Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Tecnologías de la Información y Comunicación

DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA

Ing. Mario Rolando Lara Cabrera
Asesorado por el Lic. Msc. Oscar Daniel Edelmann Amaya

Guatemala, noviembre de 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. MARIO ROLANDO LARA CABRERA
ASESORADO POR EL LIC. MSC. OSCAR DANIEL EDELMANN AMAYA

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

GUATEMALA, NOVIEMBRE DE 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Christian Moisés de la Cruz Leal
VOCAL V	Br. Kevin Armando Cruz Lorente
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Ing. Marlon Antonio Pérez Türk
EXAMINADOR	Ing. Edwin Estuardo Zapeta
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado con fecha 16 de marzo de 2019.

Ing. Mario Rolando Lara Cabrera

Decanato
Facultad de Ingeniería
24189101 - 24189102

DTG. 443.2020.

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Postgrado, al Trabajo de Graduación titulado: **DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA,** presentado por el Ingeniero **Mario Rolando Lara Cabrera,** estudiante de la **Maestría en Tecnologías de la Información y Comunicación** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Anabela Cordova Estrada
Decana

Guatemala, noviembre de 2020.

AACE/asga

Guatemala, Noviembre de 2020

EEPFI-1548-2020

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen y verificar la aprobación del Revisor y la aprobación del Área de Lingüística al Trabajo de Graduación titulado: "**DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA**" presentado por el Ingeniero **Mario Rolando Lara Cabrera** quien se identifica con Carné **009517011** correspondiente al programa de **Maestría en Artes en Tecnologías de la Información y la Comunicación** apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

Guatemala, Noviembre de 2020

Como Coordinador de la **Maestría en Artes en Tecnologías de la Información y Comunicación** doy el aval correspondiente para la aprobación del Trabajo de Graduación titulado: **"DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA"** presentado por el Ingeniero **Mario Rolando Lara Cabrera** quien se identifica con Carné **009517011.**

Atentamente,

"Id y Enseñad a Todos"

MARLON ANTONIO PEREZ TURK
INGENIERO EN CIENCIAS Y SISTEMAS
COLEGIADO No. 4492

Mtro. Ing. Marlon Antonio Pérez Turk
Coordinador de Maestría
Escuela de Estudios de Postgrado
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

Guatemala, Noviembre de 2020

EEPFI-1549-2020

En mi calidad como Asesor del Ingeniero **Mario Rolando Lara Cabrera** quien se identifica con Carné **009517011** procedo a dar el aval correspondiente para la aprobación del Trabajo de Graduación titulado: **"DESARROLLO DE UN PROTOTIPO QUE PERMITA REALIZAR EL ANÁLISIS PREDICTIVO DE DELITOS DE INVESTIGACIÓN CRIMINAL EN EL DEPARTAMENTO DE GUATEMALA"** quien se encuentra en el programa de **Maestría en Artes en Tecnologías de la Información y la Comunicación** en la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala.

Atentamente,

"Id y Enseñad a Todos"

MSc. Lic. Oscar Daniel Edelmann Amaya
Asesor

M. Sc. Oscar Daniel Edelmann Amaya
Col. 21,755

ACTO QUE DEDICO A:

Dios	Por haber colocado dentro de su propósito para mi vida este peldaño que hoy alcanzo. Gracias por su amor y sus dones inefables.
Mi esposa	Por ser esa ayuda idónea durante todo este tiempo para motivarme a finalizar cada etapa de esta meta.
Mis hijos	Para que les sirva de motivación para nunca dejar de aprehender y superarse en conocimientos y puedan aplicarlos en su vida.

AGRADECIMIENTOS A:

Dios	Por la sabiduría, inteligencia y conocimiento que me ha dado, así como el ánimo y aliento para no rendirme y alcanzar cada meta propuesta.
Mi asesor	Por su colaboración en el desarrollo de este trabajo y la motivación necesaria para finalizarlo.
Mis catedráticos	Por todo el conocimiento compartido durante las clases, lo cual fue parte importante para aplicarlo en el desarrollo de este trabajo.
Mis compañeros	Por ser miembros de una comunidad de aprendizaje, donde todos compartimos los conocimientos y las experiencias y nos ayudamos para alcanzar este triunfo.
Universidad de San Carlos de Guatemala	Por la calidad educativa que se mantiene en los salones de clases y por abrir la oportunidad para la preparación académica, la formación y acreditación que se necesita en las áreas de tecnología.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SIMBOLOS

Símbolo	Significado
CGI	Common Gateway Interface
COLLECTION+*JSON*	Formato de hipermedia basado en Json
CSV	Comma Separated Values
ETL	Extract, Transform, Load
HOM	Delitos contra la vida e integridad de las personas
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
K	Número de cluster
KDD	Knowledge Discovery in Databases
PAT	Delitos patrimoniales
REST, RESTFUL	Representational State Transfer
SEX	Delitos sexuales
SIREN	Formato de hipermedia

GLOSARIO

Análisis criminal	Conjunto de tareas en las que se analiza la información y son realizadas por las autoridades policiales con la finalidad de esclarecer los hechos delictivos.
API	Application Programming Interface
Caracterización del crimen	Establecimiento de patrones relacionados a los hechos delictivos y los modos de operación de los criminales.
Criminología	Ciencia que estudia a los delincuentes, el lugar de los hechos, el delito, las conductas desviadas, el control social, con relación al delito mismo, y la víctima, con el objetivo de entender al criminal y las distintas motivaciones que lo llevaron a cometer determinados crímenes.
***Clustering* jerárquico**	Tipo de agrupamiento que consiste en utilizar niveles asociativos por medio de jerarquías encontradas en la información que es sujeta de estudio.
DataSet	Conjunto de datos

DEIC	División Especializada en Investigación Criminal
DOCKER	Contenedor de Software
DOCKER-COMPOSE	Componente de *DOCKER* para construir aplicaciones y microservicios
Flask	*Framework* de desarrollo en lenguaje Python.
FRONT-END	Interfaz de usuario frontal.
Fuzzy C	Algoritmo de agrupamiento que utiliza en su diseño lógica difusa y que busca evitar que un miembro pueda pertenecer a varios grupos.
GIS	Sistema de información georreferencial.
HAL	Hypertext Application Language
Hot-Spot	Puntos calientes que representan las áreas más propensas o dónde se comenten más hechos delictivos.
Índice de Bouldin	Métrica de evaluación a nivel interno para los algoritmos de agrupamiento.
JEPEDI	Jefatura de planificación estratégica de la Policía Nacional Civil

Kmeans	Algoritmo de agrupamiento que clasifica objetos en k grupos basándose en sus características y se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o conglomerado.
Leatflet	Interface de JavaScript interactiva para visualización georreferencial
Machine Learning	Conjunto de herramientas de inteligencia artificial orientadas al aprendizaje automático.
Matplot-Lib	Biblioteca desarrollada en lenguaje Python para la generación de gráficos, utilizando como fuente de datos matrices o listas que contienen la información.
Modus operandi	Modo de operar del delito, el delincuente o las estructuras criminales.
Móvil	La causa o la hipótesis que se cree llevaron al delincuente a cometer el delito.
Numpy	Biblioteca desarrollada en lenguaje Python utilizada para el manejo de vectores y matrices.
Pandas	Biblioteca desarrollada en lenguaje Python utilizada para el manejo de archivos y conjuntos de datos.

PNC	Policía Nacional Civil de Guatemala.
Predicción del crimen	Conjunto de técnicas de análisis cuantitativo para identificar objetivos potenciales que requieren la intervención policial utilizando la información de crímenes pasados por medio de los pronósticos estadísticos.
Red de Bayes	Modelo probabilístico que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un gráfico a cíclico dirigido.
Scikit-Learn	Biblioteca desarrollada en lenguaje Python utilizada para implementar algoritmos de aprendizaje automático.
SEICMSJ	Secretaría Ejecutiva de la Instancia Coordinadora de la Modernización del Sector Justicia.
STCNS	Secretaría Técnica del Consejo Nacional de Seguridad.
TI	Tecnologías de la información.
UML	Lenguaje de modelado unificado.

RESUMEN

Los grandes índices de violencia en los delitos contra la vida, trata de personas, índole sexual, robos, hurtos, extorsiones, entre otros, causan desasosiego en la sociedad guatemalteca. Actualmente, la Policía Nacional civil no cuenta con los recursos humanos, tecnológicos y de competencias especializadas en áreas relacionadas al análisis e inteligencia del crimen; realizan de manera muy escaza las tareas de prevención, por lo que su visión queda enfocada únicamente a la persecución y captura de los delincuentes que comenten los delitos.

Por esta razón, en el presente trabajo se realizó la investigación, el desarrollo y la implementación de un prototipo de clasificación y predicción para los delitos contra la vida e integridad de las personas, de índole sexual y de índole patrimonial (robos y hurtos), el cual puede ser utilizada como herramienta para realizar el proceso de análisis de criminalidad de manera optimizada por las unidades de análisis de información de los departamentos de investigación criminal de la DEIC.

Considerando qué, el análisis criminal y la predicción del delito, es un enfoque sistemático (Omkar, Sayak, Raj, Suraj y Rohini, 2018), se implementó un proceso de minería de datos, utilizando algoritmos de agrupamiento jerárquico(Clustering jerárquico), Kmeans y Difuso (Fuzzy C); permite al analista, la identificación de patrones y características de los delitos, agrupándolos en conglomerados, los cuales desde un punto de vista criminológico, tienen sentido, si se analiza dentro del contexto de las áreas de incidencia criminal. Para realizar

este agrupamiento, se realizó la identificación y preparación de conjuntos de datos (Datasets) históricos (años 2017-2019), que contienen las variables necesarias que se relacionan con los hechos criminales que son sujetos de estudio y que son cargados al sistema como insumos para la clasificación y agrupamiento del crimen en los municipios y zonas del departamento de Guatemala.

Utilizando la herramienta, el analista puede caracterizar el crimen y probar con diferentes conjuntos de datos, verificando la precisión de resultados, mediante el índice de Bouldin (Bouldin, 1979), pero también puede realizar una visualización georreferencial, clasificada por conglomerados e identificados por un color dentro del mapa, para una mejor comprensión. Con este análisis, podrá cotejar los resultados con otros insumos del análisis criminal, por ejemplo: presencia de grupos criminales en el lugar, falta de presencia policial, entre otros y optimizar los modelos con sus variables y realizar la predicción de puntos calientes (*Hot-Spot*), puesto que, el sistema asigna un identificador de clúster (categoría o feature) a cada delito del conjunto de datos (que fue sujeto de estudio), mediante el algoritmo Gaussian Naive Bayes (red de Bayes), con el cual se logra un nivel de precisión de más del 90 %, y analizarlos en un mapa del departamento de Guatemala. La información de este análisis servirá a la PNC en la elaboración de estrategias de prevención del delito en el departamento de Guatemala.

Para lograr un mayor desacoplamiento y escalabilidad en el desarrollo del sistema, se utilizó una arquitectura de microservicios, mediante una capa *API REST*, desarrollada con las herramientas libres (*Open Source*) Scikit-Learn, Matplot-lib, Numpy de Python, entre otras, que soportan los algoritmos de minería de datos y aprendizaje automático y que pueden ser implementadas en un servidor CGI en cualquier sistema operativo, de preferencia Linux, con la finalidad

de ahorrar costos de licenciamiento mediante la adquisición de software propietario y que este no sea un impedimento para su implementación.

PLANTEAMIENTO DEL PROBLEMA Y FORMULACIÓN DE PREGUNTAS ORIENTADORAS

Una de las necesidades que afronta la División Especializada en Investigación Criminal (DEIC), es la escasa inteligencia criminal basada en el análisis de la información, específicamente en materia de la predicción de delitos, esto dotaría a la DEIC con información para adoptar medidas preventivas. Actualmente, la DEIC no se da abasto con la inmensa cantidad de casos de investigación a los que les tienen que dar seguimiento, según los altos índices de criminalidad de acuerdo al reporte estadístico 2017 realizado por la Secretaría Técnica del Consejo Nacional de Seguridad, donde se ve un incremento en los hechos violentos en el departamento de Guatemala (Secretaría Técnica del Consejo Nacional de Seguridad de Guatemala [STCNS], 2017), por lo que el recurso humano es dedicado a realizar estos esfuerzos, descuidando así la inteligencia criminal.

> El análisis y la predicción del delito es un enfoque sistemático utilizado para analizar e identificar diferentes patrones, relaciones y tendencias en la delincuencia. La predicción visualiza las regiones con alta probabilidad de ocurrencia de delitos e indica aquellas áreas propensas a la delincuencia (Omkar, Sayak, Raj, Suraj y Rohini, 2018, p.4).

Los sistemas de minería de datos proveen diferentes algoritmos para llevar a cabo la clasificación, estudios relacionados a la predicción del crimen en países como Estados Unidos, Colombia y España (Barreras, Díaz, Riascos y Ribero, 2016), demuestran que entre los más precisos se encuentran: K-means (partición

de conjunto de n observaciones en k grupos), Fuzzy c (Clustering difuso), Clustering jerárquico, red bayesiana.

La evaluación de los algoritmos de clasificación se puede realizar de acuerdo con los siguientes aspectos: 1) precisión (porcentaje de casos clasificados correctamente), 2) eficiencia, 3) robustez, 4) escalabilidad. Para lograr altos niveles de precisión y exactitud en la predicción, estos algoritmos consumen grandes cantidades de recursos de hardware (memoria, disco y procesamiento) en las fases de clasificación y entrenamiento; por lo que se tiene que buscar una arquitectura que sea escalable de manera horizontal o vertical, para evitar el problema de la falta de recursos y minimizar el tiempo de procesamiento.

> La diferencia en la precisión y la eficiencia de los diferentes algoritmos puede ser explicada de la manera siguiente: para el caso de Clustering jerárquico, la diferencia se da por el tamaño de los segmentos de datos que permiten identificar puntos calientes (*Hot-Spot*), si la agrupación de los datos es granular, se hace difícil identificar los clústeres (i.e., agrupaciones) de crimen, porque la visualización tendrá demasiadas zonas demarcadas como puntos calientes. Por otro lado, si los datos se agrupan de acuerdo a diferentes criterios geográficos habrá un problema de unidad de área modificable, en el que las estadísticas resultantes son altamente sensibles a la escogencia arbitraria de los límites de agrupación de los datos (Barreras, Díaz, Riascos y Ribero, 2016, p.20).

En el caso de K-Means y Fuzzy C existe menos precisión cuando no se utiliza un conjunto de datos ideal para el entrenamiento y es más preciso, cuando se utiliza un conjunto de datos más amplio (Omkar, Sayak, Raj, Suraj y Rohini, 2018); para red bayesiana en promedio, el 80 % de las características de los

delitos se predicen correctamente de acuerdo al perfil criminal y dado que cada predicción se acompaña por un nivel de confianza que es proporcional a la precisión esperada, al considerar solo predicciones con altos niveles de confianza, la exactitud promedio aumenta a 95.6 % (Baumgartner, Ferrari y Palermo, 2008).

Por lo tanto, se deben realizar diferentes análisis comparativos entre los distintos modelos de predicción del crimen según las diferencias de tipologías y estructuras criminales que operan en el país. La información estructurada almacenada actualmente en el sistema de información de documentación de casos de investigación criminal de la DEIC, será tomada para los conjuntos de datos, según el modelo de clasificación y las variables que incluye, esto determinará la precisión de los modelos, proveyendo una herramienta para la predicción del delito.

El estudio busca responder las siguientes preguntas:

- Pregunta central

¿Qué sistema de información permite realizar el proceso de análisis de criminalidad de manera optimizada?

- Preguntas auxiliares

    - ¿Cuál es el proceso de preparación de los datos granulares de criminalidad para garantizar la calidad de los datos y asegurar resultados precisos?

- ¿Cuál es el modelo predictivo más preciso utilizando K-means, Fuzzy C, Clustering jerárquico y red bayesiana?

- ¿Cuál es la arquitectura de sistemas que permita la reducción de tiempo en el procesamiento de la información para la predicción de delitos criminales?

OBJETIVOS

General

Desarrollar un prototipo de sistema de información que permita realizar el proceso de análisis de criminalidad de manera optimizada, en el departamento de Guatemala.

Específicos

- Documentar el proceso de preparación de los datos granulares de criminalidad para garantizar la calidad de los datos que se utilizarán de insumo para la predicción del delito.

- Documentar el proceso de construcción y entrenamiento que permita evaluar y determinar la precisión de los modelos predictivos utilizando K-means, Fuzzy C, Clustering jerárquico y red bayesiana.

- Diseñar y desarrollar los componentes tecnológicos de la arquitectura que permita la reducción del tiempo en el procesamiento de la información para la predicción de delitos criminales.

MARCO METODOLÓGICO

- Tipo de estudio
    - Mixto

Se utilizaron las características de la metodología cualitativa con la finalidad de estudiar el fenómeno del delito que opera en el departamento de Guatemala, para describir los patrones sobre la conducta de la criminalidad y la incidencia criminal en ciertas áreas geográficas. Con esto se buscó observar aquellas variables asociadas a los patrones criminales que servirán para el diseño de modelos predictivos, dichas variables pueden ser tipo cualitativo, como cuantitativo.

De la misma manera se utilizó la metodología cuantitativa para analizar las variables de los modelos predictivos propuestos y las relaciones que existen entre los mismos. Utilizando algoritmos matemáticos estadísticos de agrupamiento como: K-MEANS, FUZZY C, Clustering jerárquico; así como de clasificación como: red de Bayes.

Estos algoritmos utilizan conjuntos de datos asociados a los delitos contra la vida e integridad de las personas, de índole sexual, así como de índole patrimonial: robos y hurtos, los cuales fueron cometidos en el departamento de Guatemala durante los años 2017, 2018 y 2019, con la finalidad de establecer los niveles de precisión de los modelos predictivos propuestos y así obtener un prototipo que pueda predecir el delito con un porcentaje mínimo de error.

- Diseño
    - Experimental

Se realizó un diseño de investigación experimental por fases, desarrollando un prototipo de sistema de información utilizado para predecir la probabilidad de ocurrencia del delito en el departamento de Guatemala de manera precisa.

Inicialmente se diseñaron los modelos predictivos con sus variables independientes y dependientes para identificar y describir las relaciones que existen en determinado hecho delictivo que se pretende predecir, en un área con incidencia delincuencial. En este proceso se identificaron los conjuntos de datos que contienen la información de las variables asociadas a los modelos, los cuales fueron proporcionados por la DEIC y la Unidad GIS, de la base de datos histórica donde se registra la información de los casos de investigación criminal del departamento de Guatemala durante los años 2017, 2018 y 2019

Posteriormente, se desarrolló un prototipo para la definición, configuración y parametrización de los modelos predictivos diseñados y sus variables; así como la ejecución de los algoritmos: K-MEANS, FUZZY C, Clustering jerárquico y red de Bayes para realizar la clasificación del delito y la identificación de patrones criminales para tener un mecanismo de predicción, utilizando la información de los conjuntos de datos seleccionados.

Con el prototipo desarrollado se procedió a realizar la experimentación de los modelos predictivos y sus variables en los cuatro algoritmos propuestos según los siguientes objetivos:

- Afinar los modelos diseñados: Es decir verificar el grado de dependencia que existe entre las variables y la posibilidad de exclusión y agregación de otras.

- Optimizar el rendimiento en el procesamiento de los componentes de la arquitectura de sistemas propuesta.

Después del proceso de experimentación se analizaron los resultados obtenidos y se procedió a realizar una clasificación de los algoritmos según los niveles de precisión obtenidos durante la ejecución de la predicción, cabe notar, que cada algoritmo, tiene diferente propósito. Buscando la precisión y adecuado funcionamiento, cuando son sometidos a diferentes circunstancias en la variación de las variables dependientes de los modelos predictivos.

- Alcance del estudio

    - Explicativo

Se realizó un estudio explicativo de las variables relacionadas a la predicción del delito en zonas con índice delincuencial en el departamento de Guatemala, así como también de los niveles de precisión de los algoritmos K-MEANS, FUZZY C, Clustering jerárquico y red de Bayes, con la finalidad de obtener un prototipo de sistema de información que pueda predecir la probabilidad de ocurrencia de los delitos en determinados puntos, los cuales se denominan puntos calientes (*Hot-Spot*).

En la tabla I se describen las variables y subvariables objeto del estudio de investigación y los indicadores.

Tabla I. **Variables del estudio de la investigación y sus relaciones**

Variables	Definición	Subvariables	Indicadores
Zona con índice delincuencial	Identifica la zona compuesta por el área geográfica donde se predecirá la ocurrencia de hechos delictivos en el departamento de Guatemala	Punto georreferencial con incidencia criminal Tipificación del delito Tiempo • Mes • Día de la semana • Hora Características de la victima • Edad • Sexo • Ocupación	Tasa de delincuencia • Por tipificación del delito • Por características de la victima • Por tiempo Cantidad de hechos delictivos • Por tipificación del delito • Por características de la victima • Por tiempo
Precisión de los algoritmos	Mide el porcentaje de precisión de los algoritmos utilizados para la predicción	Algoritmo K-MEANS Algoritmo FUZZY C Algoritmo Clustering jerárquico Algoritmo red de Bayes Tamaño de la muestra	Porcentaje de precisión • K-MEANS • FUZZY C • Clustering jerárquico • Red de Bayes Porcentaje del error • K-MEANS • FUZZY C • Clustering jerárquico • red de Bayes

Fuente: elaboración propia.

- Técnicas de recolección de información

Las técnicas para la recolección de información y sus instrumentos se encuentran descritos en las fases del estudio.

Para la experimentación de los modelos de predicción y sus variables se seleccionaron diferentes muestras y conjuntos de datos pertenecientes a la información histórica de los casos de información de investigación criminal del departamento de Guatemala, registrada por la DEIC en los años 2017, 2018 y 2019. El ámbito geográfico está determinado por las zonas con índice delincuencial en el departamento de Guatemala y que pertenezcan a los delitos siguientes:

- Delitos contra la vida e integridad de las personas
- Delitos de índole sexual
- Delitos de índole patrimonial

- Fases del estudio

    - Revisión documental

Para esta fase se revisaron diferentes fuentes, entre ellas libros, revistas, documentos y tesis que describen el marco conceptual sobre el que se basa el análisis predictivo, específicamente aplicado a la predicción del delito.

Se logó obtener una base conceptual sólida, la cual se utilizó en el diseño de los modelos de predicción del delito en la DEIC que fueron incluidos en el desarrollo del prototipo del presente proyecto. Para ello se realizó una investigación de los siguientes tópicos:

- Minería de datos y algoritmos aplicados a la predicción.
- Análisis de modelos predictivos.
- Análisis predictivo del crimen.
- Metodología para realizar el análisis predictivo del crimen.
- Herramientas estadísticas para la implementación de los algoritmos de predicción.

Con la investigación anterior, se lograron aplicar los conceptos de análisis predictivo y su metodología a la predicción del delito.

De la misma manera se realizó una revisión de la documentación vigente de la DEIC sobre la investigación criminal en el departamento de Guatemala, la cual incluyó:

- Normativa vigente de la investigación criminal y sus delitos
- Modelo de investigación criminal
- Información recolectada durante la investigación criminal
- Información utilizada para la inteligencia criminal

En el desarrollo del prototipo se utilizó una arquitectura de microservicios, la cual integra diferentes tecnologías y servicios para la predicción del delito y la interpretación de los resultados, por lo que se revisaron libros y documentos sobre los siguientes temas:

- Arquitecturas de software utilizando microservicios.
- Diseño de componentes de arquitectura de microservicios.
- Herramientas y tecnologías para el desarrollo de software con microservicios.

- Diseño de los modelos de predicción

Esta fase cubrió la definición de los modelos matemáticos y los algoritmos, que fueron utilizados en la predicción de la probabilidad de ocurrencia del delito.

Esta fase se realizó con la ayuda de un experto en criminología, logrando delimitar el alcance de la predicción del delito, identificar las variables asociadas a los tipos de delitos fueron objeto de estudio, así como las fuentes de donde se obtuvo la información que sirvió para la experimentación del prototipo.

Para ellos se siguieron los siguientes pasos:

- Establecimiento del diseño de los modelos predictivos, así como del análisis de los resultados de los experimentos.
- Delimitación del alcance de la predicción.
  - Tipificación del delito.
  - Geográfico: Municipio, zonas, áreas.
- Identificación de variables que serán objeto de estudio.
- Selección de las fuentes de información de investigación criminal.
- Definición de los modelos.
- Limpieza y preparación de la información.
- Elaboración de las fichas de modelos de predicción.

Como resultado de esta fase se elaboró un documento con la estructura que se muestra en la tabla II.

Tabla II. **Documento de diseño de los modelos predictivos**

Documento de diseño de modelos de predicción del delito para la DEIC
Objetivos • General • Específicos
Justificación
Delimitación del alcance
Descripción de las fuentes de información de investigación criminal
Ficha de modelos matemáticos predictivos y sus variables
Conclusiones
Recomendaciones
Firmas

Fuente: elaboración propia.

La ficha de modelos predictivos y sus variables cuenta con el formato de acuerdo con la tabla III.

Tabla III. **Ficha para la definición de los modelos predictivos**

Modelo predictivo • Nombre
No. <Departamento de investigación criminal><correlativo> Donde <Departamento de investigación criminal>: Departamento de la DEIC que investiga el delito • HOM: Delitos contra la vida e integridad de las personas • SEX: Delitos sexuales • PAT: Delitos patrimoniales <correlativo>: Número correlativo de modelo predictivo
Tipo de Delito: Según la tipificación de los delitos investigados por la DEIC
Fecha de elaboración:

Continuación de la tabla III.

Variables:				
Nombre	Tipo	Tipo de dato	Fuente primaria	Repositorio
(Dependiente, Independiente)				
Personal responsable de la elaboración				
Nombres	Apellidos	Cargo		
Bitácora de revisiones y cambios				
Fecha	Personal responsable	Tipo de cambio/revisión	Observaciones	

Fuente: elaboración propia.

- Diseño del prototipo

En esta fase se realizó el diseño de los módulos o componentes que conforman el prototipo, considerando una arquitectura orientada a microservicios.

Según la solución propuesta, se diseñaron los siguientes módulos o componentes:

- Configuración y parametrización
- Validación y entrenamiento
- Clasificación e identificación de patrones
- Predicción de la probabilidad de ocurrencia en un punto caliente (Hot-Spot)
- Módulo de visualización e interpretación de resultados

Para alcanzar el objetivo propuesto en esta fase, se utilizaron patrones de diseño UML (Lenguaje unificado de modelado) en la elaboración de un documento de diseño que contenga las especificaciones siguientes:

- Representación de la arquitectura
    - Elaboración del diagrama y representación de la arquitectura
    - Elaboración del diagrama de componentes
    - Descripción de los componentes
- Vista de la implementación
    - Descripción general
    - Diagrama de la implementación
- Vista de puesta en marcha
    - Descripción general
    - Diagrama de puesta en marcha
- Anexos
    - Listado de herramientas a utilizar en el desarrollo de los diferentes componentes
    - Desarrollo del prototipo

En esta fase se desarrollaron las aplicaciones informáticas de cada uno de los componentes diseñados, utilizando las herramientas descritas en el documento de diseño, las cuales fueron seleccionadas como las más idóneas para la realización de dicha fase.

Las actividades generales realizadas para cumplir el objetivo de esta fase son los siguientes:

- Instalación y configuración del ambiente de desarrollo
    - Instalación de herramientas
    - Instalación de servicios y servidores
    - Configuración de máquinas virtuales
- Codificación de los componentes informáticos
- Sesiones de pruebas y realimentación de los componentes desarrollados

- Integración de los componentes desarrollados
- Presentación final del prototipo desarrollado con sus ajustes

El producto o entregable obtenido de esta fase fueron las aplicaciones informáticas de los componentes de la solución propuesta para el análisis de predicción del delito.

- Experimentación

Con el prototipo ya desarrollado se realizó la fase de experimentación de los modelos de predicción y sus variables, utilizando conjuntos de datos identificados en la fase de diseño de los modelos.

Los objetivos alcanzados fueron los siguientes:

- Se probaron y afinaron los modelos predictivos diseñados para garantizar un nivel óptimo de precisión.

- Se identificaron patrones de incidencia criminal.

- Se probaron y ajustaron los componentes de la arquitectura del prototipo con la finalidad de optimizar el rendimiento del procesamiento realizado durante la predicción.

Para alcanzar los objetivos propuestos se realizaron pruebas, siguiendo los siguientes pasos:

- Configuración y parametrización del prototipo
  - Registro de parámetros necesarios

- Registro de la información de los modelos de predicción

- Carga de los conjuntos de datos necesarios para la experimentación de los modelos.
- Ejecución de los modelos de predicción.
- Análisis de los resultados de la predicción y su comportamiento.
- Identificación de ajustes en la definición de los modelos.
- Desarrollo de ajustes necesarios a los componentes informáticos.

- Recolección y evaluación de resultados

En esta fase se realizó un análisis cualitativo descriptivo, tomando como base los resultados obtenidos durante la fase de experimentación, realizando un análisis más exhaustivo sobre los comportamientos criminales predichos.

El objetivo principal fue evaluar los fenómenos de predicción del delito, los comportamientos delictivos que el prototipo predice, así como las interpretaciones variables relacionadas en los modelos de la predicción. Esto con la finalidad de determinar y clasificar cuales de los algoritmos de agrupación (K-MEANS, Clustering jerárquico, FUZZY C) y clasificación (red de Bayes) son más precisos a la predicción de determinados delitos.

Para evaluar los resultados se utilizó la información histórica de los delitos cometidos en las áreas que han sido sujetas del estudio, analizando el comportamiento que arroja la predicción ejecutada durante el proceso. Este análisis comparativo determinó la clasificación correcta de los algoritmos para su posterior uso.

- Redacción del informe final

En esta fase se elaboró el informe final con los resultados obtenidos en todas las fases del estudio, la cual incluye los entregables de cada uno, de la manera siguiente:

- Diseño de los modelos predictivos
  - Descripción de los modelos predictivos
  - Listado de los conjuntos y fuentes de datos para cada modelo diseñado
- Diseño del prototipo
  - Representación y características de la arquitectura
  - Diagrama de componentes
  - Descripción de componentes
- Desarrollo del prototipo
  - Componentes informáticos desarrollados
  - Descripción general de la codificación de los componentes
- Experimentación
  - Conjunto de pruebas realizadas
  - Resultados de la validación de los modelos predictivos
  - Validación de los componentes informáticos
    - Configuración
    - Predicción
    - Análisis e interpretación de resultados
- Recolección y evaluación de los resultados
  - Clasificación de los algoritmos de agrupamiento y clasificación
  - Interpretación de la predicción realizada por el prototipo
- Conclusiones
- Recomendaciones

- Requisitos para la implementación del prototipo

INTRODUCCIÓN

Guatemala es uno de los países con índice de delincuencia más alto a nivel América Latina. El Ministerio del Interior a través de la PNC se encarga de velar por la seguridad del ciudadano, haciendo grandes esfuerzos en materia de la persecución del delito, utilizando toda una estructura organizacional y logística a través de programas de seguridad y de prevención. Sin embargo, una de las debilidades más grandes que afronta actualmente el sistema de seguridad del país es la prevención del delito.

La Subdirección de Investigación Criminal, a través de la DEIC es la dependencia encargada de la persecución y prevención del delito, la cual cuenta con protocolos de investigación creados e implementados en conjunto con expertos de seguridad, cuya finalidad principal es reducir el índice de la delincuencia en Guatemala.

Un sistema de información criminal bien diseñado, que facilite la tarea de la investigación criminal, no está completo, sino cuenta con análisis de inteligencia criminal, identificación de patrones criminales y predicción del delito.

Los avances científicos y tecnológicos relacionados a los modelos predictivos, el análisis y sus aplicaciones, han abierto la posibilidad de su utilización en la predicción de eventos en distintas áreas, en las que se menciona, por ejemplo: finanzas, administración, clima, producción y seguridad, entre otras.

De igual manera en el área de la predicción del delito, los sistemas policiales, buscan hoy en día utilizar y aplicar modelos predictivos que les ayuden

a identificar patrones de criminalidad en áreas de mucha incidencia y a la vez les permita organizar y movilizar los recursos policiales operativos para evitar que los crímenes se cometan. Esto le permite al sistema de seguridad y justicia reducir la tasa interanual de crímenes cometidos, asegurando el resguardo de la vida y patrimonio de la sociedad civil.

El presente trabajo de investigación realizó el desarrolló y experimentación de un prototipo que permita a la DEIC definir modelos de predicción del delito, en sus diferentes manifestaciones, para el departamento de Guatemala. Para ello se utilizaron modelos matemáticos y algoritmos de clasificación y agrupamiento, los cuales reducen el error en la precisión y de esta manera pueden proveer a la unidad de análisis e inteligencia criminal, la interpretación de los resultados y la coordinación con otras unidades para el aumento de la vigilancia y la presencia policial en los sectores de mucha incidencia y donde la predicción del delito lo indique.

En el capítulo uno, se describen los antecedentes de la investigación, dónde se hace énfasis sobre la necesidad que tiene la DEIC de un sistema de predicción del delito que le permita realizar actividades de coordinación para la prevención del mismo. Para ello se describe la forma en que algunos países han integrado en sus sistemas de información criminal, modelos predictivos y algoritmos que les han permitido predecir el delito de manera precisa.

En el capítulo dos, se justifica la investigación, siguiendo la línea de Tecnologías de información y comunicación para apoyo a la seguridad ciudadana y nacional, destacando que el desarrollo del prototipo de predicción del delito fortalecería a la DEIC y al Ministerio de Gobernación en la prevención del crimen y la seguridad y resguardo del ciudadano en el departamento de Guatemala.

En el capítulo tres, se detallan y se delimitan los alcances investigativos, técnicos y de resultados que cubre el presente trabajo, con la finalidad de cumplir con los objetivos propuestos y obtener los resultados esperados.

En el capítulo cuatro, se describen los conceptos relacionados al análisis predictivo, los algoritmos de agrupamiento y clasificación: K-MEANS, FUZZY C, Clustering jerárquico y red de Bayes, los cuales fueron implementados utilizando procesos de minería de datos realizando las pruebas de precisión de los modelos matemáticos definidos. Igualmente se describen los principales componentes y la metodología de análisis de predicción del crimen que se utilizó para el proceso de diseño, construcción y experimentación del prototipo. También se describe en un apartado los conceptos generales de una arquitectura de software orientada a microservicios y sus componentes, los cuales fueron aplicados en el diseño de la arquitectura del prototipo, para lograr un desacoplamiento entre los componentes que integran el sistema, disponibilidad y rendimiento en el procesamiento de los algoritmos de predicción.

En el capítulo cinco, se presentan como resultados, los diseños de los modelos predictivos aplicados a la predicción del delito, así como la descripción de cada uno de ellos y la interpretación de las variables asociadas con el conjunto de datos que se utilizó. Se presentan también las vistas del diseño de la arquitectura utilizada, así como los componentes funcionales que permiten la definición, configuración y parametrización de los modelos y sus variables, la definición de los conjuntos de datos, la experimentación, el afinamiento de los modelos, ejecución de la predicción del delito e interpretación de los resultados.

En el capítulo seis, se discuten los resultados de los experimentos realizados en el prototipo, para cada uno de los modelos predictivos definidos, realizando una interpretación y contextualización de las variables, patrón criminal

y área de incidencia, con la finalidad de establecer el comportamiento de la predicción ejecutada y clasificar los algoritmos con base en esta interpretación, estableciendo: donde, cuando y como utilizarlos para que el funcionamiento sea el más preciso.

Para finalizar se presentan las conclusiones y recomendaciones, realizando una breve interpretación del comportamiento criminal y sus patrones en el departamento de Guatemala y como puede ser utilizado el prototipo de manera óptima, así como los requisitos técnicos y funcionales necesarios para ponerlo en funcionamiento en un departamento de la DEIC.

1. ANTECEDENTES

- La investigación criminal en Guatemala

En el año 2009, por medio del Acuerdo Gubernativo número 97-2009 del uno de abril, *Reglamento de Organización de la Policía Nacional Civil*, y sus reformas en los Acuerdos Gubernativos 240-2011 y 515-2011, se estableció la Orden General que describe el manual de puestos y funciones de la Subdirección de Investigación Criminal (DEIC), así como el desarrollo integral de cada una de sus unidades (Policía Nacional Civil [PNC], 2009).

Con la Orden General No. 12-2009 y posteriormente su actualización 67-2014, se crea la DEIC, quien es: “Un órgano técnico científico dedicado a la investigación criminal siendo su principal función la de investigar y perseguir los delitos tipificados en las leyes vigentes del país y adoptar medidas urgentes para la prevención de los mismos” (División Especializada en Investigación Criminal [DEIC], 2009).

Los organismos internacionales a través de la Secretaría Ejecutiva de la Instancia Coordinadora de la Modernización del Sector Justicia (Secretaría Ejecutiva de la Instancia Coordinadora de la Modernización del Sector Justicia [SEICMSJ] 2014), ha realizado diferentes esfuerzos de cooperación para fortalecer a la DEIC en materia de capacidades, estrategias y sistematización de los procesos de investigación criminal, integración con el sector justicia del país y de la misma Policía Nacional Civil, cuyo fin principal es el de reducir la incidencia criminal, tal es el caso de las asistencias técnicas brindadas para la sistematización del proceso de investigación criminal.

A pesar de todos los esfuerzos por fortalecer a la DEIC, se sigue afrontando el problema de la muy poca o escasa inteligencia criminal basada en el análisis de la información, que permita planificar y ejecutar medidas preventivas. Esto se debe a la inmensa cantidad de casos de investigación a los que les tienen que dar seguimiento, según los altos índices de criminalidad de acuerdo al Reporte Estadístico 2017 realizado por la Secretaría Técnica del Consejo Nacional de Seguridad (STCNS, 2017), donde se menciona un incremento en los hechos violentos en el departamento de Guatemala.

Es importante mencionar que a pesar de que se cuenten con sistemas de registro de la información obtenida de las diligencias de investigación criminal, como por ejemplo el Sistema de Información Policial (SIPOL 2). Esta información no es utilizada para realizar análisis para la toma de decisiones.

El Acuerdo Gubernativo 97-2009 establece como función principal de la Subdirección General de Investigación Criminal y la DEIC, entre otras, la de combatir el crimen en sus diferentes manifestaciones, por medio de sistemas estratégicos (PNC, 2009).

Para ello es necesario contar con herramientas de análisis y predicción del delito, en las distintas zonas geográficas consideradas de alto riesgo, donde operan las estructuras criminales y de esta manera lograr la desarticulación de las mismas.

Según Omkar, Sayak, Raj, Suraj y Rohini (2018), se puede utilizar el análisis y la predicción del delito, como enfoque sistemático para analizar e identificar los patrones, relaciones y tendencias en la delincuencia. Al trabajar con los modelos predictivos, se logran resultados de visualización de las regiones con alta probabilidad de ocurrencia de delitos e indicar aquellas áreas propensas a la

delincuencia con la finalidad de fortalecer las acciones de las unidades de la PNC y la DEIC en materia de la prevención.

> El análisis predictivo puede llevarse a cabo utilizando minería de datos con algoritmos de agrupamiento, mediante un procedimiento que básicamente cuenta con los siguientes pasos: 1) entrada: definición y diseño del conjunto de datos, 2) preprocesamiento del conjunto de datos (extracción, transformación y carga), 3) selección de características, 4) clasificación algorítmica, 5) predicción (entrenamiento y ejecución) (Sreedevi, Vardhan y Krishna, 2018, p.10).

Según estudios relacionados sobre la predicción del crimen en Colombia Barreras, Díaz, Riascos y Ribero, (2016), demuestran que entre los algoritmos más precisos se encuentran: K-means (partición de conjunto de un número de observaciones y grupos específico), Fuzzy c (Clustering difuso), Clustering jerárquico, red bayesiana. Para la evaluación de los algoritmos de clasificación se toman en cuenta los siguientes aspectos: 1) precisión (% de casos clasificados correctamente), 2) eficiencia, 3) robustez, 4) escalabilidad. Para lograr altos niveles de precisión y exactitud en la predicción, estos algoritmos consumen grandes cantidades de recursos de hardware (memoria, disco y procesamiento) en las fases de clasificación y entrenamiento; por lo que se tiene que buscar una arquitectura que sea escalable de manera horizontal o vertical, para evitar el problema de la falta de recursos y minimizar el tiempo de procesamiento.

> La diferencia en la precisión y la eficiencia de los diferentes algoritmos puede ser explicada de la manera siguiente: Para el caso de Clustering jerárquico, la diferencia se da por el tamaño de los segmentos de datos que permiten identificar puntos calientes (*Hot-Spot*), si la agrupación de los datos es granular, se hace difícil identificar los clústeres (i.e. agrupaciones) de crimen, porque la visualización tendrá demasiadas zonas demarcadas

como “puntos calientes”. “Por otro lado, si los datos se agrupan de acuerdo con diferentes criterios geográficos habrá un problema de unidad de área modificable, en el que las estadísticas resultantes son altamente sensibles a la escogencia arbitraria de los límites de agrupación de los datos (Barreras, Díaz, Riascos y Ribero, 2016, p.16).

Según Omkar, Sayak, Raj, Suraj y Rohini (2018), “En el caso de K-Means y Fuzzy C existe menos precisión cuando no se utiliza un conjunto de datos ideal para el entrenamiento y es más preciso, cuando se utiliza un conjunto de datos más amplio” (p.4).

Sin embargo, Baumgartner, Ferrary y Palermo (2008), afirman:

Para red bayesiana en promedio, el 80 % de las características de los delitos se predicen correctamente de acuerdo al perfil criminal y dado que cada predicción se acompaña por un nivel de confianza que es proporcional a la precisión esperada, al considerar solo predicciones con altos niveles de confianza, la exactitud promedio aumenta a 95.6 % (p.4).

Por lo tanto, se puede ver, por un lado, la necesidad que existe en la Subdirección de Investigación Criminal, la DEIC, la Policía Nacional Civil, y en la población en general de contar con un sistema para el análisis y la predicción del delito, y, por otro lado, aunque, haya estudios y avances tecnológicos en la materia, los mismos han sido aplicados a países, regiones y poblaciones, dónde se manifiestan patrones criminales particulares.

Por lo que el presente trabajo aplica los conceptos y avances tecnológicos de análisis predictivo, basándose en los patrones criminales que operan en

Guatemala, utilizando diferentes categorías para la clasificación, según la tipología de delitos vigente, buscando maximizar la precisión.

2. JUSTIFICACIÓN

La línea de investigación que persigue este proyecto se enfoca en tecnologías de la información y comunicación para apoyo a la seguridad ciudadana y nacional.

Los modelos predictivos del delito analizan e identifican patrones, relaciones y tendencias en la delincuencia con el objetivo de visualizar las regiones con alta probabilidad de ocurrencia de delitos e indicar aquellas áreas propensas a la delincuencia con la finalidad de fortalecer las acciones de los órganos de justicia en materia de la prevención.

La escasa inteligencia criminal basada en el análisis de la información que actualmente realiza la Policía Nacional Civil de Guatemala, a través de la División Especializada en Investigación Criminal (DEIC), no le permite realizar acciones preventivas para evitar que se cometan los crímenes y eso ha ocasionado que únicamente se invierta tiempo y esfuerzo en la persecución del delito, pero no en la prevención del mismo. Contar con un sistema de predicción fortalecería a la DEIC en la inteligencia criminal logrando establecer la probabilidad de ocurrencia de los hechos delictivos e identificar las áreas geográficas más propensas y así realizar acciones de vigilancia, seguimiento y desarticulación de grupos criminales que operan en las mismas.

Estas acciones preventivas realizadas en conjunto por las diferentes unidades de la PNC, tendrían un efecto positivo para la seguridad de la población del departamento de Guatemala, ya que contribuirían a la reducción del crimen, el cual actualmente crece cada día más y aunque el Ministerio de Gobernación

ha realizado grandes esfuerzos en materia del resguardo de la seguridad, no ha logrado disminuir; por el contrario, cada día aumenta.

El análisis predictivo es llevado a cabo a través de algoritmos de agrupamiento, utilizando el concepto de minería de datos, en el que inicialmente se definen las características de agrupamiento y el conjunto de datos basados en la tipificación de los delitos, con la finalidad de realizar la clasificación algorítmica, utilizando el entrenamiento y parametrización adecuada para realizar la predicción.

Estudios realizados sobre este campo demuestran que los algoritmos de clasificación más precisos son: K-means (partición de conjunto de n observaciones en k grupos), Fuzzy c (Clustering difuso), Clustering jerárquico, red bayesiana. Cada uno debe ser configurado adecuadamente para lograr niveles máximos de precisión.

El proyecto construye un prototipo que determina cuál de estos algoritmos de clasificación maximiza la precisión, tomando como base la definición de las características de los hechos delictivos y el conjunto de datos obtenido de los registros de los casos de investigación criminal de la DEIC.

Este prototipo desarrollado, considera en su diseño la arquitectura de sistemas y sus componentes que permitan optimizar el tiempo, el procesamiento y los recursos utilizados durante el análisis predictivo, permitiendo escalar, cuando sea necesario.

La DEIC puede hacer uso de este prototipo, readecuarlo a sus necesidades e implementarlo como una herramienta de análisis para la inteligencia criminal y

de esta manera realizar acciones de coordinación y ejecución para la prevención del delito en el departamento de Guatemala.

3. ALCANCES

3.1. Investigativos

- Definir los conceptos del análisis predictivo, así como de los modelos K-MEANS, FUZZY C, Clustering jerárquico y red de Bayes, asociados a la predicción.

- Dar a conocer el proceso y la metodología de análisis predictivo del crimen utilizando Minería de Datos.

- Describir los conceptos y beneficios de utilizar una arquitectura orientada a microservicios aplicada a la predicción del delito, utilizando APIS de servicios REST para la integración de los microservicios.

3.2. Técnicos

- Diseñar los modelos de agrupamiento, clasificación, así como los conjuntos de datos para la predicción del delito, tomando como base la información de los casos de investigación criminal de la DEIC.

- Diseñar los modelos de datos y los esquemas necesarios para almacenar la información utilizada antes y después de la predicción del delito.

- Diseñar y desarrollar los procesos de extracción, transformación y carga (ETL) que serán utilizados en el análisis y predicción del delito.

- Diseñar una arquitectura orientada a microservicios para los componentes del análisis y la predicción del delito utilizando los algoritmos K-MEANS, FUZZY C, Clustering jerárquico y red de Bayes

- Diseñar y desarrollar una API de servicios REST que integre los procesos de entrenamiento, análisis e interpretación de resultados para la predicción del delito

- Instalar y configurar las herramientas necesarias para la implementación de la arquitectura de microservicios propuesta y sus componentes.

3.3. Resultados

Prototipo de sistema de información que permite realizar el proceso de análisis de criminalidad de manera optimizada, en el departamento de Guatemala, utilizando una arquitectura de microservicios.

Las funcionalidades del prototipo son las siguientes:

- Creación de los modelos predictivos: variables, relaciones, que permitan predecir los eventos priorizados.

- Carga de los conjuntos de datos para la validación.

- Validación de los modelos predictivos, utilizando los algoritmos: K-MEANS, CLUSTERING JERÁRQUICO, FUZZYC y red bayesiana.

- Carga de los conjuntos de datos para la ejecución de la predicción.

- Ejecución de la predicción por medio de los algoritmos K-MEANS, CLUSTERING JERÁRQUICO, FUZZY C y red bayesiana.

- Mapeo de puntos calientes mediante georreferenciación para realizar el análisis de la posible intervención policial en materia de la prevención del delito.

Estas funcionalidades están agrupadas modularmente, pero integradas por medio de microservicios, de la manera siguiente:

- Módulo de configuración y parametrización
- Módulo de validación y entrenamiento predictivo
- Módulo de ejecución y análisis predictivo
- Módulo de visualización e interpretación de resultados

Como resultados de la implementación de este prototipo la DEIC, puede:

- Definir los modelos para la predicción del delito en el departamento de Guatemala, utilizando una metodología funcional basada en algoritmos de agrupamiento y de clasificación.

- Utilizar un sistema de información que permite:

    - Realizar la validación de los modelos de predicción, en cuanto al nivel de precisión de los resultados obtenidos.

    - Interpretar los resultados obtenidos en un mapa geográfico de puntos calientes, agrupados por zona.

4. MARCO TEÓRICO

4.1. Minería de datos

Según la definición de Pérez (2014), es: "Un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos. Se trata de analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos" (p.3).

"Actualmente se dispone de grandes cantidades de datos y es más necesario que nunca analizarlos ordenadamente para extraer de un modo automatizado la inteligencia contenida en ellos utilizando técnicas especializadas apoyadas en herramientas informáticas" (Pérez, 2014, p.3).

"La minería de datos ha pasado de tratar de entender los datos a: tratar de entender los eventos que se encuentran detrás" (Gironés, Casas, Minguillón y Caihuelas, 2017, p.25).

La necesidad del análisis de datos y la extracción del conocimiento de los mismos ha llevado a utilizar las herramientas tecnológicas con la finalidad de minimizar el tiempo de la obtención de la información.

Según Pérez (2004), el crecimiento tecnológico facilita el uso de algoritmos estadísticos, lo que permite automatizar el tratamiento de la información con análisis multivariante de datos de una forma muy sencilla.

Las técnicas de minería de datos y la estadística son de la misma manera tan antiguas que coinciden con las técnicas estadísticas de análisis multivariante de datos, mismas que se incluyen en las herramientas de la minería de un modo ordenado y secuencial (Pérez, 2004).

Según Pérez (2004), estas técnicas se pueden clasificar de la siguiente manera:

- Modelado según la teoría: las variables se clasifican en dependientes e independientes, en el que se establece como base el conocimiento previo y después se contrasta con la realidad antes de aceptarse como válido.

- Modelado según los datos: las variables tienen inicialmente la misma clasificación, por lo que no se le asigna ningún rol predeterminado a cada variable, ni tampoco que exista un modelo previo. Estos modelos se crean basándose en el reconocimiento de patrones, con una mezcla de conocimiento previo y validado posteriormente. Un ejemplo idóneo para esto son las redes neuronales que permiten identificar modelos complejos y su respectivo ajuste durante la estimación y la validación; técnicas de clasificación de las que se extraen patrones de comportamiento o clases, arboles de decisión para dividir los datos en grupos basados en los valores de las variables del modelo.

- Auxiliares: Serie de métodos basados en estadística descriptiva e informes.

En todo modelo, sin importar la técnica utilizada, se pueden incluir análisis de regresión y asociación, varianza y covarianza, análisis discriminante, así como series temporales, estas se deben llevar acabo en las fases siguientes:

- Identificación objetiva: conjunto de reglas que llevan a justar el modelo de manera óptima, con base en los datos de entrada.

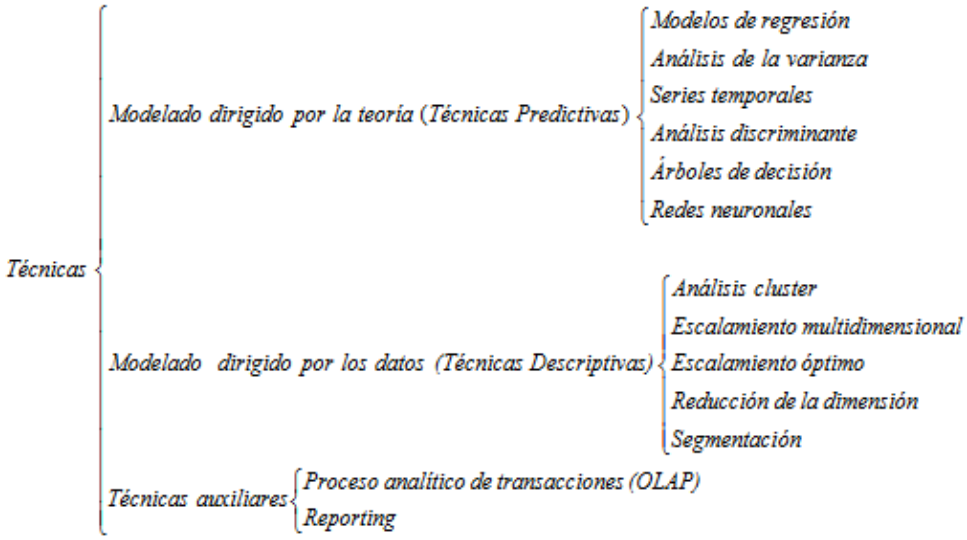
- Estimación: proceso en el que se realizan los cálculos con base en los parámetros del modelo que se ha diseñado y los datos de entrada.

- Diagnóstico: se contrasta la información del modelo, con la estimación realizada.

- Predicción: proceso en el que se utiliza el modelo ajustado, estimado y validado, con la finalidad de predecir eventos futuros, según las variables dependientes.

Gráficamente, se presenta a continuación la clasificación descrita anteriormente:

Figura 1. **Técnicas de minería de datos**



Fuente: Pérez. (2004). *Técnicas de análisis multivariante de datos.*

4.2. Algoritmos de minería de datos

Según Microsoft Corporation (2018):

Un algoritmo de minería de datos es un conjunto de heurísticas y cálculos que crea un modelo de minería de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas (párr.1).

Según Pérez (2014), estos algoritmos pueden describirse como:

- Un conjunto de clústeres que describen relaciones entre los distintos casos en los conjuntos de datos.

- Un árbol de decisión que predice un evento y como se ven afectados los distintos puntos de vista de los datos.

- Un modelo matemático predictivo para un evento específico, entre otros.

Según Microsoft Corporation (2018), estos se dividen en:

- Clasificación: "Los que predicen una o más variables discretas, basándose en los demás atributos del conjunto de datos" (párr.6).

- Regresión: "Los que predicen una o más variables continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos" (párr.7).

- Segmentación: "Los que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares" (párr.8).

- Asociación: "Los que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra" (párr.9).

- Análisis de secuencias: "Los que buscan resumir las secuencias frecuentes o episodios en los datos, como un flujo de la ruta de acceso Web" (párr.10).

El alcance de este proyecto tomará en cuenta los siguientes algoritmos, los cuales fueron seleccionados, como los más eficientes para el análisis predictivo criminal, según varios autores.

4.2.1. K-means

Es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

La agrupación del conjunto de datos es un problema computacionalmente difícil y se pueden utilizar algoritmos de refinamiento iterativo a través de mezclas de distribuciones gaussianas. Además, estos algoritmos utilizan los centros que los grupos realizan para modelar los datos, sin embargo, K-means tiende a encontrar grupos de extensión espacial comparable (MacQueen, 1967, pp 281-297).

La definición matemática para los algoritmos K-means es explicada por (MacQueen, 1967) de la manera siguiente:

Dado un conjunto de observaciones ($\mathbf{x}_1$, $\mathbf{x}_2$,…, $\mathbf{x}_n$), donde cada observación es un vector real de *d* dimensiones, *k*-means construye una partición de las observaciones en *k* conjuntos ($k \leq n$) a fin de minimizar la suma de los

cuadrados dentro de cada grupo (WCSS): **S** = {S_1, S_2,…, S_k}

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Donde $\boldsymbol{\mu}_i$ es la media de puntos en S_i (p.289).

4.2.2. Fuzzy C

El agrupamiento difuso (en inglés, *fuzzy Clustering*) es una clase de algoritmos de agrupamiento donde cada elemento tiene un grado de pertenencia difuso a los grupos.

Este tipo de algoritmos surge de la necesidad de resolver una deficiencia del agrupamiento exclusivo, que considera que cada elemento se puede agrupar inequívocamente con los elementos de su clúster y que, por lo tanto, no se asemeja al resto de los elementos. Tras la introducción de la lógica difusa por Zadeh en 1965 surgió una solución para este problema, caracterizando la similitud de cada elemento a cada uno de los grupos. Esto se logra representando la similitud entre un elemento y un grupo por una función, llamada función de pertenencia, que toma valores entre cero y uno. Los valores cercanos a uno indican una mayor similitud, mientras que los cercanos a cero indican una menor similitud. Por lo tanto, el problema del agrupamiento difuso se reduce a encontrar una caracterización de este tipo que sea óptima (Yang, 1993, pp 1-2).

Los algoritmos de agrupamiento difuso han tenido varias aplicaciones en diversas áreas como: procesamiento de imágenes, sistemas de ingeniería, estimación de parámetros, entre otras (Yang, 1993).

4.2.3. Clustering jerárquico

Este método, se basa en clúster de manera puntual, utilizando niveles jerárquicos, para ello utiliza los siguientes tipos de estrategias:

- Aglomerativas: se conocen como ascendentes, las observaciones comienzan en un grupo de concordancias similares, luego estos son mezclados entre grupos, mientras uno sube en la jerarquía de manera ascendente.

- Divisivas: se conocen como descendentes: todas las observaciones comienzan en un grupo de concordancias similares, luego estos son mezclados entre grupos, mientras uno baja en la jerarquía de manera ascendente.

Partiendo de tantos grupos iniciales como observaciones se estudian, se trata de conseguir agrupaciones sucesivas entre ellas de forma que progresivamente se vayan integrando en clústeres, los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al clúster final que contiene todos los casos analizados.

Según Perez (2004), la gráfica que usualmente se utiliza para representar el ordenamiento jerárquico es un árbol invertido, denominado dendograma.

4.2.4. Red de Bayes

Según Pearl (2000):

> Una red bayesiana, red de Bayes, red de creencia, modelo bayesiano (de Bayes) o modelo probabilístico en un grafo a cíclico dirigido es un modelo grafo probabilístico (un tipo de modelo estático) que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un grafo a cíclico dirigido (DAG por sus siglas en inglés). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades (p.280).

Para estas redes se forman grafos dirigidos en cuyos nodos se utilizan variables que pueden ser condicionalmente independientes de las otras, las cuales utilizan una función de probabilidad, según los conjuntos que se defina.

> Formalmente, las redes bayesianas son grafos dirigidos a cíclicos cuyos nodos representan variables aleatorias en el sentido de Bayes: las mismas pueden ser cantidades observables, variables latentes, parámetros desconocidos o hipótesis. Las aristas representan dependencias condicionales; los nodos que no se encuentran conectados representan variables las cuales son condicionalmente independientes de las otras. Cada nodo tiene asociado una función de probabilidad que toma como entrada un conjunto particular de valores de las variables padres del nodo y devuelve la probabilidad de la variable representada por el nodo. Por ejemplo, si por padres sonvariables booleanas entonces la función de probabilidad puede ser representada por una tabla de entradas, una entrada para cada una de las posibles combinaciones de los padres siendo

verdadero o falso. Ideas similares pueden ser aplicadas a grafos no dirigidos, y posiblemente cíclicos; como son las llamadas redes de Márkov.

Existen algoritmos eficientes que llevan a cabo la inferencia y el aprendizaje en redes bayesianas. Las redes bayesianas que modelan secuencias de variables (ej. señales del habla o secuencias de proteínas) son llamadas redes bayesianas dinámicas. Las generalizaciones de las redes bayesianas que pueden representar y resolver problemas de decisión bajo incertidumbre son llamados diagramas de influencia (Pearl, 2000, pp 283-285).

Naive Bayes (ingenuo)

Es un tipo de algoritmo de Bayes, que se utiliza mucho, pues asume que la presencia o ausencia de una característica particular que no se encuentra presente en otra, dada la clase de variable, por ejemplo: Una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Bayes ingenuo considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

Para otros modelos de probabilidad, los clasificadores de Bayes ingenuo se pueden entrenar de manera muy eficiente en un entorno de aprendizaje supervisado. En muchas aplicaciones prácticas, la estimación de parámetros para los modelos Bayes ingenuo utiliza el método de máxima verosimilitud, en otras palabras, se puede trabajar con el modelo ingenuo de Bayes sin aceptar probabilidad bayesiana o cualquiera de los métodos bayesianos.

Una ventaja del clasificador de Bayes ingenuo es que solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros (las medias y las varianzas de las variables) necesarias para la clasificación. Como las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza (Pearl, 2000, p.287).

4.3. El análisis y los modelos predictivos

A continuación, se define el análisis predictivo, los tipos de modelos y su uso.

4.3.1. El análisis predictivo

El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro.

El análisis predictivo se fundamenta en la identificación de relaciones entre variables en eventos pasados, para luego explotar dichas relaciones y predecir posibles resultados en futuras situaciones. Ahora bien, hay que tener en cuenta que la precisión de los resultados obtenidos depende mucho de cómo se ha realizado el análisis de los datos, así como de la calidad de las suposiciones (Gorunescu, 2011, p.35).

El análisis predictivo es más que realizar pronósticos sobre eventos simples, sino que se fundamenta en relacionar más variables asociadas al evento pronosticado.

En un principio puede parecer que el análisis predictivo es lo mismo que hacer un pronóstico (que hace predicciones a un nivel macroscópico), pero se trata de algo completamente distinto. Mientras que un pronóstico puede predecir cuántos helados se van a vender el mes que viene, el análisis predictivo puede indicar qué individuos es más probable que se coman un helado. Esta información, si se utiliza de la forma correcta, permite orientar los esfuerzos para ser más productivos en la consecución de los objetivos. Para llevar a cabo el análisis predictivo es indispensable disponer de una considerable cantidad de datos, tanto actuales como pasados, para poder establecer patrones de comportamiento y así inducir conocimiento (Gorunescu, 2011, p.36).

4.3.2. Modelos aplicables al análisis predictivo

Generalmente, se usa el término análisis predictivo cuando en realidad se está hablando del modelado predictivo, que realiza calificaciones mediante modelos predictivos y pronósticos. Sin embargo, cada vez se está utilizando más el término para referirse a todo lo relacionado con la disciplina analítica, como el modelado descriptivo o el modelado decisivo. Estas disciplinas implican un riguroso análisis de datos y son ampliamente utilizadas en negocios, como mecanismo de ayuda a la toma de decisiones (Gorunescu, 2011, p.36).

Los modelos predictivos son utilizados para predecir el comportamiento de un individuo o evento, con base en las características de los mismos, obteniendo una calificación basada en la probabilidad de ocurrencia.

Un modelo predictivo es un mecanismo que predice el comportamiento de un individuo, para ello utiliza la información de las características de él como

entrada y proporciona una calificación predictiva como salida. Cuanto más alta la calificación, más alta es la probabilidad de que el individuo exhiba el comportamiento predicho.

> La calificación producida por cualquier modelo predictivo es una estimación, nunca una realidad, por lo que debe de ser tenida en cuenta con especial cuidado y puede ser necesario que se cruce con otro modelo o que se produzca un análisis adicional a la hora de aplicarla a un individuo concreto. Las calificaciones hablan de tendencias y posibilidades en un grupo lo suficientemente grande, pero no garantiza que la predicción se cumpla en cada caso individual, pues una probabilidad individual por naturaleza simplifica excesivamente la cosa del mundo real que describe.
>
> El tipo de análisis que permiten los modelos predictivos valora la relación existente entre cientos de elementos para aislar los datos que informan sobre un hecho, guiando a la toma de decisiones por un camino seguro. Un paso más allá se encuentra los modelos de decisión, que tienen un modo de trabajar muy similar a la de los modelos predictivos, aunque se emplean en escenarios de mayor complejidad. Se trata de la forma más avanzada de análisis predictivo y consiste en predecir lo que sucedería si se toma una acción determinada. También se conocen como modelos prescriptivos y se basan en la cartografía de las relaciones existentes entre todos los elementos de una decisión (Gorunescu, 2011, p.36).

Según Gorunescu (2011), los modelos aplicables al análisis predictivo son los siguientes: Predictivos, descriptivos y de decisión.

4.3.3. Modelos predictivos

Los modelos predictivos son modelos de la relación entre el rendimiento específico de una unidad en una muestra y uno o más atributos o características conocidos de la unidad. El objeto del modelo es evaluar la probabilidad de que una unidad similar en una muestra diferente exhiba un comportamiento específico (Gorunescu, 2011, p.37).

El análisis predictivo crea un modelo estadístico, en el cual se utiliza un conjunto de datos con variables y relaciones de entrada y es utilizado para predecir eventos de comportamiento en el futuro.

Como ejemplo del análisis predictivo se incluyen las líneas de tendencia o la puntuación de la influencia. Para la creación del modelo predictivo se utilizan unidades de muestra disponibles con atributos conocidos y un comportamiento conocido, a este conjunto de datos se le denomina conjunto de entrenamiento. Por otro lado, se utilizará una serie de unidades de otra muestra con atributos similares, pero de las cuales no se conoce su comportamiento, a este conjunto de datos se le denomina conjunto de prueba (Gorunescu, 2011, p.37).

4.3.4. Modelos descriptivos

Los modelos descriptivos cuantifican las relaciones entre los datos de manera que es utilizada a menudo para clasificar clientes o contactos en grupos. A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular, los modelos descriptivos identifican diferentes relaciones entre los clientes y los productos. La analítica descriptiva proporciona resúmenes simples sobre la audiencia de

la muestra y sobre las observaciones que se han hecho. Estos resúmenes pueden constituir la base de la descripción inicial de los datos como parte de un análisis estadístico más amplio, o pueden ser suficientes en sí mismos para una investigación en particular (Gorunescu, 2011, p.38).

Los modelos descriptivos no utilizan la probabilidad para clasificar u ordenar a los individuos o eventos, sin embargo, pueden utilizarse para desarrollar modelos adicionales y utilizarse en los modelos predictivos.

Los modelos descriptivos no clasifican u ordenan a los clientes por su probabilidad de realizar una acción particular de la misma forma en la que lo hacen los modelos predictivos. Sin embargo, los modelos descriptivos pueden ser utilizados por ejemplo para asignar categorías a los clientes según su preferencia en productos o su franja de edad. Las aplicaciones de los modelos descriptivos pueden ser utilizados para desarrollar nuevos modelos adicionales que pueden imitar un gran volumen de agentes individuales y hacer predicciones. Entre los modelos descriptivos se pueden citar los modelos de simulación, la teoría de colas o las técnicas de previsión (Gorunescu, 2011, p.38).

4.3.5. Modelos de decisión

Los modelos de decisión describen la relación entre todos los elementos de una decisión, los datos conocidos (incluyendo los resultados de los modelos predictivos), la decisión y el pronóstico de los resultados de una decisión, con la intención de predecir los resultados de una decisión en la que se involucran gran cantidad de variables. Estos modelos pueden ser utilizados en la optimización o maximización de determinados resultados mientras minimizan otros. Estos se utilizan en general para el desarrollo de la

decisión lógica o conjunto de reglas de negocio que debería producir el resultado deseado para cada cliente o circunstancia (Gorunescu, 2011, p.38).

4.3.6. Validación de los modelos

Tras haber definido el modelo y las variables relacionados al evento que se necesita predecir, es necesario realizar un proceso de validación del mismo, para comprobar que efectivamente, el mismo funciona. Para ello se utiliza la información histórica de la cual se obtuvieron las variables.

La validación es el proceso de evaluar la precisión de predicción de un modelo. Esta se refiere a la obtención de predicciones utilizando el modelo existente, y luego comparando estos resultados con resultados ya conocidos. Este es el paso más importante en el proceso de construcción de un modelo.

El uso de un modelo que no coincide con los datos no puede producir resultados correctos para responder adecuadamente a la meta pretendida del estudio. Por lo tanto, se entiende que existe toda una metodología para validar un modelo basado en datos existentes (por ejemplo, exclusión, submuestreo aleatorio, validación cruzada, estratificado muestreo, bootstrap, entre otros).

“Finalmente, en la comprensión del modelo es importante identificar los factores que llevan a la obtención del ‘éxito’ así como el ‘fallo’ en la predicción proporcionada por el modelo” (Gorunescu, 2011, p.39).

4.4. El análisis predictivo del crimen

Para comprender adecuadamente el contexto del análisis predictivo es necesario entender el crimen y como los organismos policiales lo combaten.

4.4.1. Generalidades

Los crímenes son un problema social común que afecta la calidad de vida y el crecimiento económico de una sociedad. Se considera un factor esencial que determina si las personas se mudan o no a una nueva ciudad y qué lugares se deben evitar cuando viajan. Con el aumento de crímenes, los organismos de justicia continúan exigiendo sistemas avanzados de información geográfica y nuevos enfoques de minería de datos para mejorar la analítica del crimen y proteger mejor a sus comunidades (Tahani, Rsha y Lor, 2015, p.1).

En los últimos años las fuerzas de policía han estado mejorando su método tradicional de denuncia de delitos con nuevos avances tecnológicos para aumentar su producción registrando eficientemente los crímenes para ayudar a su investigación. Los datos no son justamente un registro de crímenes, también contiene información valiosa que podría ser utilizado para vincular escenas del crimen basadas en el modus operandi (MO) del delincuente (s), sugieren qué delincuentes pueden ser responsable del crimen y también identificar a los delincuentes que trabajan en equipos (redes de delincuentes), entre otros. En la actualidad, las computadoras juegan un papel importante en investigación de todos los tipos de delitos de aquellos que son considerados como un delito por volumen (robo, delito de vehículos, entre otros) a crimen mayor como fraude, tráfico de drogas, homicidios, entre otros (Adderley y Musgrove, 2017, p.1).

No es una tarea fácil para un analista de policía a mano desentrañar las complejidades inherentes dentro de los datos policiales y este problema se agrava cuando el análisis se lleva a cabo por un equipo. La distribución de los datos al equipo puede causar significativa información, que podría ser útil para resolver los crímenes y que se puede perder, ya que cada miembro no está en posesión de todos los hechos relevantes.

Durante mucho tiempo, los criminólogos y los estadísticos han estado aplicando sus habilidades y conocimientos tratando de predecir cuándo y dónde el próximo conjunto de crímenes ocurrirá, con diversos grados de éxito. El volumen del crimen y la conciencia actual de los criminales ponen una tensión en los métodos existentes. El razonamiento ya no es capaz de analizar el conjunto de datos, cuando son millones de registros. Por lo tanto, hay claramente un requisito para un kit de herramientas para ayudar a analizar los datos que harán el mejor uso de los recursos limitados.

“Las técnicas de descubrimiento de conocimiento en bases de datos (KDD) pueden ser usado para revelar el conocimiento que está más allá de la intuición" (Kumar y Chandrasekar, 2011, p.6).

4.4.2. Descubrimiento de conocimiento en base de datos

El “Descubrimiento de Conocimiento en Base de Datos” (KDD en inglés), es un proceso que permite a los usuarios buscar información valiosa en las bases de datos con contenido histórico.

“En este proceso se combina el modelado estadístico, el aprendizaje automático, el almacenamiento de bases de datos y las tecnologías de inteligencia artificial” (Kumar y Chandrasekar, 2011, p.6).

En la prevención del delito el objetivo del KDD es la predicción del comportamiento humano y delincuencial, la cual es una de las aplicaciones más comunes (Mena, 2003); esto puede en alguna manera aplicarse a las necesidades de las fuerzas policiales para detectar y disuadir a los delincuentes.

La predicción del comportamiento delincuencial es la capacidad de encontrar los patrones de actividades ilegales, predecir su ubicación y el tiempo probable de los delitos e identificar a los delincuentes (Mena, 2003).

4.4.3. Información del crimen

El crimen o delito se define como: la acción u omisión voluntaria, típica, antijurídica y culpable (Jirón, 2013).

La información referente a los delitos que han sido denunciados o forman parte de las investigaciones realizadas por la policía, forman parte de los registros policiales de las novedades, denuncias e investigaciones criminales de la PNC.

Esta información es de carácter relevante, pues se cuenta con una extensa base de datos histórica sobre los delitos cometidos, víctimas, sospechosos, grupos criminales, información relacionada a las denuncias y a los casos de investigación criminal, entre otros. La cual puede ser analizada, mediante el proceso KDD y llevar acabo análisis de predicción del delito.

4.4.3.1. Registro de la información del crimen

Cada vez que se comete un delito, y éste es objeto de investigación por parte de la PNC, ya sea a solicitud del Ministerio Público, flagrancia o por denuncia presentada; los oficiales de policía asignados realizan sus

investigaciones, visitando la escena del crimen o a las víctimas que han sido agredidas por un delincuente.

Como resultado de sus investigaciones, registran la información en papel, teléfonos. Entre la información que se recolecta, se encuentra la siguiente:

- Fecha y hora del delito
- Lugar del delito
- Tipo de delito
- Modus operandi: forma de operar
- Móvil
- Información de las victimas
- Información de sospechosos

Esta información es almacenada en los sistemas de registro de la PNC, para su posterior análisis y obtención de reportes estadísticos.

4.4.3.2. Ambiente de la criminología

Comprender el comportamiento de los delincuentes, tanto a nivel individual, como en estructuras criminales, tiene mucha importancia en la predicción del delito, puesto que es necesario tener experiencia en entender los patrones de criminalidad para saber definir los modelos e interpretar sus resultados (Kumar y Chandrasekar, 2011).

Este comportamiento tiene que ser analizado con la información que se registra sobre los hechos delincuenciales, sin embargo, el alcance de los modelos de predicción no debe limitarse únicamente al registro de la información policial, pues existen otros componentes que pueden influir en el comportamiento

delictivo, por ejemplo: comunidades con poca presencia policial o poca infraestructura urbana.

4.4.4. Técnicas actuales de predicción del crimen

El objetivo final en la vigilancia policial es predecir cuándo y dónde ocurrirá el próximo crimen o el conjunto de delitos, pero, esto no es del todo posible. Se han hecho varios intentos en el campo de la predicción, algunos de carácter limitado, sin embargo, el enfoque ha sido de carácter particular, un individuo, un grupo criminal, un patrón criminal o una comunidad (Kumar y Chandrasekar, 2011).

A continuación, se presentan algunas técnicas para la predicción del delito:

4.4.4.1. Métodos estadísticos

Los métodos estadísticos se basan en la probabilidad de ocurrencia de los delitos, según la información histórica. Por ejemplo: la mayoría de los delincuentes por robo (69 %) y violencia (55 %) viven a menos de una milla de la escena del crimen. Solo el 8 % de los ladrones y el 15 % de los delincuentes violentos viven a más de 5 millas de distancia de la escena del crimen. Las estadísticas también indican que la escena de un crimen es una característica clave de la dirección o la base de un delincuente. Los delitos ocurren muy cerca de la residencia del delincuente y hay un patrón de decaimiento a distancia para los viajes delictivos (Kumar y Chandrasekar, 2011).

4.4.4.2. Métodos misceláneos

Los métodos y las herramientas que provee la minería de datos son utilizados para la predicción del crimen, la prevención y la detección de patrones criminales, la cual reúne las disciplinas de estadística, aprendizaje automático, inteligencia artificial, criminología, psicología y tecnología de bases de datos (Adderley y Musgrove, 2017) han documentado el desarrollo de herramientas de investigación que aprovechan al máximo el poder de los equipos de cómputo, como un mecanismo para ayudar a la solución de los delitos mayores y de mayor volumen, cada uno de los cuales requiere una estrategia de investigación diferente.

Entre otras técnicas que han sido utilizadas por varios investigadores, se ha utilizado la extracción de entidades para descubrir los patrones que identifican los nombres de las personas, sus direcciones, vehículos y otras características. Algunos de los enfoques, como el comparador de cadenas, el análisis de redes sociales y la detección de desviaciones, para usarlos en datos de delitos para comprender el comportamiento delictivo (Kumar y Chandrasekar, 2011).

4.4.4.3. Métodos de información geográfica

Uno de los métodos utilizados para realizar prevención del delito es que se desarrollen puntos críticos de delincuencia en áreas de la comunidad que puedan ser etiquetados como generadores de delitos, como áreas de entretenimiento y centros comerciales. Sin embargo, los puntos calientes inestables probablemente sean el resultado de los delincuentes prolíficos que atacan un área a intervalos irregulares (Kumar y Chandrasekar, 2011). Las técnicas utilizadas por las fuerzas policiales para identificar puntos calientes no siempre son consistentes. Los problemas delictivos en áreas designadas como puntos críticos pueden ser

momentáneos y pueden desaparecer antes de que los recursos se asignen oficialmente a esas áreas.

Aparte de ser un generador de delitos, hay una variedad de razones por las cuales un área geográfica particular se considera un punto caliente. Por ejemplo, la tasa de criminalidad puede ser causada por un delincuente prolífico que es liberado de la prisión o debido a un evento comunitario en particular que está ocurriendo. Estos puntos calientes pueden ser usados como buenos predictores del crimen y de la delincuencia (Kumar y Chandrasekar, 2011).

El mapeo de puntos calientes es útil para dar seguimiento a la prevención del delito, sin embargo, es necesario que la experiencia policial se haga manifiesta, para identificar aquellos otros patrones, que no se puedan modelar estadísticamente en la predicción del delito.

4.5. Metodología para realizar el análisis predictivo del crimen

Considerando la naturaleza del delito, los patrones criminales, la información registrada por la policía y utilizando minería de datos y los métodos de agrupamiento y clasificación (K-MEANS, CLUSTERING JERARQUICO, FUZZY-C, CLUSTERING JERÁRQUICO), es necesario crear una metodología para diseñar un modelo de predicción del crimen.

Para cumplir con el objetivo, se presentan a continuación los pasos de esta metodología. Se presenta una descripción general del modelo de investigación criminal utilizado por la DEIC, como preámbulo para entender un poco el actuar de la investigación de los delitos en Guatemala.

4.5.1. Modelo de investigación criminal en Guatemala

Una de las principales funciones de la DEIC es la de:

> Realizar todas las diligencias necesarias para la recolección de indicios o medios de prueba que pueda demostrar la existencia de un hecho delictivo e individualice a los posibles autores, con el objetivo de fundamentar su persecución penal, siempre bajo la instrucción del Ministerio Público (DEIC, 2009, p.2).

Para llevar a cabo esta importante función la DEIC está organizada en diferentes departamentos que se dedican a la investigación de determinado tipo de delito, de la siguiente manera:

- "Departamento de investigación de delitos contra la vida e integridad de las personas: Es el encargado de la investigación de los delitos dolosos, que atenten contra la Vida e Integridad de las personas" (DEIC, 2009, p.10).

- Departamento de investigación de delitos sexuales, trata de personas, de la niñez y adolescencia y delitos conexos: es responsable de la investigación de los delitos que atenten contra la libertad y seguridad sexual, contra el pudor de las personas, conductas relacionadas con la trata de personas y conductas delictivas relacionadas con Niñez y Adolescencia, para comprobar la existencia del hecho criminal e individualizar a los responsables (DEIC, 2009, p.14)
- "Departamento de investigación de delitos patrimoniales especializados: Es responsable de la investigación de los delitos que atenten contra el patrimonio de las personas particulares y del Estado, para comprobar la

existencia del hecho criminal e individualizar a los responsables" (DEIC, 2009, p.20).

- "Departamento de investigación de delitos de organizaciones criminales: estará encargado de la investigación de los delitos que presenten indicios de la existencia de una organización criminal, para comprobar la existencia del hecho criminal e individualizar a los responsables" (DEIC, 2009, p.38).

Cada departamento se compone de brigadas de investigación, las cuales están compuestas por grupos de cinco investigadores que realizan las tareas de recolección de la información asociada a los delitos investigados por medio de diligencias (DEIC, 2009), entre las más importantes: procesamiento de la escena del crimen, entrevistas a víctimas, sospechosos y testigos, peritajes, allanamientos, inspección y registro. En las cuales se recolecta información de la denuncia, personas, objetos o evidencias, diligencias de investigación, entre otros

Con la información de los casos de investigación criminal, recolectada y registrada en los sistemas de información de la DEIC, los investigadores realizan informes del seguimiento de los casos a las agencias fiscales del Ministerio Público, quienes son los que dan los lineamientos de la investigación criminal. Entre los informes que se presentan, están los siguientes: informes de veinticuatro horas, informes de cuarenta y ocho horas, informes de setenta y dos horas e informes de seguimiento (DEIC, 2009).

4.5.2. Estandarización de la información

La estandarización de la información es muy importante para que se garantice la calidad de la información que se realizara por KDD durante las aplicaciones de la predicción.

La información es útil en primera instancia para el diseño de los modelos de predicción, pero también para realizar las pruebas, mediante una selección de muestras de la información. Para esto se necesita información de calidad, la cual se obtiene de diferentes formatos, repositorios y bases de datos.

Por esta razón se debe llevar a cabo un proceso de extracción, transformación y carga (ETL), de las fuentes de datos que hayan sido seleccionados para formar parte de la información que se utilizará en la predicción del delito (Revatthy y Satheesh, 2012).

Para fines de este proyecto se utilizará la información de los casos de investigación originada mediante las denuncias de los delitos que es registrada por la DEIC. Para ello se realizarán procesos de extracción, transformación y carga.

4.5.3. Modelado, construcción de modelos y patrones

En la fase de modelado se construye un modelo de predicción en el cual se identifican las variables que se necesitan incluir en la predicción, para ello es necesario identificar patrones delincuenciales y así elegir los atributos que formaran parte del modelo.

Es importante durante el diseño de modelos predictivos determinar el tipo de que se va a diseñar, porque esto afecta la validación y la implementación. Entre los tipos de modelos se encuentran los siguientes: Modelos de agrupamiento, modelos de clasificación, modelos de regresión, modelos de inteligencia artificial.

El presente proyecto está delimitado a la construcción de modelos de agrupamiento: Clustering jerárquico, K-MEANS, FUZZY C y clasificación: red de Bayes.

Cuando se ha seleccionado el modelo que se va a diseñar, se procede a buscar patrones y comportamientos criminales. En este punto es importante la delimitación del entorno donde se realizará la predicción, la cual puede ser: Geográfica o por tipificación de los delitos, por ejemplo: Delitos contra la vida e integridad de las personas, y dentro de esta se puede enfocar únicamente a homicidios, los que a su vez se puede enfocar específicamente al femicidio.

Una vez delimitado el alcance de la predicción e identificado los patrones de criminalidad, se procede a la identificación de variables que influyen en los patrones y el comportamiento criminal. Para ello se utiliza la información histórica de los acontecimientos o hechos que se requieren analizar, realizando análisis de la información (Ahishakiye y Niyonzima, 2017).

4.5.4. Entrenamiento y validación del modelo

Dependiendo del tipo de modelo seleccionado en el análisis, así también será la forma o el método utilizado para realizar la validación de este.

Por ejemplo, para modelos del tipo agrupamiento las técnicas más utilizadas para la validación de los modelos clúster (Jerárquico, Fuzzy C, K-MEANS) se encuentran:

- Estrategia de la distancia mínima
- Estrategia de la distancia máxima
- Estrategia de la distancia promedio
- Estrategia de la distancia promedio ponderada
- Métodos basados en el centroide

Cada método de validación se basa en el concepto de distancia o similitud, el cual es definido por Witold (2005), de la manera siguiente:

Es el componente esencial de cualquier forma de agrupación que nos ayude a navegar a través del espacio de datos y formar agrupaciones. Al computar la disimilitud, podemos sentir y articular cómo se comportan juntos dos patrones en función de esta cercanía, estos se asignan al mismo grupo. Formalmente, se considera la disimilitud d (**x, y**) entre **x** y. para ser una función de dos argumentos que cumpla las siguientes condiciones:

$$d(x, y) \geq 0 \text{ para cada x and y}$$
$$d(x, x) = 0 \text{ para cada x}$$
$$d(x, y) = d(y, x)$$

El proceso de validación de modelos debe incluir una serie de pruebas en las que se deben seleccionar conjuntos de datos (Data Set) de calidad que puedan arrojar resultados con los que se pueda comparar la predicción efectuada con los hechos históricos que se está tratando de predecir.

Witold (2005) destaca entre los aspectos más importantes de la validación de modelos, los siguientes:

- El número de clústeres es impulsado por la aplicación y el uso que se le quiere dar.

- Como el usuario está en medio del proceso de análisis de datos, es beneficioso considerar un número variable de grupos y analizar los resultados producidos.

- Obviamente, sería útil tener algo de automatización en este proceso. Esta información debe venir con una medida sintética con la que se pueda evaluar la calidad de la estructura descubierta.

- Para plantear el problema de una manera diferente, ¿cuál es el número óptimo de grupos?, o, mejor aún, ¿cuál es el número preferido de clústeres, dada la geometría subyacente impuesta en el proceso de clustering?

En cuanto al proceso de automatización de la validación y el entrenamiento, se ha tomado la decisión de utilizar la herramienta de análisis de datos y procesos estadísticos R, la cual cuenta con componentes para el uso de modelos y algoritmos de agrupamiento y clasificación.

4.5.5. Predicción e interpretación de resultados

En esta fase de la metodología los modelos predictivos deben haber sido afinados lo suficientemente para que los resultados de la predicción sean óptimos.

Los analistas encargados de la predicción e interpretación realizarán diferentes cambios en las variables de los modelos, esto a nivel de criterios o

tendencias, según la delincuencia este afectando zonas geográficas, o los índices de determinado delito este en aumento. Aquí se debe considerar mucho las necesidades actuales en materia de prevención del delito y las políticas que se tengan en las fuerzas policiales (Tahani, Rsha y Lor, 2015).

La interpretación asignada a los resultados debe estar sujeta al escrutinio de los analistas expertos en criminalística, para determinar si la misma es confiable o incluye ciertos patrones y comportamientos no planeados.

Los resultados arrojados por la predicción deben realimentar los modelos y en determinados momentos ajustarlos o incluir nuevas variables que inicialmente no fueron tomadas en cuenta, pero que afectan la predicción (Tahani *et al.,* 2015).

Los analistas deben entregar la información interpretada en formatos que puedan ser leídos por las áreas encargadas de la prevención del delito, para realizar las coordinaciones necesarias.

4.6. Arquitecturas de software utilizando microservicios

No existe una definición estándar para el término ‘arquitectura de software’. Sin embargo, se puede dar una definición aproximada de la manera siguiente:

> La arquitectura se define por la práctica recomendada como la organización fundamental de un Sistema, incorporado en sus componentes, sus relaciones entre sí y con el medio ambiente, y los principios que rigen su diseño y evolución. La arquitectura describe la estructura del sistema en términos de componentes y cómo interactúan, también define las reglas de diseño de todo el sistema y considera cómo un sistema puede cambiar (Gorton, 2006, p.20).

Las tecnologías de la información han ido evolucionando en los últimos tiempos, como resultado de las distintas necesidades en cuanto a disponibilidad, crecimiento de los volúmenes de información, nuevos giros de negocios, entre otros.

Por lo que los ingenieros arquitectos de software y las grandes compañías como Microsoft, Google, Netflix, Amazon han ido desarrollando investigaciones de nuevos modelos arquitectónicos de soluciones de software e infraestructura para dar soporte a las continuas necesidades del día a día. Como lo expresa Sam Newman: “Las nuevas compañías tecnológicas operan de diferentes maneras para crear sistemas de TI que ayudar a hacer más felices a sus clientes y sus propios desarrolladores” (Newman, 2015).

Dentro de las arquitecturas de software se pueden mencionar:

- Modular: el software está dividido en módulos con un nivel de acoplamiento muy alto.

- Cliente servidor: el software se divide en dos partes principales, generalmente la primera es la vista del usuario y la otra es la del procesamiento de transacciones.

- Arquitectura en capas: el software se diseña en diferentes componentes o capas, cada una con un alto nivel de definición y conceptualización y con un acoplamiento medio. Las capas están organizadas de tal manera que corresponden a funcionalidades específicas, permitiendo el escalamiento.

- Orientada a servicios (SOA): es un enfoque de diseño donde múltiples servicios colaboraran para proporcionar un conjunto final de capacidades.

Un servicio aquí típicamente significa un proceso del sistema de información completamente separado. La comunicación entre estos servicios se produce a través de llamadas por medio de una red en lugar de llamadas a métodos dentro de un límite de proceso. Tiene como objetivo promover la reutilización del software; dos o más aplicaciones de usuario final, por ejemplo, podrían usar los mismos servicios. También apunta hacia facilitar el mantenimiento o la reescritura del software, ya que, en teoría, se puede reemplazar un servicio (Newman, 2015).

- Arquitectura de microservicios: según Newman, la arquitectura de microservicios emerge como una tendencia, o patrón, a partir del uso en el mundo real, es decir de las necesidades de cambios en cuanto a los giros de negocio (dominios), entrega continua de cambios al software, virtualización bajo demanda, automatización de la infraestructura, equipos pequeños autónomos, sistemas a escala.

Las arquitecturas de microservicios dan una mayor libertad para reaccionar y tomar decisiones diferentes, permitiendo responder más rápido al cambio inevitable que afecta al mundo de los negocios.

4.6.1. Conceptos generales de los microservicios

Se puede definir una arquitectura de microservicios, como aquella que se orienta al servicio, es decir: la orientación al servicio significa encapsular los datos con la lógica empresarial que opera en los datos, con el único acceso a través de una interfaz de servicio publicada. No hay base de datos directa y el acceso está permitido desde fuera del servicio, por lo que no hay intercambio de datos entre los servicios (Carneiro y Schmelmer, 2016).

4.6.1.1. ¿Qué son los servicios?

Un servicio es un sistema que satisface una necesidad pública. Los servicios de software deben satisfacer las necesidades de una o más aplicaciones cliente internas o externas, por lo tanto, la interfaz de dichos servicios debe diseñarse desde la perspectiva del cliente, optimizando para que sea de utilidad al consumidor (Carneiro y Schmelmer, 2016). Si bien lo ideal es que cada servicio sea diferente, ya que se optimiza para los casos de uso de sus clientes, generalmente se puede decir que un servicio consiste en una pieza de funcionalidad y su conjunto de datos asociados.

4.6.1.2. ¿Qué son los microservicios?

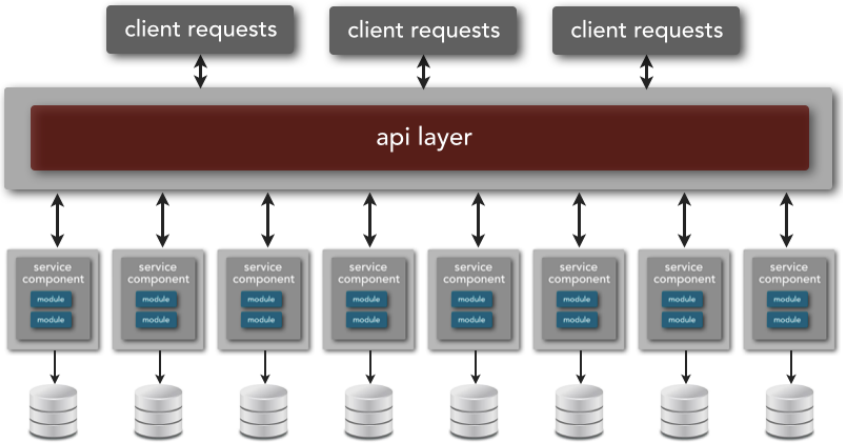
Los microservicios son entidades autónomas que se desarrollan con la finalidad de cumplir con un objetivo funcional del negocio. Un aspecto importante es evitar el acoplamiento entre las demás unidades funcionales de software, con la finalidad de mantener la autonomía. Estos microservicios deben ser capaces de cambiar independientemente entre sí y ser implementados por ellos mismos sin requerir que los consumidores cambien (Newman, 2015).

4.6.2. Componentes básicos de la arquitectura

La topología básica de una arquitectura de microservicios se divide en dos componentes principales: capa de servicios y la capa *API* (*Application Programming Inteface*). Desde el punto de vista de la implementación, puede tener otros componentes tales como un registro de servicio y un componente de descubrimiento, un componente de monitoreo de servicios, y un administrador de despliegue de servicios. Pero arquitectónicamente esos componentes serían considerados como servicios de infraestructura (Richards, 2016).

En la figura 2 se presenta la topología general de microservicios descrita anteriormente.

Figura 2. **Topología de la arquitectura de microservicios**



Fuente: Richards. (2006) *Microservices vs. Services-Oriented-Architecture.*

4.6.2.1. Capa de servicios

La capa de microservicios se compone de pequeños componentes de software que se especializan en una tarea y trabajan juntos para lograr una tarea de alto nivel. Un microservicio es una unidad de trabajo autónoma que puede ejecutar una tarea sin interferir con otras partes del sistema, similar a lo que es una posición de trabajo para una empresa. Esto tiene una serie de beneficios que se pueden utilizar en favor de la ingeniería y ayudar a escalar los sistemas de una empresa (Gonzalez, 2016).

Según González (2016), no existe una regla de oro para diseñar la capa de microservicios, sin embargo, se pueden listar los siguientes principios claves:

- Los microservicios son unidades de negocio que modelan los procesos de la empresa.

- Son puntos finales inteligentes que contienen la lógica empresarial y se comunican usando canales y protocolos simples.

- Las arquitecturas orientadas a microservicios están descentralizadas por definición. Esto ayuda a construir software robusto y resistente.

Los microservicios pueden ser desarrollados utilizando diferentes tecnologías y diferentes lenguajes de programación. Sin embargo, se debe utilizar un lenguaje estándar de comunicación.

El estándar de comunicación entre microservicios que ofrece mayores beneficios a la arquitectura es *REST* (Representational State Transfer). En el año 2000, Roy Fielding presentó su tesis doctoral, Estilos arquitectónicos y Diseño de Arquitectura de software basada en red. En él acuñó el término *REST*, un estilo arquitectónico para la distribución de sistemas hipermedia. En pocas palabras, *REST* es un estilo arquitectónico definido para ayudar a crear y organizar sistemas distribuidos. La palabra clave de esa definición debe ser estilo, porque un aspecto importante de *REST* es que es un estilo arquitectónico: no es una pauta, ni una norma, ni nada que implique que hay un conjunto de elementos difíciles. Es un conjunto de reglas a seguir para terminar teniendo una arquitectura *REST*ful (Doglio, 2015).

Según Doglio, los principales beneficios de utilizar *REST* son los siguientes:

- Rendimiento: el estilo de comunicación propuesto por *REST* está destinado a ser eficiente y simple, permitiendo un aumento de rendimiento en los sistemas que lo adoptan.

- Escalabilidad de la interacción de componentes: cualquier sistema distribuido debería poder manejar este aspecto lo suficientemente bien, y la interacción simple propuesta por *REST* en gran medida permite esto.

- Simplicidad de la interfaz: una interfaz simple permite interacciones más simples entre sistemas, que a su vez pueden otorgar beneficios como los mencionados anteriormente.

- Modificabilidad de los componentes: la naturaleza distribuida del sistema y la separación de las inquietudes propuestas por *REST*, permite que los componentes sean modificados independientemente entre sí a un costo y riesgo mínimos.

- Portabilidad: significa que puede ser implementado y consumido por cualquier tipo de tecnología.

- Confiabilidad: la restricción sin estado propuesta por *REST* permite La recuperación más fácil de un sistema después de un fallo.

- Visibilidad: una vez más, la restricción sin estado propuesta tiene el beneficio adicional de mejora de la visibilidad, porque cualquier sistema de monitoreo no necesita buscar más de un solo mensaje de solicitud para determinar el estado completo de dicha solicitud.

4.6.2.2. Capa *API*

Una *API* es un componente de software desarrollado para realizar operaciones de entradas, salidas y tipos subyacentes, las cuales definen funcionalidades que son independientes de sus respectivas implementaciones, permitiendo que las definiciones y las implementaciones varíen sin comprometer la interfaz. Una buena *API* facilita el desarrollo de un programa al proporcionar todos los bloques de construcción, que luego son ensamblados por el programador (Carneiro y Schmelmer, 2016).

Según Carneiro y Schmelmer (2016), las *API* no tienen un uso exclusivo en las arquitecturas de microservicios ni de otras aplicaciones basadas en web. Diferentes tipos de sistemas informáticos los emplean para permitir a los desarrolladores de software crear aplicaciones para éstos, utilizando especificaciones de interfaz de programación publicadas.

Las *API* son importantes en el ámbito de los microservicios porque la mayoría de las comunicaciones entre las aplicaciones pasan a través de llamadas a la *API*. Cada uno de sus microservicios expondrá sus funcionalidades a través de un conjunto de puntos finales que tienen un conjunto bien definido de entradas y salidas, y que realizan las tareas que el servicio es responsable de ejecutar.

Dependiendo del tipo o línea de negocio, la *API* puede ser el producto real que provee y expone a los clientes que utilizan los microservicios; si es así, debe tomarse todo el tiempo en el diseño y desarrollo que se necesite antes de exponer ese servicio a los clientes. Si es posible, es necesario crear un proceso que ayude a comenzar a desarrollar y exponer la *API* de manera temprana a un pequeño

subconjunto de usuarios, para que comience a validar esa API de forma iterativa. La documentación de esta es muy importante (Carneiro y Schmelmer, 2016).

Para Carneiro y Schmelmer (2016), las *API* cuentan con características generales en cuanto a diseño y desarrollo para cumplir con los objetivos principales de comunicación entre los clientes y los microservicios.

Estas características principales son las siguientes:

- Clara denominación: cada *API* debe tener un nombre obvio que tenga sentido para los usuarios que la utilizarán. Este es uno de los problemas más difíciles en la ingeniería de software, pero es especialmente importante cuando se crea un punto final en que otros desarrolladores llamarán para que sea muy fácil de recordar.

- Enfoque: las *API* deben hacer una cosa y hacerlo bien. El propósito de cada *API* debe ser claro para todos sus clientes. Si la razón de una API es fácil de entender, recordar y usar, entonces logrará este objetivo.

- Completo: una *API* necesita poder cumplir su razón de ser; la funcionalidad que anuncia debe ser adecuada para su implementación. Esto puede parecer contrario a la característica anterior, pero en realidad es complementario.

- Intuitividad: las mejores *API* son aquellas a las que puedes llamar sin tener que pensar mucho en ellas. Si un consumidor de *API* puede intuir qué hace la *API* revisando la documentación, entonces se ha logrado este objetivo.

- Consistencia: es imperativo que se tenga un conjunto de reglas consistentes sobre el desarrollo de *API*. Cuando se adoptan convenciones de nomenclatura coherentes para la forma en que llama a las *API* y cómo se esperan las entradas / salidas, éstas tendrán un uso fácil e intuitivo.

Para ser consistentes las *API*, necesitan utilizar estándares hipermedia para las entradas y salidas. Esto beneficiara el uso de los clientes que la utilizan y permitirá un lenguaje común entre los clientes y la capa de microservicios.

Los formatos más utilizados de Hipermedia son los siguientes:

- *HAL*
- *JSON*
- *JSON-LD*
- *SIREN*
- *Collection+JSON*

Cada uno de estos tipos de hipermedia tiene sus reglas que describen el formato exacto que las solicitudes y respuestas deben ajustarse a diversos grados de robustez, verbosidad y rigor (Carneiro y Schmelmer, 2016).

4.6.3. Beneficios de la arquitectura

Los beneficios de la arquitectura de microservicios son muchos y variados. Entre los más importantes, se pueden mencionar los siguientes:

- Uso de tecnologías heterogéneas: con un sistema compuesto por múltiples servicios de colaboración, podemos decidir utilizar diferentes tecnologías dentro de cada una. Esto permite elegir la herramienta adecuada para cada trabajo, en lugar de tener que seleccionar un enfoque

más estandarizado y de talla única que a menudo termina siendo el mínimo común denominador. Si una parte del sistema necesita mejorar su rendimiento, se puede decidir utilizar una tecnología diferente que es más capaz de alcanzar los niveles de rendimiento requeridos. También se puede decidir que la forma en que almacenamos los datos debe cambiar para diferentes partes del sistema. Por ejemplo, para una red social, se pueden almacenar las interacciones de los usuarios en una base de datos orientada a gráficos para reflejar la naturaleza altamente interconectada de un gráfico social, pero tal vez las publicaciones que hacen los usuarios podrían almacenarse en un almacén de datos orientado a documentos (Newman, 2015, p.15).

- Resiliencia: un concepto clave en la ingeniería de resiliencia es el mamparo. Si un componente de un sistema falla, pero esa falla no se conecta en cascada, puede aislar el problema y el resto del sistema puede seguir funcionando. Los límites del servicio se convierten en sus mamparos obvios. En un servicio monolítico, si el servicio falla, todo deja de funcionar. Con un sistema monolítico, se puede ejecutar en varias máquinas para reducir la posibilidad de fallo, pero con microservicios, se pueden construir sistemas que manejen la falla total de los servicios y degraden funcionalidad en consecuencia (Newman, 2015).

- Escalabilidad: con un servicio grande y monolítico, se tiene que escalar todos juntos. Una pequeña parte del sistema general está limitado en el rendimiento, pero si ese comportamiento está bloqueado en una aplicación monolítica gigante, sé tiene que manejar el escalamiento de todo como una pieza. Con microservicios, solo se escalan aquellos servicios que necesiten escalar, lo que permite ejecutar otras partes del sistema en hardware más pequeño y menos potente (Newman, 2015).

- Despliegue rápido: permite hacer cambios a los microservicios e implementarlos de forma independiente del resto del sistema. Esto permite desplegar el código más rápidamente. Si existe algún problema, también es fácil revertir el cambio. Por lo tanto, se pueden desplegar a los clientes los cambios más rápidamente (Newman, 2015).

- Alineamiento de la organización: la arquitectura del sistema se empareja con la arquitectura del negocio, esto permite ser más productivo, al dividir el trabajo de desarrollo en partes, según la estructura de negocio (Newman, 2015).

- Composabilidad: el acoplamiento de servicios permite crear servicios más complejos, conforme la necesidad del negocio cambia.

- Optimización y reemplazo: con una arquitectura orientada al servicio, los servicios pueden ser optimizados por separado, cuando fuere necesario o reemplazados, sin necesidad de que todo el sistema se vea afectado (Newman, 2015).

5. PRESENTACIÓN DE RESULTADOS

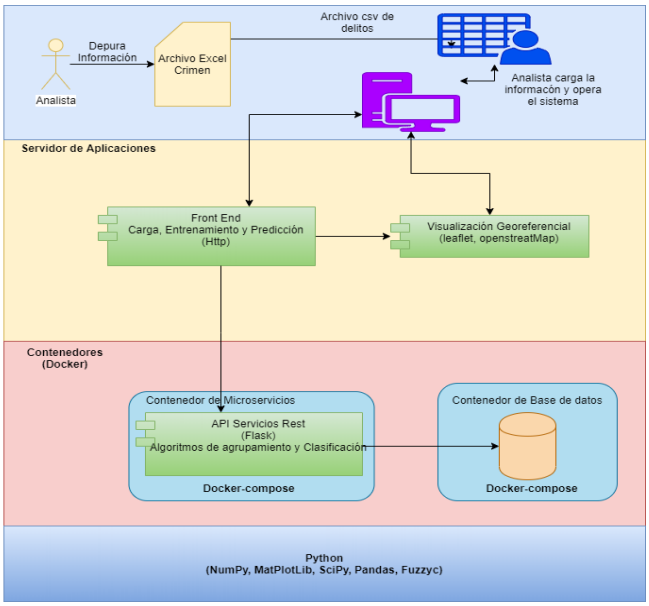
5.1. Análisis, diseño y desarrollo del prototipo

A continuación, se describen los componentes de diseño y desarrollo del prototipo.

5.1.1. Representación y descripción de la arquitectura

En la figura 3 se presenta el diagrama general de componentes de la arquitectura utilizada en el desarrollo del prototipo.

Figura 3. **Diagrama general de componentes de la arquitectura**



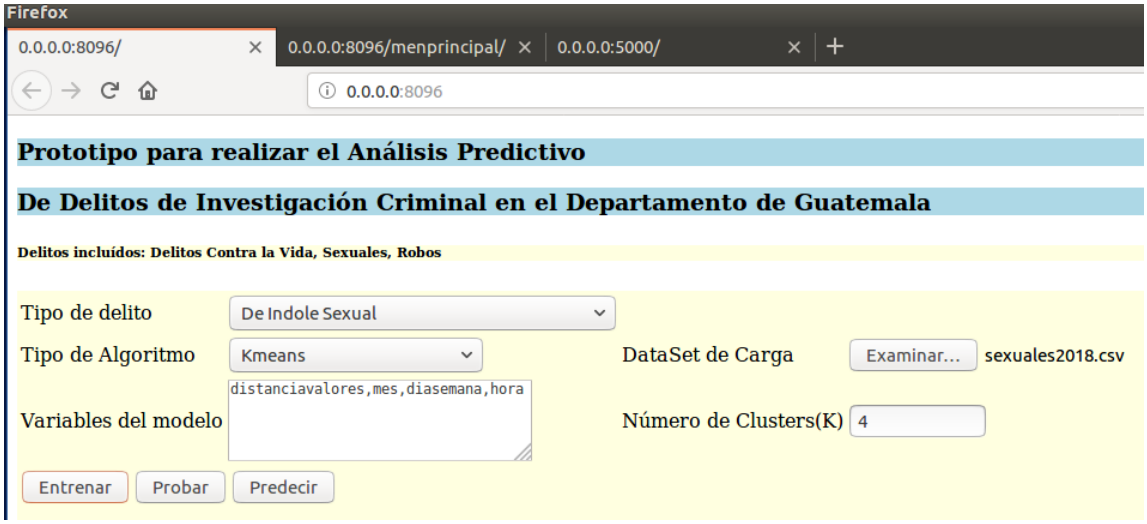
Fuente: elaboración propia, utilizando Google Draw.

El sistema de información implementa tres capas principales:

- Capa de visualización o Front-End

Esta capa contiene los componentes necesarios para que el usuario pueda realizar el análisis de la predicción de los delitos de investigación criminal, a través de los algoritmos de agrupamiento y clasificación. Para ello el usuario deberá ingresar los datasets de información, previamente depurados y que servirán de insumo de los algoritmos. Como se muestra en la Figura 4

Figura 4. **Pantalla principal del prototipo**



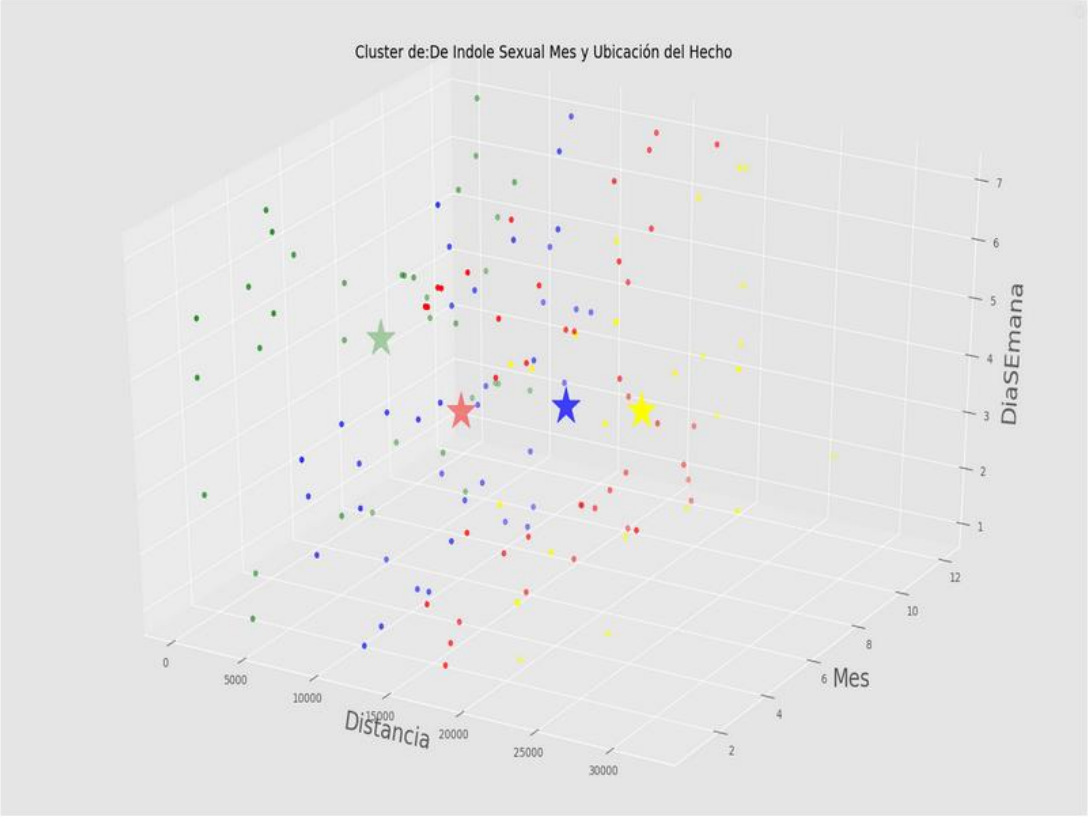
Fuente: elaboración propia utilizando Flask.

De la misma manera se desarrollaron los componentes informáticos para la visualización de los resultados, en gráficas cartesianas y de visualización georreferencial, como se muestra en las figuras 5 y 6.

Figura 5. **Visualización en coordenadas cartesianas del agrupamiento**

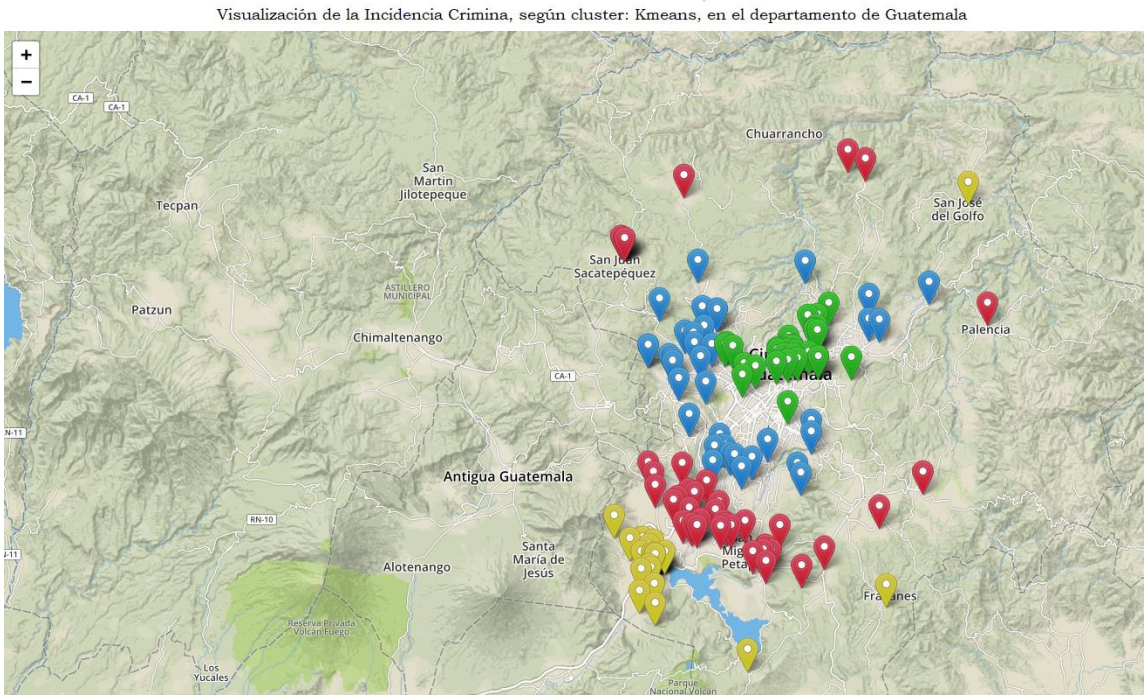
Algoritmo: Kmeans; **Número de Cluster**: 4

Cantidad x Cluster: **44** Rojos; **38** azules; **34** verdes; **28** amarillos de un total de: **144** delitos



Fuente: elaboración propia utilizando Matplot-Lib.

Figura 6. **Visualización georreferencial de los delitos, después del agrupamiento**



Fuente: elaboración propia utilizando Leatflet.

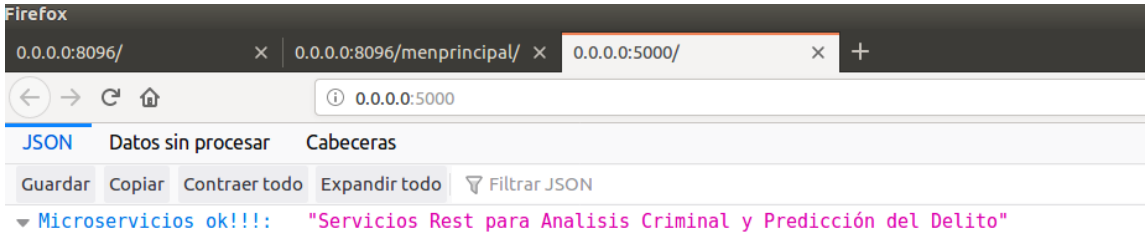
Los componentes informáticos fueron desarrollados con herramientas CGI en Python, utilizando Flask y para la visualización georreferencial, se utilizó LeatFlet y Open Street Map.

- Capa de microservicios

Cómo se muestra en la figura 3, se desarrollaron servicios Rest para el entrenamiento y la predicción del delito, utilizando para la codificación Flask de Python y para contener los servicios se utilizó Docker y para la comunicación entre ellos Docker-Compose.

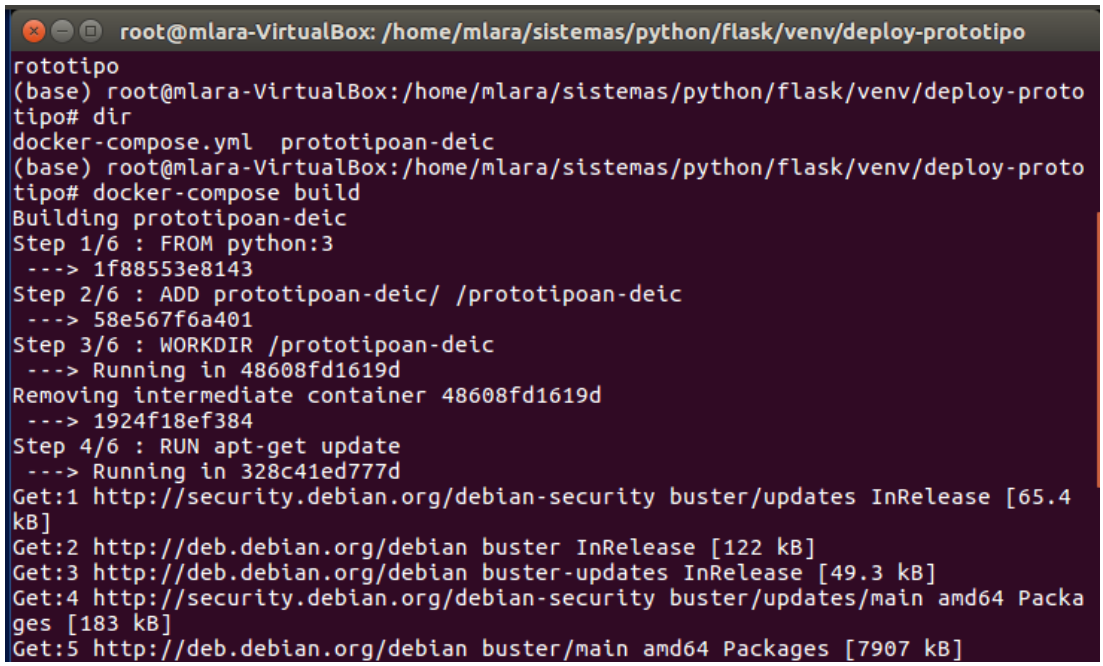
A continuación, en las figuras 7, 8 y 9, se muestra la página web, donde los servicios Rest, están alojados en un servidor CGI de Python y disponibles en un contenedor de Docker-Compose.

Figura 7. **Servidor de servicios Rest para la predicción del delito**



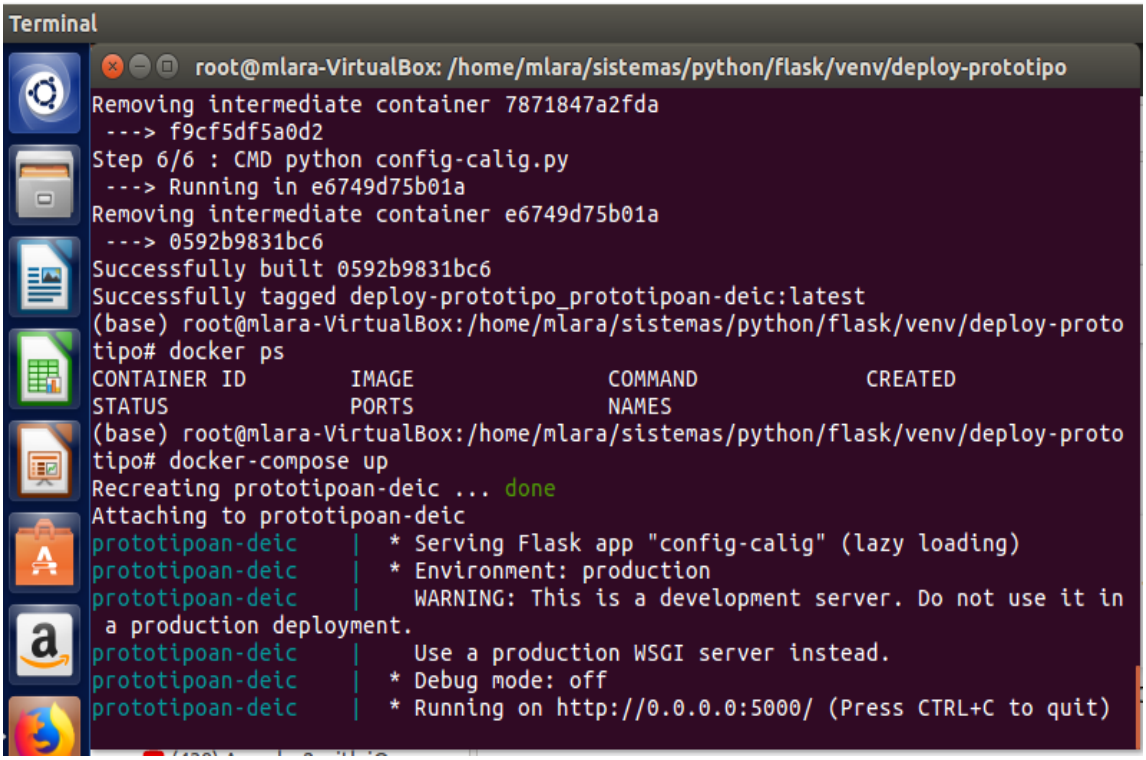
Fuente: elaboración propia utilizando Flask.

Figura 8. **Compilación y desplegado del contenedor que aloja los microservicios**



Fuente: elaboración propia utilizando Terminal.

Figura 9. **Visualización del alojamiento del contenedor de microservicios del prototipo**



Fuente: elaboración propia utilizando Terminal.

- Capa de servicios de soporte

Esta capa esta sostenida sobre las librerías de Python que prestan servicios de cálculos matemáticos, manejo de conjuntos de datos (Dataset), gráficas y algoritmos de aprendizaje automático, entre ellas: Pandas, Numpy, Matplot-lib, Scikit-Learn.

5.1.2. Proceso de recolección y preparación de los datos granulares para los modelos predictivos

Para la obtención de los modelos, se establecieron patrones de incidencia criminal con el fin de determinar la relación que existe entre las diferentes características de los hechos delictivos, por medio de los conglomerados o clústeres, ya que se cuenta con series de datos de eventos con distintas categorías o características y se buscaron agrupaciones de acuerdo a ellas, estas agrupaciones a simple vista no parecen tener coincidencia o similitud y no pueden ser inferidos de ésta manera, por la amplia cantidad de información.

El proceso para la identificación de estos patrones criminales y la obtención de las variables de los modelos sigue los siguientes pasos:

- Delimitación de los hechos delictivos: se determinó que, como parte del estudio, únicamente se analizaran los delitos contra la vida, que incluye las lesiones y los homicidios; delitos de índole sexual y delitos patrimoniales que contiene a los robos y los hurtos.

- Establecimiento de la fuente de información: después de revisar los archivos y algunos sistemas de la DEIC, se logró determinar que la fuente más confiable del registro de los hechos delictivos, se encuentra en la Jefatura de Planificación Estratégica (JEPEDI), la cual cuenta con una unidad de georreferenciación de los delitos cometidos que se obtienen del sistema de estadísticas y novedades de los hechos que ocurren a nivel república y que son reportados por las diferentes denuncias de la población a las comisarías. Estos delitos son clasificados, validados y la información es depurada y registrada en archivos de Excel, como se muestra en la tabla IV.

Tabla IV. **Extracto de hoja de cálculo de los delitos cometidos en el año 2018**

	A	B	C	D	E	G	H	I
1	NO.	POINT_X	POINT_Y	AÑO	FECHA	MES	DIA	HORA
19731	19736	-90.54610582	14.40646584	2018	10/12/2018	DICIEMBRE	LUNES	18:00
19732	19737	-91.44521981	15.36981193	2018	11/12/2018	DICIEMBRE	MARTES	23:00
19733	19738	-90.81121099	14.48561287	2018	11/12/2018	DICIEMBRE	MARTES	01:00
19734	19739	-90.67952569	14.60848926	2018	12/12/2018	DICIEMBRE	MIERCOLES	22:00
19735	19740	-89.92385485	14.29402821	2018	12/12/2018	DICIEMBRE	MIERCOLES	10:00
19736	19741	-91.09293354	15.28704304	2018	13/12/2018	DICIEMBRE	JUEVES	01:00
19737	19742	-90.19857463	14.78723406	2018	13/12/2018	DICIEMBRE	JUEVES	10:00
19738	19743	-89.96977985	14.94506624	2018	14/12/2018	DICIEMBRE	VIERNES	01:00
19739	19744	-90.57808921	14.53526570	2018	15/12/2018	DICIEMBRE	SABADO	02:00
19740	19745	-91.70946460	14.47127835	2018	16/12/2018	DICIEMBRE	DOMINGO	09:00
19741	19746	-90.29843290	14.27709114	2018	16/12/2018	DICIEMBRE	DOMINGO	11:00
19742	19747	-91.45463275	14.81094412	2018	16/12/2018	DICIEMBRE	DOMINGO	04:00
19743	19748	-90.49352757	14.64974124	2018	17/12/2018	DICIEMBRE	LUNES	06:00
19744	19749	-90.49098790	14.64919660	2018	17/12/2018	DICIEMBRE	LUNES	06:00
19745	19750	-90.55001598	14.61527475	2018	17/12/2018	DICIEMBRE	LUNES	04:00
19746	19751	-91.77880024	14.96734094	2018	17/12/2018	DICIEMBRE	LUNES	04:00
19747	19752	-91.71482666	15.40567157	2018	17/12/2018	DICIEMBRE	LUNES	12:00
19748	19753	-90.70965564	14.49511936	2018	17/12/2018	DICIEMBRE	LUNES	07:00
19749	19754	-90.80445112	14.48443989	2018	18/12/2018	DICIEMBRE	MARTES	02:00
19750	19755	-89.35118440	14.56465277	2018	19/12/2018	DICIEMBRE	MIERCOLES	02:00
19751	19756	-90.49883075	14.66044103	2018	20/12/2018	DICIEMBRE	JUEVES	20:00
19752	19757	-89.95694989	14.28777073	2018	25/12/2018	DICIEMBRE	LUNES	10:00
19753	19758	-90.45242925	15.46951921	2018	25/12/2018	DICIEMBRE	LUNES	21:00
19754	19759	-90.97630208	15.90985854	2018	25/12/2018	DICIEMBRE	LUNES	14:00
19755	19760	-90.93000194	14.73760996	2018	25/12/2018	DICIEMBRE	LUNES	12:00
19756	19761	-91.91042352	14.70025738	2018	25/12/2018	DICIEMBRE	LUNES	18:00
19757	19762	-89.54200787	14.79924311	2018	27/12/2018	DICIEMBRE	JUEVES	23:00
19758	19763	-90.30221494	14.27721286	2018	27/12/2018	DICIEMBRE	JUEVES	17:00
19759	19764	-89.54200787	14.79924311	2018	27/12/2018	DICIEMBRE	JUEVES	23:00
19760	19765	-90.52244094	14.63615597	2018	28/12/2018	DICIEMBRE	VIERNES	01:00
19761	19766	-89.91043284	16.92144238	2018	29/12/2018	DICIEMBRE	SABADO	12:00
19762	19767	-90.07030885	14.85338220	2018	29/12/2018	DICIEMBRE	SABADO	07:00
19763	19768	-89.15770326	17.06750625	2018	31/12/2018	DICIEMBRE	MARTES	03:00

Hechos Delictivos 2018

Fuente: Jefatura de Planificación Estratégica de la Policía Nacional Civil. (2018) *Hechos delictivos del año 2018*. Consultado el día 10 de noviembre de 2020. Recuperado del Sistema de Información Georreferencial (GIS)

Las variables que contiene la hoja de cálculo son las siguientes:

- No.: número correlativo de hecho delictivo.
- Point_x: longitud georreferencial del hecho.

- Point_y: latitud georreferencial del hecho.
- Año: año del hecho.
- Fecha: fecha del hecho.
- Mes: mes del hecho.
- Día: día del hecho.
- Hora: hora del hecho.
- Dirección: dirección del hecho.
- Colonia_ba: nombre de la colonia o barrió dónde ocurrió el hecho.
- Distrito: nombre del distrito que cubrió el hecho (para el departamento de Guatemala, distrito central).
- Comisaría: número de comisaría que cubrió el hecho.
- Departamento: nombre del departamento donde ocurrió el hecho.
- Municipio: nombre del municipio donde ocurrió el hecho.
- Código de municipio: código del municipio donde ocurrió el hecho.
- Zona: número de la zona donde ocurrió el hecho.
- Víctima: nombre de la víctima.
- Sexo: sexo de la víctima.
- Edad: edad de la víctima.
- Nacionalidad: nacionalidad de la víctima.
- Profesión: profesión u ocupación de la víctima.
- Delito: nombre del delito cometido.

- Creación de los conjuntos de datos: para la creación de los conjuntos de datos se siguieron los siguientes pasos:

    - Caracterización de las variables: se identificaron los siguientes tipos de variables:

- Categóricas: sexo, profesión, distrito, comisaría, departamento, municipio, nacionalidad
- Continuas: fecha y hora del hecho, latitud, longitud
- Discretas: año, mes y día del hecho, edad

- Limpieza de los datos: se eliminaron los datos que se consideraban incompletos, por ejemplo: en algunos registros de variables se encontraba la palabra ‘ignorada o ignorado’, algunas coordenadas georreferénciales no coincidían con el departamento y municipio indicado.

- Remoción de variables: se eliminaron del conjunto de datos las variables descriptivas, como, por ejemplo: la dirección del hecho, el nombre de la víctima.

- Transformación de variables: debido al análisis criminal y por cuestiones de precisión y requisitos de los algoritmos utilizados, se realizaron las siguientes transformaciones:

  - Variable fecha y hora del hecho: se dividió en las siguientes variables discretas:

    - ✓ Año del hecho.
    - ✓ Mes del hecho.
    - ✓ Día del hecho.
    - ✓ Día de la semana del hecho: número del día de la semana: 1 es domingo, 2 es lunes, entre otros
    - ✓ Hora del hecho.

- Se agregó la variable ‘categoría del hecho y tipohechovalores’, para diferenciar entre delitos contra la vida, índole sexual e índole patrimonial y sus distintos tipos, como se muestra en la tabla V.

Tabla V. **Transformación de la variable tipo de hecho**

Tipo de hecho (según fuente de información)	**Categoría del hecho**	**Valor del tipo de hecho (tipohechovalores)**
Homicidio	1	1
Lesionado	1	2
Delito sexual	2	1
Hurto a comercio	3	1
Hurto a iglesia católica	3	2
Hurto a iglesia evangélica	3	3
Hurto a residencia	3	4
Hurto de arma de fuego	3	5
Hurto de motocicleta	3	6
Hurto de vehículo	3	7
Robo a banco	3	8
Robo a comercio	3	9
Robo a iglesia católica	3	10
Robo a iglesia evangélica	3	11
Robo a peatón	3	12
Robo a residencia	3	13
Robo a turistas	3	14
Robo de arma de fuego	3	15
Robo de motocicleta	3	16
Robo de vehículo	3	17

Fuente: elaboración propia.

- Variable comisaría: se agregó la variable *bit* (patrullaje) y se extrajo del nombre de la comisaría el número de la comisaría que cubrió y verificó el hecho delictivo a través de las unidades policiales.

- Variable sexo: se transformó la variable descriptiva de sexo en: 0 que corresponde a femenino y 1 que corresponde a masculino.

- Conversión de las coordenadas georreferénciales en distancia del hecho: por motivos de estandarización de las medidas, se transformaron las coordenadas geográficas siguiendo la fórmula de Haversine (The Math Forum at NCTM, 2020):

  - ✓ Se agregó una dirección georreferencial constante compuesta por la latitud y longitud que corresponde a la ubicación del parque central de la ciudad de Guatemala, tomada como referencia.

  - ✓ Se agregó la variable ‘distanciavalores’ que contiene el valor en metros obtenido de la distancia entre dos puntos georreferénciales, la cual corresponde a la distancia del hecho a la distancia tomada como referencia.

En la tabla VI se muestra el ejemplo de la transformación descrita anteriormente:

Tabla VI. **Transformación de las variables longitud y latitud del hecho**

Tipo del Hecho	Longitud del hecho	Latitud del hecho	Longitud de referencia	Latitud de referencia	Distancia del hecho (mts) (distanciavalores)
1	-90.533458	14.6026354	14.641907	-90.513898	4852.87022
1	-90.540865	14.6031873	14.641907	-90.513898	5197.63955
1	-90.4903113	14.6504391	14.641907	-90.513898	2712.07778
1	-90.430954	14.6641899	14.641907	-90.513898	9270.92472
1	-90.4505147	14.6637678	14.641907	-90.513898	7247.07538

Fuente: elaboración propia.

- Segmentación de los datos: del total de los hechos delictivos proporcionados, se tomaron únicamente los hechos delictivos de los años 2017, 2018 y 2019, correspondientes al departamento de Guatemala para los delitos contra la vida, de índole sexual y de índole patrimonial y se obtuvieron cuatro diferentes conjuntos:

    - Conjunto de hechos delictivos de delitos contra la vida.
    - Conjunto de hechos delictivos de delitos de índole sexual.
    - Conjunto de hechos delictivos de delitos de índole patrimonial.
    - Conjunto de hechos delictivos de delitos integrado.

Del total de las variables, se seleccionaron únicamente las variables numéricas, para una mejor precisión de los algoritmos propuestos. Como se muestra la tabla VII.

Tabla VII. **Extracto de datos segmentados de homicidios en el departamento de Guatemala del año 2019**

VictimaId	categoria	longitud	latitud	distanciavalores	anio	mes	dia	diasemana	hora	bit	CODI_MUNI	zona	sexo	edad	tipohechovalc
790	1	-90.53345795	14.60263544	4852.870224	2019	4	1	2	20	13	101	13	0	40	1
791	1	-90.54086503	14.60318725	5197.639553	2019	4	1	2	20	14	101	12	1	23	1
792	1	-90.49031126	14.65043909	2712.077784	2019	4	1	2	14	12	101	6	1	34	2
794	1	-90.43095398	14.66418988	9270.924724	2019	4	1	2	14	12	101	18	1	14	1
795	1	-90.45051468	14.66376782	7247.075375	2019	4	1	2	4	12	101	18	1	27	2
800	1	-90.51768638	14.62270351	2176.30443	2019	4	1	2	1	11	101	4	0	20	2
804	1	-90.35695124	14.6690491	17170.62505	2019	4	1	2	17	12	105	0	1	25	2
805	1	-90.52231047	14.61550099	3075.973971	2019	4	2	3	21	11	101	9	1	51	1
806	1	-90.52231047	14.61550099	3075.973971	2019	4	2	3	21	11	101	9	1	27	1

Fuente: elaboración propia.

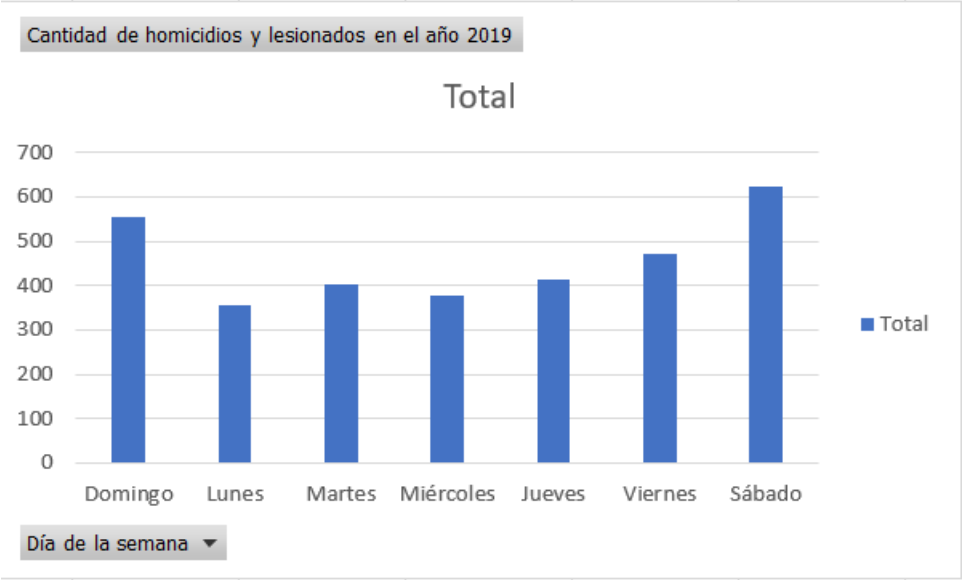
5.1.3. Modelos de clasificación y predicción del delito

Para determinar el conjunto de variables que se deben incluir en cada uno de los modelos de clasificación y predicción se realizaron los siguientes pasos:

- Análisis de frecuencia de variables: se realizó un análisis de las siguientes variables: tipo de hecho, zona donde ocurrió el hecho mes en que ocurrió el hecho, día de la semana del hecho, hora del hecho, edad de la víctima, sexo de la víctima, con la finalidad de determinar el comportamiento de esas variables en los hechos delictivos e identificar patrones de incidencia criminal.

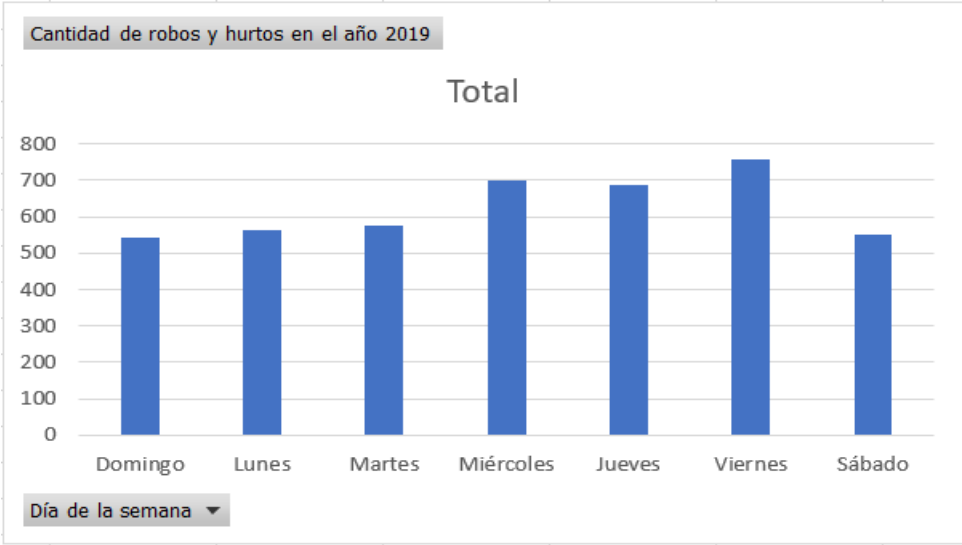
Por ejemplo, en las figuras 10 y 11 se muestra el análisis de frecuencias para la variable día de la semana en los delitos de homicidios, lesionados, robos y hurtos.

Figura 10. **Distribución de los delitos contra la vida del año 2019, en los días de la semana**



Fuente: elaboración propia utilizando Microsoft Excel.

Figura 11. **Distribución de los robos y hurtos en el año 2019, en los días de la semana**



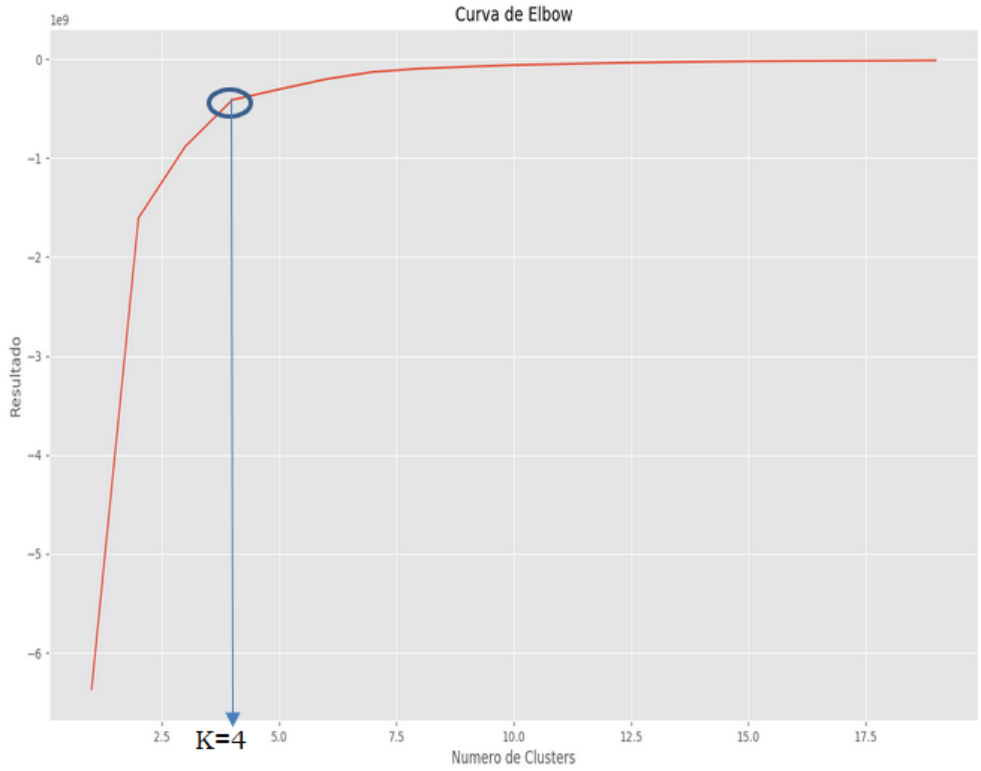
Fuente: elaboración propia utilizando Microsoft Excel.

Como se observa en las gráficas anteriores, existe mayor cantidad de homicidios, lesiones, robos y hurtos en los últimos días de la semana (jueves, viernes, sábado y domingo, que corresponden a los días 5, 6, 7 y 1).

- Establecimiento de número de agrupaciones necesarias: para determinar el número de agrupaciones óptimos en los algoritmos de agrupamiento se utilizó el método del codo (Elbow) para Kmeans y clúster difuso, y la gráfica de Dendograma, para clustering jerárquico.

En las figuras 12 y 13 se muestran los ejemplos de la obtención del número de clústeres (K).

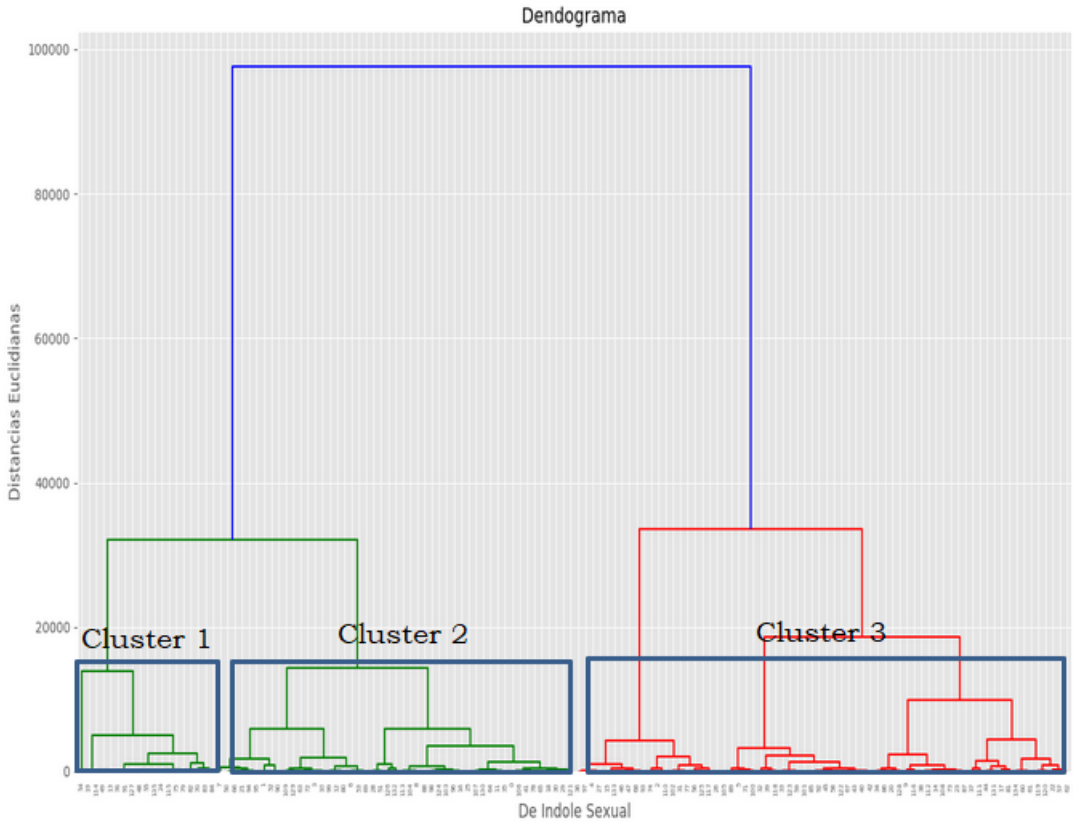
Figura 12. **Gráfica de Elbow para agrupamiento Kmeans, delitos de índole sexual**



Fuente: elaboración propia, utilizando Matplot-Lib.

En el ejemplo anterior se puede observar que el codo sucede aproximadamente en el clúster número 4, por lo que el número óptimo K es de 4 agrupamientos.

Figura 13. **Gráfica del dendograma del agrupamiento jerárquico del delito de índole sexual**



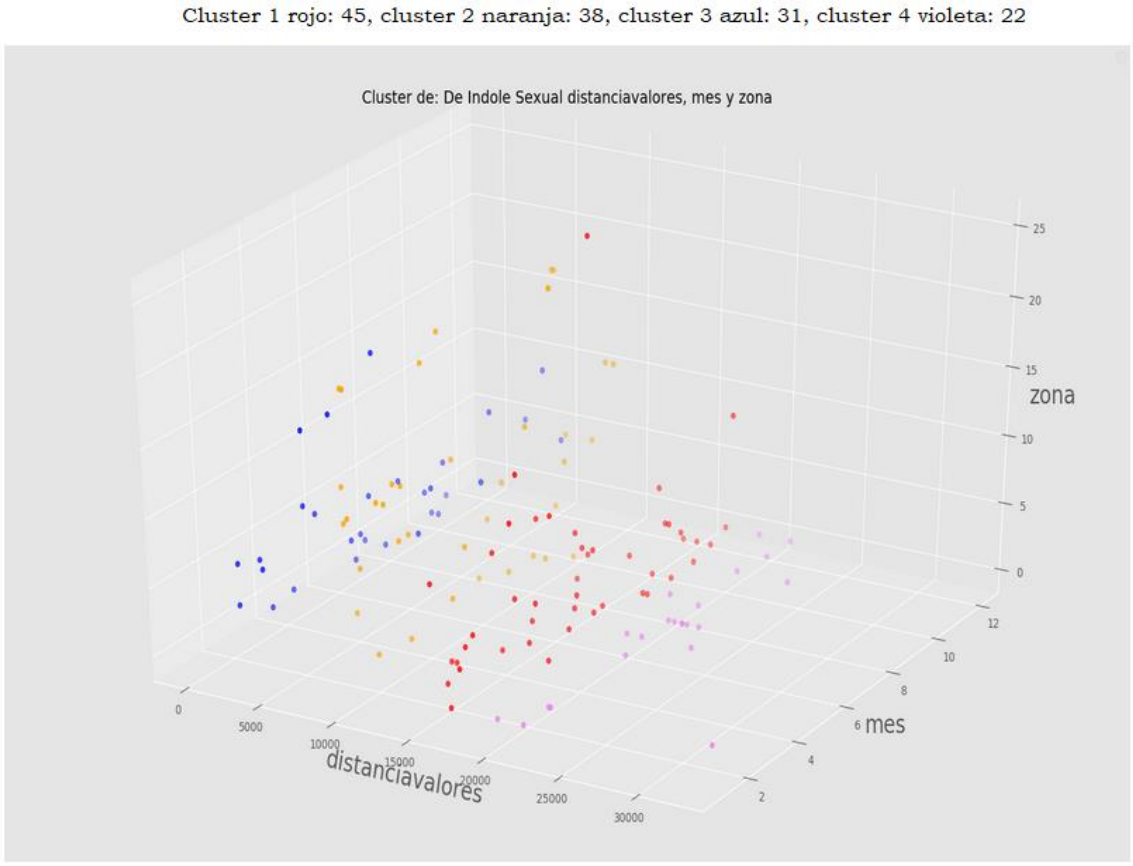
Fuente: elaboración propia, utilizando Matplot-Lib.

En la figura anterior se observa que para cubrir el tamaño total de los hechos delictivos de índole sexual se necesitan K = 3 agrupamientos.

- Obtención de clústeres diferenciados: Para obtener una buena predicción en los modelos, se necesita que los elementos en cada agrupamiento, estén lo más independientes posibles, por lo que se realizaron pruebas en

el prototipo, utilizando gráficas cartesianas de los hechos delictivos y el clúster al que corresponde, para visualizar la independencia de los elementos, así como el índice de Bouldin (Bouldin, 1979). Por ejemplo, en la Figura 15 se presenta una gráfica en tres dimensiones, donde se muestra la independencia de los hechos delictivos en las agrupaciones, para un modelo de delitos de índole sexual, que incluye las variables de mes, día de la semana, hora del hecho y zona donde ocurrió el hecho, y un índice de Bouldin = 0.2842

Figura 14. **Gráfica de clústeres diferenciados de delitos sexuales 2019 por mes, zona y día de la semana**



Fuente: elaboración propia, utilizando Matplot-Lib.

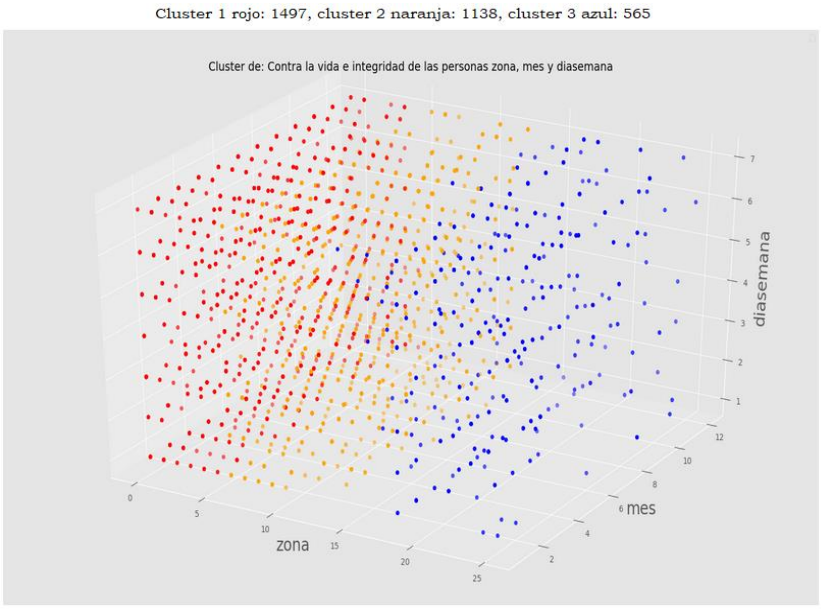
5.1.3.1. Delitos contra la vida e integridad de las personas

El modelo de delitos contra la vida e integridad de las personas, comprende aquellos hechos delictivos que terminaron en un homicidio o en una lesión grave. Después de realizar el procedimiento para la obtención de los patrones de incidencia criminal, se logró determinar que las variables que más aportan diferencia o que crean una tendencia en este delito son las siguientes:

- Mes del hecho
- Día de la semana en que ocurre el hecho
- Hora del hecho
- Ubicación del hecho (distanciavalores)
- Zona donde ocurrió el hecho

En el Figura 15 se muestra una gráfica cartesiana del agrupamiento y clasificación del delito y en el anexo No. 1 se muestran las gráficas de los agrupamientos realizados para este modelo, mediante la ejecución del prototipo para algoritmos Kmeans, Clustering jerárquico y Fuzzy C.

Figura 15. **Gráfica de agrupamiento y clasificación para delitos contra la vida 2019**



Fuente: elaboración propia utilizando Matplot-Lib.

5.1.3.2. Delitos patrimoniales

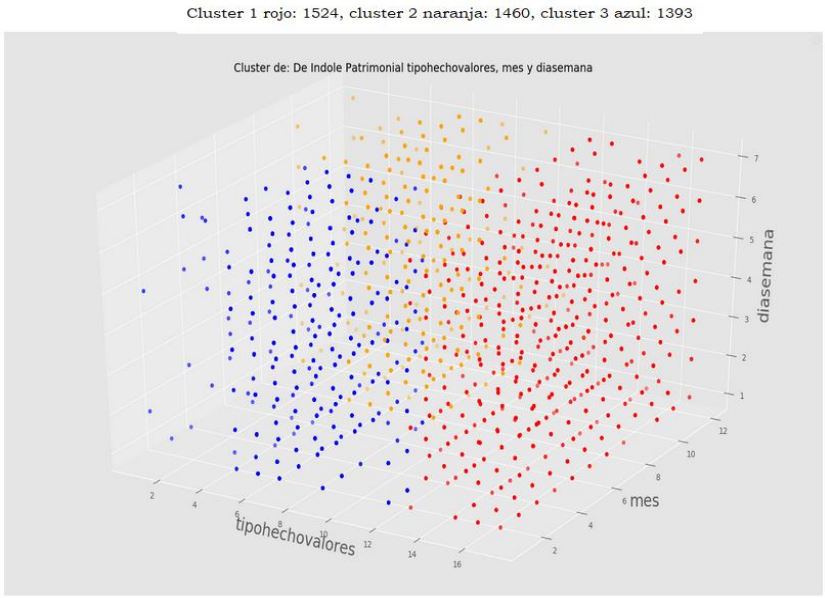
El modelo de delitos índole patrimonial, comprende aquellos hechos delictivos de robos y hurtos. Después de realizar el procedimiento para la obtención de los patrones de incidencia criminal, se logró determinar que las variables que más aportan diferencia o que crean una tendencia en este delito son las siguientes:

- Tipo de hecho
- Mes
- Día de la semana

En el Figura 16 se muestra una gráfica cartesiana del agrupamiento y clasificación del delito y en el anexo No. 2 se muestran las gráficas de los

agrupamientos realizados para este modelo, mediante la ejecución del prototipo para algoritmos Kmeans, Clustering jerárquico y Fuzzy C.

Figura 16. **Gráfica de agrupamiento y clasificación para delitos patrimoniales 2019**



Fuente: elaboración propia utilizando Matplot-Lib.

5.1.3.3. Delitos sexuales

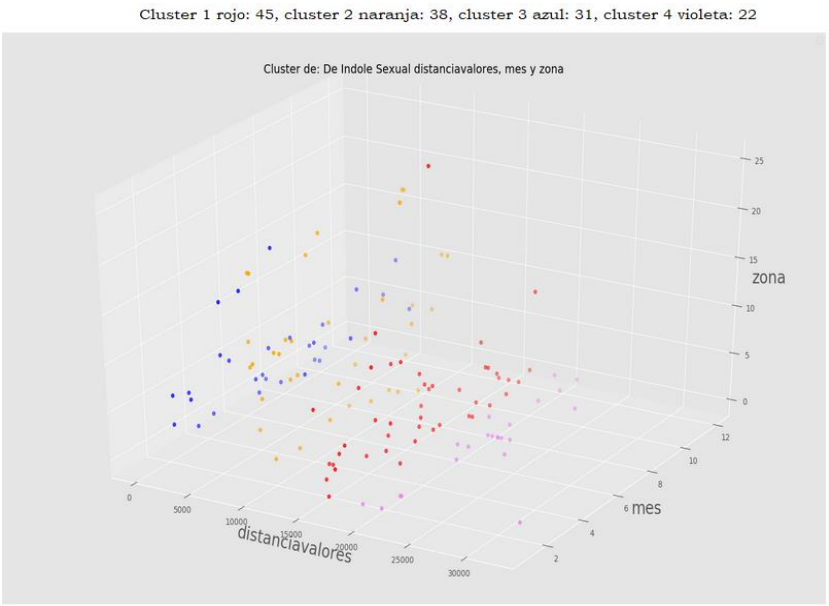
El modelo de delitos sexuales incluye a mujeres y menores de edad en su mayoría, comprende aquellos hechos delictivos que terminaron en un homicidio o en una lesión grave. Después de realizar el procedimiento para la obtención de los patrones de incidencia criminal, se logró determinar que las variables que más aportan diferencia o que crean una tendencia en este delito son las siguientes:

- Ubicación del hecho
- Mes dónde ocurrió el hecho
- Día de la semana

- Zona del hecho

En el Figura 17 se muestra una gráfica cartesiana del agrupamiento y clasificación del delito y en los apéndices se muestran las gráficas de los agrupamientos realizados para este modelo, mediante la ejecución del prototipo para algoritmos Kmeans, Clustering jerárquico y Fuzzy C.

Figura 17. **Gráfica de agrupamiento y clasificación para delitos sexuales 2019**



Fuente: elaboración propia utilizando Matplot-Lib.

5.2. Componentes funcionales del prototipo

El prototipo desarrollado cuenta con tres funcionalidades principales: Entrenamiento, predicción y visualización georreferencial, las cuales permitirán al analista, poder validar los modelos y sus variables, verificar la precisión de los algoritmos, realizar predicciones del delito y visualizar los resultados en un mapa de incidencia criminal.

Para realizar el entrenamiento y afinación de los modelos, el prototipo cuenta con la interfaz que permite realizar tres funciones principales:

- Entrenar: realiza el entrenamiento del modelo con el ochenta por ciento de los datos del conjunto de entrada y presenta los resultados.

- Probar: realiza el entrenamiento del modelo con el veinte por ciento de los datos del conjunto de entrada y presenta los resultados.

- Predecir: realiza la predicción del prototipo, mediante una red de Bayes, utilizando el algoritmo Gaussian Naive Bayes con los modelos de agrupamiento y clasificación del delito realizado con los algoritmos Kmeans, Fuzzy C y Clustering jerárquico.

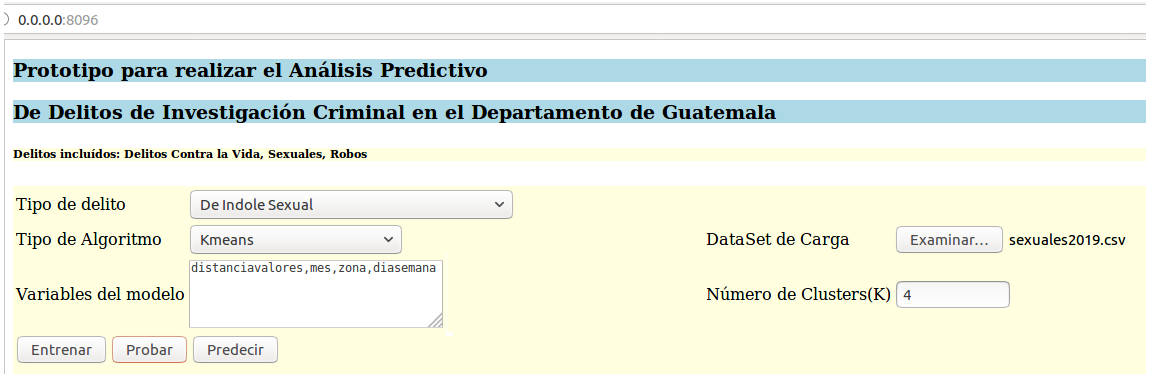
- Los insumos necesarios para realizar estas operaciones son los siguiente:
- Tipo de algoritmo: indica el tipo de algoritmo a utilizar (Kmeans, Jerárquico o Fuzzy C, red de Bayes)

- Dataset de carga: solicita el archivo que contiene el conjunto de datos en formato de texto separado por comas (CSV).

- Variables del modelo: indica el listado de variables a utilizar en el modelo, separadas por comas.

A continuación, en la figura 18 se presentan las funciones principales del prototipo.

Figura 18. **Página principal con las funcionalidades principales del prototipo**



Fuente: elaboración propia, utilizando Flask.

Después de ejecutar las operaciones seleccionadas el prototipo presenta los resultados de la manera siguiente:

- Gráfica de agrupamiento.
- Gráfica de indicadores de precisión del valor del número de agrupaciones: Codo para Kmeans y Fuzzy C, Dendograma para Clustering jerárquico.
- Visualización de la incidencia criminal, utilizando los agrupamientos.
- Índice de Bouldin que mide el nivel de eficiencia del prototipo.
- Archivo delimitado por comas, con los resultados del agrupamiento.

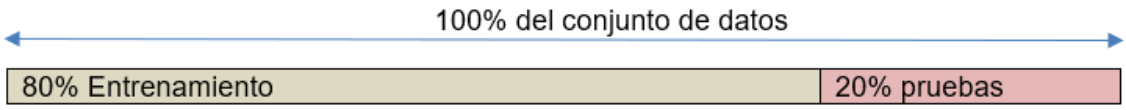
El objetivo es que el usuario final pueda probar sus modelos, con sus variables, los conjuntos de datos, agrupar y clasificar el delito, predecir y analizar los resultados, y realizar las interpretaciones criminológicas necesarias para que puedan servir de insumo, para realizar las estrategias de prevención.

A continuación, se mostrarán en detalle las funcionalidades indicadas anteriormente, realizando un experimento completo, utilizando el modelo de delitos de índole sexual, definido en la sección anterior.

5.2.1. Entrenamiento, afinación y precisión de los modelos

Como parte de las buenas prácticas recomendadas para entrenar y probar los algoritmos de agrupamiento y predicción, se dividieron los conjuntos de datos de manera aleatoria en dos partes, la primera parte con el 80 % de los hechos delictivos de determinado delito, se utilizó para el entrenamiento y el 20 % restante para las pruebas.

Figura 19. **División del conjunto de datos para entrenamiento y pruebas**



Fuente: elaboración propia utilizando Microsoft Excel.

Utilizando como ejemplo el delito de índole sexual del año 2019 del departamento de Guatemala, donde ocurrieron un total de 136 delitos y después de dividir el conjunto de datos de manera aleatoria, se obtuvieron los dos conjuntos de datos de la manera siguiente:

- Subconjunto de entrenamiento de 108 delitos
- Subconjunto de pruebas de 28 delitos

Como referencia se muestra un extracto de los conjuntos de datos en las tablas VIII y IX.

Tabla VIII. **Extracto del conjunto de datos para entrenamiento**

Victimald	categoria	longitud	latitud	distanciavalores	anio	mes	dia	diasemana	hora	bit	municipio	zona	sexo	edad	tipohecho
15308	2	-90.56227818	14.56508156	10014.98353	2019	11	21	5	9	14	115	12	0	32	1
8682	2	-90.6112575	14.65086226	10533.04659	2019	9	10	3	3	16	108	7	0	13	1
8620	2	-90.58751188	14.52655687	15092.25003	2019	7	21	1	9	15	115	4	0	18	1
13935	2	-90.49773536	14.5693875	8258.401235	2019	10	29	3	9	13	102	2	1	15	1
3345	2	-90.64087792	14.46391566	24078.54189	2019	6	1	7	1	15	114	0	0	14	1
8644	2	-90.44150772	14.73638357	13090.8395	2019	8	12	2	21	12	107	0	0	18	1
6498	2	-90.51970883	14.61358575	3214.21549	2019	3	12	3	6	11	101	9	0	55	1
16639	2	-90.52381821	14.60756989	3968.909162	2019	12	5	5	17	11	101	9	0	14	1
8616	2	-90.65072186	14.50819553	20948.65497	2019	7	20	7	11	15	114	0	0	5	1
6508	2	-90.56805029	14.57010856	9894.867576	2019	1	2	4	10	14	115	12	0	65	1
6596	2	-90.48063296	14.60929101	5101.045065	2019	2	11	2	6	13	101	16	0	153	1
6559	2	-90.61863754	14.70761882	13443.50402	2019	3	17	1	9	16	110	0	0	116	1
8604	2	-90.44103779	14.46659678	21035.14996	2019	7	14	1	10	13	113	2	0	15	1
8610	2	-90.51432406	14.6283023	1515.155008	2019	7	15	2	7	11	101	1	0	17	1
6595	2	-90.48063296	14.60929101	5101.045065	2019	3	10	1	4	13	101	16	0	152	1

Fuente: elaboración propia.

Tabla IX. **Extracto del conjunto de datos para pruebas**

Victimald	categoria	longitud	latitud	distanciavalores	anio	mes	dia	diasemana	hora	bit	municipio	zona	sexo	edad	tipohecho
13907	2	-90.55636772	14.49570003	16906.55482	2019	10	5	7	10	15	117	2	0	5	1
8624	2	-90.48063719	14.66996633	4752.659226	2019	7	25	5	9	12	101	6	0	45	1
6594	2	-90.56139228	14.55952177	10501.61821	2019	2	6	4	3	14	115	12	0	151	1
8677	2	-90.58758233	14.5273183	15024.20184	2019	9	5	5	18	15	115	1	0	15	1
6508	2	-90.56805029	14.57010856	9894.867576	2019	1	2	4	10	14	115	12	0	65	1
1894	2	-90.60128301	14.52274974	16265.716	2019	5	26	1	7	15	115	9	0	19	1
8595	2	-90.56431085	14.52001834	14615.12136	2019	7	1	2	19	15	115	5	0	22	1
6588	2	-90.5083825	14.63575922	906.2273218	2019	3	24	1	0	11	101	1	0	145	1
622	2	-90.59243017	14.65175147	8528.71584	2019	4	15	2	0	16	108	7	1	7	1
8653	2	-90.51613092	14.61929602	2528.492886	2019	8	19	2	6	11	101	4	0	14	1
8616	2	-90.65072186	14.50819553	20948.65497	2019	7	20	7	11	15	114	0	0	5	1
16661	2	-90.62664757	14.48507684	21268.70775	2019	12	29	1	20	15	114	0	0	13	1
8662	2	-90.59888779	14.54885531	13824.70951	2019	8	26	2	7	15	115	2	0	19	1
3349	2	-90.55186124	14.52478234	13664.59733	2019	6	3	2	16	15	117	7	0	13	1

Fuente: elaboración propia.

El prototipo presenta el resultado del rendimiento de los algoritmos, tomando como medida el índice de Bouldin (Bouldin, 1979), el cual mide la precisión del algoritmo, a través de la separación de los grupos, esto quiere decir que cuando el índice es cercano a cero, existe una mayor separación y por lo tanto una mejor precisión en el algoritmo.

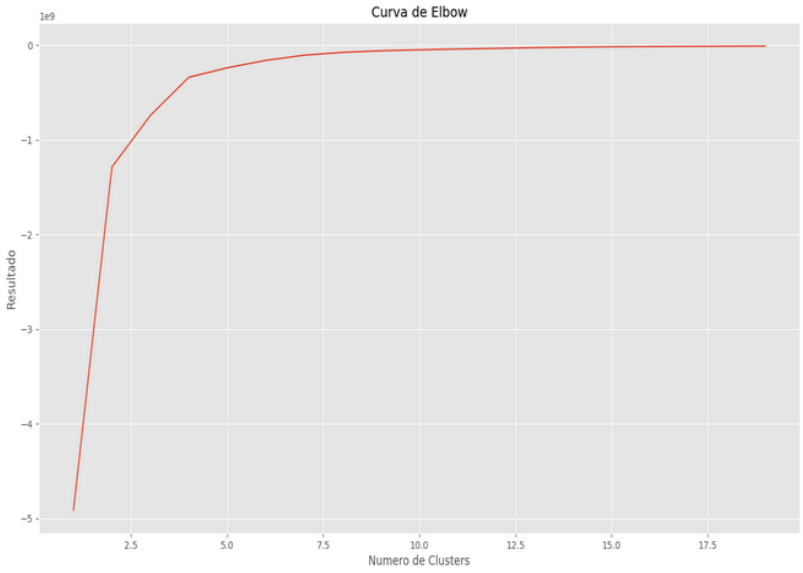
A continuación, se presentan los resultados obtenidos en el prototipo de los agrupamientos para cada uno de los algoritmos, utilizando los conjuntos de datos de delitos de índole sexual y el nivel de precisión de los algoritmos y después de realizar diferentes pruebas, con diferentes modelos y número de agrupamientos.

5.2.1.1. KMEANS

Insumos: Delitos de índole sexual cometidos en el departamento de Guatemala durante el año 2019, variables: ubicación del hecho (distancia del hecho), zona, mes y día de la semana en que ocurrió el hecho.

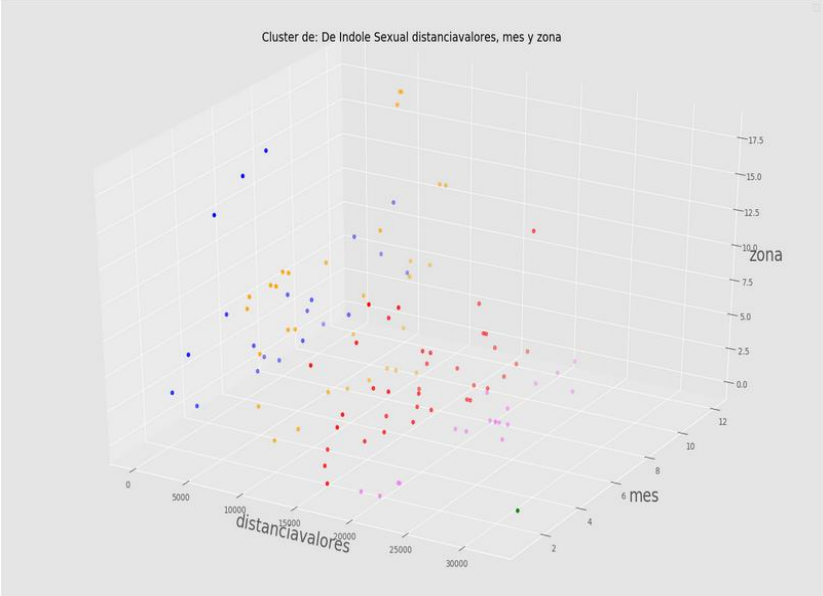
Número de conglomerados (K) = 5

Figura 20. **Gráfica de codo: entrenamiento Kmeans para delitos de índole sexual del año 2019**



Fuente: elaboración propia, utilizando Matplot-Lib.

Figura 21. **Agrupamiento Kmeans: entrenamiento para delitos de índole sexual del año 2019**



Fuente: elaboración propia, utilizando Matplot-Lib.

Precisión del algoritmo en el entrenamiento: Índice de Bouldin: 0.3505, efectividad: 0.6495 aproximadamente

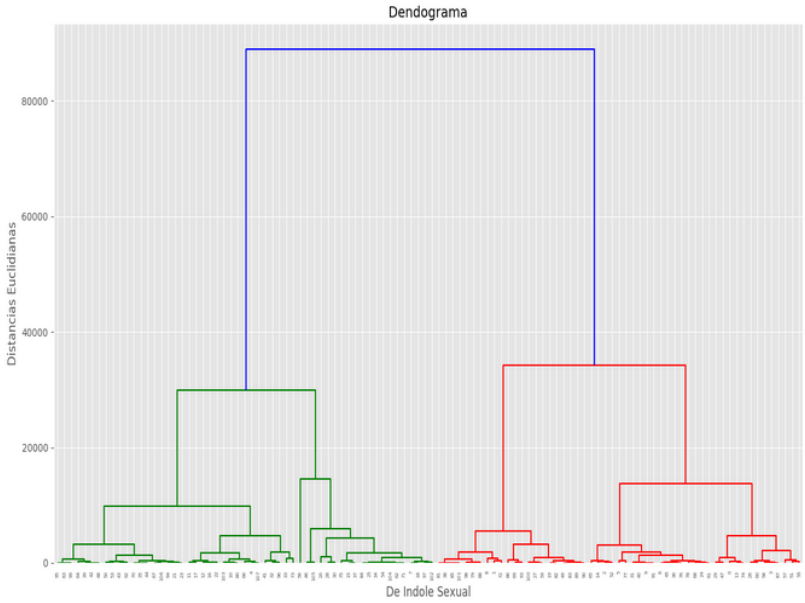
Para el ejercicio de pruebas se obtuvo una precisión del algoritmo de: 0.3519, efectividad: 0.6481 aproximadamente.

5.2.1.2. Clustering jerárquico

Insumos: Delitos de índole sexual cometidos en el departamento de Guatemala durante el año 2019, variables: ubicación del hecho (distancia del hecho), zona, mes y día de la semana en que ocurrió el hecho.

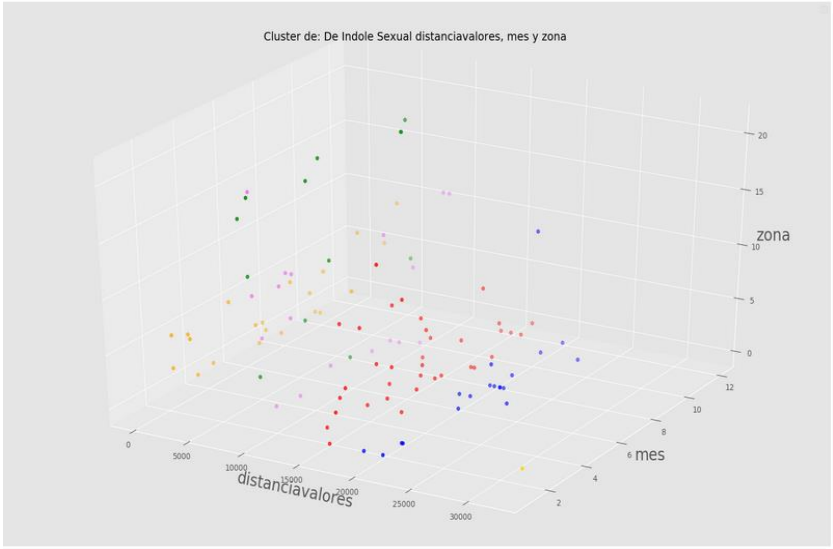
Número de conglomerados (K) = 6

Figura 22. **Dendograma: entrenamiento Clustering jerárquico para delitos de índole sexual del año 2019**



Fuente: elaboración propia, utilizando Matplot-Lib.

Figura 23. **Agrupamiento Jerárquico: entrenamiento para delitos de índole sexual del año 2019**



Fuente: elaboración propia, utilizando Matplot-Lib.

Para el entrenamiento, se obtuvo una precisión del algoritmo: Índice de Bouldin: 0.3361, efectividad: 0.6639 aproximadamente

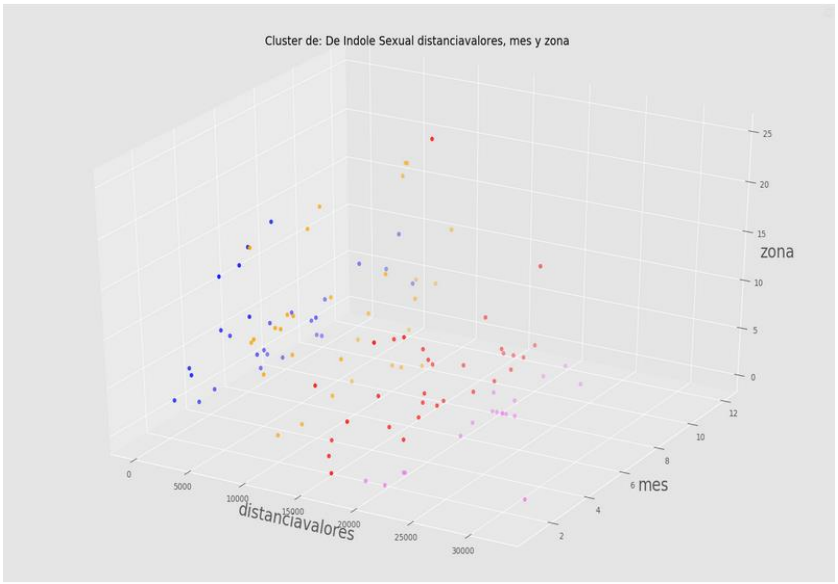
Para el ejercicio de pruebas se obtuvo una precisión del algoritmo: Índice de Bouldin: 0.2842, efectividad: 0.7158 aproximadamente

5.2.1.3. FUZZY C

Insumos: Delitos de índole sexual cometidos en el departamento de Guatemala durante el año 2019, variables: ubicación del hecho (distancia del hecho), zona, mes y día de la semana en que ocurrió el hecho.

Número de conglomerados (K) = 6, obtenido de la gráfica de codo del algoritmo Kmeans.

Figura 24. **Agrupamiento Fuzzy C: entrenamiento para delitos de índole sexual del año 2019**



Fuente: elaboración propia, utilizando Matplot-Lib.

Para el entrenamiento, se obtuvo una precisión del algoritmo: Índice de Bouldin: 0.4420, efectividad: 0.558 aproximadamente.

Para el ejercicio de pruebas, se obtuvo una precisión del algoritmo: Índice de Bouldin: 0.3777, efectividad: 0.6223 aproximadamente.

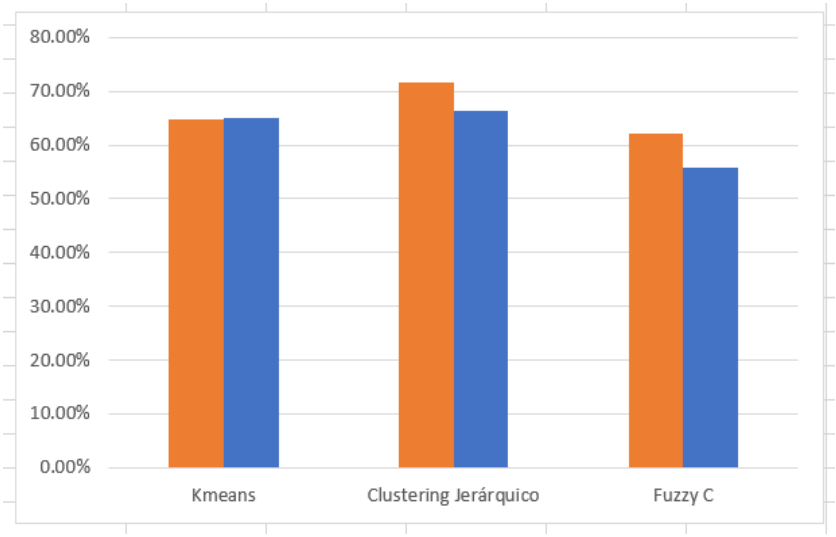
A continuación, se presenta en la tabla X y la figura 25 el porcentaje de precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, según el ejercicio realizado.

Tabla X. **Precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C para delitos de índole sexual**

Algoritmo	No. Grupos(K)	No. Delitos	Entrenamiento	Pruebas
Kmeans	5	136	64.95 %	64.81 %
Clustering jerárquico	6	136	66.39 %	71.58 %
Fuzzy C	6	136	55.80 %	62.23 %

Fuente: elaboración propia.

Figura 25. **Gráfica de barras del nivel de precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, para delitos de índole sexual**



Fuente: elaboración propia, utilizando Microsoft Excel.

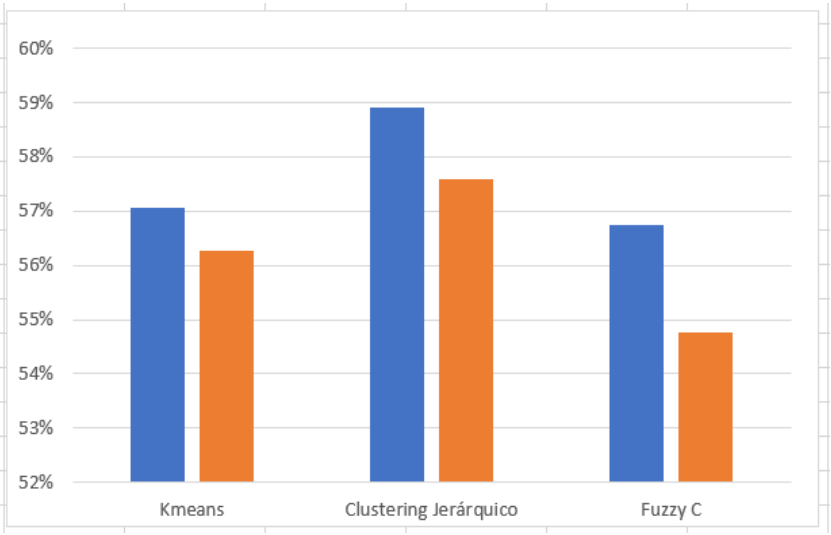
De la misma manera se realizaron los ejercicios de entrenamiento y pruebas para los delitos contra la vida y de índole patrimonial, por lo que a continuación en las Tablas XI, XII y en las Figuras 26 y 27 se muestra una comparativa de la precisión de los algoritmos de agrupamiento Kmeans, Clustering jerárquico y Fuzzy C, para estos delitos.

Tabla XI. **Precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C para delitos contra la vida**

Algoritmo	No. Grupos (K)	No. Delitos	Entrenamiento	Pruebas
Kmeans	6	3200	57 %	56 %
Clustering jerárquico	6	3200	59 %	58 %
Fuzzy C	6	3200	57 %	55 %

Fuente: elaboración propia utilizando Microsoft Excel.

Figura 26. **Gráfica de barras del nivel de precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, para delitos contra la vida**



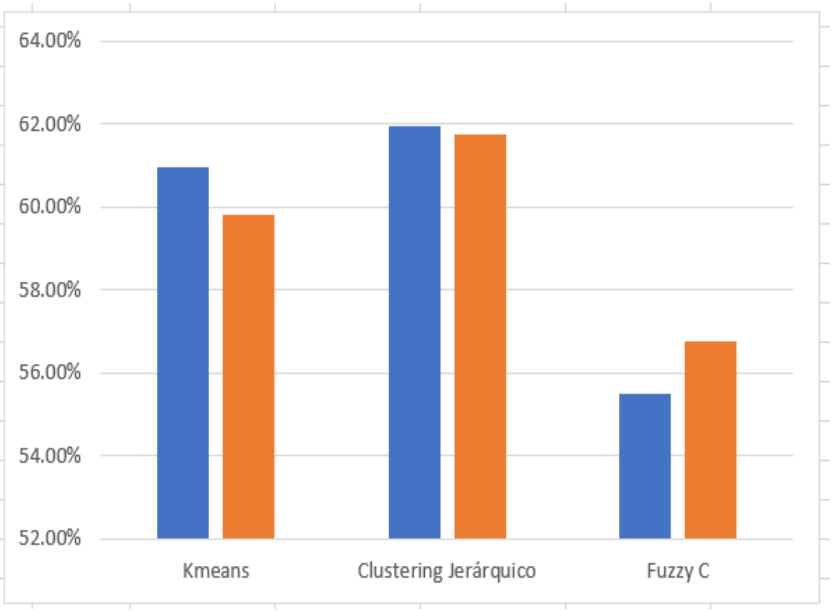
Fuente: elaboración propia, utilizando Microsoft Excel.

Tabla XII. **Precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C para delitos de índole patrimonial**

Algoritmo	No. Grupos (K)	No. de delitos	Entrenamiento	Pruebas
Kmeans	7	4377	60.95 %	59.82 %
Clustering jerárquico	6	4377	61.95 %	61.77 %
Fuzzy C	6	4377	55.50 %	56.75 %

Fuente: elaboración propia.

Figura 27. **Gráfica de barras del nivel de precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, para delitos de índole patrimonial**



Fuente: elaboración propia, utilizando Microsoft Excel.

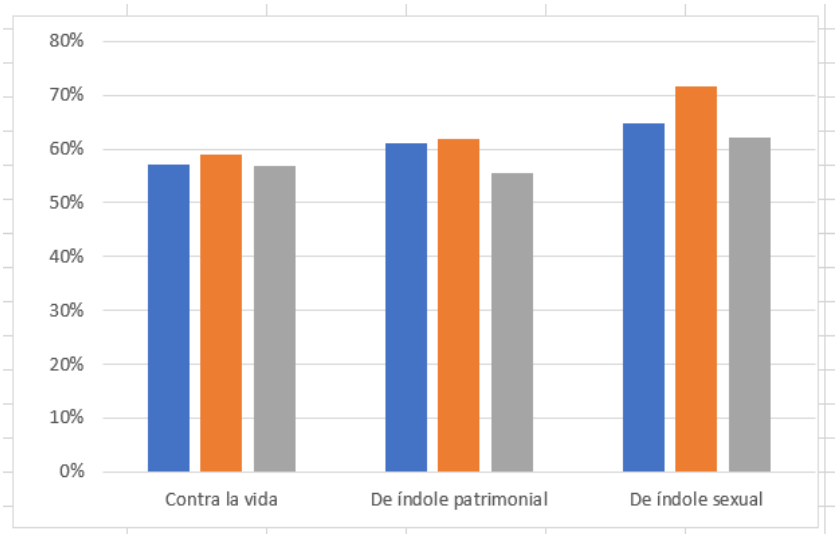
A manera de resumen, a continuación, se presenta la Tabla XIII y la Figura 28, que muestran la precisión para los tres delitos y los tres algoritmos utilizados.

Tabla XIII. **Precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C para todos los delitos**

Delitos	Algoritmos			
	No. Delitos	Kmeans	Clustering jerárquico	Fuzzy C
Contra la vida	3200	57 %	59 %	57 %
De índole patrimonial	4377	61 %	62 %	56 %
De índole sexual	136	65 %	72 %	62 %

Fuente: elaboración propia.

Figura 28. **Gráfica de barras del nivel de precisión de los algoritmos Kmeans, Clustering jerárquico y Fuzzy C para todos los delitos**



Fuente: elaboración propia utilizando Microsoft Excel.

5.2.2. Generación de insumo para predicción

El prototipo, genera un archivo separado por comas, con el conjunto de datos de entrada y adicionalmente la columna “Clúster”, que indica la categoría

del agrupamiento. Este archivo es el que se utilizará para realizar la predicción, pero también podrá ser útil para realizar el análisis nominal de los registros.

A continuación, en la Tabla XIV se muestra un extracto del resultado del agrupamiento, visto de manera nominal.

Tabla XIV. **Agrupamiento y clasificación para delitos de índole sexual**

Victima	categoria	longitud	latitud	distancia	anio	mes	dia	diasemar	hora	bit	CODI_MUNI	zona	sexo	edad	tipohec	Cluster
8705	2	-90.45278715	14.77876515	16595.05162	2019	9	29	1	2	12	107	0	0	24	1	0
8595	2	-90.56431085	14.52001834	14615.12136	2019	7	1	2	19	15	115	5	0	22	1	0
6594	2	-90.56139228	14.55952177	10501.61821	2019	2	6	4	3	14	115	12	0	151	1	5
6505	2	-90.5377281	14.55693539	9801.097384	2019	1	1	3	6	14	101	21	0	62	1	5
600	2	-90.45278715	14.77876515	16595.05162	2019	4	1	2	10	12	107	1	0	13	1	0
15305	2	-90.59584506	14.76626582	16416.33241	2019	11	17	1	23	16	111	0	0	17	1	0
8672	2	-90.59576849	14.50674145	17441.24247	2019	9	3	3	11	15	115	4	0	12	1	0
8620	2	-90.58751188	14.52655687	15092.25003	2019	7	21	1	9	15	115	4	0	18	1	0
6504	2	-90.56068696	14.57155015	9313.647371	2019	3	19	3	15	14	115	12	0	61	1	5
1890	2	-90.47560547	14.65677053	4443.640396	2019	5	22	4	23	12	101	18	0	14	1	1
13933	2	-90.61620601	14.48373055	20773.70803	2019	10	28	2	20	15	114	0	0	14	1	3
8681	2	-90.5054446	14.54199779	11158.99945	2019	9	9	2	0	15	116	0	0	24	1	5
13920	2	-90.53071851	14.61370326	3624.854141	2019	10	19	7	0	11	101	8	0	15	1	2
13935	2	-90.49773536	14.5693875	8258.401235	2019	10	29	3	9	13	102	2	1	15	1	1
15310	2	-90.5653479	14.56106393	10569.11679	2019	11	23	7	11	14	115	12	0	37	1	5
6555	2	-90.37827239	14.78219248	21380.15884	2019	2	24	1	10	12	104	0	0	112	1	3
8677	2	-90.58758233	14.5273183	15024.20184	2019	9	5	5	18	15	115	1	0	15	1	0
15295	2	-90.5975385	14.65390373	9106.599876	2019	11	10	1	1	16	108	6	0	17	1	5
16649	2	-90.49663293	14.62380066	2742.374564	2019	12	19	5	22	13	101	5	0	25	1	2
3345	2	-90.64087792	14.46391566	24078.54189	2019	6	1	7	1	15	114	0	0	14	1	3
13907	2	-90.55636772	14.49570003	16906.55482	2019	10	5	7	10	15	117	2	0	5	1	0

Fuente: elaboración propia.

5.2.3. Predicción del delito

De acuerdo a las agrupaciones obtenidas en el valor K, estas representarán las diferentes categorías de incidencia criminal o puntos calientes, ordenados según el número de delitos que han sido agrupados en cada categoría de manera descendente y representados por un color. Por ejemplo, para el modelo de delitos de índole sexual entrenado anteriormente y tomando el agrupamiento jerárquico, dónde se tienen seis agrupaciones (K=6), a continuación, en la Tabla XV se muestran categorías de incidencia criminal obtenida:

Tabla XV. **Resumen del agrupamiento y clasificación de delitos de índole sexual**

Nivel incidencia	Color	Cantidad de delitos
1	Rojo	35
2	Naranja	20
3	Azul	18
4	Violeta	17
5	Verde	17
6	Oro	1
	Total	108

Fuente: elaboración propia.

Estas categorías deben ser analizadas desde el punto de vista criminológico, para analizar los diferentes patrones de delincuencia que pertenecen a determinada zona, la cual se podrá analizar en el mapa georreferencial.

Para realizar la predicción, el prototipo utiliza el algoritmo Gaussian Naive Bayes, que es una red de Bayes, solicitando al usuario como insumo el archivo de agrupamiento realizado por los algoritmos Kmeans, Clustering jerárquico o Fuzzy C, mostrado en el apartado anterior.

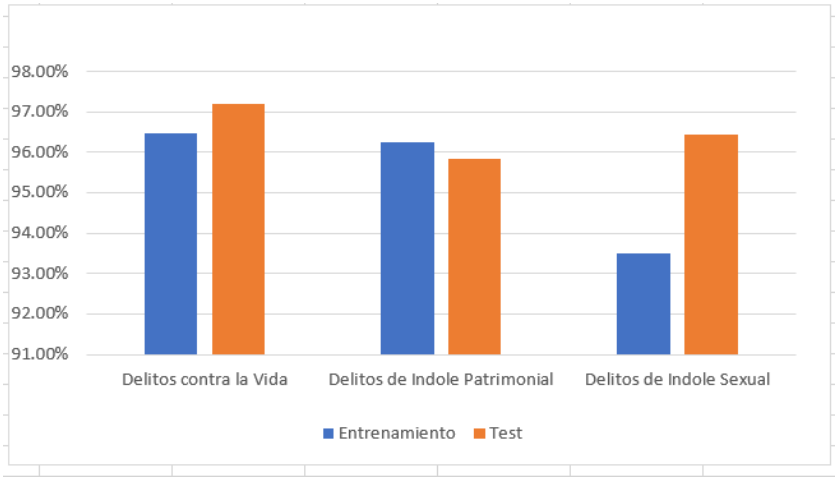
Se realizaron los ejercicios de entrenamiento y pruebas con los conjuntos de datos de agrupamiento, tomando como base el agrupamiento Clustering jerárquico, y se obtuvieron los siguientes resultados de precisión del algoritmo Naive Bayes, ver Tabla XVI y Figura 29.

Tabla XVI. **Precisión del algoritmo Gaussian Naive Bayes para la predicción del delito**

	No. De datos	Entrenamiento	Test
Delitos contra la Vida	3200	96.48 %	97.18 %
Delitos de índole Patrimonial	4377	96.25 %	95.85 %
Delitos de índole Sexual	136	93.51 %	96.42 %

Fuente: elaboración propia.

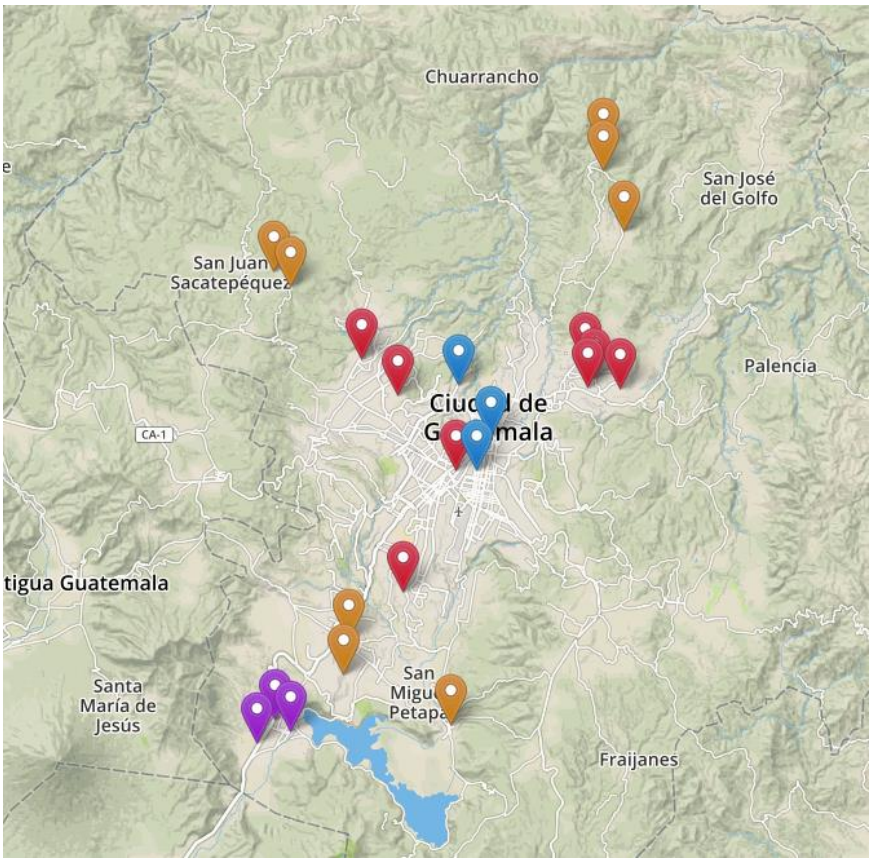
Figura 29. **Gráfica de barras del nivel de precisión del algoritmo Gaussian Naive Bayes para la predicción del delito**



Fuente: elaboración propia, utilizando Microsoft Excel.

Con el algoritmo entrenado se realizaron predicciones de delitos que no han ocurrido, utilizando las variables de: Mes, día del hecho, hora del hecho y ubicación y se logró una precisión del 95 %, como se muestra en la distribución geográfica visualizada de la Figura 30 que se presenta a continuación.

Figura 30. **Distribución de los delitos de la predicción realizada para delitos de índole sexual**



Fuente: elaboración propia, utilizando Leatflet.

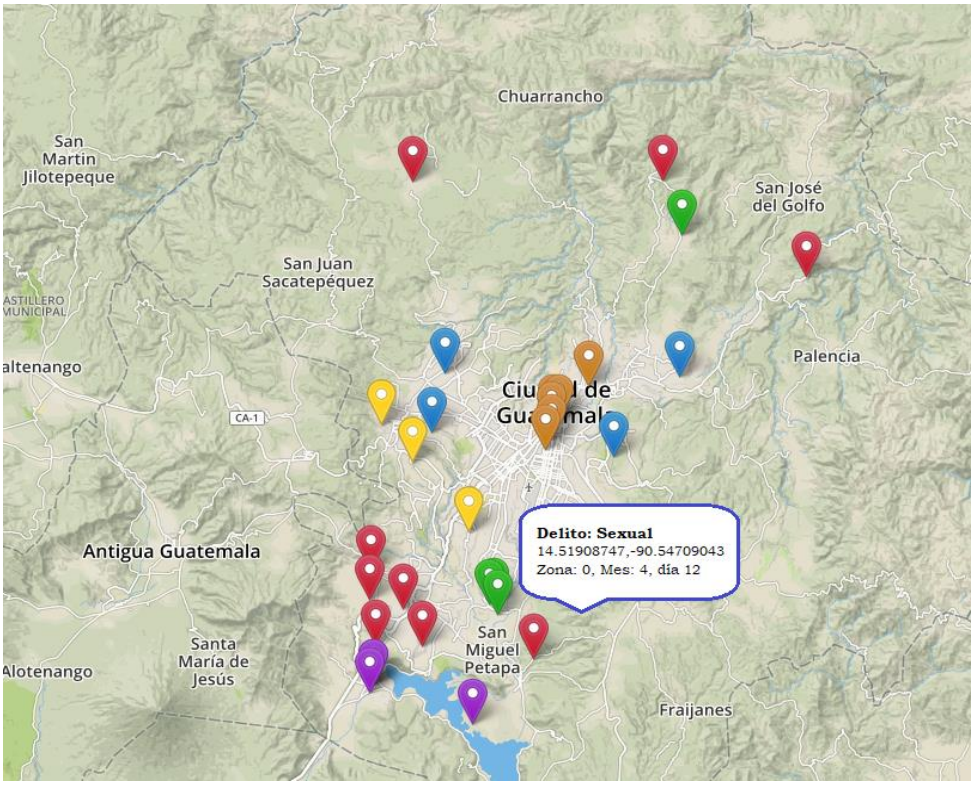
5.2.4. Visualización georreferencial de resultados

El prototipo permite que el usuario realice un análisis georreferencial de los delitos agrupados y predichos, en las distintas categorías o niveles de incidencia criminal. Esto es útil para el criminólogo y experto, para que revise los patrones de delincuencia que ocurren en las distintas zonas georreferénciales, o los grupos de delincuencia organizados que operan y así poder proveer insumos para las estrategias y equipos, encargados de la predicción del delito.

El prototipo presenta un conjunto de marcas georreferénciales agrupadas según el color de la incidencia criminal definido anteriormente, con la información del delito.

A continuación, en la Figura 31 se muestra un mapa georreferencial, con la agrupación realizada en el apartado anterior y según el conjunto de datos ingresado, en el que se puede ver la distinta distribución de los colores, según la categoría de incidencia criminal.

Figura 31. **Visualización georreferencial de las categorías predichas para delitos de índole sexual**

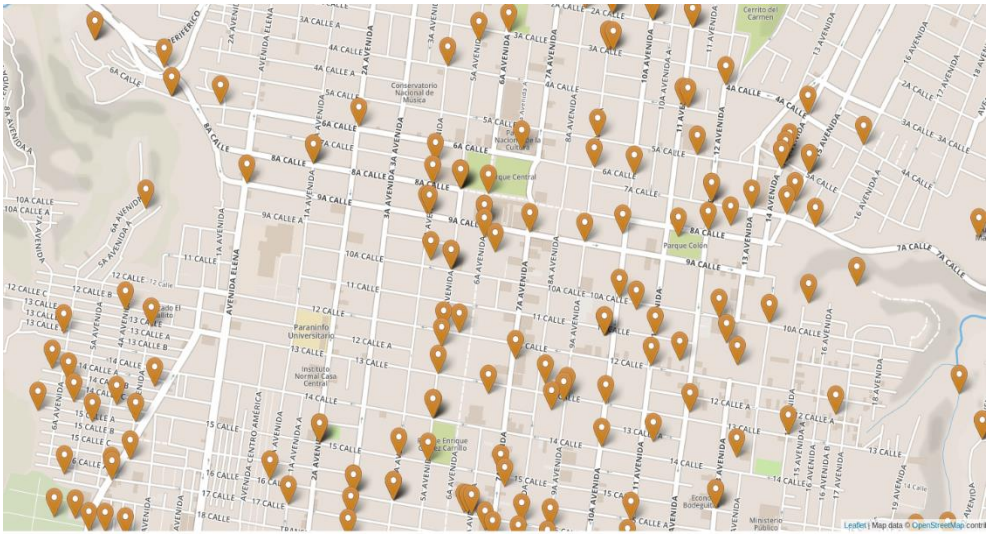


Fuente: elaboración propia, utilizando Leatflet.

5.2.5. Patrones del delito identificados

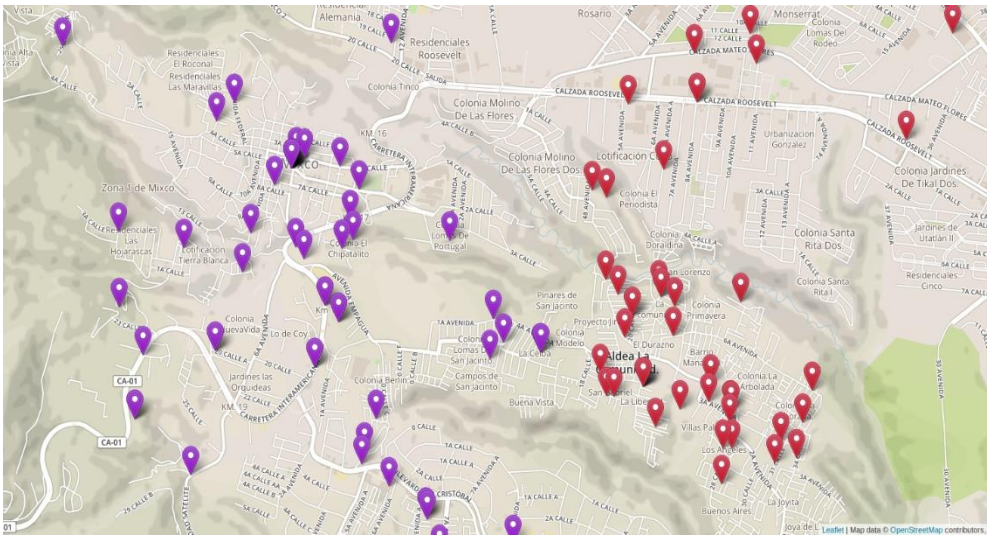
La clasificación del delito visualizada en un mapa de incidencia criminal, permite ver que los delitos de homicidios y hechos violentos contra la integridad y la vida de las personas, tienen la particularidad que se concentran en determinadas áreas geográficas del área metropolitana y de los municipios de Mixco y villa nueva, como se aprecia en las Figuras 32, 33 y 34:

Figura 32. **Distribución geográfica de los delitos contra la vida del año 2019 en la zona 1 del área metropolitana**



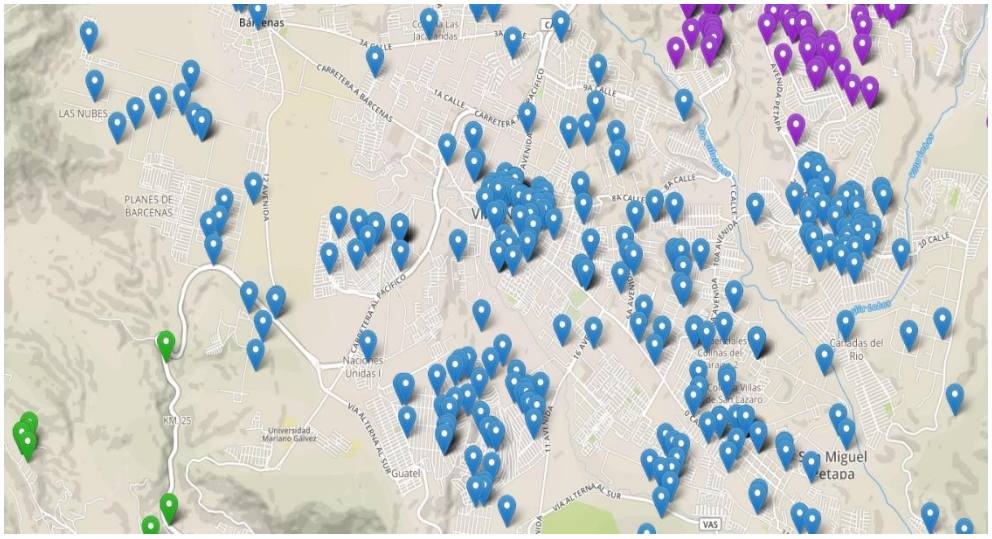
Fuente: elaboración propia, utilizando Leatflet.

Figura 33. **Distribución de los delitos contra la vida del año 2019 en el municipio de Mixco y sus alrededores**



Fuente: elaboración propia utilizando Leatflet.

Figura 34. **Distribución de los delitos contra la vida del año 2019 en el municipio de Villa nueva y sus alrededores**

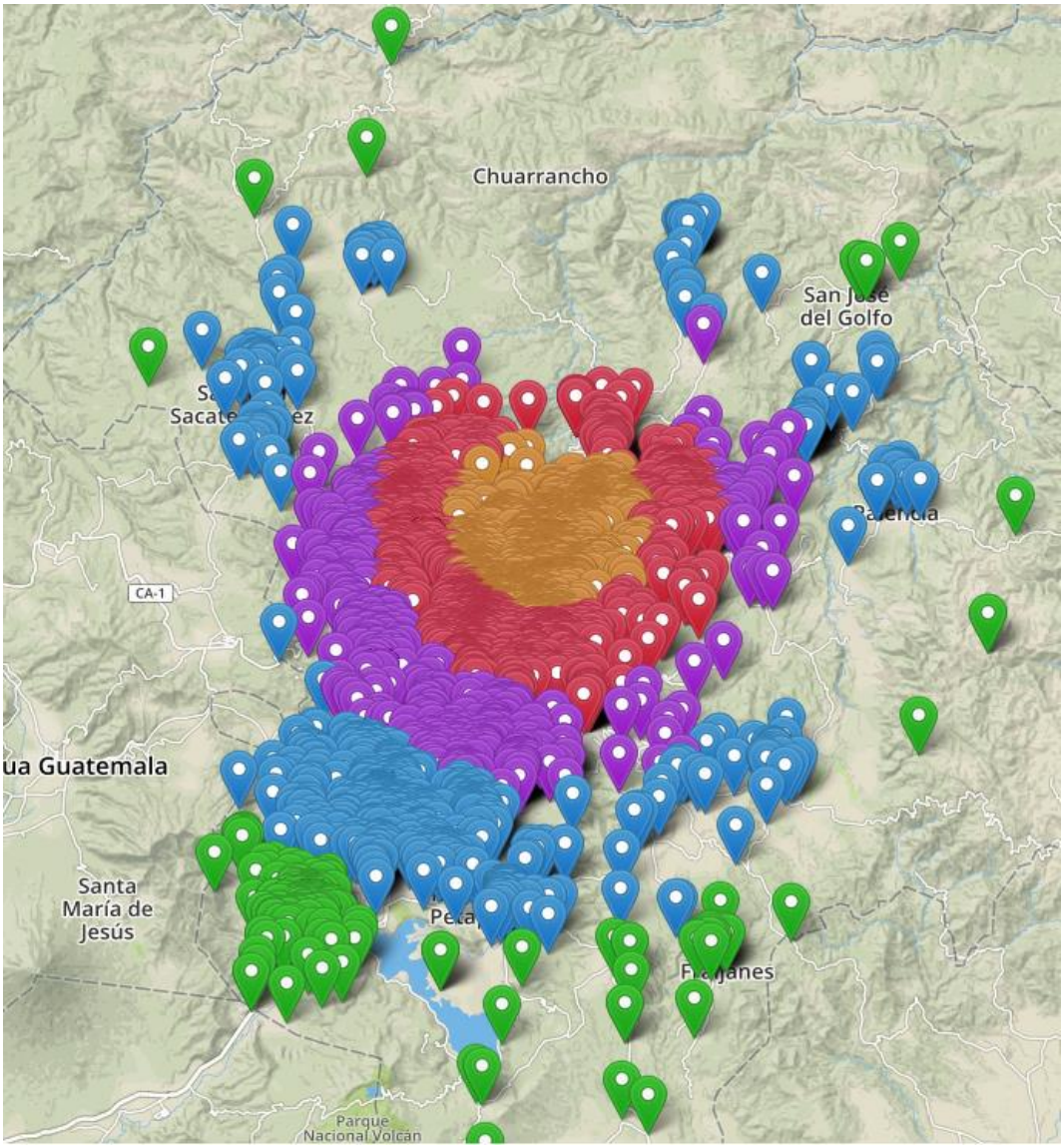


Fuente: elaboración propia, utilizando Leatflet.

De la misma manera los delitos de índole patrimonial, siguen una caracterización similar, donde la mayoría de los robos y hurtos ocurren en el

centro del área metropolitana, según el agrupamiento jerárquico realizado, como se muestra en la Figura 35:

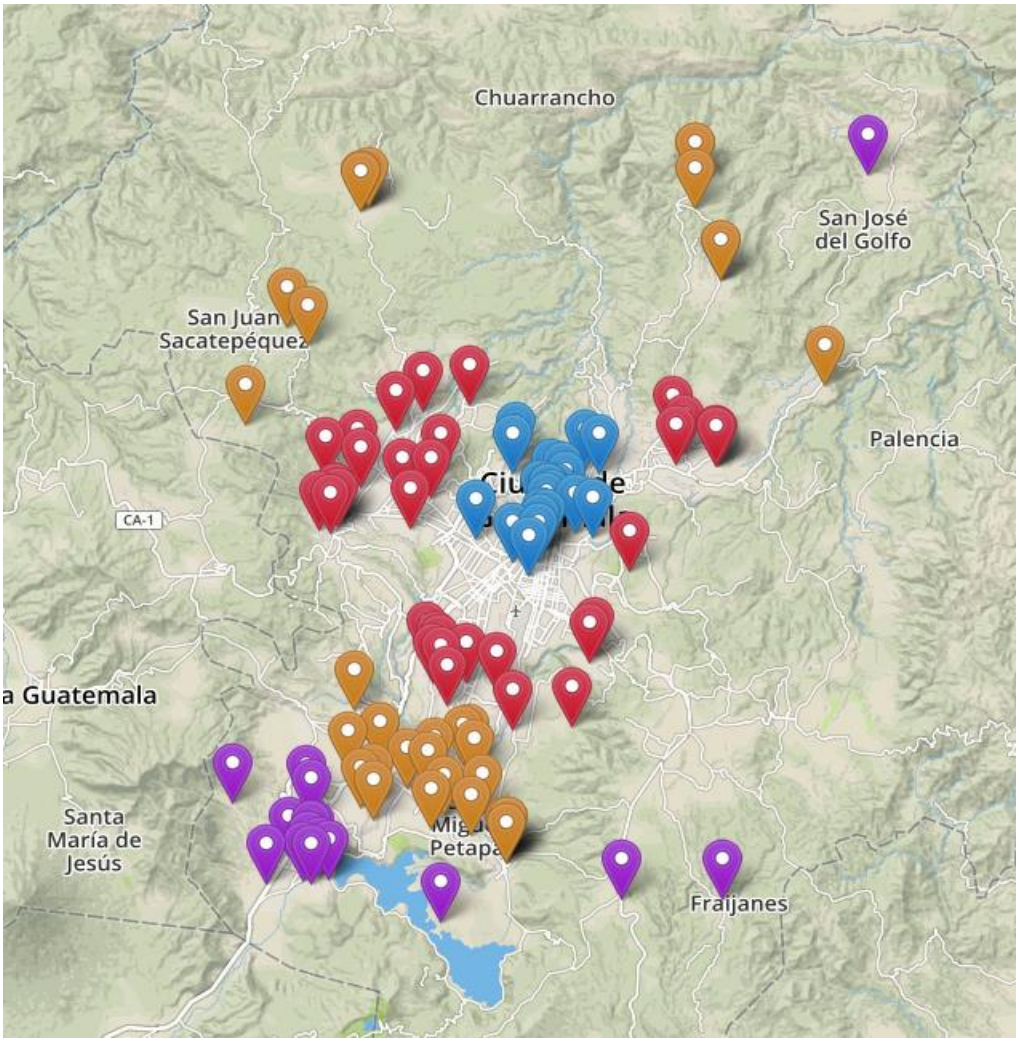
Figura 35. **Distribución de los delitos de índole patrimonial 2019 en el departamento de Guatemala**



Fuente: elaboración propia, utilizando Leatflet.

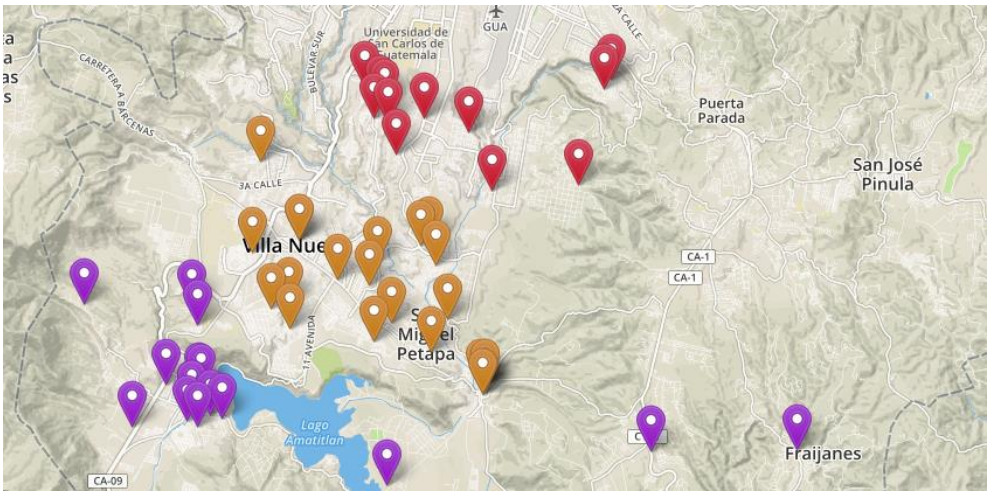
Por el contrario, en el caso de los delitos de índole sexual, estos son relativamente pocos (136 delitos) en comparación con los homicidios y lesionados (3200 delitos) y los robos y hurtos (4377 delitos), pero que se caracterizan y agrupan en los alrededores del departamento de Guatemala, municipios de Villa Nueva, San Miguel Petapa y Mixco, según el agrupamiento jerárquico realizado, como se muestra en las Figuras 36 y 37:

Figura 36. **Distribución de los delitos de índole sexual del año 2019 en el departamento de Guatemala**



Fuente: elaboración propia, utilizando Leatflet.

Figura 37. **Distribución de los delitos de índole sexual del año 2019 en los municipios de Villa Nueva, San Miguel Petapa y Amatitlán**



Fuente: elaboración propia, utilizando Leatflet.

Como se puede observar se comenten muchos más delitos en los municipios de Villa Nueva, Mixco, San Miguel Petapa y Amatitlán, que en el área metropolitana.

6. DISCUSIÓN DE RESULTADOS

6.1. La importancia de la preparación de los datos granulares en los resultados obtenidos

El proceso de preparación de los datos granulares que incluye, la caracterización de las variables, la limpieza de los datos, la remoción de variables y la transformación de las mismas, es un proceso que afecta directamente la precisión de los algoritmos y el rendimiento de los mismos, pues estos algoritmos son muy sensibles a la cantidad de variables, y son más eficientes con datos numéricos.

En la categorización y agrupamiento del delito, fue clave el transformar la ubicación del hecho en una distancia en metros como se presentó en el capítulo anterior, esto permitió a los algoritmos de agrupamiento calcular de manera eficiente la distancia euclidiana de cada uno, evitando así errores en el agrupamiento. Como se muestra en la Tabla VI “Transformación de las variables longitud y latitud del hecho”, del capítulo anterior.

Es de resaltar que la transformación de la variable categórica tipo de hecho, como se muestra en Tabla V “Transformación de la variable tipo de hecho”, del capítulo anterior, y así con las demás variables categóricas, es necesaria para la correcta funcionalidad de los algoritmos de agrupamiento Kmeans, Clustering jerárquico y Fuzzy C, que necesitan trabajar con conjuntos de datos numéricos y que como buenas prácticas se recomienda transformar estos valores en datos discretos.

Según los expertos en criminología es bien sabido que el delito a nivel general es muy cambiante y obedece a diferentes situaciones y fenómenos que no pudieron ser abarcados en el marco de este trabajo, por esta razón, se consideró segmentar la información proporcionada únicamente para los años 2017, 2018 y 2019, donde se muestra que los delitos contra la vida y de índole patrimonial, tendieron a disminuir en el transcurso del tiempo, pero muy escasamente, mientras que los delitos de índole sexual se mantuvieron constantes, como se muestra en la Tabla XVII y Figura 38.

Tabla XVII. **Cantidad de delitos cometidos durante los años 2017 al 2019 en el departamento de Guatemala, agrupados por la categoría del delito**

	Delitos			
Años	**Contra la vida**	**Sexuales**	**Patrimoniales**	**Total**
2017	4768	132	5709	10609
2018	3851	144	4746	8741
2019	3200	136	4375	7711
Total	**11819**	**412**	**14830**	**27061**

Fuente: elaboración propia.

Figura 38. **Gráfica de frecuencia de la cantidad de delitos cometidos en los años 2017 al 2019**

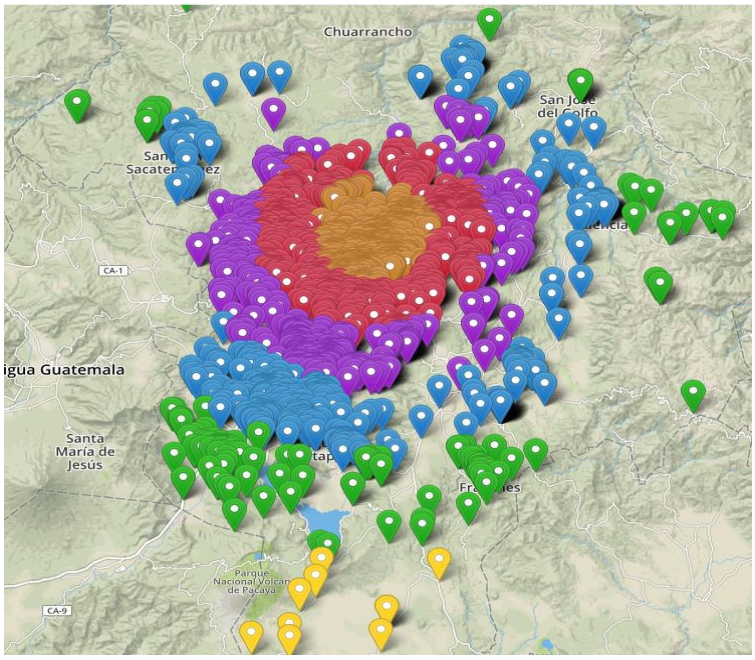
Fuente: elaboración propia, utilizando Microsoft Excel.

Por lo tanto, para la creación de los modelos predictivos, éstos se analizaron de forma separada, pues ambos tienen diferentes patrones de incidencia criminal y es necesario analizarlos y estudiarlos por separado, como, por ejemplo: los delitos de índole patrimonial, se mantienen constantes en los tres años de estudio.

6.2. Categorización del delito según los tipos de hechos seleccionados

El agrupamiento de los hechos delictivos por los experimentos realizados, utilizando las variables de: Ubicación del hecho (latitud y longitud) transformada en su equivalente en metros, el mes en que ocurrió el hecho, el día de la semana y la hora, están en concordancia con la distribución georreferencial del mapa de delitos cometidos durante el año 2019, como se muestra en la figura 39:

Figura 39. **Agrupamiento jerárquico de los delitos contra la vida del año 2019 y su distribución geográfica**



Fuente: elaboración propia utilizando Leatflet.

Los modelos para realizar el agrupamiento, clasificación del delito, realizados mediante agrupamientos jerárquicos, proveen un buen insumo para realizar predicciones sobre los puntos calientes de la incidencia criminal, en el departamento de Guatemala, como se ha visualizado en las figuras anteriores. Estos fueron obtenidos de los hechos históricos ocurridos y representan tendencias y patrones de lo que está ocurriendo en la delincuencia actualmente, como se describe en la sección: 5.2.5 Patrones del delito identificados.

Cada color significa un determinado agrupamiento y aunque están organizados por la cantidad, éstos deben ser analizados en detalle por los expertos en criminología, para reforzar los hallazgos de los datos y sus relaciones, desde el punto de vista social y conductual, mediante los algoritmos utilizados.

6.3. Evaluación del rendimiento y precisión de los algoritmos

Utilizando el índice de Bouldin como métrica de evaluación de los algoritmos de agrupamiento, Bouldin (1979), de manera interna, es decir según los datos que fueron sujetos del estudio, se puede ver que los agrupamientos realizados por Clustering jerárquico fueron más eficientes que los realizados por Kmeans y Fuzzy C, según la Tabla XIII “Precisión de los algoritmos Kmeans, Clustering jerárquico, Fuzzy C para todos los delitos”, del capítulo anterior, dónde se muestra, por ejemplo para los delitos de índole sexual, el agrupamiento jerárquico sobrepaso a Kmeans en siete puntos porcentuales y a Fuzzy C en diez puntos porcentuales, es decir Clustering jerárquico: 72 %, Kmeans 65 % y Fuzzy C 62 %.

La ventaja del agrupamiento jerárquico es que no necesita que se indique el número de agrupamientos, sino que lo hace, recorriendo todos los datos y utilizando el anidamiento de los mismos, de manera que, una vez agrupado un elemento, este no puede pertenecer a otro.

Durante los experimentos se pudo observar que el algoritmo jerárquico, aunque arroja buena diferenciación entre los elementos de cada clúster, utiliza más tiempo de procesamiento que los algoritmos Kmeans y Fuzzy C, éste último, es bastante rápido en comparación con los demás.

Una de las principales desventajas del algoritmo Kmeans y Fuzzy C es que se tuvo que indicar de antemano, el número de grupos (K) y los centroides iniciales, influyendo directamente en el agrupamiento de los datos.

La adecuada selección de las variables que se deben incluir en los modelos del agrupamiento del delito y la transformación de las mismas, garantizan la

buena calidad de los índices de precisión y desempeño de los algoritmos: Kmeans, Clustering jerárquico y Fuzzy C.

Los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, son eficientes en la agrupación del delito y las diferencias porcentuales en rendimiento son muy pocas al compararse entre sí, sin embargo, en los experimentos realizados, se logró determinar que el algoritmo Clustering jerárquico es el más preciso para llevar a cabo esta tarea, aunque no se deben desechar los algoritmos Kmeans y Fuzzy C, por lo que éstos pueden ser utilizando como alternativas para otros modelos de clasificación y validación del agrupamiento jerárquico.

Los algoritmos Kmeans, Clustering jerárquico y Fuzzy C, no son predictivos, permiten realizar análisis descriptivo, por lo que fueron utilizados para determinar las divisiones de las áreas de incidencia criminal o puntos calientes (Hot-Spot), que sirven de insumo para la predicción por medio del algoritmo red de Bayes.

Otro aspecto importante que se debe considerar, es la interpretación criminológica de los agrupamientos realizados por los algoritmos, pues, éstos deben ser considerados como una herramienta de minería de datos, que demuestran tendencias y clasificaciones de los datos, que, de forma manual, no podrían ser detectados, pero que no pueden sustituir al analista criminal, que conoce de antemano, otros factores, otras variables y otros enfoques de lo que está sucediendo con el crimen.

En el caso del algoritmo Naive Bayes (red de Bayes), utilizado para realizar la predicción de puntos calientes, zonas de incidencia criminal, o caracterización y agrupamiento del delito, se demuestra por los experimentos realizados de un promedio de 96 % de efectividad en la clasificación y predicción, según se mostró en la Tabla XVI “Precisión del algoritmo Gaussian Naive Bayes para la predicción

del delito", del capítulo anterior". Esto es debido, al cálculo probabilístico condicional de las variables del delito, que utiliza el algoritmo, por lo que su mayor ventaja radica en el buen detalle que existe de los datos históricos utilizados en el estudio y las tendencias de criminalidad en las áreas geográficas del departamento de Guatemala, así como también el uso de multi clase en la variable "Clúster" del conjunto de datos que se utilizó como insumo.

6.4. Reducción del tiempo de procesamiento de la información para la predicción de delitos criminales

Durante las visitas para la obtención de la información y conocer el contexto de la prevención del delito que actualmente realiza la PNC, se realizó una visita a la Unidad de Análisis de Fenómenos Criminales de la División de Operaciones conjuntas de la PNC, unidad que se encarga de analizar los hechos delictivos y encontrar patrones y fenómenos criminales, para elaborar mapas de prevención que le trasladan a la Subdirección de Operaciones y a la Dirección General, para que se organicen las estrategias de prevención del delito. Esta unidad cuenta con un personal aproximado de ocho analistas.

En la actualidad, le toma a esta unidad alrededor de quince días encontrar patrones de delincuencia y categorización de incidencia criminal, puesto que no cuentan con las herramientas necesarias y se han limitado a utilizar archivos de Excel para efectuar cálculos estadísticos básicos. De igual manera los mapas de incidencia los realizan de forma manual.

Si la unidad implementara el sistema propuesto en el presente trabajo y con una capacitación presencial y contextualizando a los integrantes de esta unidad en temas de análisis de datos, agrupamiento y clasificación, se reduce

notablemente el tiempo de procesamiento para el análisis y predicción del delito, aproximadamente un 80 %.

6.5. Limitaciones y debilidades de los modelos y algoritmos predictivos

Los patrones delictivos no pueden ser estáticos, ya que los mismos cambian con el tiempo. Esta es una desventaja, cuando el sistema no se realimenta de los delitos que ocurren en un determinado momento en el tiempo, lo cual puede estar reflejando una historia pasada y no la realidad del momento. Un sistema transaccional en línea de denuncias que alimente los conjuntos de datos y un sistema de investigación criminal, marcarían una diferencia notable en este punto.

Otra limitante que se encontró es que actualmente se cuenta únicamente con datos generales sobre los delitos cometidos, pero no se encuentran datos específicos referentes a las víctimas, a los sospechosos y a los sindicados, por lo que no se pueden tener modelos y precisiones más acertadas de las tendencias y patrones delincuenciales.

Aspectos importantes y variables de índole social, comunitario y económico, no fueron agregados a los modelos de predicción, siendo éstos muy a menudo los que influyen y determinan los patrones de conducta de la delincuencia en Guatemala, por ejemplo: La influencia de grupos criminales en el lugar, las cercanías de los centros de detención y de condena, los niveles de pobreza de la comunidad, la escasa vigilancia de algunas áreas geográficas, entre otros.

Estas limitantes deben ser analizadas y consideradas con las autoridades policiales para recolectar más información importante que conduzca a mejorar la prevención del delito de manera general y oportuna.

CONCLUSIONES

1. El desarrollo del prototipo y las pruebas realizadas con los datos históricos de los delitos cometidos en los años 2017 al 2019 en el departamento de Guatemala, demuestran que es posible implementar un sistema para realizar el proceso de análisis de criminalidad de manera optimizada en el departamento de Guatemala, el cual está conformado por:

- Un proceso adecuado de selección y tratamiento de la información histórica relacionada a los hechos delictivos.

- Un mecanismo de agrupamiento en conglomerados utilizando Kmeans, Clustering jerárquico y Fuzzy C, el cual permite identificar patrones criminales y traducir la información en modelos útiles para la predicción por medio de Gaussian Naive Bayes.

- Una interfaz de visualización georreferencial que permite al analista del crimen tener un panorama más amplió sobre el delito en las áreas geográficas que son objeto de estudio.

2. Se implementó un proceso de preparación de los datos granulares de criminalidad que garantiza la calidad de los datos que sirven de insumo para la predicción del delito y aumenta la precisión de los algoritmos de agrupamiento y clasificación, el cual está compuesto por los siguientes pasos:

- Selección adecuada de las fuentes de información.

- Tratamiento de las variables por medio de la limpieza, transformación y estandarización de tipos y medidas.

- Reducción de las dimensiones y la proyección de la información utilizada

3. La construcción y entrenamiento de los modelos de clasificación y predicción fue exitosa, dando como resultado una buena calidad en los índices de precisión y desempeño de los algoritmos: Kmeans: 65 %, Clustering jerárquico: 72 % y Fuzzy C: 62 %; sin embargo, se evidencia mayor precisión utilizando Clustering jerárquico, aunque las diferencias porcentuales son muy pocas. El algoritmo Gaussian Naive Bayes predice con un promedio de 96 % de precisión los puntos calientes de las zonas geográficas con mayor incidencia criminal en el departamento de Guatemala.

4. La arquitectura utilizada en el desarrollo del prototipo permite la reducción del tiempo de procesamiento de la información en un 80 % al integrar los componentes Kmeans, Clustering jerárquico, Fuzzy C y Gaussian Naive Bayes para la realización del análisis criminal. La arquitectura implementa una API de servicios REST que integra los algoritmos de agrupamiento y clasificación que provee Scikit-Learn de Python, esto permite un nivel de desacoplamiento bastante grande con la capa de Front-END y una respuesta a las solicitudes bastante ligera. Esta API Rest fue integrada a un servicio de contenedor de DOCKER-COMPOSE, proveyendo escalabilidad al sistema para un mejor rendimiento en tiempo de procesamiento.

RECOMENDACIONES

1. Construir un sistema de información formal que permita a las autoridades policiales realizar el proceso de análisis criminal de manera optimizada en el departamento de Guatemala, que cumpla con las siguientes características.

- Un proceso adecuado de preparación de los datos granulares que se transforman en insumo para los modelos de agrupamiento y predicción.

- Componentes de la arquitectura que permitan realizar el análisis de los conglomerados e identificar patrones criminales, utilizando algoritmos de agrupamiento Kmeans, Clustering jerárquico, Fuzzy C y Gaussian Naive Bayes para realizar la predicción de puntos calientes.

- Componentes de visualización georreferencial de los agrupamientos y predicciones para la correcta interpretación y análisis criminal.

- Diseño de una arquitectura que permita lograr mayor escalabilidad, la adaptación de nuevas funcionalidades y reduzca el tiempo de procesamiento.

2. Investigar más a fondo el contexto de los patrones criminales que ocurren en el departamento de Guatemala e identificar las variables necesarias para que sean incluidas en los sistemas de registro de la investigación criminal, con el fin de construir modelos de agrupamiento y predicción más precisos que puedan ser integrados al sistema de análisis de criminalidad.

3. Investigar sobre otros modelos de predicción del delito, como: Regresión lineal, redes neuronales, árboles de decisión, entre otros para medir si la precisión de los mismos supera a los algoritmos Kmeans, Clustering jerárquico, Fuzzy C y Gaussian Naive Bayes para que puedan integrarse dentro de la API de servicios y se utilicen en el análisis de criminalidad.

4. Comparar con otras arquitecturas de sistemas de información y verificar si la arquitectura utilizada en el desarrollo de los componentes tecnológicos del prototipo, permite una mayor escalabilidad, mejora el tiempo de procesamiento de la información y provee un mayor desacoplamiento de los componentes de agrupamiento y clasificación del delito.

REFERENCIAS

1. Adderley, R., y Musgrove, P(1999). *Data mining at the West Midlands Police: A Study of bogus official Burglaries*. Londres, Inglaterra: Springer.

2. Ahishakiye, E., y Niyonzima, I (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. *International Journal of Computer and Information Technology, 06*(03), 1-9. Recuperado de: https://www.researchgate.net/publication/316960839_Crime_Prediction_Using_Decision_Tree_J48_Classification_Algorithm/link/591a86760f7e9b1db652acd7/download.

3. Barreras, F., Díaz, C., Riascos, Á., y Riberto, M (2016). *Una comparación de diferentes modelos para la predicción del crimen en Bogotá*. Bogotá, Colombia: Centro de estudios sobre seguridad y drogas, Universidad de los Andes.

4. Baumgartner, K., Ferrari S., y Palermo G (2008). Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems, ELSEVIER*, 1-10. Recuperado de: https://www.academia.edu/6386968/Constructing_Bayesian_networks_for_criminal_profiling_from_limited_data

5. Bouldin, D (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 1-5. Recuperado de: https://www.researchgate.net/profile/Don_Bouldin

/publication/224377470_A_Cluster_Separation_Measure/links/02e7e52d57df1ad121000000/

6. Carneiro, C., y Schmelmer, T (2016). *Microservices from Day One, Build robust and scalable software.* Hollywood, Florida, Estados Unidos de América: Apress.

7. División Especializada en Investigación Criminal de la Policía Nacional Civil de Guatema (2009). *Organización y designación de funciones de la División Especializada en Investigación Criminal de la Policía Nacional Civil.* Recuperado de http://pnc.edu.gt/wp-content/uploads/2013/07/12-2009-DEIC.pdf

8. Doglio, F (2015). *Rest Api Development from Node.js.* New York, Estados Unidos de América: Springer.

9. Gironés, J., Casas, J., Minguillón, J., y Caihuelas, R (2017). *Minería de datos, modelos y algoritmos*. Barcelona, España: UOC.

10. Gonzalez, D (2016). *Developing Microservices with Node.js.* Birmingham, Reino Unido: Packt Publishing.

11. Gorton, I (2011). *Essential Software Architecture.* Richland, Washington, Estados Unidos de América: Springer.

12. Gorunescu, F (2011). *Data mining, conceptos, modelos y técnicas.* Australia: Springer.

13. Jirón, J (2013). *Teoría del delito*. Guatemala, Guatemala: Instituto de la Defensa Pública Penal.

14. Kumar, V., y Chandrasekar, C (2011). Evaluation of Modern Crime Prediction Techniques. *International Journal of Advanced Research in Computer Sciense*, *2*(4), 1-6. Recuperado de: http://www.ijarcs.info/index.php/Ijarcs/article/view/695/683

15. MacQueen, J (1967). *Some Methods for classification and Analysis of Multivariate Observations.* California, Estados Unidos de Ampérica: University of California Press.

16. Mena, J (2003). *Investigative Data Mining for Security and Criminal Detection.* Burlington, Estados Unidos de América: Elsevier Science.

17. Microsoft Corporation (2018). *Algoritmos de minería de datos*. Recuperado de: https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=asallproducts-allversions

18. Newman, S (2015). *Building Microservices.* Sebastopol, California, Estados Unidos de América: O´Reilly Media, Inc.

19. Omkar, V., Sayak, M., Raj, K., Suraj, C., y Rohini, P(2018). Comprehensive comparative analysis of methods for crime rate prediction. *International Research Journal of Engineering and Technology 5*, 1-4. Recuperado de: https://www.irjet.net/archives/V5/i2/

20. Pearl, J (2000). *Causality: Models, Reasoning, and Inference*. Boston, Estados Unidos de América: Cambridge University Press.

21. Perez, C (2004). *Técnicas de análisis multivariante de datos.* Madrid, España: Pearson Education S.A.

22. Pérez, M (2014), *Minería de datos a través de ejemplos*. Madrid, España: Grupo RC.

23. Policía Nacional Civil (2009). *Acuerdo Gubernativo 97-2009, Reglamento sobre la organización de la Policía Nacional Civil.* Recuperado de https://pnc.edu.gt/wp-content/uploads/2013/07/ Reglamento-Organizacion-PNC-97-20091.pdf

24. Revatthy, K., y Satheesh, K (2012). Survey of data mining techniques on crime. *International Journal of Data Mining Techniques and Applications, 1*(2), 1-4. Recuperado de: https://www.researchgate .net/publication/322469559_A_Survey_of_Data_Mining_Techniques_for_Crime_Detection

25. Richards, M (2016). *Microservices vs. Services-Oriented-Architecture.* Sebastopol, California, Estados Unidos de América: O'Reilly Media.

26. Secretaría de la Instancia Coordinadora de la Modernización del Sector Justicia (2014). *Base legal de la Secretaría Ejecutiva de la Instancia Coordinadora de la Modernización del Sector Justicia*. Recuperado de http://seij.gob.gt/base-legal/

27. Sreedevi, M., Vardhan, H., y Krishna, V (2018). Review on crime analysis and prediction using data mining techniques. *International Journal of Innovative* Research *in Science, Engineering and Technology*, *7*(4), 1-10. Recuperado de: http://www.ijirset.com/upload/2018/april/

28. Secretaría Técnica del Consejo Nacional de Seguridad de Guatemala (2017). *Reporte Estadistico Enero 2017.* Recuperado de: https:// stcns.gob.gt/docs/2017/Reportes_DMC*.*

29. Tahani, A., Rsha, M., y Lor, E (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *International Journal of Data Mining & Knowledge Managment Process*, *5*(4),1-20. Recuperado de: https://arxiv.org/ftp/arxiv/papers/1508.

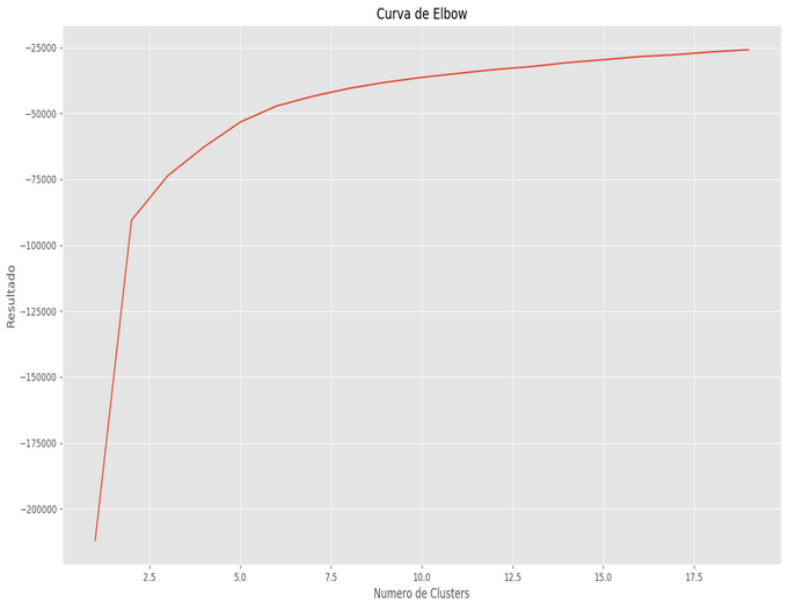
30. The Math Forum at NCTM (2020). *The Math Forum People Learning Math Together*. Recuperado de http://mathforum.org/library/drmath /view/51879.html

31. Witold, P(2005). *Knowledge-based C, From Data to Information Granules.* New Jersey, Estados Unidos de América: John Wiley & Sons, Inc., Publication.

32. Yang, M (1993). A survey of Fuzzy Clustering. *Departament of Mathematics, Chung Yuan Christian University*, *18*(11), 1-16. Recuperado de: https://www.sciencedirect.com/science/article/pii /089571779390202A

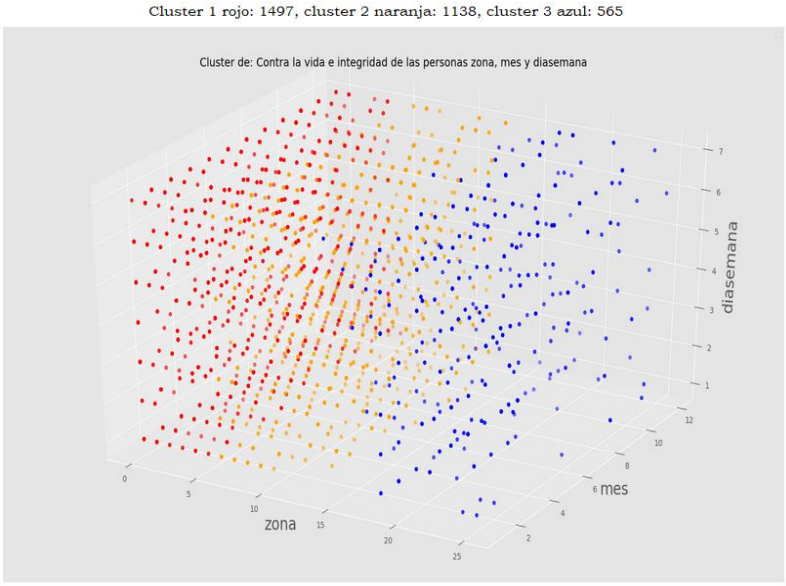
APÉNDICES

Apéndice 1. **Gráfica de codo para obtener el valor K=3**



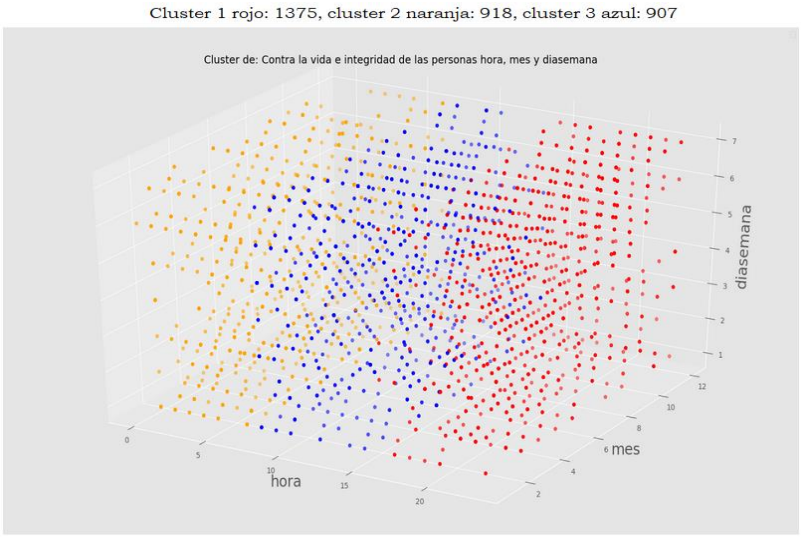
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 2. **Agrupamiento Kmeans de delitos contra la vida del año 2019, por zona, mes, día de la semana del hecho K=3**



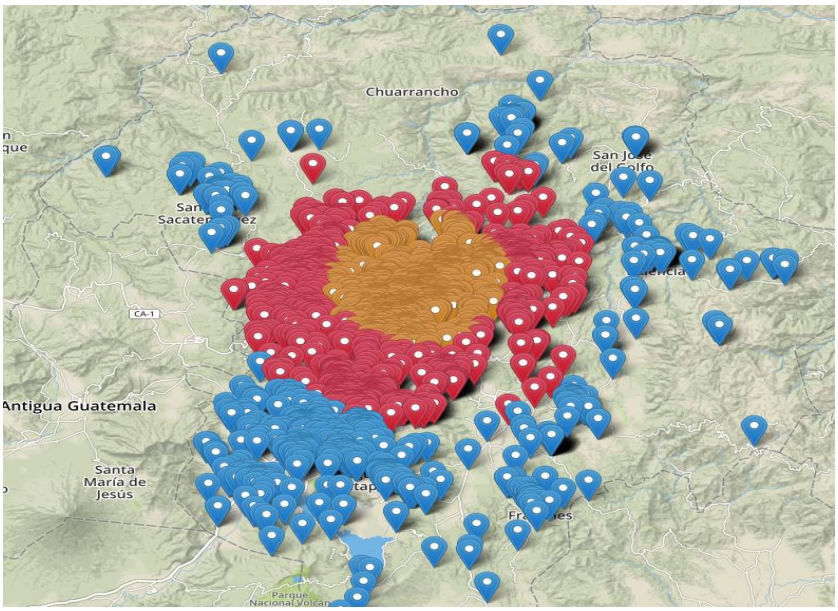
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 3. **Agrupamiento Kmeans de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho K=3**



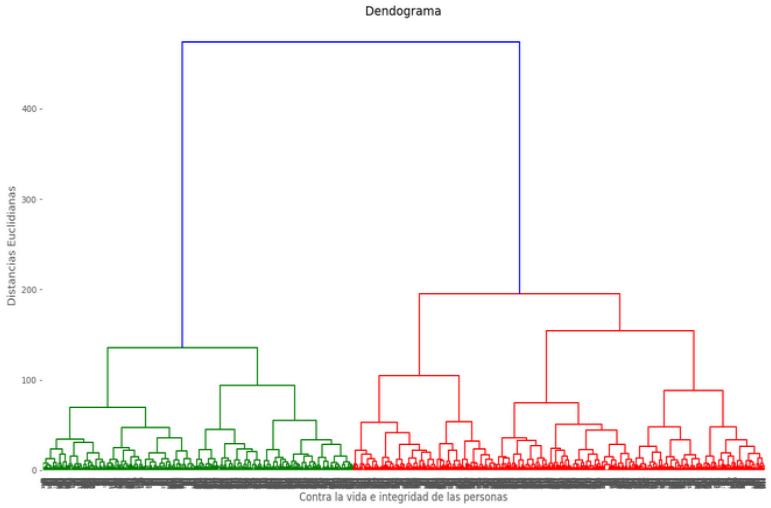
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 4. **Visualización georreferencial Kmeans de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho**



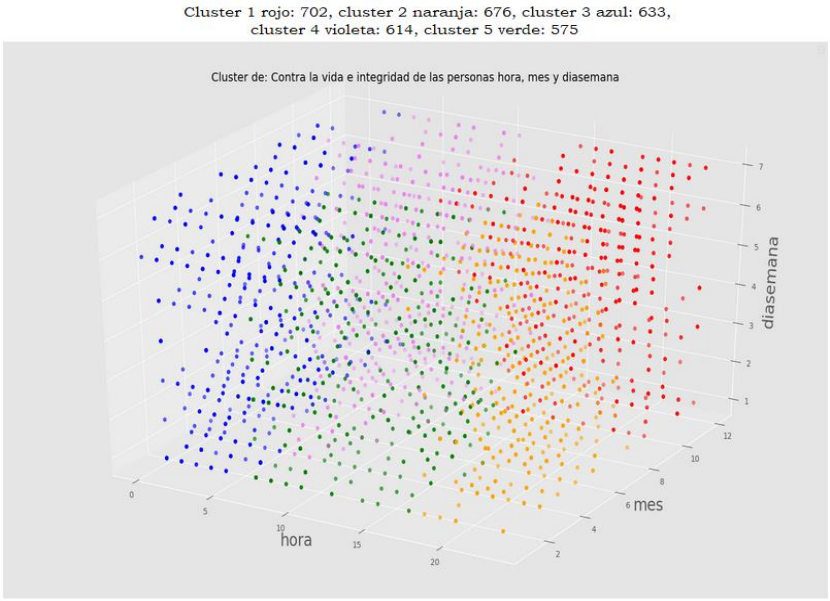
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 5. **Dendograma para obtener el valor K=5**



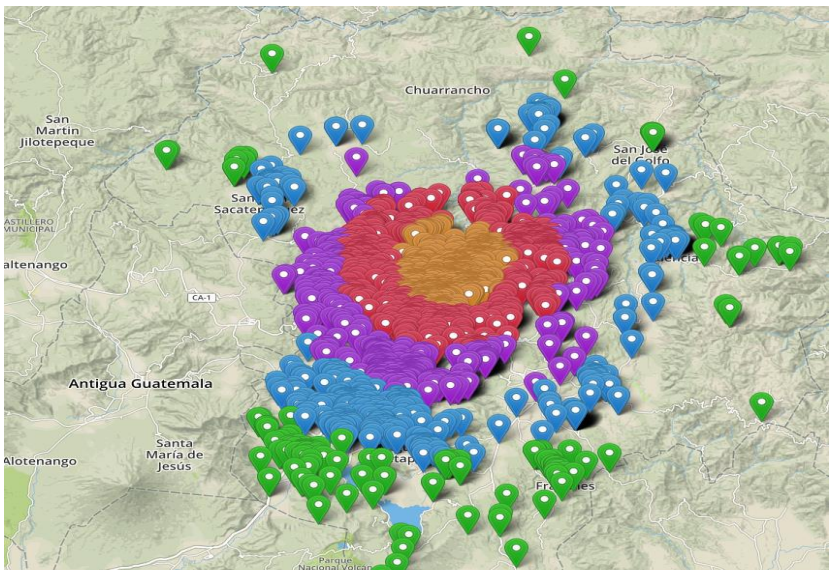
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 6. **Agrupamiento jerárquico de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho**



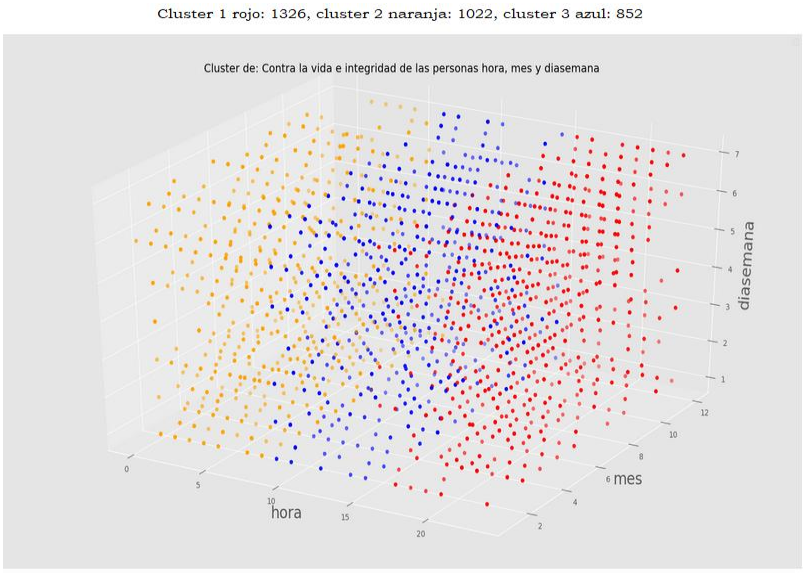
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 7. **Visualización georreferencial agrupamiento jerárquico de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho**



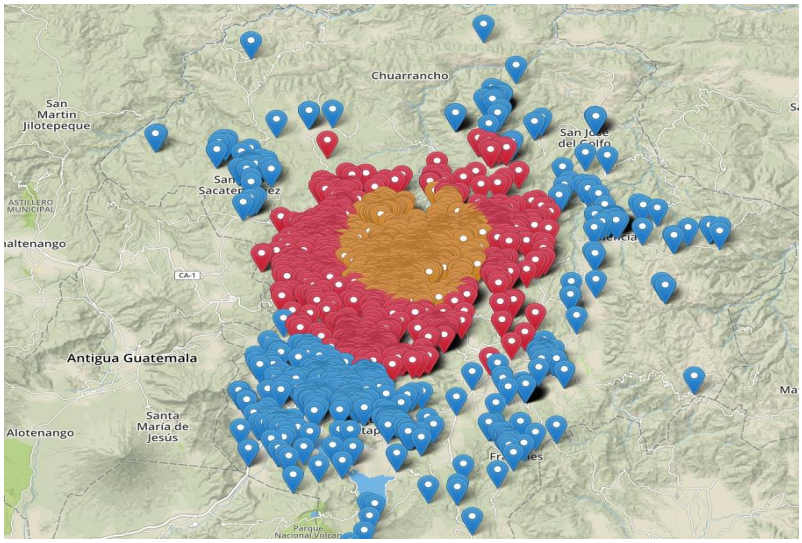
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 8. **Agrupamiento Fuzzy C de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho K=3**



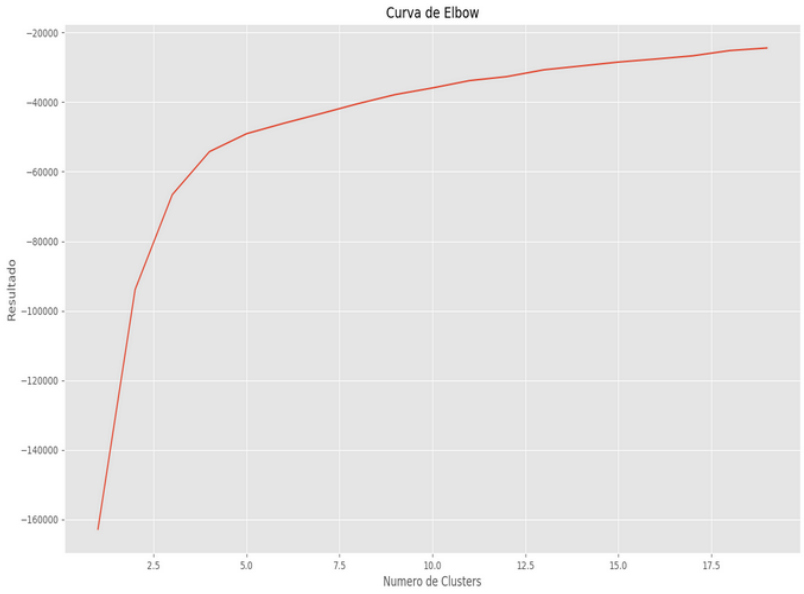
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 9. **Visualización georreferencial Fuzzy C Means de delitos contra la vida del año 2019, por hora, mes, día de la semana del hecho**



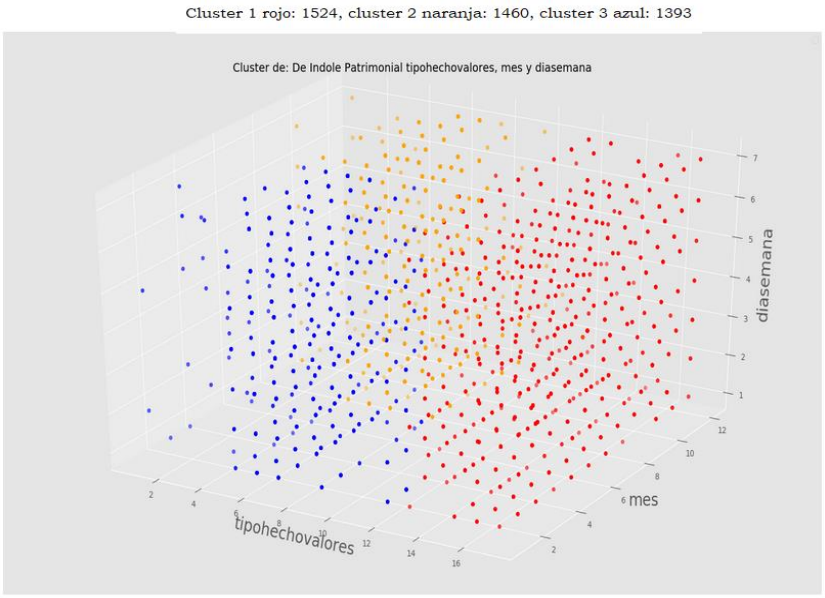
Fuente: elaboración propia utilizando Leatflet.

Apéndice 10. **Gráfica de codo para obtener el valor K=3**



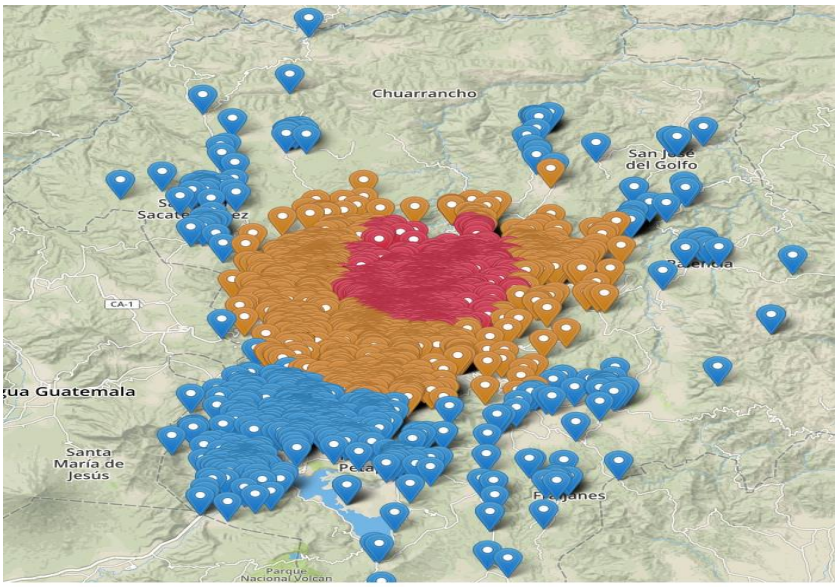
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 11. **Agrupamiento Kmeans de delitos patrimoniales del año 2019, por tipo de hecho, mes y día de la semana**



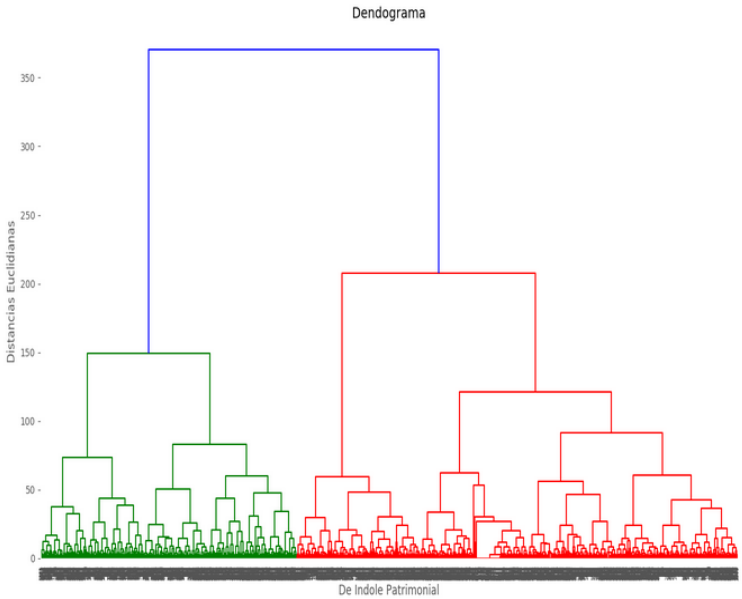
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 12. **Visualización georreferencial Kmeans de patrimoniales del año 2019, por zona, mes, día de la semana del hecho**



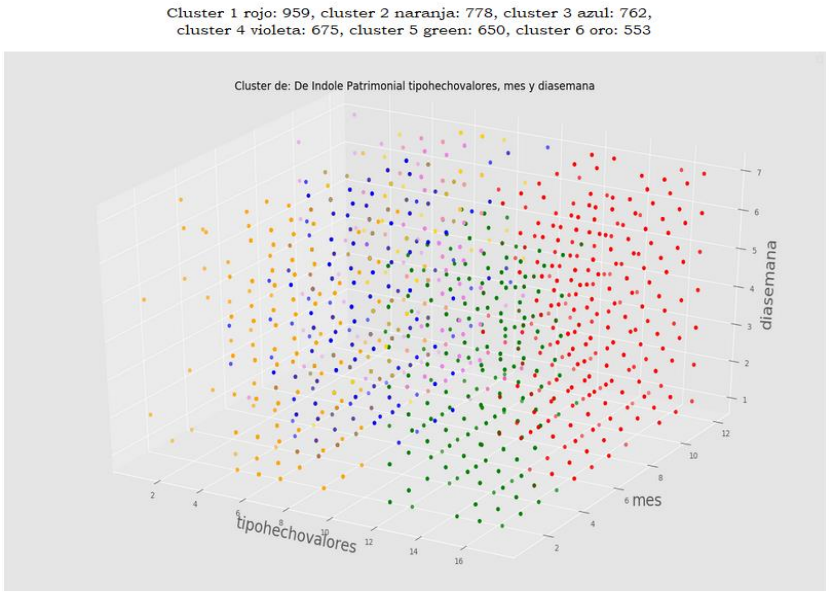
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 13. **Dendograma jerárquico K=6**



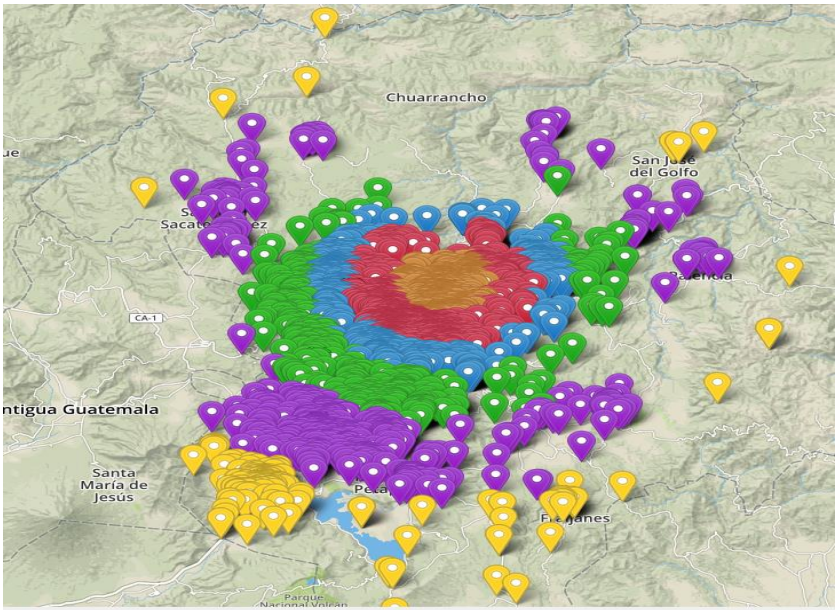
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 14. **Agrupamiento jerárquico de delitos patrimoniales del año 2019, por tipo de hecho, mes y día de la semana**



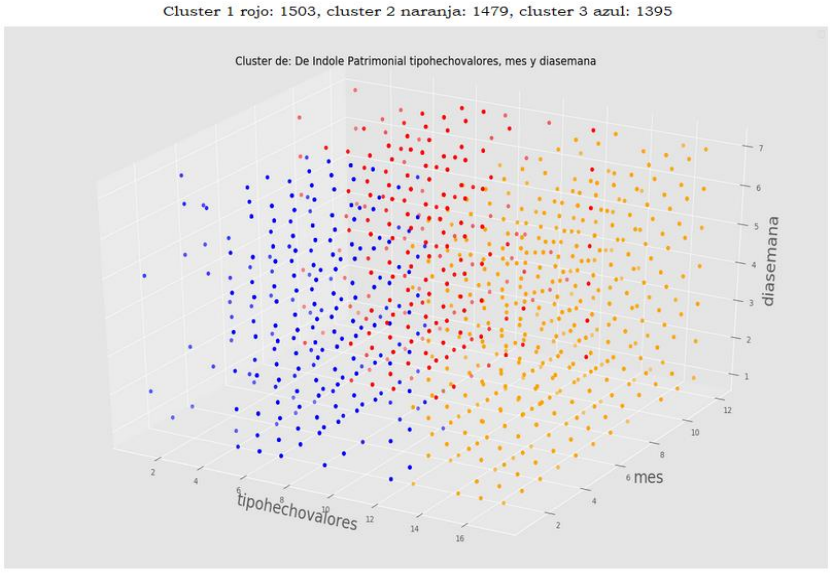
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 15. **Visualización georreferencial clustering jerárquico de delitos patrimoniales del año 2019, por zona, mes, día de la semana del hecho**



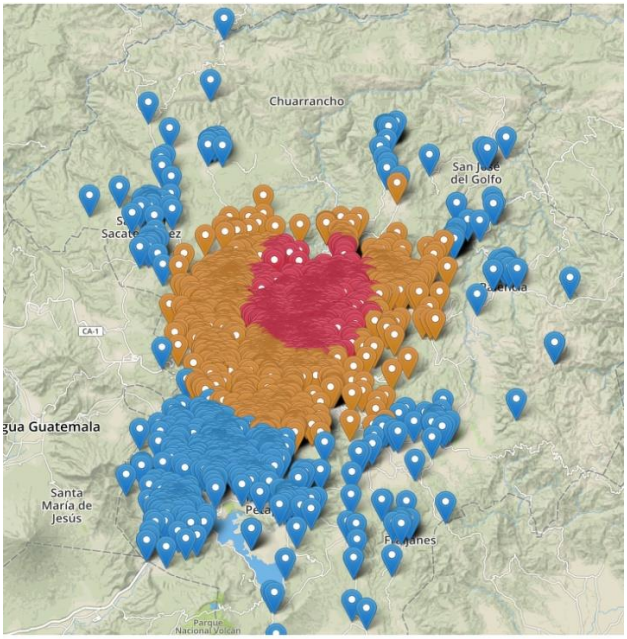
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 16. **Agrupamiento Fuzzy C de delitos patrimoniales del año 2019, por zona, mes, día de la semana del hecho K=3**



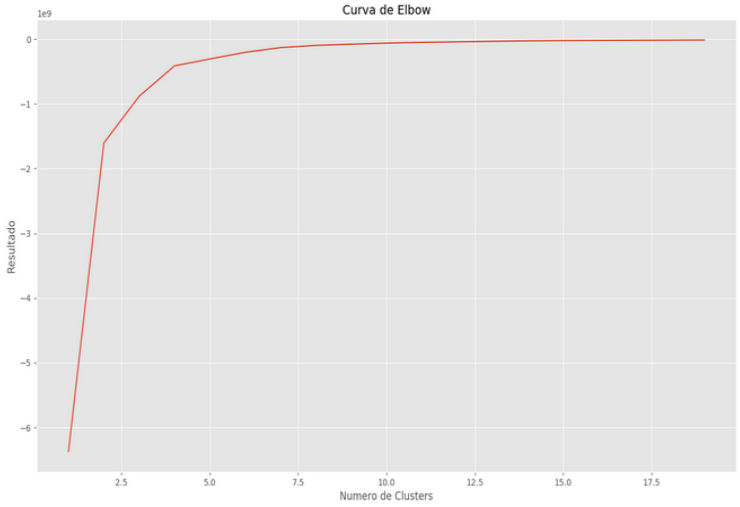
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 17. **Visualización georreferencial Fuzzy C de delitos patrimoniales del año 2019, por zona, mes, día de la semana del hecho**



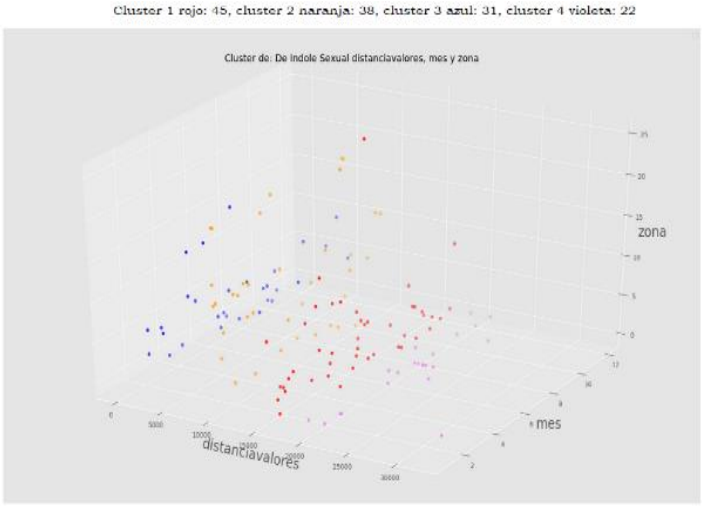
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 18. **Gráfica de codo K=4**



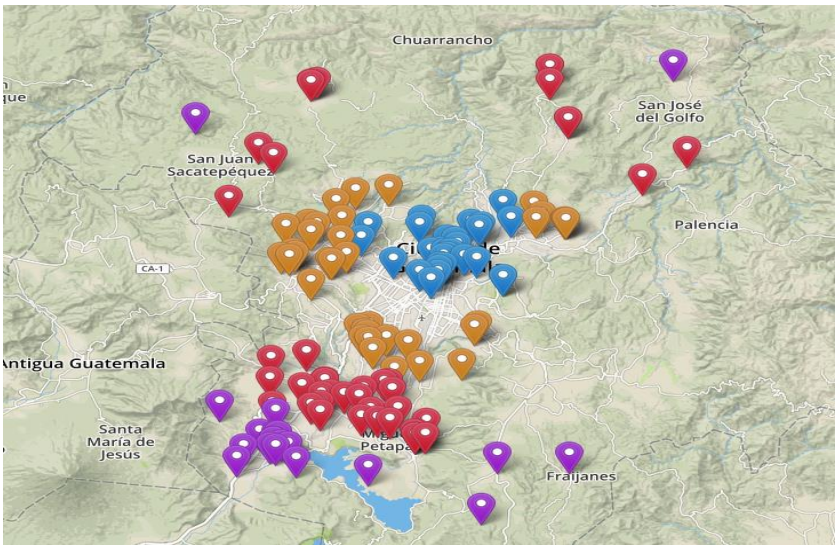
Fuente: elaboración propia utilizando Matplot-Lib.

Apéndice 19. **Agrupamiento Kmeans de delitos sexuales 2019 por ubicación del hecho, mes, zona**



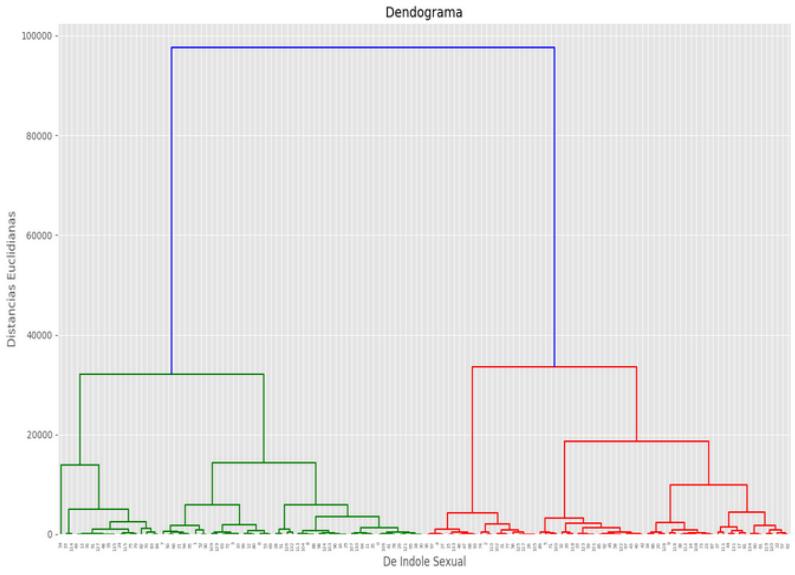
Fuente: elaboración propia, utilizando Matplot-Lib

Apéndice 20. **Visualización georreferencial Kmeans de delitos sexuales 2019 por ubicación del hecho, mes, zona**



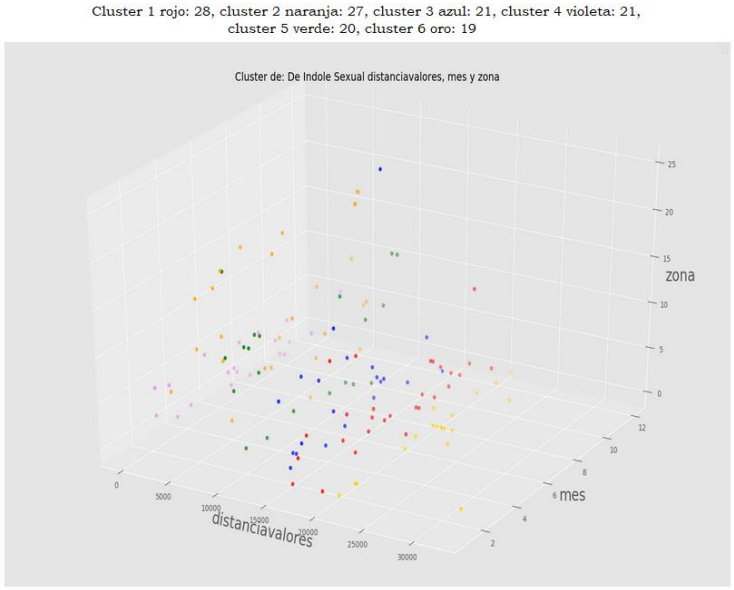
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 21. **Dendograma K= 5**



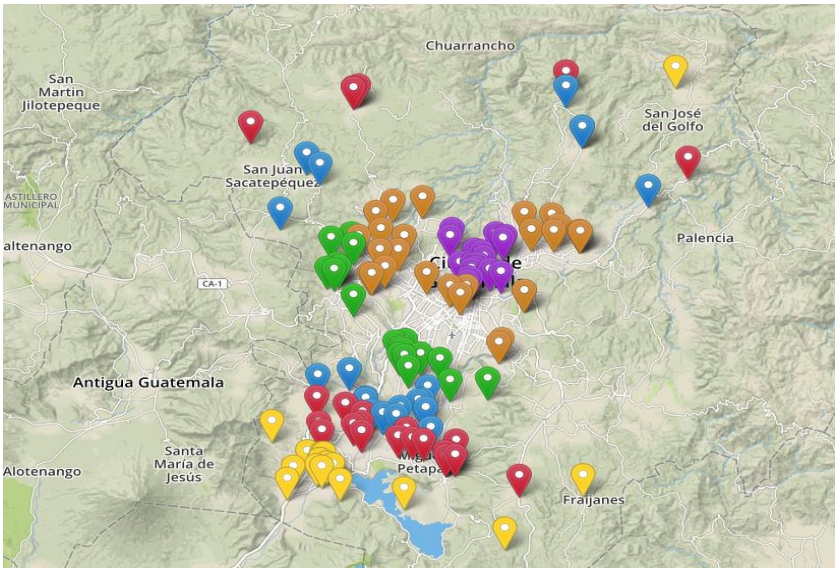
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 22. **Agrupamiento jerárquico de delitos sexuales 2019 por ubicación del hecho, mes, zona**



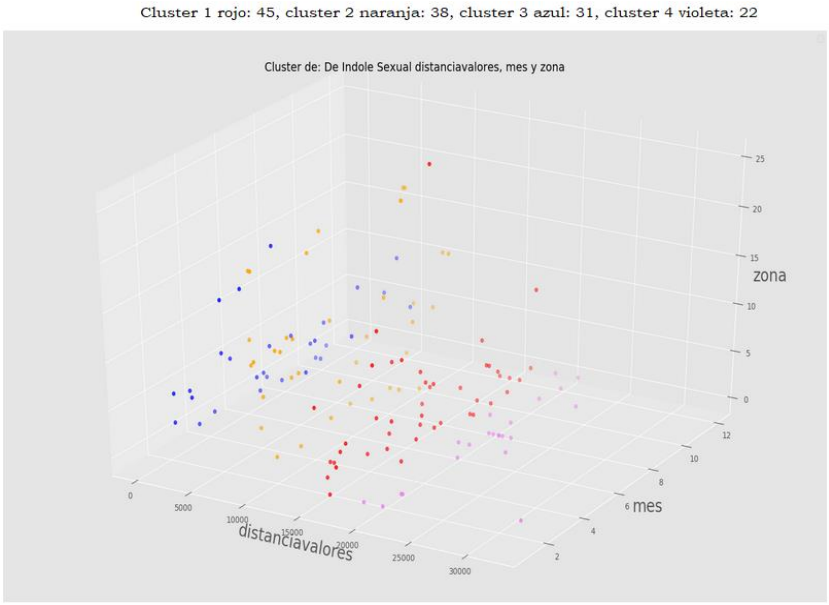
Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 23. **Visualización georreferencial Agrupamiento jerárquico de delitos sexuales 2019 por ubicación del hecho, mes, zona**



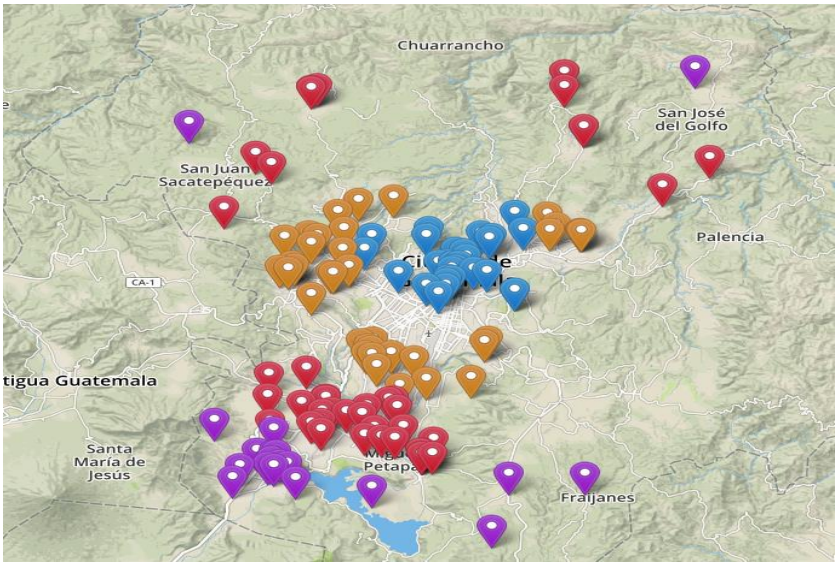
Fuente: elaboración propia, utilizando Leatflet.

Apéndice 24. **Agrupamiento Fuzzy C de delitos sexuales 2019 por ubicación del hecho, mes, zona K = 4**



Fuente: elaboración propia, utilizando Matplot-Lib.

Apéndice 25. **Visualización georreferencial Fuzzy C de delitos sexuales 2019 por ubicación del hecho, mes, zona**



Fuente: elaboración propia, utilizando Leatflet.