite hasta que el algoritmo converge en un criterio previamente definido como: Un numero límite de iteraciones, la taza de similitud total o la convergencia de ubicación del medoide [7].

De acuerdo con [7] , as principales variaciones de K-Medoid son:

- Combinación de K Medoids clustering y clasificación de Naive Bayes
- Clustering de aplicación grandes CLARA por sus siglas en inglés
- Clustering de aplicación grande basado en búsqueda aleatoria CLARANS por sus siglas en inglés.

Clustering basado en malla:

Esta técnica consiste en graficar los datos como puntos en un plano multidimensional con una malla o cuadrícula trazada, los clusters son formados mediante la densidad de puntos en la cuadrícula [20]. Los pasos para aplicar esta técnica son:

1. Se divide el plano multidimensional en *k* celdas, donde *k* es un meta parámetro del algoritmo.

2. Se eliminan las celdas con menor densidad
3. Se formal los clusters uniendo celdas adyacentes con alta densidad [21] .

Esta técnica destaca por su bajo coste computacional y su precisión depende del número de celdas *k*.

De acuerdo con [7] los algoritmos de clustering basados en malla más populares son:

- Clustering basado en malla de información estadística o STING por sus siglas en inglés
- Agrupación por búsqueda o CLIQUE (clustering in quest)
- GridDBSCAN una variante que combina DBSCAN mencionando anteriormente con el clustering basado en malla

1.5.5. Minería de reglas de asociación

La minería de reglas de asociación o ARM por sus siglas en inglés, es uno de los principales enfoques de la minería descriptiva de datos, permite el descubrimiento de patrones, correlaciones, asociaciones o estructuras existentes entre datos o conjuntos de los mismos. Aporta información importante sobre bases masivas de datos, y es frecuentemente utilizada para describir datos derivados del comportamiento humano [22] .

Una regla de asociación se compone de dos partes: antecedente (LHS) y consecuente (RHS) y de manera general tiene la forma LHS → RHS donde LHS Y RHS son conjuntos disjuntos, de esta manera si LHS ocurre en un conjunto de datos, es muy probable que RHS ocurra.

La tarea de identificación de reglas de asociación depende de medidas estadísticas de precisión y confianza para garantizar su efectividad [23], las mimas que analizaremos en las secciones subsiguientes.

De acuerdo con [24] dentro de la minería exhaustiva de reglas de asociación existen dos algoritmos principales *Apriori* y *FP-growth*, sobre los cuales profundizaremos a continuación.

Algoritmo Apriori

Es el algoritmo más utilizado para encontrar patrones frecuentes dentro de una base de datos transaccional, este encuentra todos los conjuntos de datos X, Y tales que $X \rightarrow Y$ y posteriormente realizar las siguientes etapas:

Encuentra los conjuntos X con mayor frecuencia dentro de la base de datos para después depurar a los conjuntos que no satisfacen con una cantidad de apariciones mínima determinada previamente. Posteriormente los conjuntos X se unen con otros para crear una nueva lista de conjuntos de dos dimensiones donde el proceso de eliminación se repite. Esta primera etapa se repite tantas veces como dimensiones k se haya determinado para el algoritmo

Se revisa que cada subconjunto no vacío de los conjuntos $k$dimencionales X sea igual a cualquiera de los conjuntos con dimensión $k-1$, caso contrario el conjunto k dimensional X es eliminado.

Finalmente se utilizan subconjuntos de la lista depurada de conjuntos $k$dimencionales X para construir reglas de asociación. El conjunto de reglas de asociación es filtrado en base a un parámetro denominado *mínimo de confidencia* del cual se determinado para eliminar reglas débiles. La *confidencia (1)* de una regla de asociación se es determinada por el algoritmo de la siguiente manera:

$$C = \frac{S(X \cup Y)}{S(X)} \tag{1}$$

Donde S es una función que indica la frecuencia de cada conjunto en la base de datos [24]

Algoritmo FP-*growth*

Es un algoritmo que encuentra reglas de asociación sin generar listas de conjuntos. Consiste en revisar la base de datos tan solo dos veces; en la primera revisión el algoritmo encuentra elementos frecuentes, mientras que en la segunda revisión se genera una estructura denominada *árbol de patrones frecuentes* o FP-tree, esta estructura es suficiente para el análisis posterior, en cada rama de esta estructura se encuentra un conjunto candidatos para generar una regla se asociación. El principal concepto del algoritmo es que cada elemento de la base se encuentra en al menos una rama del árbol. Este método puede descubrir conjuntos candidatos para las reglas se asociación que no se encuentran en la base de datos y disminuye el coste computacional para ejecutarlo, sin embargo, aumenta el uso de memoria pudiendo crear inconvenientes sobre todo en el análisis de grandes bases de datos [24].

El algoritmo FP-*growth* es uno de los principales algoritmos no basados en Apriori, y una base para el resto de algoritmos pertenecientes a esta categoría del enfoque exhaustivo de minería de reglas de asociación [24].

1.5.6. Herramientas

Dentro de las herramientas utilizadas se contemplaron herramientas para la limpieza y transformación de datos, junto con herramientas para el modelamiento y evaluación de las técnicas de minería descriptiva de datos mencionadas anteriormente. En la Tabla 1, se muestra ventajas y desventajas de: Python, R, Weka y RapidMiner.

Tabla 1. Ventajas y desventajas de herramientas de minería, limpieza y transformación de datos

	Ventajas	**Desventajas**
Python	• Gran cantidad de bibliotecas y frameworks disponibles para diferentes tareas de minería de datos. • Es de código abierto y tiene una gran comunidad de desarrolladores. • Flexible y versátil, adecuado para tareas de minería de datos desde simples hasta complejas. • Visualización de datos versátil con librerías como Matplotlib y Bokeh [25]	• Puede requerir más líneas de código que otras herramientas para realizar tareas específicas [25]. • Menos eficiente que herramientas especializadas para tareas muy específicas • Curva de aprendizaje más empinada para no programadores [25].
R	Especializado en estadísticas y análisis de datos, con una amplia gama de paquetes específicos para diferentes técnicas. Visualización de datos robusta con paquetes como ggplot2.	• Curva de aprendizaje empinada para usuarios sin experiencia en programación o estadísticas. • Menos adecuado para aplicaciones fuera del ámbito de la estadística

RapidMiner	Interfaz gráfica intuitiva que facilita el flujo de trabajo de minería de datos para usuarios no técnicos. Amplia gama de algoritmos de aprendizaje automático disponibles. Integración con Python y R para complementar capacidades [26].	• Limitado en comparación con lenguajes de programación completos como Python o R en términos de personalización y control. • Las versiones comerciales pueden ser costosas dependiendo de las necesidades de la empresa. • La curva de aprendizaje puede ser empinada para usuarios que prefieren el código sobre las interfaces gráficas [26].
Weka	Software gratuito y de código abierto. Interfaz gráfica fácil de usar para análisis de datos y minería de datos. Amplia gama de algoritmos de aprendizaje automático disponibles [26].	• Pocas librerías especializadas y menor capacidad de control • Menos utilizado en comparación con herramientas como Python y R, lo que puede limitar la disponibilidad de recursos de aprendizaje y la comunidad de soporte [26].
Power BI	• Se integra perfectamente con otras herramientas de Microsoft como Excel, SharePoint, SQL Server, entre otras, lo que facilita el análisis de datos desde diferentes fuentes. • Visualizaciones atractivas. Ofrece una amplia gama de visualizaciones atractivas y personalizables, como gráficos, mapas, tablas pivotantes, etc., lo que facilita la comprensión de los datos.	• Puede experimentar problemas de rendimiento al trabajar con conjuntos de datos muy grandes o complejos. • Puede haber limitaciones en algunas áreas, como el diseño de informes y visualizaciones más complejas. • Power BI tiene diferentes planes de licenciamiento, y los planes más avanzados pueden ser costosos para algunas organizaciones

	• Los informes creados son interactivos, lo que permite a los usuarios explorar los datos de manera dinámica y obtener información más detallada.	

En base a lo expuesto en la Tabla 1, en este proyecto se utilizó Python para la limpieza, transformación y modelamiento de datos. Debido a la baja curva de aprendizaje que representa para mi persona, su gratuidad y la considerable cantidad de bibliotecas existentes dentro del campo de minería de datos. Las bibliotecas utilizadas se listan y explican en la siguiente *sección.* Mientras que, para la visualización de resultados mediante dashboard se aplicó *Power BI*. Además, esta herramienta es utilizada de manera activa por la FGE y, ya que el presente trabajo pretende ser un aporte para los mismos, se decidió crear visualizaciones compatibles con las ya existentes en la FGE, facilitando la integración con las herramientas disponibles en dicha institución.

Bibliotecas de Python utilizadas

Como se habia mencionado, uno de los motivos del uso de Python es la variedad de bibliotecas encontradas para minería, limpieza y transformación de datos. A continuación, se muestra la lista de bibliotecas utilizadas:

*Bokeh: E*s una biblioteca de Python especializadas en crear visualizaciones interactivas para navegadores web. Con su ayuda se puede crear visualizaciones basadas en JavaScript sin tener que escribir JavaScript [27]. En el contexto del presente trabajo esta librería fue utilizada para la exploración y limpieza de las coordenadas entregadas por la FGE.

matplotlib: Es una librería destinada a la creación de visualizaciones estáticas y animadas con Python [28]. Fue utilizada para el análisis exploratorio de los datos y detección de datos atípicos.

missingno es una librería que permite la visualización de datos nulos [29]. Esta librería fue aplicada en la detección de datos nulos dentro de la base otorgada por la FGE.

mlxtend: es una librería que implementa una variedad de algoritmos y herramientas para aprendizaje automático y minería de datos. Con el objetivo hacer accesibles las herramientas más utilizadas dentro de estos campos a los investigadores y científicos de datos [30]. Se utilizaron las herramientas de minería de reglas de asociación proporcionadas por esta librería para la aplicación de los algoritmos *Apriori* y *FP-Growth*

pandas: Es una librería especializada en el análisis y manipulación de datos. Proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento [31].Fue aplicada durante todo el desarrollo del proyecto.

Scikit-learn: proporciona una amplia gama de herramientas para transformación, preprocesamiento de datos junto con herramientas para el desarrollo y evaluación de algoritmos de aprendizaje de máquina. Dentro del proyecto esta librería fue utilizada para el preprocesamiento de datos, aplicación de modelos de clusterización y evaluación de los mismos.

1.5.7. Revisión de la literatura

De acuerdo a la literatura revisada, el uso de técnicas de minería descriptiva de datos para el descubrimiento de patrones criminales, ha sido ampliamente utilizado en todo el mundo y con distintos enfoques. Los tres enfoques principales de la minería descriptiva de datos: Sumarización, clusterización y minería de reglas de asociación, han permitido extraer información que ayude a las entidades de justicia a resolver crimines de manera eficiente, comprender el comportamiento criminal de una ciudad o región y acelerar procesos penales y de investigación.

En la presente revisión de literatura se examinaron 32 publicaciones entre artículos científicos y revisiones sistemáticas publicadas desde 2012 al 2022 sobre la aplicación de las técnicas de minería descriptiva de datos en el descubrimiento de patrones criminales. Dichos artículos fueron encontrados en: Google Scholar, Scopus, Springer, IEEE Explorer, ACM Digital Library y Elsevier (ScienceDirect). Dentro de las técnicas utilizadas destaca el uso de *clustering* y minería de reglas de asociación para identificar patrones de crímenes, obteniendo información relevante en el contexto de cada una, además, cabe recalcar que todos los artículos revisados emplean al menos una técnica de sumarización para la exploración y comprensión de los datos, reducción de dimensiones y presentación de resultados. También, se tomaron en cuenta las herramientas, descripción de los datos utilizados y técnicas de validación de resultados de cada uno de los artículos revisados. A

partir de lo mencionado anteriormente, siete artículos destacaron por su similitud con los objetivos del presente proyecto, los cuales se presentan a continuación.

Uso de clusterización

En cuanto al uso de algoritmos de clusterización podemos observar que en [32], [33] y [34], se destaca la aplicación del algoritmo k-means para distintos propósitos, en [32] los autores lo utilizan para distinguir dos categorías, acorde al número de personas involucradas en el acto criminal. En [33] se utilizó una modificación denominada k-means updated para identificar distintas categorías a partir de historiales criminales, esto con el propósito de clasificar crímenes futuros, mientras que en [34] el algoritmo fue aplicado para la delimitación de zonas geográficas destacadas dentro del territorio de Kenia. Los autores de [35] aplicaron un algoritmo de DBSCAN para identificar zonas con mayor incidencia de crimen dentro de las ciudades de Chicago y Nueva York. Este estudio presenta varias similitudes con [34], en cuanto a la identificación de zonas geográficas significativas dentro de una región determinaday la definición de manera heurística de hiperparámentros para los modelos de clusterización aplicados. A pesar del uso de modelos diferentes para el mismo objetivo, ambos estudios obtienen resultados favorables en el contexto de su investigación.

Uso de minería de reglas de asociación

Dentro de los estudios revisados [36], [32], [37], [38], y [34] se puede observar un constante uso de los algoritmos de minería de reglas de asociación como una herramienta para identificar patrones del comportamiento criminal mediante el enlace de información que a simple vista no puede ser asociada, por ejemplo en [32], los autores utilizaron el resultado del algoritmo k-means con dos categorías, como entrada del algoritmo de minería de reglas de asociación *Apriori*, obteniendo reglas de asociación entre el número de personas involucradas en actos criminales con otros atributos como: Arrestado, absuelto, inocente y convicto. Descubriendo una relación directa entre personas arrestadas y liberadas dentro del mismo año.

En [36], [37] y [34] los autores descubren mediante el uso de los algoritmos *Apriori* y *FP-growth,* patrones criminales importantes como: Crimines con mayor frecuencia, crímenes a los que las mujeres están más expuestas de acuerdo a zonas geográficas delimitadas, comportamiento criminal y detección de nuevos tipos de crímenes. A pesar de que en los trabajos mencionados no se muestra distinción entre el uso de los algoritmos *Apriori* y *FP-growth.* En el trabajo [38] al analizar información sobre Intervalos de tiempo, locación, tipo

y frecuencia de crímenes, determina que, *FP-growth* obtiene los mismos resultados que *Apriori,* pero lo supera en términos de eficiencia.

Sumarización

Como se mencionó anteriormente, este enfoque de la minería descriptiva de datos tiene por objetivo crear una versión resumida de la base de datos de la cual se puede extraer información útil directamente, o a través de un proceso complementario. Podemos observar que los métodos más populares de sumarización en la detección de patrones criminales son la *sumarización histograming* y agregados. Los autores de [32], [33], [39] utilizan este enfoque para representar gráficamente sus resultados mediante la visualización de estadísticas y patrones descubiertos en la base de datos.

Datos utilizados

En el desarrollo de los estudios revisados se utilizan datos estructurados, almacenados en distintos formatos. En esta sección se presenta una breve descripción de los datos utilizados en cada estudio con el objetivo de comprender su naturaleza. Dicha descripción se puede observar en la Tabla 2.

Tabla 2 Descripción de los datos utilizados en trabajos similares

Autor	**Descripción**
O. E. Isafiade & A. B. Bagula [36]	Se utilizaron datos de registros de crímenes en la provincia de Cabo occidental en Sudáfrica. Dichos datos muestran: Género de la víctima, tipo de crimen y nombre de la localidad donde sucedió el crimen, siendo todas las variables cualitativas.
S. Yadav, et al. [32]	Se utilizaron las variables: Número de personas arrestadas, número de crímenes cometidos, número de personas absueltas, información demográfica, estado, ciudad, año del incidente, mostrando un balance entre variables cualitativas y cuantitativas
X. A. Inbaraj, & A. S. Rao. [33]	Se utilizaron en su mayoría variables cualitativas dentro de ellas se encuentran: Tipo de delito, Ubicación (información geoespacial), Marca de tiempo (fecha y hora del delito), Características del delincuente (edad, sexo, antecedentes penales), Gravedad de los delitos etc.

Autor	Descripción
C. Catlett, & E. Cesario, et. al. [35]	Su utilizó dos bases de datos con más de 3 millones de registros en conjunto, con información geoespacial y temporal de crímenes en las ciudades de Chicago y New York. Ambas bases incluyen: Fecha y hora, tipo de crimen: categoría o clasificación del crimen cometido, coordenadas geoespaciales donde se produjo el crimen, tipo de arma utilizada, y presencia de heridos.
Z. Jian & H Jianguo, et.al. [37]	Se utilizaron dos bases de datos: Datos de crímenes en Chicago (2012 - 2017), con 365 tipos diferentes de crímenes, y 1,456,713 eventos criminales. Estos datos se pueden encontrar en: https://www.kaggle.com/currie32/crimesinchicago .Y datos de crímenes en Nueva Gales del Sur con 63 tipos diferentes de crímenes. Ambas bases poseen información de la localidad donde sucedieron los crímenes, estado del crimen, tipo y fecha en la que ocurrió.
D. Çalişkan, K Yildiz, et. al.[38]	Se utilizó la base de datos NIBRS (National IncidentBased Reporting System), dentro de la base se contempla: Información sobre el crimen, fecha y hora del incidente, ciudad, locación exacta del crimen, número de víctimas y estación de policía donde se generó el registro.
S. Wainana & J. Njuguna. Et. al. [34]	Se utilizaron registros de crímenes en 47 provincias de Kenia en los años 2012 al 2015, dichos registros contemplan información como: Tipo de crimen, estado de la investigación, y provincia donde sucedió.

Herramientas utilizadas

Durante el desarrollo de la presente revisión literaria y en los siete estudios seleccionados para la misma, se destaca el uso recurrente de: R y Python, tanto de manera individual como combinada, por ejemplo: En [35], se denota el uso de R y Python, para el desarrollo de modelos de clustering y sumarización de los resultados mediante graficas. Los mismo se puede observar en [34] donde se utiliza R como herramienta principal, junto con Python y SPSS como apoyo para visualización, desarrollo y evaluación de modelos y resultados.

2. METODOLOGÍA

En el presente proyecto se utilizará la metodología CRISP-DM ya que toma en cuenta el entorno del negocio, los resultados y describe el ciclo de vida de un proyecto de minería de datos. Cubre fases, tareas y las relaciones entre ellas, convirtiéndola en la metodología apropiada para este tipo de proyectos [40]. Tal como se muestra en la figura 7 esta metodología presenta seis etapas, que van desde la comprensión del negocio hasta el despliegue del modelo de minería de datos.

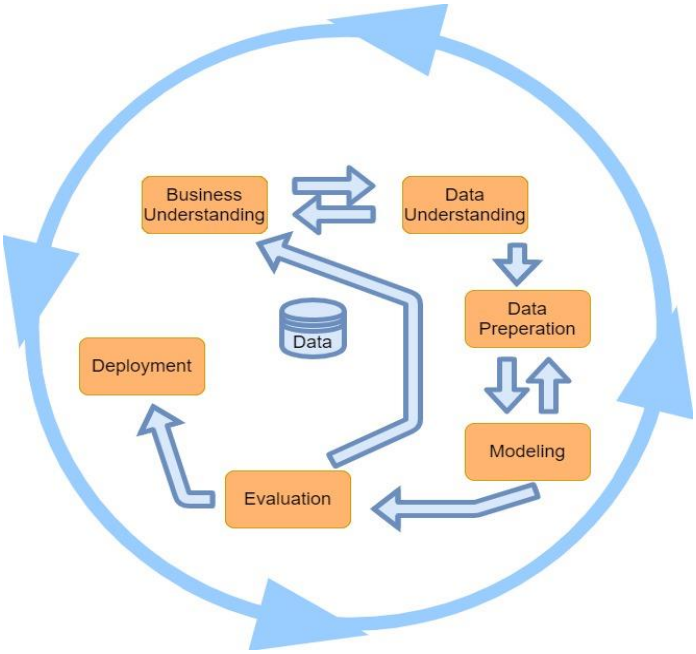


Figura 7: Metodología CRISP DM autor: Alexander Schröder [40]

A continuación, se describe cada una de las fases de la metodología elegida:

Fase 1. Comprensión del Negocio (*Business Understanding*)

En esta fase inicial, se comprende los objetivos y requisitos del negocio. Se busca identificar los problemas clave que deben abordarse y definir los criterios de éxito para el proyecto. [41]

Fase 2. *Comprensión de los Datos (Data Understanding)*

En esta etapa, se recopilan y exploran los datos disponibles para el análisis. Esto implica la identificación de fuentes de datos, la recopilación de conjuntos de datos relevantes y la evaluación de la calidad de los datos. Se busca comprender la estructura de los datos, identificar patrones iniciales y evaluar la idoneidad de los datos para abordar los objetivos del proyecto[41].

Fase 3. *Preparación de los Datos (Data Preparation)*

Una vez que se han recopilado los datos, se realiza la preparación de estos para el análisis. Esto incluye la limpieza de datos, la transformación y la integración de conjuntos de datos. La calidad y la preparación adecuada de los datos son cruciales para garantizar la precisión y la eficacia de los resultados del análisis [41].

Fase 4. *Modelado (Modeling)*

En esta fase, se seleccionan y aplican diversas técnicas de modelado para construir y entrenar modelos predictivos o descriptivos. Esto puede implicar el uso de algoritmos de aprendizaje automático, técnicas estadísticas u otros métodos de modelado. Los modelos se evalúan en función de su rendimiento y se ajustan según sea necesario [41].

Fase 5. *Evaluación (Evaluation)*

En esta etapa, se evalúan los modelos construidos en la fase anterior. Se verifica su eficacia y se comparan con los criterios de éxito definidos en la fase de comprensión del negocio. Si es necesario, se ajustan y refinan los modelos para mejorar su rendimiento [41].

Fase 6. *Despliegue (Deployment)*

Una vez que se ha seleccionado un modelo final, se implementa en el entorno operativo del negocio. Esto puede incluir la integración con sistemas existentes, la capacitación del personal y la puesta en marcha del modelo para su uso continuo [41].

Dado que la FGE no ha otorgado los permisos para la implementación de los resultados del presente proyecto, esta tesis aborda la metodología CRISP-DM hasta la fase 5.

2.1. Comprensión del negocio

La FGE, de acuerdo a sus funciones definidas en la Constitución de la República del Ecuador [42], está encargada de la investigación procesal penal de los delitos de acción pública. Esta entidad, por medio de diferentes fiscalías especializadas, recepta las denuncias colocadas por los ciudadanos para iniciar la investigación penal. El fiscal es el agente principal que orquesta la investigación a través de la consecución de diligencias que tienen la finalidad de esclarecer el fenómeno criminal investigado. En el caso de robo, las fiscalías encargadas de realizar la investigación son las fiscalías DACE (Descubrir Autores, Cómplices y Encubridores), aun cuando dependiendo de los involucrados pueden también estar las fiscalías especializadas de justicia juvenil o de crimen organizado [2].

2.2. Comprensión de los datos

Dentro de esta fase se realizó un análisis exploratorio de los datos con el fin de comprender cada variable, así como identificar posibles problemas dentro de las mismas tales como: Datos nulos, atípicos o formatos que impidan su análisis y posterior uso en los modelos propuestos.

Los datos otorgados por la FGE en un archivo CSV con 663646 registros y 17 variables, corresponden al registro de delitos de robo a partir del 1 enero del 2015 al 31 de diciembre del 2022. En la base de datos se registra diversa información sobre la ejecución del robo y su estado en los procesos jurídico penales.

Para facilitar la comprensión de los datos se han agrupado las variables en tres categorías:

- *Temporales:* Se refiere a las variables que registran información sobre el tiempo y sus características.
- *Espaciales*: Se refiere a variables con información sobre espacio en el que se registró el delito, se toma en cuenta coordenadas, cantón, provincia y ruralidad de la zona.
- *Descriptivas*: Se refiere a variables que indican las circunstancias, clasificación, registro y detalles del delito como: Tipo de arma, crimen circunstancial y número de registro.
- En la Tabla 3 se muestra el tipo de dato de cada variable junto con el grupo al que pertenece de acuerdo con el párrafo anterior.

Tabla 3 Variables otorgadas por la FGE, tipo de dato y clasificación

Variable	Tipo de dato	Clasificación
NDD_OFUSCADA	Cuantitativo	**Descriptivas**
DELITO_CIRCUNSTANCIAL	Cualitativo nominal	
TIPO_ARMA	Cualitativo nominal	
MODALIDAD_DESAGREGACION	Cualitativo nominal	
modalidad_desagregacion_comision	Cualitativo nominal	
TIPO_FLAGRANTE	Cualitativo nominal	
ETAPA_ACTUAL	Cualitativo nominal	
DELITO	Cualitativo nominal	
TIPO_DELITO	Cualitativo nominal	
FECHA_INCIDENTE	Cuantitativo discreto	**Temporales**
HORA_INCIDENTE	Cuantitativo continuo	
GRUPO_HORAINC	Cualitativo ordinal	
DIA_INCIDENTE	Cualitativo ordinal	
PROVINCIA_INCIDENTE	Cualitativo nominal	**Espaciales**
CANTON_INCIDENTE	Cualitativo nominal	
COORDENADAS_INCIDENTE	Tupla	
URB_RURAL_INCIDENTE	Cualitativo nominal	

En base a la información de la Tabla 3, la base de datos presenta 12 variables cualitativas y 4 variables cuantitativas. A continuación, describimos cada una de las variables presentadas en la Tabla 3:

NDD_OFUSCADA: Código designado al archivo de reporte del crimen a manera de ID. Con el objetivo de proteger datos delicados la FGE este campo está codificado.

DELITO_CIRCUNSTANCIAL: Especifica la acción delictiva cometida sin premeditación ni planificación sucedida durante la ejecución del delito principal (robo).

TIPO_ARMA: Muestra el objeto con el que se llevó a cabo el robo registrado mediante 10 categorías

MODALIDAD_DESAGREGACION: Indica la modalidad de robo, presenta un total de 29 categorías que dan información de la naturaleza y como se ejecutó el delito.

TIPO_FLAGRANTE: Indica si el robo fue flagrante o no.

ETAPA_ACTUAL: Con cinco categorías, da información sobre la etapa del proceso judicial del crimen registrado hasta la fecha de corte para generar la presente base.

DELITO: Indica la clasificación del delito de acuerdo con la clasificación presentada en el código integral penal del Ecuador (COIP), al ser una base de datos sobre registros de robos esta variable presenta un solo campo (ROBO) para todos los registros.

TIPO_DELITO: Con seis categorías diferentes, otorga información sobre la consumación del delito,

FECHA_INCIDENTE: Muestra el día mes y año en los que se ejecutó el robo. Los registros de esta variable siguen el siguiente formato DD-MM-AAAA.

HORA_INCIDENTE: Con el formato 24h esta variable indica la hora en la que se llevo a cabo el delito con una precisión de minutos.

GRUPO_HORAINC: Con cuatro categorías muestra la parte del día (MADRUGADA, MAÑANA, TARDE, NOCHE) en la que se realizó el robo.

DIA_INCIDENTE: Muestra el día de la semana donde se propició el delito.

PROVINCIA_INCIDENTE: Muestra la provincia donde se suscitó el delito, la variable presenta las 24 provincias del Ecuador

CANTON_INCIDENTE: Indica el cantón dende se ejecutó el robo, presenta 220 de 221 cantones distribuidos a lo largo del país

COORDENADAS_INCIDENTE: Par ordenado que presenta las coordenadas del delito en formato de Grados decimales (DD)

URB_RURAL_INCIDENTE: Con dos categorías muestra la ruralidad de la zona donde se ejecutó el delito.

De acuerdo a la descripción de las variables y la información presentada en la Tabla 3 analizamos cada una de ellas con enfoque en el tipo de datos que las conforman.

2.2.1. Variables cualitativas

Como información complementaria a la Tabla 3 y bajo la descripción de las variables presentada anteriormente, en la Tabla 4 se muestra la cantidad de categorías que posee cada variable cualitativa, su moda y valores nulos.

Tabla 4. Cantidad de categorías, moda, frecuencia modal y número de datos nulos dentro de cada variable cualitativa

Variable	Número de categorías	Moda	Frecuencia modal	Nulos
CANTON_INCIDENTE	220	GUAYAQUIL	160158	0
DELITO	1	ROBO	663646	0
DELITO_CIRCUNSTANCIAL	15	ROBO	369537	0
DIA_INCIDENTE	7	VIERNES	105703	1
ETAPA_ACTUAL	5	INVESTIGACION PREVIA	615025	0
GRUPO_HORAINC	4	NOCHE	192494	0
MODALIDAD_DESAGREGACION	29	ASALTO	312955	0
modalidad_desagregacion_comision	28	ASALTO	302134	21800
PROVINCIA_INCIDENTE	24	GUAYAS	217243	0
TIPO_ARMA	9	DESCONOCE	225539	85330
TIPO_DELITO	2	CONSUMADO	647279	0
TIPO_FLAGRANTE	2	NO FLAGRANTE	605018	0
URB_RURAL_INCIDENTE	2	URBANO	580856	41

A partir de la Tabla 4 y conociendo las descripciones de cada variable cualitativa las variables: *CANTON_INCIDENTE, DELITO, DIA_INCIDENTE, ETAPA_ACTUAL,*

GRUPO_HORAINC, PROVINCIA_INCIDENTE y URB_RURAL_INCIDENTE no presentan una cantidad de categorías incongruente con su definición, mientras que el resto de variables debe pasar por un análisis más profundo para determinar las categorías definitivas de cada una de ellas. Además, se observa que las variables: *delitos_seguimiento_comision, modalidad_desagregacion_comision, TIPO_ARMA,* muestran una cantidad significativa de datos nulos, por lo que en las secciones subsiguientes se deberá realizar un análisis complementario para su imputación o eliminación.

2.2.2. Variables cuantitativas

Para comprender el comportamiento de las variables cuantitativas analizaremos las principales medidas de estadística descriptiva (media, mediana, desviación estándar o std, rango Inter cuartil o RIC, máximo y mínimo) junto con la cantidad de datos nulos, dentro de cada una de las variables. Tenemos que hacer una distinción sobre la variable *NDD_OFUSCADA,* debido a que esta representa los identificadores de cada registro, consecuentemente solo podemos revisar la existencia de datos nulos. En cuanto a la variable *COORDENADAS_INCIDENTE*, debido a su formato de tupla, se debe hacer una transformación sobre la misma, separándola en dos nuevas variables *LATITUD* y *LONGITUD*. En la Tabla 5 podemos observar las medidas obtenidas de cada variable cuantitativa.

Tabla 5. Medidas de estadística descriptiva y cantidad de datos nulos de las variables cuantitativas

Variable	**Nulos**	**Media**	**Mediana**	**std**	**RIC**	**Min**	**Max**
FECHA_INCIDENTE	0	11/13/2018	10/9/2018	880 días 05:58 horas	1584 días	1/1/2015	12/31/2022
HORA_INCIDENTE	0	12:59:59 PM	1:25:00 PM	06:27:07 horas	10:45:00 horas	12:00:00 AM	11:59:21 PM
ARTICULO	0	189	189	0	0	189	189
LATITUD	1112	-7.05	-2.08	19.95	1.93	-81.55	0

Variable	Nulos	Media	Mediana	std	RIC	Min	Max
LONGITUD	1112	-78.58	-79.48	7.54	1.39	-90.97	1.01
NDD_OFUSCADA	0	-	-	-	-	-	-

De acuerdo con la Tabla 5 podemos obtener la siguiente información sobre las variables:

- *FECHA_INCIDENTE* y *HORA_INCIDENTE,* presentan un std y RIC significativos, lo que indica la existencia de datos atípicos.
- *ARTICULO* presenta un único valor, lo que indica que la información es congruente con su definición y dicha variable no aporta datos relevantes para el estudio
- *LATITUD* y *LONGITUD* presentan la misma cantidad de datos nulos, los mismos que deberán pasar por un análisis posterior, para su correspondiente imputación o eliminación.
- *LATITUD* muestra un valor máximo de 0, esto indica que no hay registros de robos en latitudes superiores a la línea ecuatorial. Esta información será corroborada en la etapa de preparación de los datos.
- *NDD_OFUSCADA* no presenta datos nulos.

2.3. Preparación de los datos

Una vez comprendidos los datos otorgados por la FGE procedemos a definir un subconjunto de la base de datos original con el objetivo de utilizarlo en la etapa de *modelamiento.* Este proceso se ha dividido en las fases de: Selección de variables, limpieza y transformación de los datos. En cada una de ellas se realizaron distintos enfoques dependiendo de la naturaleza de la variable y la información que se desea obtener de la misma.

2.3.1. Selección de variables

De acuerdo con la información expuesta en la sección anterior junto con lo mostrado en la Tabla 4 y Tabla 5, se ha decido no seleccionar las siguientes variables:

- *DELITO*, la variable presenta una sola categoría ROBO, la cual no aporta información relevante para este estudio ya que la base pertenece solo a registros de robo.

- *modalidad_desagregacion_comision*, esta variable presenta las mimas categorías que la variable *MODALIDAD_DESAGREGACION*, junto con una adicional denominada nan, además, a diferencia de *MODALIDAD_DESAGREGACION* que no alberga datos nulos, esta variable presenta un total de 21800 datos nulos Tabla 4. Por lo que se ha decidido no incluirla en las etapas posteriores.
- *ARTICULO*, la variable muestra un único dato que corresponde al artículo dentro del COIP que tipifica el delito, al ser una base exclusiva de registros de robo, esta variable no aporta información al estudio.
- *COORDENADAS_INCIDENTE*, debido a su formato de tupla, la variable complica su análisis y uso en los modelos posteriores, por lo que en su lugar se utilizaran las dos variables obtenidas a partir de esta *LATITUD* y *LONGITUD.*

Teniendo como resultado un conjunto de variables que aportan información relevante para el estudio, este conjunto de ahora en adelante se lo denominará como base de datos, en la Tabla 6 podemos observar las variables que la conforman.

Tabla 6. Conjunto de variables seleccionadas para su uso en etapas posteriores

Variable	Tipo de dato	Clasificación
NDD_OFUSCADA	Cuantitativo	**Descriptivas**
DELITO_CIRCUNSTANCIAL	Cualitativo nominal	
TIPO_ARMA	Cualitativo nominal	
MODALIDAD_DESAGREGACION	Cualitativo nominal	
TIPO_FLAGRANTE	Cualitativo nominal	
ETAPA_ACTUAL	Cualitativo nominal	
TIPO_DELITO	Cualitativo nominal	
FECHA_INCIDENTE	Cuantitativo discreto	**Temporales**
HORA_INCIDENTE	Cuantitativo continuo	
GRUPO_HORAINC	Cualitativo ordinal	
DIA_INCIDENTE	Cualitativo ordinal	
PROVINCIA_INCIDENTE	Cualitativo nominal	**Espaciales**
CANTON_INCIDENTE	Cualitativo nominal	
LATITUD	Cuantitativo discreto	
LONGITUD	Cuantitativo discreto	
URB_RURAL_INCIDENTE	Cualitativo nominal	

2.3.2. Transformación de datos

Procedemos a analizar cada una de las variables y realizar las transformaciones necesarias con el objetivo de obtener un conjunto de datos utilizable en la etapa de modelamiento, este proceso se muestra a continuación:

Dentro de las variables: *DELITO_CIRCUNSTANCIAL* y *MODALIDAD_DESAGREGACION,* se observaron varias categorías con el mismo significado pero que habían sido ingresadas con caracteres extra como tildes, espacios o

el uso de letras mayúsculas y minúsculas, esto hace que *pandas* las detecte como categorías distintas a pesar de tener la misma escritura y significado, además también se encontraron categorías sinónimas entre sí, la existencia de este tipo de categorías puede generar ruido dentro de la base de datos, además de afectar la fiabilidad de los modelos a utilizar. Para resolver este problema se analizó el contexto de cada una de las categorías y se las agrupo en una sola que abarque la misma información de las categorías analizadas. En las tablas 7 y 8 se muestran las categorías agrupadas de las variables *MODALIDAD_DESAGREGACION* y *DELITO_CIRCUNSTANCIAL* correspondientemente.

Tabla 7. Agrupación de categorías en la variable *MODALIDAD_DESAGREGACION*

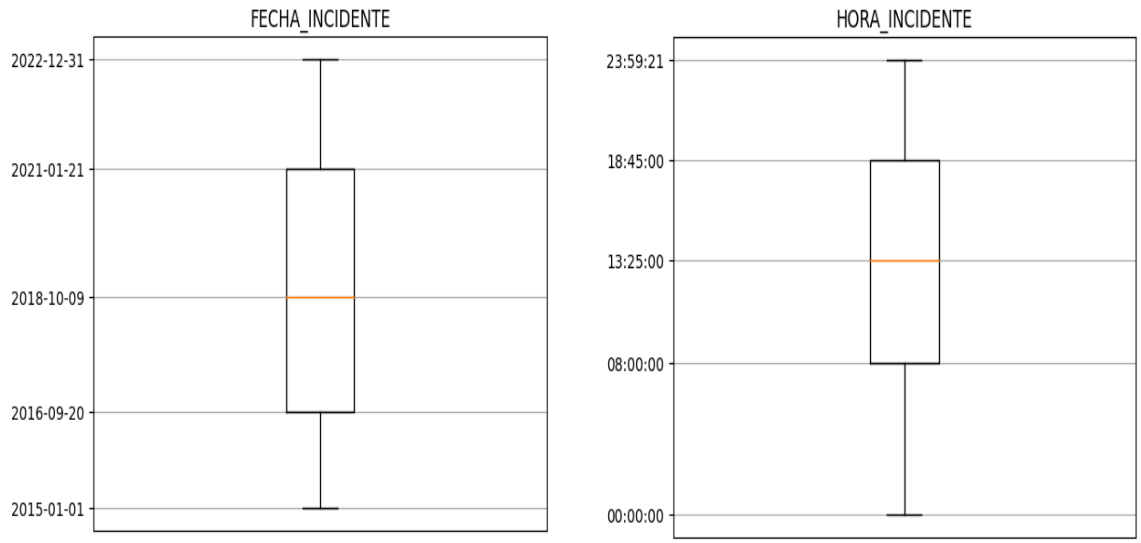
Variable: *MODALIDAD_DESAGREGACION*	
Categorías	**Categoría agrupada**
ENGAÑO POR FALSOS EMPLEADOS O FUNCIONARIOS\r\n\r\n *FALSOS FUNCIONARIOS PUBLICOS* *FALSAS EMPLEDAS DOMESTICAS* *FALSOS AGENTES CTE* *FALSOS AGENTES FF AA* *ENGAÑO POR FALSOS EMPLEADOS O FUNCIONARIOS* *ENGAÑO POR FALSOS EMPLEADOS, USUARIOS O FUNCIONARIOS* *ENGAÑO POR FALSOS FUNCIONARIOS PUBLICOS*	ENGAÑO POR FALSOS EMPLEADOS O FUNCIONARIOS
HORAMEN *FORAMEN*	FORAMEN
ESTRUCHANTE *ESTRUCHE*	ESTRUCHE

Tabla 8. Agrupación de categorías en la variable *DELITO_CIRCUNSTANCIAL*

Variable: *DELITO_CIRCUNSTANCIAL*	
Categorías	**Categoría agrupada**
ROBO. SI A CONSECUENCIA DEL ROBO SE OCASIONA LA MUERTE. *RoBO SI A CONSECUENCIA DEL ROBO SE OCASIONA LA MUERTE*	ROBO SI A CONSECUENCIA DEL ROBO SE OCASIONA LA MUERTE
ROBO. SI EL ROBO SE PRODUCE ÚNICAMENTE CON FUERZA EN LAS COSAS. *ROBO CUANDO EL ROBO SE PRODUCE ÚNICAMENTE CON FUERZA EN LAS COSAS*	ROBO CUANDO EL ROBO SE PRODUCE ÚNICAMENTE CON FUERZA EN LAS COSAS

En cuanto a las variables *MODALIDAD_DESAGREGACION, TIPO_ARMA* y *ETAPA_ACTUAL* se identificaron las categorías *NO APLICA* y *SIN INFORMACION*, dichas categorías no aportan con información a la base, por lo que fueron reemplazadas por datos nulos.

Para las variables FECHA_*INCIDENTE* y *HORA_INCIDENTE* al ser variables cuantitativas y poseer std y RIC significativos como se muestra en la Tabla 5, se realizó el diagrama de caja y bigote de ambas variables para visualizar la existencia de datos atípicos en cada una de ellas. Ver Figura 8. a) y b)



a) **b)**

Figura 8. Diagramas de caja y bigote de las variables FECHA_INCIDENTE (Figura 8. a)) y HORA_INCIDENTE (Figura 8. b))

Como se puede observar en la figura 8. a) y b) las variables *FECHA_INCIDENTE y HORA_INCIDENTE* no albergan datos atípicos, por lo tanto, no requieren de ninguna transformación o limpieza.

2.3.3. Limpieza y transformación de coordenadas

Para la limpieza y transformación de las variables *LATITUD y LONGITUD* se debe tener en cuenta el significado de dichas variables, estas deben indicar correctamente la posición geográfica del delito de robo y deben ser congruentes con la información expuesta en las variables *PROVINCIA_INCIDENTE* y *CANTON_INCIDENTE*, por lo tanto, si no existe congruencia entre la información de provincia, cantón y coordenadas del registro se considerará que las coordenadas han sido ingresadas de manera errónea. Además de acuerdo a los valores de máximo, mínimo, std y RIC mostrados en la Tabla 5, junto con los diagramas de caja y bigote mostrados en la Figura 9, se puede sospechar la existencia de registros de coordenadas fuera del territorio nacional.

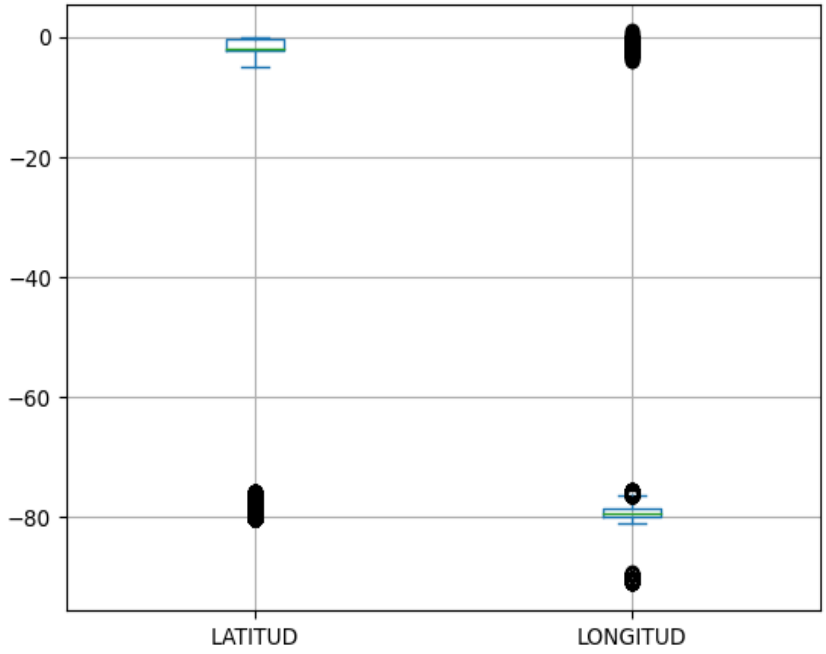


Figura 9. Diagramas de caja y bigote de las variables LATITUD y LONGITUD

Para realizar este análisis y detectar los registros de coordenadas erróneos se han planteados dos etapas:

- **Detección de coordenadas fuera del territorio nacional**

Visualizamos los puntos de coordenadas en el mapa para corroborar la existencia de registros fuera del territorio nacional. Para realizar esta visualización nos apoyamos de la librería *bokeh* de Python y transformamos las coordenadas a formato Mercator.

Como se puede observar en la Figura 10, existen varios registros de coordenadas fuera del territorio nacional y totalmente fuera del contexto de la base de datos, estos registros se concentran en localidades de la Antártica y el océano Pacifico cerca del continente africano, lo que significa que existe un problema al momento de ingresar la información geográfica de los registros, además de impedir su utilización en las siguientes fases.

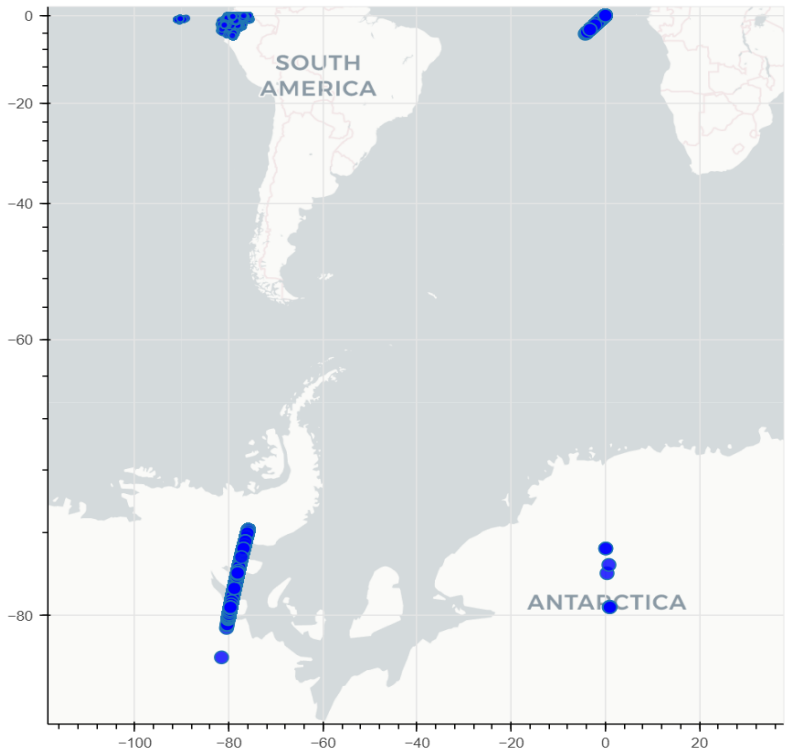


Figura 10. Visualización de coordenadas en el mapa

Por ello, con el objetivo de filtrar los registros fuera del territorio nacional, las coordenadas fueron normalizadas y se examinaron sus valores extremos descritos en la Tabla 9.

Tabla 9. Mínimo y máximo de las variables LATITUD y LONGITUD normalizadas

Variable	min	max
LATITUD	-3.743204	0.352733
LONGITUD	-1.645872	10.569153

Después, a partir del conjunto de coordenadas normalizadas se aislaron los registros mayores a 3 o menores a -3, creando así un conjunto de registros con coordenadas fuera del territorio nacional. Para determinar la efectividad del proceso descrito se visualizaron las coordenadas en el mapa, distinguiendo los datos fuera del territorio (color rojo) de que se encuentran dentro (color azul) Figura 11.

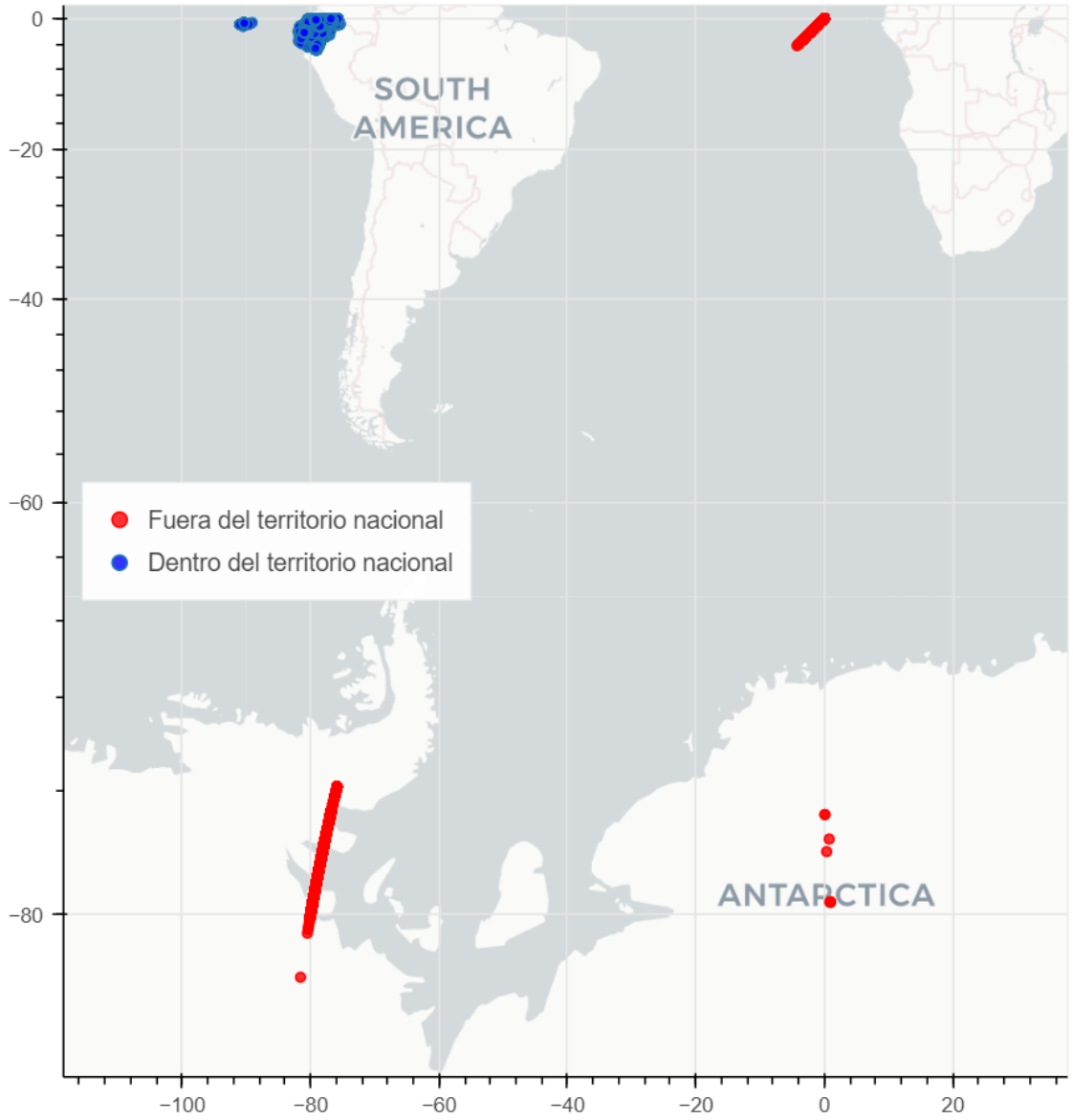


Figura 11. Registros de coordenadas dentro y fuera del territorio nacional

Como se muestra en la Figura 11 el análisis anterior ha sido efectivo y se ha logrado aislar los registros de coordenadas dentro y fuera del territorio nacional. Obteniendo un total de 51752 registros que figuran fuera del territorio nacional y un total de 591860 coordenadas dentro del territorio nacional.

Cabe resaltar que después de eliminar los registros con coordenadas fuera del territorio nacional 20 de las 24 provincias perdieron un media del 1,5 % de sus registros, mientras que los registros de las 4 provincias restantes disminuyeron del 75% al 100%, de acuerdo a la información mostrada en la Tabla 10, las provincias de Carchi, Imbabura, Esmeraldas y Sucumbíos albergaron un total de 39501 registros con coordenadas erróneas, además la provincia de Carchi no posee ningún registro de coordenadas correcto, seguido de las

provincias de Imbabura y Esmeraldas que solo muestran 1 y 4 registros correctos correspondientemente.

Tabla 10. Conteo antes y después, junto con reducción y porcentaje de perdida de las provincias con mayor pérdida de registros.

Provincia	**Conteo Antes**	**Conteo Después**	**Reducción**	**% Eliminado**
Carchi	2133	0	2133	100
Imbabura	10552	1	10551	99.99
Esmeraldas	20370	4	20366	99.98
Sucumbíos	8534	2083	6451	75.59
Total			39501	

- **Detección de coordenadas no correspondientes a la provincia del registro**

Una vez aisladas las coordenadas que figuran dentro del territorio nacional, debemos verificar que la provincia del registro coincida con la ubicación que proporcionan las coordenadas. Para realizar este análisis nos apoyamos de la librería de Python *geopandas,* la cual está dedicada al análisis de datos geográficos y a la base de polígonos de cada provincia del Ecuador proporcionada por [43].

Para determinar si las coordenadas están en la provincia que indica el registro, se verifica si se encuentran dentro del polígono que dibuja dicha provincia en el mapa, esto se puede hacer de manera visual como se muestra en la Figura 12, donde se grafica las coordenadas de los registros que pertenecen a la provincia de Bolívar y se puede observar que dos no se encuentran dentro del polígono que dibuja la provincia.

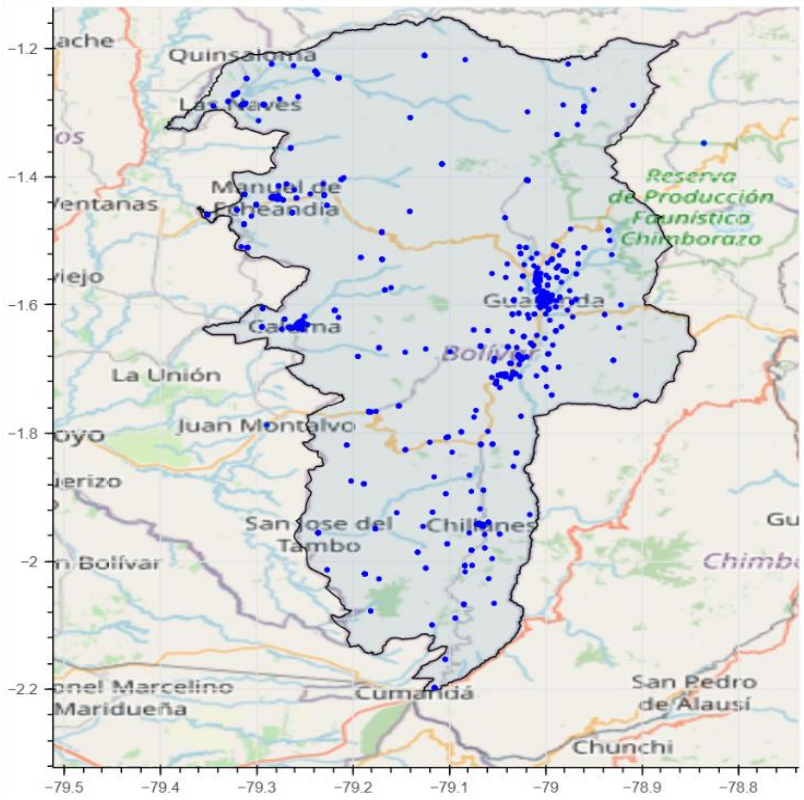


Figura 12. Visualización de coordenadas de los registros que figuran en la provincia de Bolívar.

Para identificar estos registros de manera automática se utilizó la librería *geopandas* para transformar las variables *LATITUD* y *LONGITUD* a puntos geográficos y con la ayuda del método *within* verificamos si pertenece al polígono de la provincia que se está analizando. De esta manera podemos identificar los registros y diferenciarlos para así quedarnos solo con aquellos donde los valores de las coordenadas son congruentes con la provincia registrada. En la Figura 13 se visualizan los resultados de este proceso para la provincia de Bolívar.

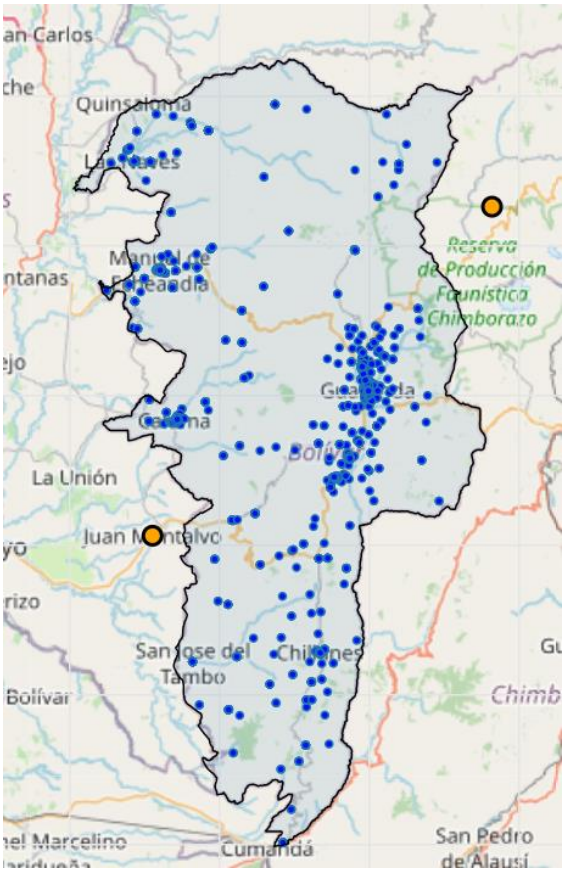


Figura 13. Visualización de coordenadas dentro y fuera de la provincia de Bolívar

El mismo análisis se realizó para cada una de las provincias del Ecuador obteniendo una nueva base de datos donde todas las coordenadas son congruentes con las provincias de su registro, esta nueva base consta de un total de 589771 registros. Cabe recalcar que después de este proceso las provincias de Esmeraldas e Imbabura han perdido el 100% de sus registros, es decir qué; en toda la base de datos proporcionada por la FGE, ningún registro de coordenadas correspondiente a robos en Esmeraldas o Imbabura ha sido ingresado correctamente.

2.3.4. Limpieza de datos nulos

Una vez realizados los procesos de transformación y limpieza de los datos, identificamos los valores nulos dentro de la base, para su posterior eliminación o imputación, de acuerdo a la relevancia de la variable y cantidad de datos nulos detectados en la misma. Con ayuda de la librería *missingno* de Python, junto con los métodos *isna* y *sum* de la librería pandas se ha identificado que las únicas variables con datos nulos son: URB_RURAL_INCIDENTE y TIPO_ARMA con 70380 y 38 datos nulos correspondientemente.

Para la variable *TIPO_ARMA* se reemplazaron los datos nulos por la categoría *DESCONOCE,* mientras que para la variable *URB_RURAL_INCIDENTE,* los datos nulos fueron reemplazados por la moda de la variable *URBANO.*

2.3.5. Resultados

Como resultado final de esta etapa se ha obtenido una base de datos con variables cualitativas y cuantitativas que albergan datos relevantes y congruentes entre sí. La mencionada base de datos contiene un total de 16 variables y 589771 registros. Lo que en comparación con la base de datos otorgada por la FGE representa una reducción del 11% de los registros. Además, no existen datos geográficos fiables de las provincias de: Carchi, Esmeraldas e Imbabura, debido a que las coordenadas registradas no son congruentes con la provincia y cantón del registro, por lo tanto, con el objetivo de no afectar la fiabilidad de los modelos propuestos en la etapa siguiente, la información correspondiente a estas provincias queda excluida del estudio.

2.4. Modelado

En la presente etapa se implementan modelos de minería descriptiva de datos, con el objetivo de descubrir noticias del delito relacionadas en los datos generados por la FGE. Además de Implementar un dashboard para la visualización y análisis de resultados, junto con información de la base de datos procesada.

Los modelos de minería descriptiva de datos utilizados, han sido seleccionados a partir de la literatura revisada y evaluados bajo el contexto de los datos estudiados.

Como se observa en los estudios [34] y [35], determinamos zonas geográficas importantes aplicando un algoritmo de clustering sobre los registros de cada provincia. Una vez identificadas estas zonas, minamos reglas de asociación dentro de cada una de ellas, para de esta manera buscar patrones de delitos de robo sobre el conjunto de reglas de asociación encontrado. Con los patrones de robo identificados, se determinaron los conjuntos de noticias de delito que siguen dichos patrones, mostrando una relación entre sí. En las siguientes secciones se muestra la ejecución de cada una de las fases mencionadas anteriormente.

2.4.1. Agrupación de robos por zonas geográficas en cada provincia

Aplicamos un algoritmo de clusterización sobre las variables de LATITUD y LONGITUD, con el objetivo de identificar las principales zonas geográficas de crímenes dentro de cada provincia. A continuación, se detalla la elección del método de clusterización y el proceso de búsqueda y definición de las zonas.

De acuerdo con las propuestas de [7], [34], [35], [44], los algoritmos de clustering más utilizados y con mejores resultados dentro de la agrupación de crímenes como robo y asesinato, mediante el análisis de sus coordenadas, son DBSCAN y K-means. La elección del método utilizado se fundamenta en la Tabla 11, donde se muestran las ventajas y desventajas de cada algoritmo, junto con los resultados obtenidos y analizados al aplicar dichos algoritmos en 3 provincias.

Tabla 11. Ventajas y desventajas de DBSCAN y K-means

Tool	**Ventajas**	**Desventajas**
DBSCAN	• Puede detectar cualquier tipo de clúster sin importar su forma • No es sensible a ruido o datos atípicos • Determina el número de clusters de manera automática	• Muy sensible al hiper parámetro épsilon • Alto coste computacional para bases de datos grandes y/o multidimensionales • Su rendimiento se ve afectado si los datos tienen varias densidades • Difícil de escalar
K-Means	• Es efectivo y escalable • Eficiente frente a bases de datos grandes y/o multidimensionales • Es fácil de implementar e interpretar • Es uno de los algoritmos de clusterización más aplicados y estudiados, por lo tanto, tiene una amplia gama de herramientas para solventar posibles problemas o desventajas al momento de aplicarlo.	• Sensible a datos atípicos • Se limita a determinar clusters con forma esférica • No determina el número de clusters de manera automática

Las provincias seleccionadas para el estudio comparativo entre los dos algoritmos mencionados fueron: Pastaza, Loja y Azuay. Dichas provincias representan los cuartiles 1, 2 y 3 respectivamente, sobre el conteo de registros de crímenes por provincia. Dentro de esta comparación se analizó:

- Número de clusters obtenidos y cantidad de registros en cada uno
- Visualización de los clusters obtenidos en cada método
- Índice de silueta e índice de Davies–Bouldin

Para el desarrollo de los modelos se utilizaron los algoritmos disponibles en la librería *scikit-learn* de Python, mientras que la definición de los parámetros de cada realizo de acuerdo a la información expuesta en la Tabla 12.

Tabla 12 Elección de parámetros para los modelos K- means y DBSCAN

Tool	Parámetro	Método de elección
DBSCAN	Épsilon (eps)	Utilizamos el método *del codo* descrito en [45], [46]
	Mínimo de vecinos (min_samples)	Al tener 2 dimensiones utilizamos el criterio de [47] que indica un valor de 4 para este parámetro.
K-Means	Número de clusters (n_clusters)	Se utilizó el método *del codo* descrito en [48] para identificar el número óptimo de clusters.
	Centroides iniciales (init)	Utilizamos el método Kmeans ++ que garantiza una rápida convergencia del algoritmo [49].

Una vez aplicados los algoritmos de K-means y DBSCAN con los parámetros indicados en la Tabla 12, sobre las provincias de Azuay, Loja y Pastaza se obtuvieron los siguientes resultados:

- Azuay

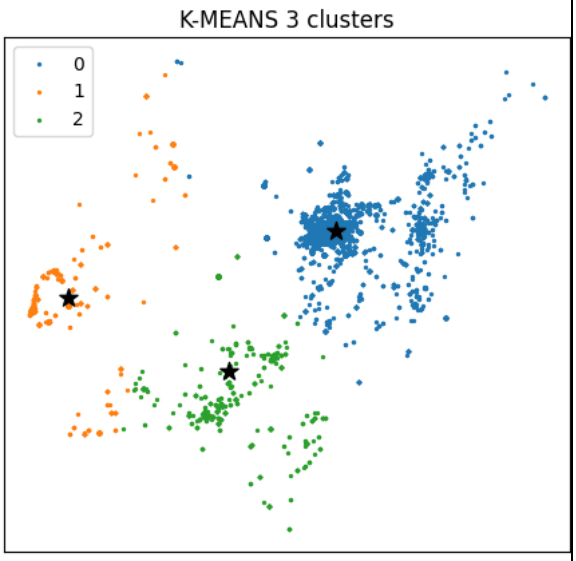
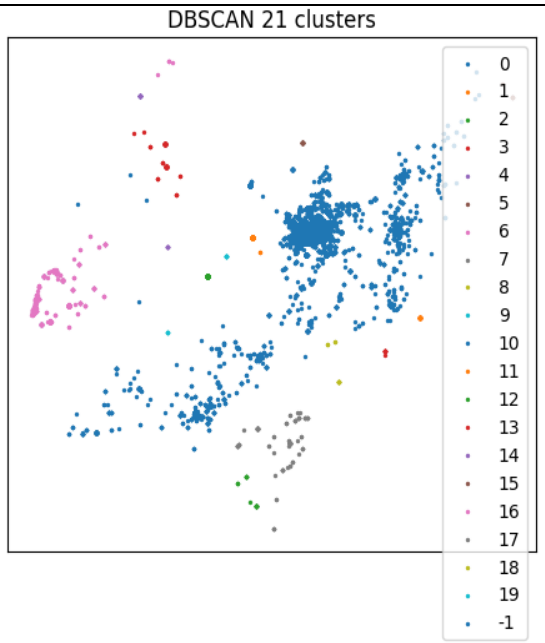
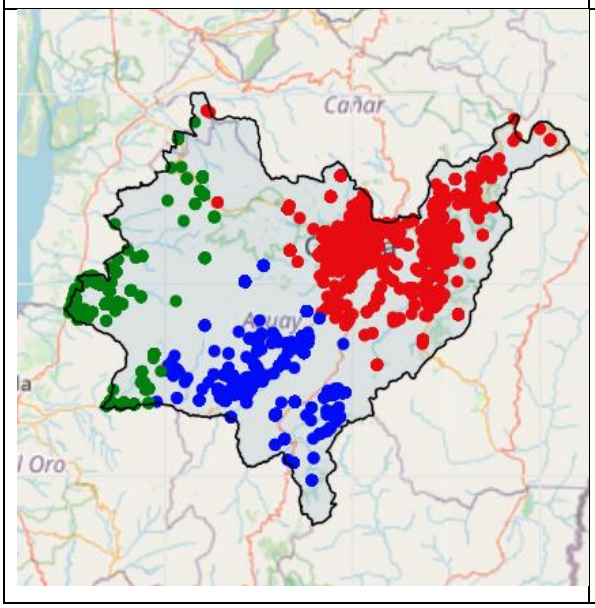
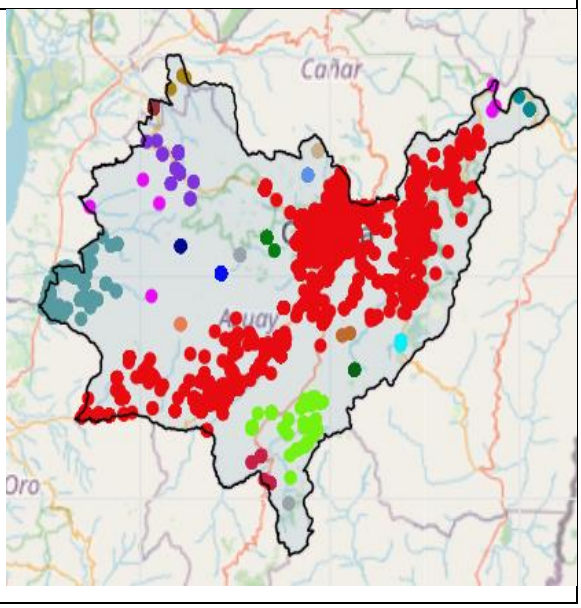
a) Clusters en el plano obtenidos con K-means	**b)** Clusters en el plano obtenidos con DBSCAN
	
c) Clusters en el mapa de la provincia de Azuay obtenidos con K-means	**d)** Clusters en el mapa de la provincia de Azuay obtenidos con DBSCAN
	

Figura 14. Visualización de los clusters obtenidos a partir de K-means (a) y c)) y DBSCAN (b) y d)), para la provincia de Azuay

- Loja:

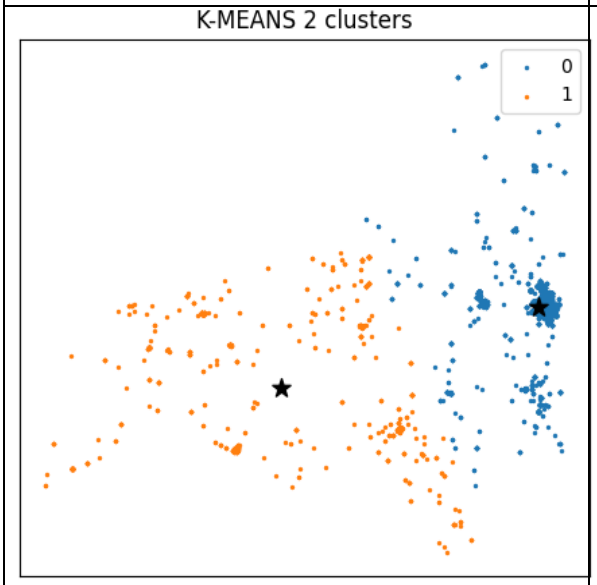
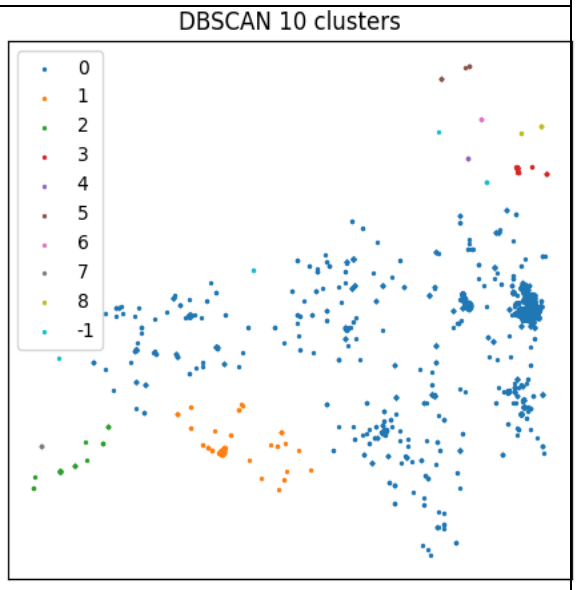
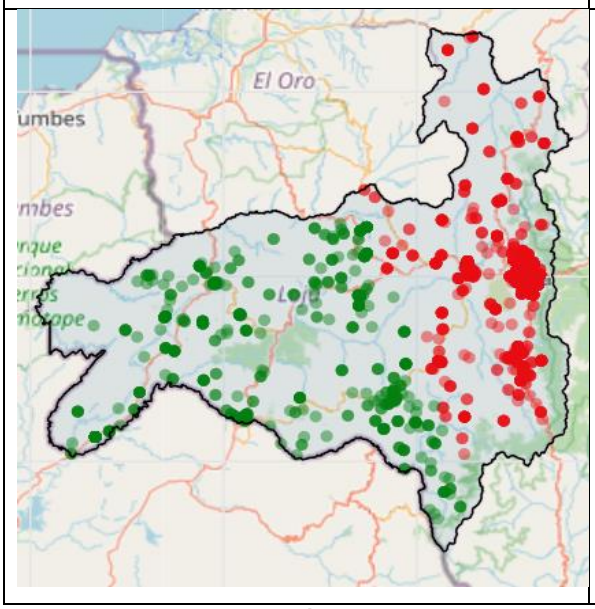
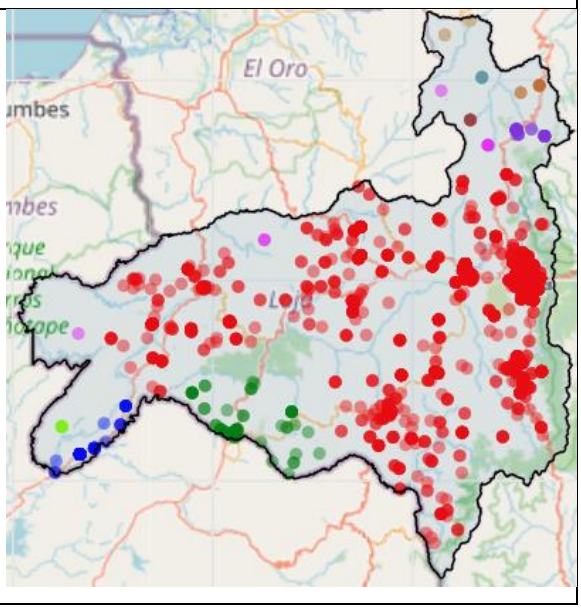
a) Clusters en el plano obtenidos con K-means	**b)** Clusters en el plano obtenidos con DBSCAN
	
c) Clusters en el mapa de la provincia de Loja obtenidos con K-means	**d)** Clusters en el mapa de la provincia de Loja obtenidos con DBSCAN
	

Figura 15 Visualización de los clusters obtenidos a partir de K-means (a) y c)) y DBSCAN(b) y d)), para la provincia de Loja

- Pastaza

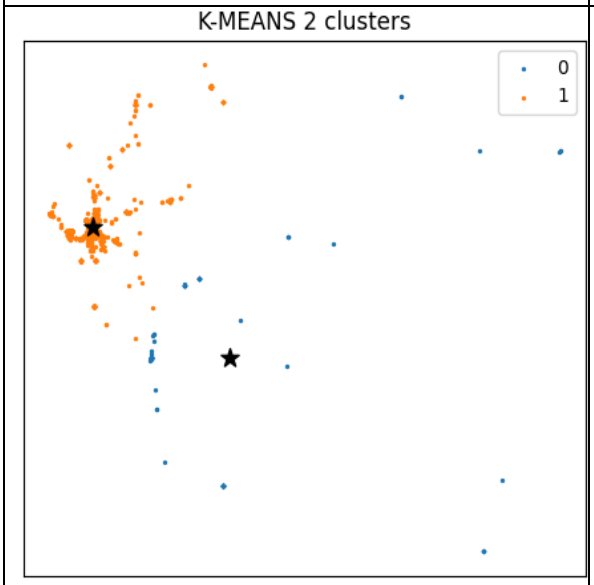
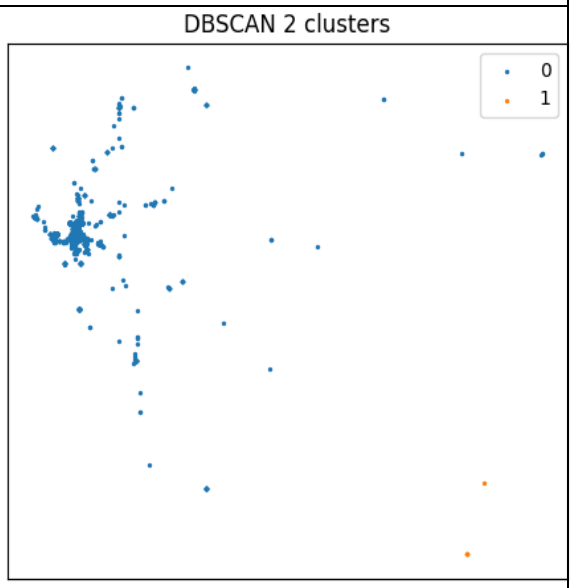
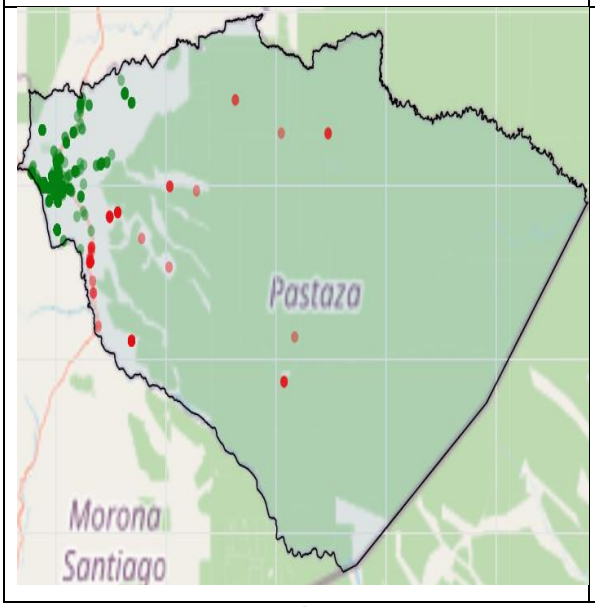
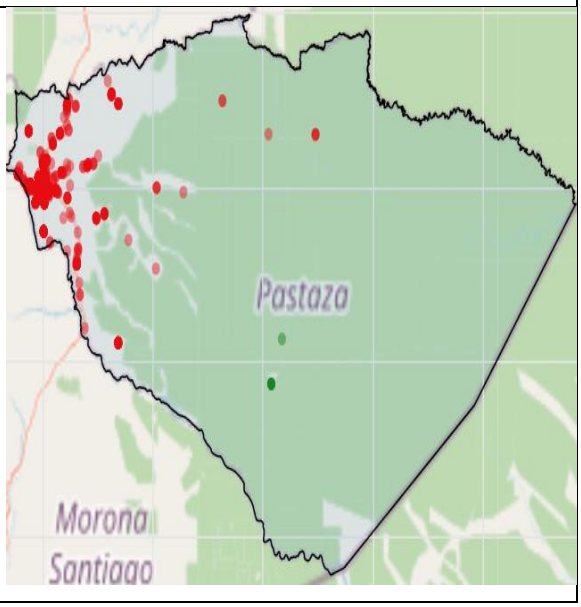
a) Clusters en el plano obtenidos con K-means	**b)** Clusters en el plano obtenidos con DBSCAN
	
c) Clusters en el mapa de la provincia de Pastaza obtenidos con K-means	**d)** Clusters en el mapa de la provincia de Pastaza obtenidos con DBSCAN
	

Figura 16.. Visualización de los clusters obtenidos a partir de K-means (a) y c)) y DBSCAN (b) y d)), para la provincia de Pastaza.

Como se observa en las Figuras 14, 15, 16, la forma y cantidad de clusters difieren de acuerdo con el método de clusterización aplicado y exceptuando por la provincia de Pastaza no hay similitud en la cantidad de clusters definidos. En la tabla 13 podemos observar la cantidad de clusters, media, mínimo y máximo número de registros por cluster en cada provincia.

Tabla 13. Número de clusters, media, máximo y mínimo de registros por método y provincia.

	Azuay		*Loja*		*Pastaza*	
	K-Means	**DBSCAN**	**K-Means**	**DBSCAN**	**K-Means**	**DBSCAN**
Número de clusters	3	21	2	10	2	2
Media de registros por cluster	8043.7	1149,09	4156.5	831.3	1449	1449
Mínimo de registros por cluster	1182	4	930	4	81	5
Máximo de registros por cluster	21257	21821	7383	7955	2817	2893

También, se puede observar en las Figuras 14 c), 15 c) y 16 c) que los clusters abarcan principales puntos de aglomeración de registros de robo junto con sus a los rededores. Por ejemplo, en la Figura 14 c) que corresponde a la provincia de Azuay, cada cluster coincide con los cantones de Cuenca, Santa Isabel, Girón y Camilo Ponce. Dándonos un indicio que el método de K-means identifica de manera más precisa zonas geográficas significativas en cada provincia. Para confirmar esta hipótesis las Tablas 14 y 15 muestran los índices de *silueta* y de *Davies–Boulding* sobres los clusters obtenidos con los métodos K-means y DBSCAN, los cuales evalúan la eficacia de los métodos de clustering en base la cohesión de los miembros dentro de cada cluster y separación de entre clusters, un índice de silueta cercano a 1 indica que los clusters estan bien definidos y compactos, mientras que un menor índice de *Davies–Boulding muestra* una agrupación más efectiva [7].

Tabla 14. Índices de silueta sobre los clusters obtenidos en cada provincia con los métodos DBSCAN y K-means

Índice de silueta	**Azuay**	**Loja**	**Pastaza**
DBSCAN	0.440	0.613	0.933
K-means	0.806	0.813	0.867

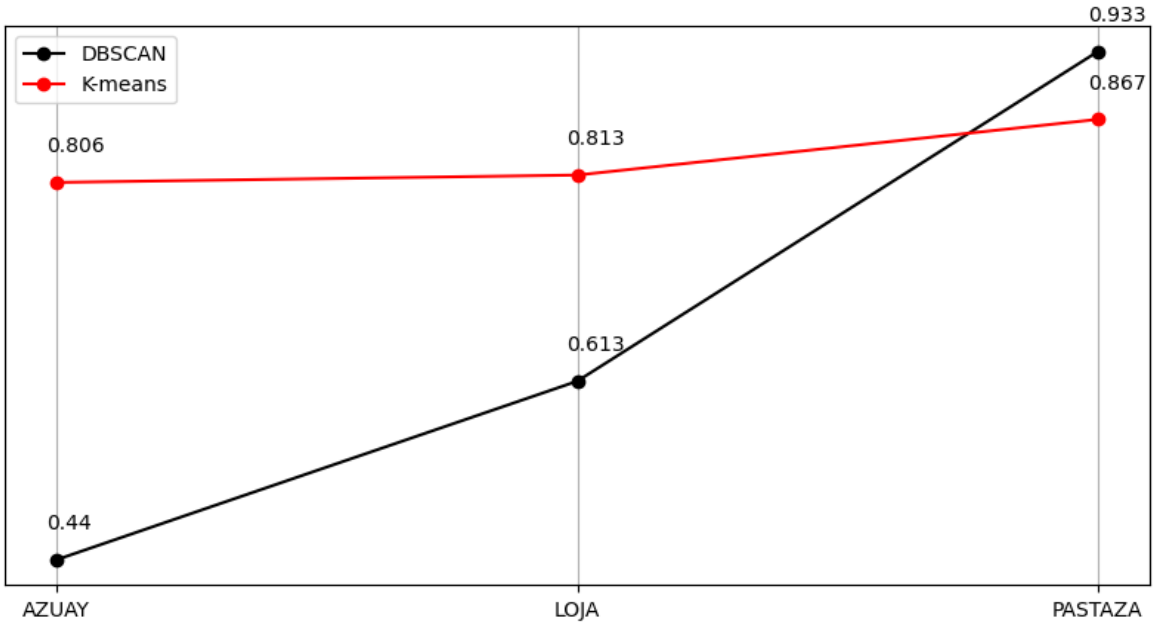


Figura 17. Comparación grafica de los índices de silueta obtenidos en cada provincia con los métodos DBSCAN y K-means

Tabla 15. Índices de *Davies–Boulding* sobre los clusters obtenidos en cada provincia con los métodos DBSCAN y K-means

Índice de Davies–Boulding	**Azuay**	**Loja**	**Pastaza**
DBSCAN	2.37	0.87	0.08
K-means	0.56	0.55	0.74

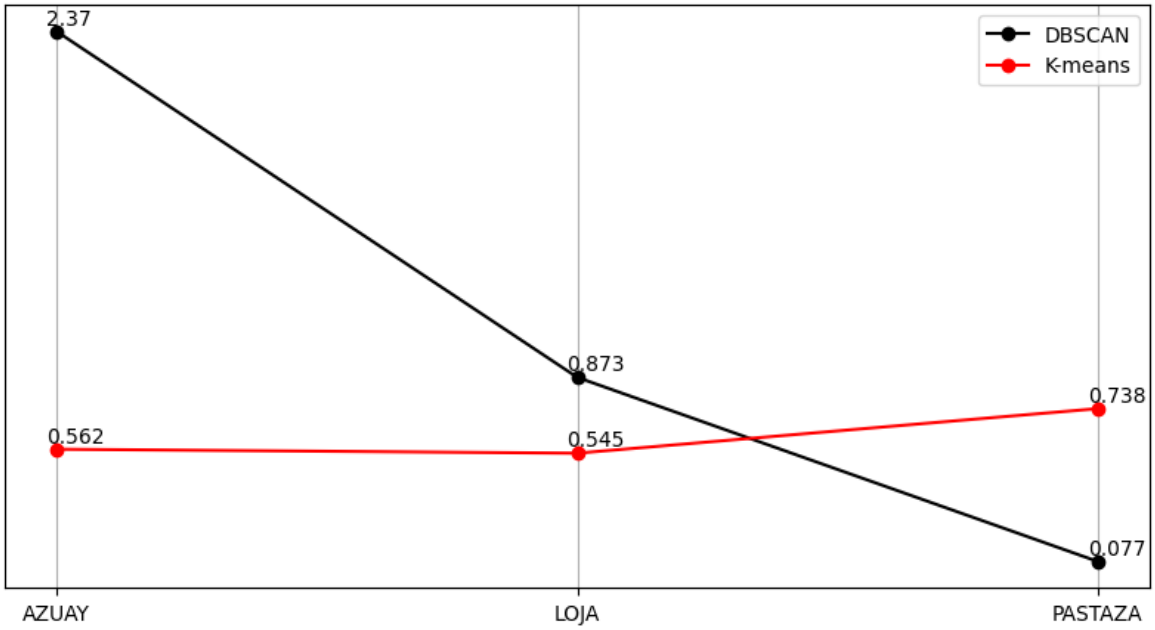


Figura 18. Comparación grafica de los índices de *Davies–Boulding* obtenidos en cada provincia con los métodos DBSCAN y K-means

De acuerdo a los resultados del índice de la silueta, expuestos en la Tabla 14 y la Figura 17, el método de K-means presenta mejores resultados sobre DBSCAN, tanto en las provincias de Azuay y Loja, mientras que, en la provincia de Pastaza, DBSCAN muestra una pequeña superioridad sobre K-means. Lo mismo se puede observar en la Tabla 15 y Figura 18, donde se observa una clara diferencia entre los índices de *Davies–Boulding* calculados sobre los clusters de las provincias de Azuay y Loja, donde el método de K-means indica ser más efectivo, mientras que en la provincia de Pastaza DBSCAN lo supera por una pequeña diferencia.

Una vez analizados los métodos de K-means y DBSCAN sobre las provincias de Azuay, Loja y Pastaza, y de acuerdo a los resultados expuestos anteriormente se ha decido utilizar el método de K-means para identificar zonas geográficas significativas dentro de cada provincia del Ecuador.

En la Figura 19 podemos observar los centroides de los clusters obtenidos en cada una de las 21 provincias analizadas. Además, en la tabla 16 se muestra la cantidad de zonas identificadas por provincia.

Figura 19. Centroides de los clusters obtenidos en cada provincia analizada

Tabla 16. Número de zonas por cada provincia analizada

Provincia	Zonas	Provincia	Zonas	Provincia	Zonas
Azuay	3	*Guayas*	6	*Pastaza*	4
Bolívar	3	*Loja*	2	*Pichincha*	3
Cañar	2	*Los ríos*	2	*Santa Elena*	3
Chimborazo	2	*Manabí*	4	*Santo domingo de los Tsáchilas*	4
Cotopaxi	2	*Morona Santiago*	3	*Sucumbíos*	3
El Oro	5	*Napo*	2	*Tungurahua*	5
Galápagos	3	*Orellana*	5	*Zamora Chinchipe*	3

2.4.2. Minería de reglas de asociación por provincia

Con el objetivo de determinar patrones de robo por provincia que apunten a noticias de robo relacionadas, y una vez identificadas sus zonas geográficas significativas, pasamos a la extracción de reglas de asociación en cada una de ellas.

De acuerdo con los estudios [34], [36], [38], [50], los métodos de extracción de reglas de asociación más utilizados y con mayor éxito en la extracción de patrones criminales son los algoritmos *Apriori* y *FP-Growth*. La elección del algoritmo se basa en el estudio [38] donde se indica que los dos algoritmos mencionados identifican las mismas reglas de asociación, pero mostrando una diferencia significativa en el rendimiento. Donde *FP-Growth* se destaca sobre *Apriori,* además comparamos los resultados obtenidos y el tiempo de ejecución sobre los registros determinados por los clusters de las provincias de: Santo Domingo, Morona Santiago y Tungurahua. Estas provincias fueron elegidas debido a que representan los cuartiles1, 2 y 3 respectivamente, sobre la cantidad de registros por provincia, sin tomar en cuenta a las provincias de Pichincha y Guayas, las que contienen una cantidad de registros atípica con respecto al número de registros en el resto de provincias.

Se utilizó un total de 7 variables en este estudio que son: GRUPO_HORAINC, DIA_INCIDENTE, URB_RURAL_INCIDENTE, MODALIDAD_DESAGREGACION, TIPO_ARMA. CANTON_INCIDENTE y MES_INCIDENTE.

Tabla 17. Comparación de tiempos de ejecución de los algoritmos A priori y FP-growth sobre los clusters de la provincia de Morona Santiago

Tiempo de ejecución	**cluster 0**	**cluster 1**	**cluster 2**
Apriori	0.04	0.648	0.064
FP-growth	0.018	0.312	0.04

Tabla 18. Comparación de las medias de soporte de los itemsets obtenidos con los algoritmos A priori y FP- Growth sobre los clusters de la provincia de Morona Santiago

Media del soporte de los itemsets obtenidos	**cluster 0**	**cluster 1**	**cluster 2**
Apriori	0.061	0.012	0.039
FP-growth	0.061	0.012	0.039

En las tablas 17 y 18 se puede observar tanto el tiempo de ejecución de los algoritmos *Apriori* y *FP-Growth*, así como la media de soporte del conjunto de itemsets obtenidos a partir de los clusters de la provincia de Morona Santiago. A partir de esta información corroboramos que FP-Growth presenta menor tiempo de ejecución y extrae las mismas reglas de asociación que Apriori, por lo tanto, lo utilizaremos para la extracción de reglas de asociación de cada una de las provincias de la base de datos estudiada.

Utilizamos el algoritmo *fpgrowth* de la librería de Python *mlxtend,* este requiere de una base de datos transaccional y un soporte mínimo para poder ser ejecutado. Para el primer requisito se empleó el método *get_dummies* de la librería pandas sobre las variables seleccionadas, este método transforma todas las variables cualitativas en binarias con los valores 1 y 0. Mientras que para la definición del soporte mínimo se realizó el siguiente análisis:

Al ser el soporte la medida que indica la proporción en que se repite un itemset dentro de una base transaccional, y debido a que el objetivo de este estudio es encontrar noticias del delito relacionadas mediate patrones de robo específicos, el soporte debe ser pequeño para no pasar por alto estos patrones, que al representar comportamientos específicos a la hora de cometer un robo, poseen un bajo soporte de manera general, pero debe tener cierta frecuencia dentro de la base, por lo tanto el soporte mínimo para el algoritmo *FP-Growth* de definió con la Ecuación (2).

$$Soporte = \frac{n}{N} \tag{2}$$

Donde n indica la frecuencia mínima que debe tener un itemset dentro del conjunto de registros perteneciente a cada cluster de la provincia analizada y N es la cantidad total de registros dentro del cluster.

Utilizando la fórmula 2 se experimentó con n en un rango de 3 a 10 dentro de los clusters de las provincias de Santo Domingo, Morona Santiago y Tungurahua. Para determinar el soporte mínimo optimo, se observó el promedio de campos que abarcan los itemsets de longitud máxima encontrados con cada soporte. Debido a que mientras más específico sea el itemset encontrado es decir mientras mayor sea su longitud, mayor probabilidad tiene este de ser un patrón de robo que puede relacionar registros dentro de la base. En la Figura 20 observamos los resultados de este experimento, donde en el eje *x* tenemos la frecuencia mínima (n) y en el eje *y* se muestra el porcentaje promedio de campos que abarcan los itemsets de mayor longitud.

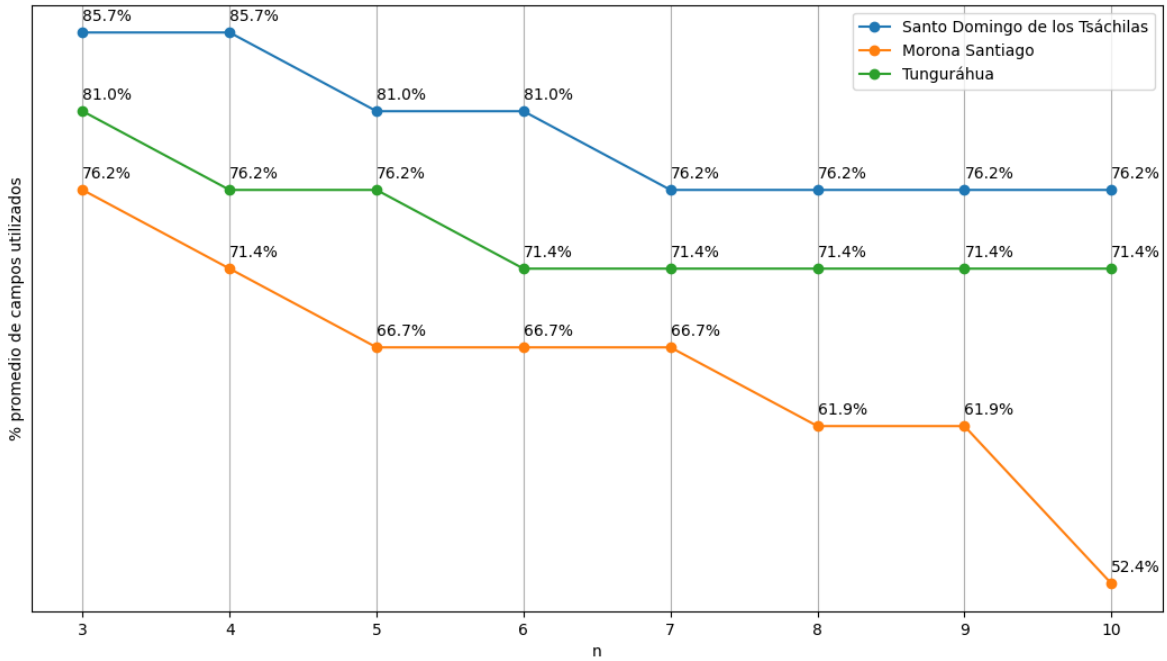


Figura 20. Porcentaje promedio de campos utilizados frente al valor *n* en los itemsets de mayor longitud en las provincias de Santo Domingo, Morona y Tungurahua

Como se muestra en la Figura 20 el soporte mínimo óptimo para encontrar itemsets de mayor longitud se calcula utilizando la fórmula 2 y con una frecuencia mínima $n = 3$.

Una vez identificado el soporte mínimo optimo, procedemos a obtener las reglas de asociación a partir de los itemsets de cada cluster utilizando la métrica *lift*, esta nos indica que un valor mayor a 1, asegura que el antecedente de las reglas de asociación obtenidas tiene una fuerte relación positiva y causal sobre su consecuente. En consecuencia, el conjunto de reglas de asociación de cada provincia se obtuvo con un *lift* mínimo de 1.

2.4.3. Extracción de patrones de robo y NDDs relacionadas por provincia y a nivel nacional

Una vez encontradas las reglas de asociación con fuerte correlación entre antecedente y consecuente en los clusters de cada provincia, podemos identificar patrones criminales específicos dentro de ellas. Para esto necesitamos seleccionar las reglas de asociación con mayor cantidad de variables involucradas en la misma.

Filtramos las reglas con los consecuentes de tamaño máximo, dichas reglas nos indicaran patrones específicos de robo dentro del cluster. Por ejemplo: en el cluster 0 de la provincia de Morona Santiago, 6% de los robos han sido bajo la modalidad de asalto, en el cantón Palora Metzera, en la madrugada, los días sábados (5) y con armas que no son contundentes, blancas, de fuego ni constrictoras. La figura 21 muestra una representación gráfica del patrón descrito.

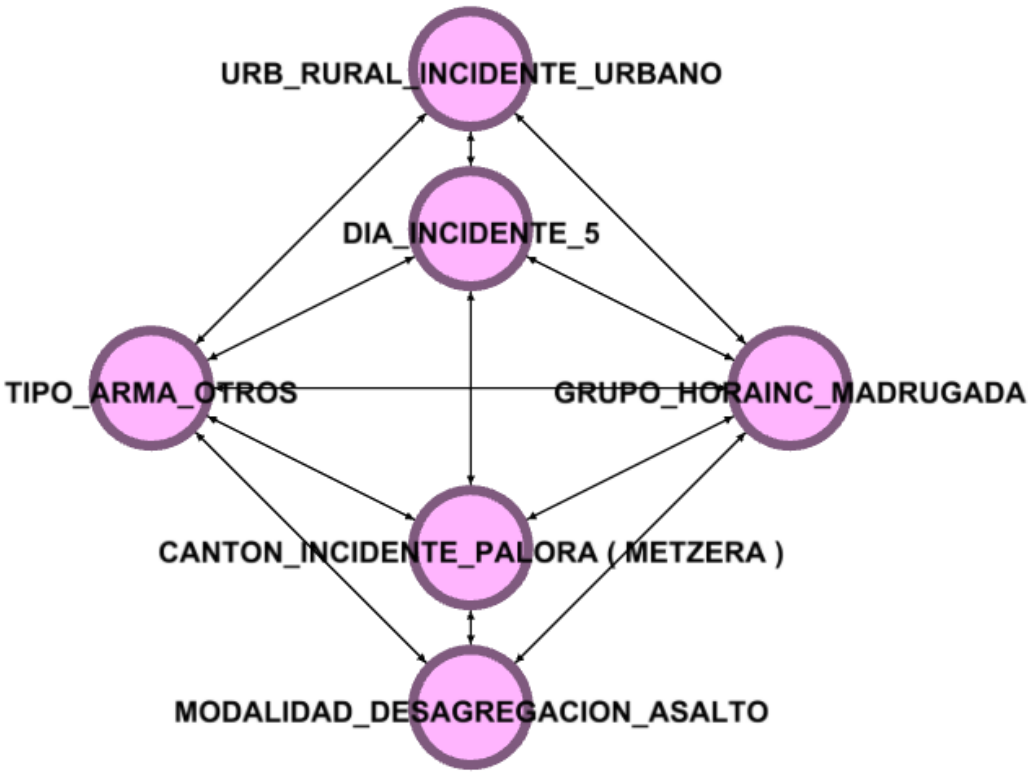


Figura 21. Patrón criminal encontrado en el cluster 0 de la provincia de Morona Santiago

Sin embargo, el filtrado de reglas de asociación con consecuentes de tamaño máximo no es suficiente para encontrar todos los patrones criminales específicos, este conjunto de reglas puede ser amplio y no todas las reglas filtradas van a representar un patrón criminal diferente, para solucionar este problema identificamos las reglas de asociación con el mismo soporte, esto nos indica que dichas reglas tienen la misma frecuencia dentro de la base de datos, dándonos un indicio de que las reglas de asociación pertenecientes a este conjunto representan, un patrón criminal.

Para determinar si las reglas de asociación con el mismo soporte pertenecen a un solo patrón criminal, estas deben cumplir las siguientes condiciones:

1. El conjunto de valores únicos de todos los antecedentes debe ser igual al conjunto de valores únicos de todos los consecuentes
2. No puede existir dos elementos pertenecientes a la misma variable.

Tomando en cuenta lo mencionado en la Tabla 19 se muestran los patrones criminales encontrados dentro de la provincia de Morona Santiago.

Tabla 19. Patrones criminales específicos encontrados en la provincia de Morona Santiago

	Patrón 1	**Patrón 2**	**Patrón 3**	**Patrón 4**
Cluster	0	1	1	2
Soporte	0.066	0.0038	0.0051	0.0175
GRUPO_HORAINC	*MADRUGADA*	*TARDE*	*MADRUGADA*	*NOCHE*
DIA_INCIDENTE	*Sábado*	*Jueves*	*Sábado*	*Sábado*
URB_RURAL_INCIDENTE	*URBANO*	*RURAL*	*URBANO*	*URBANO*
MODALIDAD_DESAGREGACION	*ASALTO*	*ASALTO*	*ASALTO*	*ESTRUCHE*
TIPO_ARMA	*OTROS*	*ARMA BLANCA*	*ARMA BLANCA*	*OTROS*
CANTON_INCIDENTE	*PALORA METZERA*	*MORONA*	*MORONA*	*GUALAQUIZA*
MES_INCIDENTE		*Julio*	*Julio*	

A partir de la información mostrada en la Tabla 19, podemos identificar los registros de crímenes relacionados con cada patrón dentro de la provincia de Morona Santiago, esta información se presenta en la Tabla 20

Tabla 20. Conteo de NDDs relacionadas por patrones criminales en la provincia de Morona Santiago

	Cantidad de NDDs relacionadas
Patrón 1	9
Patrón 2	3
Patrón 3	8
Patrón 4	5

El mismo proceso se realizó con el resto de provincias en la base de datos y con cada uno de sus clusters. Como resultado, se identificaron un total de 17784 patrones criminales que relacionan a 172905 noticias de delito en todo el país. En la Tabla 21 se muestra el total de noticias de delito relacionadas dentro de cada provincia.

Tabla 21 Conteo de NDDs relacionadas por provincia

Provincia	**NDDs Relacionadas**	**Provincia**	**NDDs Relacionadas**
Pichincha	86579	*Loja*	599
Guayas	43043	*Orellana*	540
Los Ríos	8313	*Cañar*	493
Azuay	6944	*Cotopaxi*	148
El oro	6738	*Sucumbíos*	70
Manabí	5781	*Napo*	60
Santo domingo de los Tsáchilas	5246	*Pastaza*	35
Santa elena	3374	*Morona Santiago*	25
Chimborazo	2721	*Bolívar*	6
Tungurahua	2251	*Zamora Chinchipe*	6

2.4.4. Creación de dashboards

En la siguiente sección se presenta la creación de dashboards para el análisis de los datos otorgados por la FGE y visualización las noticias de delito encontradas. Para su creación se utilizó la base de datos resultante de etapa de preparación de los datos presente en la metodología CRISP-DM, junto con los resultados de las secciones anteriores.

El dashboard se compone de 5 páginas las cuales presentan la siguiente información.

Tabla 22. Descripción del dashboard por páginas

Página	Información
1	Cantidad de registros por año, por locación urbana o rural, y por modalidad de robo
2	Registros de robo por mes, día del mes, día de la semana y grupo horario
3	Conteo de registros por provincia, cantón y modalidad
4	Registros dentro de cada zona identificada, con filtro de modalidad de robo
5	Registros de robo relacionados dentro de cada zona geográfica identificada

En la figura 22 se muestra cada página del dashboard en cuestión.

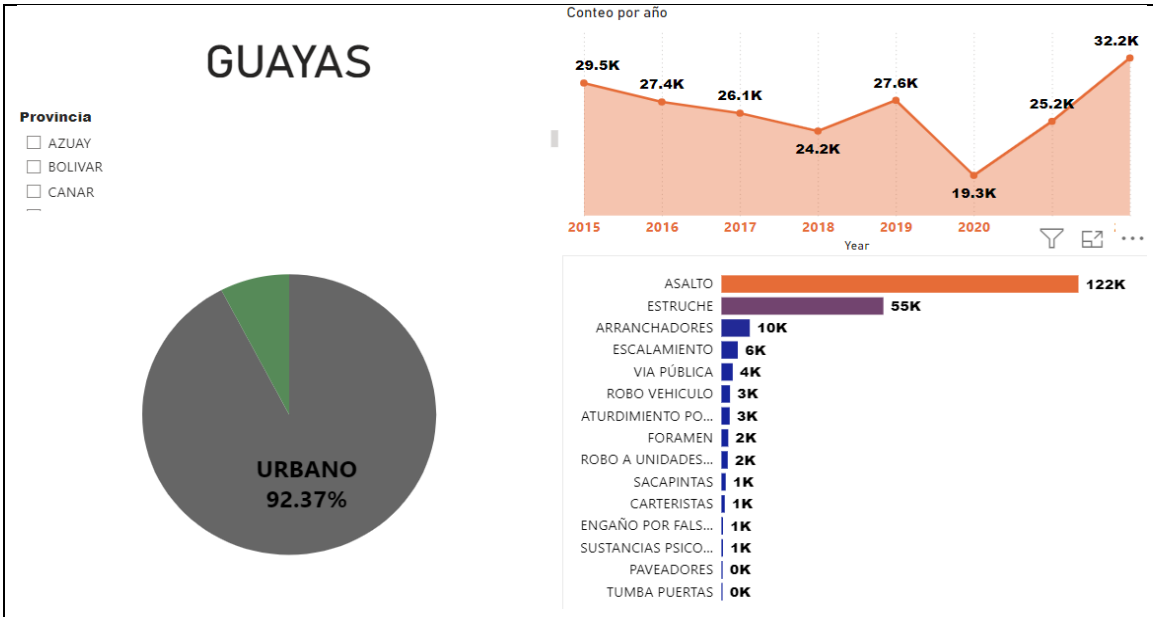


Figura 22. a) Página 1 conteo de registros por provincia, año, localidad y modalidad

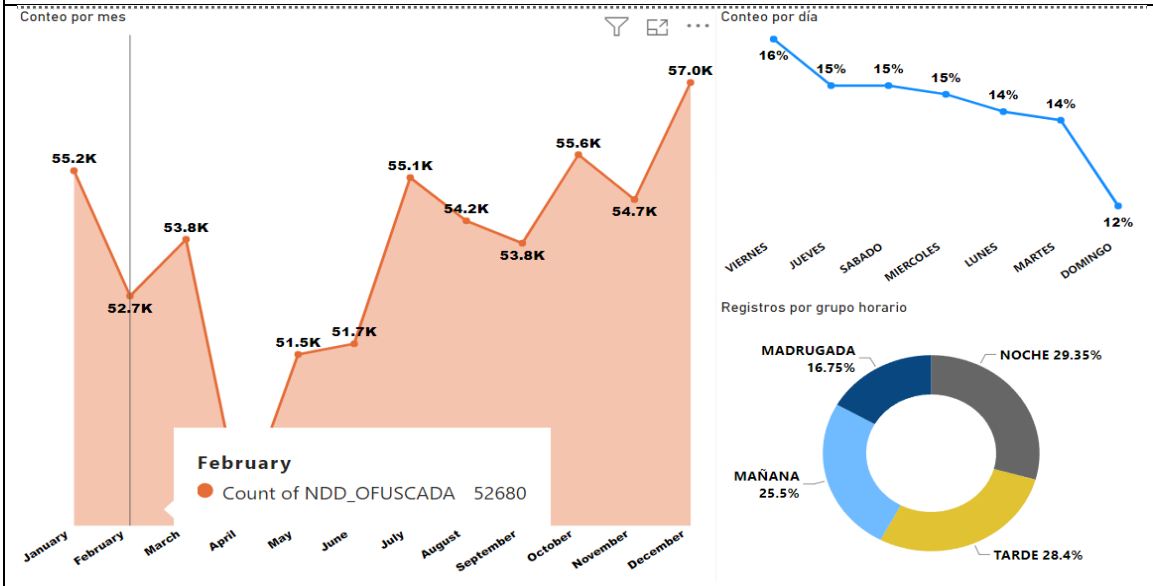


Figura 22. b) Conteo de registros, por mes, grupo horario y día de la semana

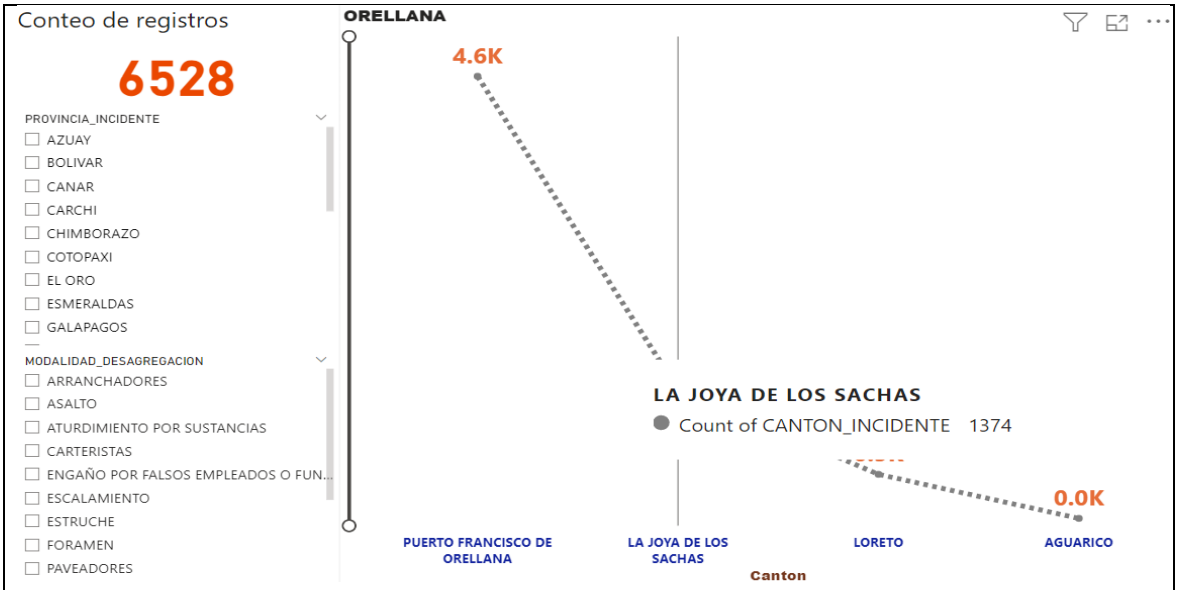


Figura 22. c) Conteo de registros por cantón y modalidad

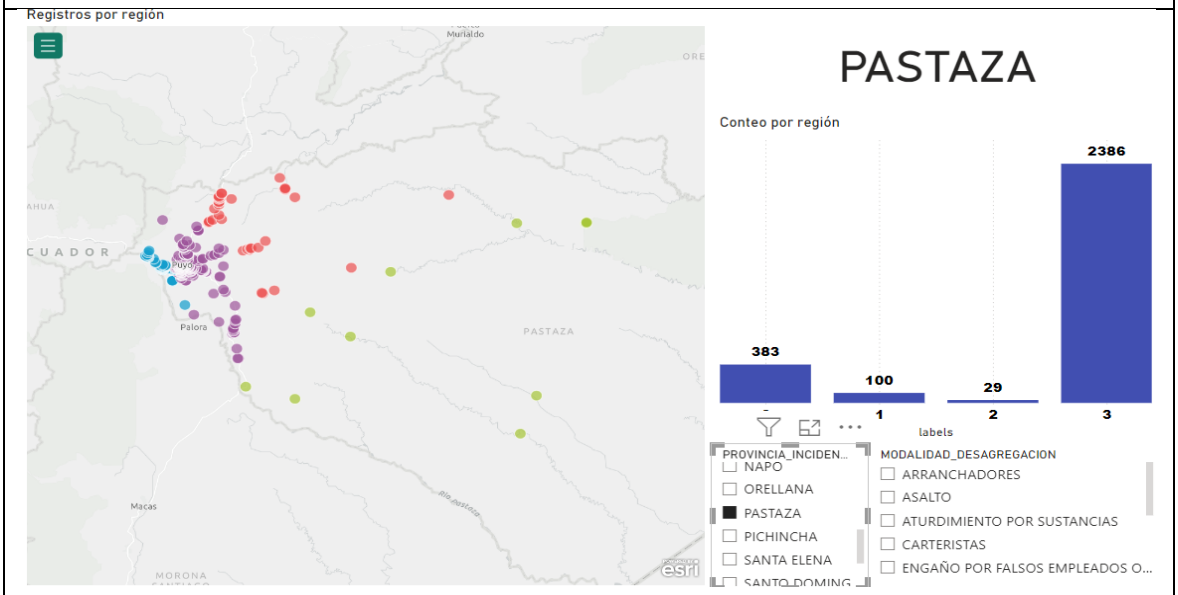


Figura 22. d) Conteo por zona dentro de cada provincia y modalidad

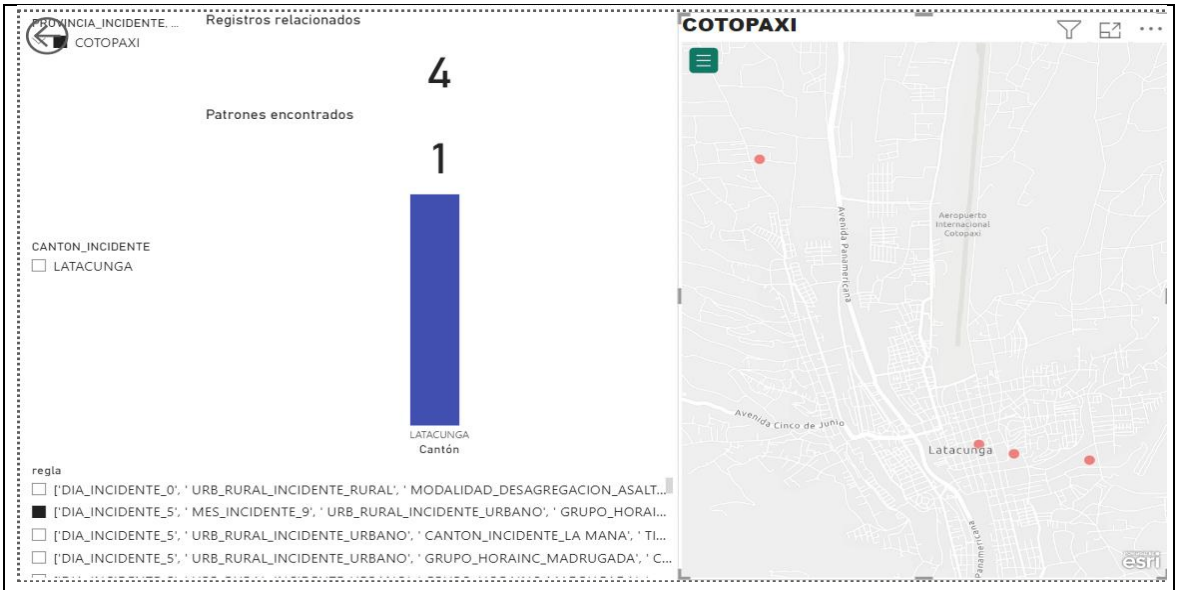


Figura 22. e) Visualización de registros de robo relacionados, cantidad de registros relacionados y reglas por provincia y cantón

Figura 22. a), b), c) y d). *Dashboard* para el análisis y visualización de resultados

Cabe recalcar que el dashboard desarrollado tiene la finalidad de analizar el comportamiento criminal dentro de cada provincia en base a los datos otorgados por la FGE además de visualizar los clusters encontrados dentro de cada provincia y las NDDs relacionadas en cada cluster.

2.5. Evaluación

Con el objetivo de evaluar el desempeño de los modelos resultantes de la etapa anterior, esta etapa se ha divido en dos secciones que se presentan a continuación.

2.5.1. Evaluación de zonas geográficas encontradas en cada provincia

En la evaluación de las zonas geográficas encontradas mediante el método de clusterización K-means, y tomando en cuenta que no existe una fuente externa con información que valide los clusters encontrados. Se realizó una validación interna de los mismos, donde se analizó la calidad de los clusters en base a: coherencia interna, separación, compactación interna, y dispersión interna de cada uno, en las 21 provincias analizadas. Con este propósito se calcularon los índices mostrados en la tabla 23.

Tabla 23. Índices utilizados para la evaluación de clusters por provincia.

Índice	Aspectos que evalúa	Interpretación
Silueta	• Coherencia interna • Separación entre clusters	Un valor cercano a 1 indica que los valores están bien agrupados, mientras que uno cercano a -1 indica lo contrario
Calinski-Harabasz	• Dispersión entre clusters • Dispersión interna	Un valor alto indica que los clusters están bien definidos y separados
Davies-Bouldin	• Separación entre clusters • Dispersión interna	Un bajo valor indica que los clusters tienen poca dispersión interna y están separados

Podemos observar en las figuras 23, 24 y 25 cada uno de los índices de la tabla 23 calculados para los clusters encontrados en las 21 provincias, dichas figuras ordenan de mayor a menor las provincias de acuerdo al valor de los índices, mientras que en el eje *Y* se pueden visualizar los valores máximo, medio y mínimo del conjunto de índices calculado.

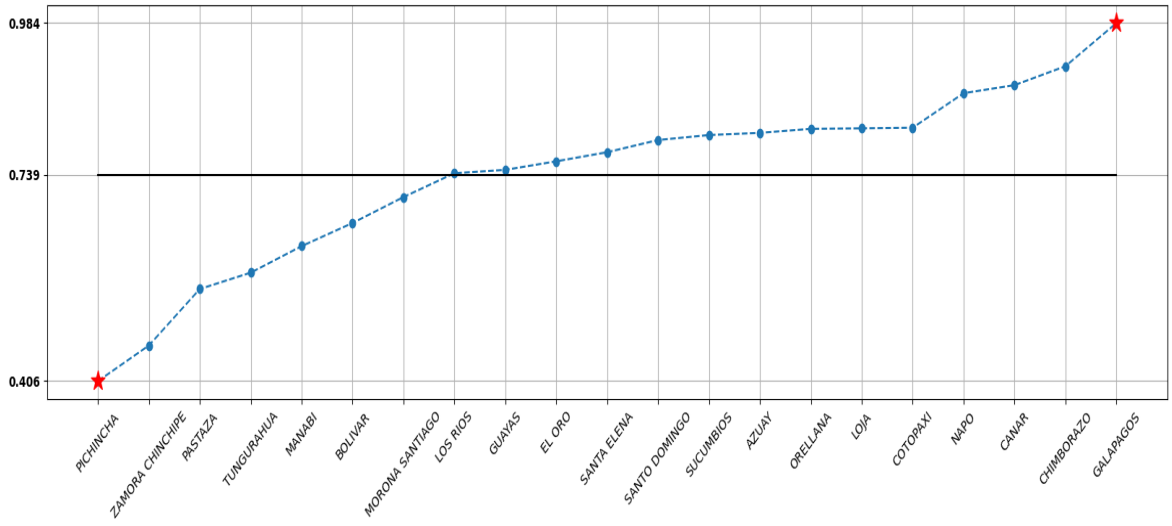


Figura 23 Índice de silueta para los clusters de las 21 provincias

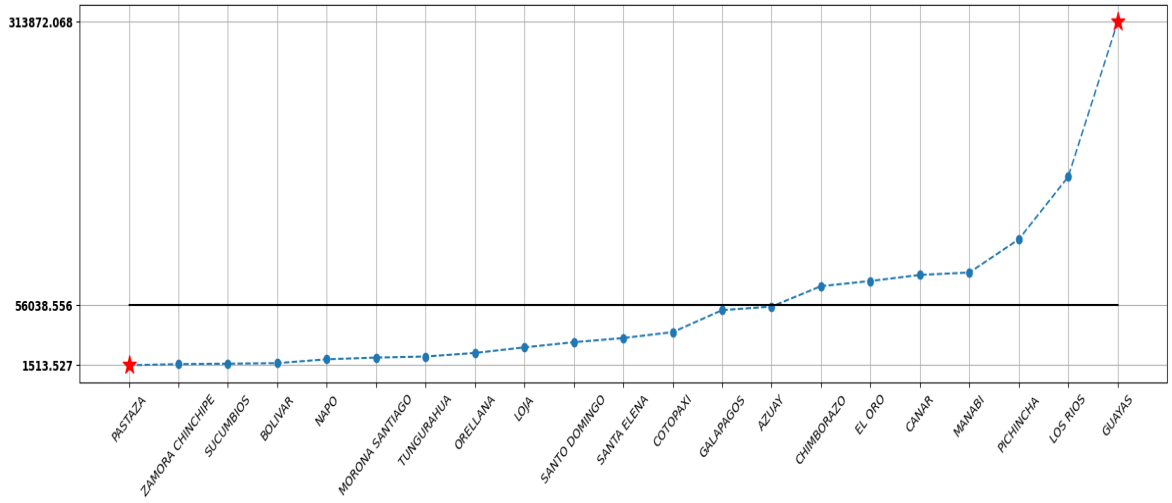


Figura 24. Índice de *Calinski-Harabasz* para los clusters de las 21 provincias

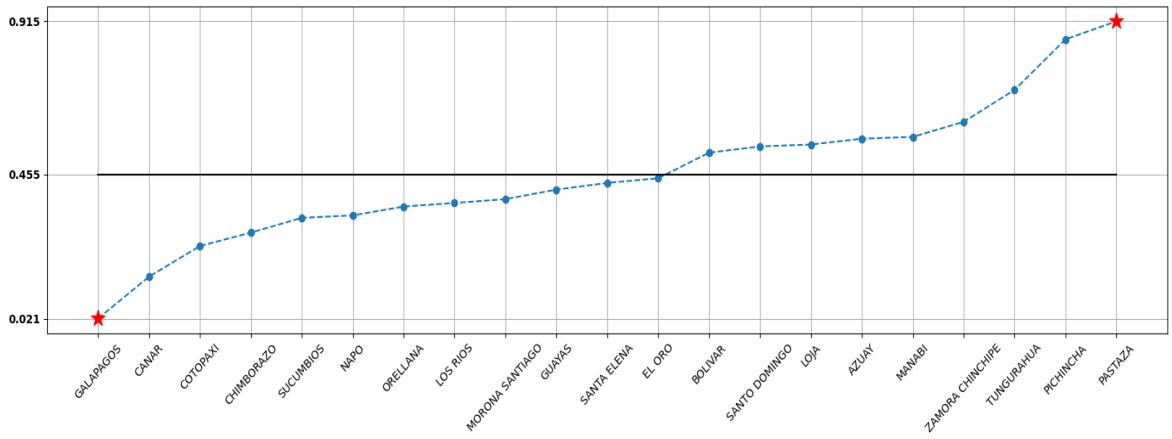


Figura 25. Índice de *Davies-Bouldin* para los clusters de las 21 provincias

En base a lo observado en las Figuras 23, 24 y 25 la provincia con mejor separación entre clusters, coherencia interna y dispersión interna es la provincia de Galápagos, mientras que de acuerdo con el índice de *Calinski-Harabasz*, la provincia con mejor dispersión entre clusters y dispersión interna es la provincia de Guayas. Además, la provincia con los peores índices de *Davies-Bouldin* y *Calinski-Harabasz* es la provincia de Pastaza, presentando el índice de *Davies-Bouldin* más alto y el menor índice de *Calinski-Harabasz,* indicando la peor separación entre clusters y dispersión interna de las 21 provincias, a pesar de tener una mejor coherencia interna que las provincias de Pichincha y Zamora Chinchipe.

A continuación, observamos las métricas de estadística descriptiva de cada uno de los clusters con el objetivo de realizar un análisis general, en la Tabla 24 se presentan los valores mínimos, máximo, junto con los cuartiles 1,2 y 3 del conjunto de índices calculados.

Tabla 24. Mínimo cuartiles 1,2 ,3 (Q1, Q2, Q3) y máximo de los índices calculados sobre los clusters de las 21 provincias del Ecuador

	Mínimo	**Q1**	**Q2**	**Q3**	**Máximo**
Índice de silueta	0.41	0.66	0.78	0.81	0.98
Calinski-Harabasz	1513.53	8482.47	26238.82	78164.04	313872.07
Davies-Bouldin	0.02	0.33	0.43	0.56	0.92

De acuerdo con la información de la Tabla 24, se puede deducir que:

- El 75% de las provincias tiene un índice de silueta mayor a 0.66, mientras que el 25% restante figura un índice entre 0.41 y 0.66. Esto nos indica que ese 75% tiene una buena coherencia interna y separación entre clusters, mientras que el 25%. restante presenta coherencia interna y separación entre clusters aceptable.
- El 100% de las provincias presenta un índice de *Davies-Bouldin* menor a 0.92, y tomando en cuenta que, mientras más bajo sea el índice de *Davies-Bouldin* mejor separación entre clusters y dispersión interna presenta la agrupación, se puede determinar que los clusters presentan una buena dispersión interna, esta conclusión se fundamenta con los resultados del índice de *Calinski-Harabasz* que evalúa los mismos aspectos, con la diferencia que mientras mayor sea el índice mejor es la agrupación.

2.5.2. Evaluación de reglas de asociación y NDDs relacionadas

En esta sección se evalúan las reglas de asociación encontradas que relacionan NDDs.

Al no tener un convenio con la FGE y dado que los datos fueron entregados con un ID (NDD) ofuscado, con el objetivo de proteger los datos personales de los involucrados, se imposibilita revisar los expedientes y relatos pertenecientes a los registros que el presente estudio indica que están relacionados. Por lo tanto, no se puede realizar un análisis de efectividad del presente proyecto para encontrar NDDs relacionadas.

Consecuentemente y al igual que en la sección anterior realizaremos un análisis interno de las reglas de asociación encontradas y los patrones que asocian las NDDs dentro de cada provincia. Para esto revisaremos los índices de: Elevación (lift), convicción (conviction) y la

métrica de Zhang de cada una de las reglas de asociación encontradas. La Tabla 25 presenta los aspectos que evalúa e interpretación de cada uno de los índices mencionados.

Tabla 25. Aspecto que evalúa e interpretación de los índices: Lift, conviction, y Zhang

Índice	Aspecto que evalúa	Interpretación
Lift	Dependencia causal entre antecedente y consecuente	Un valor mayor a 1 indica que una dependencia causal fuerte positiva entre antecedente y consecuente
Conviction	Independencia entre antecedente y consecuente	A mayor convicción más fuerte será la regla
Zhang	Muestra la naturaleza y fuerza de la correlación entre antecedente y consecuente	Está delimitado entre -1 y 1, donde un valor de 1 indica una correlación fuerte positiva, -1 correlación fuerte negativa y 0 que no existe correlación.

Calculamos los índices mostrados en la Tabla 25 sobre las reglas de asociación que describen NDDs relacionadas dentro de cada provincia. Para realizar este análisis se observaron las los promedios del conjunto de índices calculados a partir de las reglas de asociación obtenidas por provincia. Las figuras 26, 27 y 28 muestran los promedios de cada índice por provincia en el eje *y,* y la provincia a la que pertenece en el eje *x,* Con el objetivo de identificar la provincia con promedio más alto y bajo los valores fueron ordenados de manera descendente.

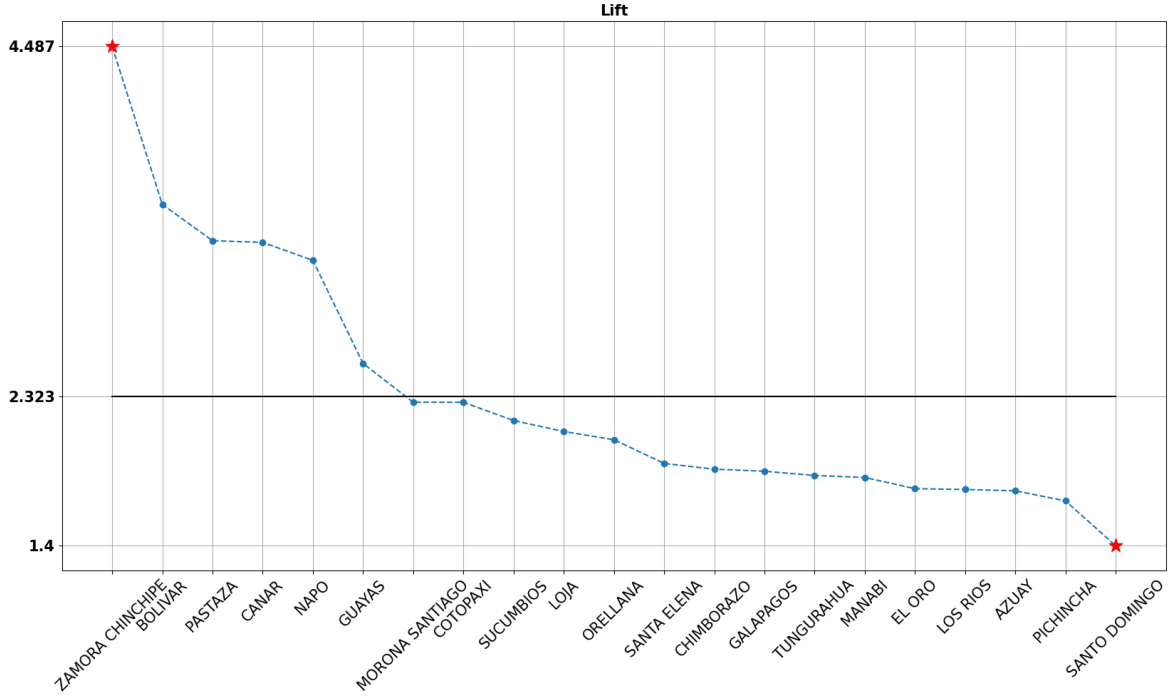


Figura 26. Promedio de índices Lift en cada provincia

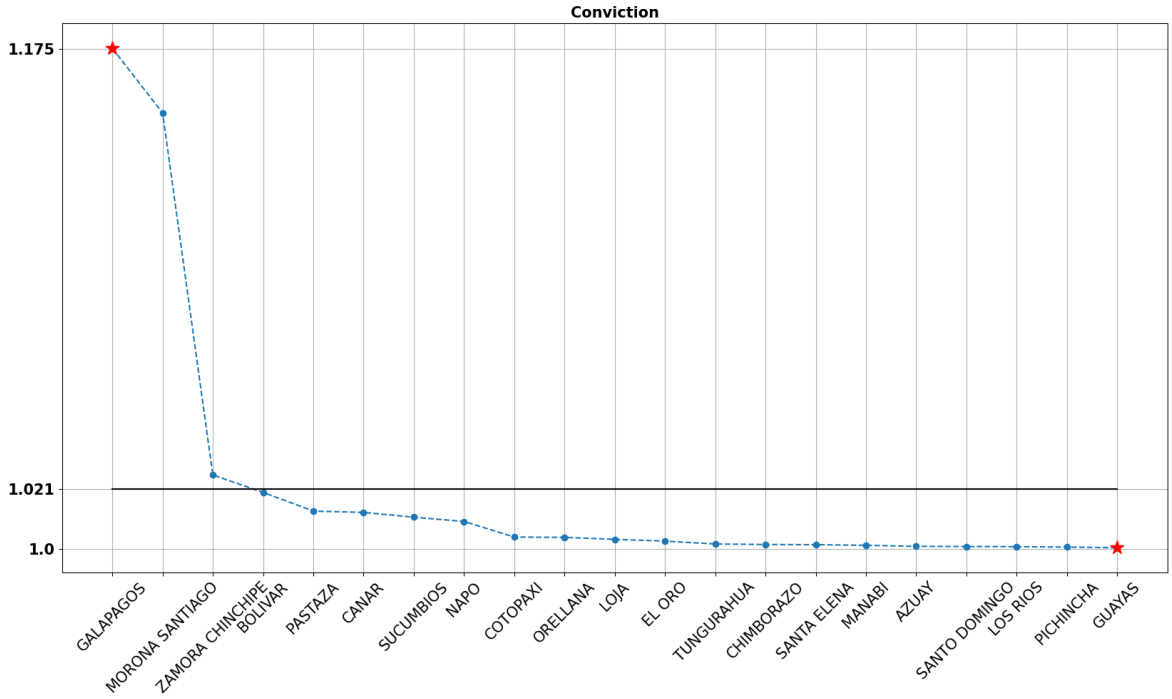


Figura 27. Promedio de índices de conviction en cada provincia

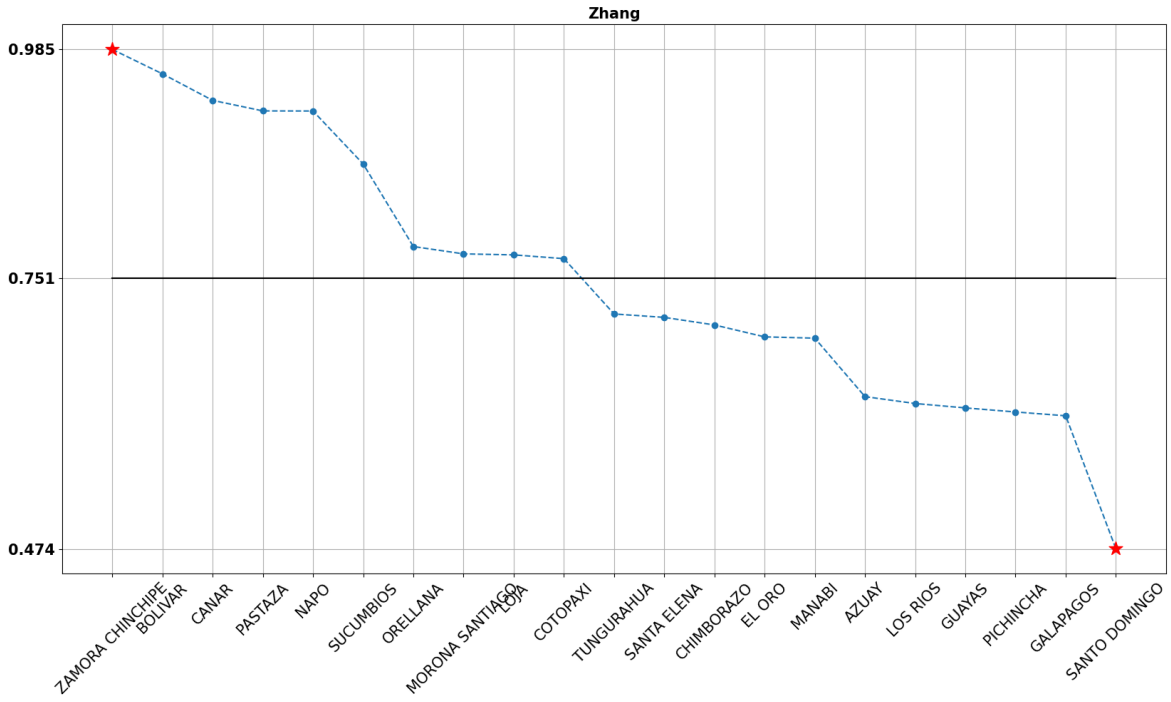


Figura 28. Promedio de índices de Zhang en cada provincia

A partir de la información mostrada en las figuras 26, 27 y 28 podemos deducir que las reglas de asociación encontradas en la provincia de Zamora poseen correlación y dependencia positiva más fuerte, mientras que las reglas encontradas en Galápagos poseen mejor índice de convicción, lo que indica, reglas de asociación más fuertes. Además, el valor mínimo del índice de *lift* es de 1.4, esto significa, que todas las reglas de asociación encontradas poseen una fuerte dependencia causal positiva.

A partir de lo mencionado anteriormente y la información expuesta en las figuras 26, 27 y 28, podemos deducir que todos los antecedentes tienen una fuerte dependencia causal positiva con sus consecuentes, además de presentar correlación entre moderada y fuerte. Al obtener las NDDs relacionadas mediante el descubrimiento de patrones que se originan del análisis de estas reglas de asociación que acabamos de evaluar en las cuales el conjunto de valores únicos de todos los antecedentes debe ser igual al conjunto de valores únicos de todos los consecuentes podemos inferir que dichas reglas forman efectivamente un patrón de robo.

3. Resultados y Discusión

3.1. Resultados

En esta sección se resume los principales resultados encontrados, luego de haber realizado las etapas de la metodología descritas en la sección anterior. Para ello se presentan las siguientes subsecciones.

3.1.1. Limpieza y selección de los datos

El primer resultado de este estudio se presenta una vez culminada la etapa de preparación de los datos, cuyo resultado es la obtención de una base de datos, a partir de los datos otorgados por la FGE, con la información y formato idóneos para su posterior uso en modelos de minería descriptiva de datos. Se destaca la selección de variables, limpieza de datos nulos, transformación y depuración de datos geográficos. La información de dicha base se presenta en la tabla 26.

Tabla 26. Descripción de la base de datos resultante de la etapa de preparación de los datos

Cantidad de registros	589771
Número total de variables	16
Número de variables cualitativas	10
Número de variables cuantitativas	6
Provincias utilizadas	21
Provincias Eliminadas	3 (Esmeraldas, Imbabura, Carchi)
Peso en MB	81.6

3.1.2. Identificación de zonas geográficas en cada provincia

Como resultado de la aplicación del algoritmo de K-means sobre los registros de coordenadas en cada provincia se identificó zonas geográficas dentro de cada una de ellas. Estas zonas geográficas fueron utilizadas posteriormente para la delimitación de la búsqueda de patrones de robo por provincia. En la figura 29 se puede observar los centroides de cada zona identificada, marcada con una cruz celeste.

Figura 29. Centroide de cada una de las zonas geográficas identificadas en las 21 provincias del Ecuador.

Además, en la tabla 27 podemos observar la cantidad de zonas identificadas por provincia junto con el número de registros que contiene.

Tabla 27. Zonas geográficas identificadas por provincia y cantidad de registros en cada una de ellas

Provincia	**Zonas**	**Cantidad de registros por zona**	**Provincia**	**Zonas**	**Cantidad de registros por zona**
Pichincha	1	70445	*Manabí*	0	13573
	0	40837		2	11638
	2	40719		1	6107
Guayas	1	169024		3	3013
	2	19549	*Los Ríos*	0	20888
	5	6705		1	16650
	4	5679	*Loja*	0	7383
	3	4445		1	930
	0	3613	*Galápagos*	0	102

Provincia	Zonas	Cantidad de registros por zona	Provincia	Zonas	Cantidad de registros por zona
Zamora Chinchipe	0	1013		1	40
	2	198		2	25
	1	172	*El oro*	2	17024
Tungurahua	0	10077		0	6195
	2	1300		3	4612
	1	689		1	3528
Sucumbíos	0	1906		4	860
	1	132	*Cotopaxi*	0	7010
	2	45		1	1107
Santo domingo de los Tsáchilas	0	18291	*Chimborazo*	0	11351
	2	975		1	845
	1	643	*Cañar*	1	3437
Santa Elena	0	10269		0	2451
	2	2088	*Bolívar*	0	1272
	1	594		1	621
Pastaza	3	2386		2	257
	0	383	*Azuay*	0	21257
	1	100		1	1692
	2	29		2	1182
Orellana	2	4242	*Napo*	0	2622
	0	1360		1	195

Provincia	Zonas	Cantidad de registros por zona	Provincia	Zonas	Cantidad de registros por zona
	1	520	*Morona Santiago*	0	2240
	4	307		1	517
	3	41		2	371

Los registros pertenecientes a cada zona se pueden visualizar en el dashboard desarrollado en la sección anterior y presente en los anexos.

3.1.3. NDDs relacionadas

Para identificar NDDs relacionadas entre sí, primero se obtuvo reglas de asociación maximáles en cada una de las zonas geográficas determinadas, a partir de estas reglas se construyeron patrones de robo que relacionan NDDs dentro de cada zona geográfica, en la tabla 28 se presentan la cantidad de patrones encontrados en cada provincia, junto con la cantidad de NDDs relacionadas y el promedio de NDDs que cada patrón de robo relaciona por provincia.

Tabla 28. Cantidad de patrones pro provincia, junto con el número de NDDs relacionadas y promedio de NDDs relacionadas por patrón y provincia.

Provincia	Cantidad de Patrones	Cantidad NDDs relacionadas	Promedio NDDs por patrón y provincia
Azuay	1409	6944	4.93
Bolívar	2	6	3
Cañar	99	493	4.98
Chimborazo	624	2721	4.36
Cotopaxi	32	148	4.62
El oro	1216	6738	5.54
Guayas	5304	43043	8.12

Provincia	Cantidad de Patrones	Cantidad NDDs relacionadas	Promedio NDDs por patrón y provincia
Loja	170	599	3.52
Los Ríos	1779	8313	4.67
Manabí	1280	5781	4.52
Morona Santiago	4	25	3.33
Napo	20	60	3
Orellana	151	540	3.58
Pastaza	11	35	3.18
Pichincha	3739	86579	23.16
Santa elena	609	3374	5.54
Santo domingo de los Tsáchilas	691	5246	7.59
Sucumbíos	21	70	3.33
Tungurahua	544	2251	4.14
Zamora Chinchipe	2	6	3
Total	*17707*	*172972*	**Promedio:** 5.41

Como se puede observar en la tabla 28 el promedio de NDDs relacionadas por patrón de robo encontrado a nivel nacional es de 5.41. Mientras que las provincias que presentan un mayor promedio de NDDs relacionadas son: Pichincha con 23.16 y Guayas con 8.12. Por otro lado, el promedio mínimo de NDDs relacionadas por patrón es de 3, presente en las provincias de: Bolívar, Napo y Zamora.

3.1.4. Dashboard

Para facilitar el análisis y visualización de resultados, se generó un dashboard que consta de 5 páginas, la descripción de las páginas y capturas de cada una de ellas se pueden

observar en la Tabla 22 y Figura 22 presentes en la sección anterior, mientras que el enlace al dashboard se encuentra en los anexos.

3.2. Discusión

En el presente trabajo se explora técnicas de minería descriptiva de datos, con el objetivo de encontrar NDDs relacionadas a partir de los datos otorgados por la FGE. Para el cumplimiento de este objetivo se utilizaron las técnicas descritas en las investigaciones utilizadas dentro de la revisión literaria de este documento. A partir de la misma se determinó el uso de técnicas de clusterización para delimitar las zonas geográficas estudiadas y técnicas de minería de reglas de asociación, para la identificación de patrones de robo que relacionen las NDDs dentro de cada zona.

Con el propósito de discutir con mayor detalle los resultados y técnicas utilizadas la presente sección se divide en tres subsecciones que son:

3.2.1. Limpieza y selección de los datos

De acuerdo a la metodología CRISP-DM una vez comprendido el negocio y los datos a analizar, se procede a la etapa de preparación de los datos, el resultado de esta etapa es una base de datos que puede ser utilizada en las etapas subsiguientes. En el contexto de este estudio la limpieza y transformación de los datos otorgados por la FGE fue exhaustiva, debido a la gran cantidad de variables cualitativas categóricas con categorías redundantes, sinónimas o que poseen espacios y símbolos que generan ruido en su interpretación, además de presentar formatos que de primera mano la librería *pandas* de Python interpreta como un objeto a pesar de contener información cuantitativa, como la variable COORDENADA_INCIDENTE que se presenta en formato de tupla.

También, se debe mencionar que, varias coordenadas no son congruentes con la información presentada en el resto de variables del registro, mostrando ubicaciones incongruentes que se encuentran fuera de la provincia y hasta fuera del territorio nacional. Para solucionar este problema la técnica de normalización y filtrado de valores demostró ser muy eficiente en la detección de coordenadas fuera del territorio nacional, ya que solo una transformación sobre las variables de LATITUD y LONGITUD, junto con el filtrado de los valores entre -3 y 3 fueron suficientes para identificar todas las coordenadas fuera del territorio nacional. En cuanto a la detección de registros de coordenadas fuera de la provincia se necesitó el apoyo de la librería *geopandas* y el uso de un archivo geoJson con los polígonos de cada provincia, lo que derivó en mayor esfuerzo y tiempo para

identificarlas, aun así, se obtuvo una base de datos donde toda la información de sus variables es congruentes y libre de datos nulos.

Como resultado de la limpieza y transformación de los registros de coordenada las provincias de IBARRA, ESMERALDAS y CARCHI fueron eliminadas del estudio al presentar cero registros de coordenadas congruentes con el resto de información. Esto puede ser un indicio errores al momento de registrar las coordenadas o una falla en las fiscalías de las provincias mencionadas.

3.2.2. Identificación de zonas geográficas por provincia.

Como se puede observar en las investigaciones [34] y [35] la delimitación de zonas geográficas o *hotspot* al momento de investigar patrones de comportamiento delictivo es fundamental. Por lo que se procedió identificar zonas geográficas dentro de cada provincia analizada.

De acuerdo a la literatura revisada, los métodos de clusterización más utilizados y con mejores resultados a la hora de analizar registros de comportamiento criminal son: DBSCAN y K-means, los mismos fueron analizados y probados en la etapa de modelamiento de este trabajo de tesis, dando como resultado que K-means es el mejor método para identificar clusters sobre las coordenadas analizadas. Este resultado se sustenta en la información expuesta en las Tablas 12-15 y las Figuras 12-18, sin embargo, el triunfo de K-means sobre DBSCAN también puede ser justificado de manera empírica. Esto se debe a que DBSCAN en un algoritmo de clusterización basado en densidad y los parámetros que necesita para trabajar consisten en la distancia entre registros vecinos y la cantidad mínima de vecinos que debe existir para definir un cluster. Al ver el comportamiento de los registros de coordenadas en las Figuras 14-16 se puede identificar que en cada provincia existe una densidad de registros diversa, este fenómeno se puede observar hasta en zonas geográficas pequeñas, lo que dificulta la definición de los parámetros para DBSCAN y a pesar del uso de métodos preestablecidos para la definición de sus parámetros, como los mostrados en [47] y [48] los resultados no son satisfactorios.

Por otro lado, las zonas geográficas identificadas con el método de K-means muestran congruencia con el comportamiento geoespacial de los registros, identificando y delimitando zonas con mayor incidencia de robo de manera efectiva y lógica. La calidad de las zonas (clusters) identificados se evalúa en la sección 2.5.1 y se sustenta con la información mostrada en la tabla 24.

3.2.3. NDDs relacionadas

Una vez definidas las zonas geográficas en cada provincia, la extracción de NDDs relacionadas consiste en tres fases que son:

1. Extracción de reglas de asociación maximáles
2. Creación de patrones de robo
3. Reconocimiento de NDDs que siguen los patrones encontrados

Para llevar a cabo la primera fase se necesita crear un itemset con soporte mínimo, el soporte mínimo fue planteado con la ecuación (1) presente en la sección 2.4.2 la cual depende del parámetro *n,* que representa la frecuencia mínima de un ítem en la base de datos. Este parámetro fue encontrado de manera experimental tomando en cuenta la mayor longitud de los ítems generados para distintos valores de *n,* concluyendo que el mejor soporte mínimo se obtiene cuando *n=3*. Definir de esta manera el soporte mínimo evita la perdida de reglas de asociación que pueden derivar en patrones criminales con baja frecuencia dentro de cada zona geográfica.

Debido a que las reglas de asociación con las que se construyen los patrones criminales poseen un soporte bajo, no se puede utilizar esta medida para su evaluación, es por eso que se utilizaron las métricas de: Lift, conviction y Zhang que se en enfocan en la dependencia, correlación, fuerza y causalidad del antecedente con el consecuente. En la sección 2.5.2 se puede observar que las reglas de asociación obtenidas posen correlación, dependencia y causalidad fuertes, lo que nos indica que los patrones generados a partir de las mismas describen de manera efectiva el comportamiento criminal registrado en la base de datos.

Una vez encontradas las reglas de asociación maximáles que pueden construir patrones de comportamiento criminal, debemos asegurarnos que el patrón resultante de las mismas relacione NDDs dentro de la zona geográfica analizada, para esto las reglas de asociación deben cumplir con los criterios expuestos en la sección 2.4.3, dichos criterios permiten identificar de manera efectiva patrones que relacionan al menos 2 NDDs en la zona geográfica. Lo mencionado se sustenta con la información presentada en la Tabla 28 de la sección 2.6.4, donde podemos observar que la media mínima de NDDs relacionadas es de 3 por patrón criminal encontrado, indicando que el método expuesto es efectivo. Además, en la misma tabla destaca el promedio de NDDs relacionadas por patrón en las provincias de Pichincha y Guayas con valores de 23.6 y 8.12 correspondientemente. Estos resultados

nos indican que en la provincia de Pichincha existe una mayor cantidad de crímenes asociados entre sí, bajo el mismo patrón delictivo que en Guayas.

Lamentablemente debido a factores externos a esta investigación como: Protección de datos personales, NDDs ofuscada, carga procesal de la institución y el hecho de que no existe un convenio con la FGE. No se puede realizar un a análisis de efectividad del método propuesto en este estudio, con relato del delito. Limitando la evaluación de los resultados a su visualización y análisis de métricas internas.

4. CONCLUSIONES

- Se revisó el estado del arte de técnicas de minería descriptiva de datos aplicadas en el reconocimiento de patrones delictivos para poder decidir que técnicas resultaban ser las más efectivas
- Se implementó un *dashboard* para el análisis y visualización de los datos otorgados por la FGE junto con los resultados obtenidos en el presente estudio.
- Se revisó y determino las técnicas de minería descriptiva de datos apropiadas para descubrir noticias de delito relacionadas.
- Se identificaron noticias del delito relacionadas a través de un patrón de robo dentro de los registros otorgados por la FGE.
- Se utilizó la metodología CRISP-DM para la creación de un modelo que identifique noticas de delito relacionadas en los datos otorgados por FGE.
- Se realizó la limpieza y transformación de los datos otorgados por la FGE, dando como resultado que los datos de las provincias de Esmeraldas, Imbabura y Carchi no poseen coordenadas congruentes con el resto de información de su registro.
- Se determinó que la aplicación de *K-means* para identificar zonas geográficas significativas en cada provincia es más efectiva que DBSCAN
- Se determinó que el algoritmo *FP-Growth* obtiene las mismas reglas de asociación que el algoritmo *Apriori*, pero con mejor tiempo de ejecución.
- Se identificó que el soporte mínimo óptimo para la extracción de reglas de asociación que derivan patrones de robo es $\frac{3}{N}$ donde N es la cantidad de registros dentro de cada zona geográfica analizada.
- Se comprobó que si un conjunto de reglas de asociación maximales con el mismo soporte, cumplen con los siguientes criterios expuestos en la sección 2.4.3, pueden formar un patrón que relaciona noticias de delito dentro de la zona geográfica analizada.
- La identificó que la media a nivel nacional de noticias de delito relacionadas es de 5,41 noticias por patrón de delito identificado.

Recomendaciones

- Realizar un convenio con la FGE con el objetivo de verificar los resultados del modelo, mediante la revisión del archivo fiscal. De esta manera poder calcular su precisión para identificar NDDs relacionadas.
- Pedir acceso al relato del delito para extraer información del mismo y así reentrenar el modelo con más variables. De esta manera obtener patrones de robo más específicos que mejorarán la precisión del modelo para identificar NDDs relacionadas.
- Desplegar el modelo en la FGE y evaluar sus resultados con nuevas NDDs.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] "State Attorney General's Office | Analyze theft figures". Consultado: el 26 de noviembre de 2023. [En línea]. Disponible en: https://www.fiscalia.gob.ec/analitica-cifras-de-robo/

[2] R. Í. Villagómez Cabezas, "El rol del fiscal en el procedimiento penal abreviado", 2008, Consultado: el 27 de noviembre de 2023. [En línea]. Disponible en: http://repositorio.uasb.edu.ec/handle/10644/484

[3] "R. Oficial Suplemento, "Código Orgánico Integral Penal, COIP", Consultado: el 27 de junio de 2024. [En línea]. Disponible en: www.lexis.com.ec

[4] "Manual de conceptualización de indicadores de seguridad ciudadana y convivencia pacífica desde el enfoque de la prevención. - PDF Free Download". Consultado: el 27 de junio de 2024. [En línea]. Disponible en: https://docplayer.es/59063202-Manual-de-conceptualizacion-de-indicadoresde-seguridad-ciudadana-y-convivencia-pacifica-desde-el-enfoque-de-la-prevencion.html

[5] G. Shmueli, P. C. Bruce, I. Yahav, N. R. (Nitin R. Patel, y K. C. Lichtendahl, *Data mining for business analytics : concepts, techniques, and applications in R*.

[6] A. Nasridinov, J. Y. Byun, N. Um, y H. S. Shin, "Application of data mining for crime analysis", en *Lecture Notes in Electrical Engineering*, Springer Verlag, 2016, pp. 503–508. doi: 10.1007/978-3-662-47895-0_61.

[7] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, y I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective", *Symmetry*, vol. 15, núm. 9. Multidisciplinary Digital Publishing Institute (MDPI), el 1 de septiembre de 2023. doi: 10.3390/sym15091679.

[8] N. Jain y V. Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER". [En línea]. Disponible en: http://www.ijret.org

[9] F. Siraj y M. A. Abdoulha, "4 Mining Enrolment Data Using Predictive and Descriptive Approaches". [En línea]. Disponible en: www.intechopen.com

[10] M. Ahmed, "Data summarization: a survey", *Knowl Inf Syst*, vol. 58, núm. 2, pp. 249–273, feb. 2019, doi: 10.1007/s10115-018-1183-0.

[11] J. Han, M. Kamber, y J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)", 2011.

[12] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[13] U. Hahn y I. Mani, "The challenges of automatic summarization", *Computer (Long Beach Calif)*, vol. 33, núm. 11, pp. 29–36, 2000.

[14] Z. R. Hesabi, Z. Tari, A. Goscinski, A. Fahad, I. Khalil, y C. Queiroz, “Data summarization techniques for big data-a survey”, en *Handbook on Data Centers*, Springer New York, 2015, pp. 1109–1152. doi: 10.1007/978-1-4939-2092-1_38.

[15] J. Oyelade *et al.*, “Data Clustering: Algorithms and Its Applications”, en *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, Institute of Electrical and Electronics Engineers Inc., jul. 2019, pp. 71–81. doi: 10.1109/ICCSA.2019.000-1.

[16] “Dendrograma - Minitab”. Consultado: el 17 de enero de 2024. [En línea]. Disponible en: https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistical-modeling/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/

[17] H. Pirim, *Recent Applications in Data Clustering*. Rijeka: IntechOpen, 2018. doi: 10.5772/intechopen.71315.

[18] N. and T. S. M. Bagirov Adil and Karmitsa, “Heuristic Clustering Algorithms”, en *Partitional Clustering via Nonsmooth Optimization: Clustering via Optimization*, Cham: Springer International Publishing, 2020, pp. 135–163. doi: 10.1007/978-3-030-37826-4_5.

[19] “Clusters de partición”. Consultado: el 19 de enero de 2024. [En línea]. Disponible en: https://rstudio-pubs-static.s3.amazonaws.com/974095_b91a4b6ea3aa4fa397755fb2814492eb.html

[20] M. Tareq, E. A. Sundararajan, A. Harwood, y A. A. Bakar, “A Systematic Review of Density Grid-Based Clustering for Data Streams”, *IEEE Access*, vol. 10. Institute of Electrical and Electronics Engineers Inc., pp. 579–596, 2022. doi: 10.1109/ACCESS.2021.3134704.

[21] M. Derdour, M. Ahmim, A. Benjedou, Jāmiʻat ʻAnnābah, Institute of Electrical and Electronics Engineers. Algeria Section, y Institute of Electrical and Electronics Engineers, *Proceedings, ICNAS 2019 : 4th International Conference on Networking and Advanced Systems : 26-27 June 2019*.

[22] S. K. Solanki y J. T. Patel, “A survey on association rule mining”, en *International Conference on Advanced Computing and Communication Technologies, ACCT*, Institute of Electrical and Electronics Engineers Inc., abr. 2015, pp. 212–216. doi: 10.1109/ACCT.2015.69.

[23] Ubon Thongsatapornwatana, “A Survey of Data Mining Techniques for Analyzing Crime Patterns”, *2016 Second Asian Conference on Defence Technology (ACDT)*, 2016.

[24] S. M. Ghafari y C. Tjortjis, “A survey on association rules mining using heuristics”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, núm. 4. Wiley-Blackwell, el 1 de julio de 2019. doi: 10.1002/widm.1307.

[25] M. Rahmany, A. Mohd Zin, y E. A. Sundararajan, “COMPARING TOOLS PROVIDED BY PYTHON AND R FOR EXPLORATORY DATA ANALYSIS”.

[26] C. F. Mohd Foozy, R. Ahmad, M. A. Faizal Abdollah, y C. C. Wen, “A Comparative Study with RapidMiner and WEKA Tools over some Classification Techniques for SMS Spam”, en *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, ago. 2017. doi: 10.1088/1757-899X/226/1/012100.

[27] “Bokeh documentation — Bokeh 3.4.1 Documentation”. Consultado: el 7 de mayo de 2024. [En línea]. Disponible en: https://docs.bokeh.org/en/latest/

[28] “Matplotlib — Visualization with Python”. Consultado: el 7 de mayo de 2024. [En línea]. Disponible en: https://matplotlib.org/

[29] “Python | Visualize missing values (NaN) values using Missingno Library - GeeksforGeeks”. Consultado: el 7 de mayo de 2024. [En línea]. Disponible en: https://www.geeksforgeeks.org/python-visualize-missing-values-nan-values-using-missingno-library/

[30] “Journal of Open Source Software: MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack”. Consultado: el 7 de mayo de 2024. [En línea]. Disponible en: https://joss.theoj.org/papers/10.21105/joss.00638

[31] “pandas - Python Data Analysis Library”. Consultado: el 7 de mayo de 2024. [En línea]. Disponible en: https://pandas.pydata.org/about/

[32] Rohit Vishwakarma y Nikhilesh Yadav, *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017) : date: 20,21,22, April 2017*.

[33] X. Alphonse Inbaraj y A. Seshagiri Rao, *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies : 01-03, March 2018*.

[34] S. Mangara Wainana, J. Njuguna Karomo, R. Kyalo, y N. Mutai, “Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya”, *International Journal of Data Science and Analysis*, vol. 6, núm. 1, p. 20, 2020, doi: 10.11648/j.ijdsa.20200601.13.

[35] C. Catlett, E. Cesario, D. Talia, y A. Vinci, “Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments”, *Pervasive Mob Comput*, vol. 53, pp. 62–74, feb. 2019, doi: 10.1016/j.pmcj.2019.01.003.

[36] O. E. Isafiade y A. B. Bagula, “CitiSafe: Adaptive spatial pattern knowledge using Fp-growth algorithm for crime situation recognition”, en *Proceedings - IEEE 10th International Conference on Ubiquitous Intelligence and Computing, UIC 2013 and IEEE 10th International Conference on Autonomic and Trusted Computing, ATC 2013*, 2013, pp. 551–556. doi: 10.1109/UIC-ATC.2013.72.

[37] Z. Zhang, J. Huang, J. Hao, J. Gong, y H. Chen, “Extracting relations of crime rates through fuzzy association rules mining”, *Applied Intelligence*, vol. 50, núm. 2, pp. 448–467, feb. 2020, doi: 10.1007/s10489-019-01531-3.

[38] S. Yazar, / Corresponding, D. Çalişkan, K. Yildiz, B. Doğan, y A. Aktaş, “Crime Data Analysis with Association Rule Mining Bi rli kteli k Kural Çikarimi ile Suç Veri Analizi”, doi: 10.292228/porta.1.

[39] S. K. Pani *et al.*, “1 Crime Pattern Detection Using Data Mining”, 2021.

[40] “CRISP-DM: La metodología para poner orden en los proyectos - Sngular”. Consultado: el 22 de noviembre de 2023. [En línea]. Disponible en: https://www.sngular.com/es/data-science-crisp-dm-metodologia/

[41] J. Clinton, “CRISP-DM 1.0 Step-by-step data mining guide”, DaimlerChrysler, 1999.

[42] “Código Orgánico de la Función Judicial Actualizado 2024”. Consultado: el 12 de junio de 2024. [En línea]. Disponible en: https://www.lexis.com.ec/biblioteca/codigo-organico-funcion-judicial

[43] “provincias”. Consultado: el 18 de mayo de 2024. [En línea]. Disponible en: https://sgr-ecuador.carto.com/tables/provincias/public

[44] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, y H. H. R. Sherazi, “Spatiooral crime hotspot detection and prediction: A systematic literature review”, *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 166553–166574, 2020. doi: 10.1109/ACCESS.2020.3022808.

[45] “Ejemplo de uso de DBSCAN en Python para eliminación de outliers – Exponentis”. Consultado: el 11 de abril de 2024. [En línea]. Disponible en: http://exponentis.es/ejemplo-de-uso-de-dbscan-en-python-para-deteccion-de-outliers

[46] Tara Mullin, “DBSCAN Parameter Estimation Using Python | by Tara Mullin | Medium”. Consultado: el 11 de abril de 2024. [En línea]. Disponible en: https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

[47] M. Ester, H.-P. Kriegel, J. Sander, y X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, 1996. [En línea]. Disponible en: www.aaai.org

[48] “Elbow Method for optimal value of k in KMeans - GeeksforGeeks”. Consultado: el 20 de mayo de 2024. [En línea]. Disponible en: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

[49] D. Arthur y S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding”.

[50] Sunil Yadav, Meet Timbadia, y Ajit Yadav, *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017) : date: 20,21,22, April 2017*.

[51] “Enajenación mental. Diccionario médico. Clínica Universidad de Navarra.” Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://www.cun.es/diccionario-medico/terminos/enajenacion-mental

[52] “Capacidad volitiva y cognitiva | VIU España”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://www.universidadviu.com/es/actualidad/nuestros-expertos/capacidad-volitiva-y-cognitiva

[53] “Urbano - Qué es, objetivos, definición y concepto”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://definicion.de/urbano/

[54] “Rural - Qué es, definición y concepto”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://definicion.de/rural/

[55] “¿Conoces las etapas del Procedimiento Penal Ordinario? | FEXLAW”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://fexlaw.com/boletin-legal/conoces-las-etapas-del-procedimiento-penal-ordinario/

[56] “Definición de impugnación - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/impugnaci%C3%B3n

[57] “Definición de delito consumado - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/delito-consumado

[58] “Definición de tentativa de delito - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/tentativa-de-delito

[59] “Definición de delito flagrante - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/delito-flagrante

[60] “escruche | Diccionario de americanismos | ASALE”. Consultado: el 27 de febrero de 2024. [En línea]. Disponible en: https://www.asale.org/damer/escruche

[61] “¿Quiénes son los sacapintas y cómo proteger tu dinero?” Consultado: el 1 de marzo de 2024. [En línea]. Disponible en: https://www.pichincha.com/blog/sacapintas-como-protegerte

[62] “Definición de escalamiento - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 1 de marzo de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/escalamiento

[63] “Mejoran las medidas de seguridad ante una nueva modalidad de robos | Comunidad | Guayaquil | El Universo”. Consultado: el 1 de marzo de 2024. [En línea]. Disponible en: https://www.eluniverso.com/noticias/2016/08/11/nota/5736009/mejoran-medidas-seguridad-ante-nueva-modalidad-robos/

[64] “«bujiazo», término válido | FundéuRAE”. Consultado: el 1 de marzo de 2024. [En línea]. Disponible en: https://www.fundeu.es/recomendacion/bujiazo-bujiero/

[65] “Definición de arma blanca - Diccionario panhispánico del español jurídico - RAE”. Consultado: el 1 de marzo de 2024. [En línea]. Disponible en: https://dpej.rae.es/lema/arma-blanca

[66] “Categoría:Armas contundentes - Wikipedia, la enciclopedia libre”. Consultado: el 2 de marzo de 2024. [En línea]. Disponible en: https://es.wikipedia.org/wiki/Categor%C3%ADa:Armas_contundentes

6. ANEXOS

Anexo I

Glosario de términos encontrados en la base de datos.

- **Madrugada:** Se refiere al intervalo entre las 00:00:00 y las 05:59:00 horas.
- **Mañana:** Se refiere al intervalo entre las 06:00:00 y las 11:59:55 horas.
- **Tarde:** Se refiere al intervalo entre las 12:00:00 y las 17:59:59 horas.
- **Noche:** Se refiere al intervalo entre las 18:00:00 y las 23:59:21 horas.
- **Enajenación mental:** De acuerdo con [51] el termino describe un estado de encontrase fuera de sí, perturbado en el uso de la razón o ajeno a sí mismo.
- **Capacidad volitiva:** Capacidad de una persona de actuar en base a su comprensión, habilidad que otorga a la persona de aceptar o rechazar una inclinación y controlar sus actos [52].
- **Urbano:** Referente a la ciudad o área con alta densidad poblacional cuyos habitantes no se dedican a actividades agrícolas [53].
- **Rural:** Opuesto a lo urbano, se refiere a la vida de campo o en zonas donde se practica actividad agrícola y tienen baja densidad poblacional [54].
- **Investigación previa:** Fase previa al proceso judicial penal que tiene como finalidad descubrir y manifestar un motivo suficiente para poder abrir el propio proceso penal [55]**.**
- **Instrucción fiscal:** Etapa en la que el fiscal presenta elementos que probatorios que permitan al poder judicial conocer sobre la responsabilidad del acusado en la comisión del delito [55]**.**
- **Preparatoria de juicio:** se refiere a la *Etapa de Evaluación y Preparatoria de Juicio*, en la misma el juez establece la validez proceso, valora y evalúa los elementos de en qué se sustenta la acusación fiscal [55]**.**
- **Impugnación:** Formalización de un recurso contra una resolución administrativa o judicial [56].
- **Delito consumado:** Delito que ha sido realizado por completo [57].
- **Tentativa:** Se refiere a “tentativa de delito”: Acción puesta por el sujeto con el objetivo de cometer un delito, pero que no se consuma por causas ajenas a su voluntad [58]**.**

- **Flagrante:** Evidente, que no admite refutación, la base de datos hace referencia a *Delito flagrante:* Delito que se comete cuando el delincuente es sorprendido en el momento de cometer la infracción. Se produce no solo cuando el delincuente es detenido en el momento de cometer el delito, sino también cuando es detenido o perseguido inmediatamente después de consumado este [59].
- **No flagrante:** Contrario a delito flagrante.
- **Estruche:** Robo en una vivienda después de haber forzado los accesos a ella y generalmente en ausencia de los propietarios [60]**.**
- **Sacapintas:** Bandas criminales organizadas que se infiltran en una agencia bancaria para identificar a clientes que retiran altas sumas de dinero para después asaltarlas [61].
- **Escalamiento:** Medio utilizado para realizar el robo con fuerza en las cosas, que consiste en ascender o trepar para acceder o abandonar el lugar donde se encuentra la cosa mueble ajena que es objeto de sustracción [62].
- **Foramen:** Modalidad de robo que consiste en crear huecos en cerramientos para robar el interior de las propiedades [63].
- **Paquetazo:** Timo, hurto, especialmente el que consiste en engañar a alguien con un vigésimo de la lotería grande supuestamente premiado o con un fajo de billetes, la mayoría falsos.
- **Robo express:** En contexto de la base de datos se trata del robo que sufre una víctima de secuestro exprés.
- **Bujiazo:** Técnica delictiva que consiste en romper las lunas de los vehículos con una bujía para robar a sus conductores y pasajeros [64].
- **Arma blanca:** Arma constituida por una hoja metálica u otro material de características físicas semejantes, cortante o punzante [65].
- **Arma constrictora:** Arma utilizada para estrangular como: cuerdas, alambres, cables, etc.
- **Arma contundente:** Es un instrumento o herramienta que fue diseñada con el fin de golpear y permite atacar o defenderse puede lastimar físicamente o hasta matar a otra persona, puede distinguirse por su aspecto físico, suele confundirse con el termino objeto contundente [66].

Anexo II

Dashboard para la visualización de resultados y análisis de datos otorgados por la FGE

Visualización.pbix